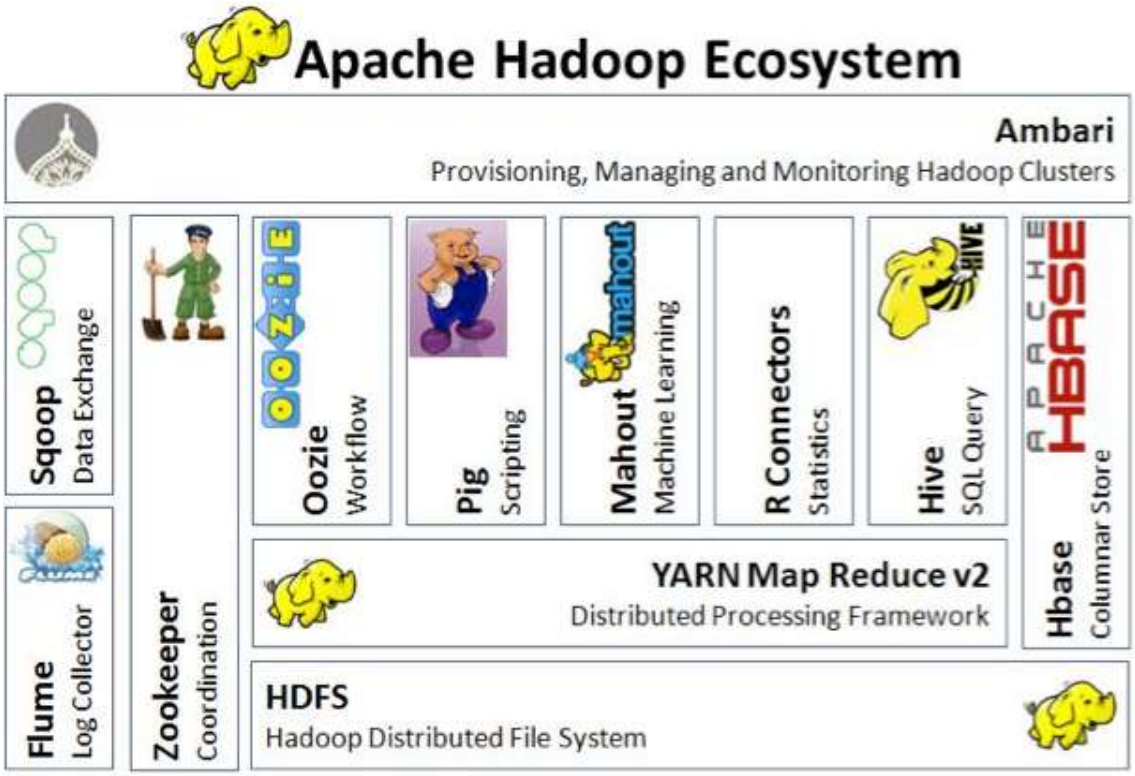
Herramientas de Hadoop

Apache Hadoop es un framework que permite el procesamiento de grandes volúmenes de datos a través de clusters, usando un modelo simple de programación. Además su diseño permite pasar de pocos nodos a miles de nodos de forma ágil. Hadoop es un sistema distribuido usando una arquitectura Master-Slave, usando para almacenar su Hadoop Distributed File System (HDFS) y algoritmos de MapReduce para hacer cálculos.



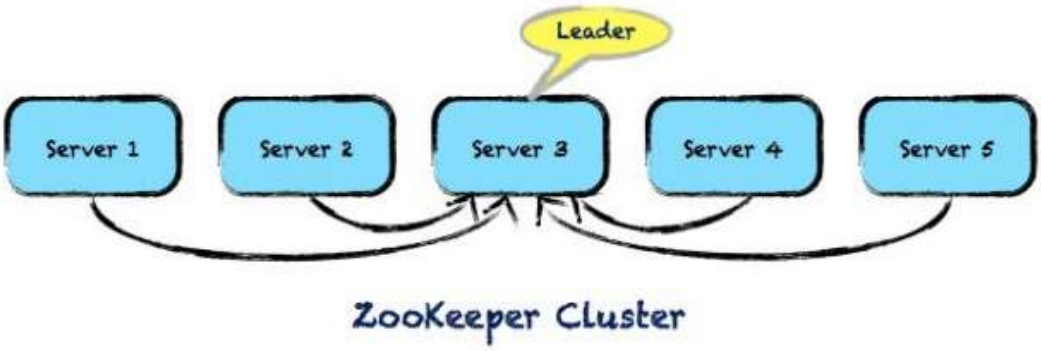
En Hadoop tenemos un ecosistema muy diverso, que crece día tras día, por lo que es difícil saber de todos los proyectos que interactúan con Hadoop de alguna forma. A continuación se detallarán los más comunes.

1 Apache Avro

Es un sistema de serialización de datos. En los proyectos en Hadoop, suele haber grandes cuantidades de datos, la serialización se usa para procesarlos y almacenar estos datos, de forma que el rendimiento en tiempo sea efectivo. Esta serialización puede ser en texto en plano, JSON, en formato binario. Con Avro podemos almacenar y leer los datos fácilmente desde diferentes lenguajes de programación. Está optimizado para minimizar el espacio en disco necesario para nuestros datos.

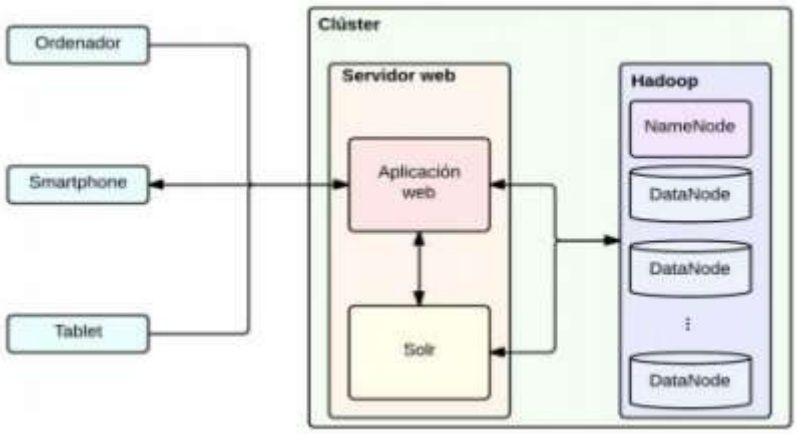
2 ZooKeeper

Apache ZooKeeper es un proyecto de software de la Apache Software Foundation, que provee un servicio de configuración centralizada y registro de nombres de código abierto para grandes sistemas distribuidos. La arquitectura de ZooKeeper soporta alta disponibilidad a través de servicios redundantes. Los clientes pueden así preguntar a otro maestro ZooKeeper si el primero falla al responder. Los nodos ZooKeeper guardan sus datos en un espacio de nombres jerárquico, como hace un sistema de archivos. Los clientes pueden leer y escribir desde/a los nodos y de esta forma tienen un servicio de configuración compartido.



3 SOLR

Apache Solr es un motor de búsqueda basado en el Apache Lucene, escrito en Java y que facilita a los programadores el desarrollo de aplicaciones de búsqueda. Lucene ofrece indexación de información, tecnologías para la búsqueda así como corrección ortográfica, resaltado y análisis de información, entre otras muchas características Una arquitectura típica de Solr Cuenta con un servidor web, para que los usuarios puedan interactuar y realizar distintos tipos de búsquedas con conexión directa con Solr y que consulta datos mediante este en Hadoop.



4 Chukwa

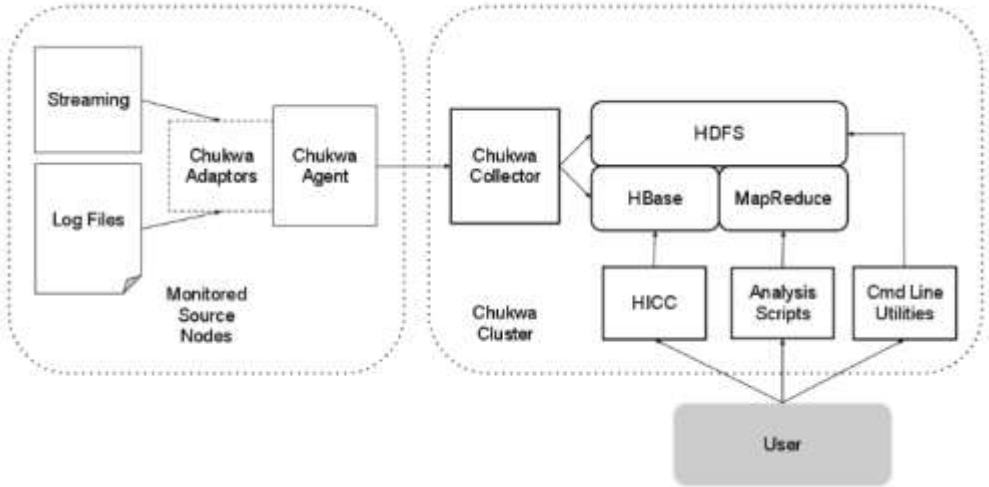
Es un sistema de captura de datos y framework de análisis que trabaja con Hadoop para procesar y analizar grandes volúmenes de logs. Incluye herramientas para mostrar y monitorizar los datos capturados. La arquitectura de Chukwa se compone de cuatro componentes:

- Agentes: los procesos que se encargan de capturar datos.

- Colectores: reciben los datos de los agentes y lo escriben en un almacenamiento permanente.

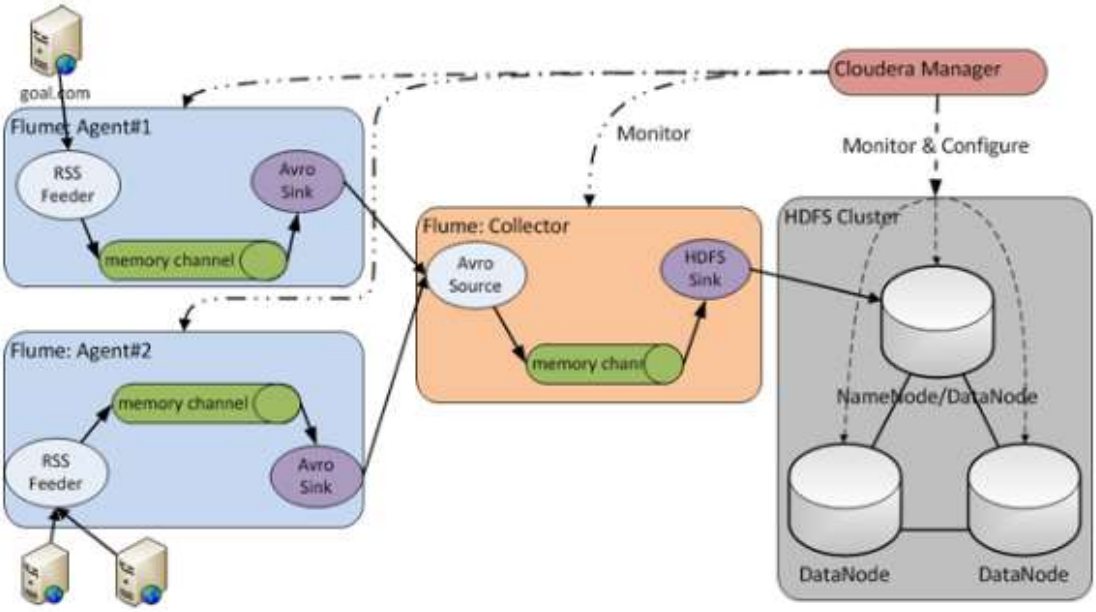
-Trabajos MapReduce para trabajar con los datos.

- HICC: es una interfaz web para visualizar datos.



5 FLUME

Apache Flume es un sistema distribuido para capturar de forma eficiente, agregar y mover grandes cantidades de datos log de diferentes orígenes (diferentes servidores) a un repositorio central, simplificando el proceso de recolectar estos datos para almacenarlos en Hadoop y poder analizarlos. Flume y Chukwa son proyectos parecidos, la principal diferencia es que Chukwa está pensado para ser usado en Batch.



6 Hive

Es una herramienta para data warehousing que facilita la creación, consulta y administración de grandes volúmenes de datos distribuidos en forma de tablas relacionales. Cuenta con un lenguaje derivado de SQL, llamado Hive QL, que permite realizar las consultar sobre los datos. A su vez, Hive QL está construido sobre MapReduce, de manera que se aprovecha de las características de éste para trabajar con grandes cantidades de datos almacenados en Hadoop. Esto también provoca que Hive no ofrezca respuestas en tiempo real.

7 MAHOUT

Mahout es una librería Java que contiene básicamente funciones de aprendizaje y que está construida sobre MapReduce. Al usar MapReduce está pensada para trabajar con grandes volúmenes de datos y en sistemas distribuidos. Aqui vemos un pequeño ejemplo:

8 OOZIE

Oozie es un planificador de workflows para sistemas que realizan trabajos o procesos Hadoop. Proporciona una interfaz

de alto nivel para el usuario no técnico o no experto y que gracias a su abstracción permite a estos usuarios realizar flujos de trabajo complejos.

9 PIG

Es una herramienta para analizar grandes volúmenes de datos mediante un lenguaje de alto nivel -PigLatin- que está diseñado para la paralelización del trabajo. Permite a los usuarios de Hadoop centrarse más en el análisis de los datos y menos en la creación de programas MapReduce. (Veremos un pequeño ejemplo más adelante).

10 HUE

Hue es una herramienta que proporciona a los usuarios y administradores de las distribuciones Hadoop una interfaz web para poder trabajar y administrar las distintas herramientas instaladas. De esta manera Hue ofrece una serie de editores gráficos, visualización de datos y navegadores para que los usuarios menos técnicos puedan usar Hadoop sin mayores problemas.

11 Sqoop

Es una herramienta diseñada para transferir de forma eficiente bulk data entre Hadoop y sistemas de almacenamiento con datos estructurados, como bases de datos relacionales. Permite importar tablas individuales o bases de datos enteras a HDFS, genera clases Java que permiten interactuar con los datos importados, además, permite importar de las bases de datos SQL a Hive.

12 UIMA

Es un framework para analizar grandes volúmenes de datos no estructurados, como texto, video, datos de audio, etc… y obtener conocimiento que sea relevante para el usuario final. Por ejemplo a partir de un fichero plano, poder descubrir que entidades son personas, lugares, organizaciones, etc…

13 HDFS

HDFS es el sistema de almacenamiento, es un sistema de ficheros distribuido. Fue creado a partir del Google File System (GFS). HDFS se encuentra optimizado para grandes flujos y trabajar con ficheros grandes en sus lecturas y escrituras. Su diseño reduce la E/S en la red. La escalabilidad y disponibilidad son otras de sus claves, gracias a la replicación de los datos y tolerancia a los fallos. Los elementos importantes del cluster:

- NameNode: Sólo hay uno en el cluster. Regula el acceso a los ficheros por parte de los clientes. Mantiene en memoria la metadata del sistema de ficheros y control de los bloques de fichero que tiene cada DataNode.
- DataNode: Son los responsables de leer y escribir las peticiones de los clientes. Los ficheros están formados por bloques, estos se encuentran replicados en diferentes nodos.

14 HBase

Se trata de la base de datos de Hadoop. Es el componente que debemos usar cuando se requieren escrituras/lecturas en tiempo real y acceso aleatorio para grandes conjuntos de datos.

Distribuciones Hadoop

1 Amazon EMR

Es un servicio web que facilita el procesamiento rápido y rentable de grandes cantidades de datos, fue uno de los primeros productos comerciales Hadoop en el mercado, y lidera en presencia de mercado global. Amazon EMR simplifica el procesamiento de big data, al proporcionar un marco de trabajo de Hadoop gestionado que facilita la distribución y el procesamiento de grandes cantidades de datos entre instancias de Amazon EC2 dinámicamente escalables de manera sencilla y rápida.

Ventajas

- AWS es considerado el proveedor de almacén de los principales datos en la nube como un servicio. Logra una adopción fuerte, impulsada por su amplia aceptación de la nube, flexibilidad y agilidad, tanto desde el técnico y el punto de vista financiero.
- Facilidad de uso. Puede lanzar un clúster de Amazon EMR en cuestión de minutos. No hay que preocuparse por el aprovisionamiento de nodos, la configuración del clúster, la configuración de Hadoop ni el ajuste del clúster, ya que Amazon EMR se encarga que todas estas tareas.
- AWS es compatible con una amplia variedad de casos de uso cuando se combina con otras soluciones de gestión de datos.
- Buena experiencia del cliente y rápida penetración en el mercado.

Inconvenientes

- Una competencia cada vez mayor en las capacidades funcionales y opciones en la nube para elegir.
- Como AWS es un proveedor en la nube, carece de soporte para las combinaciones de almacenamiento de datos.
- AWS maduran en el uso de Redshift, y se están empezando a reportar limitaciones en relación a sus expectativas para la gestión mixta de carga de trabajo.

2 Cloudera

Se dedica únicamente a ofrecer soluciones Hadoop para Big Data y es una de las compañías líderes y punteras en este campo. Aparte de las soluciones Big Data, Cloudera también se dedica a ofrecer soporte para sus productos y cuentan con un sistema de entrenamiento y certificaciones profesionales llamado Cloudera University.

Cloudera ofrece su motor de consultas SQL, Impala, que fue creado para realizar análisis de datos almacenados en Hadoop a través de herramientas SQL. El resultado es que el procesamiento de datos a gran estala y las consultas interactivas se pueden hacer en el mismo sistema utilizando los mismos datos y metadatos, eliminando la necesidad de migrar los conjuntos de datos a sistemas especializados o formatos propietarios solo para realizar el análisis.

Además Cloudera manager es la plataforma de administración de Cloudera para la gestión de clústeres Hadoop. Este tipo de frameworks facilita la gestión manual que supone la administración de un clúster.

Ventajas

- Cloudera se diferencia de otros proveedores de distribución de Hadoop en que sigue invirtiendo en capacidades específicas, tales como mejoras adicionales como Cloudera Navigator (que proporciona la gestión de metadatos, linaje y la auditoría), mientras que al mismo tiempo trata de mantenerse al día con el proyecto de código abierto Hadoop.
- Cloudera ha colocado con éxito su solución como complemento al almacén de datos tradicional y ha hecho uso de sus relaciones con los proveedores de DBMS tradicionales, sobre todo de Oracle.
- Cloudera ha continuado su expansión geográfica.
- La modularidad Hadoop permite añadir nuevos componentes fácilmente, y Cloudera continúa expandiendo su conjunto de componentes para cumplir con los nuevos casos y las necesidades de uso. Esto permite a Cloudera ofrecer nuevas capacidades sin interrumpir los sistemas existentes.

Inconvenientes

- A pesar de que las organizaciones tienen un interés cada vez mayor en implementaciones en la nube, Cloudera se refiere a la nube utilizando un enfoque de infraestructura como un servicio que no ofrece soporte de servicio escalable, elástico y administrado. Sin embargo, Cloudera se dirige a estas necesidades con mejoras para facilitar la implementación de clusters elásticos en la nube.

- Se considera que la disponibilidad de recursos de apoyo de servicios se está reduciendo. Cloudera ha reconocido que esto es un problema, y ha trabajado para abordar estos puntos en 2015.

3 HORTONWORKS

Hortonworks es una compañía de software empresarial. La compañía se centra en el desarrollo y apoyo de Hadoop, un marco que permite el procesamiento distribuido de grandes conjuntos de datos a través de cluster de ordenadores.

Ventajas

- Hortonworks ha ganado fuerza de mercado con un mayor número de socios reconocidos, incluyendo proveedores tradicionales de DBMS.
- Hortonworks con la plataforma Open Data es compatible con el crecimiento de nuevos proyectos de la Fundación Apache. Hortonworks se diferencia de otros proveedores de distribución mediante la adopción de código abierto.

Inconvenientes

- Será un reto para Hortonworks mantener la integración con un ecosistema de análisis más amplio, como Teradata.
- Incluso con la deriva de la demanda hacia soluciones de código abierto, no seleccionan sólo proveedores o soluciones en base a este único aspecto.

4 IBM InfoSphere BigInsights

IBM BigInsights extiende los componentes básicos de Hadoop para mejorar la usabilidad. Se añaden características a escala empresarial de IBM para ofrecer un procesamiento y análisis de datos masiva scale-out con una función de la resistencia y tolerancia a fallos. Capacidades de administración y gestión simplificadas, ricas herramientas para desarrolladores y potentes funciones analíticas reducen la complejidad de Hadoop.

Ventajas

- IBM ha lanzado dashDB y DataWorks como ofertas en la nube. Estos ofrecen la oportunidad de desplegar rápidamente modelos analíticos y de datos en un entorno elástico. Se dirigen a la creciente demanda de soluciones en la nube.
- IBM ha introducido dispositivos de conexión a fuentes relacionales y NoSQL. Se permite el acceso al procesamiento a través de una amplia variedad de entornos.
- IBM con el proyecto de código abierto Apache Spark añade valor a los productos de IBM por streaming que permite, aprendizaje automático y análisis avanzados. Puede ayudar a una maduración más rápida de Spark.

Inconvenientes

- El nivel de adopción de dashDB es incierta.

- La reducción de personal de IBM reposiciona su oferta para el mercado de soluciones en la nube y de gestión de datos modernos.
- IBM queda en el tercio inferior de relación calidad-precio. Los productos en la nube de IBM no tienen un precio competitivo.

6 MapR Technologies

MapR es una compañía de software empresarial, que desarrolla y vende software Hadoop-derived. La compañía contribuye a proyectos Apache Hadoop como HBase, Pig, Hive and Zookeeper. Pretende ofrecer una protección completa de datos, sin puntos únicos de fallo, un mejor desempeño, y facilidad.

Ventajas

- MapR posee alta disponibilidad y gestión de clusters. MapR ha mejorado estas capacidades con la adición de funciones de autorización y auditoría.
- MapR se centra en abordar una amplia gama de casos de uso. Es compatible con streaming, casos de uso operativos y analíticos, todo desde la misma plataforma, con el apoyo y capacidades de SQL.
- MapR se ha estado expandiendo por todo el mundo. Dispone de soluciones conjuntas con los principales actores, como AWS, Google, HPE, IBM, Microsoft, SAP, SAS y Teradata.

Inconvenientes

- Todavía carece de una falta de visibilidad en el mercado.
- En general poco avanzado. Los usuarios crean los análisis que se implementan como productos de datos completos para su uso en producción.
- Problemas con actualizaciones e instalaciones. Para hacer frente a estos retos, MapR ahora ofrece a los instaladores interfaz gráfica de usuario, libros de utilización, un instalador de parches y paquetes de actualización.

7 Pivotal Software

Es una compañía de software que está centrada en soluciones Big Data. Pivotal fue el primer proveedor de EDW en proporcionar un appliance de grado empresarial con todas las características; también fue la primera en lanzar una familia de appliances que integra su Hadoop, EDW y capas de administración de datos en un solo rack.