FACULTAD DE INGENIERÍA Y CIENCIAS APLICADAS

IMPLEMENTACIÓN DE UNA INSTANCIA DE BIG DATA CON HADOOP PARA LA ADQUISICIÓN DE INFORMACIÓN DE CONSUMO ENERGÉTICO DE UN DATA CENTER EXPERIMENTAL

Trabajo de titulación presentado en conformidad con los requisitos establecidos para optar por el título de Ingeniero en redes y telecomunicaciones

Profesor Guía

Mg. Iván Patricio Ortiz Garcés

Autor

Carlos Iván Chicaiza Tamayo

Año

2019

DECLARACIÓN DEL PROFESOR GUÍA

“Declaro haber dirigido este trabajo, Implementación de una instancia de Big Data con Hadoop para la adquisición de información de consumo energético de un data center experimental, a través de reuniones periódicas con el estudiante, Carlos Iván Chicaiza Tamayo, en el semestre 201920, orientando sus conocimientos y competencias para un eficiente desarrollo del tema escogido y dando cumplimiento a todas las disposiciones vigentes que regulan los trabajos de titulación”.

Iván Patricio Ortiz Garcés

Magister en Redes de Comunicaciones

CI: 0602356776

DECLARACIÓN DEL PROFESOR CORRECTOR

“Declaro haber revisado este trabajo, Implementación de una instancia de Big Data con Hadoop para la adquisición de información de consumo energético de un data center experimental, de Carlos Iván Chicaiza Tamayo, en el semestre 201920, dando cumplimiento a todas las disposiciones vigentes que regulan los trabajos de titulación”.

William Eduardo Villegas
Magister en Redes de Comunicaciones
CI: 1715338263

DECLARACIÓN DE AUTORÍA DEL ESTUDIANTE

“Declaro que este trabajo es original, de mi autoría, que se han citado las fuentes correspondientes y que en su ejecución se respetaron las disposiciones legales que protegen los derechos de autor vigentes”.

_____

Carlos Iván Chicaiza Tamayo

CI: 1720778735

AGRADECIMIENTOS

Agradezco principalmente a Dios por estar conmigo en cada paso que doy y darme las fuerzas para seguir adelante, a mi familia por estar ahí desde el día uno, por brindarme su apoyo y aliento en todos estos años de estudio, a mis amigos por esos momentos de distracción, a mis maestros por todo el conocimiento transmitido, a todos muchas gracias por el apoyo, y este logro es gracias a muchos de ustedes.

DEDICATORIA

A mi esfuerzo y perseverancia en todos estos años de estudio, en donde mantuvimos una lucha interminable por superarnos día a día, a mi madre por enseñarme que la base es el trabajo y la constancia por su apoyo transparente y puro que me alentó a seguir adelante en mis días más complejos y tristes, a mis hermanos que pese a la distancia siempre estuvieron dispuestos ayudarme de cualquier manera y a los que ya no están, pero siguen presentes en mí.

RESUMEN

Actualmente, la información generada día a día mediante las conexiones entre dispositivos que se interconectan está aumentando el volumen de generación de datos. Por lo cual, la información está disponible pero no cuenta con un valor que permita realizar la toma de decisiones en momentos precisos dentro de las organizaciones. Para todo lo previsto en la toma de decisiones con información disponible desde cualquier fuente, nace el término de Big Data que ayuda a identificar valor en un conjunto de datos estructurados o no estructurados.

En el transcurso de este proyecto de titulación, se realizó una recopilación de información de los equipos disponibles del centro de datos experimental de la UDLA, para conocer las especificaciones técnicas provenientes del consumo eléctrico que ayuden a realizar una comparación del uso actual al especificado en su ficha técnica.

Continuando con el proyecto de tesis se abarcó temas de la tecnología Hadoop como base para implementar una solución de Big Data, se detalló sus módulos, funcionamiento y ventajas de esta tecnología para el procesamiento de datos.

Como parte central del proyecto se realizó la implementación de un nodo maestro y varios nodos esclavos que en conjunto formaron la instancia de Big Data. La misma que está procesando información adquirida mediante el monitoreo de equipos.

Finalmente, para concluir con el proyecto de tesis se generó un monitoreo de los equipos disponibles de la red y se realizó un análisis con la información adquirida en base al consumo eléctrico de los equipos disponibles en el centro de datos. Todo con el fin de poder encontrar información que identifique los consumos excesivos realizado por los equipos que se encuentran en el centro de datos experimental.

ABSTRACT

Currently, the information generated day by day through connections between devices that are interconnected, is increasing the volume of data generation. Therefore, the information is available but does not have a value that allows making decisions at precise times within organizations. For everything planned in the decision-making process with information available from any source, the Big Data term is born that helps to identify value in a set of structured or unstructured data.

In the course of this titling project, a compilation of information on the equipment available from the experimental data center of the UDLA was carried out, in order to know the technical specifications coming from the electrical consumption that help to make a comparison of the current use to the one specified in its file technique.

Continuing with the thesis project, topics of Hadoop technology were covered as a basis for implementing a Big Data solution, detailing its modules, operation and advantages of this technology for data processing.

As a central part of the project, the implementation of a master node and several slave nodes that together formed the Big Data instance was carried out. The same one that is processing information acquired by monitoring equipment.

Finally, to conclude the thesis project, a monitoring of the available equipment of the network was generated and an analysis was made with the information acquired based on the electrical consumption of the equipment available in the data center. All in order to be able to find information that identifies the excessive consumption made by the teams that are in the experimental data center.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. CAPÍTULO I. INTRODUCCIÓN

1.1. Antecedentes

La información generada día a día ha llevado a la búsqueda de mejorar las herramientas de análisis de información, un estudio de IBM en el 2012 muestra que se generan 2.5 billones de gigabytes de datos por día, las interacciones digitales y la conectividad entre dispositivos, han hecho que el conjunto de aplicaciones y dispositivos estén generando volúmenes de información grandes y complejos, por lo que se necesite de tecnologías de procesamiento de datos específicos como Hadoop, el cuál permita procesar volúmenes de información de manera óptima e interprete la información de forma eficiente y comprensible.

En estos días, el poder de la información que tiene una empresa puede incrementarse debido a su fiabilidad, volumen y accesibilidad que la empresa pueda darle a la información en tiempo real. Big Data surge para ayudar a procesar y analizar grandes volúmenes de datos los cuales permitan descubrir patrones y otros aspectos fundamentales para la toma de decisiones.

Hadoop surge como una opción para resolver los problemas asociados a Big Data, sus componentes HDFS (Hadoop Distributed File System) que es un sistema de archivos distribuido y MapReduce que trabaja sobre grandes colecciones de datos en grupos de computadores, han hecho posible el tratamiento de datos distribuidos en clúster de computadoras posibilitando la eficiencia en el procesamiento de datos.

Hadoop un proyecto open source desarrollado por Google, ha transformado la manera de adoptar Big Data, permitiendo el uso de arquitecturas que disminuyan el costo que normalmente se tendría en un sistema de analítica de datos, proporcionando gran escalabilidad y manejo de infraestructura reducida para proyectos de recolección y procesamiento de información.

La tecnología avanza a ritmos insostenibles y cada día incrementa el uso de aplicaciones y dispositivos, los cuales son consumidos de manera excesiva aumentando de forma exponencial el consumo de energía, esto nos lleva a concientizar y abrir una brecha para contribuir con la sostenibilidad del planeta, mediante la incorporación de conceptos de ahorro energético, que de acuerdo al

termino global es el uso eficiente de energía, que busca proteger el medio ambiente, mediante la reducción de intensidad energética y habituando al usuario a consumir los necesario y nomas.

1.2. Alcance

Se implementará una instancia de Big Data con Hadoop para la recopilación de datos referente al consumo de energía eléctrica de los equipos activos del centro de datos experimental, con el fin de tener un ambiente de procesamiento de información que permita el análisis de los datos.

Se realizará una recolección de datos sobre el consumo de energía de los diferentes equipos de red correspondientes al centro de datos experimental.

Se analizará las ventajas de utilizar Hadoop como una tecnología que permita implementar una solución de Big Data robusta que cumpla con las necesidades del proyecto.

Finalmente, se generará un monitoreo con la información del consumo energético de los equipos del centro de datos experimental, con el fin de conocer y mejorar el consumo de energía.

1.3. Justificación

Big Data se ha posicionado como una solución a los problemas que se tiene al momento de realizar análisis con grandes volúmenes de información, por ende contar con una tecnología de procesamiento de información como Hadoop permitirá recopilar datos desde cualquier fuente, como puede ser el consumo eléctrico en equipos de red y así poder descubrir patrones de consumo excesivos lo cual no permita llevar un eficiente dispendio eléctrico en una época en donde toda arquitectura debe ser amigable con el medio ambiente.

Como estudiantes de una carrera tecnológica conocemos de la importancia de los datos generados en un centro de datos y de la problemática que se tiene con la cantidad de información que es generada día a día siendo imposible tratar este tipo de información con herramientas de ofimática actuales debido al volumen y complejidad de los datos.

En el transcurso del tiempo y para cubrir las necesidades se han desarrollado herramientas de Big Data de pago y algunas muy costosas, que permiten el

procesamiento de información en centro de datos, con esto se ha podido conocer el comportamiento del tráfico de la red y el consumo excesivo de energía en los diferentes equipos, sin embargo la necesidad sigue latente por contar con soluciones y herramientas open source, que ayuden al procesamiento de datos, es por todo lo mencionado que el desarrollo del tema planteado con tecnología Hadoop puede ayudar a descubrir la causa y patrones de consumo eléctrico excesivos y dejar como base una instancia para el procesamiento de datos.

1.4. Objetivo General

Implementar una instancia de Big Data con Hadoop para recopilar información del consumo eléctrico de los equipos del centro de datos experimental, con el fin de analizar los datos que ayuden a mejorar el consumo energético.

1.5. Objetivos Específicos

1. Evaluar la información de consumo eléctrico de los equipos de red configurados en el centro de datos experimental.
2. Analizar las ventajas de utilizar Hadoop como tecnología de recopilación de datos sobre otras herramientas de Big Data.
3. Implementar una instancia de Big Data en el centro de datos experimental para el procesamiento de datos.
4. Generar un monitoreo con la información del consumo eléctrico de los equipos del centro de datos experimental.

1.6. Metodología

Para desarrollar el proyecto de titulación, se utilizó los siguientes métodos: Método Descriptivo, deductivo y experimental.

Mediante el método descriptivo se realizó una descripción del consumo eléctrico de equipos del centro de datos experimental.

Con el método deductivo se recopilo la información de los equipos de red y se pasó almacenar la información en el ambiente de Big Data implementado.

Finalmente, se trabajó con el método experimental que permitió realizar el análisis y las pruebas con la información recopilada a fin de favorecer el consumo adecuado de energía en equipos de red del centro de datos experimental.

2. CAPÍTULO II. MARCO TEÓRICO

2.1. Introducción

Los datos generados por diferentes dispositivos están cambiando la manera de cómo las empresas analizan la información, el volumen, variedad y la cantidad de datos tanto estructurados como no estructurados, los cuales provienen de redes sociales, celulares, sensores, datos científicos entre otros, están revolucionando la forma de como las industrias acceden y analizan la información.

El concepto de Big Data como tendencia surge, cuando la industria se percata que no puede almacenar ni manejar la información de manera convencional, empieza adoptar nuevas herramientas que dan inicio al concepto del Big Data que puede ser adoptada en cualquier industria y supone entender de manera dinámica las fuentes de datos internas y externas.

El conjunto de datos de los cuales podemos disponer y la interacción que se puede realizar entre ellos, permiten que los datos tomen valor sea este económico o científico aportando al entendimiento de problemas en diferentes sectores como en el aeronáutico en el cual mayor ventaja se tiene, seguido por la banca, seguros, el sector médico o el agrícola, con esto se debe entender que el Big Data forma parte de un proceso de negocio, que el uso masivo de los datos ofrecen nuevas posibilidades de orientar los modelos de negocio de acuerdo a cada industria.

2.2. Big Data

En el portal Gartner, el concepto de Big Data se atribuye a sistemas de información que manejan conjunto de datos de gran volumen, procesos de alta velocidad, de veracidad, de valor y gran variedad de recursos que demanda formas de procesamiento de información innovadoras y rentables que permiten una visión mejorada de la información y atribuye a la mejora en la toma de decisiones y automatización de procesos permitiendo una compresión eficiente. (Glossary Gartner IT, 2019)

Big Data es la solución al crecimiento exponencial de los datos, en el momento en que se hace difícil su administración con respecto al almacenamiento, procesamiento y acceso.

“Optimizar el cálculo y la precisión algorítmica para reunir, analizar, enlazar y comparar conjuntos de grandes datos”. (Christof, 2019)

“Identificar patrones para la toma de decisiones en los ámbitos económico, social, técnico y legal” (Forrester, 2019)

De acuerdo con la mayoría de las definiciones que se encuentran sobre el Big Data, estas visualizan su enfoque al volumen de los datos, de esto se puede expresar que el volumen importa pero que también existen otros atributos importantes de Big Data que son la velocidad, la variedad, la veracidad de las diferentes fuentes de datos y el valor, este último como enfoque central para conocer la calidad de información que se podría obtener en una solución con Big Data.

Estos cinco aspectos constituyen una definición comprensiva y además destruyen el mito acerca de que Big Data se trata únicamente del volumen.

En la tabla 1 se detalla los aspectos mencionados y se le atribuyen características a cada término:

Tabla 1.

Atributos de Big Data.

Volumen	Velocidad	Variedad	Veracidad	Valor
Almacenamiento en terabytes	Por lotes	Estructurada	Integridad y autenticidad	Estadística
Registros	Tiempo corto	No estructurado	Origen y reputación	Eventos
Tablas y archivos	Procesos	Probabilística	Responsabilidad	Hipótesis

2.2.1. Las cinco V de Big Data

Como se visualizó en la tabla 1 es común que al hablar de Big Data se haga referencia a grandes cantidades de datos, pero esta filosofía es más que eso, para describir mejor lo que representa debemos decir que Big Data es un entorno que engloba a las cinco V, IBM como pioneros en el desarrollo y soluciones de Big Data fue uno de las compañías que empezó definiendo tres V y en el transcurso de estos años se han añadido las otras dos dependiendo de la fuente y la utilidad que se le otorga a cada solución de Big Data. A continuación, se detalla el concepto de las 5 V de Big Data.

2.2.1.1. Volumen

Un sistema de Big Data es capaz de almacenar una gran cantidad de datos mediante infraestructuras escalables y distribuidas. Los sistemas de almacenamiento actuales empiezan a tener problemas en su rendimiento al tener cantidades de datos en proporción de Petabytes o superiores. Big Data está hecho para trabajar con estos volúmenes de datos.

2.2.1.2. Velocidad

Una de las principales características es el tiempo de procesado y respuesta sobre estos grandes volúmenes de datos, obteniendo resultados en tiempo real y procesándolos en tiempos muy reducidos. No sólo se trata de procesar sino también de recibir, en la actualidad las fuentes de datos pueden llegar a generar mucha información cada segundo, obligando al sistema que recepta dicha información a tener la capacidad para almacenar la información de manera muy veloz.

2.2.1.3. Variedad

La infinidad de nuevas fuentes de datos proporcionan nuevos y distintos tipos de formatos de información a los ya conocidos hasta ahora, como son los datos no estructurados, que un sistema de Big Data es capaz de almacenar y procesar sin tener que realizar un reproceso para estructurar o indexar la información.

2.2.1.4. Variabilidad

Todas las tecnologías que componen una arquitectura Big Data deben ser flexibles a la hora de adaptarse a nuevos cambios en el formato de los datos,

tanto en la obtención como en el almacenamiento al igual que su procesado. Se puede decir que la evolución es una constante en la tecnología de tal manera que los nuevos sistemas deben estar preparados para admitirlos.

2.2.1.5. Valor

El objetivo final es generar valor de toda la información almacenada a través de distintos procesos de manera eficiente y con el costo más bajo posible.

De tal manera, un sistema Big Data debe extraer ese valor que tienen los datos en forma de nueva información, por ejemplo, sobre grandes volúmenes de datos y de diferentes fuentes de la manera más rápida y eficiente posible, adaptándose a todos los formatos estructurados o no existentes.

2.3. Arquitectura Big Data

En la Figura 1, se puede observar el flujo que la información tendría en una arquitectura Big Data, con orígenes de datos diversos, documentos o datos recibidos en manera de Streaming, los cuales se reciben y almacenan a través de la capa de recolección de datos, con herramientas específicamente desarrolladas para tal función. Los datos recibidos pueden procesarse, analizarse y/o visualizarse tantas veces como haga falta y lo requiera el caso de uso específico.

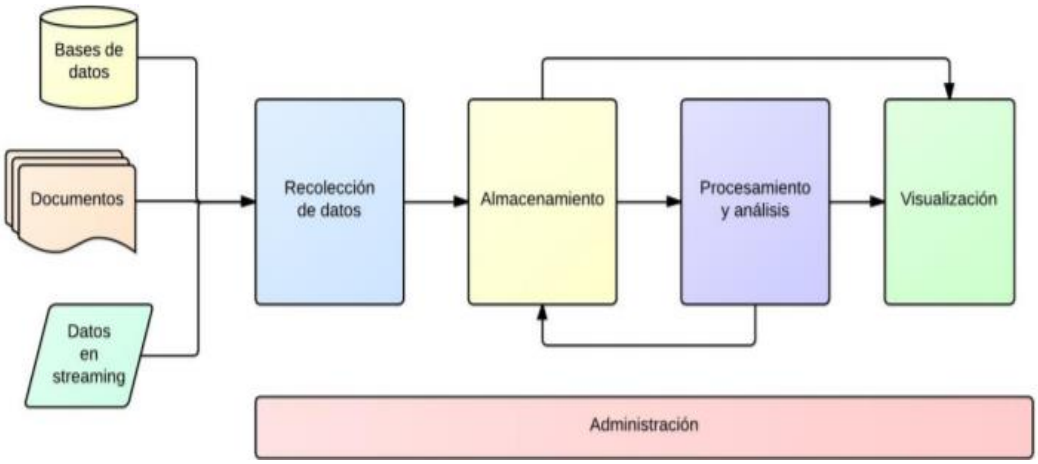


Figura 1. Modelo de Big Data por capas

Tomado de (Talend, 2019)

La arquitectura Big Data está compuesta básicamente por cinco capas: la recolección de datos, el almacenamiento, procesamiento de datos, la visualización y administración, en la Figura 2 se puede visualizar de manera más

clara estas etapas. Esta arquitectura viene desde su inicio lo cual implica que no es nueva, sino que ya es algo generalizado en las soluciones de inteligencia de negocios que existen hoy en día. Sin embargo, debido a las nuevas necesidades cada uno de estos pasos ha ido adaptándose y aportando nuevas tecnologías a la vez que abriendo nuevas oportunidades.

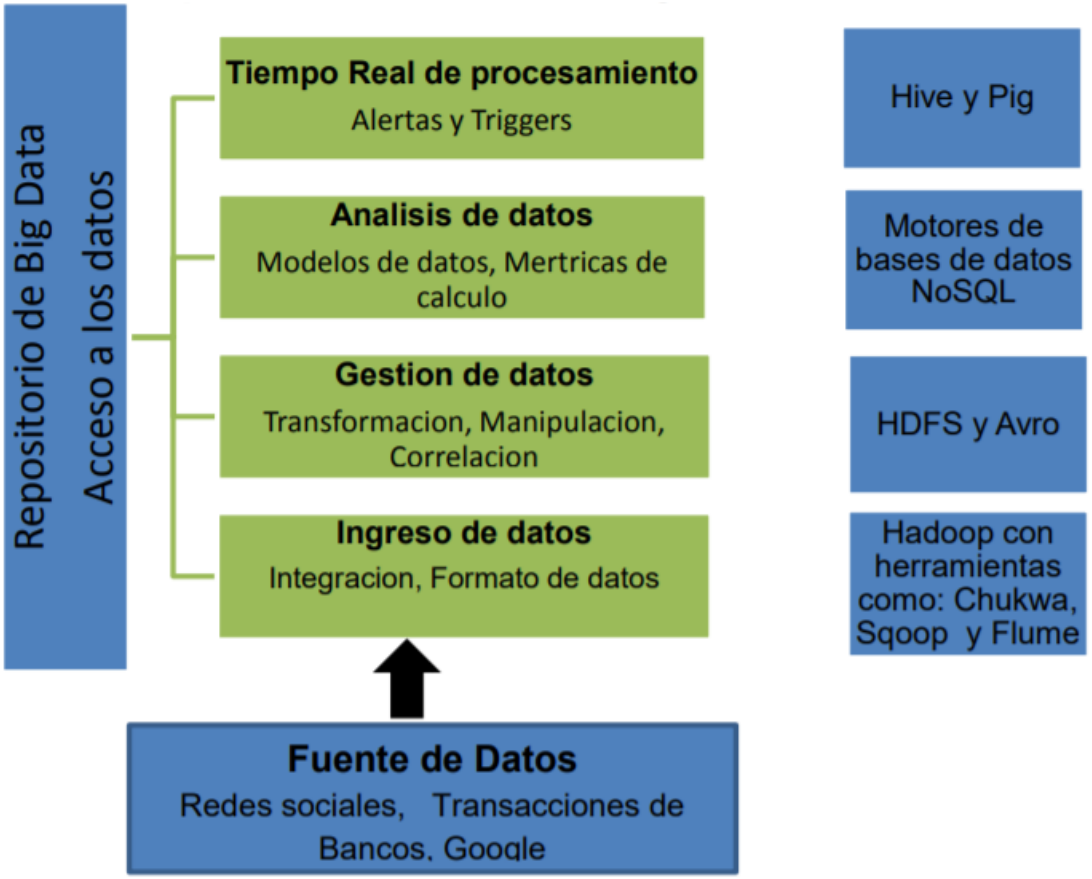


Figura 2. Arquitectura de Big Data.

Tomado de (Fernández, 2017)

La eficiencia de una arquitectura robusta y el paso del tiempo han fomentado la creación de nuevas estrategias para la toma de decisiones, dando un importante lugar al análisis predictivo, debido a que esto ha podido determinar diversos tipos de patrones entre la sociedad, generando como consecuencia gran cantidad de beneficios consistentes en la innovación, investigación y desarrollo de nuevas soluciones.

- La red social Facebook almacena, registra y analiza diariamente 30 Petabytes de datos, 94% de los usuarios de Hadoop realiza análisis de

grandes volúmenes de información que antes no se podía analizar. (IBM, 2019)

- Descifrar el genoma humano tardó cerca de 10 años, actualmente ese proceso se puede realizar en una semana. (IBM, 2019)

2.4. Evolución de Big Data

Big Data ha demostrado tener un crecimiento exponencial en los últimos años. “Su historia se remonta al nacimiento de las primeras herramientas informáticas que llegaron en 1940. En esta misma década comenzaron a aparecer programas que eran capaces de predecir posibles escenarios futuros. Por ejemplo, el equipo del Proyecto Manhattan (1944) que realizaba simulaciones por ordenador para predecir el comportamiento de una reacción nuclear en cadena”. (NIST, 2017).

Según Artaza, no fue hasta la década de los 70 en la que se popularizó el análisis de datos. En 1978 se crea Black-Scholes, un modelo matemático que permitía predecir el precio de acciones futuras. Con la llegada de Google en 1998 y el desarrollo de algoritmos para mejorar las búsquedas en la web, se produce el estallido de Big Data. (NIST, 2017).

“Con la entrada del nuevo siglo, este concepto se acuña y recoge todo el significado que se le otorga en la actualidad. Según los analistas, hoy en día se generan 2,5 trillones de Bytes relaciones con el Big Data. Además, cada vez son más demandados aquellos perfiles profesionales que sean capaces de gestionar herramientas de análisis”. (NIST, 2017).

La evolución de Big Data ha permitido incorporar estas tecnologías como apoyo a las labores de los profesionales de TI, sin embargo, la incorporación de soluciones con Big Data, llevan a estudiar de forma más clara, como los datos son ingresados, gestionados y procesados, para un entendimiento mejor se especifica a continuación estos conceptos.

2.5. Ingreso de datos

El ingreso de datos es el procedimiento de obtener información para posterior uso o almacenamiento en una base de datos relacional o no. Se basa en coleccionar datos de diferentes fuentes con el propósito de realizar un análisis basado en modelos de programación.

2.5.1. Gestión de datos

La gestión o administración de datos es el desarrollo y ejecución de arquitecturas, políticas, prácticas y procedimientos a fin de gestionar las necesidades del ciclo de vida de información de una compañía de una manera eficaz. Es un planteamiento para administrar el flujo de datos de un sistema a través de su ciclo de vida, desde su invención hasta el momento en que son eliminados. La administración de Big Data es la forma en que se organizan y gestionan grandes cantidades de datos, sea de información estructurada como no estructurada para incrementar estrategias con el fin de ayudar con los conjuntos de datos que crecen rápidamente, donde se ven involucrados desde Bytes pasando por Terabytes y hasta Petabytes de información con variedad de tipos.

2.5.2. Tiempo de procesamiento

Es un proceso que automatiza e incorpora el flujo de datos en la toma de decisiones, este aprovecha el movimiento de los datos para acceder a la información estática y así lograr responder preguntas a través de análisis dinámicos. Los sistemas de procesamiento de flujo se han construido con un modelo centrado que funciona con datos estructurados tradicionales, así como en aplicaciones no estructuradas, como vídeo e imágenes.

“El procesamiento de flujos es adecuado para aplicaciones que tienen tres características: calcular la intensidad (alta proporción de operaciones de E/S), permitir paralelismo de datos y por último la capacidad de aplicar los datos que se introducen de forma continua”. (Cook, 2009).

2.5.3. Análisis de datos

“Es el proceso de examinar grandes cantidades de datos para descubrir patrones ocultos, correlaciones desconocidas y otra información útil”. Esta información puede proporcionar ventajas competitivas y resultar en beneficios para el negocio, como el marketing para generar mayores ingresos. (Techtarget, 2019).

2.6. Procesamiento de datos

Se debe comprender que, sin el procesamiento de datos, las empresas se limitan de acceder a sus propios datos los cuales pueden afinar su ventaja competitiva

y ofrecer información empresarial crítica. Debido a todo eso, que es crucial para todas las empresas conocer y entender la necesidad de procesar todos sus datos y sobre todo saber cómo hacerlo, dicho esto a continuación de describe un concepto de procesamiento de datos.

El procesamiento de datos se produce cuando los datos se recopilan y se traducen en información útil. Habitualmente esta tarea es realizada por un científico de datos (un experto en la ciencia de datos) o un equipo de científicos de datos, es importante mencionar que el procesamiento de datos debe ser realizado correctamente para no afectar de forma negativa el producto final o la salida de datos.

El procesamiento de datos inicia con los datos en su forma sin procesar y los convierte en un formato más legible que pueden ser (gráficos, documentos, etc.), lo que le proporciona la forma y el contexto necesarios para ser interpretados por las computadoras y utilizados por los empleados en las organizaciones.

2.6.1. Etapas del procesamiento de datos

2.6.1.1. Recopilación de datos

La recopilación de datos es el paso inicial en el procesamiento de datos. Los datos se obtienen de las diferentes fuentes disponibles, incluidas las diferentes fuentes de redes sociales y sus almacenes de datos. Es primordial que las fuentes de datos disponibles sean confiables y estén bien construidas, por lo que los datos recopilados (que luego se utilizarán como información) sean de la mejor calidad posible.

2.6.1.2. Preparación de datos

Una vez que se realiza el primer pasó que trata de recopilar los datos, estos entran a la etapa de preparación de los datos. La preparación de datos, a menudo denominada “pre procesamiento”, es la etapa en la que los datos sin procesar se limpian y organizan para la siguiente etapa de procesamiento de datos. Durante la preparación, los datos sin procesar se comprueban diligentemente para detectar cualquier novedad o error. El propósito de este paso es eliminar los datos con fallas que pueden ser (datos redundantes,

incompletos o incorrectos) e iniciar a crear datos de alta calidad para la mejor inteligencia empresarial.

2.6.1.3. Entrada de datos

Cumplido el segundo paso, los datos limpios se ingresan en su destino (tal vez un CRM que puede ser cualquiera en el mercado como Salesforce o un almacén de datos como Redshift), y se traducen a un idioma que pueda entenderse. La entrada de datos es la primera etapa en la que los datos sin procesar comienzan a tomar la forma de información utilizable.

2.6.1.4. Procesamiento

En esta etapa, los datos ingresados en la computadora en la etapa anterior se procesan para su interpretación. El procesamiento se realiza mediante algoritmos de aprendizaje automático, aunque el proceso en sí puede variar ligeramente dependiendo de la fuente de datos que se procesa, siendo (redes sociales, dispositivos conectados, etc.) y su uso previsto (desarrollo de patrones de publicidad, diagnóstico de uso de dispositivos conectados, determinando las necesidades del cliente, etc.).

2.6.1.5. Salida e interpretación de datos

La etapa de salida e interpretación es la etapa en la que los datos son finalmente utilizables por cualquier persona. La misma que puede ser traducida, se puede leer y, a menudo, en forma de gráficos, videos, imágenes, texto sin formato, etc.).

Los miembros de la empresa o institución ahora pueden comenzar a auto administrar los datos para sus propios proyectos de análisis de datos.

2.6.1.6. Almacenamiento de datos

La etapa final del procesamiento de datos es el almacenamiento. Una vez que todos los datos se procesan, se almacenan para su uso futuro. Si bien es posible que cierta información se use de inmediato, gran parte de ella tendrá un propósito más adelante. Por consiguiente, los datos almacenados correctamente son una necesidad para cumplir con la legislación de protección de datos como GDPR. Cuando los datos se almacenan correctamente, los miembros de la organización pueden acceder a ellos rápida y fácilmente cuando sea necesario. (Talend, 2019).

2.6.2. Tipo de información

La información se la puede clasificar en varios tipos de datos según sea su naturaleza u origen, esta clasificación ayuda a comprender mejor el porqué de la evolución de los sistemas de procesamiento de la información hacia entornos de Big Data, a continuación, se detalla los tipos de datos más comunes:

2.6.2.1. Datos estructurados

Es conocida como la información que ya está procesada es decir ya cuenta con un formato estructurado y puede ser filtrada de forma simple. Es el tipo de datos más utilizado en estos días por organizaciones (una tabla dentro de una base de datos relacionales puede ser como el ejemplo de la tabla 2). (IBM, 2013).

Tabla 2.

Tabla personas de una base de datos relacional.

nombre	Apellido1	Apellido2	teléfono	Mail
José	Torres	Pérez	0956254621	jtorres@mcs.com
Iván	Robles	Torres	0968061154	irobles@dcm.com
Víctor	López	Obrador	0895642256	vlopez@ibo.com
Lorena	Rodríguez	Sánchez	0958458545	lsanchez@sos.com

Adaptado de (w3schools, 2018)

2.6.2.2. Datos semiestructurados

Se puede decir que es información procesada y con un formato definido o descrito, pero no estructurado. De este método se puede tener la información definida, pero con una estructura variable. Como ejemplos se pueden definir dos tipos, las bases de datos basadas en columnas y los ficheros con información en un lenguaje de etiquetas (HTML o XML) en la figura 3 se visualiza un ejemplo de datos en formato XML.

```xml
<personas>
    <persona>
        <nombre orden="primero"> Valentina </nombre>
        <nombre orden="segundo"> Lizeth </nombre>
        <apellido orden="primero"> Jaramillo </apellido>
        <apellido orden="segundo"> Obrador </apellido>
        <nacionalidad>Ecuatoriana</nacionalidad>
    </persona>
    <persona>
        <nombre orden="primero"> Itzel </nombre>
        <apellido orden="primero"> Torres </apellido>
        <apellido orden="segundo"> Smith </apellido>
        <nacionalidad>Panameña</nacionalidad>
        <nacionalidad>Española</nacionalidad>
    </persona>
</personas>
```

Figura 3. Ejemplo de un fichero XML con información semiestructurada.

2.6.2.3. Datos no estructurados

Es información sin procesar y que puede tener cualquier estructura. Se puede encontrar en cualquier formato: vídeo, texto, imagen, código, etc. Los directorios de registros (Logs) de aplicaciones o la información subida en las redes sociales son ejemplos específicos de datos no estructurados.

Actualmente la forma de trabajar en el día a día ya incluye almacenar datos de tipo estructurados, no estructurados, o semiestructurados, obligando a pasar por un proceso de filtrado y transformación de los datos no estructurados. Esta gestión o proceso radica en un coste adicional y en una pérdida inevitable de datos que cada vez es más difícil ignorar, esto implica adoptar nuevas maneras de procesar información que permitan encontrar las cinco V comentadas anteriormente y que en un sistema de explotación de la información busca obtener especialmente la variabilidad, veracidad y velocidad de la recolección de información.

Por lo detallado en los párrafos anteriores sobre los tipos de información, se puede mencionar que una de las características principales de un sistema Big Data es trabajar con datos no estructurados, permitiendo aumentar la variedad y la variabilidad. De esta manera también se entiende que el sistema debe poder almacenar y trabajar con un gran volumen de información.

Con lo detallado sobre los datos no estructurados y siendo este una estructura básica en soluciones de Big Data se detallará de mejor manera el concepto no relacional o bases de datos NoSQL termino definido en tecnologías que permiten almacenar y procesas información no estructurada.

2.7. NoSQL

En la actualidad, la manera en la que las aplicaciones web tratan los datos ha cambiado de forma importante durante los últimos 15 años. Cada vez se coleccionan más datos y a su vez son más los usuarios que acceden a estos datos al mismo tiempo. Esto indica que la escalabilidad y el rendimiento se han convertido en auténticos retos para las bases de datos relacionales basadas en esquemas.

Con la aparición del término NoSQL en los años 90 y su uso desde principios de este siglo por empresas como Facebook, Google o Amazon empresas que palpitando la necesidad de mejorar el rendimiento a sus esquemas de datos relacionales buscaron nuevas soluciones de almacenamiento, dando lugar a la creación de bases no relacionales o NoSQL como BigTable, DynamoDB y Cassandra, estas investigaciones que posterior se convirtieron en tecnologías que están dando una posibilidad de abordar la forma de gestionar la información de una manera distinta a como se venía realizando. (MongoDB, 2019).

Para tener una definición adecuada de las bases de datos NoSQL se debe tener en cuenta las siguientes características que se asemejan a esta tecnología.

Distribuido. Base de datos NoSQL a menudo son distribuidos en donde varias máquinas físicas o virtuales contribuyen en grupos para ofrecer a los usuarios datos. Cada entidad que almacena los datos se replica normalmente entre varias máquinas para la redundancia y alta disponibilidad.

Escalabilidad horizontal. Usualmente se pueden agregar nodos de manera dinámica, sin restricción de tiempo de inactividad de cada nodo, permitiendo capacidades de procesamiento general.

Desarrollado para grandes volúmenes. La mayoría de los sistemas NoSQL fueron desarrollados para ser capaces de almacenar y procesar enormes

cantidades de datos de forma rápida contribuyendo a las arquitecturas de Big Data.

Modelo y arquitectura de datos no relacionales. Cada modelo de datos puede variar, y casi siempre no son relacionales. Lo que permite estructuras más complejas y no son tan estructurales o rígidas como en el modelo relacional.

Esquemas diferentes. La estructura de los datos en sistemas NoSQL generalmente no se definen a través de algún esquema explícito que la base de datos maneje. En su lugar, los usuarios almacenan datos como necesiten o deseen, sin tener que cumplir con alguna estructura predefinida.

En la figura 4 se visualiza una comparación sobre las bases de datos relacionales y las NoSQL.

Feature	NoSQL Databases	Relational Databases
Performance	High	Low
Reliability	Poor	Good
Availability	Good	Good
Consistency	Poor	Good
Data Storage	Optimized for huge data	Medium sized to large
Scalability	High	High (but more exponsive)

Figura 4. Comparación de bases de datos NoSQL con base relacional
Tomado de (MongoDB, 2019).

En el contexto de base de datos NoSQL se pueden encontrar cuatro categorías para el almacenamiento de datos NoSQL.

2.7.1. Almacenamiento Key Value

Estas son las bases de datos más simples en cuanto a su uso (la implementación puede ser muy complicada), debido a que en el tipo de almacenamiento Key Value, utiliza una tabla hash en la que una clave única apunta a un elemento, es decir que simplemente almacena valores identificados por una clave.

Normalmente, el valor guardado se almacena como un arreglo de Bytes (BLOB) y es todo. De esta forma el tipo de contenido no es importante para la base de datos, solo la clave y el valor que tiene asociado.

Las claves pueden ser organizadas por grupos, claves lógicas, necesitando solamente estas claves para ser únicas dentro de su propio grupo. Esto permite tener claves idénticas en diferentes grupos lógicos. La tabla 3 muestra un ejemplo de un almacén de valores clave, la clave es el nombre de la ciudad y el valor es la dirección de la universidad de las Américas, sede principal.

Tabla 3.

Ejemplo de almacenamiento Key Value en una base NoSQL

KEY	VALUE
QUITO	{“Universidad de las Américas, de los Colimes esquina y avenida granados, Quito 170125 Ecuador”}

Mediante el ejemplo anterior se puede mencionar que lo que se necesita conocer para acceder a los elementos almacenados en la base de datos: es la clave. En este tipo de almacenamiento los datos se almacenan en una forma de cadena JSON con diferentes atributos, como se visualizó en la tabla anterior.

2.7.2. Base de datos columnares

Estas bases de datos guardan la información en columnas en lugar de filas ("como se lo llevaría en la mayoría de los sistemas de gestión de bases de datos relacionales"), con esto se logra una mayor velocidad en realizar la consulta debido a que el almacenamiento en columnas permite el acceso rápido de lectura y escritura.

Esta solución de almacenamiento por columnas es conveniente en ambientes donde se presenten muchas lecturas como en Data Warehouse y sistemas de inteligencia de negocios conocido también con sus siglas en inglés BI (Business

Intelligence), las bases de datos más conocidas que usan el almacenamiento por columnas incluyen a Cassandra, Hbase y Google BigTable.

2.7.3. Bases de datos orientadas a documentos

Se asemejan al tipo de almacén Key-Value, diferenciándose a que la información no se guarda en binario, sino como un formato que la base de datos pueda leer, como XML, o cualquier otro leguaje de etiquetado.

En el tipo de almacenamiento orientado a documentos, los valores proporcionan codificación XML, JSON o BSON (JSON codificado binario) para los datos almacenados.

2.7.4. Bases de datos orientados a grafos

Estas bases de datos manejan la información en forma de grafo, en una gráfica de una base de datos NoSQL, se utiliza una “estructura de gráfica dirigida” para representar los datos, el gráfico está compuesto por bordes y nodos. Entregando una mayor importancia a la relación que tienen los datos, con esto se logra que las consultas puedan ser realizadas de forma óptima que en un modelo relacional, en la actualidad InfoGrid e InfiniteGraph son las bases de datos gráficas más populares.

Tabla 4.

Comparación de bases de datos NoSQL.

	Performance	**Scalability**	**Flexibility**	**Complexity**	**Functionality**
Key-Value stores	Hight	Hight	Hight	None	Variable
Column stores	**Hight**	**Hight**	**Moderate**	**Low**	Minimal
Document stores	Hight	Variable (Hight)	Hight	Low	Variable
Graph databases	**Variable**	**Variable**	**Hight**	**Hight**	**Graph theory**
Relational databases	variable	variable	Low	Moderate	Relational algebra

Adaptado de (Christof, 2019)

En los párrafos descritos anteriormente se puede ver que las bases de datos relacionales dan más importancia a la consistencia y a la disponibilidad, en contraste con las NoSQL, que dan mayor prioridad a la tolerancia y en muchas ocasiones a la disponibilidad, estos aspectos se definen de mejor manera revisando el teorema de CAP.

2.8. Teorema de CAP

En ciencias de la computación, el teorema CAP, también llamado Conjetura de Brewer, menciona que es imposible para un sistema de cómputo distribuido garantizar a la misma vez consistencia, disponibilidad y tolerancia a particiones, es decir que no podrán todos los nodos ver la información al mismo tiempo.

El teorema de CAP define tres requerimientos a tener en cuenta al implementar un sistema distribuido, estos son:

- Consistencia: se refiere a la integridad de la información. Al realizar una consulta o inserción todos los nodos del sistema deben ver la misma información en todo momento.
- Disponibilidad: la aplicación debe estar siempre disponible, si falla algún nodo los demás pueden seguir operando sin inconvenientes y el usuario debe poder leer y escribir, aunque se haya caído uno de los nodos.
- Tolerancia al particionamiento: el sistema continúa funcionando a pesar, de que se pierdan mensajes.

2.8.1. Clasificación de requerimientos según el teorema de CAP

AP: Garantizan disponibilidad y tolerancia a particiones, pero no la consistencia, al menos de forma total. Algunas de ellas consiguen una consistencia parcial a través de la replicación y la verificación.

CP: Garantizan consistencia y tolerancia a particiones. Para lograr la consistencia y replicar los datos a través de los nodos, sacrifican la disponibilidad.

CA: Garantizan consistencia y disponibilidad, pero tienen problemas con la tolerancia a particiones. Este problema lo suelen gestionar replicando los datos. (Genbeta, 2019)

En la figura 5 se visualiza los tres requerimientos a tener en cuenta al implementar un sistema distribuido según el teorema de CAP.

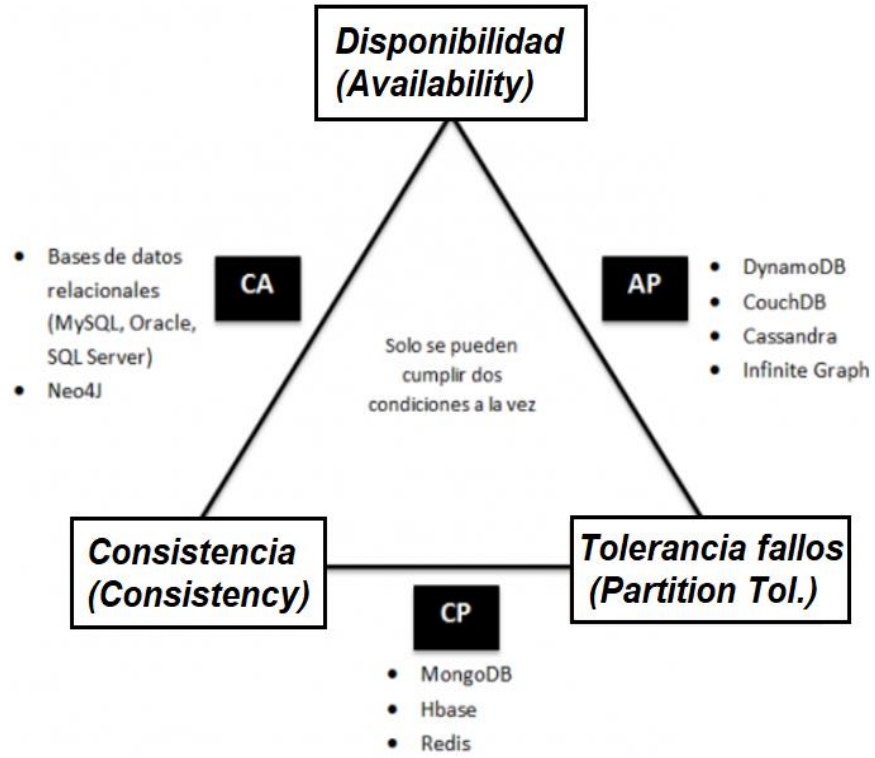


Figura 5. Clasificación de bases de datos según el teorema de CAP. Tomado de (Genbeta, 2019).

Las diferencias que pueden existir dentro de las bases de datos relaciones y las NoSQL, se puede entender revisando dos modelos importantes: BASE y ACID, los cuales se utilizan respectivamente para el manejo de las transacciones.

John D. Cook menciona que la clave de los sistemas basados en NoSQL, es que nada debe ser compartido, para cumplir el propósito de replicar y particionar los datos sobre muchos servidores, debido a que esto permite soportar un gran número de operaciones de escritura y lectura por segundo, esto conlleva a que los sistemas NoSQL generalmente no proveen las propiedades transaccionales del modelo ACID las cuales son atomicidad, consistencia, aislamiento y durabilidad, pero si pueden proveer su propio modelo el cual es llamado BASE por sus propiedades donde las transacciones deben ser básicamente disponibles, eventualmente consistentes y de estado ligero. (Cook, 2009).

En la tabla 5 se realiza una comparación entre las propiedades BASE y ACID para diferenciar los sistemas basados en NoSQL.

Tabla 5.

Comparación de propiedades BASE y ACID.

BASE	ACID
Basically Available: esto quiere decir básicamente disponible, aquí se utiliza la replicación para reducir la probabilidad de que se presente la indisponibilidad de los datos o la fragmentación de la información a través de los distintos servidores de almacenamiento.	Atomicidad: toda la secuencia de acciones debe ser completada o abortada, es decir que la operación no puede ser parcialmente completada, debe haber una operación de commit cuando se completa, o una de rollback en el momento en que la transacción no pueda ser completada.
Soft State: mientras los sistemas que utilizan el acrónimo ACID, en las bases de datos relacionales asumen la consistencia de la información como requerimiento de alta prioridad, los sistemas NoSQL permiten que los datos sean inconsistentes y relegan de diseño estructural.	Consistencia: esto manifiesta que la transacción toma los recursos de un estado valido o real para llevar a la base de datos a otro estado valido.

Tomado de (Cook, 2009).

3. CAPÍTULO III. EVALUAR LA INFORMACIÓN DEL CONSUMO ELÉCTRICO DE LOS EQUIPOS DE RED CONFIGURADOS EN EL CENTRO DE DATOS EXPERIMENTAL

3.1. Centro de datos

Se define a un centro de datos como una ubicación física dentro de una empresa en donde se realiza el procesamiento de datos y se albergan equipos activos y pasivos de telecomunicaciones y sistemas de almacenamiento todo en un ambiente controlado mediante fuentes de alimentación, sistemas de climatización permitiendo que los equipos tengan el mejor nivel de rendimiento con la máxima disponibilidad de los sistemas.

Los centros de datos se categorizan de acuerdo a su fiabilidad y son certificados por el Uptime Institute el cual valida disponibilidad, fiabilidad y continuidad de negocio entre las características más principales, por esta razón los nombra con el termina de TIER que va desde el uno al cuatro dependiendo el porcentaje de disponibilidad. (Uptime, 2019).

3.2. Clasificación de los centros de datos

Debido a la importancia de la interconexión de dispositivos y el valor de la información que estos trasmiten por medio de los centros de datos, este ya es un recurso crítico de negocio, debido a esto el Uptime Institute entidad certificadora a nivel mundial los clasifica de acuerdo con su disponibilidad de la siguiente manera.

3.2.1. Tier I: Centro de datos Básico

- Disponibilidad del 99.671%.
- El servicio puede interrumpirse por actividades planeadas o no planeadas.
- No hay componentes redundantes en la distribución eléctrica y de refrigeración.
- Puede o no puede tener suelos elevados, generadores auxiliares o UPS.
- Tiempo medio de implementación, 3 meses.
- La infraestructura del centro de datos deberá estar fuera de servicio al menos una vez al año por razones de mantenimiento y/o reparaciones. (Guilarte, 2013)

3.2.2. Tier II: Centro de datos Redundante

- Disponibilidad del 99.741%.
- Menos susceptible a interrupciones por actividades planeadas o no planeadas.
- Componentes redundantes (N+1)
- Tiene suelos elevados, generadores auxiliares o UPS.
- Conectados a una única línea de distribución eléctrica y de refrigeración.
- De 3 a 6 meses para implementar.
- El mantenimiento de esta línea de distribución o de otras partes de la infraestructura requiere una interrupción del servicio. (Guilarte, 2013)

3.2.3. Tier III: Centro de datos Concurrentemente Mantenibles

- Disponibilidad del 99.982%.
- Permite planificar actividades de mantenimiento sin afectar al servicio de computación, pero eventos no planeados pueden causar paradas no planificadas.
- Componentes redundantes (N+1)
- Conectados múltiples líneas de distribución eléctrica y de refrigeración, pero únicamente con una activa.
- De 15 a 20 meses para implementar.
- Hay suficiente capacidad y distribución para poder llevar a cabo tareas de mantenimiento en una línea mientras se da servicio por otras. (Guilarte, 2013)

3.2.4. Tier IV: Centro de datos Tolerante a fallos

- Disponibilidad del 99.995%.
- Permite planificar actividades de mantenimiento sin afectar al servicio de computación crítico, y es capaz de soportar por lo menos un evento no planificado del tipo ‘peor escenario’ sin impacto crítico en la carga.
- Conectados múltiples líneas de distribución eléctrica y de refrigeración con múltiples componentes redundantes (2 (N+1) significa 2 UPS con redundancia N+1).
- De 15 a 20 meses para implementar. (Guilarte, 2013)

En la figura 6 se visualiza el sello que identifica a un centro de datos certificado como Tier IV.

Figura 6. Sello de un centro de datos certificado Tier IV.
Tomado de: (Uptime, 2019)

3.3. Tendencias actuales de los centros de datos

En la actualidad ya se está inmerso en las tecnologías del IoT o más conocido como el IoE (Internet del todo), la filosofía de Big Data esta día a día incorporándose más en las empresas de Latinoamérica y sin dejar a un lado el termino de cloud computing, que hoy en día es un agregado al momento de elegir una prestación, en donde las empresas puedan alojar sus servicios en un centro de datos que ofrezcan soluciones a nivel de (IaaS, PaaS, Sagas) todo orientado al SDN o mejor conocido como software definido por red.

Todo esto implica recopilar nuevas tendencias que ya no están solo ligadas a sistemas convergentes o hyperconvergentes los cuales hoy en día ya están inmersos en los centros de datos actuales, las nuevas tendencias ya cuentan como soluciones que proyectan el uso de modelos de eficiencia energética para lograr un mejor consumo de energía en equipos, por lo cual en este apartado presentamos soluciones innovadoras y nuevas tendencia que en los próximos años serán indispensables al momento de administrar un centro de datos.

3.3.1. Edge Computing

Es un concepto que se apoya en seguir avanzando en acercar la información almacenada a los puntos de generación de los datos en las nuevas aplicaciones derivadas del IoT, existe una gran oportunidad de negocio en Micro Data Center.

Estos permiten desplegar soluciones seguras, íntegras y gestionadas en muchas ubicaciones de forma estandarizada, a fin de disminuir la latencia y lograr un mejor performance de las aplicaciones. (Fraga, 2018).

3.3.2. Consolidación del Cloud

En los últimos años el desarrollo de diferentes tecnologías ha permitido que el usuario tenga una mejor madurez al momento de escoger la mejor forma de consumir sus datos. Según el informe FutureScape 2018, realizado por Schneider Electric y la consultora IDC, a finales del 2018 el 65% de los activos de TI de las empresas alojarán sus servicios en entornos cloud, en un centro de datos de algún proveedor, mientras que un tercio del personal TI será contratado por proveedores de servicios Cloud.

Las empresas operarán entornos diversificados que incluirán distintos tipos de despliegue (dentro y fuera de las localizaciones físicas de la organización) y un amplio portafolio de servicios cloud (PaaS, IaaS, SaaS). (Fraga, 2018).

3.3.3. IoT y equipos conectados

Con la aceleración de internet y la telefonía móvil, las organizaciones invertirán más en sistemas Big Data, analítica y sistemas que permitan el control de internet de las cosas, deseando mejores infraestructuras de comunicaciones con el exterior. El centro de datos se convertirá en parte fundamental del desarrollo del internet de las cosas, que permita una mayor eficiencia de utilización de recursos y de energía. (Fraga, 2018).

3.3.4. Convergencia e Hiperconvergencia

Es real que en la actualidad ya se encuentre inmiscuido tecnologías como convergencia e Hiperconvergencia en la mayoría de centro de datos, también es verdad que a un no se ha desplegado como debería. Estas tendencias se terminarán de adaptar en los siguientes años debido a que la convergencia implica simplificar la gestión y reduzca las tareas manuales, y la Hiperconvergencia busque integrar hardware y software para operar eficientemente los centros de datos, esto será un aporte diferenciador al momento de administrar un centro de datos. (Fraga, 2018).

En el pasado 2018 se dio a conocer el Hyper-Pod, una solución que permite encerrar o encapsular el aire caliente de la mayor cantidad de racks posibles en un mismo espacio, de manera que se canalice más rápido hacia el sistema de enfriamiento y consuma menos energía.

3.3.5. Blockchain como sistema biológico del nuevo CPD

Una de las amenazas, o tal vez una oportunidad y reto de seguridad para los centros de datos tiene nombre y es el Blockchain.

En los siguientes años surgirán mecanismos para gestionar los datos de forma inmutable, es decir verdaderamente distribuida y con toda confianza (es decir, descentralizado, sin una autoridad central) que van a tener un impacto profundo en los centros de datos. (Fraga, 2018).

La Blockchain es un buen ejemplo de esto, Mark Bregman, responsable tecnológico de NetApp. Menciono lo siguiente “Los mecanismos descentralizados representan un reto al concepto tradicional de la protección y la gestión de datos; Como no hay un punto central de control, como un servidor centralizado, es imposible cambiar o eliminar la información en un Blockchain y, además, todas las transacciones son irreversibles”. (Fraga, 2018).

Esto quiere decir que un centro de datos construido sobre una Blockchain será más seguro y el flujo de información siempre mantendrá su integridad.

Detalladas las tendencias que a criterio del autor serán de suma importancia en los próximos años, se debe decir que en la actualidad los centros de datos se han convertido en una pieza fundamental en la industrialización de las empresas siendo vital tener un centro de procesamiento disponible, lo que implica un coste alto en energía, por eso el objetivo actual de construir sistemas que incluyan modelos de eficiencia energética amigables con el ambiente, a fin de poder conseguir que el PUE (Eficacia del uso de energía) que es una medida utilizada para determinar la eficiencia energética de un centro de datos.

Debido al tema del proyecto de tesis que abarca el análisis de datos del consumo de energía en equipos del centro de datos experimental de la UDLA, se especifica el PUE (Eficacia del uso de energía) y como esto ayuda a distinguir el uso correcto de energía en los equipos activos y pasivos de un centro de datos.

3.4. Eficacia del uso de energía PUE (Power Usage Effectiveness)

Con la incorporación del Big Data y tecnologías como el cloud computing, la creación de centro de datos eficientes y comprometidos con el medio ambiente y el ahorro de energía se convierte en una necesidad de primer orden. Todo esto ya no solo implica una solución a nivel ambiental sino también en el económico, debido a que una empresa que albergue en sus instalaciones un centro de datos gasta más de un 30% de su presupuesto en energía y gastos de mantenimiento. (DCiE, 2018).

Existe un baremo que mide lo eficiente o no que es un centro de datos, y este es el PUE por sus siglas en inglés, también es conocido como (Power Usage Effectiveness), que es una medida utilizada para determinar la eficiencia energética de un centro de datos.

El PUE calcula dividiendo la energía total del centro de datos entre la energía consumida por los servidores. Cuanto más se aproxime la cifra resultante a 1, más eficiente será el centro de datos, lo que se traducirá en un mayor ahorro para las empresas. (Rojas, 2013).

En un ejemplo donde se determine que un centro de datos tiene un PUE de 2.0, esto se interpretará que por cada vatio de energía que alimenta a los servidores, otro vatio va para la refrigeración, la iluminación y otros sistemas. Es decir, un PUE de 2.0 en un centro de datos no resultaría rentable para una empresa.

En la figura 7 se visualiza los valores matemáticos y parámetros técnicos que se toman en cuenta al momento de realizar el cálculo de PUE correspondiente a un centro de datos.

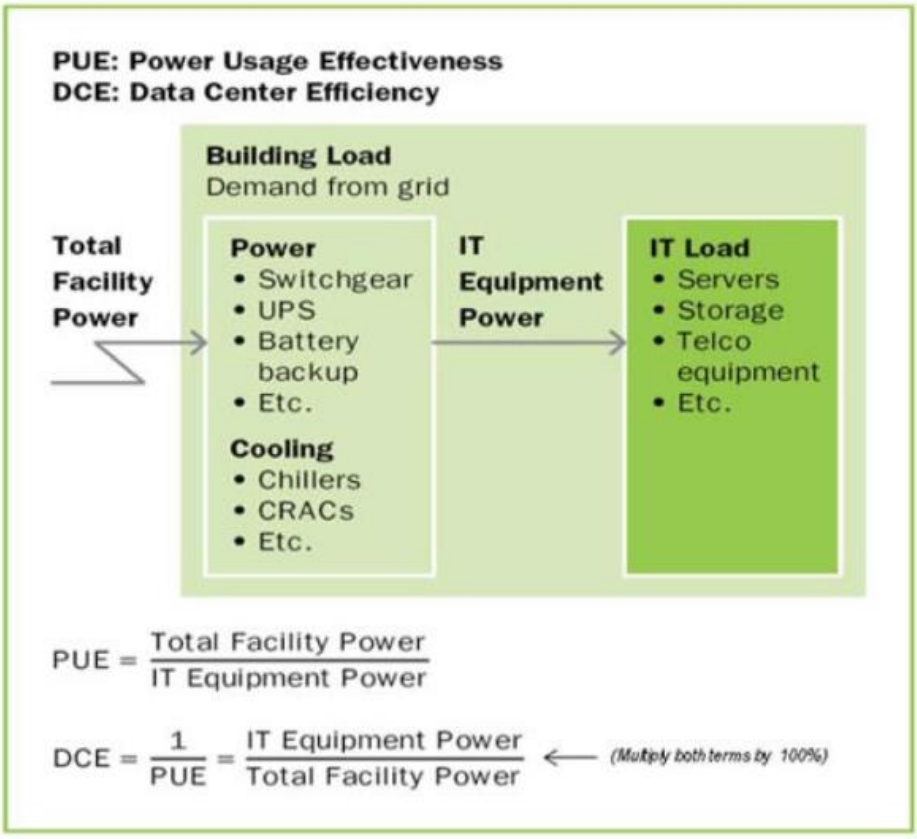


Figura 7. Especificaciones para el cálculo de PUE en centro de datos. Tomado de: (DCiE, 2018)

3.5. Detalle de la infraestructura y consumo actual de equipos del centro de datos experimental

3.5.1. Infraestructura actual

La universidad de las Américas cuenta con un centro de datos experimental en el cual los alumnos pueden poner en práctica lo estudiado en clases, mediante prácticas físicas y el uso de software de simulación de red.

El centro de datos actualmente está equipado con una infraestructura básica que se puede categorizar en tres grandes grupos: equipos utilizados para networking, subsistema de cómputo y un subsistema de almacenamiento, adicional cuenta con un sistema de alimentación ininterrumpida UPS.

3.5.2. Componentes de red

En la tabla 6 se especifica las características de los equipos de red que se encuentran en el centro de datos experimental.

Tabla 6.

Equipos de red (centro de datos experimental UDLA).

EQUIPOS DE RED				
Centro de datos experimental del campus Query				
Cantidad	Modelo	Serie	Detalle	Observación
2	Cisco Nexus 3524	N3K-C3524P-10GX	Switch Licenciamiento LAN Basic 24 puertos licenciados SFP+ Sistema Operativo NX-OS	Dos fuentes de poder y ventiladores, se aprecia conectores C13 y C14

Tomado de (Cisco, 2018)

3.5.2.1. Cisco Nexus 3524

El Cisco Nexus 3524 integra la familia de switches de la serie 3000 son una cartera completa de 1, 10 y 40 Gigabits creados a partir de la arquitectura SoC (Switch on a Chip). Esta serie de switches pertenece a la serie que ofrece un rendimiento de nivel 2 y 3 de velocidad de línea y es adecuada a la arquitectura ToR (Top of the Rack), es decir, conexión punto a punto dentro del rack. (Cisco, 2019).

Cuenta con 24 puertos, es un switch raqueable de 1 unidad de rack (1UR) con una latencia ultra baja que se pueden configurar en tres modos distintos los cuales detallamos a continuación.

- Modo normal: es un modo para ambientes que requieren baja latencia y alta disponibilidad. La latencia es tan baja como 250 ns (nano segundos) y se pueden emparejar con valores de escalado de capa 2 y 3.
- Modo Warp: para soluciones en ambientes pequeños y que necesitan las latencias lo más bajas posibles. La latencia es tan baja como 200 ns.
- Warp SPAN: permite que todo el tráfico ingrese en un solo puerto del Switch y se replique en cualquier interfaz con una latencia tan baja como 50 ns.

En la figura 8 se puede observar los puertos de comunicación de administración de un Nexus 3548.

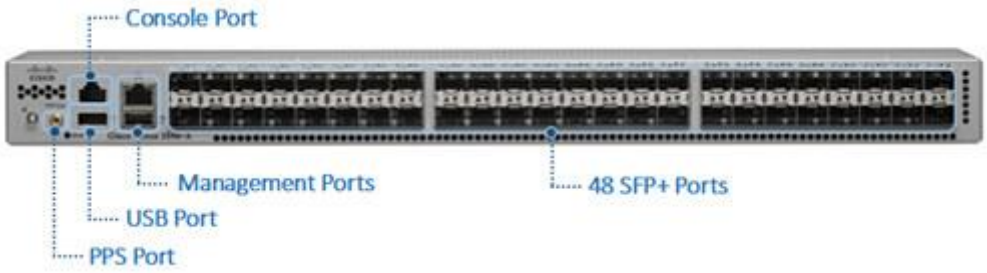


Figura 8. Cisco Nexus 3548.

Tomado de (Cisco, 2019)

2.5.2.1.1. Características de Cisco Nexus 35487

Traducción de direcciones de red NAT

- Pueden realizar NAT para paquetes enrutados de unidifusión IPv4 sin aumentar tiempos de espera.

Monitoreo de latencia

- Se puede identificar la latencia en un puerto de salida específico a través de la interfaz de línea de comandos.

Monitoreo activo del búfer

- Permite la recopilación de datos de utilización del búfer en hardware en muestreos de 10 ns o menos.

Reflejo del tráfico avanzado

- Permite la supervisión de la red y monitoreo de cada puerto.
- IEEE 1588 PTP con salida de pulso por segundo
- Crea y mantiene la solución sincronizada para un óptimo funcionamiento.

Monitoreo del tráfico de red con Cisco Nexus Data Broker

- Se puede crear una conexión de red o SPAN simple, escalabre y rentable para el monitoreo y análisis del tráfico de red.

Una de las ventajas de la robustez del Nexus 3548 se debe a la potencia del hardware que cuenta con las siguientes especificaciones.

- 48 puertos SFP + fijos (1 o 10 Gbps); El Cisco Nexus 3524-X habilita solo 24 puertos.
- Dos fuentes de alimentación intercambiables en modo caliente y redundantes
- Cuenta con cuatro ventiladores individuales redundantes intercambiables en caliente.
- Un puerto de temporización 1-PPS, con el tipo de conector Quick Connect RF1.0 / 2.3
- Un puerto de gestión 10/100/1000
- Un puerto de consola serie RS-232
- Dos puertos USB
- LED localizador (Cisco, 2019).

A nivel de software el Cisco Nexus 3524 cuenta con un sistema operativo Cisco NX-OS desarrollado con flexibilidad, modularidad y capacidad de servicios en su base. Este sistema operativo ayuda a garantizar la disponibilidad continua con lo que es idóneo para centros de datos con misión crítica, su diseño modular y autorreparable permite que las operaciones de impacto cero se conviertan en realidad y proporcionan flexibilidad. A continuación, se procede a detallar sus principales características:

- Software común en los centros de datos Cisco: se ejecuta en todas las plataformas de switches de Cisco.
- Diseño de software modular: lo que implica que sea más robusto, con mayor tolerancia a fallos, mayor disponibilidad de red y mayor escalabilidad.
- Facilidad de administración: cuenta con una interfaz XML basada en el estándar de la industria NETCONG lo que proporciona un rápido desarrollo y creación de herramientas para una mejor gestión.
- Control de acceso basado en roles: aumenta la seguridad del centro de datos.

3.5.3. Evaluación técnica del consumo eléctrico

Para la evaluación técnica de los equipos de red, cómputo y almacenamiento, se instaló la herramienta de monitoreo zabbix para extraer información actual de los equipos Nexus mediante uso de sus MIB las cuales son proporcionadas por el fabricante y se puede descargar desde el sitio oficial del equipo.

3.5.4. Evaluación de consumo eléctrico de los quipos Nexus 3524

Al implementar el monitoreo mediante protocolo SNMP para conocer los parámetros actuales de consumo eléctrico, se realizó los siguientes pasos.

- Para la extracción de información de los equipos se necesitó configurar una comunidad SNMP con la que se pudo extraer información del equipo correspondiente a sus parámetros de consumo actuales, en la figura 9 se puede verificar que los equipos Nexus no tenían configurado ninguna comunidad SNMP por lo cual se procedió a configurar este parámetro, en la figura 10 se puede observar la configuración de la comunidad en modo public, la figura 11 muestra los parámetros que fueron analizados mediante el protocolo SNMv2 en los Nexus.

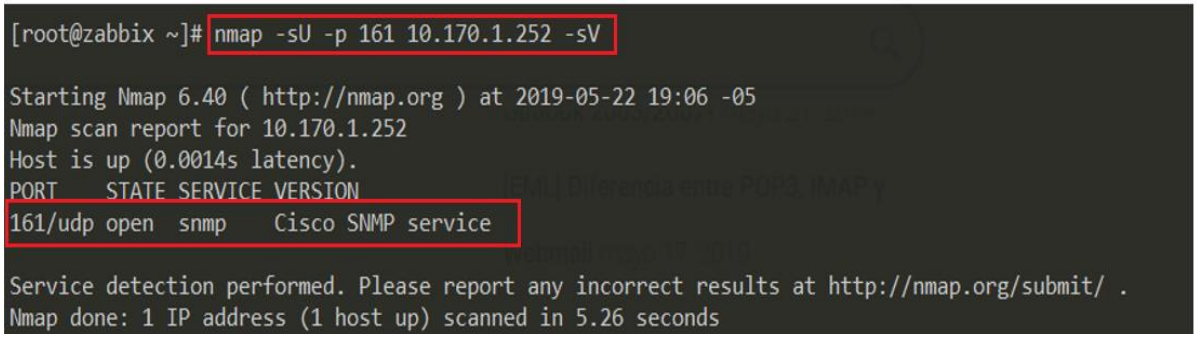


Figura 9. Validación protocolo SNMP Cisco Nexus 3524 (IP 10.170.1.252).

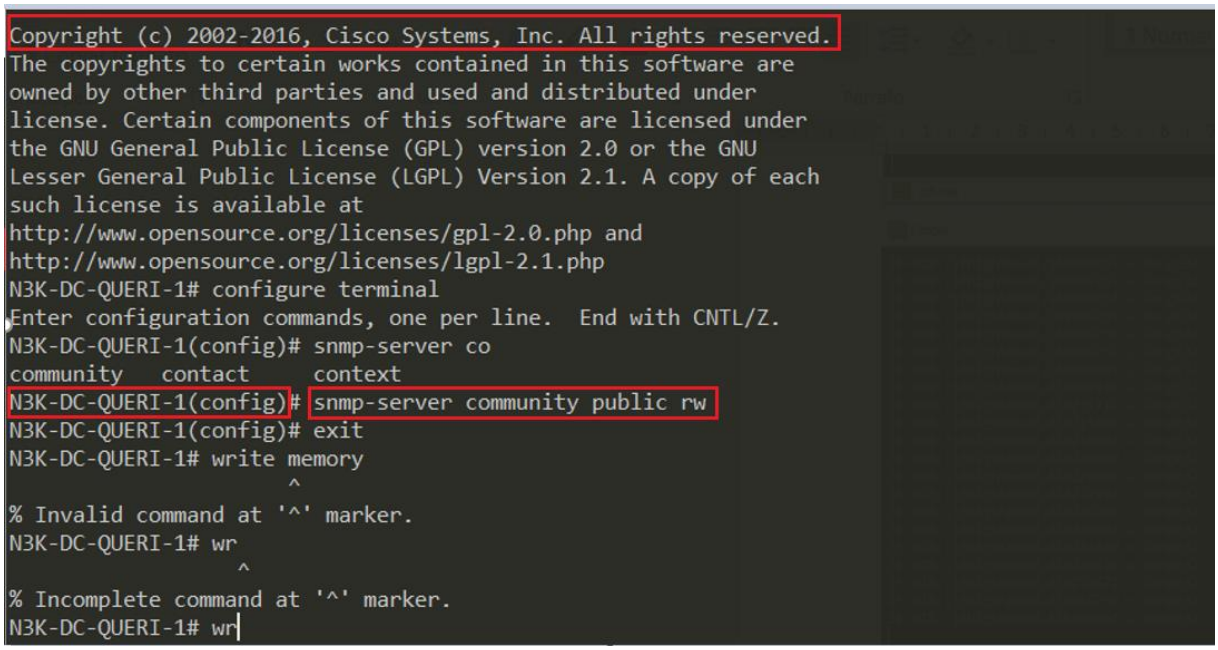


Figura 10. Configuración comunidad SNMP, equipo Nexus 3524.

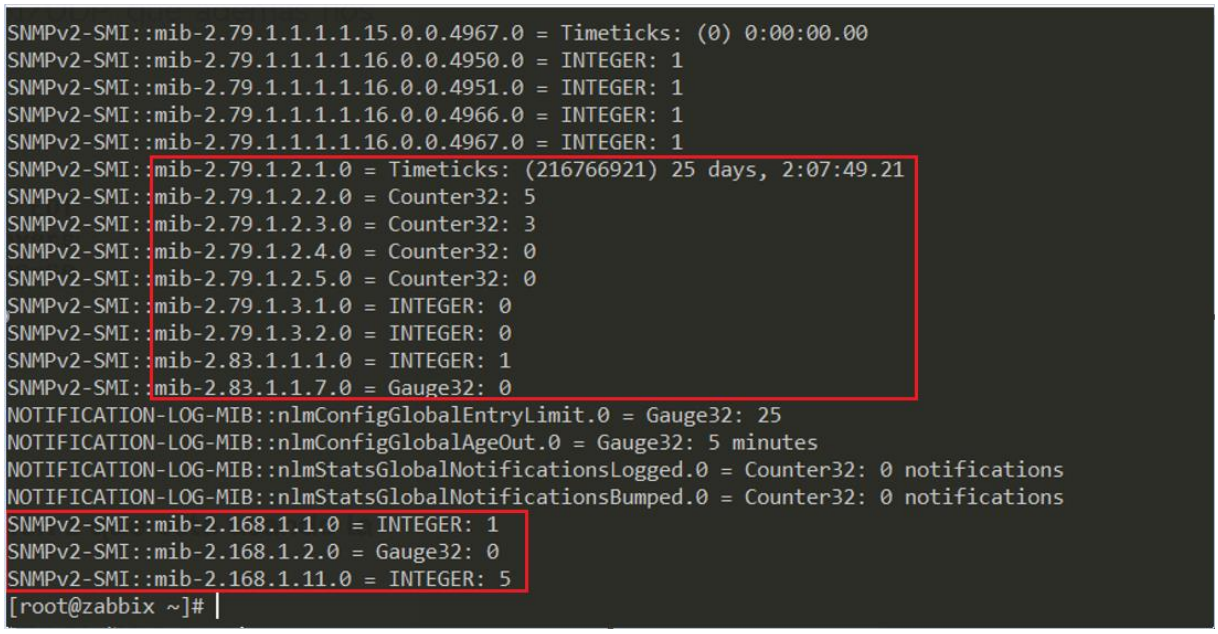


Figura 11. Parámetros analizados con protocolo SNMP Cisco Nexus.

- Los parámetros de voltaje fueron capturados en la herramienta de monitoreo zabbix mediante la configuración del template y la comunidad, en la figura 12 se visualiza los parámetros configurados en el software zabbix.

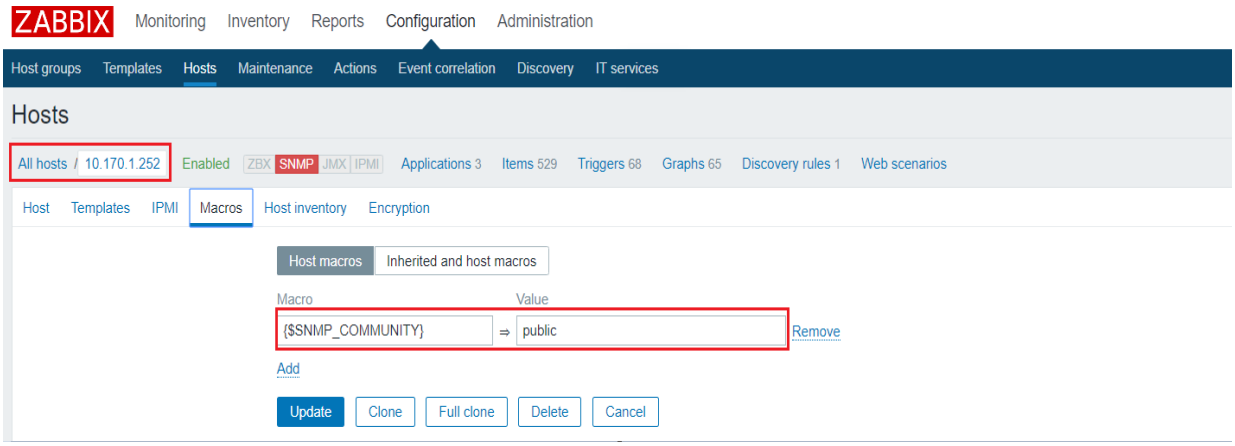


Figura 12. Configuración del template de monitoreo equipos Nexus.

- En la figura 13 se verifica que los resultados mediante SNMP están activos para los dos Nexus con las direcciones de red 10.170.1.252 y 10.170.1.253.
- Los resultados obtenidos mediante el monitoreo de su protocolo SNMP en los equipos de red serán analizados en el capítulo de evaluación de resultados.

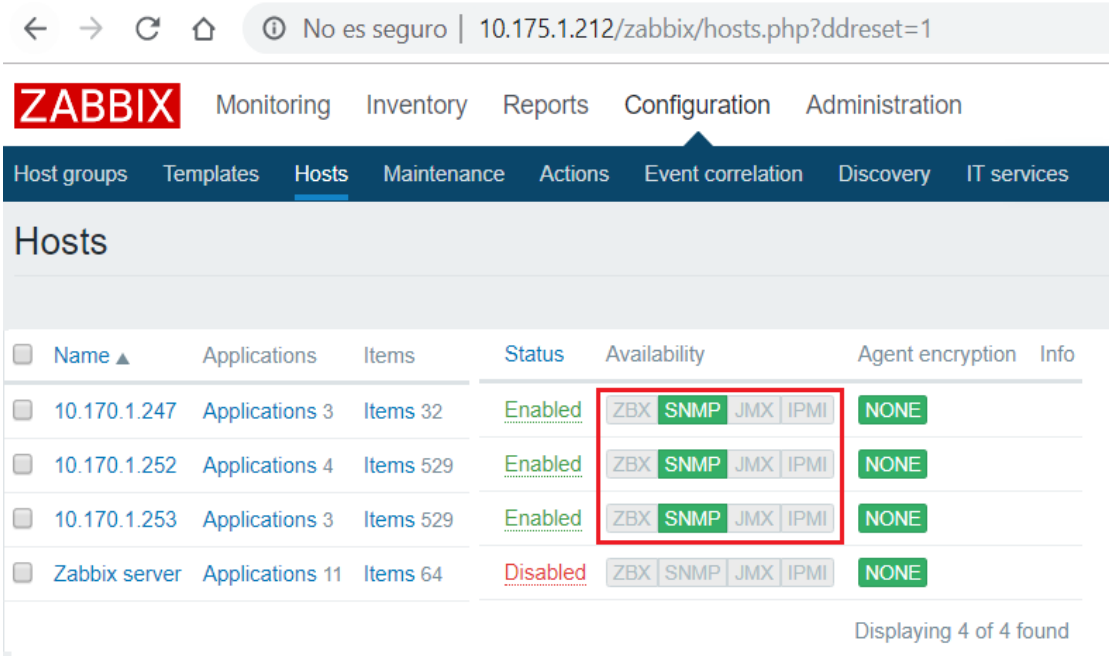


Figura 13. Resultados de monitoreo mediante SNMP Cisco Nexus 3524.

3.5.5. Componentes de cómputo

En la tabla 7 se especifica las características de los equipos de cómputo que se encuentran en el centro de datos experimental.

Tabla 7.

Equipos de cómputo (Centro de datos experimental UDLA).

EQUIPOS DE COMPUTO				
Centro de datos experimental del campus Query				
Cantidad	**Modelo**	**Serie**	**Detalle**	**Observación**
1	Cisco USC Chasis 5108	UCS-SPL-5108-AC2	Chassis: 2 Fabric interconnect 6324	Fuentes de poder y ventiladores redundantes, conectores C19 y C20
5	Cisco UCS B200M4	UCSB-B200-M4-U	Servidor Blade 64 Gb RAM 2 CPU (6 Cores a 1.9 GHz) Tarjeta VIC 1340	Servidores Blade para aprovechar el espacio, tarjeta VIC permite virtualizar NIC y HBA según la necesidad.

Tomado de (Cisco, 2018)

3.5.5.1. Cisco UCS 5108 Blade Server Chassis

El Cisco Unified Computing System (Cisco UCS) es una plataforma de centros de datos de última generación, que permite unificar la red, almacenamiento, virtualización y computación en un único sistema diseñado para reducir el coste y aumentar la productividad del negocio.

El sistema cuenta con una red unificada 10/40 Gigabit Ethernet con una latencia baja y sin pérdidas para servidores de arquitectura x86. Por otro lado, se trata de una plataforma de múltiples chasis escalable en la que todos los equipos participan como uno y se administran unificadamente, en la figura 14 se visualiza la parte frontal del equipo UCS 5108. (Cisco, 2018).

Figura 14. Frontal Cisco UCS 5108.
Tomado de (Cisco, 2018)

Con la incorporación del sistema de computación unificada de Cisco se consigue una arquitectura de alta disponibilidad, por otro lado, la facilidad de administración y la reducción de cables en la solución UCS la hace estar disponible en un solo chasis. Esto hace posible que un solo chasis de Cisco UCS se administre de la misma forma que una solución Cisco UCS de gran tamaño lo que brinda una ventaja a empresas pequeñas o sitios remotos que lo adquieren.

En la figura 15 se presenta el sistema de computación unificada de Cisco con equipos del modelo UCS.

Figura 15. Parte frontal del equipo UCS.
Tomado de (Cisco, 2018)

2.5.5.1.1 Descripción del Cisco UCS 5108

El UCS 5108 de Cisco ofrece un chasis de servidor Blade escalable y flexible para centros de datos y ayuda a reducir el coste final de la solución.

El chasis UCS 5108 está formado por seis unidades de rack (6UR) de alto y puede montarse en un rack de 19 pulgadas estándar. Un chasis tiene la posibilidad de alojar hasta ocho servidores Blade de la serie B de Cisco UCS de ancho medio.

Del mismo modo, se puede acceder a cuatro fuentes de alimentación las cuales son intercambiables en caliente, es decir, no es necesario apagar para poder realizar el cambio. Se cuenta también con fuentes de alimentación monofásicas de 2500 W CA, 2500 W -48 VDC y 2500 W 200 -380 VDC, estas fuentes tienen una eficiencia del 94 por ciento. Adicionalmente, se puede configurar como sistema redundante N + 1 y redundante de red.

Además, en su parte trasera está equipado con ocho ventiladores que se pueden cambiar en caliente, cuatro conectores de alimentación uno por cada fuente de alimentación.

Finalmente cuenta con un plano medio pasivo de hasta 80 Gbps de ancho de banda de E / S por ranura de servidor y hasta 160 Gbps de ancho de banda de E / S para dos ranuras.

En la figura 16 se visualiza la parte trasera de un equipo UCS de Cisco.

Figura 16. Cisco UCS 5180 vista frontal y trasera.
Tomado de (Cisco, 2018)

El Cisco UCS 5108 utiliza menos componentes físicos, por otro lado, no necesita una administración independiente, también permite mayor eficiencia energética

si se compara con las infraestructuras tradicionales, es decir, se elimina la administración de chasis dedicada y de conmutadores de tipo Blade, por lo tanto, se reduce drásticamente el cableado necesario.

Además, tiene la ventaja arquitectónica de no tener la necesidad de encender y enfriar los interruptores de exceso en cada chasis.

2.5.5.1.2 Características del UCS 5108

Gestión por Cisco UCS Manager

- Reduce el costo para administrar los servidores, redes y almacenamiento debido a que se encuentra todo centralizado y su configuración recae en una sola interfaz.

Unificación

- Reducen los costos debido a que se minimizan la cantidad de tarjetas de interfaz de red, adaptadores de bus de host, interruptores y cables.

Compatibilidad

- Es compatible con equipos UCS de los diferentes modelos como el 2100, 2200 o 2300 permitiendo que el sistema se amplié sin necesidad de agregar costos.

Soporte para Cisco UCS 6324 Fabric Interconnect

- Permite la consistencia y simplicidad de una solución administrada por Cisco UCS.

Autodescubrimiento

- El chasis de Cisco reconoce y se configura de manera automática.

Plano medio de alto rendimiento

- Permite hasta 2 x 40 Gigabit Ethernet para cada ranura de servidor Blade, además proporciona ocho Blades con 1.2 Tb de rendimiento.

Configuración de cuchillas mixtas

- Permite hasta ocho servidores Blade de ancho medio o cuatro servidores de ancho completo.

Instalación simple

- No se requiere de herramientas especiales para el montaje del chasis ya que proporciona rieles de montaje para una fácil instalación.

Eficiencia del flujo de aire desde adelante hacia atrás

- Debido al chasis descubierto permite que disminuya el consumo de aire y aumenta la confiablidad de los componentes.

Monitoreo integral

- Realiza una vigilancia ambiental en cada chasis y permite el uso de umbrales de usuario para optimizar la gestión ambiental del chasis.

Servidores Blade de acoplamiento activo

- Apoya a mantener un servicio ininterrumpido durante el mantenimiento del centro de datos.

Ventiladores y fuentes de alimentación redundante e intercambiable en caliente

- Da una tasa de alta disponibilidad mediante su sistema de redundancia ya que aumenta la capacidad de trabajo y brinda un servicio ininterrumpido durante los mantenimientos. (Cisco, 2018)

3.5.5.2. Cisco UCS B200 M4

El servidor Cisco UCS B200 M4 Blade da un rendimiento óptimo gracias a la capacidad de expansión y configuración pudiendo dar cargas de trabajo desde infraestructuras de TI hasta bases de datos distribuidas.

Este tipo de servidor ofrece rendimiento, flexibilidad y optimización para centros de datos y sitios remotos, es un servidor de clase empresarial que da un rendimiento y versatilidad sin comprometer las cargas de trabajo gracias al Cisco UCS Manager. (Cisco, 2019).

En la figura 17 se visualiza un equipo UCS B200 M4 desde su parte posterior.

Figura 17. Cisco UCS B200 M4.

Tomado de (Cisco, 2019)

El servidor Cisco UCS B200 M4 Blade brinda un rendimiento óptimo gracias a la capacidad de expansión y configuración permitiendo dar cargas de trabajo desde infraestructuras de TI hasta bases de datos distribuidas.

Este tipo de servidor ofrece rendimiento, flexibilidad y optimización para centros de datos y sitios remotos, es un servidor de clase empresarial que da un rendimiento y versatilidad sin comprometer las cargas de trabajo gracias al Cisco UCS Manager.

Su arquitectura está basada en Intel Xeon en la familia de productos E5-2600 v4 y v3 que ofrecen hasta 1.5 Tb de memoria cuando se usa DIMM de 64 Gb.

Hay que destacar que gracias a su arquitectura no es necesario enfriar y encender los switches en exceso. (Cisco, 2019)

2.5.5.2.1 Descripción del Cisco UCS B200 M4

El Cisco UCS B200 M4 viene equipado con:

- Hasta 2 CPU con procesador Intel Xeon multinivel E5-2600 v4 y v3 y ofrece hasta 44 núcleos de procesamiento.
- 24 ranuras DIMM para memoria DDR4 con velocidades de hasta 2400 MHz y da hasta 1.5 Tb de memoria cuando se utiliza DIMM de 64 Gb.
- Es un servidor Blade de ancho medio adaptable a cualquier unidad de rack vertical.
- Almacenamiento de disco Local Flex Storage de Cisco que da capacidades flexibles de arranque y almacenamiento que soportan hasta NVIDIA M6 GPU.

2.5.5.2.2 Características Cisco UCS B200 M4

Cisco UCS Flex Storage technology

- Ayuda a elegir el almacenamiento Blade y el controlador de almacenamiento que necesita.

Autodescubrimiento

- No requiere configuración ya que reconoce y configura automáticamente los servidores Blade y rack.

Monitoreo extensivo

- Proporciona un amplio monitoreo ambiental para cada servidor Blade.

Tarjeta de interfaz virtual Cisco UCS 1340

- Crea hasta 256 adaptadores e interfaces PCIe independientes y funcionales sin necesidad de virtualización.
- Permite visualizar la máquina virtual desde la red física.

Adaptadores mezzanine

- Dan mayor flexibilidad y rendimiento gracias a la compatibilidad con los estándares.

Cisco Flex Flash

- Tiene ranuras para tarjetas flash SDHC dobles en la parte izquierda del servidor.

Almacenamiento local opcional

- Permite disco duros y SSD opcionales SAS, SATA o PCIe.

Procesador Intel Xeon familia de productos E5-2600 v3

- Permite mayor rendimiento tecnología de memoria DDR4
- Incorpora las ultimas características de virtualización y seguridad del procesador Intel Xeon E5-2600 v4 v v3. (Cisco, 2019)

3.5.6. Evaluación de consumo de equipos de cómputo

Para la evaluación técnica de los equipos de cómputo se necesitó especificaciones precisas correspondiente a su MIB (Base de información para

gestión) y poder realizar su monitoreo mediante el protocolo de SNMP, esto no se pudo realizar debido a que el fabricante no disponía de la base de datos o OID (identificador de objetos) específicos para la adquisición de voltajes, temperatura o potencia de estos equipos.

Debido a la dificultad de no poder obtener valores de voltaje en los equipos de cómputo, se trabajó en otra solución la cual fue realizar el monitoreo del sistema de alimentación interrumpida (UPS) en donde se registra la carga total de voltaje de los equipos cuando existe un corte o problema eléctrico en la infraestructura.

Con la configuración del UPS se puede tener un consumo más preciso de los equipos y en sí de toda la infraestructura del centro de datos pudiendo solo con este equipo realizar el análisis de consumo eléctrico, por lo cual esta solución es más que factible.

3.5.7. Evaluación de consumo eléctrico de los quipos UCS 5108

Los equipos de cómputo no entraron al monitoreo por las restricciones técnicas especificadas por el proveedor, sin embargo, para el análisis se tomó las especificaciones técnicas detalladas en su ficha técnica, en la tabla 8 se visualiza los parámetros de voltaje de entrada las cuales servirán de revisión.

Tabla 8.

Consumo estándar de voltaje de equipos UCS 5108.

Fuentes de alimentación		Fuente de alimentación de CA	Fuente de alimentación de -48V DC	Fuente de alimentación de 200 a 380V DC
	Voltaje de entrada	100 a 120V AC 200 a 240 VCA	-40 a -62V DC	200 a 380V DC
	Potencia de salida máxima	1300 vatios (W) de 100 a 120 V de entrada	2500W	2500W
	Frecuencia	50 a 60 Hz	-	-
	Eficiencia	94%	92%	94%

Tomado de (Cisco, 2019)

3.5.8. Equipo de almacenamiento

En la tabla 9 se especifica las características de los equipos de almacenamiento que se encuentran en el centro de datos experimental.

Tabla 9.

Equipos de almacenamiento (Centro de datos experimental UDLA).

EQUIPO DE ALMACENAMIENTO				
Centro de datos experimental del campus Query				
Cantidad	Modelo	Serie	Detalle	Observación
1	VNXe3200	V32D12AN5PS6	Controladoras redundantes Tres discos SD con 100 Gb Seis Discos SAS de 300 Gb Seis discos SAS de 1.2 Tb	

Tomado de (EMC, 2019)

3.5.8.1. EMC VNXe3200

El sistema de almacenamiento VNXe3200 de EMC es una solución de almacenamiento destinada para pequeña y mediana empresa, tiene un diseño simple y está enfocado para negocios con infraestructura física y para negocios que quieren pasar a la virtualización. Además, cuenta con una capacidad de hasta 1000 usuarios.

En la figura 18 se visualiza la parte frontal de equipo DELL EMC Data Domain que se encuentra instalado en el centro de datos experimental de la universidad de las Américas.

Figura 18. Frontal del equipo EMC VNXe3200.
Tomado de (IBM, 2013)

VNXe3200 permite el uso de protocolos iSCSI, CIFS, NFS y Fibre Channel, lo que permite que sea compatible con la mayoría de las infraestructuras de red en la actualidad.

Su administración es mediante Unisphere que cuenta con una interfaz sencilla e intuitiva facilitando la configuración.

Los beneficios que presente este equipo de almacenamiento son expuestos a continuación.

2.5.8.1.1 Características EMC VNXe3200

- Alta disponibilidad: Gracias a la redundancia de hardware y la protección RAID.
- Snapshots unificados: Tiene la capacidad de tomar Snapshots de punto en el tiempo de datos de bloques o archivos sin usar espacio de almacenamiento adicional.
- Rendimiento y eficiencia: Usa de mejor manera el almacenamiento flash gracias a las tecnologías MCx y FASTVP.
- Replicación en bloques: Con lo que se garantiza la redundancia y la integridad de datos.
- Compatibilidad de conectividad ampliada: VNXe3200 admite dispositivos ópticos, Fibre Channel y conectividad mediante cobre.
- Integración de VMware: Permite la integración completa con los hosts de VMware vCenter y ESXi.
- Almacenamiento simple y unificado eficiente: VNXe3200 da un aprovisionamiento de almacenamiento unificado en un único rack.

- Plataforma de almacenamiento compacta: Tienen un diseño de procesador de almacenito de dos controladores que se instalan en solo 2U de un rack. (EMC, 2019).

En cuanto a hardware el VNXe3200 está equipado con Intel Xeon E5 a 2.2 GHz con una arquitectura MCx que garantiza el uso correcto de los cores, esto se visualiza en la figura 19 que muestra la parte frontal del equipo y sus diferentes enclosure.

Figura 19. Vista frontal del DPE/DAE de 12 unidades.
Tomado de (EMC, 2019)

En su diseño se puede apreciar que cuenta con indicadores LED que señalizan tanto la perdida de potencia como los fallos que puedan existir en el chasis. Además, todos los VNXe3200 cuentan con dos procesadores de almacenamiento, que se pueden extraer de manera individual lo que ayuda al mantenimiento y reemplazo. En la figura 20 se puede observar un procesador de almacenamiento.

Figura 20. Procesador de almacenamiento (Parte superior extraída).
Tomado de (EMC, 2019)

3.5.9. Evaluación de consumo de equipos de almacenamiento

Igual que los equipos de cómputo, para ingresar al monitoreo se necesita las especificaciones correspondientes a OIDS que registren los voltajes de entrada y salida de los equipos, de acuerdo al proveedor de estos equipos estos datos no se lograron encontrar por temas de firmware, por lo cual para el análisis de datos servirá con el monitoreo del UPS que contiene todos los datos de voltaje de la infraestructura del centro de datos.

Como se detalló anteriormente mediante la configuración del UPS se puede tener un consumo más preciso de los equipos y en sí de toda la infraestructura del centro de datos permitiendo solo con este equipo realizar el análisis de consumo eléctrico, por tal razón esta solución es viable.

Al finalizar de este capítulo se detalla la gestión que se realizó para configurar el UPS y poder ingresarle al monitoreo mediante el protocolo SNMP.

3.5.10. Evaluación de consumo eléctrico de los equipos de almacenamiento

Los equipos de almacenamiento no entraron al monitoreo por las restricciones sobre las especificaciones de sus MIB correspondiente a voltajes, sin embargo, el análisis se lo realizará mediante los datos estándares que se tienen en el ficha técnica del equipo.

En la tabla 10 se muestras los datos de consumo eléctrico del equipo VNXe3200.

Tabla 10*.*

Valores estándares de equipo VNXe3200.

Requerimiento	**VNXe3200 (3.5” Drives)**	**VNXe3200 (2.5” Drives)**	**VNXe3200 (12*3.5” Drives)**
AC Line Voltage	100 to 240 V Ac 10%, single-phase, 47 to 63 Hz	100 to 240 V Ac 10%, single-phase, 47 to 63 Hz	100 to 240 V Ac 10%, Single-phase, 47 to 63 Hz

Tomado de (Cisco, 2018)

3.5.11. Sistema de alimentación ininterrumpida (UPS)

UPS, la fuente de suministro eléctrico que posee el centro de datos experimental de la universidad de las Américas, es el equipo que brinda alimentación mediante el uso de su batería con el fin de seguir dando energía a los dispositivos en el caso de interrupción eléctrica.

El UPS de modelo AP9215RM fabricado por Schneider Electric es el único UPS que proporciona energía a los equipos instalados en el centro de datos experimental, una vez se pierda energía, por ende, configurarlo en un entorno de monitoreo ayuda a conocer el consumo de equipos y validar su consumo eléctrico.

En la tabla 11 se muestras las características del UPS instalado en el centro de datos experimental.

Tabla 11*.*

Características del equipo UPS.

EQUIPO DE ALIMENTACIÓN				
Centro de datos experimental del campus Query				
Cantidad	Modelo	Serie	Detalle	Observación
1	AP9215RM		UPS - APC	
			APC Symmetra LX 16kVA Escalable to 16kVA N+1 Rack-mount, 208/240V	

Tomado de (Schneider Electric, 2019)

3.5.12. Evaluación técnica del consumo eléctrico de UPS (Sistema de alimentación ininterrumpida)

Para realizar una mejor evaluación técnica de los equipos de red, cómputo y almacenamiento se consiguió configurar e instalar la herramienta de monitoreo zabbix para extraer información actual del consumo que pasa por el UPS el cual abastece de energía cuando existe un corte o falla eléctrica.

A continuación, se detalla la configuración especial que se realizó al UPS (Sistema de alimentación ininterrumpido),

3.5.13. Evaluación de consumo eléctrico del UPS (sistema de alimentación ininterrumpida).

Para la evaluación del consumo de este equipo se tuvo que realizar diferentes configuraciones a nivel de capa física en su direccionamiento, así como la habilitación del protocolo de monitoreo SNMP.

El equipo no contaba con un direccionamiento físico a nivel de dirección IP, tampoco con una ficha de administración por lo cual se procedió a consultar sus datos técnicos a través de la página oficial de APC marca oficial de equipos que brindan confiabilidad en potencia critica.

- Se procedió a descargar el software de gestión de equipos de marca APC Device IP Configuration Utility, este software realiza una consulta mediante protocolo ARP, realiza un escaneo a nivel de red buscando la dirección MAC de los equipos APC.
- Encontrada la MAC del equipo se procede a configurar la dirección IP, mascara de red y puerta de enlace, en la figura 21 se visualiza el descubrimiento de MAC del UPS mediante escaneo de MAC.

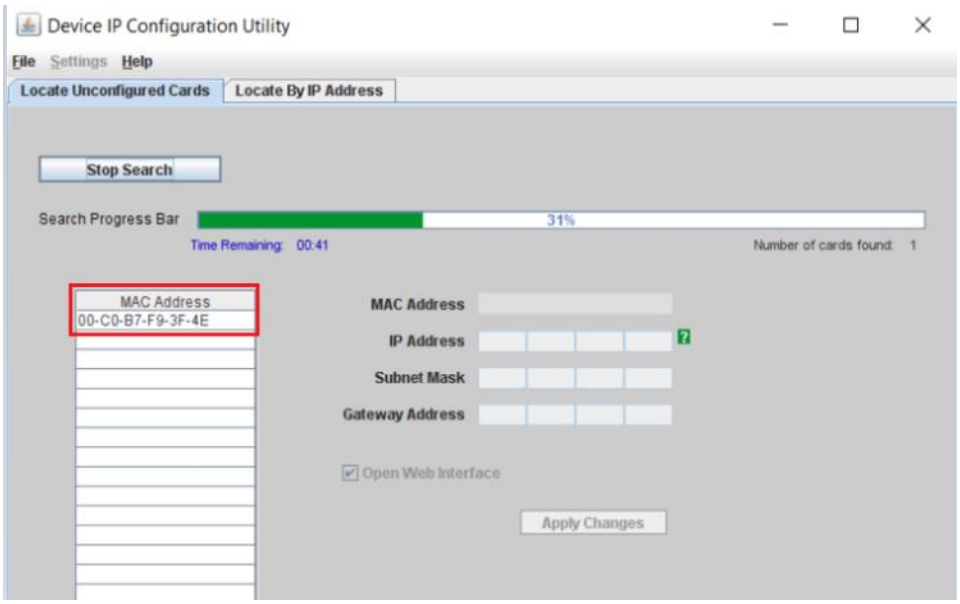


Figura 21: Escaneo de red mediante protocolo ARP.
Tomado de software APC, sf

En la figura 22 se muestra la primera asignación de IP al UPS para poder ingresar a su módulo de gestión vía navegador.

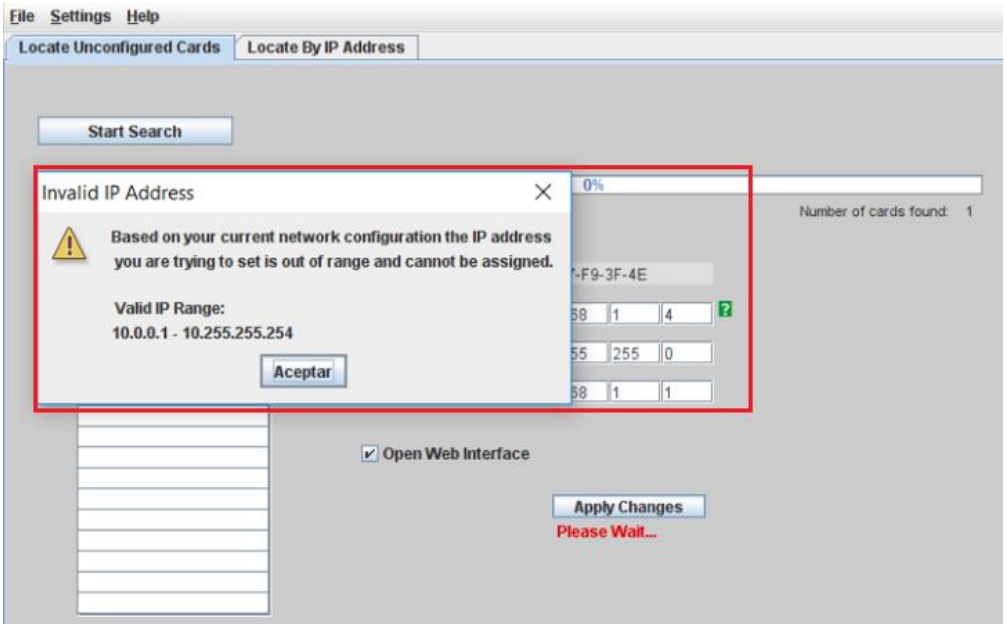


Figura 22. Configuración de la dirección IP.
Tomado de software APC, sf

En la figura 23 se verifica la pantalla de ingreso del UPS y la figura 24 muestra la dirección IP asignada al equipo, todo como parte de la configuración necesitada para el análisis de consumo eléctrico.

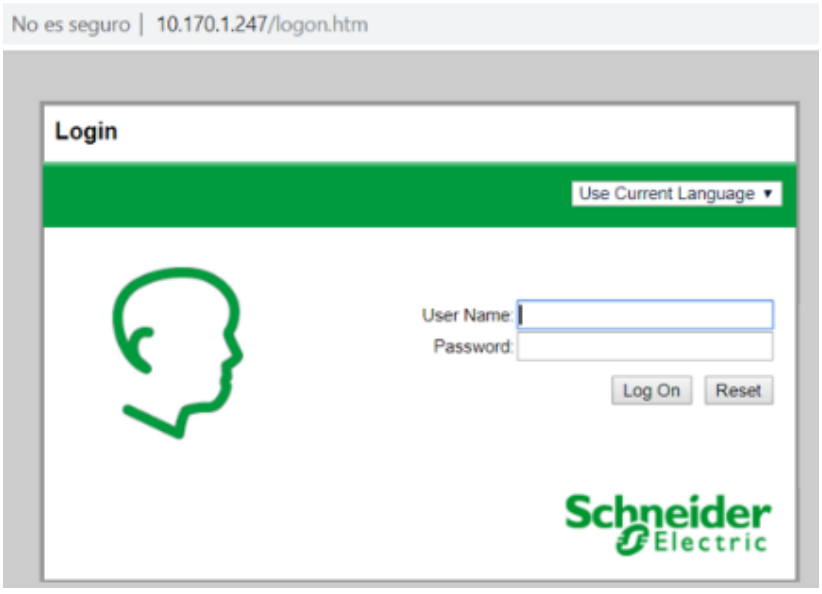


Figura 23. Pantalla de ingreso UPS AP9215RM.
Tomado de software APC, sf

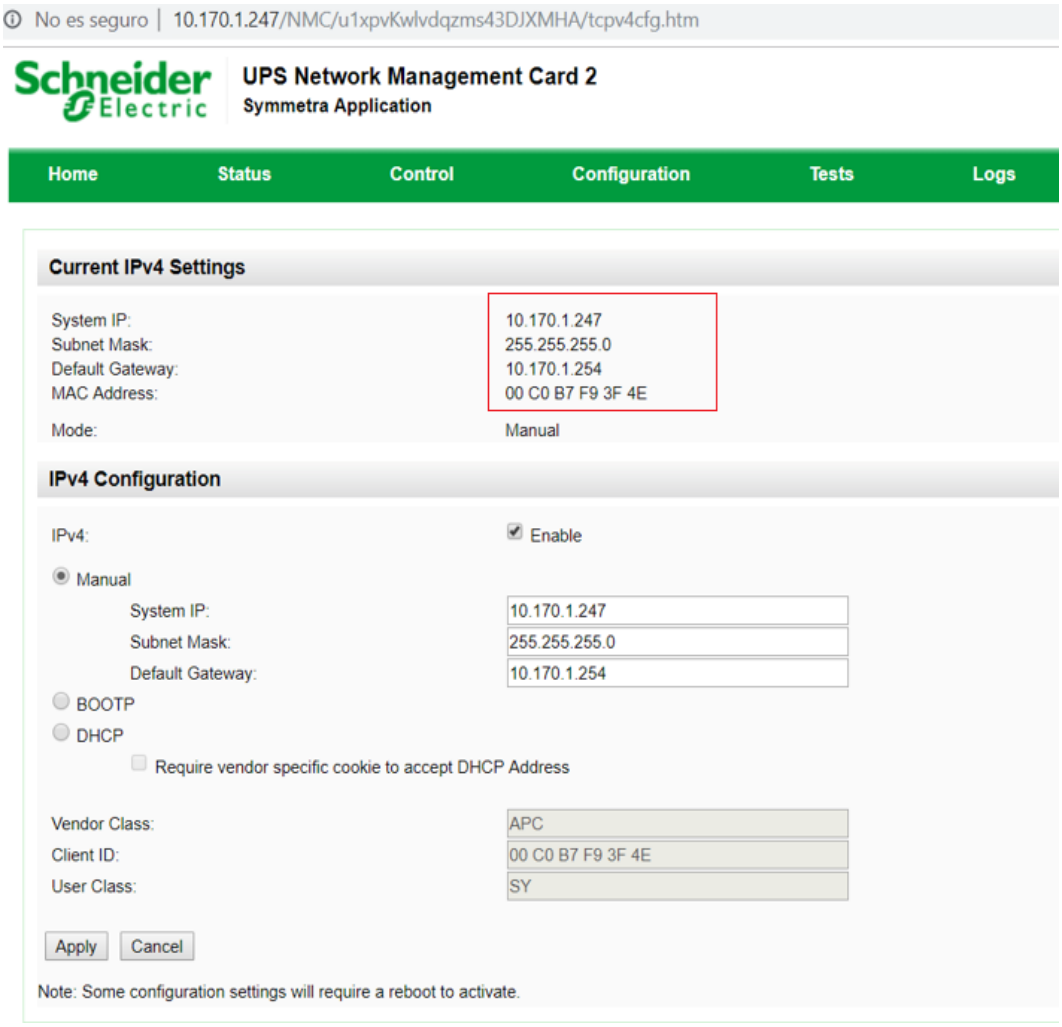


Figura 24. Página principal UPS APC, direccionamiento IP.
Tomado de (Schneider Electric, 2019)

- Se habilito el direccionamiento a nivel de red, mediante una asignación de dirección IP y mascara de red.
- Se habilito el protocolo de monitoreo SNMP, ingresando al UPS mediante su plataforma web, se marcó el protocolo de monitoreo SNMP y se ingresó la comunidad public, esto se realizó mediante una configuración manual desde la API.

En la figura 25 se visualiza la configuración realizada para habilitar el servicio de monitoreo mediante el protocolo SNMP y se procedió asignarle el nombre public como comunidad SNMPv1.

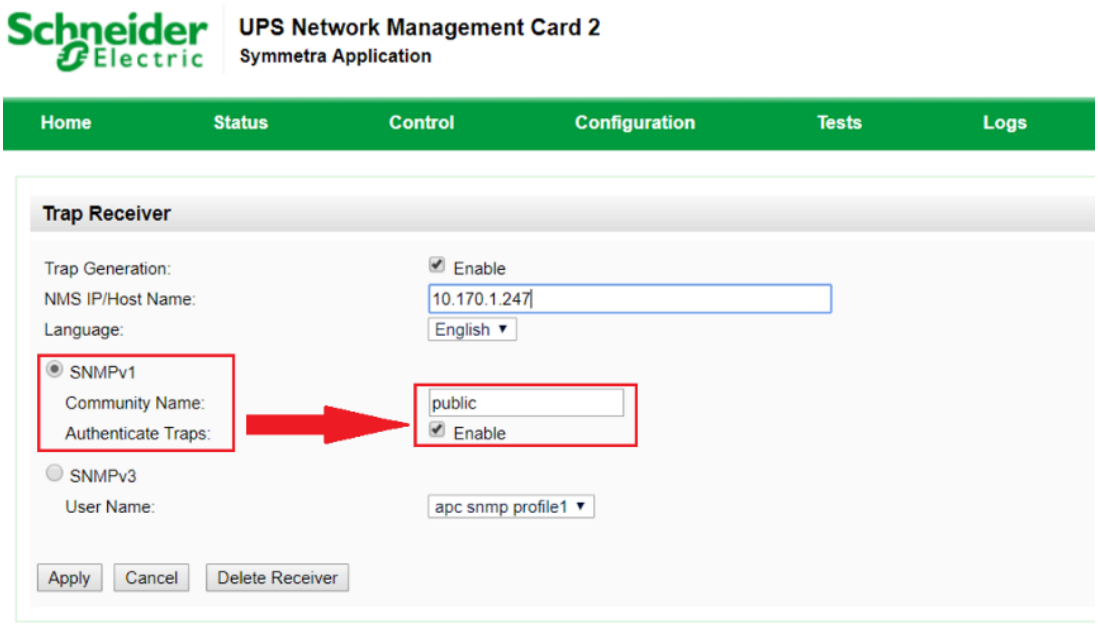


Figura 25. Habilitación del monitoreo SNMP.
Tomado de (Schneider Electric, 2019)

En la figura 26 se visualiza los parámetros de monitoreo habilitados y se verifica mediante el servidor de monitoreo zabbix que los mismo están disponibles para empezar con la recopilación de datos.

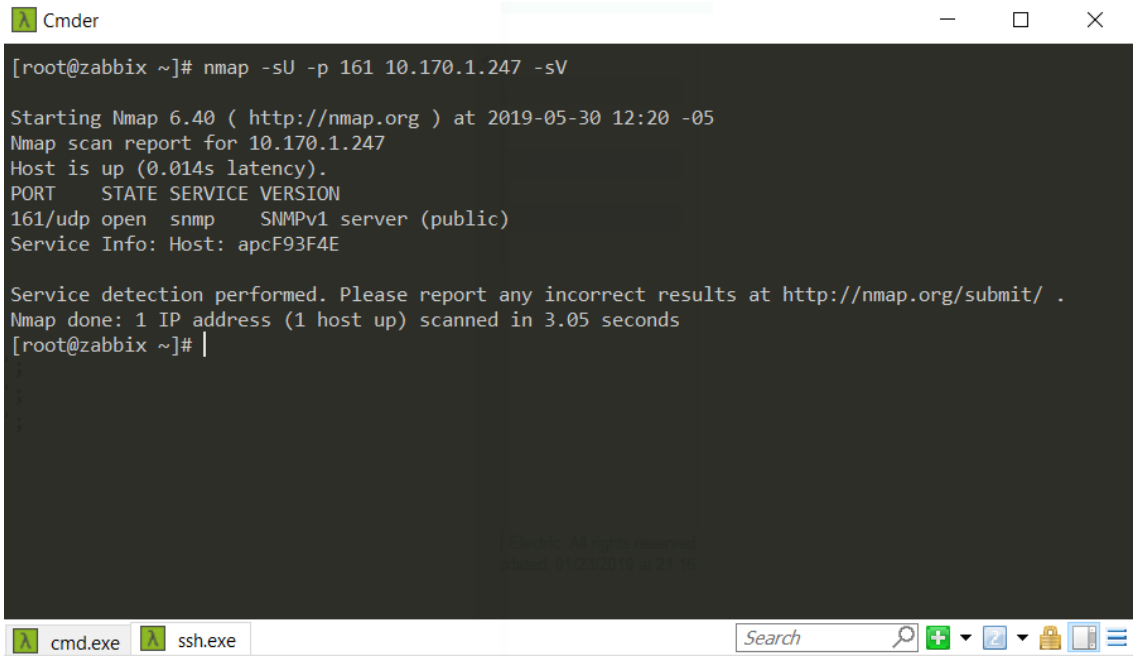


Figura 26. Parámetros del UPS configurados mediante zabbix.

En la figura 27 se puede verificar que los equipos y en especifico el UPS APC con dirección IP 10.170.1.247 ya se encuentra habilitado en el monitoreo mediante SNMP.

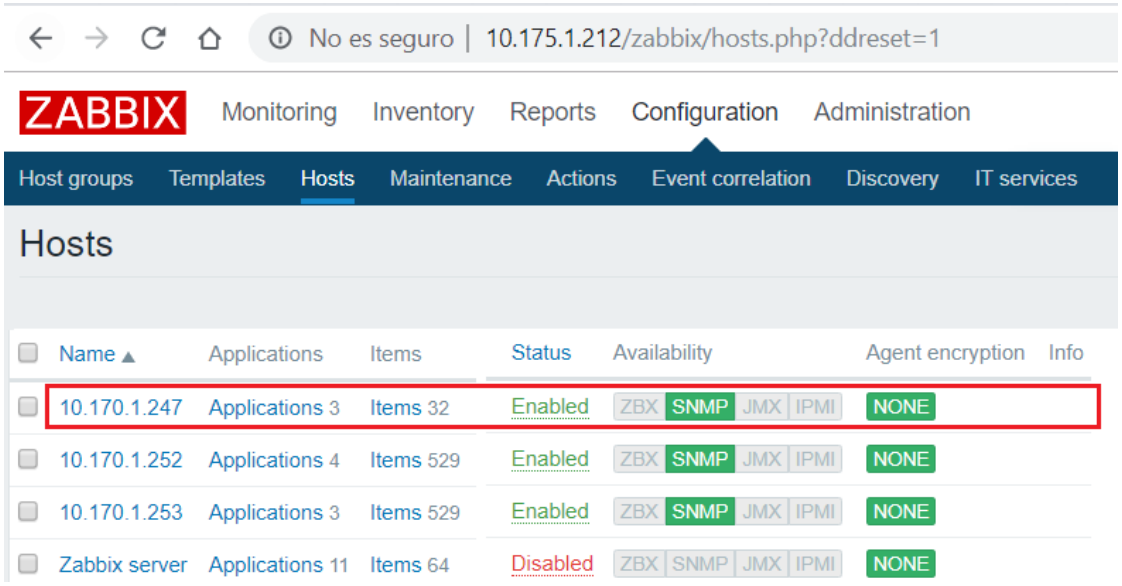


Figura 27. Protocolo SNMP habilitado UPS.

En la figura 28 se visualiza los parámetros de voltaje de entrada y de salida del UPS, los cuales se configuraron en el servidor de monitoreo zabbix, con estos datos se podrán realizar los análisis de consumo y verificar posibles datos de eficiencia o perdía de consumo eléctrico.

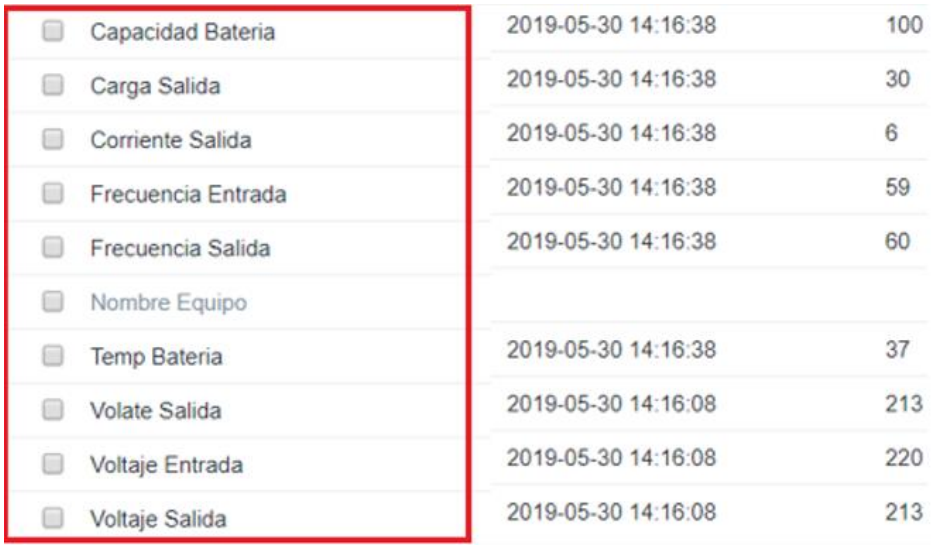


Figura 28. Parámetros configurados para el monitoreo del UPS.

Finalmente, con las directrices realizadas en este capítulo se puede realizar el monitoreo de los equipos que cuenten con el protocolo SNMP habilitado y toda

información que se adquiera puede procesarse en el ambiente de Big Data, a fin de gestionar la información mediante un sistema de visualización de datos.

4. CAPÍTULO IV. ANALIZAR LAS VENTAJAS DE UTILIZAR TECNOLOGÍA HADOOP

4.1. Hadoop

Es un sistema de código abierto que se utiliza para almacenar, procesar y analizar grandes volúmenes de datos, Hadoop también es un entorno de tipo High Performance que se puede escalar horizontalmente con hardware relativamente barato conocido como Commodity Hardware.

Uno de los puntos más fuertes de Hadoop es que está diseñado para ejecutarse en servidores de bajo coste y que dispone de una gran tolerancia a fallos, debido al crecimiento simple de nodos y está diseñado para escalar desde unos pocos a cientos de máquinas.

El entorno de Hadoop es funcional en cualquier solución de Big Data debido a su gran comportamiento escalable y sobre todo es adaptable a la lógica de negocio y almacenamiento a nivel local y distribuido.

Hadoop funciona sobre un entorno que suministra librerías open source para la computación distribuida y desde su versión 2.0 realiza el uso de varios componentes o módulos los cuales se pueden visualizar de mejor forma en la figura 29. (Fernández, 2017).

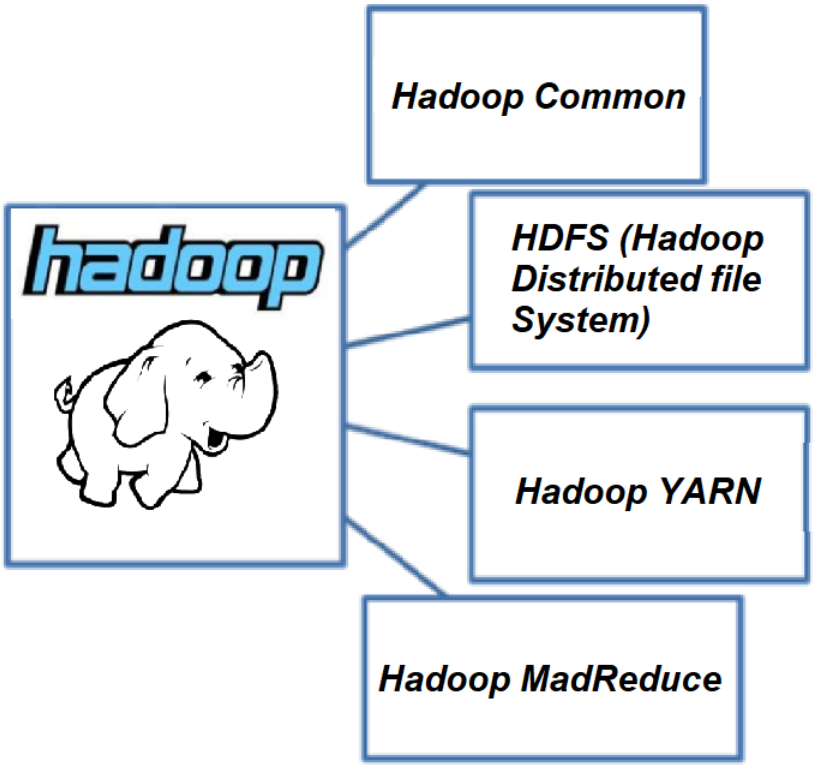


Figura 29. Tecnología módulos equipos.
Tomado de (Hadoop Apache, 2018)

Hadoop permite crear aplicaciones para procesar grandes volúmenes de información distribuida a través de un modelo de programación sencillo.

Está diseñado para ser escalable debido a que trabaja con almacenamiento y procesamiento local (pero distribuido), de forma que funciona tanto para clústeres de un solo nodo como para los que estén formados por cientos. Otra característica especial de Hadoop es la detección de errores a nivel de aplicación, permitiendo gestionar los fallos en los distintos nodos y ofreciendo un buen nivel de tolerancia a errores.

El desarrollo de Hadoop ha sido tal que la mayoría de las implementaciones de MapReduce con sistemas distribuidos usan Hadoop como base. En la actualidad existe un buen número de estas distribuciones comercializadas por las compañías más importantes del sector tecnológico.

Hadoop está construido básicamente de dos módulos, los cuales se detallan a continuación.

- Hadoop Distributed File System (HDFS): el sistema de ficheros sobre el que se ejecutan la mayoría de las herramientas que conforman el ecosistema Hadoop.
- Hadoop MapReduce: el principal Framework de programación para el desarrollo de aplicaciones y algoritmos.

Actualmente existen tres versiones de Hadoop 1.0, 2.0 y 3.0 que están siendo usadas por las distintas distribuciones. Las dos últimas versiones tienen diferencias en su arquitectura y se las detallará en esta sección, las mismas que son las más utilizadas hoy en día, la versión 1.0 actualmente está en desuso debido a su carencia de soporte y sobre todo al desarrollo de las últimas dos versiones.

En este capítulo se detalla la tecnología Hadoop y como base al proyecto de tesis se especificará la versión 3.0, de igual manera se realizará un comparativo con la versión 2.0.

4.1.1. Hadoop 2.0

La segunda versión de Hadoop parte con la base de Hadoop 1.0 añade y modifica algunas características de sus módulos para tratar de resolver algunos de los problemas que tenía y mejorar el rendimiento del sistema.

El proyecto Hadoop 2.0 está dividido esta vez en cuatro módulos:

- Hadoop Common
- Hadoop Distributed File System (HDFS)
- Hadoop YARN: un Framework para la gestión de aplicaciones distribuidas y de recursos de sistemas distribuidos.
- Hadoop MapReduce: el sistema de procesamiento principal, que esta vez se ejecuta sobre YARN.

Para este comparativo se especificará la versión 2.9.2 la última versión estable antes de la liberación de la 3.0. (Hadoop Apache, 2018)

4.1.2. HDFS 2.0

Los cambios introducidos en HDFS de la versión 1.0 han sido pocos, pero significativos. Se permite combatir la principal debilidad de la primera versión: el NameNode como punto de fallo único en el sistema. Esto evita que un sistema

HDFS tenga alta disponibilidad, debido a que un fallo en el NameNode hace que el sistema deje de funcionar. Otra novedad introducida es la HDFS Federation que permite tener múltiples espacios de nombres en HDFS.

El resto de las características que ofrece la primera versión de HDFS se han mantenido prácticamente intactas, desde la arquitectura con NameNode y DataNode a la monitorización a través de una interfaz web o la ejecución de comandos por consola.

4.1.2.1. Alta disponibilidad

La alta disponibilidad del NameNode se puede conseguir de dos maneras: a través del Quorum Journal Manager o usando Network File System.

Quorum Journal Manager

En este tipo de arquitectura se configura, aparte del NameNode principal, un segundo NameNode que está en modo espera o Standby, llamado precisamente Standby NameNode. Este servicio permanece inactivo a la espera de un fallo en el NameNode activo, que es el encargado de realizar las tareas de gestión y administración del sistema.

Para mantener la coherencia de los datos entre los dos NameNodes y mantenerlos sincronizados se crea un grupo de servicios, llamados JournalNodes, cuya función es la de actuar como diario de todas las operaciones que el NameNode activo va realizando. Este conjunto de JournalNodes se llama Quorum Journal Manager.

El funcionamiento de un sistema HDFS con Quorum Journal Manager es el que se muestra en la Figura 30. El NameNode comunica a un grupo de JournalNodes (no hace falta que lo haga con todos ya que entre ellos se sincronizan) todos los cambios que se van realizando en el sistema de ficheros, es decir, en los DataNodes.

El Standby NameNode, por su parte, va leyendo el estado del sistema a través de los JournalNodes de manera que cuando se produce un evento de fallida pueda actuar rápidamente como NameNode activo.

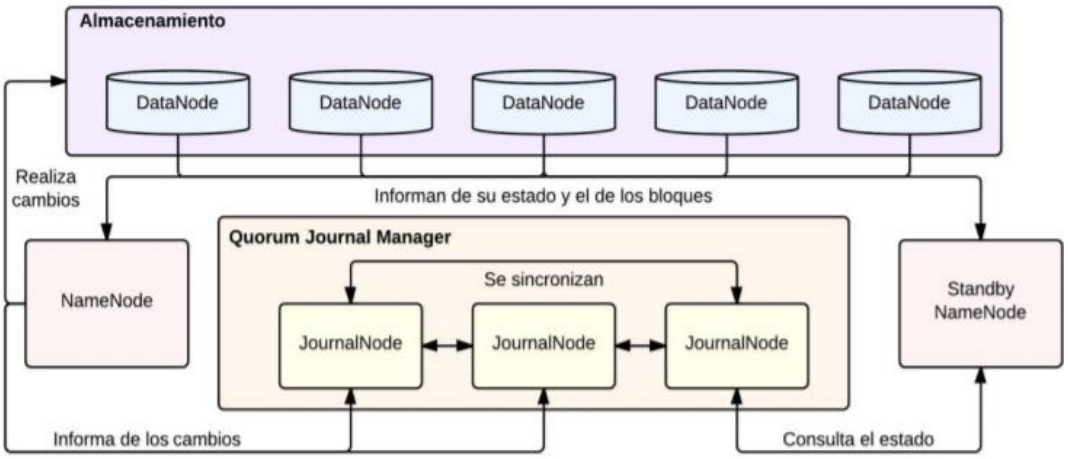


Figura 30. Esquema de los servicios de HDFS Quorum Journal Manager. Tomado de (Common, 2016)

El uso del Quorum Journal Manager es totalmente transparente para el usuario y la transición del Standby NameNode a servicio NameNode principal puede ser configurada de forma automática o manual.

Para que esta transición sea automática se requiere de una herramienta como ZooKeeper que sirve para detectar los fallos en el NameNode y monitorizar su estado.

Network File System, este método para lograr la alta disponibilidad en HDFS también cuenta con un NameNode principal y un Standby NameNode, realizando las mismas funciones que en el caso del Quorum Journal Manager.

La diferencia principal es que en lugar de usar un Quorum para la sincronización entre los NameNodes se utiliza un dispositivo de almacenamiento conectado mediante un Network File System. NFS es un protocolo de sistemas de ficheros en red que permite a diferentes ordenadores acceder a ficheros en remoto.

4.1.2.2. HDFS Federación

Se determina que es una división por capas y explicación simplificada de la arquitectura de HDFS, generalizando para cualquier sistema de ficheros, se describiría de la siguiente forma.

- Espacio de nombres: es la parte lógica del sistema y la que el usuario más acostumbra a ver. El conjunto de directorios, ficheros y bloques, así como

su estructura jerárquica. El espacio de nombres de HDFS permite las operaciones para crear, borrar, modificar y conseguir la localización de los bloques.

- Servicio de almacenamiento: es la parte que se podría conocer como física a nivel de software está compuesta por dos partes:
  - Administración de bloques: realizado por el NameNode.
  - Almacenamiento: realizado por el DataNode.

La primera versión de HDFS trabajaba con un espacio de nombres único e igual al de un sistema de ficheros Unix (con la excepción de la operación para conseguir los bloques). Con la aparición de HDFS Federation esto cambia, ahora se puede tener varios espacios convirtiendo HDFS en escalable a nivel horizontal.

4.1.3. MapReduce 2.0

MapReduce 2.0 o también conocido como MRv2 es la capa que más cambios ha sufrido con la segunda versión de Hadoop. Se ha mantenido todas las características que identifican a MapReduce a nivel de usuario, pero se ha renovado por completo su arquitectura y los servicios que la componen como por ejemplo YARN el cual lo detallamos a continuación.

4.1.3.1. Arquitectura YARN

YARN es un motor de gestión de recursos y aplicaciones o procesos distribuidos y es la principal adición de Hadoop 2.0. Actualmente solo lo usa MapReduce, pero en un futuro puede ser usado para otros Framework para gestionar las aplicaciones.

La principal idea detrás de YARN es la de hacer una gestión óptima de los recursos de un clúster a la hora de realizar un proceso MapReduce.

Hasta el momento el JobTracker realizaba dos funciones destacadas: la gestión de recursos y la planificación y monitorización de los trabajos. Con YARN, el JobTracker deja paso a dos servicios nuevos que se encargan cada uno de una de estas tareas y aparece también el NodeManager. Los servicios trabajan sobre container, un concepto abstracto que se utiliza para agrupar los diversos

recursos de un sistema como procesadores, memoria, disco, red, etc. Toda en una misma arquitectura.

- ResourceManager: es un servicio global para todo el clúster y que se encarga de gestionar los recursos del sistema. Este servicio está formado por dos partes:
  - Scheduler: Responsable de asignar los recursos a cada aplicación, está diseñado para realizar tareas de planificación, por lo que no se encarga de monitorizar ni relanzar aplicaciones caídas. Se encarga de dividir los recursos del clúster a través de diferentes colas, aplicaciones.
  - Applications Manager: se responsabiliza de aceptar las peticiones de creación de trabajos y asignar un Application Master a cada una de ellas.
- Application Master: es un servicio único para cada aplicación MapReduce. Su principal cometido es el de negociar con el Resource Manager la gestión de los recursos y la de trabajar con los NodeManager para ejecutar y monitorizar los trabajos.
- NodeManager: es un agente que se ejecuta en cada máquina y es el responsable de los containers. Se encarga de gestionar los recursos asignados al container y reportar su estado al Scheduler.

4.1.4. Hadoop 3.0

El lanzamiento de Hadoop 3 en diciembre del 2017 marcó el comienzo de una nueva era para la ciencia de datos. El marco de Hadoop es el núcleo de todo el ecosistema de Hadoop y el eje principal de este proyecto de tesis.

A continuación, se realizará la comparación de Hadoop 3 con Hadoop 2. También se explica las diferencias entre Hadoop y Apache Spark para un mejor entendimiento de la decisión de utilizar Hadoop para este proyecto, en la figura 31 de visualiza la evolución de Hadoop y las áreas donde se podría aplicar esta tecnología.

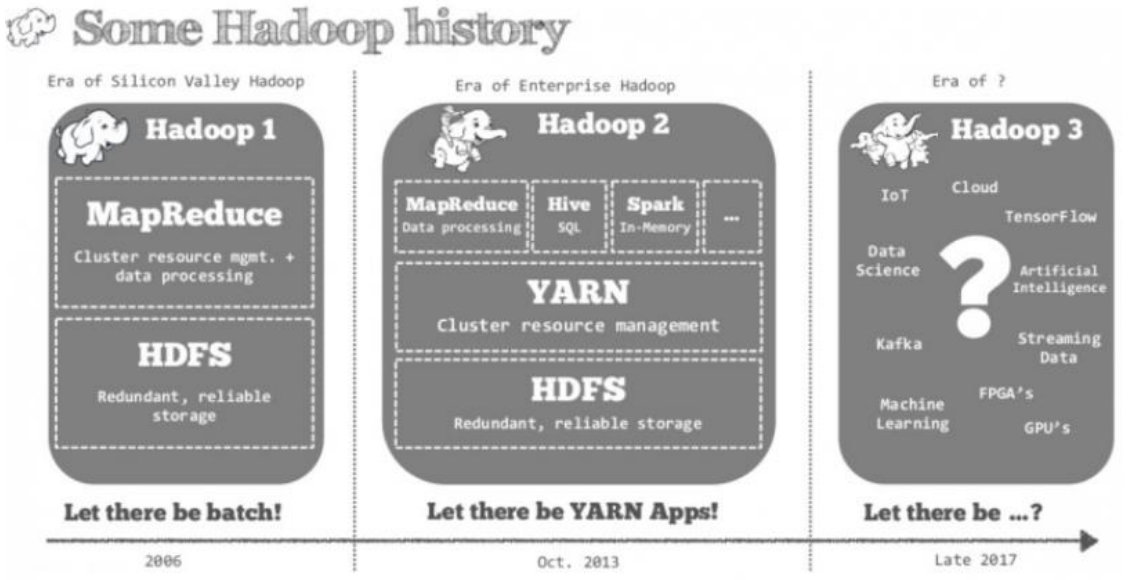


Figura 31. Evolución de Hadoop y sus áreas de uso.
Tomado de (Toppertips, 2018)

Comparación en procesamiento (Real Time)

Hadoop 2 y Hadoop 3 son motores de procesamiento de datos desarrollados en Java y lanzados en 2013 y 2017 respectivamente. Hadoop se creó con el objetivo principal de mantener el análisis de datos de un disco, conocido como procesamiento por lotes. Por lo tanto, Hadoop nativo no admite el análisis en tiempo real y la interactividad.

Hadoop es utilizado de mayor forma para procesamiento en Batch, es decir el mejor uso que se le puede dar es para procesar grandes cantidades de información historizada, esto no quiero decir que no pueda trabajar en tiempo real, claro ejemplo es que Apache Spark se basa en Hadoop para el proceso Batch y utiliza Streaming para procesamiento en tiempo real, que en concepto puede ser, pero en la práctica es una carga en información en el tiempo de manera incremental.

Para el desarrollo de este proyecto viene bien trabajar con Hadoop debido a que la información que se analizará es historizada y el monitoreo se lo realiza mediante otra aplicación la cual si trabaja en tiempo real.

Spark 2.X es un motor de procesamiento y análisis desarrollado en Scala y lanzado en 2016. El análisis en tiempo real de la información se estaba volviendo

crucial, debido a que muchos servicios de internet gigantescos confiaban en la capacidad de procesar los datos de inmediato.

En consecuencia, Apache Spark fue creado para el procesamiento de datos en vivo y ahora es popular porque puede manejar de manera eficiente las transmisiones en vivo de información y procesar datos en un modo interactivo.

Para el proyecto de tesis la información será procesada una vez sea almacenada en una instancia de ficheros (HDFS), por ende, no se necesita tener la información en tiempo real lo que se busca es mantener la información integra y bien tratada mediante una gran tolerancia a fallos por lo cual en este aspecto Hadoop en versión 3 es la mejor opción por alcance y robustez.

Nivel de abstracción y aprendizaje

Se podría decir que una de las principales diferencias entre estas tecnologías es el nivel de abstracción que es bajo para Hadoop y alto para Spark. Es decir, Hadoop es más difícil de aprender y usar, ya que los desarrolladores deben saber cómo codificar muchas operaciones básicas. Hadoop es solo el motor central, por lo que el uso de una funcionalidad avanzada requiere complementos de otros componentes, lo que hace que el sistema sea en un momento más complejo, pero más robusto, todo dependiendo de la necesidad y experiencia que se tenga.

A diferencia de Hadoop, Apache Spark es una herramienta completa para el análisis de datos. Tiene muchas funciones de alto nivel integradas útiles que operan con el conjunto de datos distribuido resistente (RDD).

Spark ya cuenta con diferentes librerías que ayudan a una mejor gestión de los datos, una de esas es Spark SQL que puede utilizarse para realizar consultas SQL, sin tener que acudir a otras herramientas de consulta de datos.

De acuerdo con lo planteado, Spark tiene una línea de aprendizaje más corta y se podría decir que es más simple de implementar debido a que todos sus componentes vienen incorporados, en cambio Hadoop es el motor central para el procesamiento de datos y para cada entorno como la extracción de datos, carga, manipulación o visualización se puede utilizar diferentes herramientas.

Mediante lo mencionado anteriormente se debe decir que, al tratarse de un proyecto de investigación e implementación, Hadoop es la tecnología ideal debido a que permite acoplar diferentes herramientas para cada acción, lo cual implica generar conocimiento y en su complejidad demostrar las habilidades de investigación y desarrollo adquiridas por el estudiante sobre nuevas tecnologías.

Ventajas

Agilidad

- La contenedorización de Hadoop 3 brinda agilidad y la incorporación de tecnología Docker que permite crear aplicaciones robustas, agiles y compatibles en diferentes escenarios.

Carga de almacenamiento

- Con la codificación de borrado en Hadoop 3, existen 6 bloques, tres más a los por defecto en la versión 2.0, los cuales contarán con 3 bloques para la paridad, lo que se traducirá en una menor sobrecarga de almacenamiento.
- Esto permite reducir a la mitad el costo de almacenamiento de HDFS al tiempo que conserva la durabilidad de los datos y la sobrecarga de almacenamiento se puede reducir del 200% al 50%.

Escalabilidad y disponibilidad

- Hadoop 2 y Hadoop 1 solo usan un único NameNode para administrar todos los espacios de nombres. Hadoop 3 tiene múltiples NameNodes para múltiples espacios de nombres, lo que mejora la escalabilidad.
- En Hadoop 2, solo hay un NameNode en espera. Hadoop 3 soporta múltiples NameNodes en espera. Si un nodo en espera se apaga durante el fin de semana, tiene el beneficio de otros NameNodes en espera para que el clúster pueda continuar operando, esta característica le da una condición de servicio más larga.
- Hadoop 2 usa un servicio de línea de tiempo antiguo que tiene problemas de escalabilidad. Hadoop 3 mejora el servicio de línea de tiempo v2 y mejora la escalabilidad y confiabilidad del servicio de línea de tiempo. (Hortonworks, 2018)

4.1.5. Hadoop Common

Es un módulo que proporciona las herramientas en java que son necesarias para los sistemas informáticos de los usuarios en Windows, Linux, Unix, entre otros, también se le puede otorgar el concepto de base tipo Framework donde están todas utilidades más comunes en las que se apoyan los distintos módulos de Hadoop. Lo cual permite proporcionar abstracciones a nivel de sistema de archivos o sistema operativo. Contiene los archivos .jar (Java Archive) y scripts necesarios para leer los datos almacenados en el sistema de archivos Hadoop. (Common, 2016)

4.1.6. Hadoop MapReduce

Hadoop MapReduce: Es un sistema basado en YARN que permite el procesamiento en paralelo de grandes volúmenes o conjuntos de datos.

MapReduce en concepto de experiencia se puede definir mencionando que lleva el nombre de las dos operaciones básicas que lleva a cabo este módulo: leer datos de la base de datos y colocarlos en un formato adecuado para el análisis y así realizar operaciones matemáticas, como un ejemplo se podría decir, contar el número de ventas de más de 5 años en una base de datos de facturas.

Para entender de mejor forma el concepto de MapReduce se debe conocer que Hadoop se ejecuta sobre la máquina virtual de Java (Java Runtime Enviroment (JRE). Las secuencias de scripts de inicio (start) y parada (stop) se ejecutan mediante SSH (Secure Shell) en los nodos que forman el clúster, la comunicación inicia o se establece mediante protocolo OSI TCP/IP.

Se debe tener en cuenta que para que tome el nombre de un sistema Hadoop, este debe correr en varias máquinas físicas o virtuales que a su vez deben estar conectadas entre sí, las mismas deben almacenar grandes archivos de datos.

Los sistemas de archivos deben proporcionar su ubicación (la del rack y, principalmente, la del switch), el termino rack en una solución de Big Data con Hadoop se asemeja al término utilizado en centro de datos, es decir el rack es el que contiene los scripts o DataNode (nodo de datos) de cada nodo, como si fuera un equipo en una unidad de rack en términos de networking y el switch viene a cumplir la función de conectar los diferentes nodos, cumple la misma función que

la realiza en una arquitectura de red, para un mejor entendimiento de conceptos visualizar la figura 32.

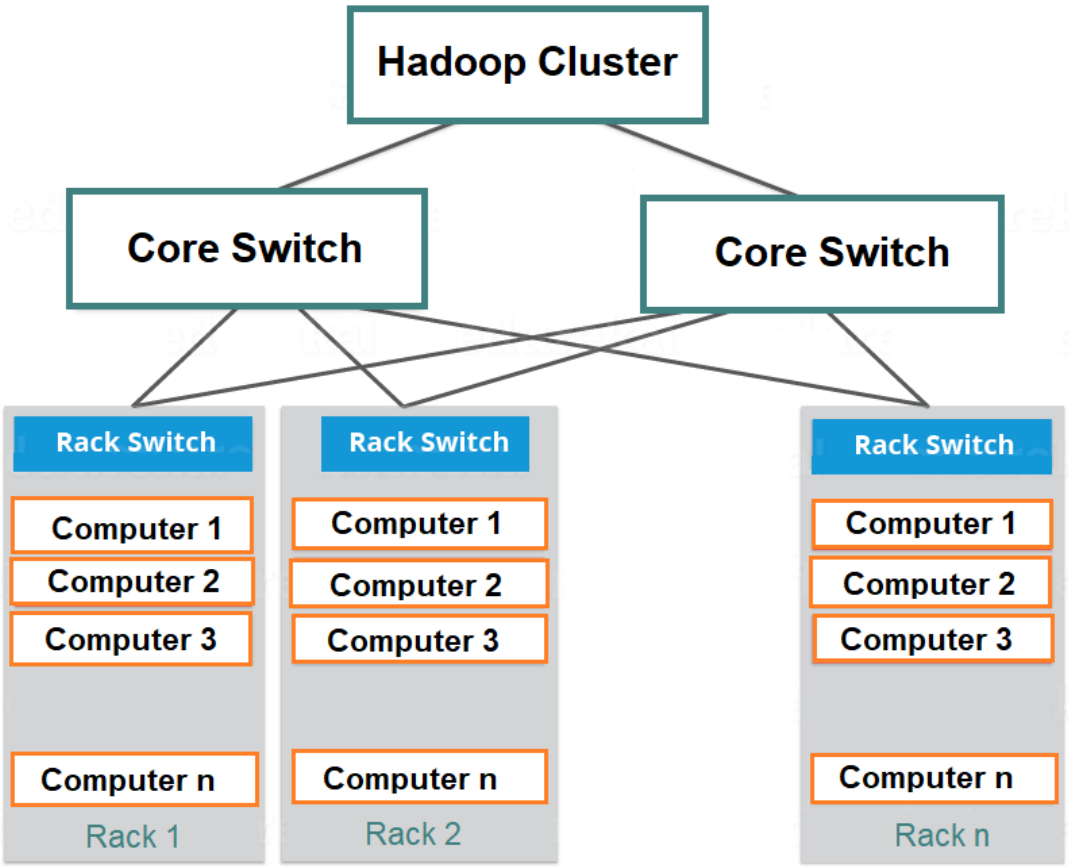


Figura 32. Funcionamiento Clúster Hadoop.
Tomado de (Hortonworks, 2018)

Las aplicaciones Hadoop pueden utilizar información para ejecutar trabajos en el nodo donde se encuentran los datos, reduciendo el tráfico de red troncal. HDFS duplica los datos para tratar de mantener diferentes copias en diferentes racks aumentando la tolerancia a fallos.

Para entender y conocer de mejor forma el módulo MapReduce, se debe adentrar en el funcionamiento de un clúster Hadoop en el cual encontramos:

MasterNode (nodo maestro). En el MasterNode encontramos un JobTracker (rastreador de trabajos), un TaskTracker (rastreador de tareas) y un NameNode (nodo de nombres) y un DataNode (nodo de datos).

4.1.7. Hadoop Distributed File System (HDFS)

En principio se debe entender que HDFS se ejecuta a través de un servidor NameNode dedicado para alojar el índice de sistema de archivos y un NameNode secundario destinado a generar instantáneas de estructuras de memoria del NameNode principal. Elaborándolo de esa forma, se evita la corrupción del sistema de archivos y la reducción de pérdida de datos.

Por otro lado, un servidor JobTracker independiente puede ejecutar la planificación de tareas.

En los clústers de Hadoop encontramos un solo NameNode (con opciones de redundancia por su criticidad) y un grupo de DataNodes, que sirven bloques de datos, permitiendo fiabilidad mediante la réplica de datos a través de múltiples hosts.

Los nodos de datos pueden hablar entre sí para balancear los datos, mover copias y mantener una réplica alta de datos, por otro lado, existe comunicación entre JobTracker y TaskTracker de manera que el primero establece un MAP (formato adecuado para el análisis) o un Reduce al segundo como tarea mediante el conocimiento de la ubicación de los datos.

La funcionalidad de MapReduce se puede entender como un JobTracker envía las tareas al nodo TaskTracker intentando mantener el trabajo tan cerca de los nodos como sea posible todo para mejorar los niveles de atenuación en cada ejecución del proceso MapReduce.

La tarea del JobTracker es conocer en todo momento qué nodos contienen información y cuáles están cerca, por lo que, si es preciso, puede enviar la información a los nodos adyacentes en el mismo rack, si el nodo actual no puede almacenar los datos, el TaskTracker envía la tarea a una JVM (Java Virtual Machine) y cada poco tiempo informa al JobTracker de su estado.

Si existe un fallo de un TaskTracker, el JobTracker asigna la tarea desde donde se dejó, es decir, cada TaskTracker contiene un número de espacios disponibles llamados ranuras o slots que es ocupado por un MAP o un Reduce, la arquitectura se la puede visualizar en la figura 33 para su mejor entendimiento.

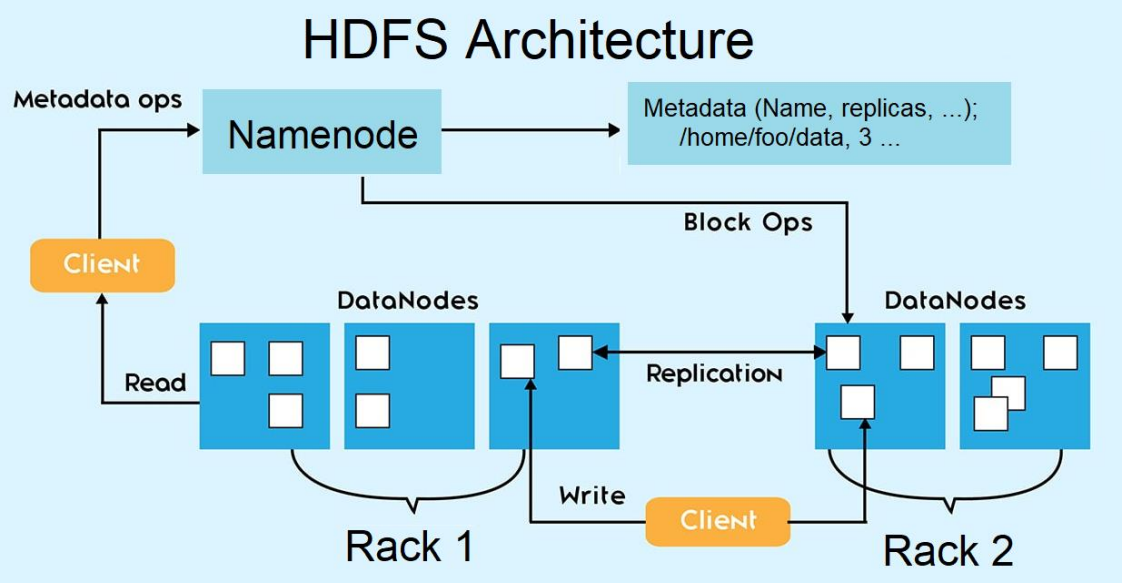


Figura 33. Arquitectura HDFS.

Tomado de (Toppertips, 2018)

4.1.8. Hadoop YARN

Es un Framework o marco de trabajo para la planificación de tareas y gestión de recursos del clúster, o también conocido como administrador de clúster, el cual gestiona los recursos de los sistemas que almacenan los datos y ejecutan el análisis, el cual es una característica clave de la segunda generación de la versión Hadoop 2 en adelante.

YARN se caracteriza actualmente como un sistema operativo distribuido, a gran escala, para aplicaciones de Big Data. Inicio de una reescritura de software que desacopla las capacidades de gestión de recursos y planificación de MapReduce del componente de procesamiento de datos, permitiendo a Hadoop soportar enfoques más variados de procesamiento, y una gama más amplia de aplicaciones.

Un ejemplo actual es que los clústers Hadoop ahora pueden ejecutar consultas interactivas y transmisiones de aplicaciones de datos de forma simultánea con los trabajos por lotes de MapReduce.

YARN combina un administrador central de recursos que reconcilia la forma en que las aplicaciones utilizan los recursos del sistema de Hadoop, con los agentes de administración de tareas que monitorean las operaciones de procesamiento de nodos individuales del clúster. (Iglesias, 2014).

4.1.9. Alegación de la tecnología Hadoop para el proyecto

Mediante lo descrito en el desarrollo del capítulo se tiene una definición real de las ventajas y funcionalidades del clúster bajo tecnología Hadoop y se puede comprender de mejor forma el funcionamiento de esta solución de Big Data.

Como se lo menciono en las guías del capítulo la nueva versión de Hadoop 3.0 permite trabajar con procesos Streaming que sería conceptual o ideal para el procesamiento en tiempo real, sin embargo, lo fuerte de esta solución se debe a que se puede acoplar a cualquier herramienta de procesamiento, analítica o visualización de datos todo dependiente de la experiencia y dominio del arquitecto que arme la solución para la necesidad que requiera.

Finalmente, este proyecto de tesis adapta la solución Big Data con tecnología Hadoop para el análisis de datos que se adquirieron en el centro de datos experimental de la UDLA los cuales serán procesados bajo Batch es decir con información almacenada.

5. CAPÍTULO V. IMPLEMENTACIÓN DE UNA INSTANCIA DE BIG DATA

En este capítulo se detalla las herramientas de software que se utilizó en la instancia de Big Data y como se utiliza el ambiente para almacenar información correspondiente al consumo eléctrico.

5.1. Herramienta de virtualización

Virtual box es un software de código abierto, bajo licencia GPL, utilizado para crear máquinas virtuales, en la figura 34 se visualiza las tres máquinas virtuales creadas para el ambiente de Big Data, requisitos básicos para la instalación son los siguientes.

Tabla 12.

Requerimiento de herramienta de virtualización Virtual Box.

Componentes	**Requerimientos**
Sistema operativo	Windows, Linux, Mac, Solaris, etc.
Arquitectura	x86 – x64
Memoria RAM	512 MB o superior

Disco Duro	5 GB Libre
Procesados	Intel – AMD de 2600 MHz
Sistema Anfitrión	Windows Server 2013

Tomado de (Virtualbox, 2019)

En la figura 34 se visualiza los nodos principales para el ambiente de Big Data.

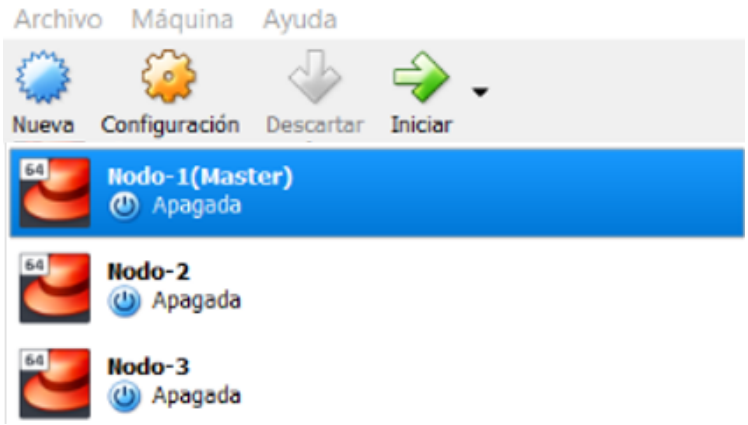


Figura 34. Instalación de VM en Virtual Box:

Tomado de (Virtualbox, 2019)

5.2. Instalación sistema operativo CentOS

CentOS es un sistema operativo descendiente de la familia de Linux Red Hat en su versión Community, este sistema operativo fue la base para implementar el clúster de Hadoop.

Se instalaron tres máquinas virtuales, nombradas nodo 1, 2 y 3 respectivamente, la máquina virtual nodo1 controla todo el clúster del ambiente de Hadoop, el mismo que está conformado por los dos nodos esclavos y el nodo maestro.

Requisitos de instalación para cada nodo Hadoop dentro de la solución de Big Data.

Tabla 13.

Requisitos para cada nodo Hadoop.

Componentes	**Requerimientos**
Software Virtualización	Virtual Box 5.2.6
Arquitectura	x64 Bits

Sistema operativo	CentOS 7
Memoria RAM	4 GB por virtual machine
Disco Duro	40 GB por virtual machine
Procesador	Intel x64 2.4 GHz o superior
Socket	1 con dos núcleos

Tomado de (Virtualbox, 2019)

En la figura 35 se visualiza la pantalla de instalación del sistema operativo CentOS 7.

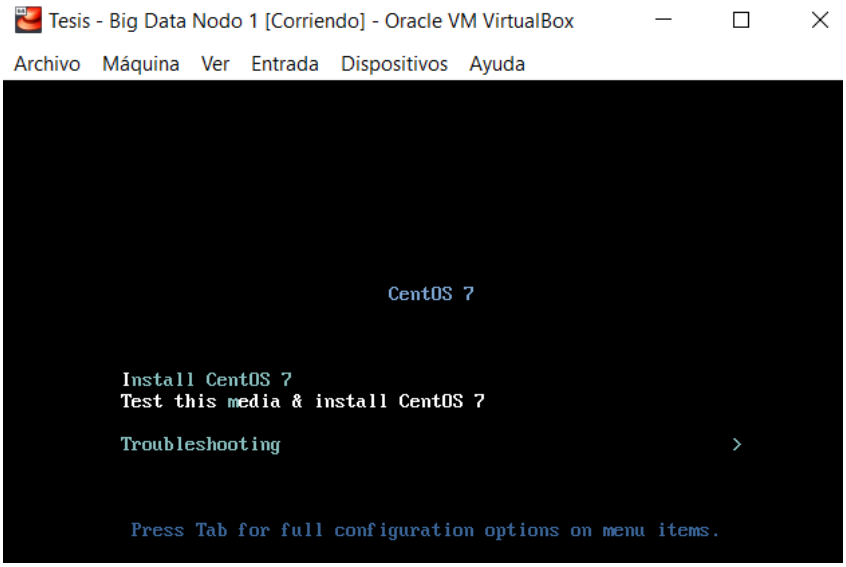


Figura 35. Instalación sistema operativo CentOS.

5.3. Instalación y configuración del entorno de Hadoop

El ambiente de Big Data debe ser tolerante a fallos, por lo cual debe proporcionar alta disponibilidad, y de acuerdo con el concepto de Big Data esto se puede asegurar mediante la arquitectura de un clúster con mínimo tres nodos, un nodo maestro y dos nodos esclavos, esto permite adicional a la alta disponibilidad un gran performance y gran escalabilidad.

En este proyecto de tesis que está enfocado a la demostración de funcionalidad de un ambiente de Big Data, se implementó un clúster compuesto de dos nodos esclavos y un nodo maestro para la gestión y procesamiento de datos, a continuación, se describe la arquitectura del ambiente de Hadoop mediante la

configuración realizada para el sistema de datos HDFS, el sistema de procesos YARN y la estructura del clúster con los tres nodos.

5.3.1. Configuración Hadoop 3.0

En cada nodo se descargó Hadoop 3.0 en su versión estable, una vez descargado el software se procedió a descomprimir el archivo en el directorio creado como entrada, esto se realiza en cada nodo del ambiente Hadoop.

- Descomprimimos

  # tar xvf /home/Hadoop/Descargas/Hadoop-3.2.0.tar.gz

- Creamos un directorio en donde almacenamos los servicios y ficheros de Hadoop 3.0 y copiamos todo el contenido del archivo descomprimido.

  # mkdir /tmp/entrada

  # cp /opt/Hadoop/etc/Hadoop/*.xml /tmp/entrada/

Creado el directorio que contiene todos los ficheros, servicios y procesos de Hadoop se empieza con la configuración de las variables de sesión de Java y Hadoop en cada máquina, esto permite arrancar los servicios en cada nodo y mandar a ejecutar cada servicio desde el nodo maestro.

5.4. Configuración de variables de entorno

- Instalación de la máquina virtual de Java JDK
  #rpm ivh jdk-8u211-linux-x64.rpm
- En cada nodo se instaló el JDK de Java mediante el paquete rpm
- En cada nodo se procedió a configurar las variables de entorno del software Java.

En la figura 36 se visualiza la instalación de los paquetes JDK de Java en su versión estable 8.0.

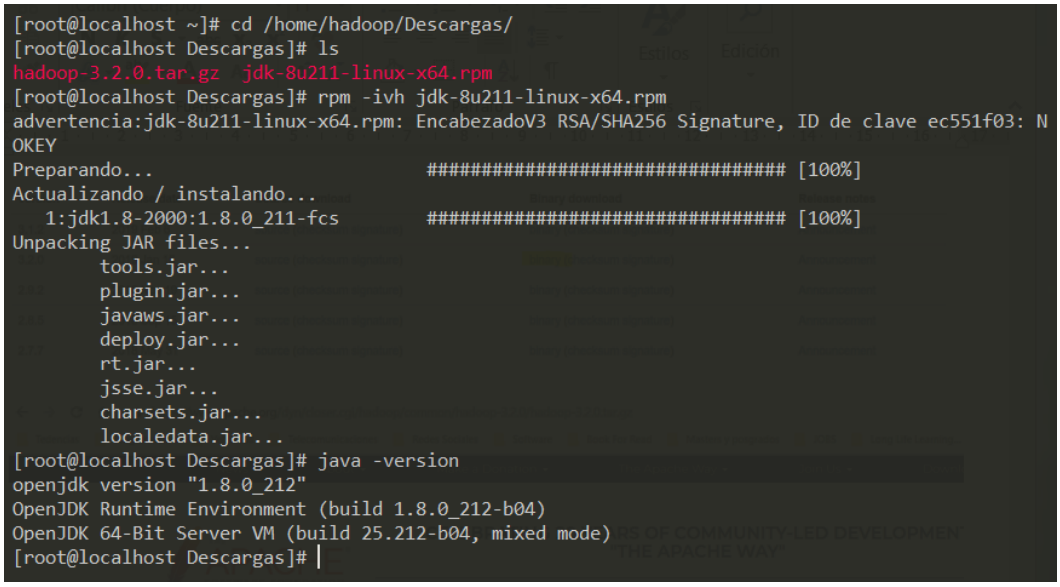


Figura 36. Instalación del Java Runtime.

- Configuración de variables de sesión del Java y Hadoop en el nodo maestro.

  # vi. bashrc

  Añadimos la dirección de las librerías de java y Hadoop

  export HADOOP_HOME=/opt/Hadoop

  export JAVA_HOME=/usr/java/jdk1.8.0_211-amd64

  PATH=$PATH:HADOOP_HOME/bin:$HADOOP_HOME/sbin

5.5. Sistema de ficheros HDFS

HDFS es la parte de almacenamiento de datos de Hadoop, conocido como el sistema de ficheros, es tolerante a fallos y almacena gran cantidad de datos, trabaja de forma incremental y lo más destacado es que puede sobrevivir a fallos de hardware sin perder datos.

Para este proyecto de tesis, HDFS gestiona el almacenamiento en el clúster, dividiendo los ficheros en bloques y almacenando copias duplicadas a través de los dos nodos, en la figura 37 se ilustra como la información se distribuye en cada nodo, en nuestro caso 3 nodos un maestro y dos esclavos.

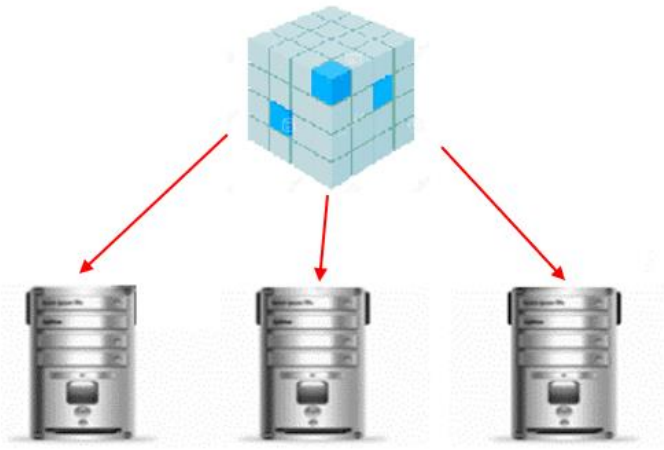


Figura 37. La información distribuida en tres nodos.

En HDFS siempre vamos a tener un nodo que actúa como máster o principal, el cual no tiene datos, solo contiene Metadatos es decir tiene la información de administración de cómo el clúster está formado, así como la información de cuáles son los nodos esclavos, en nuestro proyecto tenemos un nodo maestro y dos esclavos desde el maestro se controla a los dos nodos esclavos.

En la figura 38 se ilustra la arquitectura del proyecto para el modo de almacenamiento HDFS.

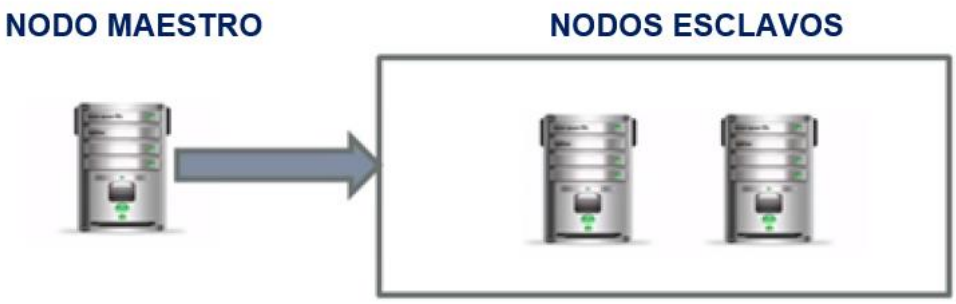


Figura 38. Arquitectura Hadoop para HDFS.

5.5.1. Configuración de ficheros HDFS

Todos los nodos del clúster deben tener la configuración del HDFS, y una parte esencial es la configuración de los directorios que contendrán tanto los ficheros de ejecución como de administración del clúster.

- NameNode: en este directorio se almacenan todas las librerías de administración del clúster que se crearon al momento de ejecutar el servicio

de HDFS, en este proyecto que tiene un clúster con tres nodos solo se necesita un NameNode para gestionar el ambiente de Hadoop.

- DataNode: en este directorio se almacenan los resultados que se obtengan al ejecutar las diferentes funciones del HDFS cada nodo esclavo tiene creado este directorio.

Para este proyecto el nodo maestro llamado nodo1 tiene creado el repositorio NameNode que sirve para guardar las directrices que permiten administrar de clúster.

En la figura 39 se visualiza creado el repositorio NameNode en el nodo maestro llamado nodo1.

Figura 39. Directorio NameNode en el nodo maestro.

Para este proyecto los dos nodos esclavos tienen creado el repositorio DataNode, en donde se guardan los resultados obtenidos al ejecutar las funciones del HDFS.

En la figura 40 se puede observar los directorios creados en el nodo2 y nodo3 respectivamente.

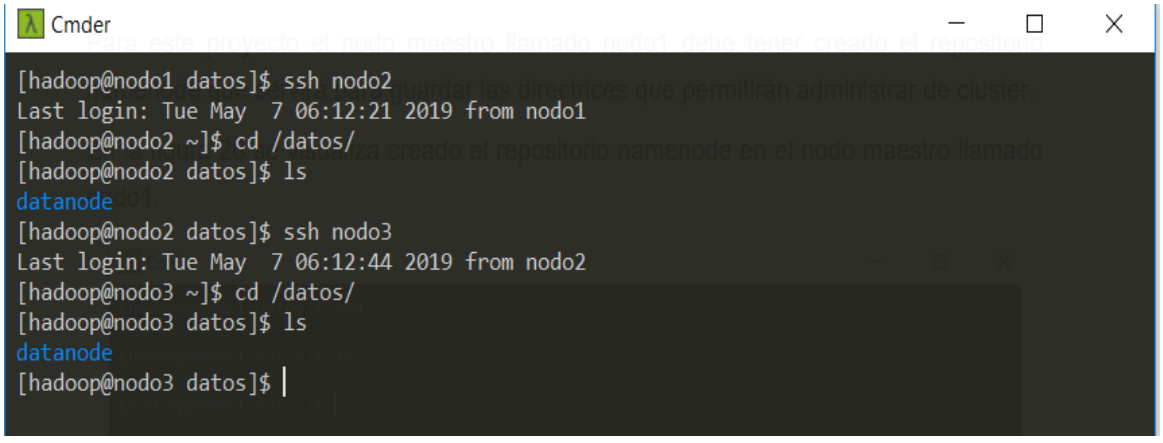


Figura 40. Directorios DataNode creados para el nodo2 y nodo3.

Una vez creado el NameNode y los DataNode se continúa con la configuración del HDFS el cual debe ser replicado para los dos nodos restantes.

Para configurar el HDFS se modificó dos archivos específicos, los descritos a continuación.

- Core-site.xml
- Hdfs-site.xml

Accedemos al directorio que contiene las dos librerías mencionadas, en este caso configuramos el core-site.xml

cd /opt/Hadoop/etc/Hadoop

vi core-site.xml

Asignamos el nombre del nodo maestro, el cual administra el sistema de archivos como se muestra en la figura 41.

```
<configuration>
        <property>
                <name>fs.defaultFS</name>
                <value>hdfs://nodo1:9000</value>
        </property>
</configuration>
```

Figura 41. Configuración de core-site.xml.

Configuramos el fichero hdfs-site.xml asignando los valores del clúster, añadimos la dirección del NameNode que es el repositorio que contiene la administración del clúster y esta creado en el nodo principal y los DataNode que corresponden al repositorio donde se guardan los datos cuando estos se procesan, los mismos que están creados en cada nodo esclavo.

En la figura 42 se visualiza la configuración del número de réplicas que en nuestro caso es dos y la dirección del NameNode y los DataNode.

Cmder

```
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
        <property>
                <name>dfs.replication</name>
                <value>2</value>
        </property>
        <property>
                <name>dfs.namenode.name.dir</name>
                <value>/datos/namenode</value>
        </property>
        <property>
                <name>dfs.datanode.data.dir</name>
                <value>/datos/datanode</value>
        </property>
</configuration>
```

Figura 42. Configuración del fichero hdfs-site.xml.

Realizada la configuración de los archivos para el HDFS, se crearon los Metadatos que son las directrices para que el clúster conozca y funcione con la arquitectura planteada.

En la figura 43 visualizamos como se crearon los Metadatos en el DataNode mediante la ejecución del siguiente comando.

hdfs namenode -format

Cmder

```
[hadoop@nodo1 namenode]$ pwd
/datos/namenode
[hadoop@nodo1 namenode]$ ls
current
[hadoop@nodo1 namenode]$ ls current/
edits_0000000000000000001-0000000000000000002  edits_0000000000000000225-0000000000000000234
edits_0000000000000000003-0000000000000000004  edits_0000000000000000235-0000000000000000236
edits_0000000000000000005-0000000000000000006  edits_0000000000000000237-0000000000000000238
edits_0000000000000000007-0000000000000000008  edits_0000000000000000239-0000000000000000240
edits_0000000000000000009-0000000000000000010  edits_0000000000000000241-0000000000000000242
edits_0000000000000000011-0000000000000000012  edits_0000000000000000243-0000000000000000259
edits_0000000000000000013-0000000000000000014  edits_0000000000000000260-0000000000000000261
```

Figura 43. Creación de los Metadatos.

Finalmente, levantamos el servicio del HDFS mediante la siguiente instrucción.

start-dfs.sh

En la figura 44 se puede observar el servicio HDFS corriendo sobre el servidor mediante su servicio web.

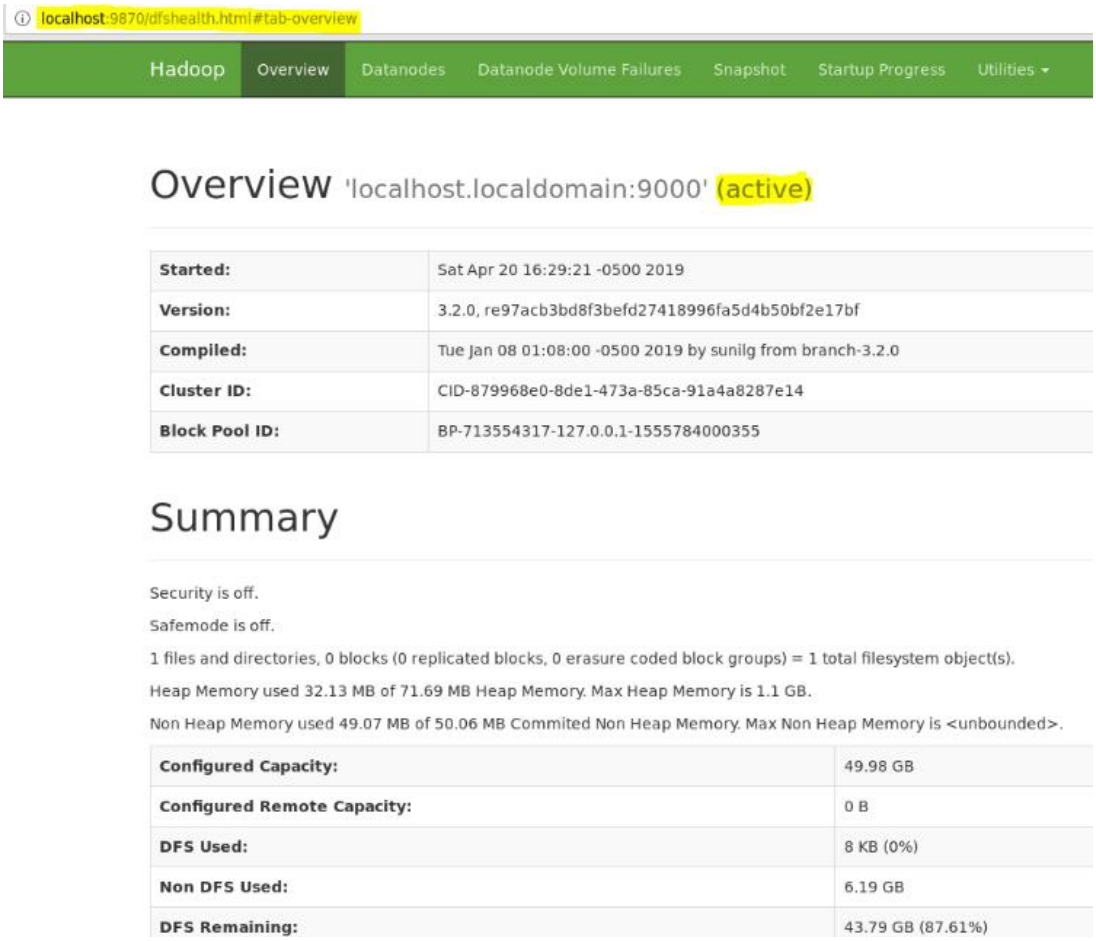


Figura 44. Servicio HDFS ejecutándose.

5.6. Configuración del sistema de procesos YARN

En Hadoop YARN permite la gestión de recursos y aplicaciones distribuidas dónde se pueden implementar múltiples aplicaciones de procesamiento de datos totalmente personalizadas, en el proyecto se implementó YARN para realizar el procesamiento de datos extraídos desde la herramienta de monitoreo zabbix.

Para iniciar se configuraron dos archivos que se ubicaron en el directorio

$ cd /opt/Hadoop/etc/Hadoop/

$ vi MapReduce-site.xml

El primer archivo de configuración es MapReduce-site.xml este define el Framework que trabaja para el procesamiento de datos que en este proyecto es MapReduce, el archivo quedo configurado de la siguiente manera.

<configuration>

```
    <property>
        <name>MapReduce.framework.name</name>
        <value>yarn</value>
    </property>
</configuration>
```

El segundo archivo es MapReduce-site.xml en este se configuro el ResourceManager quien administra las peticiones de los procesos en los diferentes nodos, para la versión 3.0 de Hadoop se añadió la dirección de distintas librerías de YARN las cuales sirven para que al compilar el servicio no arroguen excepciones de java correspondientes a seguridad y compatibilidad con la versión 3.0.

El archivo queda configurado de la siguiente manera.

```
$ vi MapReduce-site.xml
<property>
    <name>yarn.resourcemanager.nodo1</name
    <value>nodo1</value>
</property>
<property>
    <name>yarn.nodemanager.aux-services</name>
    <value>MapReduce_shuffle</value>
</property>
<property>
    <name>yarn.nodemanager.aux-services.MapReduce_shuffle.class</name>
    <value>org. apache.Hadoop.mapred.ShuffleHandler</value>
<property>
<property>
```

```
    <name>yarn.application.classpath</name>
  <value>
    /opt/Hadoop/etc/Hadoop,
    /opt/Hadoop/share/Hadoop/common/*,
    /opt/Hadoop/share/Hadoop/common/lib/*,
    /opt/Hadoop/share/Hadoop/hdfs/*,
    /opt/Hadoop/share/Hadoop/hdfs/lib/*,
    /opt/Hadoop/share/Hadoop/MapReduce/*,
    /opt/Hadoop/share/Hadoop/MapReduce/lib/*,
    /opt/Hadoop/share/Hadoop/yarn/*,
    /opt/Hadoop/share/Hadoop/yarn/lib/*
  </value>
</property>
```

Finalizada la configuración se ejecuta el servicio y se visualiza el funcionamiento de los procesos YARN, en la figura 45 se visualiza que el ResourceManager se encuentra activo en el nodo 1.

```
$ start-yarn.sh
```

Figura 45. Proceso de administración de YARN ejecutándose.

En la figura 46 se puede visualizar que el servicio está disponible vía web, como se mencionó anteriormente el sistema de procesamiento YARN permite la administración de las aplicaciones MapReduce lanzadas desde el clúster de

Hadoop por lo que se puede verificar que mediante el acceso web se pude gestionar todos los procesos MapReduce.

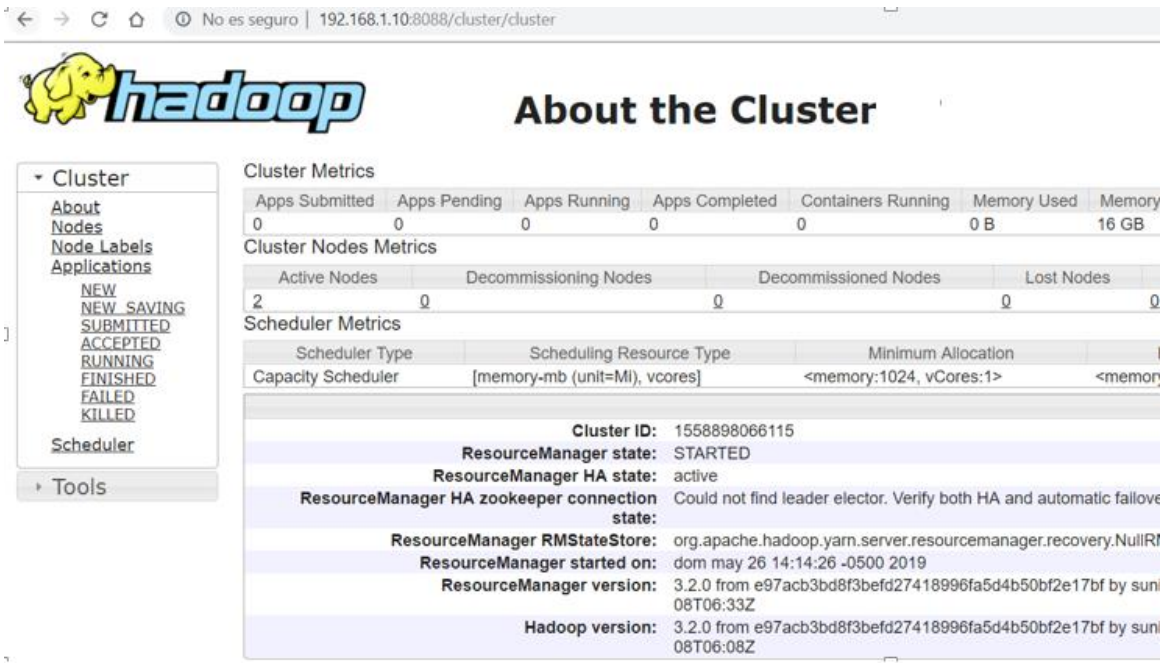


Figura 46. Servicio de MapReduce YARN ejecutándose.

5.7. Configuración del clúster Hadoop

5.7.1. Configuración de seguridades

El acceso remoto entre los equipos que componen el clúster debe ser transparente, es decir debe existir comunicación sin prohibiciones entre los tres nodos, para que esto tome la funcionalidad necesitada, se configuro todos los parámetros se seguridades correspondientes a la creación de claves privadas para el servicio remoto de ssh.

En la figura 47 se puede visualizar las claves generadas para los tres nodos, nodo1, nodo2 y nodo3 esto permite que el clúster trabaje como uno solo.

```
[hadoop@nodo1 .ssh]$ ls
authorized_keys  id_rsa  id_rsa.pub  known_hosts
[hadoop@nodo1 .ssh]$ cat authorized_keys
ssh-rsa AAAAB3NzaC1yc2EAAAADAQABAAABAQDX8vFkS4IU8a76Sq17KzwL4jOmz3MJ888Y5gMZXRdrRkkAkcmoiNKzXkl3VHV/oVE6
dzcce2nQbnxcFlCSLafppKvY61GIpq7jFZ2MfetI1NaExAkg/Cuf9zHvFruP2bkkW8g6rNKnyHVDQ0RrdMzVWrqnr4uQTS9PXXrzctKO
ToEVhMqm+JQxXj36itIf/XUw6zdD2YzCUH4uY6tZAIMUbFn6mOvGyJOXFFSGkHDTWH hadoop@nodo1
ssh-rsa AAAAB3NzaC1yc2EAAAADAQABAAABAQDPw9KJo81fW+Zo3XAHWk+VxhHCJpPH9iNsFC90k2bGnDKryKo97NPhVTV4LE1t3S43
OV1iTaxRdKCw59mwFXtXqKUZKzyobSqmR8WgYLf2mYz3eAUvEHgKIN6mqHLU7fhbf8uLec5XQW2vsMfzeeNiIjSwJRNp6i6l1zrl9Zen
nY/nD29DUBSipI8/MZ0qcYVjqXui0uOAjPKinXV9CJhLyAt5boH0BnhAP+rpke7jDb hadoop@nodo2
ssh-rsa AAAAB3NzaC1yc2EAAAADAQABAAABAQC8hQVJrXsjea8P9cI9Ce0vWH10RFjvN/+alPq6jiQxWdzyXdmETR9i8UlhQAuD79Lj
ZkMsQxEW/aqQ9Rpdg3jbwkHN2/oBD2xMpMU3XIXflB3FbZGR2VlFPwzgXxZQSJoE3XpstYQkiIGjn7vNfJh/xm4qOP2RPMNRjzOgF2p9
dl5xIBZgZts16SlCibwT2XcXxtfYa9u3uZ7k5ai3f6emYFPfKahh15EIh7v4qCB17t hadoop@nodo3
[hadoop@nodo1 .ssh]$
```

Figura 47. Claves ssh generados tres nodos.

Otra especificación que se realizo fue añadir las direcciones IP de cada nodo en el archivo que se encuentra en la siguiente dirección.

$ vi /etc/hosts

En la figura 48 se puede observar los nodos con el direccionamiento de red del centro de datos experimental de la UDLA.

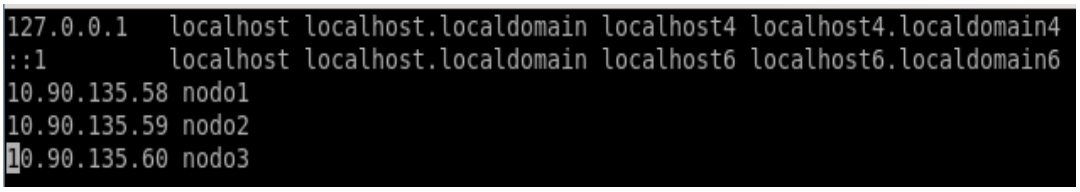


Figura 48. Direccionamiento físico de los nodos.

5.7.2. Configuración del sistema de ficheros HDFS

En el clúster el sistema de ficheros HDFS deben ser distribuidos, se debe recalcar que este fichero toma el nombre de DataNode y es donde se almacenan los datos que se suben al ambiente de Hadoop.

En la figura 49 se visualiza el número de nodos y la dirección del DataNode como del NameNode, estos directorios se encuentran en cada nodo, para la configuración se accedió a la siguiente dirección.

$ cd /opt/Hadoop/etc/Hadoop

$ vi hdfs-site.xml

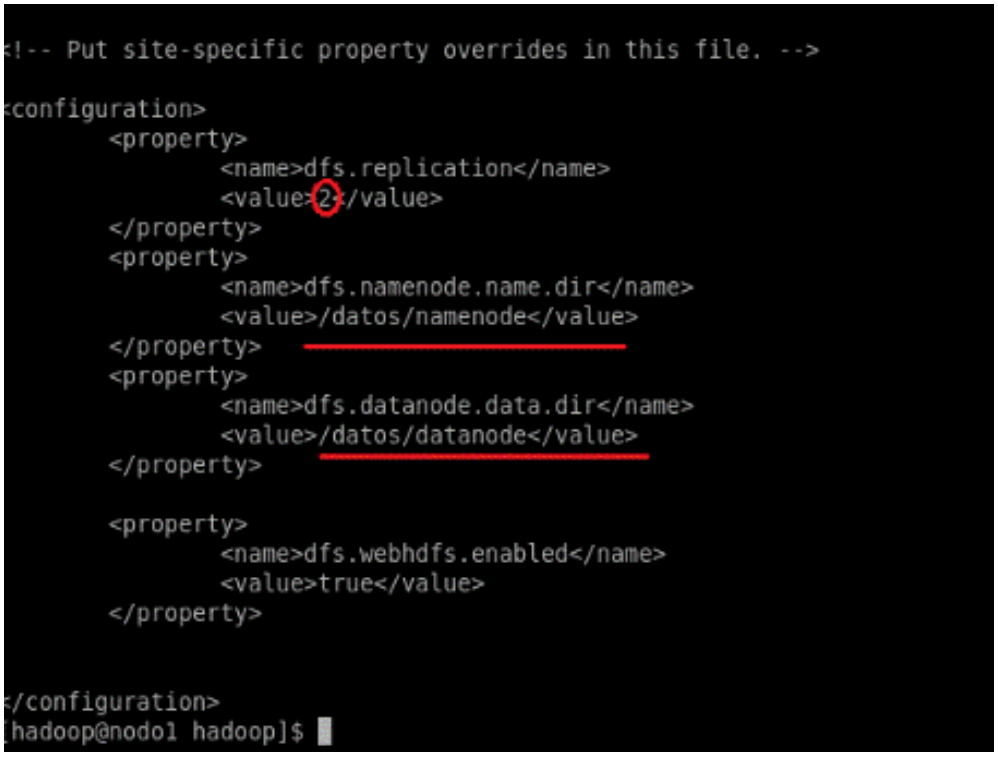


Figura 49. Configuración del HDFS para el clúster.

En este ítem se debe especificar que el NameNode solo debe estar creado en el nodo principal que trabaja como coordinador de los demás nodos, y los DataNode deben ser creados en los nodos esclavos, si esto no se cumple el servicio HDFS no se ejecuta.

5.7.3. Configuración del negociador de recurso YARN

El YARN al igual que el servicio de HDFS debe estar distribuido en el clúster, el principal servicio que administra YARN, es el ResourceManager el cual permite administrar el clúster a nivel de procesos MapReduce, la configuración de este fichero se lo hizo en el nodo principal accediendo a la siguiente dirección.

$ cd /opt/Hadoop/etc/Hadoop

$ vi yarn-site.xml

En la figura 50 se observa la configuración realizada al fichero para que el servicio funcione dentro del ambiente de Hadoop en modo clúster.

- La sección marcada como uno, especifica el nombre del nodo maestro en donde se ejecuta el ResourceManager.

- La sección marcada como dos son librerías que permiten ejecutar el servicio en la versión 3.0 de Hadoop.

```
<!-- Site specific YARN configuration properties -->
        <property>
                <name>yarn.resourcemanager.hostname</name>
                <value>nodo1</value>
        </property>

        <property>
                <name>yarn.nodemanager.aux-services</name>
                <value>mapreduce_shuffle</value>
        </property>

        <property>
                <name>yarn.nodemanager.aux-services.mapreduce_shuffle.class</name>
                <value>org.apache.hadoop.mapred.ShuffleHandler</value>
        </property>

        <property>
                <name>yarn.application.classpath</name>
                <value>
                /opt/hadoop/etc/hadoop,
                /opt/hadoop/share/hadoop/common/*,
                /opt/hadoop/share/hadoop/common/lib/*,
                /opt/hadoop/share/hadoop/hdfs/*,
                /opt/hadoop/share/hadoop/hdfs/lib/*,
                /opt/hadoop/share/hadoop/mapreduce/*,
                /opt/hadoop/share/hadoop/mapreduce/lib/*,
                /opt/hadoop/share/hadoop/yarn/*,
                /opt/hadoop/share/hadoop/yarn/lib/*
                </value>
        </property>
```

1

2

Figura 50. Configuración de administrador recursos.

5.7.4. Validación del funcionamiento del entorno Hadoop

En la validación del clúster todos los servicios deben estar ejecutándose en los nodos correspondientes, es decir los servicios del DataNode, NameNode y ResourceManager deben ejecutarse en el nodo maestro como en los nodos esclavos.

En la figura 51 se observa que los servicios DFS y YARN se están ejecutando de manera normal en el nodo maestro.

```
[hadoop@nodo1 ~]$ start-dfs.sh
Starting namenodes on [nodo1]
nodo1: Warning: Permanently added the ECDSA host key for IP address '10.90.135.58' to the list of
known hosts.
Starting datanodes
Starting secondary namenodes [nodo1]
[hadoop@nodo1 ~]$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
[hadoop@nodo1 ~]$ jps
8323 ResourceManager
8645 Jps
7816 NameNode
8027 SecondaryNameNode
[hadoop@nodo1 ~]$
```

Figura 51. Servicios DFS y YARN ejecutándose en nodo maestro.

En la figura 52 se puede observar que el nodo 2 tiene levantado el servicio del DataNode y NodeManager los únicos servicios que deben ejecutarse en los nodos esclavos, esto permite concluir que la configuración esta correcta para el ambiente de Hadoop con tres nodos.

Figura 52. Servicios ejecutándose desde nodos esclavos.

5.7.5. Almacenamiento con Hive

En el proyecto se implementó Hive como herramienta de Data Warehousing para facilitar la creación de consultas y administración de los datos que se almacenan en Hadoop.

Descargamos Hive de la página oficial www.hive.apache.org realizamos las siguientes directrices.

$ cd /opt/Hadoop

$ tar xvf /home/Hadoop/Descargas/apache-hive-2.3.4-bin.tar.gz

$ mv apache-hive-2.3.4-bin/ hive

Añadimos las variables de entorno

export HIVE_HOME=/opt/Hadoop/hive

PATH=$PATH:$HADOP_HOME/bin:$HADOOP_HOME/sbin:$HIVE_HOME/bin

Creamos el directorio bbdd donde se crearon los metastore, librerías y funciones relacionadas al uso de Hive dentro de un entorno Hadoop.

$ cd /opt/Hadoop/hive

$ mkdir bbdd

$ cd bbdd

$ schematool -dbType derby -initSchema

Creamos el Data Warehousing en el sistema de ficheros HDFS

$ hdfs dfs -mkdir /tmp

$ hdfs dfs -chmod g+w /tmp

$ hdfs dfs -mkdir -p /user/hive/warehouse

$ hdfs dfs -chmod g+w /user/hive/warehouse

Una vez finalizada las configuraciones se levantó el servicio de hiveserver2 como se observa en la figura 53.

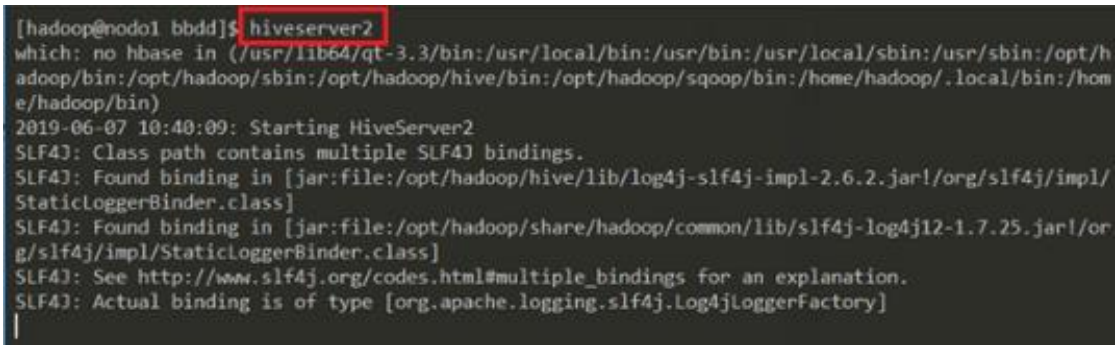


Figura 53. Se ejecuta el servicio de hiveserver2.

Hive como herramienta de Data Warehousing para Hadoop tiene una ventaja que sobre sale del resto de herramientas que es tener los drivers de conexiones como JDBC/ODBC los cuales permiten conectarnos con sistemas de inteligencia de negocios (BI), que para el proyecto es de gran utilidad debido a que se utiliza la herramienta Qlik Sense en su versión gratuita para realizar el análisis de información el que se visualiza en el capítulo final de evaluación de resultados.

Antes de finalizar con la configuración de Hive, se debe tener presente que al contar con drivers para conexiones externas este debe ejecutar el agente de conexiones remotas Beeline.

En la figura 54 se muestra la conexión mediante su agente remoto Beeline

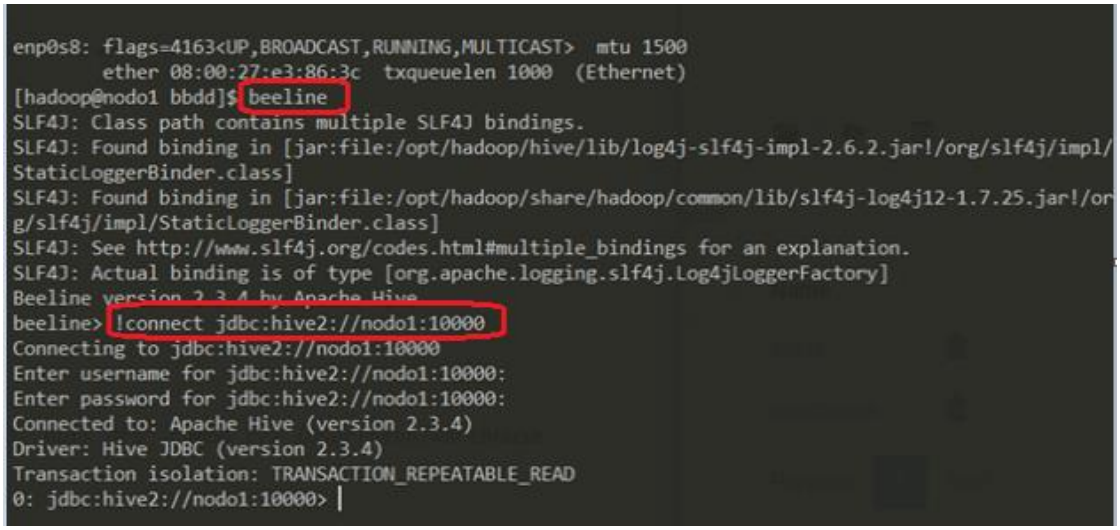


Figura 54. Servicio remoto Beeline ejecutándose.

Una vez que el servicio Hiveserver2 se encuentra ejecutándose se puede acceder de forma remota al entorno de Big Data el cual trabaja con base de datos o tablas que simulen un modelo relacional todo esto gracias a Hive que se implementó como herramienta para tratamiento de datos.

En la figura 55 se visualiza la creación de la base de datos (consumo_dcexpe_udla) en donde se almacena la información de los equipos de red más el UPS.

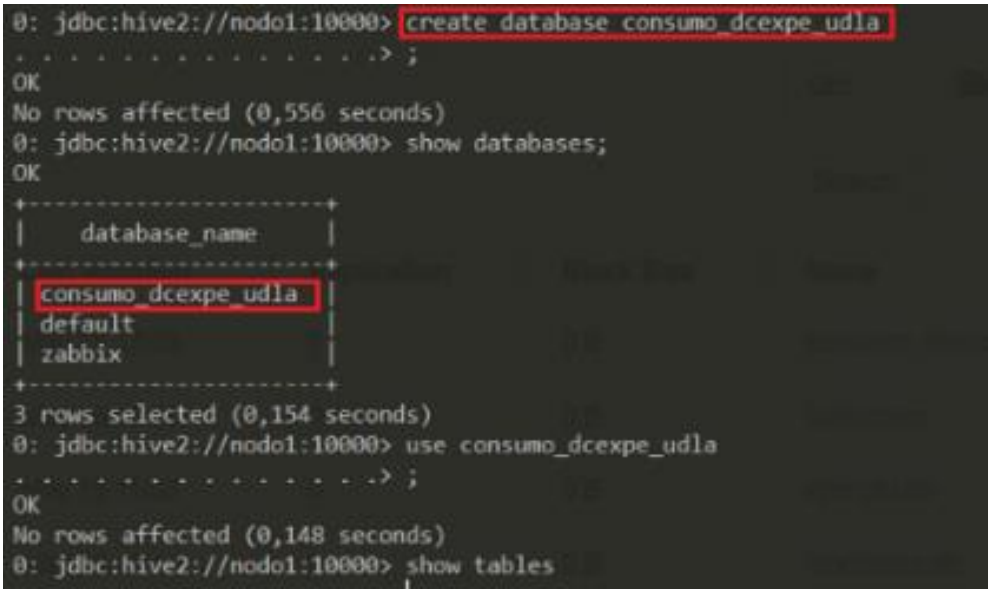


Figura 55. Creación de la base de datos NoSQL para consumo eléctrico.

En la figura 56 se observa la tabla de monitoreo creada, la cual contiene los datos de consumo de todo el equipamiento del centro de datos.

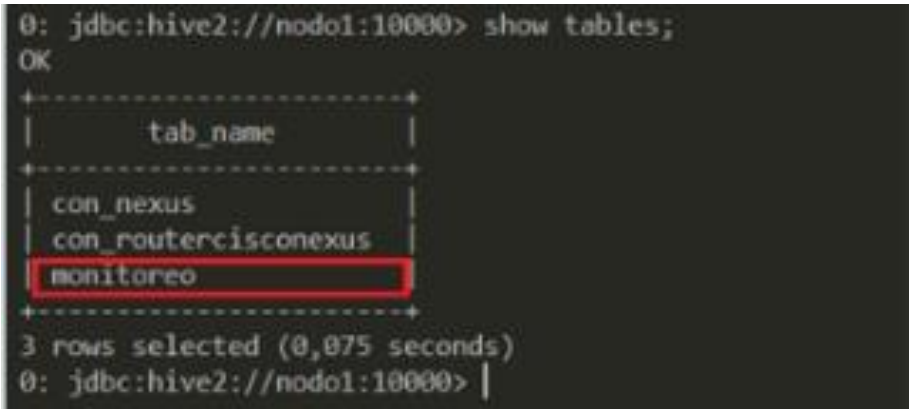


Figura 56. Tabla de monitoreo creada para almacenar consumo eléctrico.

En la figura 57 se muestra como la tabla está en un sistema de archivos HDFS dentro del entorno de Hadoop y está lista para procesarse o analizarse desde cualquier herramienta de visualización de datos.

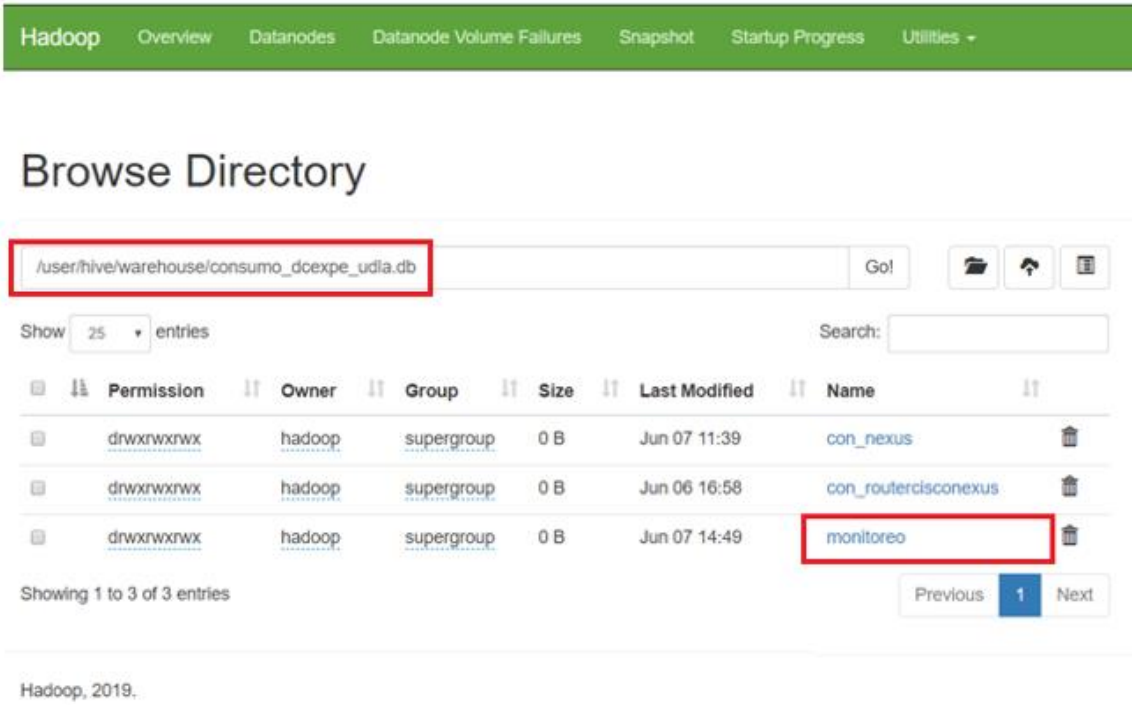


Figura 57. Entorno HDFS con la información procesada.

Finalmente, en la figura 58 se puede visualizar el sistema de archivos HDFS habilitado en los dos nodos y de manera similar la figura 59 muestra el gestionador de recursos YARN ejecutándose para los dos nodos que forman el ambiente de Hadoop.

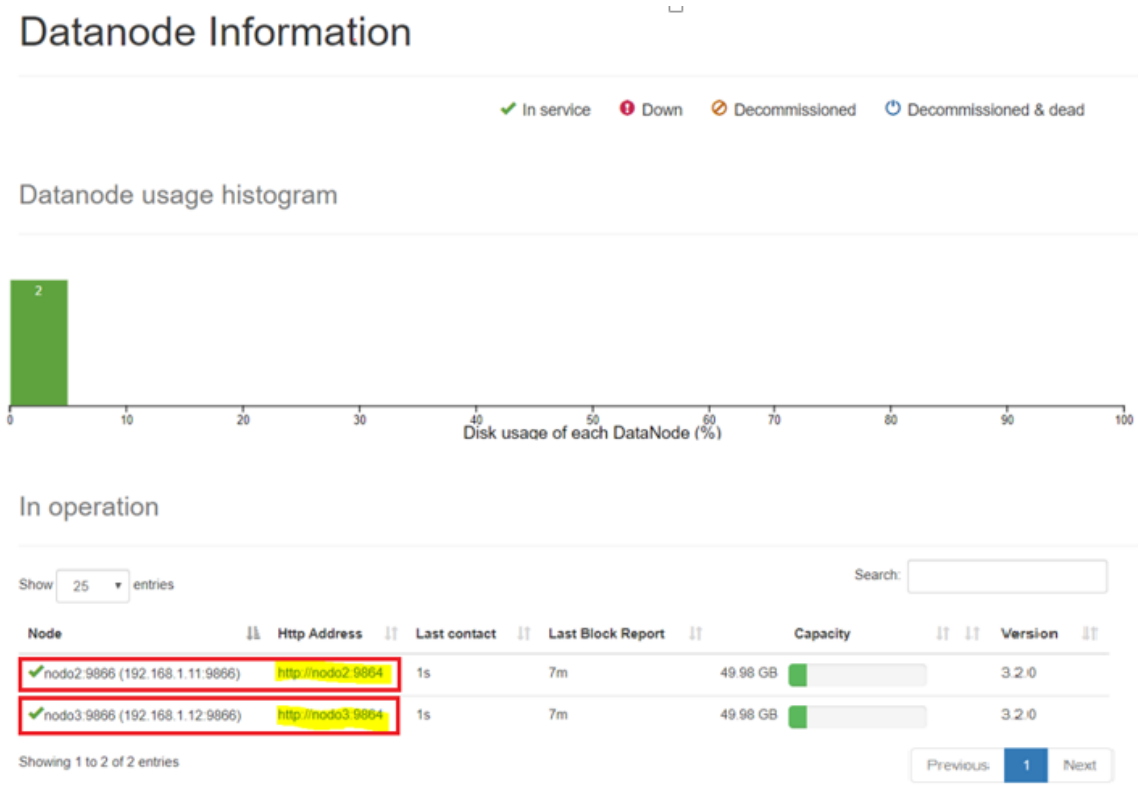


Figura 58. Servicio HDFS en modo clúster.

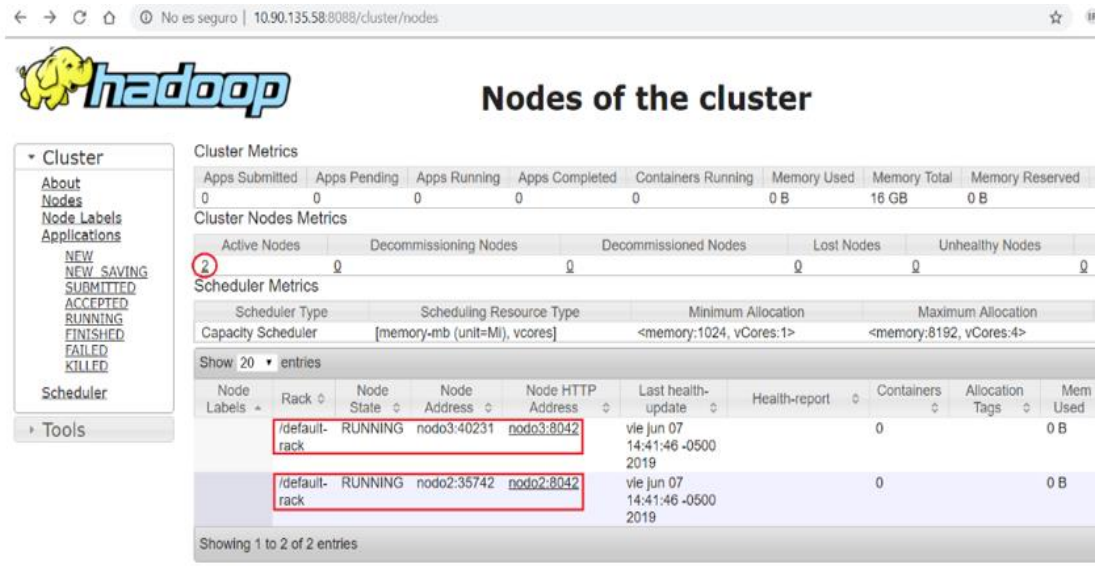


Figura 59. Servicio YARN en modo clúster.

En este capítulo se detalló las configuraciones realizadas para la implementación del ambiente de Big Data con un ecosistema de Hadoop y una arquitectura de dos nodos esclavos y uno maestro.

En trabajo paralelo se cargó la información adquirida del monitoreo correspondiente al consumo eléctrico de los equipos disponibles en el centro de datos experimental de la universidad de las Américas.

La información adquirida por el monitoreo la cual esta almacenada en el ambiente de Big Data, es analizada en el capítulo siguiente en donde se detalla los resultados adquiridos.

6. CAPÍTULO VI. EVALUACIÓN DE RESULTADOS

Finalmente, en este capítulo se evalúan los resultados y se describe los temas más destacados, correspondiente a la configuración e investigación realizada.

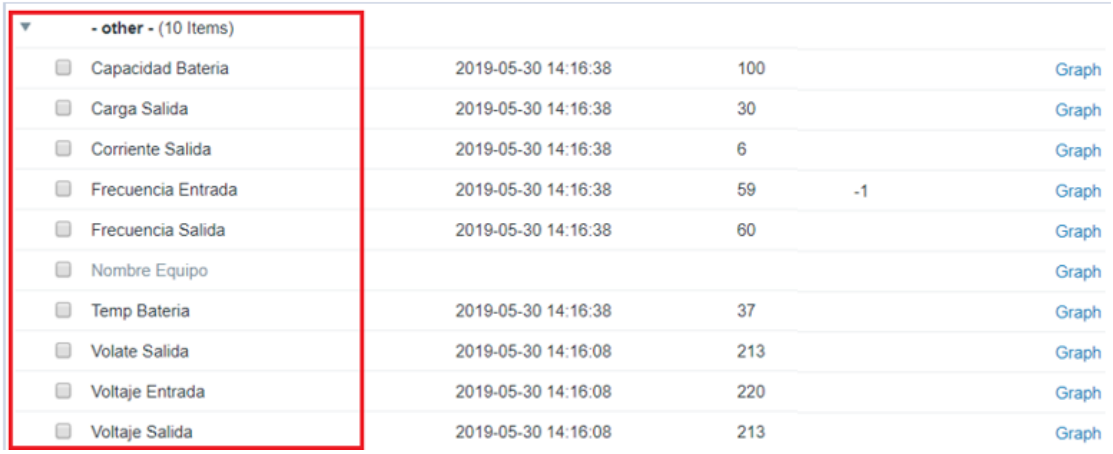
6.1. Evaluación equipos disponibles

Para iniciar una evaluación se debe destacar que se realizaron tareas específicas correspondientes a las configuraciones en cada equipo, adicional se realizó un monitoreo especial para la evaluación de consumo de los equipos que se tuvieron acceso dentro del centro de datos experimental de la UDLA.

Los equipos que entraron al monitoreo fueron los Nexus de la categoría n3500 y el UPS marca APC modelo AP9215RM, los ítems que fueron monitoreados son los siguientes.

- Velocidades en cada interfaz (Nexus)
- Etiquetas de administración (Nexus - UPS)
- Voltajes de entrada (Nexus - UPS)
- Voltajes de salida (Nexus - UPS)
- Temperatura batería (UPS)
- Corriente salida (UPS)
- Frecuencia de salida y entrada (UPS)
- Capacidad de batería (UPS)

Los parámetros que se tomaron en cuenta para el análisis son los que se especifican en la imagen 60, los mismos que se subieron al entorno de Hadoop para su análisis correspondiente.



- other - (10 Items)				
Capacidad Bateria	2019-05-30 14:16:38	100		Graph
Carga Salida	2019-05-30 14:16:38	30		Graph
Corriente Salida	2019-05-30 14:16:38	6		Graph
Frecuencia Entrada	2019-05-30 14:16:38	59	-1	Graph
Frecuencia Salida	2019-05-30 14:16:38	60		Graph
Nombre Equipo				Graph
Temp Bateria	2019-05-30 14:16:38	37		Graph
Volate Salida	2019-05-30 14:16:08	213		Graph
Voltaje Entrada	2019-05-30 14:16:08	220		Graph
Voltaje Salida	2019-05-30 14:16:08	213		Graph

Figura 60. Parámetros de monitoreo.

El ambiente de Big Data implementado está preparado para procesar cantidades enormes de información, por lo cual para conocer el volumen de información que actualmente se va a procesar se realizó la siguiente formula que encuentra la cantidad de información que se está subiendo y permite respaldar el beneficio de tener una instancia de Big Data que permita el procesamiento de información.

6.1.1. Fórmula matemática para descubrir el volumen de datos adquiridos

Variables establecidas para la configuración de adquisición de datos

n = Tiempo de recopilación de datos (Poleo)

y = Cantidad de ítems monitoreados (Tags)

t = Días, meses de monitoreo (Tiempo)

r = proyección (6 meses)

p = Tamaño (MB, GB, TB, PB, EB, ZB, YB)

s = segundos por minuto

m = minutos por hora

h = horas por día

Valores establecidos para el análisis.

n = 30’’

y = 11 (UPS) 102 (Nexus) Total 113 ítems monitoreados.

t = 78 horas

r = 4320 horas

x = ? (Número de registros).

Fórmula

$$\boldsymbol{x = \left\{ \left(\frac{\left(\frac{n * 2}{s} \right) * (m)}{h} \right) * y \right\} * t}$$

$$x = \left\{ \left(\frac{\left(\frac{30 * 2}{30} \right) * 60}{1} \right) * \ (113) \right\} * 78$$

X = 1.057.680 (Total de registros generados en el monitoreo mediante un tiempo establecido).

Proyección

Para esta simulación se toma los seis meses que transformado a horas seria la siguiente cantidad.

N = cantidad meses

H = horas por día

D = días del mes

Y = cantidad de registros

Proyección = [(H*D) * N] * Y

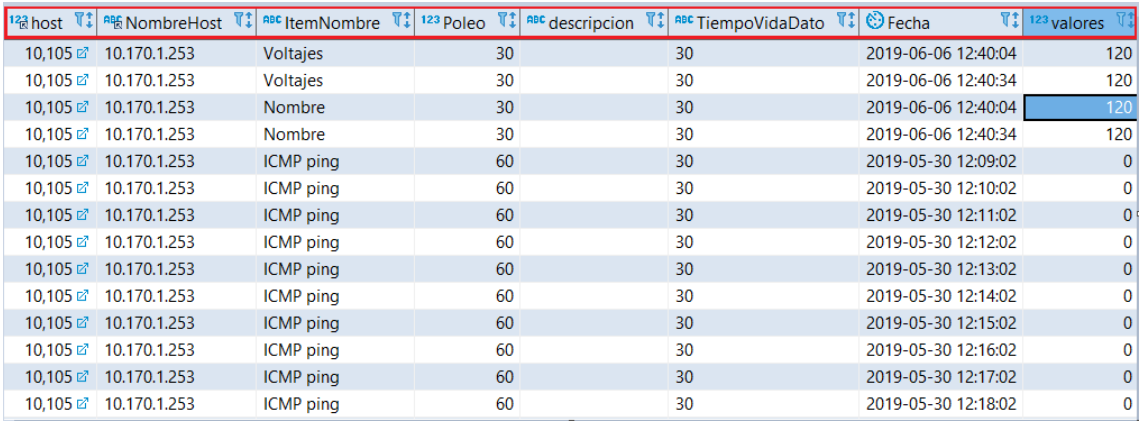
Proyección = [(24*30) * 6] * 1.057.680

Proyección = 761.529.600 (Cantidad de registros proyectados hasta el final de año).

Una vez que conocemos la cantidad de registros, necesitamos saber el peso que estos registros tiene en su equivalente en MB/GB de todos los registros y así poder tener la idea del volumen de información que procesará el ambiente de Big Data.

Para este cálculo necesitamos conocer los campos que se consultaron y el tipo de campo debido a que esto nos permite conocer el peso por tipo de dato.

En la figura 61 se visualiza los campos consultados y en la tabla 14 se especifica el tipo de dato y su valor en Bytes.



host	NombreHost	ItemNombre	Poleo	descripcion	TiempoVidaDato	Fecha	valores
10,105	10.170.1.253	Voltajes	30		30	2019-06-06 12:40:04	120
10,105	10.170.1.253	Voltajes	30		30	2019-06-06 12:40:34	120
10,105	10.170.1.253	Nombre	30		30	2019-06-06 12:40:04	120
10,105	10.170.1.253	Nombre	30		30	2019-06-06 12:40:34	120
10,105	10.170.1.253	ICMP ping	60		30	2019-05-30 12:09:02	0
10,105	10.170.1.253	ICMP ping	60		30	2019-05-30 12:10:02	0
10,105	10.170.1.253	ICMP ping	60		30	2019-05-30 12:11:02	0
10,105	10.170.1.253	ICMP ping	60		30	2019-05-30 12:12:02	0
10,105	10.170.1.253	ICMP ping	60		30	2019-05-30 12:13:02	0
10,105	10.170.1.253	ICMP ping	60		30	2019-05-30 12:14:02	0
10,105	10.170.1.253	ICMP ping	60		30	2019-05-30 12:15:02	0
10,105	10.170.1.253	ICMP ping	60		30	2019-05-30 12:16:02	0
10,105	10.170.1.253	ICMP ping	60		30	2019-05-30 12:17:02	0
10,105	10.170.1.253	ICMP ping	60		30	2019-05-30 12:18:02	0

Figura 61. Nombre de los campos consultados.

En la tabla 14 se escriben los nombres de los campos que se seleccionaron para el monitoreo con su tipo de dato y peso en Bytes.

Tabla 14.

Tamaño de los campos para el monitoreo.

Nombre	Tipo de dato	Valor/peso
Host	Bigint (4)	8 bytes
Nombremos	Varchar (125,6)	125 bytes
itemNombre	Varchar (125,6)	125 bytes
Poleo	Bigint (4)	8 bytes
Descripción	Varchar (125,6)	125 bytes
TiempoVidaDato	Bigint (4)	8 bytes
Fecha	Bigint (12)	8 bytes
valores	Bigint (4)	8 bytes

Adaptado de (w3schools, 2018)

Con los valores especificados en la tabla 14 se realizó la fórmula matemática para encontrar el peso de la consulta por registro, todo con base al peso de cada campo el mismo que corresponde al ítem que se genera en el monitoreo.

v = (peso en Bytes por tipo de dato)

n = número de registros

m = tipo de medida (MB)

Valores establecidos

v = byte

n = 1.507.680

m = 1048576

x = peso de total de registros

$$d = \frac{(v1 + v2 + v3 + v4 + v5 + v6 + v7 + v8) * n}{m}$$

Remplazamos los valores para obtener el tamaño de los registros en donde:

x = 418,60 MB (Es el tamaño en Megabyte de los registros monitoreados actualmente).

Para calcular el tamaño total con una proyección de seis meses se cambió el valor de los registros (n) en donde:

n = 761.529.600

Proyección = 301.394,256591796875 MB

Proyección = 294.33 GB

6.1.2. Análisis de la solución de Big Data

Con la predicción realizada se puede indicar que en un ambiente de monitoreo de 6 meses se estaría procesando un equivalente a 300 GB de información que con herramientas y computo tradicional sería imposible tratarlos.

Una condición importante a tener en cuenta es que si se aumenta los ítems de monitoreo por dispositivo o ingresan más equipos al monitoreo este tamaño se dispararía enormemente hasta llegar a los Terabytes o rodear los Petabytes de información.

La solución de Big Data con tecnología Hadoop que se orienta a reducir costos a nivel de cómputo y procesar información historizada (Batch) es una de las

mejores opciones para este tipo de panorama, en donde permite crecer de forma exponencial a nivel de cómputo y trabajar con información en tiempo real y almacenada.

Finalmente, se debe aclarar que estos datos son adquiridos de los equipos de red y alimentación (UPS) y la información es tratada en un lapso de 24 horas, por ende, si se necesita en algún momento analizar información anterior se debe procesar información que fue almacenada en un lapso de tiempo específico.

6.1.3. Análisis del consumo eléctrico

En esta sección se detalla la adquisición de datos correspondiente a voltajes y sus valores establecidos en los ítems de monitoreo, se realiza un análisis de la información adquirida mediante visualizadores gráficos que permiten un mejor entendimiento de la analítica propuesta.

6.1.3.1. Análisis del consumo Cisco Nexus

Para realizar este análisis se trabajó con los datos de voltaje de entrada en los Nexus de la familia 3000 y voltajes de entrada y salida del UPS APC modelo AP9215RM.

En la tabla 15 se unifica los valores de voltaje de los equipos de red, cómputo, almacenamiento y el equipo de alimentación ininterrumpida UPS, que proporciona su ficha técnica.

Tabla 15.

Valores de voltaje de equipos del centro de datos experimental

Nombre	Voltage Input	Voltage Output	Eficiencia del suministro de energía
Cisco Nexus 3124(1)	100 to 240 VAC		89 a 91 % at 220 VCA
Cisco Nexus 3124(2)	100 to 240 VAC		89 a 91 % at 220 VCA
Cisco USC Chasis 5108	100 a 120 VAC 200 a 240 VAC		94% to 240 VCA
Cisco UCS B200M4	100 to 240 VAC 90 to 264 VAC		92% to 95 VCA

EMC VNXe3200	100 to 240 V		92% to 95 VCA
UPS - APC AP9215RM	200, 208, 220, 230, 240 Vac; 60 or 50 Hz,	200, 208, 220, 230, 240 Vac; 50, 60 Hz,	Approximately 89%

Tomado de (Cisco, 2019)

El análisis que se realiza para el equipo Cisco Nexus 3124 abarca el consumo de entrada entre los valores establecidos en su ficha técnica con respecto a lo adquirido en el monitoreo actual.

El análisis estableció los valores por defecto detallados en la ficha técnica del equipo y si estos valores mantienen una línea continua en su voltaje de entrada, la figura 62 muestra el mejor escenario de consumo de voltaje de entrada que es 120 voltios.

Para la comparación se tomó la muestra del 30 de mayo en la cual se monitorearon los equipos por cerca de 10 horas.

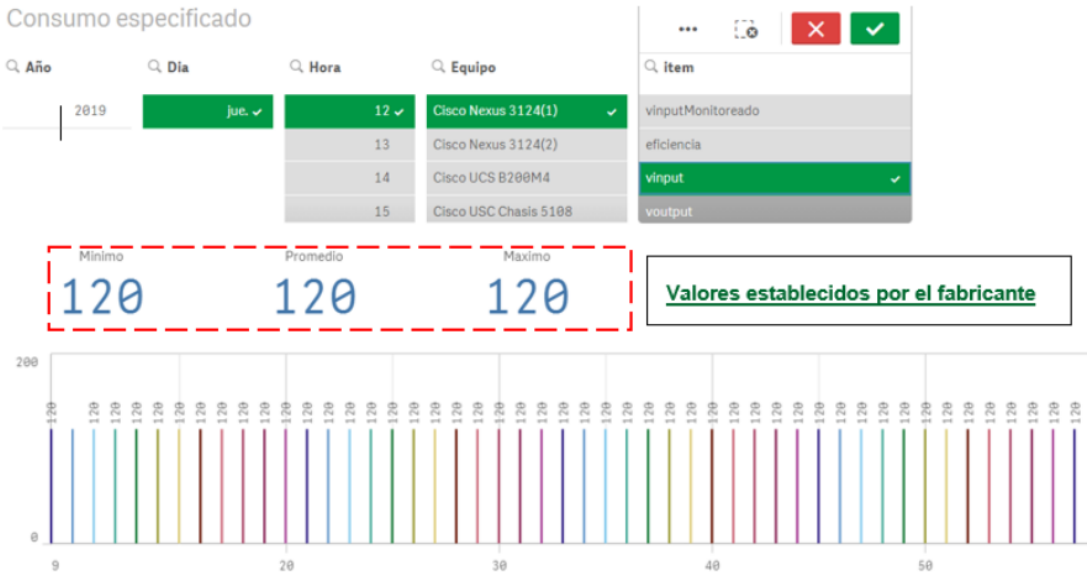


Figura 62. Comparación Cisco Nexus 3124 (1) mejor escenario.

Adaptado de Qlik Sense, sf

En la figura 63 se puede observar el escenario real, el que toma valores con relación a lo adquirido en el monitoreo que como se observa oscila entre un mínimo de 124 voltios y un máximo de 128 voltios, dando como promedio 126,8 voltios sobre el equipo Cisco Nexus 3124.

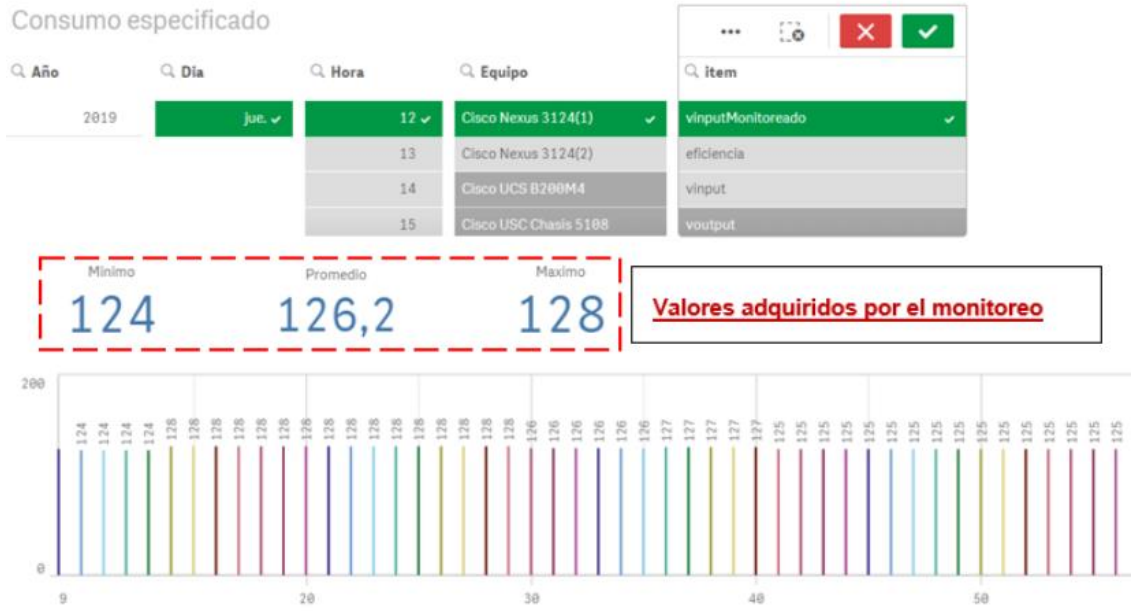


Figura 63. Comparación Cisco Nexus 3124 (1) escenario real.
Adaptado de Qlik Sense, sf

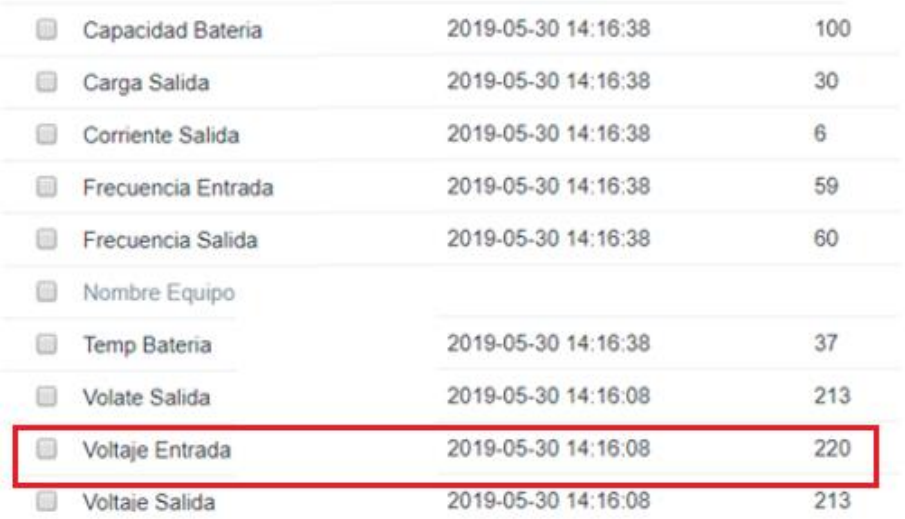
- El análisis del Nexus 3124 con IP (10.170.1.253) muestra que el equipo tiene un voltaje de 128 voltios, superior a los especificado en su ficha técnica, sin embargo los valores se mantiene en un rango validado por el fabricante, por lo que se deduce que con los datos adquiridos, si existe una eficiencia energética correspondiente al rango de 89 a 91%, esto implica que los equipos en el momento del monitoreo tiene un rendimiento enfocado a sustentar el tema de eficiencia energética y aportan a un PUE menor o igual a uno.
- En este análisis se especifica que se tomaron muestras de 6 a 10 horas de monitoreo y se encontraron variaciones por minuto de 2 hasta 8 voltios, lo que quiere decir que los equipos sufren un tipo de sobrecarga en periodos cortos en especial el equipo Cisco Nexus 1 con IP (10.170.1.253), por lo que se recomienda seguir con un monitoreo constante y verificar valores como frecuencia, potencia, memoria, para así poder detallar una posible causa de la variación que de acuerdo al fabricante están dentro de los parámetros establecidos pero que con el pasar del tiempo podría causar percances en el equipo.

6.1.3.2. Análisis de consumo UPS

El equipo de alimentación interrumpida ingreso al monitoreo para tener una aproximación más real del consumo de carga total de todos los equipos que están conectados a esta fuente de alimentación alterna y verificar el voltaje total que el equipo tiene cuando está en uso a su máxima capacidad de consumo.

En la tabla 15 visualizada anteriormente se verifica los parámetros de voltaje que corresponden al UPS - APC AP9215RM y en la figura 64 se puede observar los parámetros que se configuraron en el UPS para su monitoreo.

En el análisis realizado para este equipo se tomaron los voltajes de entrada.



Capacidad Bateria	2019-05-30 14:16:38	100
Carga Salida	2019-05-30 14:16:38	30
Corriente Salida	2019-05-30 14:16:38	6
Frecuencia Entrada	2019-05-30 14:16:38	59
Frecuencia Salida	2019-05-30 14:16:38	60
Nombre Equipo		
Temp Bateria	2019-05-30 14:16:38	37
Volate Salida	2019-05-30 14:16:08	213
Voltaje Entrada	2019-05-30 14:16:08	220
Voltaje Salida	2019-05-30 14:16:08	213

Figura 64. Parámetros de monitoreo del UPS.

En la figura 65 podemos visualizar los datos del monitoreo en un determinado lapso de tiempo, los parámetros que se analizaron corresponden al voltaje de entrada en donde se puede verificar un mínimo de 214 y un máximo de 226 en determinados minutos, sin embargo, los valores están dentro del rango establecido por el fabricante.

También se puede destacar que el voltaje de entrada varía en unos 11 voltios y tiene un promedio de 220, el mismo que se encuentra dentro del rango establecido para mantener una determinada eficiencia energética.

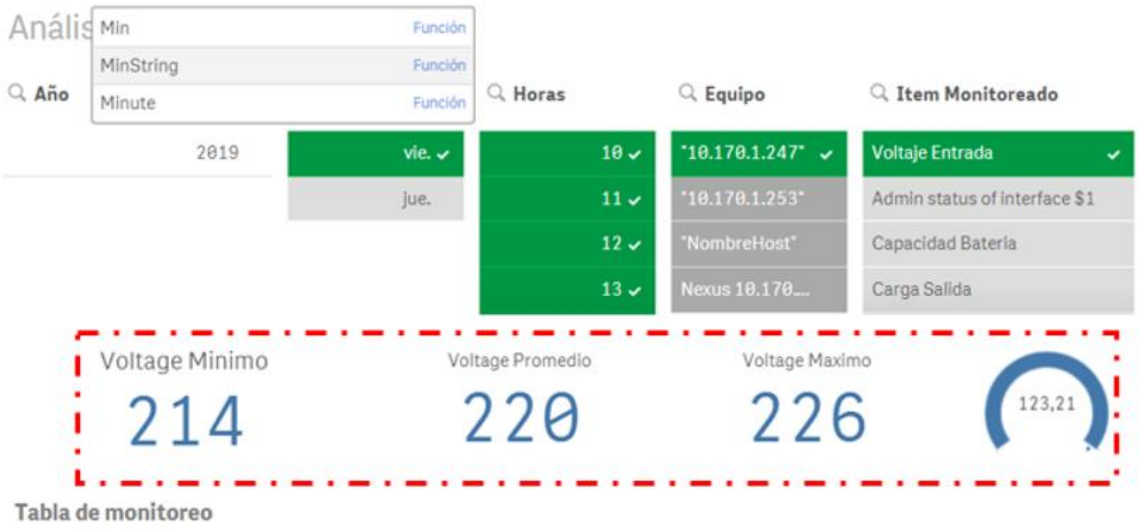


Figura 65. Valores del voltaje de entrada UPS.
Adaptado de Qlik Sense, sf

Destacamos que la medida fue tomada del monitoreo de un día, en la cual la carga del equipo no fue mayor a sus valores estándares, se podría a futuro analizar el equipo cuando exista un corte de energía y todos los dispositivos se conecten al UPS y poder validar una cargar superior y ver los mínimos y máximos de voltaje.

En la figura 66 se puede observar unos picos entre voltajes mínimos y máximos que detallan variaciones significativas en su voltaje.

Variación de voltajes

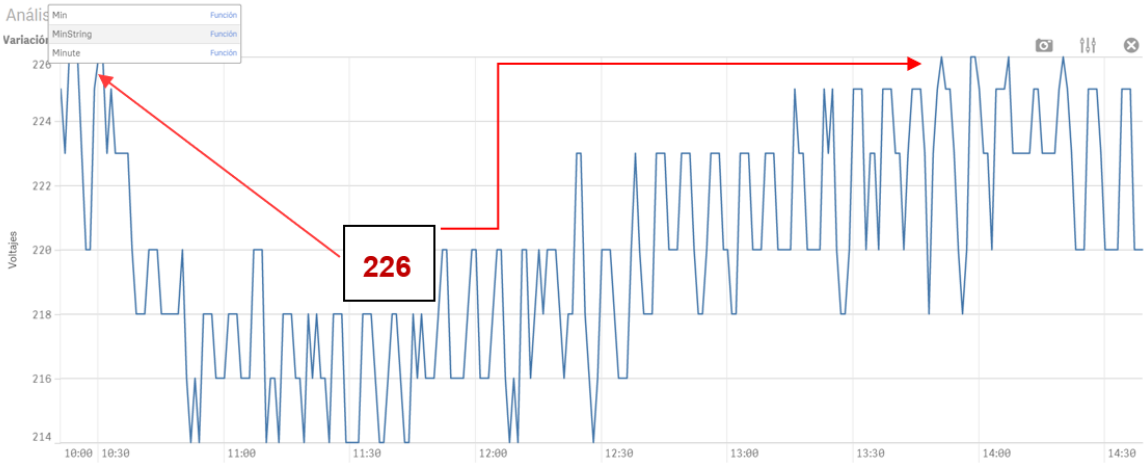


Figura 66. Picos de voltaje de entrada UPS.
Adaptado de Qlik Sense, sf

La figura 66 nos muestra el monitoreo del 7 de junio de 2019 con dos picos altos en dos determinados lapsos de tiempo, el primero entre las 10:00 y 10:30 el segundo entre las 13:45 a 14:30.

El primer periodo se puede decir que se debe a que se empezó a conectar equipos adicionales a las fuente de poder (laptops) y se empezó a prender las computadoras del centro de datos el cual estaba cerrado sin mayor consumo eléctrico, este tipo de anomalía se puede entender que a mayor tráfico eléctrico, entiéndase equipos encendidos en el centro de datos los cuales pueden estar conectados a la fuente externa UPS, el voltaje se eleva y de similar se va estabilizando en el transcurso de los minutos.

Una buena práctica es monitorear el equipo en horas picos para verificar los mínimos y máximos de voltaje, de esa forma validar que los parámetros oscilen en el rango que determina el fabricante, caso contrario se deberá realizar un balanceo en el consumo eléctrico para salvaguardar posibles fallos en la infraestructura.

El segundo periodo de 13:45 a 14:30 se debe al encendido de máquinas dentro del centro de datos, debido a que a esa hora el aula fue ocupada por estudiantes para una determinada clase y de similar el voltaje se fue estabilizando en el transcurso de los minutos.

Como se mencionó anteriormente sería un gran aporte realizar el mismo análisis en una hora específica para verificar los picos máximos de voltaje y establecer medidas que ayuden a mejorar o mantengan la eficiencia de consumo eléctrico en los equipos.

7. CONCLUSIONES Y RECOMENDACIONES

7.1. CONCLUSIONES

Para la información no estructurada que se generó en el monitoreo y el volumen de datos que se pueden obtener en el transcurso del año fue oportuno trabajar en un ambiente de Big Data que para este proyecto utilizó Hadoop como Framework de procesamiento y almacenamiento de datos con sus respectivos módulos de HDFS y MapReduce.

Con los cálculos realizados para obtener el número de registros y el peso de cada uno, se puede concluir que en un ambiente de monitoreo de 6 meses se estaría procesando un equivalente a 300 GB de información que con herramientas y computo tradicional es imposible procesarlos.

Se debe tener en cuenta que si se aumentan los ítems de monitoreo por dispositivo o ingresan más equipos al monitoreo el volumen de registros y peso aumentaría enormemente hasta llegar a los Terabytes o rodear los Petabytes de información que con una herramienta de monitoreo que maneje una base de datos relacional no sería posible procesar gran cantidad de información.

Si en algún instante se nenecita analizar información almacenada del monitoreo que pertenezca a un lapso de tiempo específico, implicaría procesar cantidades inmensas de información a nivel de Terabytes que para un modelo relacional de datos y hardware tradicional sería imposible, por lo cual la solución de Big Data con tecnología Hadoop que se orienta a reducir costos a nivel de cómputo y procesar información almacenada es una de las mejores opciones para este tipo de panorama, en donde permita crecer de forma exponencial a nivel de cómputo y trabajar con información en modo Batch.

El uso de Hive como herramienta de Data Warehousing posibilitó ser más precisos con los cálculos al momento de realizar el análisis de datos, debido a que el leguaje Hive SQL permitió realizar un mejor tratamiento de datos en un modelo NoSQL.

En los equipos Nexus 3124 los voltajes varían entre cada equipo, siendo el Nexus 1 el equipo que tiene un pico más alto en voltaje, 128 para ser exactos,

aun así los valores se mantienen en un rango especificado por el fabricante, por lo que se concluye que con los datos adquiridos, si existe una eficiencia energética correspondiente al rango de 89 a 91% por equipo todo en base a la ficha técnica del fabricante, esto implica que los equipos en el momento del monitoreo tiene un rendimiento enfocado a sustentar el tema de eficiencia energética y aportan a un PUE lo más cercano a uno.

Se puede concluir que cuando mayor cargo eléctrica tenga el UPS su voltaje de entrada varía de forma exponencial desde los 2 voltios hasta llegar a un pico de 11 voltios, los mismo que se normalizan en unos minutos después.

Con los datos generados en el monitoreo se puede validar que los equipos Nexus y el UPS APC modelo AP9215RM están en los parámetros de eficiencia energética de 89 a 92% lo cual permiten llevar una eficiencia energética en el centro de datos experimental.

7.2. RECOMENDACIONES

Una buena práctica para esta solución de Big Data seria agregar más nodos a los tres que se utilizaron en el clúster, a fin de verificar si los procesos MapReduce trabajan de mejor manera y validar si la tolerancia a fallos y escalabilidad continúa mejorando.

Realizar pruebas de conexiones remotas desde diferentes sistemas de visualización de datos sea de pago u open source, con el fin de validar la herramienta de Data Warehousing implementada en esta solución.

Existen picos de voltaje que deben ser controlados mediante un monitoreo de 24/7 para poder conocer las causas exactas de dichos excesos y así poder mitigar futuras fallas en los equipos.

Aumentar los ítems de monitoreo como pueden ser las interfaces de los equipos, procesadores, chasis, memorias RAM, discos, frecuencias en las que trabajan, temperatura entre otros parámetros, con el fin de conocer el comportamiento de los equipos dentro de la red.

Para mantener un análisis eficiente se recomienda mantener un monitoreo constante de toda la infraestructura del centro de datos experimental o por lo menos de los equipos que tienen mayor carga de trabajo.

Se debería trabajar en levantar el protocolo de monitoreo SNMP para los equipos de almacenamiento y computo mediante la actualización de su Framework de la mano del proveedor, con el fin de poder acceder a monitorearlos y conocer su comportamiento dentro de la red.

Se debería trabajar en realizar pruebas de carga al UPS, simulando cortes de energía para que todo el consumo de equipos del centro de datos experimental pasen por el UPS y poder tener un valor exacto de los mínimos y máximos de voltaje lo cual ayudaría a mejorar parámetros que aporten a una mejor eficiencia energética.

REFERENCIAS

Christof. (2019). *Christof Strauch*. Recuperado el 07 de abril de 2019, de https://www.christof-strauch.de/nosqldbs.pdf

Cisco. (2018). *Cisco UCS 5100 Series Blade Server Chassis Data Sheet*. Recuperado el 14 de abril de 2019, de https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-5100-series-blade-server-chassis/data_sheet_c78-526830.htm

Cisco. (2019). *Cisco UCS 5100*. Recuperado el 05 de abril de 2019, de Cisco: https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-5100-series-blade-server-chassis/data_sheet_c78-526830.html

Cisco. (2019). *Cisco UCS B200 M4 Blade Server*. Recuperado el 14 de abril de 2019, de https://www.cisco.com/c/en/us/products/servers-unified-computing/ucs-b200-m4-blade-server/index.html

Cisco. (2019). *community*. Recuperado el 15 de mayo de 2019, de https://community.cisco.com/t5/network-management/sfp-dom-snmp-monitoring/td-p/3229265

Cisco. (2019). *Products Services*. Recuperado el 14 de abril de 2019, de Cisco Nexus 3548 Switch - Cisco Nexus 3548-X, 3524-X, 3548-XL, and 3524-XL Switches Data Sheet.

Common, H. (2016). *El entorno de Hadoop*. Recuperado el 19 de marzo de 2019, de https://elentornodehadoop.wordpress.com/tag/hadoop-common/

Cook, J. D. (2009). *ACID versus BASE for database transactions*. Recuperado el 07 de abril de 2019, de https://www.johndcook.com/blog/2009/07/06/brewer-cap-theorem-base/

DCiE, L. T. (2018). *Soluciones en Data Center*. Recuperado el 04 de abril de 2019, de PUE y DCiE, Realidad de los Data Centers en México: https://teksar.mx/pue-y-dcie-lo-que-debes-de-conocer/

Desarrolloweb. (2017). *Desarrolloweb*. Recuperado el 09 de junio de 2019, de https://desarrolloweb.com/articulos/2336.php

Domenech, J. (2018). *Silicon*. Recuperado el 12 de junio de 2019, de https://www.silicon.es/volumen-datos-creados-mundo-se-multiplicara-10-ano-2025-2333472

EMC. (2019). *Introduction to the EMC VNXe3200*. Recuperado el 14 de abril de 2019, de https://www.emc.com/collateral/white-papers/h13058-vnxe3200-intro-wp.pdf

Fernández, E. (2017). *Big Data eje estratégico en la industria audivisual*. Recuperado el 20 de marzo de 2019, de https://dialnet.unirioja.es/servlet/articulo?codigo=6116687

Forrester. (2019). *The Pragmatic Definition Of Big Data*. Recuperado el 31 de marzo de 2019, de https://go.forrester.com/blogs/12-12-05-the_pragmatic_definition_of_big_data/

Fraga, A. (2018). *Tendencias en centro de datos para 2018*. Recuperado el 04 de abril de 2019, de TICbeat: https://www.ticbeat.com/tecnologias/tendencias-en-centro-de-datos-para-2018/

Genbeta. (2019). *NoSQL clasificación de las bases de datos según el teorema CAP*. Recuperado el 07 de abril de 2019, de https://www.genbeta.com/desarrollo/nosql-clasificacion-de-las-bases-de-datos-segun-el-teorema-cap

Glossary Gartner IT. (2019). *Gartner IT Glossary Big Data*. Recuperado el 31 de marzo de 2019, de https://www.gartner.com/it-glossary/big-data

Guilarte, M. (2013). *MuyComputerPRO*. Recuperado el 17 de abril de 2019, de https://www.muycomputerpro.com/2013/03/14/que-es-un-tier

Hadoop Apache. (2018). *Apache Fundation*. Recuperado el 21 de abril de 2019, de https://hadoop.apache.org/

Hortonworks. (2018). *Apache Hadoop 3 Adds Value Over Apache Hadoop 2*. Recuperado el 21 de abril de 2019, de https://es.hortonworks.com/blog/hadoop-3-adds-value-hadoop-2/

IBM. (2013). *Almacenamiento de datos estructurados con Big Data*. Recuperado el 04 de abril de 2019, de https://www.ibm.com/developerworks/ssa/library/bd-almacenamiento-datos/index.html

IBM. (2019). *IBM Developer*. Recuperado el 31 de marzo de 2019, de https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

Iglesias, A. (2014). *YARN la solución de Hadoop para modelos de procesamiento*. Recuperado el 21 de marzo de 2019, de https://www.beeva.com/beeva-view/tecnologia/yarn-la-solucion-de-hadoop-para-la-coexistencia-de-modelos-de-procesamiento/

Microsoft. (2019). Recuperado el 09 de junio de 2019, de Microsoft: https://docs.microsoft.com/

MongoDB. (2019). *NoSQL Databases Explained*. Recuperado el 07 de abril de 2019, de https://www.mongodb.com/nosql-explained?lang=es-es

NIST. (2017). *NIST Big Data Working Group (NBD-WG)*. Recuperado el 31 de marzo de 2019, de Bigdatawg.nist.gov: https://bigdatawg.nist.gov/

Rojas, E. (2013). *El PUE la medida de la eficiencia de los datacenters*. Recuperado el 14 de abril de 2019, de https://www.muycomputerpro.com/2013/02/26/pue-medida-eficiencia-datacenters

Schneider Electric. (2019). *APC*. Recuperado el 02 de junio de 2019, de https://www.apc.com/shop/us/en/products/APC-Symmetra-LX-16kVA-Scalable-to-16kVA-N-1-Rack-mount-208-240V/P-SYA16K16RMP

Talend. (2019). *Talend Real Time Open Source Data Integration Software*. Recuperado el 31 de marzo de 2019, de https://www.talend.com/resources/what-is-data-processing/

Techtarget. (2019). *Análisis de big data y su entorno*. Recuperado el 31 de marzo de 2019, de https://searchdatacenter.techtarget.com/es/definicion/Analisis-de-big-data

Toppertips. (2018). *Hadoop 3.0 Architecture*. Recuperado el 21 de abril de 2019, de http://toppertips.com/hadoop-3-0-architecture/

Uptime. (2019). *The Global Data Center Authority*. Recuperado el 13 de abril de 2019, de https://uptimeinstitute.com

Virtualbox. (2019). *Oracle VM VirtualBox*. Recuperado el 30 de abril de 2019, de https://www.virtualbox.org/

w3schools. (2018). *w3schools*. Recuperado el 09 de junio de 2019, de w3schools: https://www.w3schools.com/sql/sql_datatypes.asp

Wire Business. (2019). *Global Business*. Recuperado el 12 de junio de 2019, de https://www.businesswire.com/news/home/20170403006056/en/Seagate-Advises-Global-Business-Leaders-Entrepreneurs-Sharpen

Zerif Lite. (2019). *Zerif Lite*. Recuperado el 20 de mayo de 2019, de Medida del PUE en el Datacenter: http://blog.aodbc.es/2011/11/22/medida-del-pue-en-el-datacenter/

ANEXOS

ANEXO 1

Instalación Zabbix

Zabbix es un sistema de monitorización de redes, diseñado para monitorizar y registrar el estado de varios servicios de red, servidores y hardware en general, es un sistema open source de código abierto que se instala en un sistema anfitrión de las mismas características.

Para instalar el sistema se debe descargar los paquetes mediante la siguiente instrucción.

```
# rpm -Uvh https://repo.zabbix.com/zabbix/4.2/rhel/7/x86_64/zabbix-release-4.2-1.el7.noarch.rpm
# yum clean all
```

El sistema zabbix utiliza tres instancias para su funcionamiento, las cuales son, server, frontend y agente, de debe descargar los paquetes correspondientes mediante la siguiente instrucción.

```
# yum -y install zabbix-server-mysql zabbix-web-mysql zabbix-agent
```

Los datos almacenados en el monitoreo utilizan una base de datos creada en MySQL, a la cual se debe acceder y proceder a crear utilizando las siguientes instrucciones.

```
# mysql -uroot -p
password
mysql> create database zabbix character set utf8 collate utf8_bin;
mysql> grant all privileges on zabbix.* to zabbix@localhost identified by 'password';
mysql> quit;
```

Creada la base de datos se debe inicializar el esquema correspondiente, con los datos creados al momento de creación de la base de datos, esto se consigue mediante la siguiente instrucción.

```
# zcat /usr/share/doc/zabbix-server-mysql*/create.sql.gz | mysql -uzabbix -p zabbix
```

Configuración de la base de datos del servidor zabbix

Editar el archivo /etc/httpd/conf.d/zabbix.conf y añadir la zona horaria correspondiente.

PHP_VALUE DATE.TIMEZONE AMERICA/GUAYAQUIL

Iniciar el servidor zabbix y los procesos del agente

Los servicios y procesos del servidor zabbix se inician mediante las siguientes instrucciones.

```
# systemctl restart zabbix-server zabbix-agent httpd
# systemctl enable zabbix-server zabbix-agent httpd
```

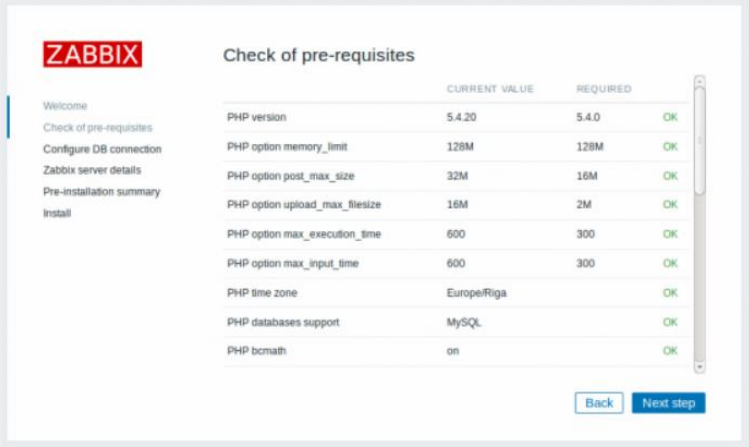
Configuración de la interfaz zabbix

Abrimos una página web con la dirección IP del servidor zabbix

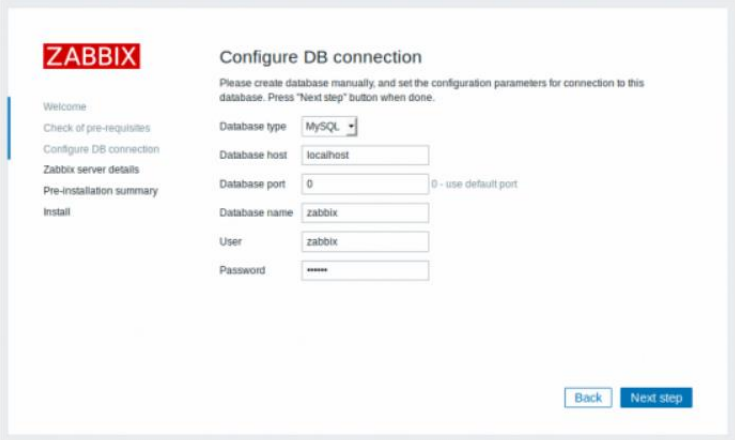
http://10.175.1.211/zabbix

Se muestra el asistente de instalación para comenzar pulsamos el botón Next Step.

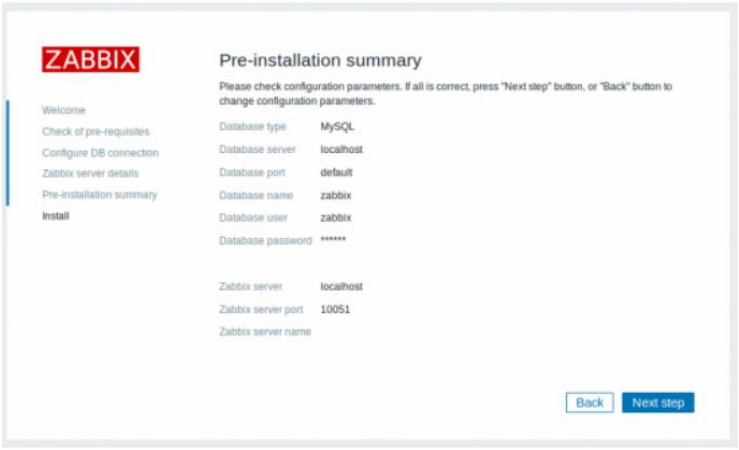
En la siguiente ventana validar que se cumplan los requisitos previos del software, si todo está correcto pulsamos el botón Next Step.



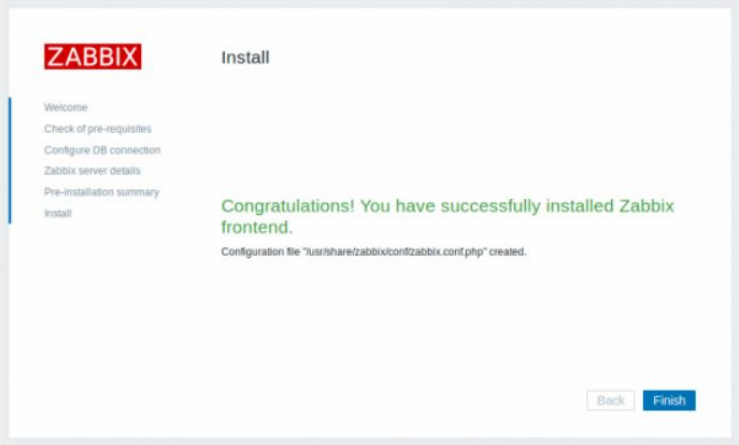
En la siguiente ventana ingresamos los datos correspondientes a la base de datos, usuario y contraseña, los datos que fueron creados en los pasos anteriores, una vez ingresado lo solicitado pulsamos el botón Next Step.



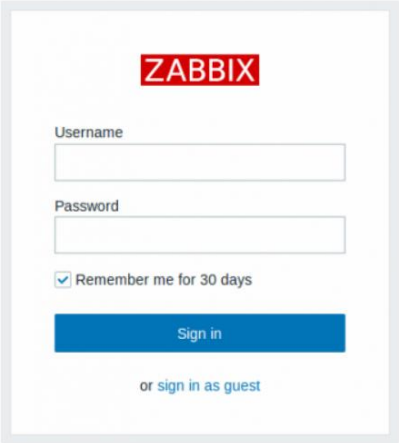
En la siguiente ventana nos muestra los parámetros configurados, si todo está bien pulse el botón Next Step y el asistente empezará con la instalación de la interface web del servidor zabbix.



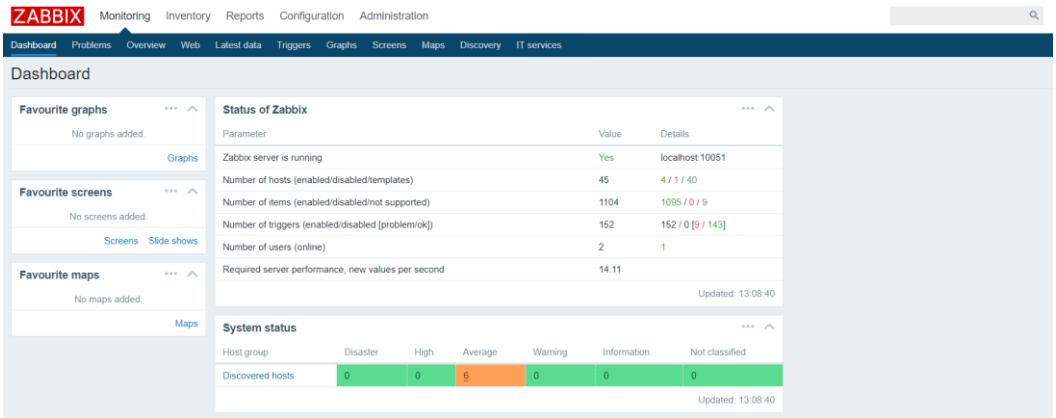
La siguiente ventana mostrará el fin de la instalación siempre y cuando no encontró fallos en los parámetros configurados.



En la siguiente ventana se muestra la página de ingreso, los datos de acceso son los por defecto, administrador y zabbix como contraseña.



La siguiente ventana se muestra la pantalla principal del sistema de monitoreo zabbix y se puede empezar con el uso del mismo.



ANEXO 2

A continuación, se adjunta la ficha técnica de los equipos de red Nexus Cisco 3524

Table 6. Cisco Nexus 3548 and 3524 specifications

Specification	Cisco Nexus 3548	Cisco Nexus 3524
Physical	• 48 fixed SFP+ ports (1 or 10 Gbps) • Dual redundant hot-swappable power supplies • Four individual redundant hot-swappable fans • One 1-PPS timing port, with the RF1.0/2.3 QuickConnect connector type* • Two 10/100/1000-Mbps management ports • One RS-232 serial console port • One USB port • Locator LED • Locator LED button	• 24 fixed SFP+ ports (1 or 10 Gbps); expandable to 48 ports • Dual redundant hot-swappable power supplies • Four individual redundant hot-swappable fans • One 1-PPS timing port, with the RF1.0/2.3 QuickConnect connector type* • Two 10/100/1000-Mbps management ports • One RS-232 serial console port • One USB port • Locator LED • Locator LED button
Performance	• 960-Gbps switching capacity • Forwarding rate of 720 million packets per second (mpps) • Line-rate traffic throughput (both Layer 2 and 3) on all ports • Configurable maximum transmission units (MTUs) of up to 9216 bytes (jumbo frames)	• 480-Gbps switching capacity • Forwarding rate of 360 mpps • Line-rate traffic throughput (both Layer 2 and 3) on all ports • Configurable MTUs of up to 9216 bytes (jumbo frames)
Typical operating power	• 152 watts (W)	• 142W
Maximum power	• 265W	• 245W
Typical heat dissipation	• 519 BTUs per hr	• 484 BTUs per hr
Maximum heat dissipation	• 904 BTUs per hr	• 835 BTUs per hr

	Mode	Normal mode	Warp mode
Hardware tables and scalability	Number of MAC addresses	64,000	8000
	Number of IPv4 unicast routes	24,000	4000
	Number of IPv4 hosts	64,000	8000
	Number of IPv4 multicast routes	8000	8000
	Number of VLANS	4096	
	Number of ACL entries	4096	
	Number of spanning-tree instances	Rapid Spanning Tree Protocol (RSTP): 512 Multiple Spanning Tree (MST) Protocol: 64	
	Number of EtherChannels	24	
	Number of ports per EtherChannel	24	
	Buffer size	6 MB shared among 16 ports; 18 MB total	
	System memory	2 GB (3524 and 3548 models) 4 GB (3524-X and 3548-X models) 16 GB (3524-XL and 3548-XL models)	
	Boot flash memory	2 GB (3524 and 3548 models) 4 GB (3524-X and 3548-X models) 16 GB (3524-XL and 3548-XL models)	
Power	Number of power supplies	2 (redundant)	
	Power supply types	• AC (forward and reversed airflow) • DC (forward and reversed airflow)	
	Input voltage	100 to 240 VAC	
	Frequency	50 to 60 Hz	
	Power supply efficiency	89 to 91% at 220V	
Cooling	Forward and reversed airflow schemes • Forward airflow: Port-side exhaust (air enters through fan tray and power supplies and exits through ports) • Reversed airflow: Port-side intake (air enters through ports and exits through fan tray and power supplies) Four individual, hot-swappable fans (3+1 redundant)		
Environment	Dimensions (height x width x depth)	1.72 x 17.3 x 18.38 in. (4.36 x 43.9 x 46.7 cm)	
	Weight	17.4 lb (7.9 kg)	
	Operating temperature	32 to 104° F (0 to 40°C)	
	Storage temperature	-40 to 158° F (-40 to 70°C)	
	Relative humidity (operating)	• 10 to 85% noncondensing • Up to 5 days at maximum (85%) humidity • Recommend ASHRAE data center environment	
	Relative humidity (nonoperating)	5 to 95% noncondensing	
	Altitude	0 to 10,000 ft (0 to 3000m)	

A continuación, se adjunta la ficha técnica de los equipos de cómputo Cisco UCS 5100 Series.

Table 2. Product Specifications

Item	Specification			
Height	10.5 in. (26.7 cm); 6RU			
Width	17.5 in. (44.5 cm); fits standard 19-inch square-hole rack			
Depth	32 in. (81.2 cm)			
Blade server half-width slots	8			
I/O slots	2			
Fabric extenders	• Cisco UCS 2204XP with 4 x 10 Gigabit Ethernet external ports and 16 x 10 Gigabit Ethernet internal ports • Cisco UCS 2208XP with 8 x 10 Gigabit Ethernet external ports and 32 x 10 Gigabit Ethernet internal ports • Cisco UCS 2304 with 4 x 40 Gigabit Ethernet external ports and 4 x 40 Gigabit Ethernet internal ports • All ports Fibre Channel over Ethernet (FCoE) capable			
Fabric interconnect	Cisco UCS 6324 with 4 x 10-Gbps uplinks, 1 x 40-Gbps Enhanced Quad Small Form-Factor Pluggable (QSFP+) expansion port, and 16 x 10-Gbps internal ports • All ports Fibre Channel over Ethernet (FCoE) capable			
Power supplies		AC power supply	-48V DC power supply	200 to 380V DC power supply
	Input voltage	100 to 120V AC 200 to 240V AC	-40 to -62V DC	200 to 380V DC
	Maximum output power	1300 watts (W) at 100 to 120V input 2500W at 200 to 240V input	2500W	2500W
	Frequency	50 to 60 Hz	-	-
	Efficiency	94%	92%	94%
	Redundancy	Nonredundant, N+1 redundant, and N+N grid redundant		
Fans	8 hot-swappable fans			

Management	• Cisco UCS 6200 Series Fabric Interconnects provide management for mutichassis configurations • Cisco UCS 6300 Series Fabric Interconnects provide management for mutichassis configurations • Cisco UCS 6324 Fabric Interconnect provides management for single/dual-chassis implementations
Backplane	1.2 Tbps of aggregate throughput; supports 10BASE-KR connections for 8 blades
Temperature: Operating	50 to 95°F (10 to 35°C) (as altitude increases, maximum temperature decreases by 1°C per 300m)
Temperature: Nonoperating	-40 to 149°F (-40 to 65°C); maximum altitude is 40,000 ft
Humidity: Operating	5 to 93% noncondensing
Humidity: Nonoperating	5 to 93% noncondensing
Altitude: Operating	0 to 10,000 ft (3000m); maximum ambient temperature decreases by 1°C per 300m
Altitude: Nonoperating	40,000 ft (12,000m)

Specification	Description
Regulatory compliance	Products comply with CE Markings per directives 2004/108/EC and 2006/108/EC
Safety	• UL 60950-1 • CAN/CSA-C22.2 No. 60950-1 • EN 60950-1 • IEC 60950-1 • AS/NZS 60950-1 • GB4943
EMC: Emissions	• 47CFR Part 15 (CFR 47) Class A (FCC Class A) • AS/NZS CISPR22 Class A • CISPR2 2 Class A • EN55022 Class A • ICES003 Class A • VCCI Class A • EN61000-3-2 • EN61000-3-3 • KN22 Class A • CNS13438 Class A
EMC: Immunity	• EN50082-1 • EN61000-6-1 • EN55024 • CISPR24 • EN300386 • KN 61000-4 Series

A continuación, se adjunta la ficha técnica de los equipos de almacenamiento EMC VNXe3200.

TheVNXe3200 is the most affordable unified hybrid flash and unified all flash storage system, bringing the power of EMC's VNX® to the IT generalist.

The VNXe3200™ delivers industry-recognized affordability, simplicity, and efficiency along with support for MCx™ multicore optimization, FAST™ Cache SSD caching, FAST VP auto-tiering, and Fibre Channel host connectivity. These enterprise-class features were previously reserved for higher-end storage systems but are standard with both the hybrid flash and all flash systems.

With the award-winning ease of use of Unisphere™ Management Software, the VNXe3200's deep integration with VMware and Microsoft for simplified provisioning and deploying virtualized applications, and EMC's legendary support, there is no need to be a storage expert to take advantage of these new and powerful storage systems.

The VNXe3200

Specifications

VNXe ELECTRICAL SPECIFICATIONS

Requirement	VNXe3200 Processor Enclosure (3.5" Drives)	VNXe3200 Processor Enclosure (2.5" Drives)	VNXe3200 Expansion Enclosure (12 x 3.5" Drives)	VNXe3200 Expansion Enclosure (25 x 2.5" Drives)
AC Line Voltage	100 to 240 V ac± 10%, single-phase, 47 to 63 Hz	100 to 240 V ac± 10%, single-phase, 47 to 63 Hz	100 to 240 V ac± 10%, single-phase, 47 to 63 Hz	100 to 240 V ac± 10%, single-phase, 47 to 63 Hz
AC Line Current	5.2A max at 100 V ac, 2.6 A max at 200 V ac	4.93A max at 100 V ac, 2.47A max at 200 V ac	2.5 A max at 100 V ac, 1.3 A max at 200 V ac	2.5 A max at 100 V ac, 1.3A max at 200 V ac
Power Consumption	520 V ac (470 W) max	493 V ac (443 W) max	250 V ac (240 W) max	250 V ac 230 W) max
Power Factor	0.98 min at full load, low voltage	0.98 min at full load, low voltage	0.98 min at full load, low voltage	0.98 min at full load, low voltage
Heat Dissipation	1.69 x 10^6 J/hr, (1604 Btu/hr) max	1.59 x 10^6 J/hr, (1512 Btu/hr) max	8.64 x 10^5 J/hr, (820 Btu/hr) max	8.28 x 10^5 J/hr, (785 Btu/hr) max
AC Protection	15 A fuse on each power supply, both phases	15 A fuse on each power supply, both phases	15 A fuse on each power supply, both phases	10 A fuse on each power supply, both phases
AC Inlet Type	IEC320-C14 appliance coupler, per power supply	IEC320-C14 appliance coupler, per power supply	IEC320-C14 appliance coupler, per power supply	IEC320-C14 appliance coupler, per power supply
Ride-through Time	12 ms min	12 ms min	30 ms min	30 ms min
Current Sharing	± 5 percent of full load, between power supplies	± 5-percent of full load, between power supplies	± 15 percent of full load, between power supplies	± 10 percent of full load, between power supplies

VNXe PHYSICAL DIMENSIONS (APPROXIMATE)

	VNXe3200 Processor Enclosure (3.5" Drives)	VNXe3200 Processor Enclosure (2.5" Drives)	VNXe3200 Expansion Enclosure (12 x 3.5" Drives)	VNXe3200 Expansion Enclosure (25 x 2.5" Drives)
Dimension (H/W/L)	3.40 in x 17.5 in x 20.0 in/ 8.64 cm x 44.45 cm x 50.8 cm	3.40 in x 17.5 in x 17.0 in/ 8.64 cm x 44.45 cm x 43.18 cm	3.40 in x 17.5 in x 20.0 in/ 8.64 cm x 44.45 cm x 50.8 cm	3.45 in x 17.5 in x 13 in/ 8.76 cm x 44.45 cm x 33.02 cm
Weight (max)	61.8lb/28.1kg	51.7 lb/23.5 kg	52.0 lb/23.6 kg	48.1 lb/21.8 kg

A continuación, se adjunta la ficha técnica del sistema de alimentación interrumpida UPS APC Symmetra RM

APC Symmetra RM

Installation Manual

English

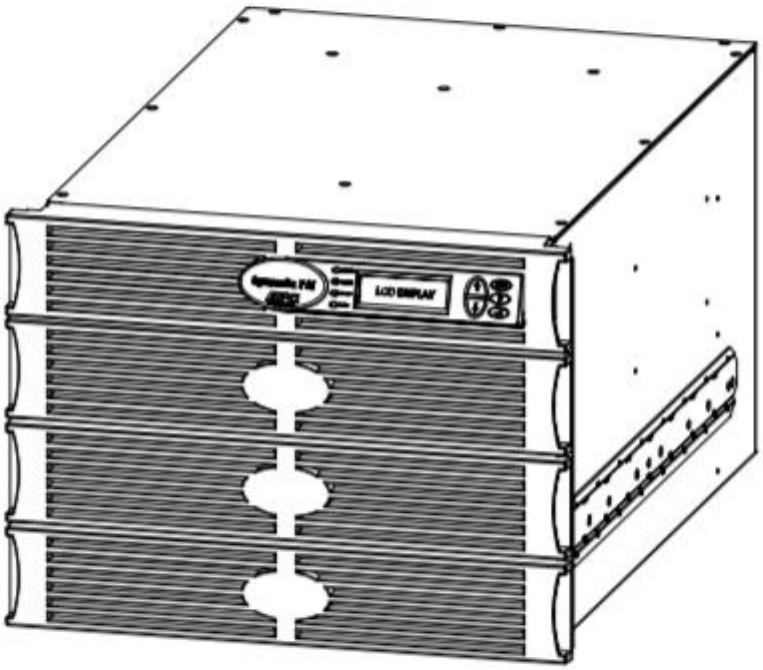
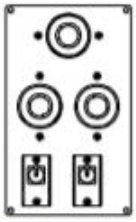
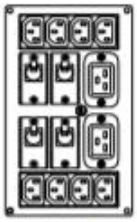
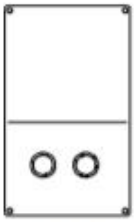


Table 3: Optional Accessories

	Model Number	Description	North America and 208/240V (Ø- Ø -G)	Europe and 230V (Ø- N -G)	Japan and 200 V (Ø- Ø -G)
Power Distribution Options	SYTF2	208 Vac to 120 Vac, 5 kVA Step-down Transformer with (12) 5-20 receptacles	Yes		
	SYTF2J	200 Vac to 100 Vac, 3.5 kVA Step-down Transformer with (12) 5-20 receptacles			Yes
	SYPD3	(2) L6-20 and (1) L6-30 receptacles	Yes		Yes
	SYPD4	(8) IEC320-C13 and (2) IEC320-C19 receptacles		Yes	
	SYPD5	(8) IEC320-C13 and (2) IEC320-C19 receptacles	Yes		Yes
	SYPD6	Output hardwiring kit	Yes	Yes	Yes
	SYPD7	(3) L6-20 receptacles	Yes		Yes
 SYPD3/SYPD7  SYPD4/SYPD5  SYPD6					
Extended Run Options	SYRMXR4	UPS rack mount 4U extended run battery cabinet (holds up to 4 battery modules)	Yes		
	SYRMXR4I	UPS rack mount 4U extended run battery cabinet (holds up to 4 battery modules)		Yes	
	SYRMXR4J	UPS rack mount 4U extended run battery cabinet (holds up to 4 battery modules)			Yes
Smart Slot Management Options	AP9608	Out-of band management card	Yes		
	AP9612TH	Environmental monitoring card			
	AP9610	Relay I/O card			
	AP9615	5-port 10Base-T hub			

UPS Specifications

This section contains operation, input, output, physical, and compliance specifications for the UPS.

Operational Specifications	
System	Power Array with hot-swappable modules that are redundant, scalable self-diagnosing, and fault-tolerant
Topology	On-line, double conversion with input power factor correction
Power Capacity	2 – 6 kVA N+1
Battery Type	Hot-swappable, sealed, maintenance-free, lead acid, 3 to 5 years life
Battery Charger	Automatic float, equalize high frequency PWM charger
Inverter	IGBT, high frequency PWM, microprocessor controlled
Battery Recharge Time	< 4 hours with standard supplied packs in the frame
Extended Battery Option	Yes
Ambient Temperature	0 – 40 °C
Relative Humidity	95% non-condensing
Elevation	0 – 10,000 ft
Input Specifications	
Nominal Input Voltage	200, 208, 220, 230, 240 Vac; 60 or 50 Hz, 1 phase, 3 wire
Input Voltage Range	155 to 276 Vac with batteries charging & supporting full load
Input Frequency Range	47 – 63 Hz
Input Power Factor	Approximately. 0.98 @ full load
Input Current THD	Approximately. 6% @ full load
Input Inrush Current	Maximum 150% of full load current
Input Generator Sizing	1.5 x frame capacity – feeder friendly, no significant oversizing; allow for battery charging and system efficiency

Output Specifications	
Nominal Output Voltage	200, 208, 220, 230, 240 Vac; 50 or 60 Hz, 1 phase, 3 wire
Output Power kVA	2 – 6 kVA
Output Power kW	1.4 – 4.2 kW
Load Power Factor	0 – 1
Output Frequency	60 or 50 Hz nominal
Output Voltage Regulation Steady State	< ± 3% for no load to full load, min ac input to max ac, min dc to max dc, linear or non-linear load or any combination
Output Voltage Regulation Transient/Dynamic	< ± 5% for 100% load application or removal, linear or non-linear load
Recovery Time	< 10 milliseconds (i.e. half cycle to steady state)
Total Harmonic Distortion	< 2% for linear loads; 5% for retention loads.
Load Crest Factor Supported	< 5% for 100% non-linear loads up to 5:1
Overload Capacity	130% for 10 minutes. With N+1
Efficiency	Approximately 89% @ full load—linear or non-linear loads
Physical Specifications	
Audible Noise	< 62 dBA
Dimensions (H x W x D)	14 in x 19 in x 28.25 in (with bezel) (35.6 cm x 48.3 cm x 71.6 cm)
Weight – Fully Loaded	Approximately 294 lb (133.6 kg)
Heat Dissipation (Full Load)	1290 BTUs typical – Batteries charged 3300 BTUs typical – Batteries charging
Compliance Specifications	
VDE-GS Certifications	EN 60950, EN 50091-1-1, EN 50091-2, IEC 60950, IEC 146-4, VDE 0558. and VDE 0805
UL Listing	UL 1778
CSA Certification	CSA 107.1