Implementación de una arquitectura de Big Data usando Hortonworks

Flor Sanabria Mata, Jorge Loría Solano y Luis Sánchez Segura

Escuela de Ingeniería,
Universidad Latinoamericana de Ciencia y Tecnología,
ULACIT, Urbanización Tournón, 10235-1000
San José, Costa Rica
[fsanabriam532,jlorias739,lsanchezs803]@ulacit.ac.cr
http://www.ulacit.ac.cr

Resumen Hoy el uso de tecnologías han hecho que la generación de datos se incremente de forma exponencial, dando cabida al concepto de Big Data. Sin embargo el problema en la actualidad no es la generación exagerada de datos, sino la forma en que se almacenan, la velocidad en que se analizan y se obtienen resultados. El siguiente trabajo mostará cómo se diseña la herramienta Hortonworks al lograr poner a disposición una plataforma que permitirá el análisis de datos masivos. Se explicará qué pruebas funciona para comprobar el funcionamiento y capacidad de la herramienta utilizando datos de muestras que se logran obtener, con resultados satisfactorios.

Keywords: BigData, Hortonworks, Tecnologías

1. Introducción

El origen de la disponibilidad de información se debe en gran parte a las tecnologías de comunicación. Las grandes empresas y organizaciones por lo general llevan a cabo sus operaciones de forma distribuida, en diferentes países y regiones. En este contexto, el uso de las tecnologías de comunicación han hecho posible el acceso en tiempo real a la información que es producida durante las operaciones. Lo cual facilita la toma decisiones de forma más rápida y precisa, a partir del análisis de esa información.

Con base en lo anterior, es posible afirmar que el auge de las tecnologías relacionadas con Big Data[1] es una consecuencia de la mayor disponibilidad de información. Esa disponibilidad es posible por el surgimiento de nuevas tecnologías de comunicación y dispositivos móviles que permiten acceder, generar y transmitir información sin importar la localización de los usuarios.

Los datos que actualmente se generan y analizan, denominados como Big Data, provienen de diferentes fuentes e incluyen detalles de preferencias del

[1] Big Data es la forma en que comúnmente se le llama al deposito de grandes cantidades de datos. Su traducción corresponde a Datos Masivos.

mercado, tendencias en redes sociales y campos más específicos, como la bioinformática. Conviene, por tanto, considerar las cinco características de Big Data: volumen, variedad, velocidad, valor y veracidad (Saporito, 2013).

El análisis de enormes cantidades de datos demanda un gran número de recursos, de procesamiento, almacenamiento y memoria. Pero las necesidades de los grupos de investigación, exigen resultados confiables y en menos tiempo. La experimentación por lo general requiere contar con el conocimiento de los resultados previos para hacer modificaciones a los experimentos (i.e., algoritmos) y volver a ejecutarlos. Por lo que el resultado de los análisis permite acelerar el proceso de experimentación en el campo.

Como consecuencia, el objetivo de este trabajo es diseñar, implementar y probar una infraestructura de Big Data basada en Hortonworks para el análisis de grandes datos, para lo cual se lleva a cabo la instalación de los elementos necesarios y se realiza un caso de uso simple con datos de pruebas. En línea con lo anterior, esta investigación busca responder la siguiente pregunta de investigación:

¿Cómo diseñar e implementar una infraestructura de Big Data para el análisis de datos masivos utilizando Hortonworks?

Desde este punto es necesario considerar los esfuerzos que diferentes organizaciones están realizando para crear componentes, herramientas y frameworks[2] y así facilitar el procesamiento y análisis de datos. Con este orden se trabaja de forma colaborativa con el objetivo de implementar herramientas que sean acogidas de forma exitosa por la comunidad de usuarios. Es relevante mencionar, que la confiabilidad de las herramientas que se producen son un elemento crítico para la investigación conjunta de muchas personas, por lo cual las características de análisis, escalabilidad, seguridad y confiabilidad son principios que se deben tener en consideración. Tomando en cuenta esto y la madurez del desarrollo de Hortonworks, se tomó la decisión de probarlo para diseñar una infraestructura de Big Data y asi analizar datos de prueba, dado la necesidad de apoyar a los profesionales que requieren este tipo de tecnologías a su alcance.

Durante el proceso de esta investigación se implementó una infraestructura virtual, utilizando los diferentes componentes de Hortonworks, que permitió desarrollar pruebas para evaluar la funcionalidad de sus componentes, y además se probó su funcionamiento mediante un caso de uso simple al utilizar datos de pruebas. Por lo que en el presente artículo se muestran los principales resultados de la implementación del framework y del procesamiento de los datos reales.

En síntesis, el resto de este trabajo lleva a cabo una revisión de varias investigaciones relacionadas (sección 2), presentan los detalles de la implementación de la infraestructura utilizando Hortonworks (sección 1), discuten los resultados (sección 1) y realizan las conclusiones (sección 1).

[2] Traducido significa marco de trabajo. Es un conjunto de herramientas relacionadas que ayudan a desarrollar diferentes objetos o componentes de un proyecto

2. Antecedentes

El diseño e implementación de una infraestructura para Big Data requiere tener en consideración elementos como la capacidad de procesamiento y almacenamiento de los dispositivos, el ancho de banda disponible, la escalabilidad, eficiencia, flexibilidad, confiabilidad y seguridad (Merelli, Pérez-Sánchez, Gesing, y D'Agostino, 2014) . Debido a lo anterior, el interés de las organizaciones por utilizar servicios de computación en la nube se incrementa día con día, por razones de escalabilidad y simplificación de la administración de la infraestructura y procesos (Shangyun Xia1 y cols., 2013). Por lo que conviene tener en cuenta que en el contexto de computación en la nube y Big Data, un aliado importante es la virtualización, porque en conjunto permiten sacar máximo provecho a los recursos y su rendimiento ("Data Virtualization for Big Data: How to Choose the Right Integration Model.", 2012).

Sin embargo, se debe considerar que la migración de servicios o elementos de infraestructura a la nube tiene diferentes implicaciones de acuerdo con el sector al cual pertenecen las organizaciones, los tipos de información que gestionan y la legislación que deben acatar de acuerdo con el país o países en que operan (Oppenheim, 2012) (Andrikopoulos, Binz, Leymann, y Strauch, 2013) (Anderson, 2010). A lo cual se debe agregar los cambios en las políticas de protección de información de los países (Crown, 1998) (de Protección de Datos, 2014) (de Elecciones Normativa, 2013).

De acuerdo con varias investigaciones, la computación en la nube no ofrece garantías suficientes de seguridad para hacer una migración de datos sensibles o elementos críticos de infraestructura. Sobre este particular, varios estudios han señalado problemas de seguridad y desconfianza de parte de las organizaciones en relación con su confiabilidad (Goyal y Supriya, 2013).

En relación con Big Data, el uso de algunas herramientas asociadas como Hadoop y bases de datos NoSQL se han extendido en poco tiempo, lo cual se contrapone con el uso de tecnologías que en el pasado que tardaban años en ser adoptadas. **En el trabajo publicado por Mitchell se previene sobre las tecnologías que tienen poco tiempo en el mercado y que no han madurado lo suficiente para ser adoptadas con confianza (Mitchell, 2014). Esta es una consideración que es conveniente resaltar no solo en este contexto, sino en general cuando de técnica se refiere: en las organizaciones de gran tamaño resulta conveniente adoptar tecnologías maduras y estables que produzcan resultados que no causen inestabilidad.**

Las herramientas de análisis de Big Data son de utilidad en un gran número campos, como la biotecnología, en la cual se realizan simulaciones computacionales en lugar de pruebas de laboratorio con el uso de probetas, como en el pasado (Greengard, 2014).

En esta investigación se considera que la gestión de datos masivos por medio de tecnología es imprescindible, debido a la necesidad de contar con métodos que permitan almacenar grandes volúmenes de datos (e.g. métodos de compresión para reducir el espacio requerido por los datos) y reducir el tiempo de análisis

para obtener resultados más precisos. Esto es de importancia para contribuir con el desarrollo de investigaciones médicas, agrícolas y la ciencia de los alimentos (Marx, 2013).

En este contexto el uso de computación en la nube resulta de gran interés para la implementación de infraestructuras de Big Data, por las posibilidades que ofrece para desarrollar soluciones escalables y el uso óptimo de recursos financieros (Shangyun Xia1 y cols., 2013). De forma particular, en el campo médico se han documentado varios esfuerzos para recolectar y analizar datos de genomas con el fin de encontrar patrones, que permitan comprender enfermedades como el cáncer mediante el uso de computación en la nube. Lo anterior con el fin de facilitar la realización de terapias y tratamientos, y además entender los efectos que pueden tener en los pacientes (Savage, 2014).

En este último punto se logran visualizar muchos esfuerzos en investigación, incluso grandes esfuerzos económicos de instituciones alrededor del mundo que permitan el desarrollo e investigación de herramientas y técnicas. Para mencionar un ejemplo las empresas de microprocesadores más populares a nivel global están desarrollando productos dirigidos específicamente a la heterogénea, es decir, sin importar la marca o tipo de procesador, incluso la infraestructura de la plataforma tecnológica que se esté utilizando, logrando que las empresas no se vean aferradas a la exclusiva de una marca o plataforma y busquen incrementar la potencia de procesamiento donde los científicos implicados en el proceso se vean beneficiados a la hora de realizar sus análisis y obtención de resultados (Merelli y cols., 2014).

3. Resultados

El desarrollo del presente trabajo requirió la realización de una serie de tareas usando una infraestructura virtual con VMware ESXi 5.5 como hipervisor o sistema operativo base y dos servidores marca HP que funcionan como nodos. Los recursos de los servidores son administrados por VMware y sobre este se ejecutan las maquinas virtualizadas. Esta infraestructura no cuenta con sistema de almacenamiento compartido, por lo que la administración de este debe realizarse por separado, limitando algunas funcionalidades que la virtualización nos brinda, por ejemplo mover máquinas de un nodo a otro, permitiendo la continuidad de funcionamiento.

En la figura 1 se observa la infraestructura utilizada durante el desarrollo de este trabajo, donde se implementaron dos nodos de Hortonworks, uno en cada servidor de la plataforma de virtualización. Estos servidores están conectados por 3 interfaces de red cada uno a un switch marca Cisco, además cada servidor cuenta con una interfaz adicional para administración. Esta infraestructura es proporcionada por la Universidad Latinoamericana de Ciencia y Tecnología para el desarrollo del presente trabajo.

En la figura 2 se observan las tareas realizadas para implementar y probar las funcionalidades de Hortonworks sobre la infraestructura virtual antes mencionada, usando como base el Sandbox de este framework. El proceso inicial

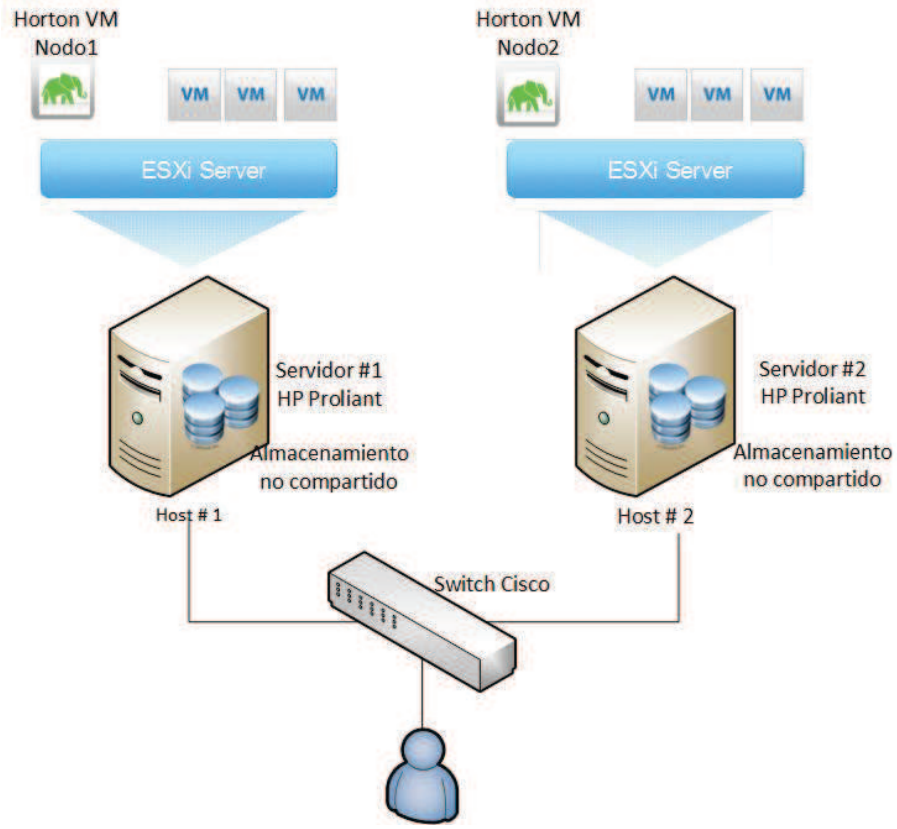

Figura 1. Diagrama de plataforma virtual donde se implementó Hortonworks.

consistió en descargar de la página de Hortonworks el Sandbox con la versión 2.2.4.2, la cual provee una imagen de tipo OVA[3] que se utiliza para empaquetar y compartir máquinas virtuales independientemente del hipervisor que se utilice.

Seguidamente se importó la imagen OVA en el ambiente VMware utilizando un cliente de "VMware vSphere", por defecto la máquina del Sandbox viene con 4 GB de memoria RAM, pero durante las pruebas realizadas en el Sandbox se observó que sin carga o con poca carga el sistema operativo realizaba paginación de hasta 4 GB, por lo que se optó por cambiar la cantidad de memoria RAM a 8 GB para el Sandbox.

Como siguiente paso se inicia la máquina del Sandbox e inician todos los servicios que componen el framework de Hortonworks ya que por defecto vienen deshabilitados. A este momento el Sandbox es funcional y se pueden realizar cargas de datos, pero es conveniente entender como agregar nuevos nodos para utilizar la escalabilidad que permite el framework.

[3] Dispositivo de Virtualizacion abierta por sus siglas en ingles

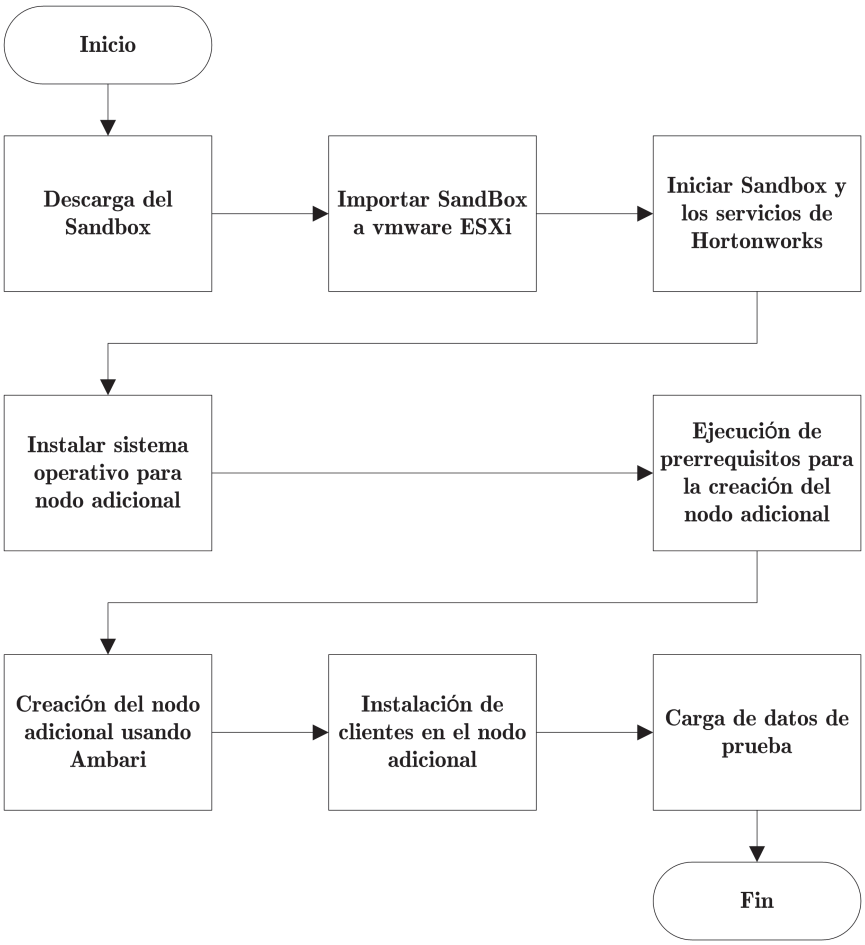

Figura 2. Tareas para la implementación de una infraestructura de Big Data al usar Hortonworks.

Dado lo anterior se procede a instalar un nuevo nodo, para lo cual se requiere instalar un sistema operativo Linux dado que el administrador de cluster de Hortonworks conocido como Ambari posee la capacidad de instalar todos los componentes requeridos para un nuevo nodo de forma automática siempre que el mismo este en un sistema operativo Linux como Red Hat, SUSE LES o CentOS, en el caso particular de este trabajo se utilizó CentOS 6.6 ya que es la misma versión de sistema operativo que posee el Sandbox.

Para la correcta creación de un nuevo nodo el mismo debe cumplir una serie de prerrequisitos los cuales se listan a continuación, como parte de este trabajo se encontrará como anexo un detalle del paso a paso de cómo realizar las tareas a continuación mencionadas:

1. Deshabilitar el firewall del nuevo nodo.
2. Deshabilitar el SELinux del nuevo nodo.
3. Asegurar que el DNS resuelva los nombres del Sandbox y el nuevo nodo, de lo contrario modificar los archivos de host correspondientes para que los nombres sean resueltos.
4. Configurar la comunicación ssh utilizando llaves privadas para que el Sandbox pueda realizar conexiones directas al nuevo nodo sin requerir password.
5. Instale Open Java JDK 1.7 en el nuevo nodo.
6. Inicie el servicio de NTP para asegurarse de que el reloj del nuevo nodo se encuentre sincronizado.
7. Configure la zona horario del nuevo nodo igual que el Sandbox.
8. Aumente el límite de archivos abiertos para todos los usuarios al menos en 63536 archivos.
9. Aumente el tiempo de espera de ejecución de comandos de Ambari, esto evitará que una conexión a internet menor a 2 Mbps repercuta en agotar el tiempo de espera de la instalación del nuevo nodo.

Una vez cumplidos los requisitos anteriores se procede a crear el nuevo nodo utilizando la interfaz web de Ambari, este proceso puede variar dependiendo de la conexión a internet, ya que Ambari configura en el sistema operativo los repositorios de Hortonworks para descargar e instalar el software, por ejemplo en el laboratorio se tardó alrededor de 14 minutos con una conexión de internet de 40 Mbps y en otra prueba con una conexión de 2 Mbps se tardó alrededor de 40 minutos.

Es importante destacar que uno de los procesos conocido como HCAT no está instalado en el sandbox, en su lugar viene un proceso llamado webHCAT que cumple la misma funcionalidad dando una interfaz web, dado esto cada vez que se intenta instalar un nuevo nodo y si se selecciona la opción de instalar los clientes, Ambari intenta instalar todos los clientes disponibles incluyendo el cliente de HCAT el cual dará error indicando que el servicio HCAT no está instalado, esto repercutirá en que la instalación del nuevo nodo falle, por eso la instalación de los clientes debe hacerse de forma manual luego de creado el nodo, esto para evitar instalar el cliente de HCAT.

De forma seguida se realizaron pruebas de cargas de datos para verificar la funcionalidad del ambiente se ejecutaron varios de los tutoriales, entre los que

destacan carga de datos de estadísticas de baseball. Luego de la ejecución de los tutoriales se observa que la ejecución de transformaciones usando Pig, son más eficientes si se ejecutan utilizándolo en combinación con Tez.

La figura3 nos puede ayudar a entender un poco más todo este proceso de implementación que llevamos a cabo, ya que nos permite ver cada uno de los componentes de Hortonworks en el proceso de ETL[4] y como es que este proceso sigue su flujo normal. Parte de la implementación es la planificación que se debe tener, ya que no son todos los componentes de la plataforma Hortonworks los que se deben instalar siempre, esto va a depender de las funcionalidades que se deseen tener en la infraestructura para Big Data.

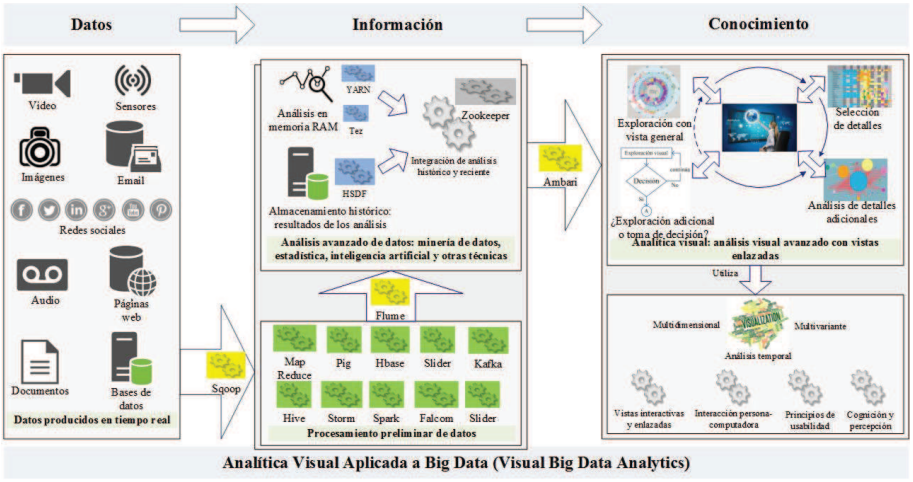

Figura 3. Analítica Visual Aplicada a Big Data. Incluye los componentes de Hortonworks.

También se creo la tabla1, donde se describen todos los componentes que poseen Hortonworks uno por uno y que forman parte del proceso de ETL. La descripción en esta tabla se realiza de forma muy general.

[4] Proceso que por sus siglas en ingles significa Extraer, Transformar y cargar, utilizado en el campo de procesamiento de Big Data

Cuadro 1. Componentes del Framework HortonWorks

Componente	**Descripción**
HDFS	Sistema de archivos compartidos de Hadoop por sus siglas en ingles. Ofrece almacenamiento escalable y tolerante a fallos. Administra y almacena los ficheros en bloquees pequeños (64 MB), consiguiendo minimizar el esfuerzo en las búsquedas.
MapReduce2	Es un proceso en bloques, que permite de forma simple procesar en paralelo grandes volúmenes de información.
YARN	Otro negociador de recursos por las siglas en inglés, es una tecnología que maneja los recursos de manera muy eficiente del clúster.
Tez	Es una herramienta para escribir aplicaciones YARN nativas. Permite que las aplicaciones de acceso a datos puedan trabajar con Petabytes[5] de datos a través de miles de nodos.
Hive	Es la herramienta por la cual se le puede hacer consultas SQL a MapReduce.
HBase	HBase es una base de datos NoSQL[6] de código abierto que ofrece acceso de lectura-escritura en tiempo real, a los grandes conjuntos de datos.
Pig	Es una plataforma creada de lenguaje menos estructurado, y se utiliza para programar en MapReduce.
Sqoop	Es el componente que permite la transferencia eficiente de grandes volúmenes datos de forma paralela entre Hadoop y otras plataformas de datos estructuradas.
Oozie	Es un sistema de flujos de trabajo que permite planificador las tareas de Hadoop. Es un motor basado en servidores especializados en la gestión de flujos de trabajo con las acciones que se ejecutan Hadoop MapReduce y Pig. Se implementa como una aplicación web que se ejecuta en Java.
Zookeeper	Es el componente que provee configuración centralizada y registro de nombres para sistemas distribuidos. Soporta alta disponibilidad mediante servicios redundantes.
Falcon	Es el componente que define horarios y supervisa las políticas de gestión de datos. Simplifica la configuración de movimiento de los datos.
Storm	Es un sistema de computación en tiempo real distribuido para procesar grandes volúmenes de datos a alta velocidad.
Flume	Nace de la necesidad de subir datos de las diferentes aplicaciones a HDFS. Se basa en flujos de streaming[7] de datos.
Slider	Es un set de herramientas para análisis de datos. Ayuda la interacción con los recursos para aplicar YARN.
Spark	Es un una librería de herramientas de código abierto para computo distribuido y análisis de datos, todo esto en memoria. Es más rápido que MapReduce.
Kafka	Streaming de datos que viene dentro del framework de Spark.
Ambari	Este componente es el que nos permite la administración y seguridad del clúster de Hadoop.
Sandbox	Es una distribución de HortonWorks virtual, la cual permite configurar de forma expedita un ambiente para pruebas.

4. Conclusiones

Antes de comenzar con la implementación de una infraestructura para Big-Data como la expuesta en el trabajo, se debe tener claro los requerimientos mismos de la aplicaciones cuanto al nivel de procesamiento se necesite, quiénes van a participar de la manipulación de estos datos y cómo los van a acceder o compartir; esto debido a que se debe tener claro qué componentes son necesarios implementar de HortonWorks, ya que la aplicación como tal trae muchos componentes, pero no necesariamente todos se deben utilizar.

La implementación debe ser ejecutada por personal que maneje las áreas de conocimiento de conectividad y redes de forma básica, manejo de bases de datos y la ejecución de consultas, administración de ambientes basados en sistemas operativos Linux, conexiones remotas SSH. Agregar que si la infraestructura se implementa en ambientes virtuales como en el caso nuestro que se utilizó VM-Ware como Hipervisor o sistema operativo base, para ello, se debe tener la capacidad de poder administrar el ambiente por medio de las distintas herramientas de administración que la plataforma provee.

Después de haber probado la plataforma de Big Data HortonWorks, con en el procedimiento de uso, anteriormente expuesto, y logramos demostrar que la herramienta cumple su objetivo. Permite la manipulación de grandes cantidades de datos, además se comprobó la escalabilidad que ofrece la plataforma y su poder de análisis. Esta escalabilidad se puede traducir también en un buen esquema de contingencia, si se toma en cuenta que hay varios nodos y alguno sufre de algún inconveniente que no le permita seguir trabajando, la plataforma se podría ver afectada en cuanto a su rendimiento, pero el servicio se seguiría brindando. Sin embargo, si el sistema operativo base de la virtualización no cuenta con las características de alta disponibilidad, nada de esta escalabilidad puede detener que un incidente en la plataforma virtual, detenga la disponibilidad del servicio Hortonworks.

5. Recomendaciones

Existen plataformas en la nube como Azure que ofrecen en pocos minutos la implementación de una plataforma utilizando Hortonworks, lo cual puede ayudar a reducir el hardware requerido para una solución de este tipo, especialmente se recomienda analizar la opción de utilizar un esquema híbrido donde los nodos adicionales se encuentren en la nube, así se puede realizar carga y procesamiento de grandes cantidades de datos bajo demanda.

---

[5] Medida utilizada en informática que que sirve para dimensionar el tamaño del almacenamiento y que corresponde a la mil Terabytes

[6] Tipo de Base de Datos diferente al modelo relacionales que todos conocemos. Como aspectos principal los datos almacenados no requieren estructuras permanentes como tablas y no garantizan ACID completamente.

[7] Es la transmisión o difusión de datos en flujo continuo, sin interrupción.

Ademas se recomienda que la implementación de esta herramienta debe ser ejecutada por personal con características técnicas, conocimientos en virtualización, Linux, nociones de protocolos como SSH, HTTP. También se debe manejar un nivel de lectura y comprensión del inglés intermedio, ya que la mayoría de referencias están escritas en este idioma.

Asimismo Se requiere recomienda que la plataforma de virtualización utilizada para alojar esta infraestructura de Big Data se configure con todos los componentes de alta disponibilidad que la virtualización cuenta. Con esto se asegura la continuidad de servicio y la no interrupción de análisis por problemas de plataforma, como por ejemplo los respectivos mantenimientos de equipos, cortes de luz o problemas de funcionamiento que pueda presentar el sistema operativo base de esta arquitectura, afecten los tiempos, los resultados o los procedimientos que puedan ejecutarse.

Asimismo se requiere un soporte de virtualización utilizada para alojar esta infraestructura de Big Data, que se configure con todos los componentes de alta disponibilidad, con la que cuente la virtualización . Con esto se asegura la continuidad de servicio y la no interrupción de análisis, por problemas de plataforma, como por ejemplo, los respectivos mantenimientos de equipos, cortes de luz o problemas de funcionamiento que pueda presentar el sistema operativo, base de esta arquitectura,como: los tiempos, los resultados o los procedimientos, los cuales, puedan afectarse.

Referencias

Anderson, W. L. (2010). Increased çloud.ªdoption accelerates the need for privacy legislation reform. *Franklin Business & Law Journal*(4), 16 - 20. pages 3

Andrikopoulos, V., Binz, T., Leymann, F., y Strauch, S. (2013). How to adapt applications for the cloud environment. *Computing*, *95*(6), 493 - 535. pages 3

Crown. (1998). Data protection act 1998. Descargado de `http://www.legislation.gov.uk/ukpga/1998/29/pdfs/ukpga_19980029_en.pdf` pages 3

Data virtualization for big data: How to choose the right integration model. (2012). *Database Trends & Applications*, *26*(1), 28. pages 3

de Elecciones Normativa, T. S. (2013). Reglamento a la ley de protecciÓn de la persona frente al tratamiento de sus datos personales. Descargado de `http://www.tse.go.cr/pdf/normativa/reglamentoleyproteccionpersona.pdf` pages 3

de Protección de Datos, A. E. (2014). Reglamento de la lopd. Descargado de `http://www.agpd.es/portalwebAGPD/canaldocumentacion/informes_juridicos/reglamento_lopd/index-ides-idphp.php` pages 3

Goyal, K., y Supriya. (2013). Security concerns in the world of cloud computing. *International Journal of Advanced Research in Computer Science*, *4*(2), 230 - 234. pages 3

Greengard, S. (2014). How computers are changing biology. *Communications of the ACM*, *57*(5), 21 - 23. pages 3

Marx, V. (2013). Biology: The big challenges of big data. *Nature*, *498*(7453), 255 - 260. pages 4

Merelli, I., Pérez-Sánchez, H., Gesing, S., y D'Agostino, D. (2014). Managing, analysing, and integrating big data in medical bioinformatics: Open problems and future perspectives. *BioMed Research International*, 1 - 13. pages 3, 4

Mitchell, R. L. (2014). 8 big trends in big data analytics. *Computerworld Digital Magazine*, *1*(3), 21 - 26. pages 3

Oppenheim, C. (2012). Cloud law and contract negotiation. *El Profesional de la Información*, *21*(5), 453 - 457. pages 3

Saporito, P. (2013). The 5 v's of big data. *Best's Review*, *114*(7), 38. pages 2

Savage, N. (2014). Bioinformatics: Big data versus the big c. *Nature*, *509*(7502), S66 - S67. pages 4

Shangyun Xia1, s., Jiang Xie2, j., Dongbo Dai1, d., Huiran Zhang1, h., Qing Nie3, q., Shigeo Kawata4, k.-u., y Wu Zhang2, w. (2013). Kvm combined with hadoop application based-on cpse-bio. *Journal of Next Generation Information Technology*, *4*(3), 160 - 166. pages 3, 4