**What is Big data?**

Data which are very large in size is called Big Data. Normally we work on data of size MB(WordDoc ,Excel) or maximum GB(Movies, Codes) but data in Peta bytes i.e. 10^15 byte size is called Big Data. It is stated that almost 90% of today's data has been generated in the past 3 years.

This is a term related to extracting meaningful data by analyzing the huge amount of complex, variously formatted data generated at high speed, that cannot be handled, processed by the traditional system.

**Source of Big Data:**

1. Social Media: Today’s world a good percent of the total world population is engaged with social media like Facebook, WhatsApp, Twitter, YouTube, Instagram, etc. Each activity on such media like uploading a photo, video, sending the message, making comment, putting like, etc create data.
2. Sensor placed on the various place: Sensor placed in various place of the city that gathers data on temperature, humidity, etc. A camera placed beside road gather information about traffic condition, creates data. Security camera placed in a sensitive area like airport, railway station, shopping mall create a lot of data.
3. Customer feedback on the product or service of the various company on their website creates data. For Example, a retail commercial site like Amazon, Walmart, Flipkart, Myntra gather customer feedback on the quality of their product, delivery time. Telecom company, other service provider organization seek customer experience with their service. These create a lot of data.
4. IoT Appliance: Electronic devices that are connected to the internet create data for their smart functionality, examples are a smart TV, smart washing machine, smart coffee machine, smart AC, etc. It is machine-generated data that are created by sensor kept in various devices.  
   For Example, Smart printing machine – it is connected to the internet. A number of such printing machines connected to a network can transfer data within each other. So, if anyone loads a file copy in one printing machine, system store that file content, another printing machine kept in another building or another floor can print out that file hard copy. Such data transfer between various printing machines generates data.
5. In an e-commerce transaction, business transaction, banking, and the stock market, lots of records stored considered as one of the sources of big data. Payment through credit card, debit card or by another electronic way, all these are kept recorded as data.
6. GPS in the vehicle that helps in monitoring movement of the vehicle to shorten the path for a destination to cut fuel, time consumption. This system creates huge data of vehicle position and movement.

the three Vs; volume, velocity, and variety, these are key to understanding how we can measure big data and just how very different ‘big data’ is to old fashioned data.

**Volume**

The most obvious one is where we’ll start. Big data is about volume. Volumes of data that can reach unprecedented heights in fact. It’s estimated that 2.5 quintillion bytes of data is created each day

**Velocity**

The growth of data, and the resulting importance of it, has changed the way we see data. There once was a time when we didn’t see the importance of data in the corporate world, but with the change of how we gather it, we’ve come to rely on it day to day. Velocity essentially measures how fast the data is coming in

**Variety**

Data was once collected from one place and delivered in one format. Once taking the shape of database files - such as, excel, csv and access - it is now being presented in non-traditional forms, like video, text, pdf, and graphics on social media, as well as via tech such as wearable devices. Data is structured as well as unstructured. Log file, CCTV footage is unstructured data. Data which can be saved in tables are structured data like the transaction data of the bank.

### How does big data analytics work?

Data analysts, [data scientists](https://www.techtarget.com/searchenterpriseai/definition/data-scientist), predictive modelers, statisticians and other analytics professionals collect, process, clean and analyze growing volumes of structured transaction data as well as other forms of data not used by conventional BI and analytics programs.

the four steps of the big data analytics process:

1. Data professionals **collect** data from a variety of different sources. Often, it is a mix of [semistructured](https://www.techtarget.com/whatis/definition/semi-structured-data) and unstructured data. While each organization will use different data streams, some common sources include:
   1. internet clickstream data;
   2. web server logs;
   3. cloud applications;
   4. mobile applications;
   5. social media content;
   6. text from customer emails and survey responses;
   7. mobile phone records; and
   8. machine data captured by [sensors](https://internetofthingsagenda.techtarget.com/definition/sensor-data) connected to the internet of things (IoT).
2. Data is prepared and **processed**. After data is collected and stored in a [data warehouse or data lake](https://www.techtarget.com/searchdatamanagement/feature/Beyond-the-RDBMS-Data-warehouse-vs-data-lake-vs-data-mart), data professionals must organize, configure and partition the data properly for analytical queries. Thorough [data preparation](https://www.techtarget.com/searchbusinessanalytics/definition/data-preparation) and processing makes for higher performance from analytical queries.
3. Data is**cleansed** to improve its quality. Data professionals scrub the data using scripting tools or data quality software. They look for any errors or inconsistencies, such as duplications or formatting mistakes, and organize and tidy up the data.
4. The [collected, processed and cleaned data](https://www.techtarget.com/searchdatamanagement/feature/Big-data-collection-processes-challenges-and-best-practices) is **analyzed** with analytics software. This includes tools for:
   1. [data mining](https://searchsqlserver.techtarget.com/definition/data-mining), which sifts through data sets in search of patterns and relationships
   2. predictive analytics, which builds models to forecast customer behavior and other future actions, scenarios and trends
   3. machine learning, which taps various algorithms to analyze large data sets
   4. deep learning, which is a more advanced offshoot of machine learning
   5. [text mining](https://www.techtarget.com/searchbusinessanalytics/definition/text-mining) and statistical analysis software
   6. artificial intelligence (AI)
   7. mainstream business intelligence software
   8. data visualization tools

### Big data analytics technologies and tools

Many different types of tools and technologies are used to support big data analytics processes. Common technologies and tools used to enable big data analytics processes include:

* [**Hadoop**](https://www.techtarget.com/searchdatamanagement/definition/Hadoop)**,**which is an open source framework for storing and processing big data sets. Hadoop can handle large amounts of structured and unstructured data.
* **Predictive analytics** hardware and software, which process large amounts of complex data, and use [machine learning and statistical algorithms to make predictions](https://www.techtarget.com/searchbusinessanalytics/tip/Big-data-vs-machine-learning-How-they-differ-and-relate) about future event outcomes. Organizations use predictive analytics tools for fraud detection, marketing, risk assessment and operations.
* **Stream analytics**tools, which are used to filter, aggregate and analyze big data that may be stored in many different formats or platforms.
* **Distributed storage**data, which is replicated, generally on a non-relational database. This can be as a measure against independent node failures, lost or corrupted big data, or to provide low-latency access.
* [**NoSQL**](https://www.techtarget.com/searchdatamanagement/definition/NoSQL-Not-Only-SQL)**databases**, which are non-relational data management systems that are useful when working with large sets of distributed data. They do not require a fixed schema, which makes them ideal for raw and unstructured data.
* **A data lake**is a large storage repository that holds native-format raw data until it is needed. Data lakes use a flat architecture.
* **A**[**data warehouse**](https://www.techtarget.com/searchdatamanagement/definition/data-warehouse)**,**which is a repository that stores large amounts of data collected by different sources. Data warehouses typically store data using predefined schemas.
* **Knowledge discovery/big data mining** tools, which enable businesses to mine large amounts of structured and unstructured big data.
* **In-memory data fabric**, which distributes large amounts of data across system memory resources. This helps provide low latency for data access and processing.
* **Data virtualization**, which enables data access without technical restrictions.
* **Data integration software,**which enables big data to be streamlined across different platforms, including Apache, Hadoop, MongoDB and Amazon EMR.
* **Data quality software**, which cleanses and enriches large data sets.
* [**Data preprocessing**](https://searchsqlserver.techtarget.com/definition/data-preprocessing)**software,** which prepares data for further analysis. Data is formatted and unstructured data is cleansed.
* **Spark,** which is an open source cluster computing framework used for batch and stream data processing.

### Big data analytics benefits

The benefits of using big data analytics include:

* Quickly analyzing large amounts of data from different sources, in many different formats and types.
* Rapidly making better-informed decisions for effective strategizing, which can benefit and improve the supply chain, operations and other areas of strategic decision-making.
* Cost savings, which can result from new business process efficiencies and optimizations.
* A better understanding of customer needs, behavior and sentiment, which can lead to better marketing insights, as well as provide information for product development.
* Improved, better informed [risk management](https://www.techtarget.com/searchsecurity/answer/Risk-assessment-vs-risk-analysis-vs-risk-management) strategies that draw from large sample sizes of data.

### Big data analytics challenges

Despite the wide-reaching benefits that come with using big data analytics, its use also comes with challenges:

* **Accessibility of data.**With larger amounts of data, storage and processing become more complicated. Big data should be stored and maintained properly to ensure it can be used by less experienced data scientists and analysts.
* **Data quality maintenance.**With high volumes of data coming in from a variety of sources and in different formats, [data quality management](https://www.techtarget.com/searchenterpriseai/feature/9-data-quality-issues-that-can-sideline-AI-projects) for big data requires significant time, effort and resources to properly maintain it.
* **Data security.**The complexity of big data systems presents unique security challenges. Properly addressing security concerns within such a complicated big data ecosystem can be a complex undertaking.
* **Choosing the right tools.**Selecting from the vast array of big data analytics tools and platforms available on the market can be confusing, so organizations must know how to pick the best tool that aligns with users' needs and infrastructure.
* With a potential lack of internal analytics skills and the high cost of hiring experienced data scientists and engineers, some organizations are finding it hard to fill the gaps.

# What is Hadoop

Hadoop is an open source framework from Apache and is used to store process and analyze data which are very huge in volume. Hadoop is written in Java and is not OLAP (online analytical processing). It is used for batch/offline processing.It is being used by Facebook, Yahoo, Google, Twitter, LinkedIn and many more. Moreover it can be scaled up just by adding nodes in the cluster.

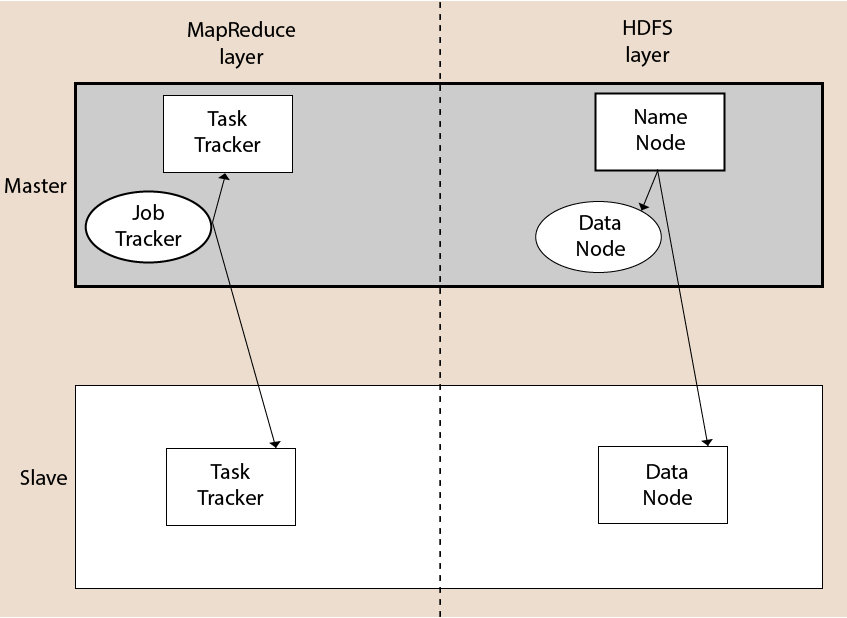
## Modules of Hadoop

1. **HDFS:** Hadoop Distributed File System. Google published its paper GFS and on the basis of that HDFS was developed. It states that the files will be broken into blocks and stored in nodes over the distributed architecture.
2. **Yarn:** Yet another Resource Negotiator is used for job scheduling and manage the cluster.
3. **Map Reduce:** This is a framework which helps Java programs to do the parallel computation on data using key value pair. The Map task takes input data and converts it into a data set which can be computed in Key value pair. The output of Map task is consumed by reduce task and then the output of reducer gives the desired result.
4. **Hadoop Common:** These Java libraries are used to start Hadoop and are used by other Hadoop modules.

## Hadoop Architecture

The Hadoop architecture is a package of the file system, MapReduce engine and the HDFS (Hadoop Distributed File System). The MapReduce engine can be MapReduce/MR1 or YARN/MR2.

A Hadoop cluster consists of a single master and multiple slave nodes. The master node includes Job Tracker, Task Tracker, NameNode, and DataNode whereas the slave node includes DataNode and TaskTracker.



## Hadoop Distributed File System

The Hadoop Distributed File System (HDFS) is a distributed file system for Hadoop. It contains a master/slave architecture. This architecture consist of a single NameNode performs the role of master, and multiple DataNodes performs the role of a slave.

Both NameNode and DataNode are capable enough to run on commodity machines. The Java language is used to develop HDFS. So any machine that supports Java language can easily run the NameNode and DataNode software.

### NameNode

* It is a single master server exist in the HDFS cluster.
* As it is a single node, it may become the reason of single point failure.
* It manages the file system namespace by executing an operation like the opening, renaming and closing the files.
* It simplifies the architecture of the system.

### DataNode

* The HDFS cluster contains multiple DataNodes.
* Each DataNode contains multiple data blocks.
* These data blocks are used to store data.
* It is the responsibility of DataNode to read and write requests from the file system's clients.
* It performs block creation, deletion, and replication upon instruction from the NameNode.

### Job Tracker

* The role of Job Tracker is to accept the MapReduce jobs from client and process the data by using NameNode.
* In response, NameNode provides metadata to Job Tracker.

### Task Tracker

* It works as a slave node for Job Tracker.
* It receives task and code from Job Tracker and applies that code on the file. This process can also be called as a Mapper.

## MapReduce Layer

The MapReduce comes into existence when the client application submits the MapReduce job to Job Tracker. In response, the Job Tracker sends the request to the appropriate Task Trackers. Sometimes, the TaskTracker fails or time out. In such a case, that part of the job is rescheduled.

## Advantages of Hadoop

* **Fast:** In HDFS the data distributed over the cluster and are mapped which helps in faster retrieval. Even the tools to process the data are often on the same servers, thus reducing the processing time. It is able to process terabytes of data in minutes and Peta bytes in hours.
* **Scalable:** Hadoop cluster can be extended by just adding nodes in the cluster.
* **Cost Effective:** Hadoop is open source and uses commodity hardware to store data so it really cost effective as compared to traditional relational database management system.
* **Resilient to failure:** HDFS has the property with which it can replicate data over the network, so if one node is down or some other network failure happens, then Hadoop takes the other copy of data and use it. Normally, data are replicated thrice but the replication factor is configurable.

\* **The Hadoop was started by Doug Cutting and Mike Cafarella in 2002. Its origin was the Google File System paper, published by Google.**

Hadoop Eco system

Apache Hadoop is an open source framework intended to make interaction with [big data](https://write.geeksforgeeks.org/geek/the-world-of-big-data/) easier. Hadoop Ecosystem is a platform or a suite which provides various services to solve the big data problems. It includes Apache projects and various commercial tools and solutions. There are four major elements of Hadoop i.e. HDFS, MapReduce, YARN, and Hadoop Common. Most of the tools or solutions are used to supplement or support these major elements. All these tools work collectively to provide services such as absorption, analysis, storage and maintenance of data etc.

the components that collectively form a Hadoop ecosystem: 

HDFS: Hadoop Distributed File System

YARN: Yet Another Resource Negotiator

MapReduce: Programming based Data Processing

Spark: In-Memory data processing

PIG, HIVE: Query based processing of data services

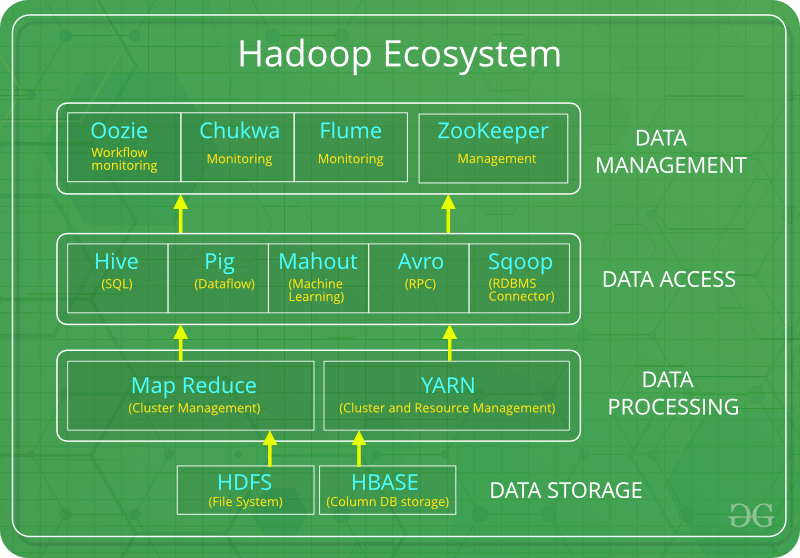
HBase: NoSQL Database

Mahout, Spark MLLib: [Machine Learning](https://www.geeksforgeeks.org/machine-learning/)algorithm libraries

Solar, Lucene: Searching and Indexing

Zookeeper: Managing cluster

Oozie: Job Scheduling



HDFS: 

HDFS is the primary or major component of Hadoop ecosystem and is responsible for storing large data sets of structured or unstructured data across various nodes and thereby maintaining the metadata in the form of log files.

HDFS consists of two core components i.e.

Name node

Data Node

Name Node is the prime node which contains metadata (data about data) requiring comparatively fewer resources than the data nodes that stores the actual data. These data nodes are commodity hardware in the distributed environment. Undoubtedly, making Hadoop cost effective.

HDFS maintains all the coordination between the clusters and hardware, thus working at the heart of the system.

YARN:   
Yet Another Resource Negotiator, as the name implies, YARN is the one who helps to manage the resources across the clusters. In short, it performs scheduling and resource allocation for the Hadoop System.

Consists of three major components i.e.

Resource Manager

Nodes Manager

Application Manager

Resource manager has the privilege of allocating resources for the applications in a system whereas Node managers work on the allocation of resources such as CPU, memory, bandwidth per machine and later on acknowledges the resource manager. Application manager works as an interface between the resource manager and node manager and performs negotiations as per the requirement of the two.

MapReduce: 

By making the use of distributed and parallel algorithms, MapReduce makes it possible to carry over the processing’s logic and helps to write applications which transform big data sets into a manageable one.

MapReduce makes the use of two functions i.e. Map() and Reduce() whose task is:

Map() performs sorting and filtering of data and thereby organizing them in the form of group. Map generates a key-value pair based result which is later on processed by the Reduce() method.

Reduce(), as the name suggests does the summarization by aggregating the mapped data. In simple, Reduce() takes the output generated by Map() as input and combines those tuples into smaller set of tuples.

PIG:

 Pig was basically developed by Yahoo which works on a pig Latin language, which is Query based language similar to SQL.

It is a platform for structuring the data flow, processing and analyzing huge data sets.

Pig does the work of executing commands and in the background, all the activities of MapReduce are taken care of. After the processing, pig stores the result in HDFS.

Pig Latin language is specially designed for this framework which runs on Pig Runtime. Just the way Java runs on the [JVM](https://www.geeksforgeeks.org/jvm-works-jvm-architecture/).

Pig helps to achieve ease of programming and optimization and hence is a major segment of the Hadoop Ecosystem.

HIVE:   
 With the help of SQL methodology and interface, HIVE performs reading and writing of large data sets. However, its query language is called as HQL (Hive Query Language).

It is highly scalable as it allows real-time processing and batch processing both. Also, all the SQL datatypes are supported by Hive thus, making the query processing easier.

Similar to the Query Processing frameworks, HIVE too comes with two components: JDBC Drivers and HIVE Command Line.

JDBC, along with ODBC drivers work on establishing the data storage permissions and connection whereas HIVE Command line helps in the processing of queries.

Mahout:   
Mahout, allows Machine Learnability to a system or application. [Machine Learning](https://write.geeksforgeeks.org/geek/ml-what-is-machine-learning-2/), as the name suggests helps the system to develop itself based on some patterns, user/environmental interaction or on the basis of algorithms.

It provides various libraries or functionalities such as collaborative filtering, clustering, and classification which are nothing but concepts of Machine learning. It allows invoking algorithms as per our need with the help of its own libraries.

Apache Spark:   
It’s a platform that handles all the process consumptive tasks like batch processing, interactive or iterative real-time processing, graph conversions, and visualization, etc.

It consumes in memory resources hence, thus being faster than the prior in terms of optimization.

Spark is best suited for real-time data whereas Hadoop is best suited for structured data or batch processing, hence both are used in most of the companies interchangeably.

Apache HBase: 

It’s a NoSQL database which supports all kinds of data and thus capable of handling anything of Hadoop Database. It provides capabilities of Google’s BigTable, thus able to work on Big Data sets effectively.

At times where we need to search or retrieve the occurrences of something small in a huge database, the request must be processed within a short quick span of time. At such times, HBase comes handy as it gives us a tolerant way of storing limited data

Other Components: Apart from all of these, there are some other components too that carry out a huge task in order to make Hadoop capable of processing large datasets. They are as follows:

Solr, Lucene: These are the two services that perform the task of searching and indexing with the help of some java libraries, especially Lucene is based on Java which allows spell check mechanism, as well. However, Lucene is driven by Solr.

Zookeeper: There was a huge issue of management of coordination and synchronization among the resources or the components of Hadoop which resulted in inconsistency, often. Zookeeper overcame all the problems by performing synchronization, inter-component based communication, grouping, and maintenance.

Oozie: Oozie simply performs the task of a scheduler, thus scheduling jobs and binding them together as a single unit. There is two kinds of jobs .i.e Oozie workflow and Oozie coordinator jobs. Oozie workflow is the jobs that need to be executed in a sequentially ordered manner whereas Oozie Coordinator jobs are those that are triggered when some data or external stimulus is given to it.

Hadoop Installation Cluster

[Hadoop](https://www.geeksforgeeks.org/hadoop-an-introduction/) Can be installed in two ways. The first is on a single node cluster and the second way is on a multiple node cluster.

Single Node Cluster – It Has one DataNode running and setting up all the NameNode, DataNode, Resource Manager, and NodeManager on a single machine. This is used for studying and testing purposes.

Multi-Node Cluster – Has more than one DataNode running and each DataNode is running on different machines.

Steps for Installing Single Node Cluster Hadoop on Windows as follows.

Prerequisite:

JAVA-Java JDK (Pre-Requiste)

HADOOP-Hadoop package (To be [Downloaded](https://hadoop.apache.org/releases.html))

Step1:- Extract Hadoop Directory in Windows C:\Hadoop

Step2:- Set up the HADOOP\_HOME Variable Path in the windows Environment.

Step3:- Set up the Hadoop bin directory path.

Step4:- Update 6 Hadoop Configuration files Mentioned Below

1. Core-site.xml

<configuration>

<property>

<name>fs.defaultFS</name>

<value>hdfs://localhost:9000</value>

</property>

</configuration>

1. Mapred-site.xml

<configuration>

<property>

<name>mapreduce.framework.name</name>

<value>yarn</value>

</property>

</configuration>

1. Hdfs-site.xml

<configuration>

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

<name>dfs.namenode.name.dir</name>

<value>C:\hadoop-2.8.0\data\namenode</value>

</property>

<property>

<name>dfs.datanode.data.dir</name>

<value>C:\hadoop-2.8.0\data\datanode</value>

</property>

</configuration>

1. Yarn-site.xml

<configuration>

<property>

<name>yarn.nodemanager.aux-services</name>

<value>mapreduce\_shuffle</value>

</property>

<property>

<name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>

<value>org.apache.hadoop.mapred.ShuffleHandler</value>

</property>

</configuration>

1. Hadoop-env.cmd

Set "JAVA\_HOME=C:\Java" (On C:\java this is path to file jdk.18.0)

6. Create two folders datanode and namenode

1. Create folder "data" under "C:\Hadoop-2.8.0"

2. Create folder "datanode" under "C:\Hadoop-2.8.0\data"

3. Create folder "namenode" under "C:\Hadoop-2.8.0\data"

Step 5: -Format the namenode folder

Open command window (cmd) and typing commend “hdfs namenode –format”

Step 6:- Test the setup

Open command window (cmd) and typing command “start-all.cmd”

1. Ensure that namenode, datanode, and Resource manager are running
2. Open: http://localhost:8088