HADOOP INTRODUCTION

Hadoop

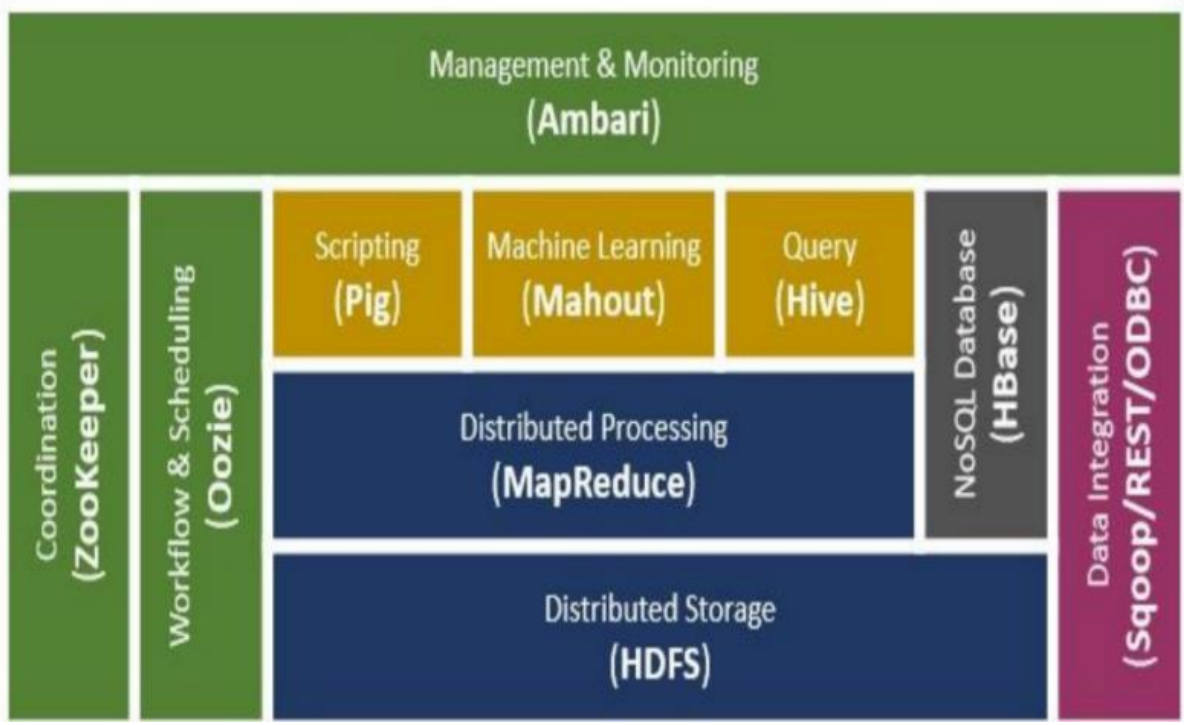
It is the technology to store massive datasets on a cluster of cheap machines in a distributed manner. It provides Big Data analytics through distributed computing framework. It is an open-source software developed as a project by Apache Software Foundation. Doug Cutting created Hadoop.

In the year 2008 Yahoo gave Hadoop to Apache Software Foundation. Since then two versions of Hadoop has come. Version 1.0 in the year 2011 and version 2.0.6 in the year 2013. Hadoop comes in various flavors like Cloudera, IBM BigInsight, MapR and Hortonworks.

Core Components of Hadoop

- **Hadoop Distributed File System (HDFS)** – It is the storage layer of Hadoop.
- **Map-Reduce** – It is the data processing layer of Hadoop.
- **YARN** – It is the resource management layer of Hadoop.

Hadoop Ecosystem and components,



How Hadoop Works?

- Hadoop does distributed processing for huge data sets across the cluster of commodity servers and works on multiple machines simultaneously. To process any data, the client submits data and

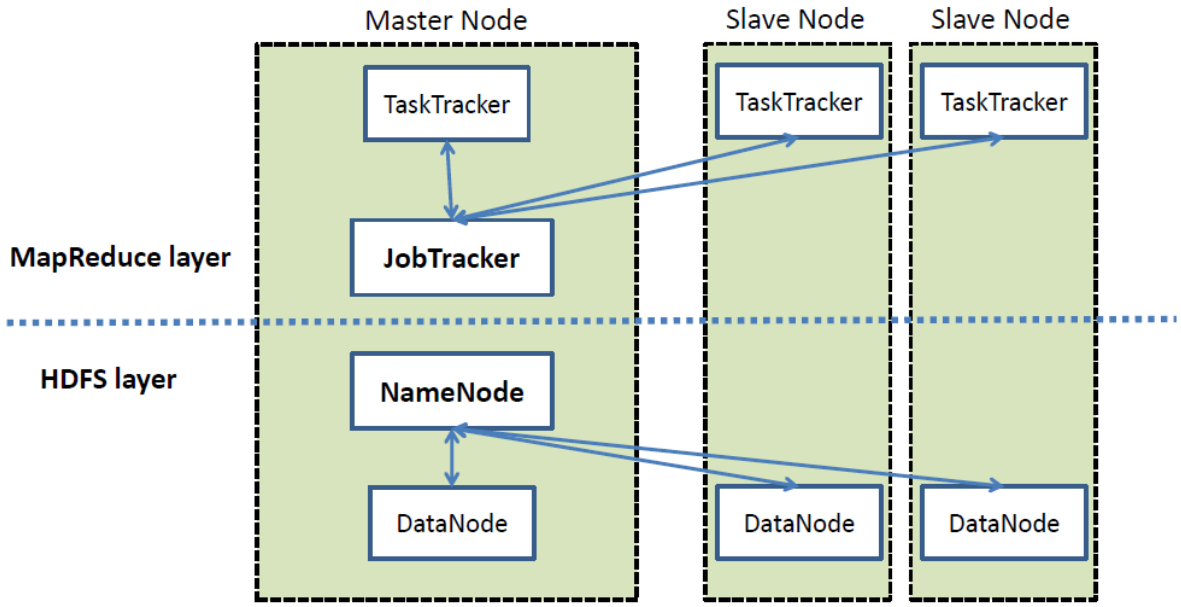
program to Hadoop. HDFS stores the data while MapReduce process the data and Yarn divide the tasks.

Why Hadoop?

- Apache Hadoop is not only a storage system but is a platform for data storage as well as processing.
- It is scalable (as we can add more nodes on the fly), Fault-tolerant (Even if nodes go down, data processed by another node).
- Flexibility to store and mine any type of data whether it is structured, semi-structured or unstructured. It is not bounded by a single schema.
- Excels at processing data of complex nature. Its scale-out architecture divides workloads across many nodes. Another added advantage is that its flexible filesystem eliminates ETL bottlenecks.
- Scales economically, as discussed it can deploy on commodity hardware. Apart from this its open-source nature guards against vendor lock.

Hadoop Architecture

Hadoop works in master-slave fashion. There is a master node and there are n numbers of slave nodes where n can be 1000s. Master manages, maintains and monitors the slaves while slaves are the actual worker nodes.
Master stores the metadata while slaves are the nodes which store the data. Distributed data stores in the cluster. The client connects with the master node to perform any task.



Key Features of Hadoop

1. Reliability

- In the Hadoop cluster, if any node goes down, it will not disable the whole cluster. Instead, another node will take the place of the failed node. Hadoop cluster will continue functioning as nothing has happened. Hadoop has built-in fault tolerance feature.

2. Scalable

- Hadoop gets integrated with cloud-based service. If you are installing Hadoop on the cloud you need not worry about scalability. You can easily procure more hardware and expand your Hadoop cluster within minutes.

3. Economical

- Hadoop gets deployed on commodity hardware which is cheap machines. This makes Hadoop very economical. Also, as Hadoop is an open system software there is no cost of license too.

4. Distributed Processing

- In Hadoop, any job submitted by the client gets divided into the number of sub-tasks. These sub-tasks are independent of each other. Hence, they execute in parallel giving high throughput.

5. Distributed Storage

- Hadoop splits each file into the number of blocks. These blocks get stored distributed on the cluster of machines.

6. Fault Tolerance

- Hadoop replicates every block of file many times depending on the replication factor. Replication factor is 3 by default. In Hadoop suppose any node goes down then the data on that node gets recovered. This is because this copy of the data would be available on other nodes due to replication. Hadoop is fault tolerant.

Hadoop Flavors

- Apache – Vanilla flavor, as the actual code is residing in Apache repositories.
- Hortonworks – Popular distribution in the industry.
- Cloudera – It is the most popular in the industry.
- MapR – It has rewritten HDFS and its HDFS is faster as compared to others.
- IBM – Proprietary distribution is known as Big Insights.

Advantage:

1. Varied Data Sources

- Hadoop accepts a variety of data. Data can come from a range of sources like email conversation, social media etc. and can be of structured or unstructured form. Hadoop can derive value from diverse data. Hadoop can accept data in a text file, XML file, images, CSV files etc.

2. Cost-effective

- Hadoop is an economical solution as it uses a cluster of commodity hardware to store data. Commodity hardware is cheap machines hence the cost of adding nodes to the framework is not much high. In Hadoop 3.0 we have only 50% of storage overhead as opposed to 200% in Hadoop2. x. This requires less machine to store data as the redundant data decreased significantly.

3. Performance

- Hadoop with its distributed processing and distributed storage architecture processes huge amounts of data with high speed. Hadoop even defeated supercomputer the fastest machine in 2008. It divides the input data file into several blocks and stores data in these blocks over several nodes. It also divides the task that user submits into various sub-tasks which assign to these worker nodes containing required data and these sub-tasks run in parallel thereby improving the performance.

4. Fault-Tolerant

- In Hadoop 3.0 fault tolerance is provided by erasure coding. For example, 6 data blocks produce 3 parity blocks by using erasure coding technique, so HDFS stores a total of these 9 blocks. In event of failure of any node the data block affected can be recovered by using these parity blocks and the remaining data blocks.

5. Highly Available

- In Hadoop 2.x, HDFS architecture has a single active NameNode and a single Standby NameNode, so if a NameNode goes down then we have standby NameNode to count on. But Hadoop 3.0 supports multiple standby NameNode making the system even more highly available as it can continue functioning in case if two or more NameNodes crashes.

6. Low Network Traffic

- In Hadoop, each job submitted by the user is split into a number of independent sub-tasks and these sub-tasks are assigned to the data nodes thereby moving a small amount of code to data rather than moving huge data to code which leads to low network traffic.

7. High Throughput

- Throughput means job done per unit time. Hadoop stores data in a distributed fashion which allows using distributed processing with ease. A given job gets divided into small jobs which work on chunks of data in parallel thereby giving high throughput.

8. Open Source

- Hadoop is an open source technology i.e. its source code is freely available. We can modify the source code to suit a specific requirement.

9. Scalable

- Hadoop works on the principle of horizontal scalability i.e. we need to add the entire machine to the cluster of nodes and not change the configuration of a machine-like adding RAM, disk and so on which is known as vertical scalability. Nodes can be added to Hadoop cluster on the fly making it a scalable framework.

10. Ease of use

- The Hadoop framework takes care of parallel processing, MapReduce programmers does not need to care for achieving distributed processing, it is done at the backend automatically.

11. Compatibility

- Most of the emerging technology of Big Data is compatible with Hadoop like Spark, Flink etc. They have got processing engines which work over Hadoop as a backend i.e. We use Hadoop as data storage platforms for them.

12. Multiple Languages Supported

- Developers can code using many languages on Hadoop like C, C++, Perl, Python, Ruby, and Groovy.

Disadvantage:

1. Issue with Small Files

Hadoop is suitable for a small number of large files but when it comes to the application which deals with many small files, Hadoop fails here. A small file is nothing but a file which is significantly smaller than Hadoop's block size which can be either 128MB or 256MB by default. These large number of small files overload the Namenode as it stores namespace for the system and makes it difficult for Hadoop to function.

2. Vulnerable by Nature

Hadoop is written in Java which is a widely used programming language hence it is easily exploited by cyber criminals which makes Hadoop vulnerable to security breaches.

3. Processing Overhead

In Hadoop, the data is read from the disk and written to the disk which makes read/write operations very expensive when we are dealing with tera and petabytes of data. Hadoop cannot do in-memory calculations hence it incurs processing overhead.

4. Supports Only Batch Processing

At the core, Hadoop has a batch processing engine which is not efficient in stream processing. It cannot produce output in real-time with low latency. It only works on data which we collect and store in a file in advance before processing.

5. Iterative Processing

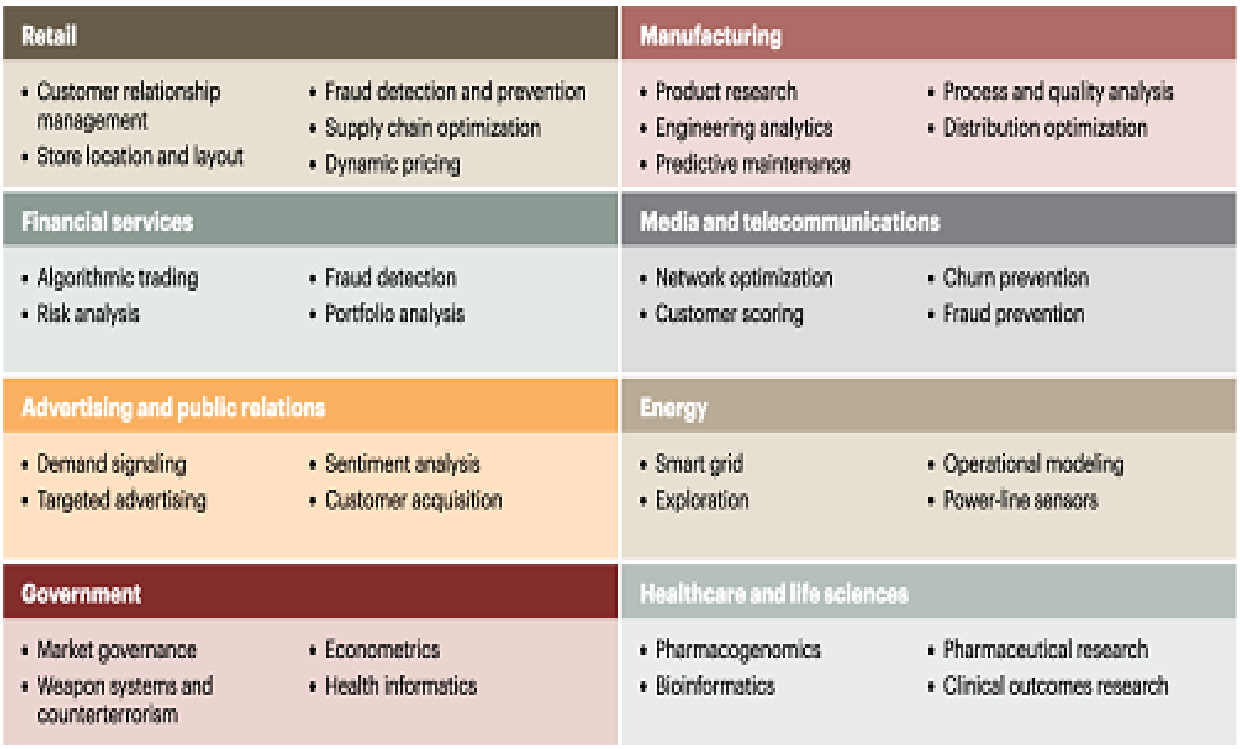
Hadoop cannot do iterative processing by itself. Machine learning or iterative processing has a cyclic data flow whereas Hadoop has data flowing in a chain of stages where output on one stage becomes the input of another stage.

6. Security

For security, Hadoop uses Kerberos authentication which is hard to manage. It is missing encryption at storage and network levels which are a major point of concern.

So, this was all about Hadoop Pros and Cons. Hope you liked our explanation.

Hadoop Industry Usage



Top 15 companies using Hadoop

	Company	Business	Technical Specs	Uses
1	*Facebook*	Social Site	*8 cores and 12 TB of storage*	Used as a source for reporting and machine learning
2	Twitter	Social site		Hadoop is used since 2010 to store and process tweets, log files using LZO compression technique as it is fast and also helps release CPU for other tasks.
3	*LinkedIn*	Social site	2X4 and 2X6 cores – 6X2TB SATA 4100 nodes	LinkedIn's data flows through Hadoop clusters.User activity, server metrics, images,transaction logs stored in HDFS are used by data analysts for business analytics like discovering people you may know.
4	*Yahoo!*	Online Portal	4500 nodes – 1TB storage, 16 GB RAM	Used for scaling tests
5	*AOL*	Online portal	*ETL style processing and statistics generation*	Targets machines and dual processors
6	*EBay*	Ecommerce	*4K+ nodes cluster*	With 300+ million users browsing more than 350 million products

				listed on their website, eBay has one of the largest Hadoop clusters in the industry that run prominelty on MapReduce Jobs. Hadoop is used at eBay for Search Optimization and Research.
7	*Alibaba*	E-Commerce	Processes *15-node cluster business data*	Analyzes vertical search engine
8	*Cloudspace*	IT developer		Specializes in designing and building web applications
9	*FOX Audience Network*	News TV Channel	30-70 machine clusters	Used for log analysis and machine learning
10	*Adobe*	Publishing and editing software	*30 nodes running HDFS, 5 to 14 nodes* HBase	*Social services to structured data storage*
11	*Infosys*	IT Consulting	Per client requirements	Client projects in finance, telecom and retail.
12	*Cognizant*	IT Consulting	Per client requirements	Client projects in finance, telecom and retail.
13	*Accenture*	IT Consulting	Per client requirements	Client projects in finance, telecom and retail.
14	*Hulu*	Video Delivery	13 machine clusters – 8 cores, 4 TB	Used for analysis and log storage

15	*Last.fm*	Online FM Music	100 nodes, 8 TB storage	Calculation of charts and data testing

Reference:

1. Data-flair
2. https://www.dezyre.com/article/top-10-industries-using-big-data-and-121-companies-who-hire-hadoop-developers/69