Salvo Nicotra

What is Hadoop ?

Source

Doug Cutting:

> The name my kid gave a stuffed yellow elephant. Short, relatively easy to spell and pronounce, meaningless, and not used elsewhere: those are my naming criteria. Kids are good at generating such. Googol is a kid's term" Hadoop is hardly the first unusual name to be attached to a tech company, of course. Google was born from a misspelling of "googol" (1 followed by 100 zeros), which itself was invented when a mathematician was playing with his nephew and together they came up with a name for really big numbers.

An Operating System for Big Data

▶

Hadoop Web Site

The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.
The Apache Hadoop software library is a framework that allows for the **distributed processing** of **large data sets** across clusters of computers using **simple programming models**. It is designed to **scale up** from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver **high-availability**, the library itself is designed to detect and handle failures **at the application layer**, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.
https://hadoop.apache.org/

Distributed Processing

It’s a generic concept, derives from distribuited systems and involves:

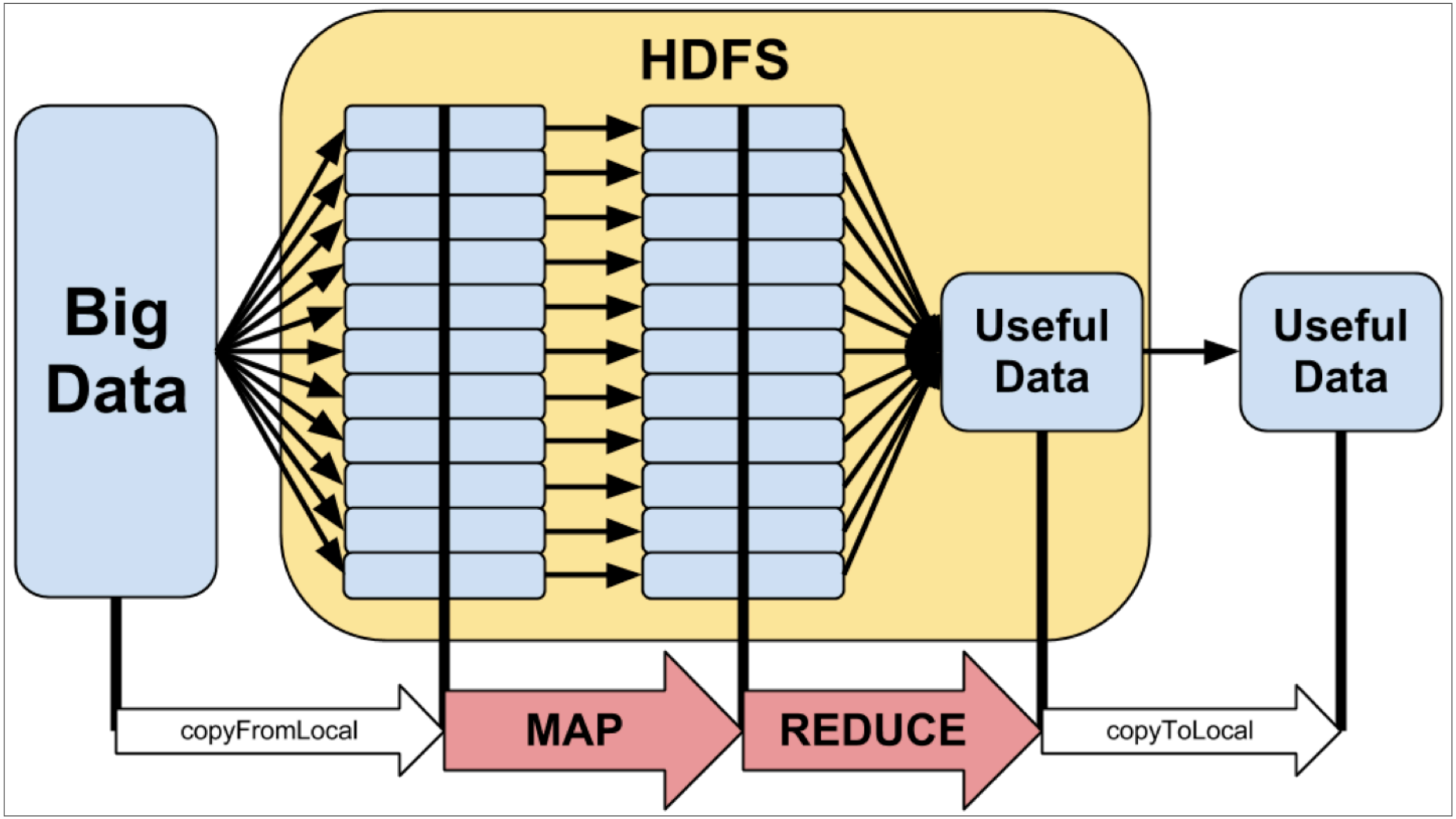
- networking
- message exchange / protocols
- fault tollerance
- data distribution
- optimization

Architectures:

- client-server
- 3-tier / n-tier
- peer to peer

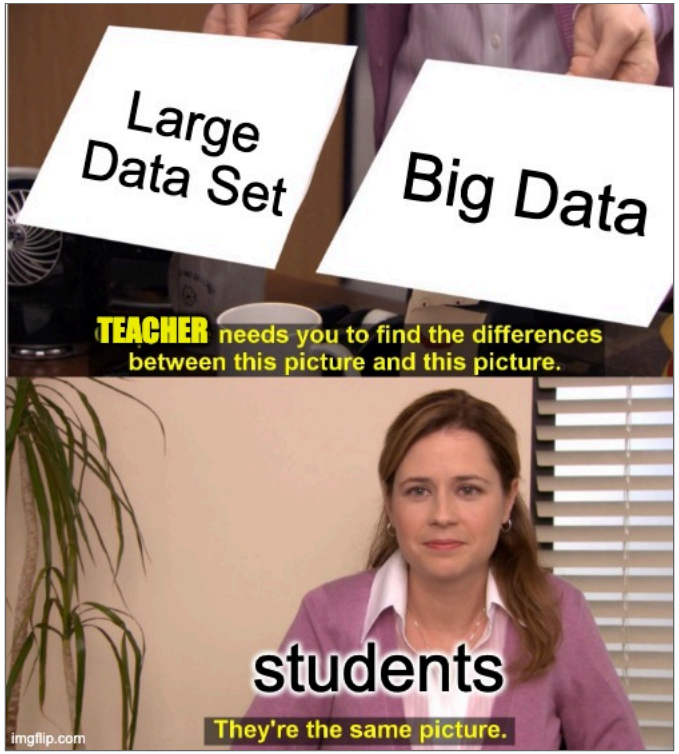
In Hadoop

Distributed processing means Map Reduce



https://www.glennklockwood.com/data-intensive/hadoop/overview.html

Large Data Set



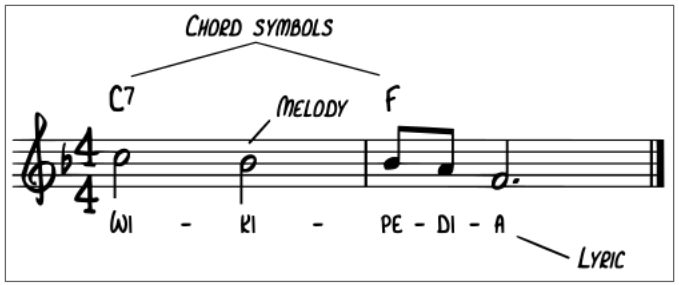
Simple Programming Models

Starting from programming language

A programming language is made by
syntax
Set of grammar rules to write correctly the source code,
including: - symbols (variables, reserved words) - expressions
(constructs)
Wikipedia

execution model
specifies the behaviour of the elements, necessary to understand
what the code does including: - order of execution - interaction
with runtime systems
Wikipedia

It’s like playing a song



Example Python programming language

A Python program is read by a parser. Input to the parser is a stream of tokens, generated by the lexical analyzer.

Doc Python

> A Python program is constructed from code blocks. A block is a piece of Python program text that is executed as a unit. The following are blocks: a module, a function body, and a class definition. Each command typed interactively is a block.
>
> **Doc Python**

To programming model

A programming model is an

- execution model
- coupled to an API or a particular pattern code.

So there can be two execution model

- the one of the base programming language
- the one of the programming model

and they can be different!

Examples

Name	Base programming language	execution model
Spark	Java,Python,Scala	Spark
Hadoop	Java	Map Reduce
Thread	C, C++, Java, Python	POSIX Thread

Scaling

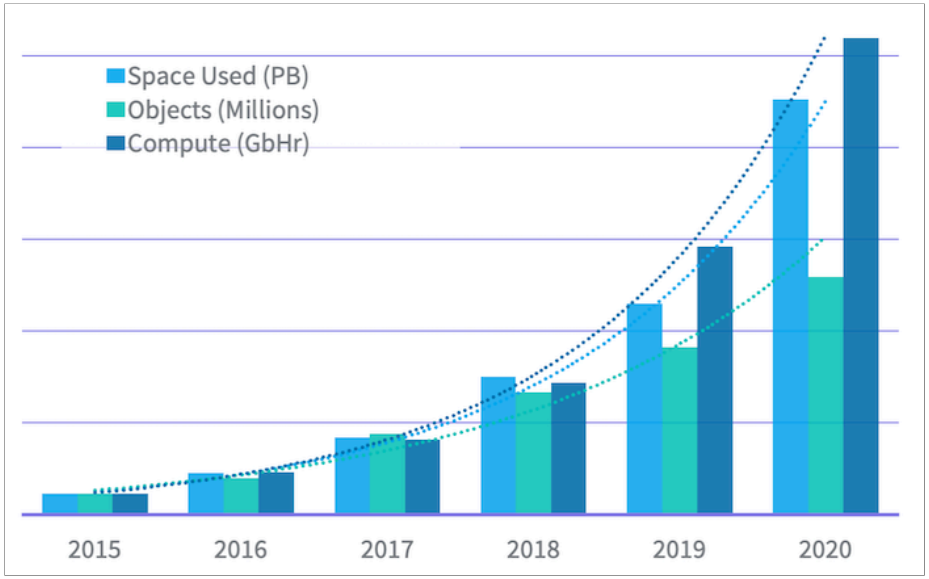
Why Does The High Scalability Site Exist?

This site tries to bring together all the lore, art, science, practice, and experience of building scalable websites into one place so you can learn how to build your website with confidence.
When it becomes clear you must grow your website or die, most people have no idea where to start. It's not a skill you learn in school or pick up from a magazine article on a plane flight home. No, building scalable systems is a body of knowledge slowly built up over time from hard won experience and many failed battles. Hopefully this site will move you further and faster along the learning curve of success.
http://highscalability.com/start-here/

HADOOP RUNS INTO CLUSTERS OF NODES

Installing a Hadoop cluster typically involves unpacking the software on all the machines in the cluster or installing it via a packaging system as appropriate for your operating system. It is important to divide up the hardware into functions.
Typically one machine in the cluster is designated as the NameNode and another machine as the ResourceManager, exclusively. These are the masters. Other services (such as Web App Proxy Server and MapReduce Job History server) are usually run either on dedicated hardware or on shared infrastructure, depending upon the load.
The rest of the machines in the cluster act as both DataNode and NodeManager. These are the workers.

Ref



Ref

Note

- Installing from scratch Hadoop is hard !
- That’s why hadoop vendors first and cloud provider later did start package Hadoop
- Now it’s time for Hadoop in Kubernetes!

High Avalability at application layer

> High availability (HA) is a characteristic of a system which aims to ensure an agreed level of operational performance, usually uptime, for a higher than normal period.

Principles There are three principles of systems design in reliability engineering which can help achieve high availability.

1. Elimination of single points of failure. This means adding or building redundancy into the system so that failure of a component does not mean failure of the entire system.
2. Reliable crossover. In redundant systems, the crossover point itself tends to become a single point of failure. Reliable systems must provide for reliable crossover.
3. Detection of failures as they occur. If the two principles above are observed, then a user may never see a failure – but the maintenance activity must.

In hadoop

At Storage level: Data Replication

HDFS is designed to reliably store very large files across machines in a large cluster. It stores each file as a sequence of blocks; all blocks in a file except the last block are the same size. The blocks of a file are replicated for fault tolerance. The block size and replication factor are configurable per file. An application can specify the number of replicas of a file. The replication factor can be specified at file creation time and can be changed later. Files in HDFS are write-once and have strictly one writer at any time.

The NameNode makes all decisions regarding replication of blocks. It periodically receives a Heartbeat and a Blockreport from each of the DataNodes in the cluster. Receipt of a Heartbeat implies that the DataNode is functioning properly. A Blockreport contains a list of all blocks on a DataNode.

At computing level: Application Master

When the application master is notified of a task attempt that has failed, it will reschedule execution of the task. The application master will try to avoid rescheduling the task on a node manager where it has previously failed. Furthermore, if a task fails four times, it will not be retried again.

Hadoop Components

Modules

Hadoop Common
The common utilities that support the other Hadoop modules.
Hadoop YARN
A framework for job scheduling and cluster resource management.
Hadoop Distributed File System (HDFS™)
A distributed file system that provides high-throughput access to application data.
Hadoop MapReduce
A YARN-based system for parallel processing of large data sets.

Related Project

- Ambari™: A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig and Sqoop. Ambari also provides a dashboard for viewing cluster health such as heatmaps and ability to view MapReduce, Pig and Hive applications visually alongwith features to diagnose their performance characteristics in a user-friendly manner.
- Avro™: A data serialization system.
- Cassandra™: A scalable multi-master database with no single points of failure.
- Chukwa™: A data collection system for managing large distributed systems.
- HBase™: A scalable, distributed database that supports structured data storage for large tables.
- Hive™: A data warehouse infrastructure that provides data summarization and ad hoc querying.
- Mahout™: A Scalable machine learning and data mining library.
- Ozone™: A scalable, redundant, and distributed object store for Hadoop.
- Pig™: A high-level data-flow language and execution framework for parallel computation.
- Spark™: A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.
- Submarine: A unified AI platform which allows engineers and data scientists to run Machine Learning and Deep Learning workload in distributed cluster.
- Tez™: A generalized data-flow programming framework, built on Hadoop YARN, which provides a powerful and flexible engine to execute an arbitrary DAG of tasks to process data for both batch and interactive use-cases. Tez is being adopted by Hive™, Pig™ and other frameworks in the Hadoop ecosystem, and also by other commercial software (e.g. ETL tools), to replace Hadoop™ MapReduce as the underlying execution engine.
- ZooKeeper™: A high-performance coordination service for distributed applications.

Hadoop ecosystem table (retired)

https://hadoopecosystemtable.github.io/

The Good and the Bad of Hadoop Big Data Framework

https://www.altexsoft.com/blog/hadoop-pros-cons/

> **Note**
>
> What does it take to store all New York Times articles published between 1855 and 1922? Depending on how you measure it, the answer is 11 million newspaper pages or just one Hadoop cluster and one tech specialist who can move 4 terabytes of textual data to a new location in 24 hours.

Hadoop organizations benefits

Hadoop can be useful for a wide range of organizations and industries that work with large amounts of data. Here are a few examples of who might benefit from using Hadoop:

- E-commerce companies that need to process large volumes of customer data to gain insights into purchasing patterns, product recommendations, and customer behavior.
- Healthcare organizations that need to analyze patient data for medical research, disease diagnosis, and treatment.
- Financial institutions that need to analyze large amounts of transaction data to detect fraud and identify trends in financial markets.
- Government agencies that need to process large amounts of data for public safety, security, and policy-making purposes.
- Social media companies that need to process large amounts of user-generated data for personalized recommendations, targeted advertising, and trend analysis.

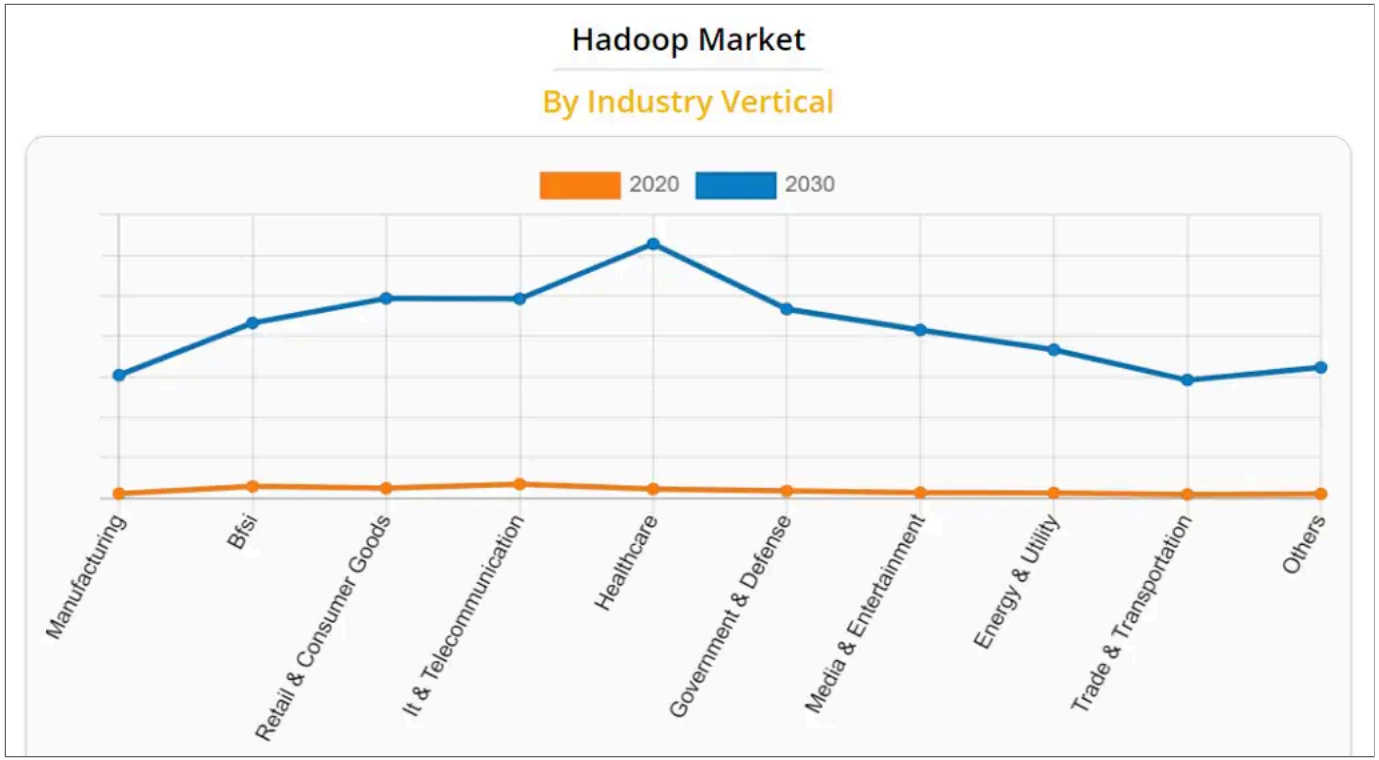
In general, any organization that needs to process and analyze large volumes of data can benefit from using Hadoop.

According to a study by the Business Application Research Center (BARC), Hadoop found intensive use as

- a runtime environment (sandbox) for classic business intelligence (BI), advanced analysis of large volumes of data, predictive maintenance, and data discovery and exploration;
- a store for raw data;
- a tool for large-scale data integration; and a suitable technology to implement data lake architecture.

Industries

Many industries, from manufacturing to banking to transportation, take advantage of what Hadoop can offer. And the number of companies adopting the platform is projected to increase by 2030. According to the latest report by Allied Market Research, the Big Data platform will see the biggest rise in adoption in telecommunication, healthcare, and government sectors.



Why Hadoop has been so important ?

To be able to compute big data

> According to an article published by IBM, traditional computing systems are designed to handle data sets that can fit into a single machine's memory. As the size of the data set grows beyond the capacity of a single machine, traditional systems become inefficient, slow, and expensive. In contrast, Hadoop is designed to handle data sets that are too large to fit into a single machine's memory by distributing the data and processing across a cluster of machines. This approach allows for parallel processing of large data sets, making Hadoop faster and more efficient than traditional systems for large-scale data processing. Additionally, Hadoop's fault-tolerant architecture enables it to recover from hardware failures, which is essential for large data sets that may span multiple machines.

IBM. "What is Hadoop?". IBM.com. https://www.ibm.com/cloud/learn/hadoop

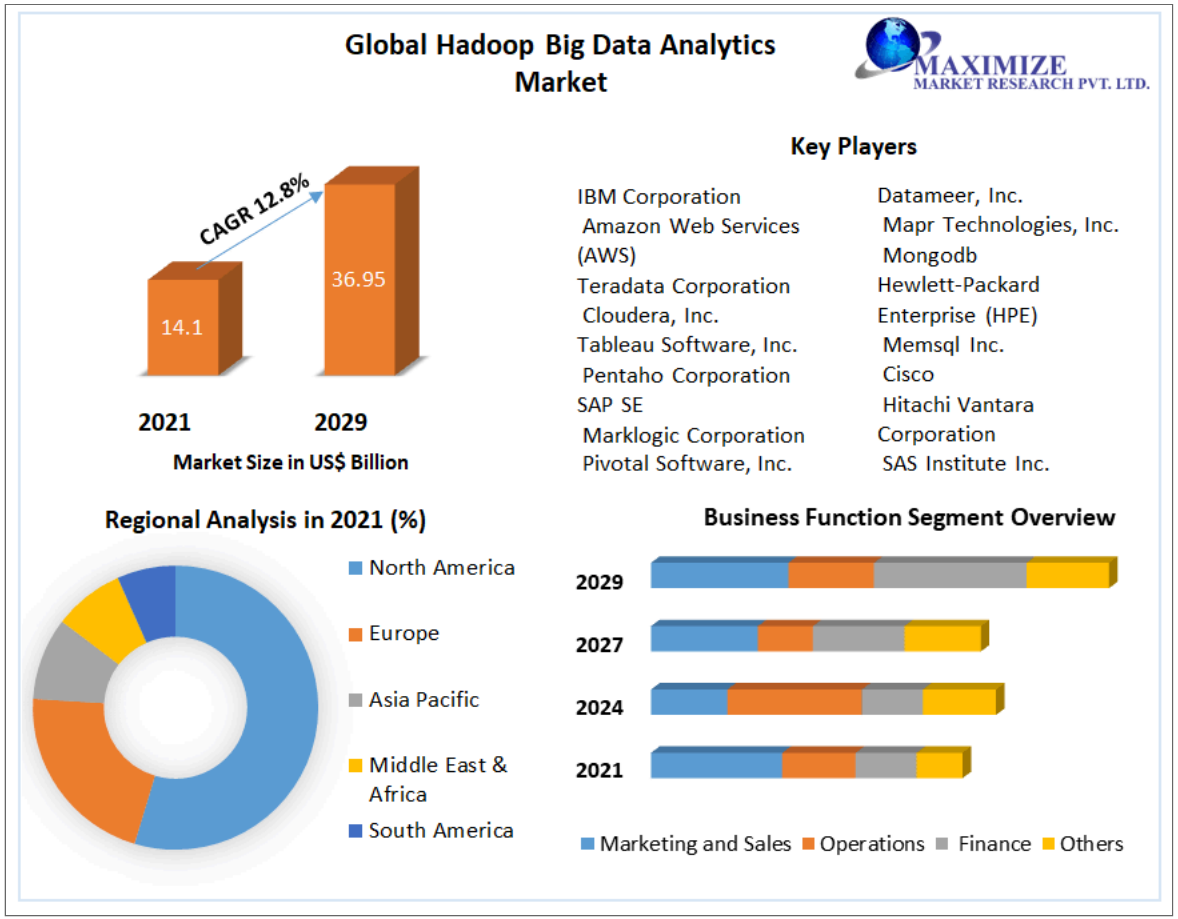
Is still valid ?

Hadoop is still a valid technology in 2024. While there are other big data technologies that have emerged in recent years, such as Apache Spark and Apache Flink, Hadoop remains a popular and widely used platform for distributed computing and big data processing.

Hadoop has a large and active community of users and developers who continue to improve and maintain the platform. In addition, many companies have invested heavily in Hadoop infrastructure and continue to use it as a core component of their big data processing pipelines.

However, it’s worth noting that Hadoop has also evolved over time to incorporate newer technologies and approaches, such as the integration of Spark as a processing engine and the use of cloud-based Hadoop offerings. As such, it’s important to stay up-to-date with the latest developments in the Hadoop ecosystem to ensure that you are making the most of this powerful technology.

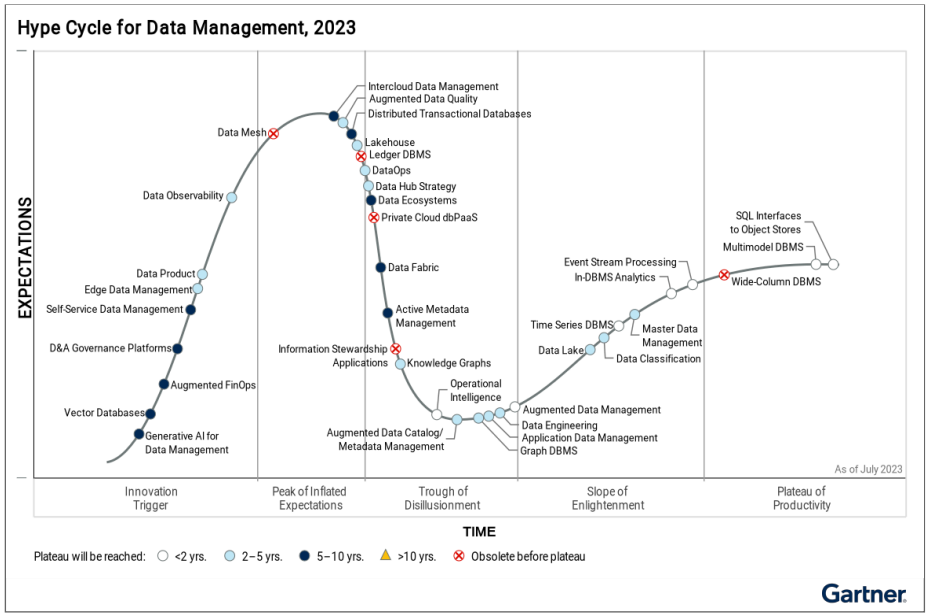
Some figures

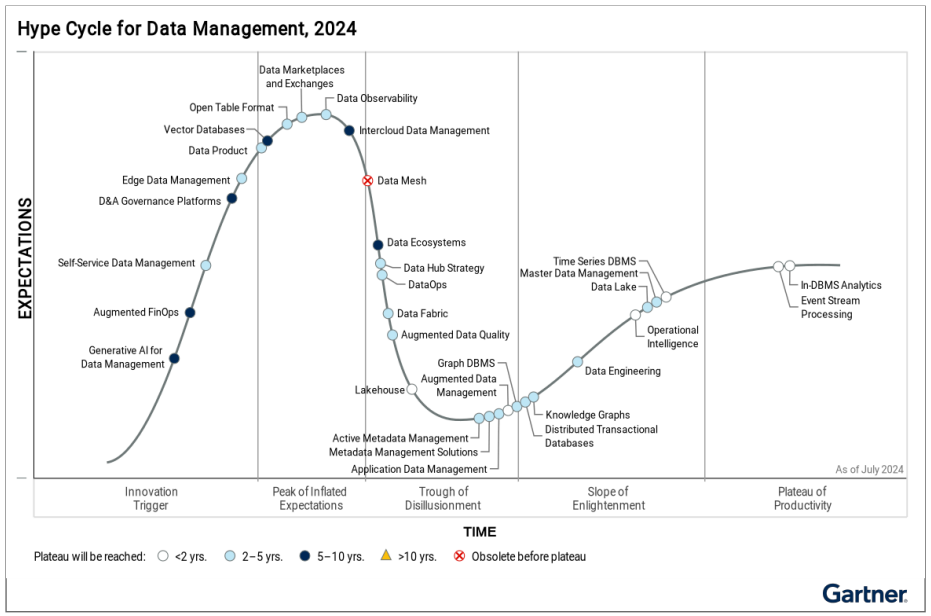


https://www.maximizemarketresearch.com/market-report/global-hadoop-big-data-analytics-market/6866/

https://www.acumenresearchandconsulting.com/hadoop-market

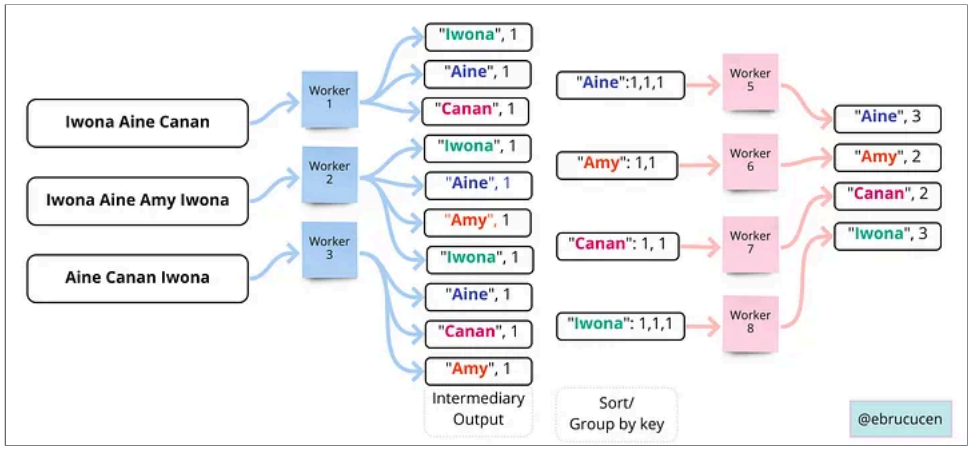
Hype Cycle for Data Management





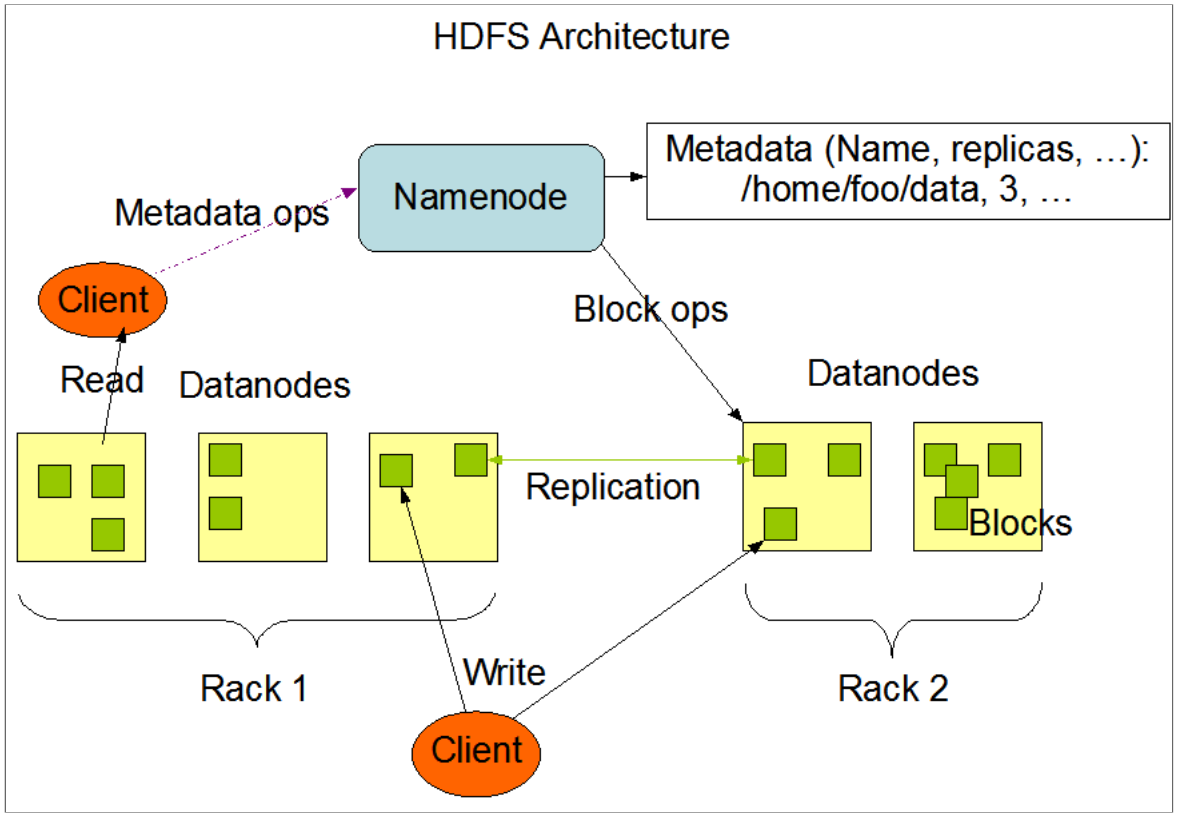
How it works ?

Map Reduce



https://towardsdatascience.com/series-on-distributed-computing-1-mapreduce-fcc3cc2dfb5

HDFS

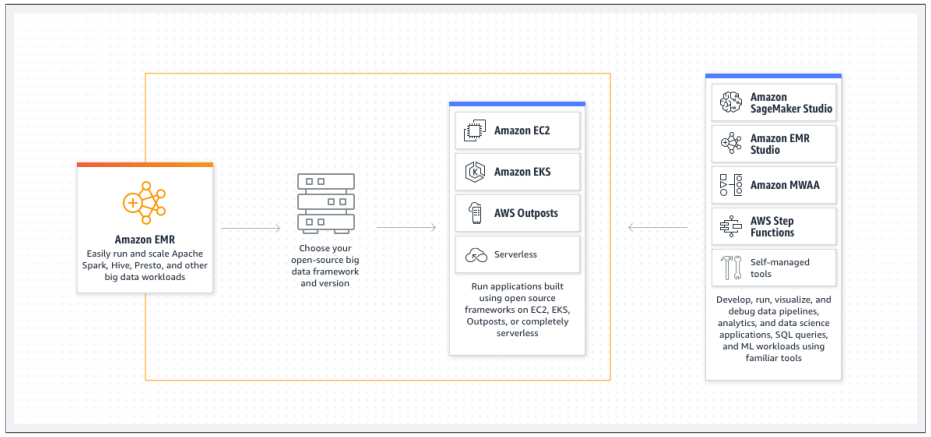


https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html

Where to use ?

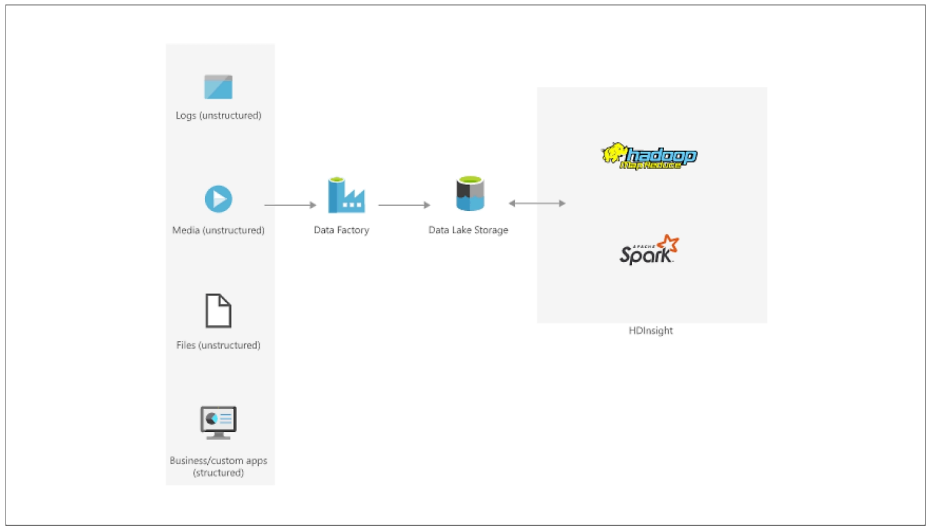
AWS EMR

Easily run and scale Apache Spark, Hive, Presto, and other big data workloads
https://aws.amazon.com/emr/



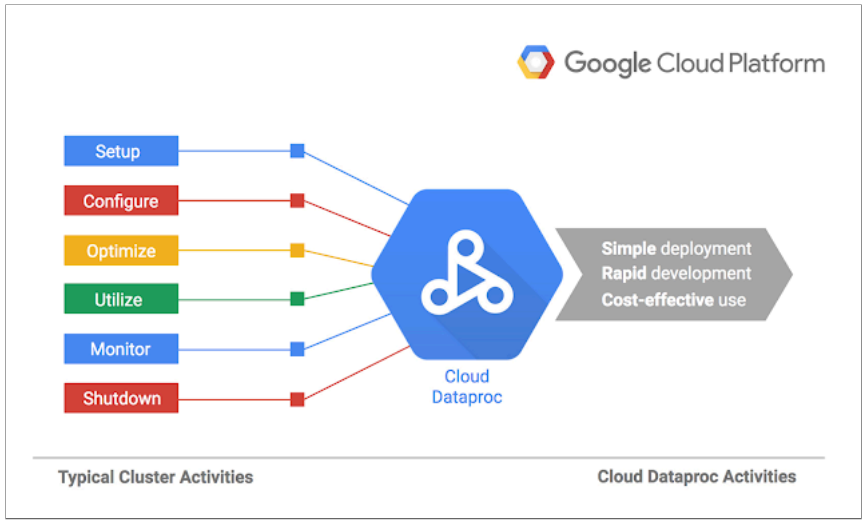
Azure HD Insight

Run popular open-source frameworks—including Apache Hadoop, Spark, Hive, Kafka, and more—using Azure HDInsight, a customizable, enterprise-grade service for open-source analytics. Effortlessly process massive amounts of data and get all the benefits of the broad open-source project ecosystem with the global scale of Azure. Easily migrate your big data workloads and processing to the cloud.
https://azure.microsoft.com/en-gb/products/hdinsight/



Google Cloud - Dataproc

Dataproc is a fully managed and highly scalable service for running Apache Hadoop, Apache Spark, Apache Flink, Presto, and 30+ open source tools and frameworks. Use Dataproc for data lake modernization, ETL, and secure data science, at scale, integrated with Google Cloud, at a fraction of the cost.
https://cloud.google.com/dataproc



Time to do practice

A Word Count with Map Reduce on DataProc

- Dasbhoard: **https://miro.com/app/board/uXjVMcdLcug=/?share_link_id=121090137939**
- Source Example: **https://github.com/apache/hadoop/blob/trunk/hadoop-mapreduce-project/hadoop-mapreduce-examples/src/main/java/org/apache/hadoop/examples/WordCount.java**

Spoiler of next lessons

https://6sense.com/tech/big-data-analytics/databricks-vs-apachehadoop

Comparing the customer bases of Databricks and Apache Hadoop

Comparing the customer bases of Databricks and Apache Hadoop, we can see that Databricks has 12,184 custo
has 12,133 customer(s). In the Big Data Analytics category, with 12,184 customer(s) Databricks stands at 1st plac
Hadoop with 12,133 customer(s), is at the 2nd place.

All

Databricks

12,184 Customer

Apache Hadoo

12,133 Customer

Tutorials

- https://www.cloudskillsboost.google/focuses/672?catalog_rank=%7B%22rank%22%3A9%2C%22num_filters%22%3A0%2C%22has_search%22%3Atrue%7D&parent=catalog&search_id=228990
- https://cloud.google.com/composer/docs/tutorials/hadoop-wordcount-job
- https://www.geeksforgeeks.org/introduction-to-apache-pig/
- https://towardsdatascience.com/the-charm-of-apache-pig-fdc92b5cc3b4
- https://codelabs.developers.google.com/codelabs/intro-cloud-composer#0
- https://www.freecodecamp.org/news/what-is-google-dataproc/
- https://github.com/apache/hadoop/tree/trunk/hadoop-mapreduce-project/hadoop-mapreduce-examples/src/main/java/org/apache/hadoop/examples
- https://medium.com/edureka/mapreduce-tutorial-3d9535ddbe7c

References

- https://hadoop.apache.org/
- https://www.linkedin.com/learning/learning-hadoop-2/getting-started-with-hadoop?autoplay=true
- https://www.projectpro.io/article/hadoop-vs-spark-not-mutually-exclusive-but-better-together/235
- https://towardsdatascience.com/series-on-distributed-computing-1-mapreduce-fcc3cc2dfb5
- https://www.databricks.com/glossary/hadoop-ecosystem
- https://medium.com/geekculture/mapreduce-with-python-5d12a772d5b3
- https://engineering.linkedin.com/blog/2021/scaling-linkedin-s-hadoop-yarn-cluster-beyond-10-000-nodes
- https://www.uber.com/en-IT/blog/scaling-hdfs/
- https://engineering.linkedin.com/blog/2021/the-exabyte-club–linkedin-s-journey-of-scaling-the-hadoop-distr
-