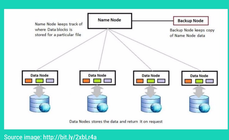
10 Examples of Open Source Projects to manage big data

Hadoop Distributed File System (HDFS)

The Hadoop Distributed File System (HDFS) is a distributed file system designed to store very large data sets reliably, and to stream those data sets at high bandwidth to user applications. Different from other distributed file systems, HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware.

HDFS uses a master/slave architecture where master consists of a single NameNode that manages the file system metadata and one or more slave DataNodes that store the actual data.



Source image: http://bit.ly/2xbLr4a

Apache Hadoop YARN

YARN is the architectural center of Hadoop that allows multiple data processing engines such as interactive SQL, real-time streaming, data science and batch processing to handle data stored in a single platform, unlocking an entirely new approach to analytics.

In Hadoop V2, YARN serves as a resource manager for job scheduling and cluster resource management.

The resource manager (RM) in a Hadoop cluster keeps track of available resources (primarily CPU, memory and data location) on each node of the cluster. The RM communicates with the NameNode that keeps track of the location of data in the cluster. When a job request is submitted the RM determines the best Data Nodes on which to run the job - best being determined by data being resident on the same node that the job is executed on, in addition to the node having enough CPU and memory.

Apache Cassandra

Apache Cassandra is a free and open-source distributed NoSQL database management system designed to handle large amounts of data across many commodity servers, providing high availability with no single point of failure.

Source:
https://en.wikipedia.org/wiki/Apache_Cassandra

Source image: http://bit.ly/2x6loIA

Apache HBase

HBase is an open-source, non-relational, distributed database modeled after Google's Bigtable and is written in Java. It is developed as part of Apache Software Foundation's Apache Hadoop project and runs on top of HDFS (Hadoop Distributed File System), providing Bigtable-like capabilities for Hadoop.

Source:
https://en.wikipedia.org/wiki/Apache_HBase

Apache HBase is a column-oriented key/value data store built to run on top of the Hadoop Distributed File System (HDFS).

Source: http://bit.ly/2jht5X3

HBase is designed to support high table-update rates and to scale out horizontally in distributed compute clusters. Its focus on scale enables it to support very large database tables – for example, ones containing billions of rows and millions of columnsDEFINITION Apache HBase.

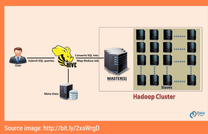
Source: http://bit.ly/2jht5X3

Currently, one of the most prominent uses of HBase is as a structured data handler for Facebook's basic messaging infrastructure.

Apache Hive

Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data summarization, query, and analysis. Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop.

Source: https://en.wikipedia.org/wiki/Apache_Hive



Source image: http://bit.ly/2xaWrgD

Apache Mahout

Apache Mahout is a project of the Apache Software Foundation to produce free implementations of distributed or otherwise scalable machine learning algorithms focused primarily in the areas of collaborative filtering, clustering and classification. Many of the implementations use the Apache Hadoop platform. Mahout also provides Java libraries for common maths operations (focused on linear algebra and statistics) and primitive Java collections.

Source:
https://en.wikipedia.org/wiki/Apache_Mahout

It provides a simple and extensible programming environment and framework for building scalable algorithms.

Apache Pig

Apache pig is a programming tool to create programs that runs on Apache Hadoop. The language for this platform is called Pig Latin. Pig can execute its Hadoop jobs in MapReduce, Apache Tez, or Apache Spark. Pig Latin abstracts the programming from the Java MapReduce idiom into a notation which makes MapReduce programming high level, similar to that of SQLfor relational database management systems.

Source:
https://en.wikipedia.org/wiki/Pig_(programming_tool)

Apache Spark

Apache Spark is an open source big data processing framework built around speed, ease of use, and sophisticated analytics.

It was developed in response to limitations in the MapReduce cluster computing paradigm, which forces a particular linear dataflow structure on distributed programs: MapReduce programs read input data from disk, map a function across the data, reduce the results of the map, and store reduction results on disk.

The availability of RDDs facilitates the implementation of both iterative algorithms, that visit their dataset multiple times in a loop, and interactive/exploratory data analysis, i.e., the repeated database-style querying of data.

Source:
https://en.wikipedia.org/wiki/Apache_Spark

Apache Storm

Apache Storm is a free and open source distributed realtime computation system. Storm makes it easy to reliably process unbounded streams of data, doing for realtime processing what Hadoop did for batch processing. Storm is simple, can be used with any programming language, and is a lot of fun to use!

Storm has many use cases: realtime analytics, online machine learning, continuous computation, distributed RPC, ETL, and more. Storm is fast: a benchmark clocked it at over a million tuples processed per second per node. It is scalable, fault-tolerant, guarantees your data will be processed, and is easy to set up and operate.

Source:
http://storm.apache.org/

Apache Tez

Apache™ Tez is an extensible framework for building high performance batch and interactive data processing applications, coordinated by YARN in Apache Hadoop.

Apache Tez provides a developer API and framework to write native YARN applications that bridge the spectrum of interactive and batch workloads. It allows those data access applications to work with petabytes of data over thousands nodes.

Source:
https://hortonworks.com/apache/tez/