revista digital universitaria
rev'sta.unam.mx

1 de noviembre de 2016 | Vol. 17 | Núm. 11 | ISSN 1607 - 6079

ARTÍCULO

INFRAESTRUCTURA PARA *BIG DATA*

(http://www.revista.unam.mex/vol.17/num11/art77)

Javier Salazar Argonza
(Facultad de ingeniería, UNAM)

Dirección General de Cómputo y de Tecnologías de Información y Comunicación - UNAM
Departamento de Acervos Digitales

INFRAESTRUCTURA PARA *BIG DATA*

> […] la infraestructura requerirá no solo de una inversión importante sino también de una serie de nuevas habilidades y conocimientos […]

Resumen

Se realiza un acercamiento a la infraestructura necesaria para implementar una plataforma de Big Data en una institución. Se explican los principales elementos de hardware y software que integran una solución arquitectónica de Big Data. Se habla sobre las principales diferencias entre las distribuciones comerciales de Apache Hadoop para su adopción y uso. Se dan sugerencias generales para la adquisición de los componentes.

Palabras clave: *Big Data*, infraestructura, TIC, marco de trabajo, analítica, *hadoop, clústeres*, sistemas distribuidos, arquitectura de hardware.

Infrastructure for Big Data

Abstract

It explores the infrastructure needed to implement a Big Data platform in an institution. The main elements of hardware and software for integrating an architecture solution for Big Data are explored. It also explains the main differences between the commercial distributions of Apache Hadoop for its adoption and use. General suggestions are given for the acquisition of its components.

Keywords: *Big Data, Infrastructure, TIC, work frame, analytics, Hadoop, Clusters, Distributed system, Hardware architecture*

INFRAESTRUCTURA PARA *BIG DATA*

Introducción

El término *Big Data* se empleó por primera vez en 1997 en un artículo de los investigadores de la NASA Michael Cox y David Ellsworth y se define como: "La gestión y análisis de enormes volúmenes de datos que no pueden ser tratados de manera convencional, ya que éstos superan los límites y capacidades de las herramientas de software comúnmente utilizadas para su captura, gestión y procesamiento". De hecho involucra el uso de infraestructuras, tecnologías y servicios especiales que han sido creados para dar solución especifica al procesamiento de estos enormes conjuntos de datos provenientes de múltiples fuentes tales como archivos, redes, sensores, micrófonos, cámaras, escáneres, imágenes, videos, entre otros.

El objetivo de *Big Data*, al igual que los sistemas analíticos convencionales, es convertir los datos en información útil que facilite la toma de decisiones. Esto inclusive en tiempo real, para brindar más oportunidades de negocio. El poder de éste sistema radica en que permite descubrir nueva información sobre las cadenas de valor de las instituciones o empresas para abordar problemas antes irresolubles.

Algunas empresas están utilizando *Big Data* para entender el perfil, las necesidades y el sentir de sus clientes respecto a los productos y/o servicios que ofrecen. Esto les permite adecuar la forma en que interactúan con sus clientes y como prestan sus servicios. No obstante las predicciones son aplicables a todas las ramas del quehacer humano.

Figura 1. Ejemplo de Sistema Distribuido/*Clúster* de Datos. Piso de servidores del Google *Data Center Lenoir NC USA*. Fuente: https://www.youtube.com/watch?v=avP5d16wEp0&feature=youtu.be

La evolución de la tecnología y los menores costos de almacenamiento han hecho posible que las aplicaciones de *Big Data* estén aumentando. Sin embargo, definir la infraestructura para un proyecto no es una tarea sencilla, recordemos que una plataforma tecnológica para esta actividad debe facilitar muy rápidamente la recopilación, el almacenamiento y el análisis de grandes volúmenes de datos, los cuales además pueden estar en diferentes formatos ó inclusive generándose en tiempo real, y que a diferencia de los "sistemas tradicionales" -por razones de eficiencia- la forma de tratar y analizar la información debe ser trasladada directamente a los datos sin precargarlos en memoria.

[1] Un sistema distribuido se define como una colección de computadoras separadas físicamente y conectadas entre sí por una red de comunicaciones; cada máquina posee sus componentes de *hardware* y *software* que el programador percibe como un solo sistema. En los sistemas distribuidos si un componente del sistema se descompone otro componente es capaz de reemplazarlo. (Tolerancia a fallos). El tamaño de un sistema distribuido puede ser muy variado, ya sean decenas (red de área local), centenas (red de área metropolitana) o miles y millones de hosts (Internet); esto se denomina escalabilidad. Fuente:https://es.wikipedia.org/wiki/Computaci%C3%B3n_distribuida

[2] El término *clúster* (del inglés *cluster*, "grupo" o "raíz") se aplica a los conjuntos o conglomerados de computadoras unidas entre sí normalmente por una red de alta velocidad y que se comportan como si fuesen una única computadora. Los clústeres son usualmente empleados para mejorar el rendimiento y/o la disponibilidad por encima de la que es provista por un solo computador. Un *clúster* brinda los siguientes servicios: Alto rendimiento, alta disponibilidad, balanceo de carga y escalabilidad. Fuente:https://es.wikipedia.org/wiki/Cl%C3%BAster_(inform%C3%A1tica)

Razón por la que deben considerarse sistemas distribuidos o basados en *clústeres*[2] tanto para el procesamiento como el almacenamiento de la información (Ver fig 1).

En lo referente al software requerido para administrar los recursos de una plataforma de Big Data, debido a que estamos hablando de trabajar con arreglos de computadoras (servidores) y clústeres de almacenamiento -que deben operarse en conjunto como un solo sistema– resulta evidente que se requiere de un entorno de trabajo "Framework"[3], capaz de administrar, distribuir, controlar y procesar rápidamente los datos dentro de los arreglos de sistemas computacionales y de almacenamiento.

Hoy en día el principal framework utilizado para *Big Data*, es *Hadoop*[4] cuyo desarrollo pertenece a: *The Apache Software Foundation*, misma que otorga el permiso para utilizar sus programas sin costo.

El proyecto original de *Apache Hadoop* incluye los siguientes módulos funcionales:

- *Hadoop Distributed File System (HDFS™)*: Sistema distribuido y creado para trabajar con archivos de gran tamaño escrito en Java con un muy alto desempeño. Véase: http://www.happyminds.es/apache-hadoop-introduccion-a-hdfs/#sthash.2M4rIyxS.dpbs.

- *HadoopMapReduce*: Sistema para escribir aplicaciones de procesamiento en paralelo para grandes cantidades de datos en sistemas de procesamiento distribuido o clústeres. Véase: http://www.tutorialspoint.com/es/hadoop/hadoop_mapreduce.htm.

- *Hadoop YARN: (Yet Another Resource Negotiator)*. Plataforma de trabajo que permite la programación de las tareas y la gestión de los recursos de clústeres. (Es básicamente una nueva generación del software de MapReduce MRv2 para la administración de clústeres). Véase: http://searchdatacenter.techtarget.com/es/definicion/Apache-Hadoop-YARN-Yet-Another-Resource-Negotiator

- https://unpocodejava.wordpress.com/2013/07/25/que-es-yarn/.

- *HadoopCommon*: Utilerías necesarias para soportar al resto de los módulos de Hadoop. Estas proporcionan acceso a los sistemas de archivos soportados. De hecho contiene los archivos en Java (".jar") y los scripts necesarios para hacerlo correr. Véase: https://es.wikipedia.org/wiki/Hadoop.

Otros proyectos de la Fundación de Software Apache relacionados con Hadoop son los siguientes:

- *Ambari™*: Herramienta con una interfaz web que permite la administración, aprovisionamiento y monitoreo de clústeres bajo Apache Hadoop. Véase: https://unpocodejava.

- *Avro™*: Sistema de serialización de datos. Serializa los datos en un formato binario compacto por medio del formato JSON que facilita la comunicación entre los nodos de Hadoop. Avro permite almacenar datos y acceder a ellos fácilmente desde varios

[3] *Framework*: (Entorno de trabajo). En el desarrollo de software, un framework o infraestructura digital, es una estructura conceptual y tecnológica con módulos concretos de software, que puede servir de base para la organización y desarrollo de software. Típicamente, puede incluir soporte de programas, bibliotecas y un lenguaje interpretado, entre otras herramientas, para así ayudar a desarrollar y unir los diferentes componentes de un proyecto.

[4] *Apache Hadoop*: Es un framework de software que soporta aplicaciones distribuidas bajo una licencia libre. Permite a las aplicaciones trabajar con miles de nodos y petabytes de datos. (En términos más simples, es un marco para el manejo de grandes conjuntos de datos en un entorno de computación distribuida).

lenguajes de programación. Está diseñado para minimizar el espacio que nuestros datos ocuparán en disco. Véase http://www.datasalt.es/2011/06/avro-hadoop/.

- *Cassandra™*: Base de Datos NO SQL, distribuida y basada en un modelo de almacenamiento de Clave-Valor, escrita en Java. Véase https://www.adictosaltrabajo.com/tutoriales/cassandra/.

- *Chukwa™*: Sistema de recolección de datos para la gestión de grandes sistemas distribuidos. Permite recolectar logs (bitácoras) de grandes sistemas para su control, análisis y visualización. Véase: https://unpocodejava.wordpress.com/2012/07/17/hadoop-chuwka-procesamiento-de-logs-de-grandes-sistemas-en-hadoop/.

- *HBase™*: Base de datos distribuida y escalable para el almacenamiento de tablas muy grandes de información de miles de millones de filas por millones de columnas que se encuentran alojadas en sistemas distribuidos o clústeres de almacenamiento. Véase: http://www.franciscojavierpulido.com/2013/09/bigdata-hadoop-iv-hbase_17.html.

- *Hive™*: Apache Hive es un sistema de almacenamiento de datos "Data Warehouse", de código abierto para la consulta y el análisis de grandes conjuntos de datos que se encuentran en los archivos de Hadoop. Este básicamente realiza tres funciones: La consulta, sumarización y análisis de los datos. Véase: http://searchdatamanagement.techtarget.com/definition/Apache-Hive.

- *Mahout™*: Es una librería escalable para minería de datos y aprendizaje automatizado (aprendizaje de máquinas), escrita en Java y optimizada para funcionar sobre Hadoop HDFS y MapReduce. Dispone de un gran número de algoritmos implementados para trabajar con técnicas de filtrado colaborativo (Collaborative filtering), Agrupación (Clustering) y Clasificación (Classification). Véase: https://unpocodejava.wordpress.com/2012/10/29/mahout-machine-learning-en-hadoop/.

- *Pig™*: *Apache Pig* es una plataforma para el análisis de grandes conjuntos de datos. Consta de un lenguaje de alto nivel para la expresión de programas junto con la infraestructura necesaria para la evaluación de los programas. Véase: https://pig.apache.org/ y http://www.ibm.com/developerworks/ssa/data/library/bigdata-apachepig/.

- *Spark™*: Es un motor para el procesamiento de datos a gran escala. Permite escribir rápidamente aplicaciones en los lenguajes de programación *Java*, *Scala*, *Pyton* y R. Ejecuta los programas en la memoria hasta 100 veces más rápido que *Hadoop MapReduce*, o 10 veces más rápido en el disco. Adicionalmente puede combinar comandos de SQL, flujo de datos (streaming) y análisis complejos. Véase: *http://spark.apache.org/ y https://geekytheory.com/apache-spark-que-es-y-como-funciona/*

- *Tez™*: Tez es un nuevo marco de ejecución distribuido para Hadoop. Esta herramienta convierte el modelo de *MapReduce* en una plataforma más potente

que es útil para una gran cantidad de casos de uso en donde el procesamiento de las consultas requiere de un rendimiento casi en tiempo real. Véase: http://www.infoq.com/articles/apache-tez-saha-murthy y http://tez.apache.org/

- *ZooKeeper™*: Es un servicio centralizado para mantener la información de configuración, denominación, sincronización distribuida y de servicios de grupo que requieren las aplicaciones distribuidas. Véase: http://zookeeper.apache.org/ y https://unpocodejava.wordpress.com/2010/11/19/zookeeper-se-ha-convertido-en-un-proyecto-apache-top/

Para mayor información sobre el Apache Hadoop, sus proyectos relacionados e incluso descargar el software puede acceder al siguiente sitio web: *http://hadoop.apache.org/y http://inlab.fib.upc.edu/es/blog/que-herramientas-necesitas-para-iniciarte-en-big-data*.

Distribuciones comerciales de *Hadoop*

Aunque *Hadoop* es un *framework* libre y por ende no tiene costo, en la práctica existen varias distribuciones comerciales de *Hadoop* que permiten implementar un medio ambiente de trabajo más robusto para *Big Data*. Estas distribuciones ofrecen diferentes mejoras en la funcionalidad, rendimiento, facilidad de configuración, gestión e integración con otras plataformas externas, producto de años de investigación de las empresas que los distribuyen. Incluyen además herramientas adicionales, soporte técnico, diferentes niveles de servicios y mantenimiento.

Como señala David Loshin, en su artículo Explorando distribuciones *Hadoop para gestionar big data*:

> El mercado incluye tres proveedores independientes que se especializan en *Hadoop*: *Cloudera Inc*., *Hortonworks Inc*. y *MapR Technologies Inc*. Otras empresas que ofrecen distribuciones o capacidades de Hadoop incluyen *Pivotal Software Inc*., *IBM*, *Amazon Web Services* y *Microsoft*[5].

Debemos prestar mucha atención entre las diferencias y las distribuciones de Hadoop ya que pueden requerir de una gran inversión para su funcionamiento (principalmente por contratos de licenciamiento muy caros), aportar un pobre desempeño o limitaciones en los resultados esperados, ofrecer poco soporte tecnológico o inclusive requerir de *hardware* propietario para su instrumentación *Appliances*.

Entre los principales aspectos que deberemos revisar se tiene:

- Las mejoras o beneficios que aportan a la solución. Recordemos que las variantes en los *frameworks* facilitan abordar los problemas de analítica con diferentes enfoques, adiciones y/o mejoras y por ende resultados.

- Facilidades de monitoreo y administración para *Hadoop*.

[5] David Loshin , *Explorando distribuciones Hadoop para gestionar big data*, en linea: http://searchdatacenter.techtarget.com/es/cronica/Explorando-distribuciones-Hadoop-para-gestionar-big-data.

- Los componentes específicos o propietarios de las distribuciones.
- Su compatibilidad con otras plataformas.
- El *hardware* requerido.
- Las facilidades de recuperación en caso de desastre.
- Modelo de servicio y soporte ofertado.
- Consultoría.
- Su costo.

¿Cuáles son estas distribuciones líderes en el mercado?, en términos generales que es lo que ofrece cada una de ellas. La tabla que se presenta a continuación lo resume.

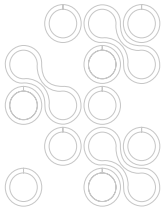
Distribución Comercial de *Hadoop*	***Ambiente operativo***	***Ofrece***
Amazone Web Servicves AWS	*Unix/Linux*	• *ElasticMapReduce* (*EMR*): Servicio web que facilita el procesamiento rápido y rentable de grandes cantidades de datos. *EMR* es *Hadoop* en la nube. Aprovecha *Amazon* EC2 para cómputo y *Amazon S3* para almacenamiento. Véase: https://aws.amazon.com/es/elasticmapreduce/. • *Amazon Kinesis* para el procesamiento de datos de streaming en tiempo real. • Integración con el *data warehouse Amazon Redshift* y otras fuentes de datos; *autoscaling* que modificará el tamaño de los *clústeres* en base a políticas; soporte para bases de datos *NoSQL* adicionales sobre *Hadoop*; y más integración de inteligencia de negocios con proveedores externos.
Cloudera	*Unix/Linux*	• *Cloudera Manager*: Herramienta de monitoreo y administración para Hadoop. • *Impala: Motor SQL* para *Hadoopmuy* rápido con una arquitectura de procesamiento paralelo masivo (*MassivelyParallelProcessing - MPP*).

Tabla 1. Principales distribuciones comerciales de Hadoop. Fuente: http://cioperu.pe/fotoreportaje/15543/hadoop-como-se-encuentran-las-distribuciones-lideres/

Hortonworks	*Unix/Linux*	• *Apache Ambari*: (Proporciona una consola de administración de *clústeres Hadoop*). • *Aliansas* con *Microsoft*, *Teradata*, *SAP*, *Red Hat* y otras empresas.
IBM	*Unix/Linux*	• *BigInsightsfor Apache Hadoop*: (Código de *Apache Hadoop* integrado con activos de *IBM* para la analítica avanzada *SPSS*, administración de cargas de trabajo para computación de alto desempeño, herramientas de inteligencia de negocio y herramientas de administración y modelamiento de datos).
MapRTechnoligies	*Unix/Linux*	• *Soporte para Network File System* (*NFS*), correr código arbitrario en el *clúster*. • *Hbase* mejorado (Base de datos para *Hadoop*). • Características de alta disponibilidad y recuperación en caso de desastre.
Pivotal	*Unix/Linux*	• *HAWQ* (Motor de consulta *SQL* paralelo de *Pivotal*). • Familia de *appliances* que integran su *Hadoop*, *EDW* (*Enterprise Data Warehouse*), y capas de administración de datos. • Sus innovaciones se enfocarán en mejorar su motor *HAWQ SQL* y la integración con otros productos de Pivotal. • Tecnología *Greenplum Database de EMC*. • *Expertis de EMC y VMware*.
Teradata	*Unix/Linux*	• *Hadoop* como *appliance*. • *SQL-H* (Motor *SQL* para realizar consultas en *Hadoop*). • Herramientas de administración propietarias de *Teradata*. • *Aster*[6]. (*Software* para realizar analítica con *Hadoop*). • *Appliances* de alto desempeño.

[6] *Aster* es un software basado en el procesamiento masivo en paralelo para el análisis y manejo de datos desarrollado por *Teradata Corp*. https://en.wikipedia.org/wiki/Aster_Data_Systems

Intel	*Unix/Linux*	• Tecnología de sus *Chips Xeon* optimizados para *Hadoop*. • Capacidades de desempeño y seguridad mejoradas por *hardware* para *Hadoop*. • *Lustre*[7]. • Optimización nativa de tareas y analítica gráfica.
Microsoft Windows Azure HDInsight Service	*Windows*	• *Polybase*[8]: Permite que los clientes de *SQL* ejecuten consultas que incluyan datos almacenados en Hadoop. • Esfuerzos en el desarrollo de la siguiente generación de *Hive* y de otros proyectos de la comunidad de código abierto de *Hadoop*. • Herramientas de desarrollo y colaboración para los clientes de *Microsoft*.



La arquitectura del *hardware* para *Big Data*

En términos arquitectónicos, las soluciones de hardware para Big Data disponibles en el mercado parecerían más complejas de lo que en realidad son, tienen muchas cosas en común, todas utilizan:

- Nodos computacionales.
- Almacenamiento.
- *Frameworks* y aplicaciones programadas en *Java* (Ejemplo: *Hadoop*, *Hbase*, *Spark*, etc.).
- Redes de alta velocidad de tecnología *gigabit* o superior.
- Sistemas de soporte vital para (alimentación de energía eléctrica y el control de temperatura y humedad).

Podemos apreciar diferencias significativas, como:

- Un número diferente de nodos y de configuraciones entre éstos, que producen rendimientos diferentes en la ejecución de los procesos.
- Tecnologías de almacenamiento diferentes, que influyen de manera determinante en el tiempo de acceso a la información y por ende en la obtención de resultados.

[7] *Lustre* es un sistema de archivos distribuido *Open Source,* normalmente utilizado en clusters a gran escala. El nombre es una mezcla de *Linux* y *clusters*. El proyecto intenta proporcionar un sistema de archivos para *clusters* de decenas de miles de nodos con petabytes de capacidad de almacenamiento, sin comprometer la velocidad o la seguridad, y está disponible bajo la *GNU GPL*. https://es.wikipedia.org/wiki/Lustre_(sistema_de_archivos)

[8] *Polybase* es una tecnología de *Microsoft* que permite a los usuarios de *SQL Server PDW* ejecutar consultas sobre los datos almacenados en *Hadoop*. Véase: http://francescsanchez-bi.webnode.es/news/uso-de-hadoop-en-sql-server-con-pdw-y-polybase-/

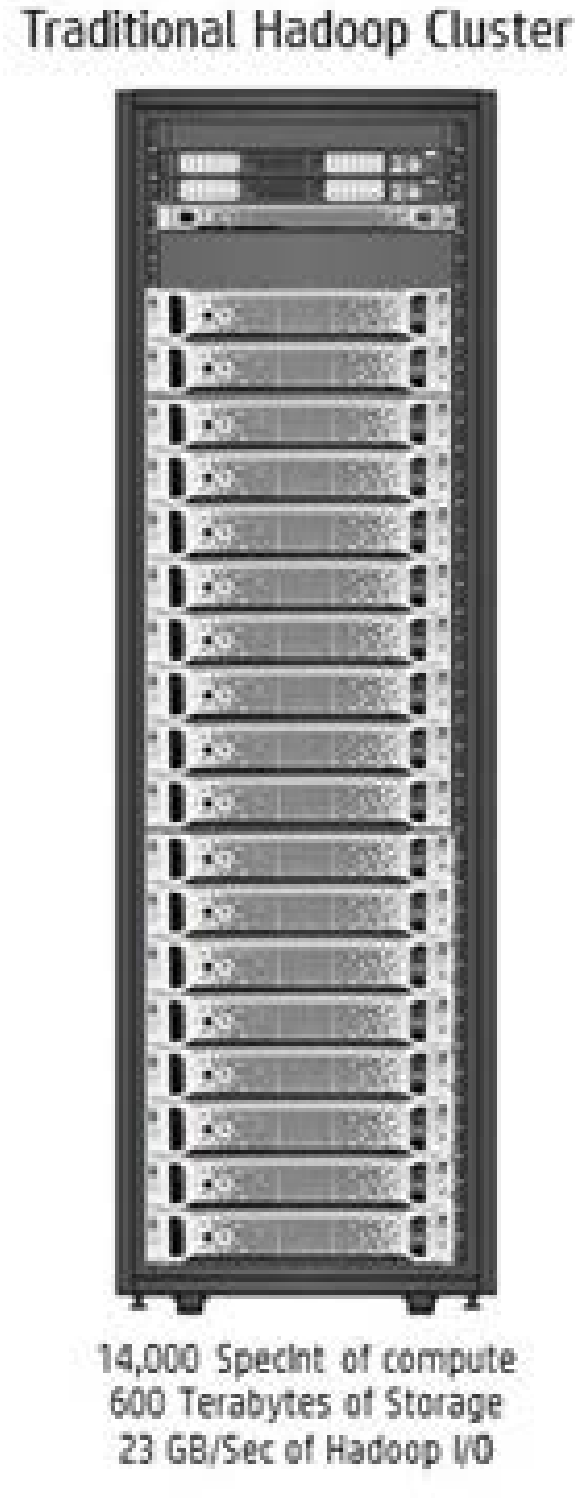


Figura 2, Ejemplo de un clúster para Big Data basado en Hadoop. Fuente: http://www.nextplatform.com/2015/07/03/building-a-better-hadoop-cluster/

• Redes de datos de diferentes velocidades que influyen en el desempeño global de la solución.

El análisis de las cargas de trabajo será un factor determinante para lograr el mejor equilibrio entre rendimiento y ahorro del hardware, por ejemplo:

a)Nodos computacionales

Los nodos computacionales son servidores para *rack* con una configuración preestablecida de acuerdo a su función. Múltiples nodos conforman un *clúster*, eso sí, en ningún momento se recomiendan servidores de tipo *Blade*, debido a que éstos comparten componentes en un chasis que los puede limitar en su desempeño, capacidad de crecimiento y posibilidades de conectividad. La cantidad de núcleos y de memoria RAM disponible en cada uno de ellos, así como su velocidad de interconexión a la red, influirá de manera determinante en el rendimiento global del *clúster*.

Los clúster basados en el Sistema de Archivos de *Hadoop* "*HDFS*", básicamente utilizará dos tipos de nodos diferentes de acuerdo a la función que desempeñarán:

• Nodo de Nombres (Rastreador de trabajos *Namenode JobTracker*): Responsable de la topología de todos los demás nodos y de gestionar el espacio de nombres. Solo hay uno por *clúster*. El Nodo de nombres se encarga de distribuir los archivos en los nodos de datos que rea-

lizarán el procesamiento de los datos. Véase: http://www.happyminds.es/apache-hadoop-introduccion-a-hdfs/#sthash.1A9onPIj.dpuf

- Nodo de Datos (Rastreador de tareas *Datanode TaskTracker*): Acceden a los datos. Almacenan los bloques de información y los recuperan bajo demanda.

Velocidad de procesamiento

En términos generales los cálculos requeridos para los procesos de *Big Data*, comúnmente estarán más limitados por la velocidad de respuesta de las operaciones de entrada y salida a disco y de la red que por la velocidad del procesador. Eso no significa que los procesadores que deban emplearse en la solución requieran ser de bajo costo o pobre rendimiento. Sino más bien que éstos cuenten con la mayor cantidad de núcleos operando a la mayor frecuencia posible para poder manejar más tareas en paralelo de forma ágil.

En este sentido también se tiene que cuidar el consumo de energía y la disipación de calor asociado a la frecuencia y altos voltajes de operación de los microprocesadores que sin duda tendrán impactos negativos en el medio ambiente y el costo de operación de la plataforma completa.

Memoria RAM

La memoria RAM necesaria para un *clúster* de *Big Data* normalmente deberá ser calculada tomando en consideración varios aspectos. Entre los más importantes destacan:

- La cantidad de nodos de datos que contenga el *clúster.*
- El número máximo de tareas que ejecutará cada nodo de datos de forma simultánea.
- La memoria necesaria para los demonios que se ejecutan en los nodos de datos, de tareas y del sistema operativo.
- El número de bloques almacenados en el sistema de archivos distribuidos.
- Las necesidades específicas del software a utilizar en la plataforma.

Los detalles sobre su tecnología, por supuesto que también influirán en el desempeño y configuración de la solución.

b) Almacenamiento

En cuanto a las características técnicas del almacenamiento necesario para una plataforma de *Big Data*, recordemos que éstas deberán ajustarse a los requerimientos del *Framework* y las aplicaciones que serán utilizadas. Desde esta óptica nos encontraremos

que los sistemas de archivos paralelos como el *HDFS* (*Hadoop Distributed File System*) están diseñados para expandirse en múltiples nodos y manejar grandes volúmenes de datos y que por ende generan réplicas para garantizar la tolerancia a fallos. Además tendrán acceso a todos los discos de forma independiente, razón por la cual no se suelen utilizar RAID[9] o LVM[10] para agrupar los discos (ni por tolerancia a fallos ni por rendimiento). Por otra parte las soluciones de almacenamiento deben ofrecer una tecnología robusta y confiable que soporte tanto datos estructurados como no estructurados con seguridad, coherencia y credibilidad. Asimismo deben:

- Disponer de una gran capacidad para almacenar datos. (De acuerdo a los requerimientos del proyecto).
- Deben poder escalar de forma modular fácilmente desde unos cuantos *terabytes* hasta múltiples petabytes de información.
- Contar con un muy alto rendimiento en IOPS (Operaciones de entrada y salida por segundo). Adecuado en tiempo real.
- Soportar y controlar todo tipo de datos.
- Contar con una tecnología que asegure una gran protección para los datos.

Recordemos también que el tema del respaldo de la información se volverá crítico cuando hablamos de terabytes o petabytes de información. Esto si no contamos con una infraestructura suficientemente robusta que asegure la preservación de la información al paso del tiempo[11].

c) Redes de datos

En términos generales el *software* para *Big Data* hace un uso muy intensivo de la red. Esto debido a que todos los nodos intercambian grandes volúmenes de datos, tanto en un esquema normal de operación, como en el caso de fallo de algún nodo.

La razón de este aumento de tráfico en la red causado por la falla en algún nodo del clúster, se debe principalmente a que Hadoop obliga al resto de los nodos a que se pongan a replicar los datos para mantener la integridad de la información. Por estos motivos es recomendable utilizar switches preferentemente de alta velocidad y dedicados para el uso exclusivo del clúster. Dada la naturaleza del *Big Data* se requiere que la red a utilizar cuente con las siguientes características funcionales:

- Gran velocidad.
- Alta disponibilidad.
- Redundancia.

[9] En informática, el acrónimo *RAID* (del inglés *Redundant Array of Inexpensive Disks* o, más común a día de hoy, *Redundant Array of Independent Disks*), traducido como «conjunto redundante de discos Independientes», hace referencia a un sistema de almacenamiento de datos en tiempo real que utiliza múltiples unidades de almacenamiento de datos (discos duros o *SSD*) entre los que se distribuyen o replican los datos. Dependiendo de su configuración (a la que suele llamarse «nivel»), los beneficios de un *RAID* respecto a un único disco son uno o varios de los siguientes: mayor integridad, mayor tolerancia a fallos, mayor *throughput* (rendimiento) y mayor capacidad. Fuente: https://es.wikipedia.org/wiki/RAID

[10] *Logical Volume Management* (*LVM*) en español "administrador de volúmenes lógicos" es una opción de gestión de discos que todas las principales distribuciones de GNU/Linux presentan. Ya sea que necesite para configurar grupos de almacenamiento, o simplemente se necesite para crear dinámicamente las particiones LVM. Véase: https://parbaedlo.wordpress.com/2012/01/05/que-es-lvm-como-se-administra-y-utiliza/

[11] Véase: http://www.computerweekly.com/feature/Big-data-storage-choices

- Resistencia a fallas.
- Alta capacidad de resolver situaciones de congestionamiento en la red.
- Consistencia.
- Fácil de escalabilidad.
- Altamente administrable.
- Resilencia. (Capacidad de continuar funcionando dentro de parámetros aceptables ante distintos tipos de problemas).

Resulta claro que a mayor velocidad, el costo de la red se incrementará. Por tal motivo se puede iniciar con velocidades de conectividad a 1 Gbit y jugar con opciones de agregación de puertos para mejorar el desempeño[12]. No obstante el mejor desempeño se logrará con la tecnología de 10 Gbit o superior.

Por otra parte a mayor número de nodos en el *clúster,* la velocidad de la red será un factor clave para su buen desempeño[13].

Como podremos darnos cuenta, recomendar una configuración de hardware ideal para un proyecto de Big Data no es una tarea sencilla. Algunos pasos para comenzar a definirla son:

- Identificar las preguntas que deberá responder la plataforma de *Big Data*.
- La urgencia de respuesta.
- Identificar las fuentes de información de donde se deberá extraer la información a procesar. Incluso pueden ser a partir de programas de TV en tiempo real o el flujo de mensajes proveniente de redes sociales.
- Cuantificar la cantidad de información con que se deberá trabajar.

Con esta información los especialistas en soluciones de *Big Data* podrán ayudarle con facilidad. No olvide que la infraestructura requerirá no solo de una inversión importante sino también de una serie de nuevas habilidades y conocimientos en el personal para desarrollar las aplicaciones. No obstante brindará oportunidades reales de valor de negocio en nuestras empresas o instituciones.

[12] *Consiste en el uso de varios puertos de red o enlaces para incrementar la velocidad de comunicación. Véase: https://es.wikipedia.org/wiki/Agregaci%C3%B3n_de_enlaces*

[8] *Véase: http://searchdatacenter.techtarget.com/es/cronica/Seis-consideraciones-para-redes-de-big-data.*

Bibliografía

[1] ALFOCEA, J. *Detenido por amenazas a su alcalde en Redes Sociales, Delitos informáticos*. 2015. Revista legal. [en línea]. Fecha de actualización 22 septiembre 2015. [Consulta: 6 de octubre de 2016]. Disponible en: http://www.delitosinformaticos.com/09/2015/delitos/amenazas-2-delitos/detenido-amenazar-alcalde-redes-sociales

[2]CACHEIRO, Javier. *Requisitos Hardware Big Data*. Fundación Pública Galega Centro Tecnolóxico de Supercomputación de Galicia, CESGA. [en Línea]. Fecha de actualización: 16 de Marzo de 2015. [Consultado: 25 de Mayo de 2016]. Disponible en: https://www.cesga.es/es/biblioteca/downloadAsset/id/775

[3] *Enauro engineering services, BIG DATA*. [En Línea]. [Consultado: 14 de Noviembre de 2016]. Disponible en: http://www.enauro.com/2016/06/10/big-data/

[4] HURTADO, Gork. Big Data y Hadoop. Cloudera vs Hortonworks. Mondragon Unibertsitatea, Investigación en TICs, [en Línea]: [Consultado: 3 Mayo de 2016]. Disponible en: http://mukom.mondragon.edu/ict/big-data-y-hadoop-cloudera-vs-hortonworks/

[5] JCASANELLA. *Introducción a Hadoop y su ecosistema, Ticout Outsourcing Center, Tutoriales de Yellowfin y BI*. [En Línea]: Fecha de actualización: 01 de Abril de 2013 [Consultado: 19 Mayo de 2016]. Disponible en: http://www.ticout.com/blog/2013/04/02/introduccion-a-hadoop-y-su-ecosistema/

[6] LOSHIN, David. *Explorando distribuciones Hadoop para gestionar big data*. TechTarget S.A. de C.V., Guía Escencial. [En Línea]: Enero de 2016. [Consultado: 14 Abril de 2016]. Disponible es: http://searchdatacenter.techtarget.com/es/cronica/Explorando-distribuciones-Hadoop-para-gestionar-big-data

[7] LURIE, Marty. *Big data de código abierto para el impaciente*, Parte 1: Tutorial Hadoop: Hello World con Java, Pig, Hive, Flume, Fuse, Oozie, y Sqoop con Informix, DB2, y MySQL. IBM, IBM deveoperWorks. [en línea] 20 de Marzo de 2013 [Consultado: 14 Abril de 2016]. Disponible en: https://www.ibm.com/developerworks/ssa/data/library/techarticle/dm-1209hadoopbigdata/

[8] MATTHEW, Mayo . *Top Big Data Processing Frameworks*. KdnuggetsTM, KDnuggets News. [en Línea]: Marzo de 2016. [Consultado: 14 Abril de 2016]. Disponible en: http://www.kdnuggets.com/2016/03/top-big-data-processing-frameworks.html

[9] O´DELL, Kevin. *How-to: Select the Right Hardware for Your New Hadoop Cluster*. Cloudera Inc., Cloudera Engineering Blog. [en Línea]: 28 de Agosto de 2013. [Consultado: 28 Mayo de 2016]. Disponible en: https://blog.cloudera.com/blog/2013/08/how-to-select-the-right-hardware-for-your-new-hadoop-cluster/

[10] STEVE, Dertien. Defining the Infrastructure for Big Data Analytics. PTC Inc., El Divan Digital. [en Línea]: 22 de Mayo de 2015 . [Consultado: 11Abril de 2016]. Disponible en: http://blogs.ptc.com/2015/05/22/defining-the-infrastructure-for-big-data-analytics/

[11] The apache software foundation Hadoop. *Web site the Apache Software Foundation, Apache Hadoop Project*. [en Línea]: 13 de Febrero de 2016. [Consultado: 17 Mayo de 2016]. Disponible en: http://hadoop.apache.org/

[12] TIRADOS, Marible. *¿Es Hadoop el fin del almacenamiento de datos tradicional?* Big Data Hispano. [en Línea]. 10 de febrero de 2014. [Consultado: 07 Mayo de 2016]. Disponible en: http://www.bigdatahispano.org/noticias/es-hadoop-el-fin-del-almacenamiento-de-datos/

[13] SÁNCHEZ, José Manuel . *¿QUÉ ES BIG DATA?* T2O AdMedia Services, S.L. 2015 [en Línea]. 02 de julio de 2016 [Consultado: 14 Noviembre de 2016]. Disponible en: http://www.t2omedia.com/ideas/actualidad/que-es-big-data-2/ *en maricultura*. México: Secretaría de Medio Ambiente, Recursos Naturales y Pesca, 1997. 192-207pp.