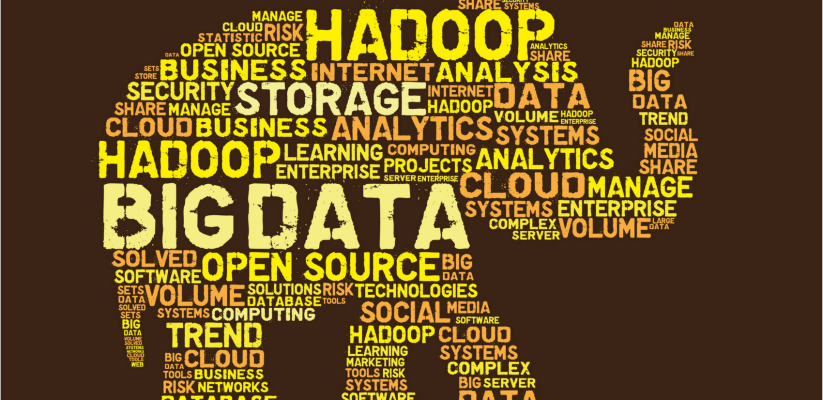
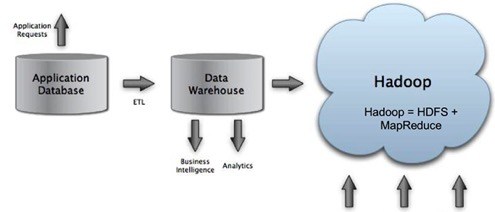
[](https://blog.auriboxtraining.com/big-data/hadoop-y-su-ecosistema/)

**[Hadoop y su ecosistema](https://blog.auriboxtraining.com/big-data/hadoop-y-su-ecosistema/)**

Existen varias soluciones para la gestión de  grandes cantidades de datos almacenados, pero sin duda alguna Hadoop se superpone a varias de estas, con su ecosistema de código abierto creado por la comunidad de Apache Software Foundation que incluye soluciones de almacenamiento, interacción, análisis y visualización de datos y con su sistema de archivos distribuido (HDFS) y su motor de procesamiento de tareas MapReduce, el corazón de Hadoop, proporciona a múltiples empresas la diversidad de sus aplicaciones para cada una de las necesidades que surgen para la gestión de datos masivos.



Como sabemos, **Hadoop** es un framework basado en Java que sigue dos sencillos conceptos: el almacenamiento de datos en Hadoop Distributed File System (HDF) y su procesamiento a través de MapReduce, también llamado modelo de programación para el procesamiento distribuido de datos. Sin duda alguna Hadoop es una de las mejores soluciones para trabajar con Big Data, pero veamos por que ha crecido exponencialmente tanto tecnológica como económicamente por que según Gartner, el mercado del ecosistema Hadoop se espera que llegue a los **50 mil millones** en el año 2020.

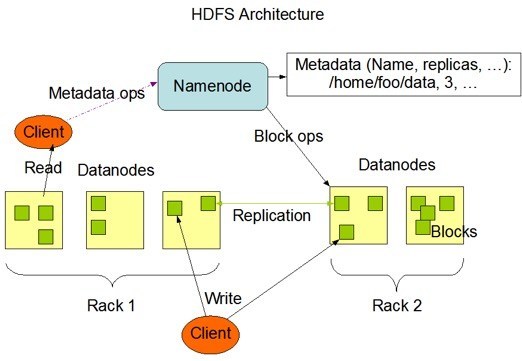
Como lo he venido mencionando HDFS y MapReduce, conforman la arquitectura principal de **Hadoop**.

Arquitectura HDFS

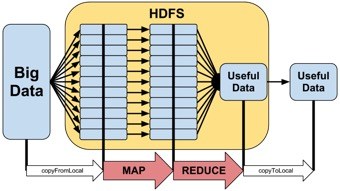
El Sistema de Archivos Distribuidos Hadoop (HDFS) es un sistema diseñado para ejecutarse en hardware común y corriente, económico o conocido como commodity hardware, tiene muchas similitudes con los sistemas de archivos distribuidos existentes, sin embargo, a diferencia de otros, HDFS es altamente tolerante a fallos y está diseñado para ser desplegado en hardware de bajo costo. HDFS proporciona un alto rendimiento de acceso a los datos de la aplicación y es adecuado para aplicaciones que tienen grandes conjuntos de datos. HDFS no es tan exigente con algunos requisitos de POSIX para permitir el acceso de streaming a los datos del sistema de archivos.

HDFS tiene una arquitectura maestro/esclavo. Un clúster HDFS consta de un único NameNode, un servidor maestro que gestiona el espacio de nombres del sistema de archivos y regula el acceso a los archivos por parte de los clientes. Además, hay un número de DataNodes, por lo general uno por nodo en el clúster, que gestionan el almacenamiento adjunto a los nodos en los que se ejecutan. HDFS expone un espacio de nombres de sistema de archivos y permite almacenar datos de usuario en archivos.

Internamente, un archivo se divide en uno o más bloques y estos bloques se almacenan en un conjunto de DataNodes. NameNode ejecuta las operaciones del espacio de nombres del sistema de archivos como abrir, cerrar y cambiar el nombre de archivos y directorios. También determina la asignación de bloques a DataNodes. Los DataNodes son responsables de servir las peticiones de lectura y escritura de los clientes del sistema de archivos. Los **DataNodes** también realizan la creación, eliminación y replicación de bloques a partir de la instrucción del NameNode.



El core de Hadoop MapReduce



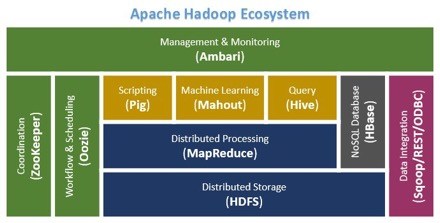
Un trabajo MapReduce normalmente divide el conjunto de datos de entrada en bloques independientes que son procesados ​​por las tareas de mapeo de una manera completamente paralela. El framework ordena las salidas de los mapas, que luego se introducen en las tareas de reducción. Normalmente tanto la entrada como la salida del trabajo se almacenan en un sistema de archivos. El framework se encarga de programar tareas, supervisarlas y volver a ejecutar las tareas fallidas.

En el año 2005, **MapReduce de Google** se vio implementado por Hadoop, aunque no fue hasta dos años después cuando se liberó como Open Source.

[](https://i0.wp.com/blog.auriboxtraining.com/wp-content/uploads/2017/06/mapreduce.jpg)

Ecosistema de Apache Hadoop

HDFS y MapReduce son los dos componentes principales del Ecosistema Hadoop y están en el corazón del framework Hadoop. A continuación listo algunos de los otros proyectos de Apache que se construyen alrededor del framework de Hadoop que son parte de su ecosistema.

[](https://i1.wp.com/blog.auriboxtraining.com/wp-content/uploads/2017/06/apache-hadoop.jpg)

Pig Apache

Es un framework que ofrece un ambiente para la ejecución de trabajos de MapReduce en un Clúster Hadoop a través de un lenguaje de scripting de alto nivel llamado Pig Latin.

Las consultas o comandos de Pig Latin se compilan en uno o más jobs de MapReduce y luego se ejecutan en un clúster de Hadoop. Apache Pig puede operar con casi cualquier tipo de datos.

Apache Hive

Facilita la consulta y administración de grandes conjuntos de datos que residen en un sistema de archivos distribuidos como Hadoop Distributed File System (HDFS), con un lenguaje llamado HiveQL, similar al SQL.

Apache Mahout

Apache Mahout es una biblioteca escalable de **Machine Learning** y **Data Mining**. Implementa los algoritmos machine learning y minería de datos usando MapReduce. Mahout tiene 4 categorías principales de algoritmos: filtrado colaborativo, clasificación, agrupación y reducción dimensional; también contiene dos tipos de algoritmos: los que se pueden ejecutar en modo local y los otros que pueden ejecutarse de forma distribuida.

Apache HBase

Apache HBase es una base de datos distribuida de **Hadoop**, escalable, útil para escrituras y lecturas en tiempo real y acceso aleatorio de los datos. HBase se basa en el concepto **BigTable de Google**. Soporta miles de millones de filas y millones de columnas. Se ejecuta en cluster y se escala de manera lineal. Ofrece APIs en Java fáciles de usar para el acceso de clientes.

Apache Sqoop

Apache Sqoop es una herramienta diseñada para transferir eficientemente los datos entre Hadoop y Bases de Datos Relacionales (RDBMS). Sqoop utiliza MapReduce para importar y exportar datos de forma eficaz utilizando las características de paralelismo y tolerancia a fallos de Hadoop.

Apache Oozie

Apache Oozie es un gestor de programación y coordinación de workflow para gestionar los trabajos ejecutados en Hadoop; puede incluir trabajos MapReduce y Non-MapReduce. Oozie está integrado con Hadoop y es una parte integral del Ecosistema de Hadoop.

Apache ZooKeeper

Apache ZooKeeper es un servicio de coordinación de código abierto para aplicaciones distribuidas, está diseñado para ser un servicio centralizado; es responsable de mantener la información de configuración, ofrecer coordinación de manera distribuida; simplifica el desarrollo de aplicaciones distribuidas. ZooKeeper está siendo utilizado por algunos de los proyectos de Apache como HBase para ofrecer alta disponibilidad y alto grado de coordinación en un entorno distribuido.

Apache Ambari

Apache Ambari es un framework de código abierto para el aprovisionamiento, la gestión y la supervisión de **clústeres Hadoop**. Ambari es útil para instalar servicios Hadoop a través de diferentes nodos del clúster y gestionar la configuración de **Hadoop Services** en el clúster. Ambari ofrece un panel de control para supervisar la salud general del clúster. Ambari ofrece alertas y mecanismo de correo electrónico para obtener la atención necesaria cuando sea necesario. Ambari ofrece las API REST a los desarrolladores para la integración de aplicaciones.

Apache Spark

Es un sistema que procesa los datos a una velocidad superior a MapReduce en tiempo real y también puede implementarse de forma independiente.

Apache Avro

Es un sistema de serialización de datos para procesarlos y almacenarlos con el fin de leerlos fácilmente desde distintos lenguajes de programación.

Apache Cassandra

Es una base de datos para [**Big Data**](http://blog.auriboxtraining.com/big-data/que-es-big-data/) distribuida [NoSQL](http://blog.auriboxtraining.com/bases-de-datos/las-bases-de-datos-nosql/).

Sin duda alguna, el crecimiento exponencial de datos generados de Web y Social Media (Facebook, Twitter, LinkedIn, blogs, medios de comunicación), de máquina a máquina (dispositivos como sensores o medidores que capturan algún evento en particular como velocidad, temperatura, presión, variables meteorológicas, variables químicas como la salud), Biométricas (huellas digitales, escaneo de la retina, reconocimiento facial, genética, etc) o la de autos inteligentes, mensajes, correo electrónico, audios, videos, aplicaciones, de IoT, geolocalizadores, internet, etc., aunado a la necesidad de recolección de datos en tiempo real provenientes de diferentes fuentes y formatos, está llevando a las empresas a que ***Big Data***, es decir, el proceso de gestión de esos grandes volúmenes de información tan heterogéneos y en algunos casos con datos estructurados y no estructurados, sea un requisito indispensable; así como, los ***profesionistas*** especialistas con conocimientos en Inteligencia de negocios, Minería de Datos, Estadísticas, Matemáticas, o los ya conocidos Científicos de Datos; y por supuesto la ***tecnología*** a utilizar.

Al día de hoy la mayoría de las grandes organizaciones empresariales o educativas utilizan la ***tecnología*** Apache Hadoop quien es líder en soluciones para Big Data para sacar el máximo valor a los datos.

Para aprovechar el potencial de los grandes datos y simplificar el trabajo con Hadoop, se han creado versiones empresariales como Cloudera y Hortonworks para tal efecto.

Cloudera

Cloudera Inc. fue fundada por personas que trabajaron en Facebook, Google, Oracle y Yahoo en 2008. Fue la primera empresa en desarrollar y distribuir software basado en Apache Hadoop y todavía tiene la mayor base de usuarios con la mayoría de los clientes. Aunque el núcleo de la distribución se basa en Apache Hadoop, también proporciona una suite propietaria Cloudera Management para automatizar el proceso de instalación y proporcionar otros servicios para mejorar la comodidad de los usuarios, que incluyen la reducción del tiempo de implementación, la visualización de nodos en tiempo real, etc.

Hortonworks

Hortonworks, fundada en 2011, ha surgido rápidamente como uno de los principales proveedores de Hadoop. La distribución proporciona una plataforma de código abierto basada en Apache Hadoop para analizar, almacenar y administrar grandes datos. Hortonworks es el único proveedor comercial que distribuye Apache Hadoop de código abierto completo sin software propietario adicional. Este se puede descargar directamente de su sitio web libre de costo y es fácil de instalar. Los ingenieros de Hortonworks están detrás de la mayoría de las innovaciones recientes de **Hadoop**, incluyendo Yarn, que es mejor que MapReduce en el sentido de que permitirá la inclusión de más frameworks de procesamiento de datos.

Empresas que utilizan Hadoop

* *Amazon*
* [*Adobe*](http://www.adobe.com/)
* [*Alibaba*](http://china.alibaba.com/)
* [*AOL*](http://aol.com/)
* [*EBay*](http://www.ebay.com/)
* [*Facebook*](http://www.facebook.com/)
* [*Google*](http://www.google.com/)
* [*Hadoop Korean User Group*](http://www.hadoop.co.kr/)*, a Korean Local Community Team Page.*
* [*IBM*](http://www.ibm.com/)
* [*Legolas Media*](http://www.legolas-media.com/)
* [*LinkedIn*](http://www.linkedin.com/)
* [*com*](http://www.mercadolibre.com/)
* [*The New York Times*](http://nytimes.com/)
* [*com*](http://www.socialmedia.com/)
* [*Spotify*](http://www.spotify.com/)
* [*Twitter*](http://www.twitter.com/)
* [*University of Maryland*](http://www.umiacs.umd.edu/~jimmylin/cloud-computing/index.html)
* [*University of Nebraska Lincoln, Holland Computing Center*](http://hcc.unl.edu/)
* [*Yahoo!*](http://www.yahoo.com/)
* *Y un largo etc., etc.*