hadoop

Ercan Karaçelik

Big Data

Big data is larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software just can’t manage them. But these massive volumes of data can be used to address business problems you wouldn’t have been able to tackle before.

SHARE
MANAGE
CLOUD RISK
STATISTIC
OPEN SOURCE
HADOOP
BUSINESS INTERNET ANALYSIS
SECURITY STORAGE INTERNET HADOOP DATA
SHARE MANAGE
CLOUD BUSINESS ANALYTICS VOLUME
SYSTEMS
HADOOP LEARNING COMPUTING ANALYTICS
ENTERPRISE PROJECTS
BIG DATA
CLOUD MANAGE
SYSTEMS ENTERPRISE
COMPLEX VOLUME
SERVER
HADOOP
BIG
DATA
TREND
SOCIAL
MEDIA
SHARE
SOLVED OPEN SOURCE BIG DATA
SOFTWARE
VOLUME SOLUTIONS RISK TECHNOLOGIES
DATABASE
SYSTEMS COMPUTING SOCIAL MEDIA
TREND HADOOP CLOUD
BIG CLOUD LEARNING SYSTEMS
TOOLS BUSINESS MARKETING COMPLEX
RISK NETWORKS TOOLS RISK BIG SERVER
DATABASE SYSTEMS DATA
SOLVED SOFTWARE SECURITY
HADOOP COMPLEX OPEN SOURCE

EXECUTION TIME

Data access rate

+

Program computation time (~60 mins)

+

Network Bandwidth.. etc..

Average Data Access Rate - 122 MB/sec

1 TB file will take 2 1/2 hours to read from disk

1 Sec - 122 MB
x Secs - 1048576 MB (1 TB)

x = 1048576 / 122 = 8595 secs
2 hr 22 mins

> 3 hrs 🙁

Hadoop

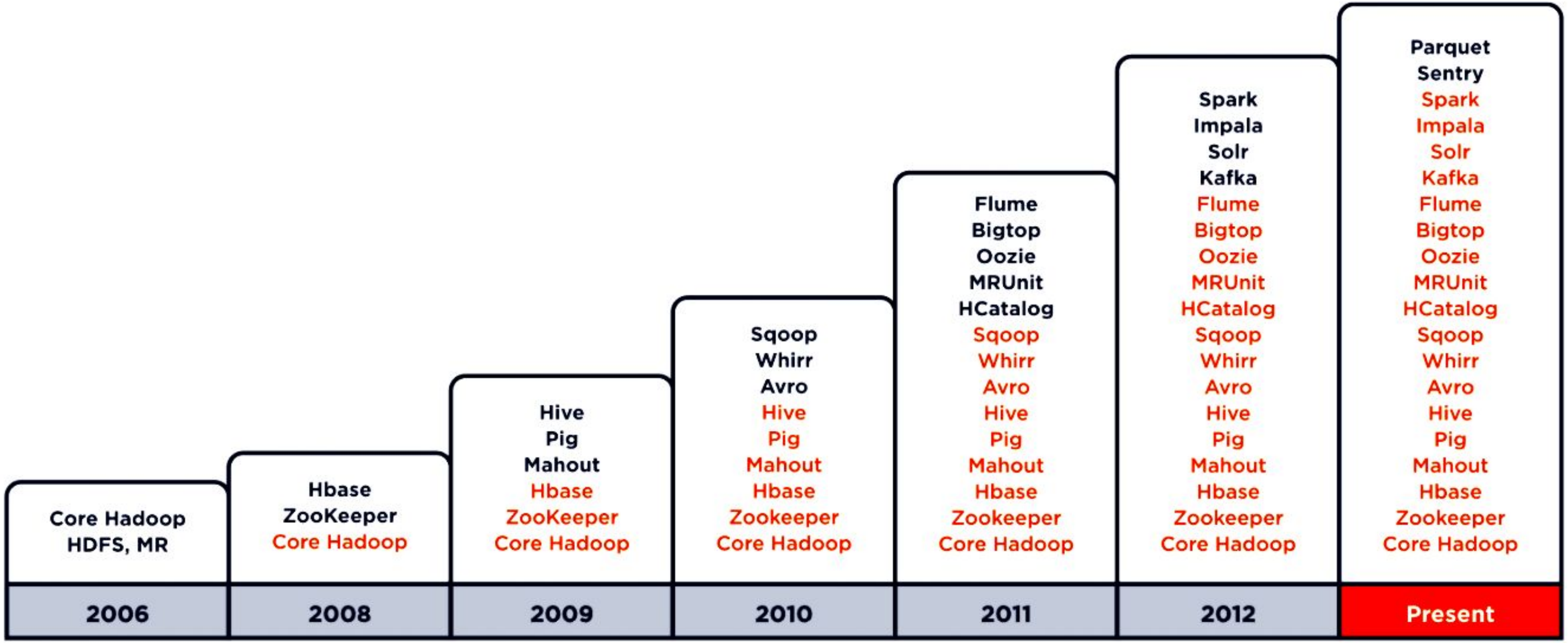
Put simply, Hadoop can be thought of as a set of open source programs and procedures (meaning essentially they are free for anyone to use or modify, with a few exceptions) which anyone can use as the "backbone" of their big data operations.

Bernard Marr

Apache Hadoop is an open source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data. Instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyze massive datasets in parallel more quickly.

Amazon

Hadoop Ecosystem Evolution



2006	2008	2009	2010	2011	2012	Present
Core Hadoop HDFS, MR	Hbase ZooKeeper Core Hadoop	Hive Pig Mahout Hbase ZooKeeper Core Hadoop	Sqoop Whirr Avro Hive Pig Mahout Hbase Zookeeper Core Hadoop	Flume Bigtop Oozie MRUnit HCatalog Sqoop Whirr Avro Hive Pig Mahout Hbase Zookeeper Core Hadoop	Spark Impala Solr Kafka Flume Bigtop Oozie MRUnit HCatalog Sqoop Whirr Avro Hive Pig Mahout Hbase Zookeeper Core Hadoop	Parquet Sentry Spark Impala Solr Kafka Flume Bigtop Oozie MRUnit HCatalog Sqoop Whirr Avro Hive Pig Mahout Hbase Zookeeper Core Hadoop

Functions Of File System

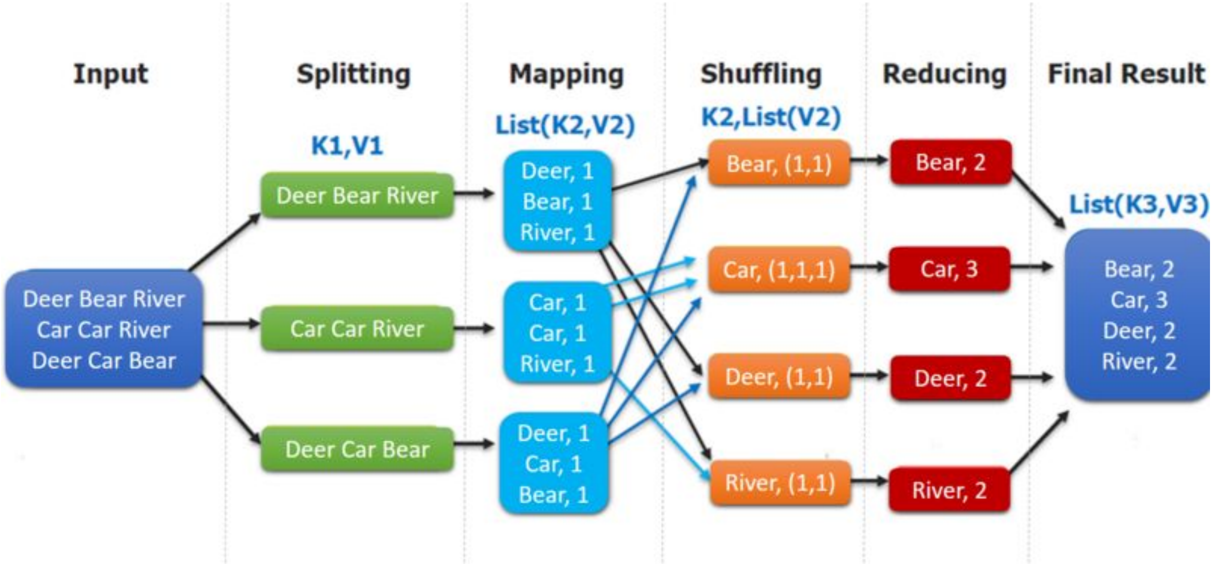
- Control how data is stored and retrieved
- Metadata about the files and folders
- Permissions and security
- Manage storage space efficiently

HDFS

HDFS (Hadoop Distributed File System): A file system that is distributed amongst many networked computers or nodes. HDFS is fault tolerant because it stores multiple replicas of files on the file system, the default replication level is 3. HDFS was designed to survive failures.

MapReduce

MapReduce is a programming model for processing large data sets with a parallel , distributed algorithm on a cluster (source: Wikipedia). Map Reduce when coupled with HDFS can be used to handle big data.



Yarn

Coordinates tasks running on the cluster and assigns new nodes in case of failure. Comprised of 2 subcomponents: the resource manager and the node manager. The resource manager runs on a single master node and schedules tasks across nodes. The node manager runs on all other nodes and manages tasks on the individual node.

Pig

High level scripting language (Pig Latin) that enables writing complex data transformations. It pulls unstructured/incomplete data from sources, cleans it, and places it in a database/data warehouses. Pig performs ETL into data warehouse while Hive queries from data warehouse to perform analysis (GCP: DataFlow).

Hive

Data warehouse software built o top of Hadoop that facilitates reading, writing, and managing large datasets residing in distributed storage using SQL-like queries (HiveQL). Hive abstracts away underlying MapReduce jobs and returns HDFS in the form of tables (not HDFS).

Sqoop

Transferring framework to transfer large amounts of data into HDFS from relational databases (MySQL)

Kafka

Apache Kafka is an open-source stream-processing software platform developed by LinkedIn and donated to the Apache Software Foundation, written in Scala and Java. The project aims to provide a unified, high-throughput, low-latency platform for handling real-time data feeds.

Spark

Framework for writing fast, distributed programs for data processing and analysis. Spark solves similar problems as Hadoop MapReduce but with a fast in-memory approach. It is an unified engine that supports SQL queries, streaming data, machine learning and graph processing. Can operate separately from Hadoop but integrates well with Hadoop. Data is processed using Resilient Distributed Datasets (RDDs), which are immutable, lazily evaluated, and tracks lineage.

Beam

Programming model to define and execute data processing pipelines, including ETL, batch and stream (continuous) processing. After building the pipeline, it is executed by one of Beam’s distributed processing backends (Apache Apex, Apache Flink, Apache Spark, and Google Cloud Dataflow). Modeled as a Directed Acyclic Graph (DAG).

Oozie

Workflow scheduler system to manage Hadoop jobs.

Hue

Hue is a web-based interactive query editor that enables you to interact with data warehouses.

Ambari

The Apache Ambari project is aimed at making Hadoop management simpler by developing software for provisioning, managing, and monitoring Apache Hadoop clusters. Ambari provides an intuitive, easy-to-use Hadoop management web UI backed by its RESTful APIs.

Docker Installation

windows
https://www.youtube.com/watch?v=_9AWYIt86B8

mac
https://www.youtube.com/watch?v=MU8HUVIJTEY

After Installation:

Step 1

```
docker -v
docker run hello-world
```

Step 2

```
docker pull cloudera/quickstart:latest
```

Step 3

```
docker run --hostname=quickstart.cloudera --privileged=true -t -i -p  8888:8888 -p 80:80 cloudera/quickstart /usr/bin/docker-quickstart
```

Step 4

Browse hue localhost:8888