Unit I

What is Big Data

Big data is a term that describes large, hard-to-manage volumes of data – both structured and unstructured – that inundate businesses on a day-to-day basis. Big data can be analyzed for insights that improve decisions and give confidence for making strategic business moves.

What is Big Data

“Big data” is high-volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.”

This definition clearly answers the “What is Big Data?” question – Big Data refers to complex and large data sets that have to be processed and analyzed to uncover valuable information that can benefit businesses and organizations.

However, there are certain basic tenets of Big Data that will make it even simpler to answer what is Big Data:

- It refers to a massive amount of data that keeps on growing exponentially with time.
- It is so voluminous that it cannot be processed or analyzed using conventional data processing techniques.
- It includes data mining, data storage, data analysis, data sharing, and data visualization.
- The term is an all-comprehensive one including data, data frameworks, along with the tools and techniques used to process and analyze the data.

BIG DATA EXAMPLES

- Personalized e-commerce shopping experiences
- Financial market modelling
- Compiling trillions of data points to speed up cancer research
- Media recommendations from streaming services like Spotify, Hulu and Netflix
- Predicting crop yields for farmers
- Analyzing traffic patterns to lessen congestion in cities
- Data tools recognizing retail shopping habits and optimal product placement
- Big data helping sports teams maximize their efficiency and value
- Recognizing trends in education habits from individual students, schools and districts.

What Comes Under Big Data?

Big data involves the data produced by different devices and applications. Given below are some of the fields that come under the umbrella of Big Data.

- **Black Box Data** − It is a component of helicopter, airplanes, and jets, etc. It captures voices of the flight crew, recordings of microphones and earphones, and the performance information of the aircraft.
- **Social Media Data** − Social media such as Facebook and Twitter hold information and the views posted by millions of people across the globe.
- **Stock Exchange Data** − The stock exchange data holds information about the ‘buy’ and ‘sell’ decisions made on a share of different companies made by the customers.

- **Power Grid Data** − the power grid data holds information consumed by a particular node with respect to a base station.
- **Transport Data** − Transport data includes model, capacity, distance and availability of a vehicle.
- **Search Engine Data** − Search engines retrieve lots of data from different databases.

Thus Big Data includes huge volume, high velocity, and extensible variety of data. The data in it will be of three types.

- **Structured data** − Relational data.
- **Semi Structured data** − XML data.
- **Unstructured data** − Word, PDF, Text, Media Logs.

Types of Big Data

Now that we are on track with what is big data, let's have a look at the types of big data:

Structured

Structured is one of the types of big data and By structured data, we mean data that can be processed, stored, and retrieved in a fixed format. It refers to highly organized information that can be readily and seamlessly stored and accessed from a database by simple search engine algorithms. For instance, the employee table in a company database will be structured as the employee details, their job positions, their salaries, etc., will be present in an organized manner.

Unstructured

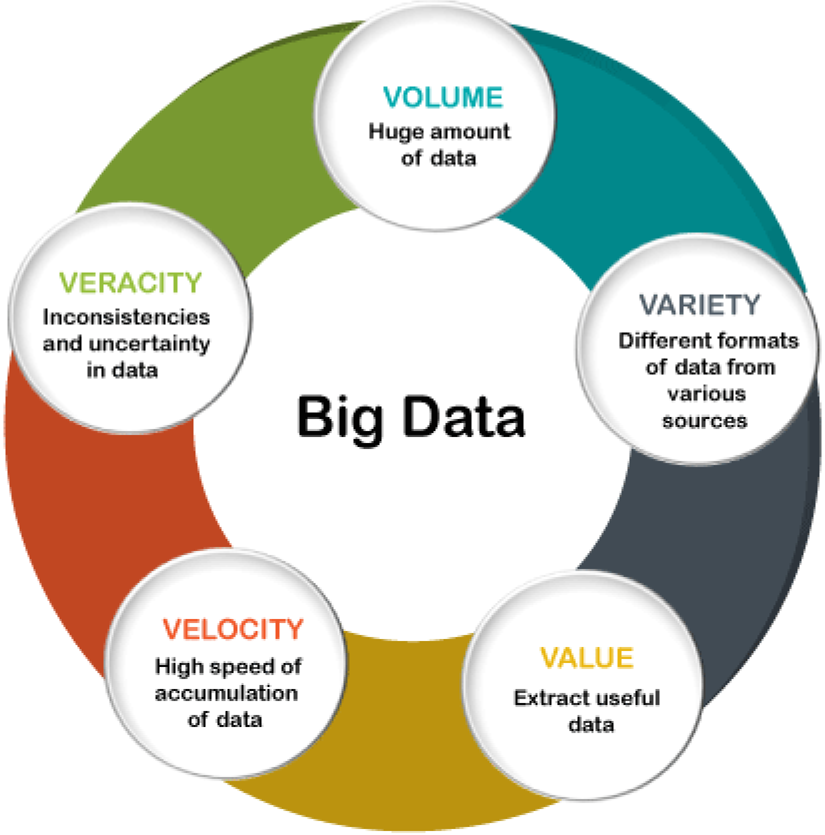
Unstructured data refers to the data that lacks any specific form or structure whatsoever. This makes it very difficult and time-consuming to process and analyze unstructured data. Email is an example of unstructured data. Structured and unstructured are two important types of big data.

Semi-structured

Semi structured is the third type of big data. Semi-structured data pertains to the data containing both the formats mentioned above, that is, structured and unstructured data. To be precise, it refers to the data that although has not been classified under a particular repository (database), yet contains vital information or tags that segregate individual elements within the data.

History of Big Data

Big data refers to data that is so large, fast or complex that it's difficult or impossible to process using traditional methods. The act of accessing and storing large amounts of information for analytics has been around for a long time. But the concept of big data gained momentum in the early 2000s when industry analyst Doug Laney articulated the now-mainstream definition of big data as the three V's:


VOLUME
Huge amount of data
VARIETY
Different formats of data from various sources
VALUE
Extract useful data
VELOCITY
High speed of accumulation of data
VERACITY
Inconsistencies and uncertainty in data
Big Data

Volume. Organizations collect data from a variety of sources, including transactions, smart (IoT) devices, industrial equipment, videos, images, audio, social media and more. In the past, storing all that data would have been too costly – but cheaper storage using data lakes, Hadoop and the cloud have eased the burden.

Velocity. With the growth in the Internet of Things, data streams into businesses at an unprecedented speed and must be handled in a timely manner. RFID tags, sensors and smart meters are driving the need to deal with these torrents of data in near-real time.

Variety. Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, emails, videos, audios, stock ticker data and financial transactions.

At SAS, we consider two additional dimensions when it comes to big data:

Variability

In addition to the increasing velocities and varieties of data, data flows are unpredictable – changing often and varying greatly. It's challenging, but businesses need to know when something is trending in social media, and how to manage daily, seasonal and event-triggered peak data loads.

Veracity

Veracity refers to the quality of data. Because data comes from so many different sources, it's difficult to link, match, cleanse and transform data across systems. Businesses need to connect and correlate relationships, hierarchies and multiple data linkages. Otherwise, their data can quickly spiral out of control. [https://www.sas.com/en_in/insights/big-data/what-is-big-data.html]

Benefits of Big Data

Big Data can help create pioneering breakthroughs for organizations that know how to use it correctly. Big Data solutions and Big Data Analytics can not only foster data-driven decision making, but they also empower your workforce in ways that add value to your business.

The benefits of Big Data Analytics and tools are –

- Data accumulation from multiple sources, including the Internet, social media platforms, online shopping sites, company databases, external third-party sources, etc.
- Real-time forecasting and monitoring of business as well as the market.
- Identify crucial points hidden within large datasets to influence business decisions.
- Promptly mitigate risks by optimizing complex decisions for unforeseen events and potential threats.
- Identify issues in systems and business processes in real-time.
- Unlock the true potential of data-driven marketing.
- Dig in customer data to create tailor-made products, services, offers, discounts, etc.
- Facilitate speedy delivery of products/services that meet and exceed client expectations.
- Diversify revenue streams to boost company profits and ROI.
- Respond to customer requests, grievances, and queries in real-time.
- Foster innovation of new business strategies, products, and services.

Now, we will expand on the most significant advantages of Big Data:

1. Cost optimization

One of the most significant benefits of Big Data tools like Hadoop and Spark is that these offer cost advantages to businesses when it comes to storing, processing, and analyzing large amounts of data. Not just that, Big Data tools can also identify efficient and cost-savvy ways of doing business.

The logistics industry presents an excellent example to highlight the cost-reduction benefit of Big Data. Usually, the cost of product returns is 1.5 times greater that of actual shipping costs. Big Data Analytics allows companies to minimize product return costs by predicting the likelihood of product returns. They can estimate which products are most likely to be returned, thereby allowing companies to take suitable measures to reduce losses on returns.

2. **Improve efficiency**

   Big Data tools can improve operational efficiency by leaps and bounds. By interacting with customers/clients and gaining their valuable feedback, Big Data tools can amass large amounts of useful customer data. This data can then be analyzed and interpreted to extract meaningful patterns hidden within (customer taste and preferences, pain points, buying behavior, etc.), which allows companies to create personalized products/services.

   Big Data Analytics can identify and analyze the latest market trends, allowing you to keep pace with your competitors in the market. Another benefit of Big Data tools is that they can automate routine processes and tasks. This frees up the valuable time of human employees, which they can devote to tasks that require cognitive skills.

3. **Foster competitive pricing**

   Big Data Analytics facilitates real-time monitoring of the market and your competitors. You can not only keep track of the past actions of your competitors but also see what strategies they are adopting now. Big Data Analytics offers real-time insights that allow you to –

- Calculate and measure the impact of price changes.
- Implement competitive positioning for maximizing company profits.
- Evaluate finances to get a clearer idea of the financial position of your business.
- Implement pricing strategies based on local customer demands, customer purchasing behavior, and competitive market patterns.
- Automate the pricing process of your business to maintain price consistency and eliminate manual errors.

4. **Boost sales and retain customer loyalty**

   Big Data aims to gather and analyze vast volumes of customer data. The digital footprints that customers leave behind reveal a great deal about their preferences, needs, buying behavior, and much more. This customer data offers the scope to design tailor-made products and services to cater to the specific needs of individual customer segments. The higher the personalization quotient of a business, the more it will attract customers. Naturally, this will boost sales considerably.

   Personalization and the quality of product/service also have a positive impact on customer loyalty. If you offer quality products at competitive prices along with personalized features/discounts, customers will keep coming back to you time and again.

5. **Innovate**

   Big Data Analytics and tools can dig into vast datasets to extract valuable insights, which can be transformed into actionable business strategies and decisions. These insights are the key to innovation. The insights you gain can be used to tweak business strategies, develop new products/services (that can address specific problems of customers), improve marketing techniques, optimize customer service, improve employee productivity, and find radical ways to expand brand outreach.

6. **Focus on the local environment**

   This is particularly relevant for small businesses that cater to the local market and its customers. Even if your business functions within a constrained setting, it is essential to understand your competitors, what they are offering, and the customers.

   Big Data tools can scan and analyze the local market and offer insights that allow you to see the local trends associated with sellers and customers. Consequently, you can leverage such insights to gain a competitive edge in the local market by delivering highly personalized products/services within your niche, local environment.

Big Data Technologies Operational vs Analytical Systems

Operational and Analytical Data Systems are both very similar in how they provide information on your organization, company, or non-profit, but the two are very structurally different, and provide different types of insights.

Operational Data Systems

Operational Data is exactly what it sounds like - data that is **produced by your organization's day to day operations**. Things like customer, inventory, and purchase data fall into this category. This type of data is pretty straightforward and will generally look the same for most organizations. If you want to know the most up to date information on something - you're using Operational Data! **Operational Data Systems** support high-volume low-latency access, called **Online Transactional Processing** tables, or OLTP, where you want to create, read, update, or delete one piece of data at a time.

Analytical Data Systems

Analytical Data is a little more complex and will look different for different types of organizations; however, at it's core is an organization's **Operational Data**. Analytical Data is **used to make business decisions**, as opposed to recording the data from actual operational business processes. Examples include grouping customers for market segmentation or changes in purchase volume over time. Every organization will have different questions to answer and different decisions to make, so Analytical Data is definitely not one-size-fits-all by any stretch of the imagination! Analytical Data is best stored in a Data System designed for heavy aggregation, data mining, and ad hoc queries, called an **Online Analytical Processing** system, OLAP, or a Data Warehouse!

Operational Data Systems, consisting largely of transactional data, are built for quicker updates. **Analytical Data Systems**, which are intended for decision making, are built for more efficient analysis. Hopefully now you have a better understanding of the difference between Operational and Analytical Data and their corresponding Data Systems! As you can see, both are very important for maintaining and growing an organization, business, or non-profit.

Big Data Challenges

In a broad range of application areas, data is being collected at a unique scale. Decisions that previously were based on guesswork, or on the basis of reality, at present now decision to be made using data-driven mathematical models. Such Big Data analysis now drives nearly every aspect of society, including mobile services, retail, manufacturing, financial services, life sciences, and physical sciences.

Timeliness and heterogeneity

When there is a lot of information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogeneous data, and cannot understand nuance. In consequence, data must be carefully structured prior to data analysis. Even after data cleaning and error correction, some incompleteness and some errors in data are likely to remain. This incompleteness and these errors must be managed during data analysis. Doing this correctly is a challenge.

Scalability

Scalability is the major challenge with the big data. You want to be able to scale very rapidly and elastically, whenever and wherever you want. There is a need of effective solution to enable the cost-effective, feasible, scalable storage and processing of large volume of data. Most NoSQL solutions like MongoDB or HBase have their own scaling limitations.

Performance

In the world of internet big data must move at extremely high velocities no matter how much you scale or what workloads your database must perform. You need that big data analysis is performed within time constraints as require. The data handling hoops of RDBMS and most NoSQL solutions put a serious drag on performance.

Continuous Availability

When you rely on big data to feed your essential, revenue-generating 24/7 business applications, even high availability is not high enough. Your data can never go down. The capabilities of existing system to process streaming information and answer queries in real-time and for thousands of concurrent users are limited. People expect real-time or near real-time responses from the systems they interact with.

Workload Diversity

Big data comes in all shapes, colors and sizes. Rigid schemas have no place here; instead you need a more flexible design. You want your technology to fit your data, not the other way around. And you want to be able to do more with all of that data – perform transactions in real-time, run analytics just as fast and find anything you want in an instant from oceans of data, no matter what from that data may take.

Data Security

Security is the big concern with the big data. As larger amount of data is processed and transfer among the organizational boundaries and this big data carries some big risks when it contains credit card data, personal ID information and other sensitive assets. Now the challenge is how to protect this sensitive data and how to keep private. Most NoSQL big data platforms have few if any security mechanisms in place to safeguard your big data. Security concerns about data protection are a major obstacle preventing companies from taking full advantage of their data.

Identifying Right Data

Identifying the right data from the vast amount of data is the big challenge. Since there are large number of sources such as social networking sites, blogs, different types of content such as articles, comments. Companies have difficulty identifying the right data and determining how to best use it. Therefore, there is the need to find out the rules that will help in identifying the right data Building data-related business cases often means thinking outside of the box and looking for revenue models that are very different from the traditional business.

Identifying right talent

Companies are struggling to find the right talent capable of both working with new technologies and of interpreting the data to find meaningful business insights

Identifying right Platform

Data access and connectivity can be an obstacle. A majority of data points are not yet connected today, and companies often do not have the right platforms to aggregate and manage the data across the enterprise. In order to address the growing volume of data created as a part of power grid operation

Identify right architecture

The technology landscape in the data world is evolving extremely fast. Leveraging data means working with a strong and innovative technology partner that can help create the right IT architecture that can adapt to changes in the landscape in an efficient manner.

Collaborating across functions and businesses.

Leveraging big data often means working across functions like IT, engineering, finance and procurement and the ownership of data is fragmented across the organization. To address these organizational challenges means finding new ways of collaborating across functions and businesses.
According to SAS following challenges are outlined in terms of data visualization.

Meeting the need for speed

In today’s hypercompetitive business environment, companies not only have to find and analyze the relevant data they need, they must find it quickly. Visual-ization helps organizations perform analyses and make decisions much more rapidly, but the challenge is going through the sheer volumes of data and accessing the level of detail needed, all at a high speed. The challenge only grows as the degree of granularity increases. One possible solution is hardware. Some vendors are using increased memory and powerful parallel processing to crunch large volumes of data extremely quickly. Another method is putting data in-memory but using a grid computing approach, where many machines are used to solve a problem. Both approaches allow organizations to explore huge data volumes and gain business insights in near-real time.

Understanding the data

It takes a lot of understanding to get data in the right shape so that you can use visualization as part of data analysis. For example, if the data comes from social media content, you need to know who the user is in a general sense – such as a customer using a particular set of products – and understand what it is you’re trying to visualize out of the data. Without some sort of context, visualization tools are likely to be of less value to the user.

One solution to this challenge is to have the proper domain expertise in place. Make sure the people analyzing the data have a deep understanding of where the data comes from, what audience will be consuming the data and how that audience will interpret the information.

Addressing Data Quality

Even if you can find and analyze data quickly and put it in the proper context for the audience that will be consuming the information, the value of data for decision-making purposes will be jeopardized if the data is not accurate or timely. This is a challenge with any data analysis, but when considering the volumes of information involved in big data projects, it becomes even more pronounced. Again, data visualization will only prove to be a valuable tool if the data quality is assured. To address this issue, companies need to have a data governance or information management process in place to ensure the data is clean. It’s always best to have a pro-active method to address data quality issues so problems won’t arise later.

Displaying Meaningful Results

Plotting points on a graph for analysis becomes difficult when dealing with extremely large amounts of information or a variety of categories of information. For example, imagine you have 10 billion rows of retail SKU data that you’re trying to compare. The user trying to view 10 billion plots on the screen will have a hard time seeing so many data points. One way to resolve this is to cluster data into a higher-level view where smaller groups of data become visible. By grouping the data together, or “binning,” you can more effectively visualize the data manager.

UNIT - II

Introduction to HADOOP

Apache Hadoop is a collection of open-source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model.

Hadoop 1 vs Hadoop 2

1**. Components:** In Hadoop 1 we have MapReduce but Hadoop 2 has YARN(Yet Another Resource Negotiator) and MapReduce version 2.

Hadoop 1	**Hadoop 2**
HDFS	HDFS
Map Reduce	YARN / MRv2

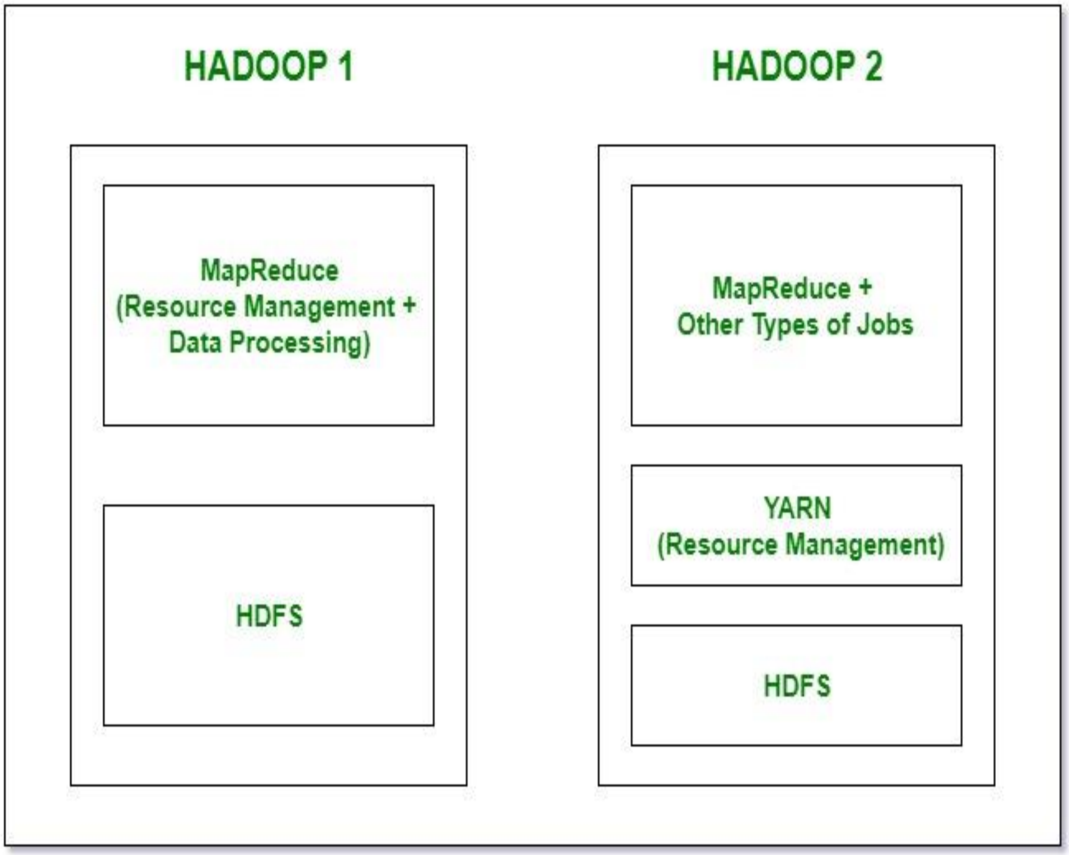
2**. Daemons:**

Hadoop 1	**Hadoop 2**
Namenode	Namenode
Datanode	Datanode
Secondary Namenode	Secondary Namenode
Job Tracker	Resource Manager
Task Tracker	Node Manager

3. Working:

In Hadoop 1, there is HDFS which is used for storage and top of it, Map Reduce which works as Resource Management as well as Data Processing. Due to this workload on Map Reduce, it will affect the performance.

In Hadoop 2, there is again HDFS which is again used for storage and on the top of HDFS, there is YARN which works as Resource Management. It basically allocates the resources and keeps all the things going on.

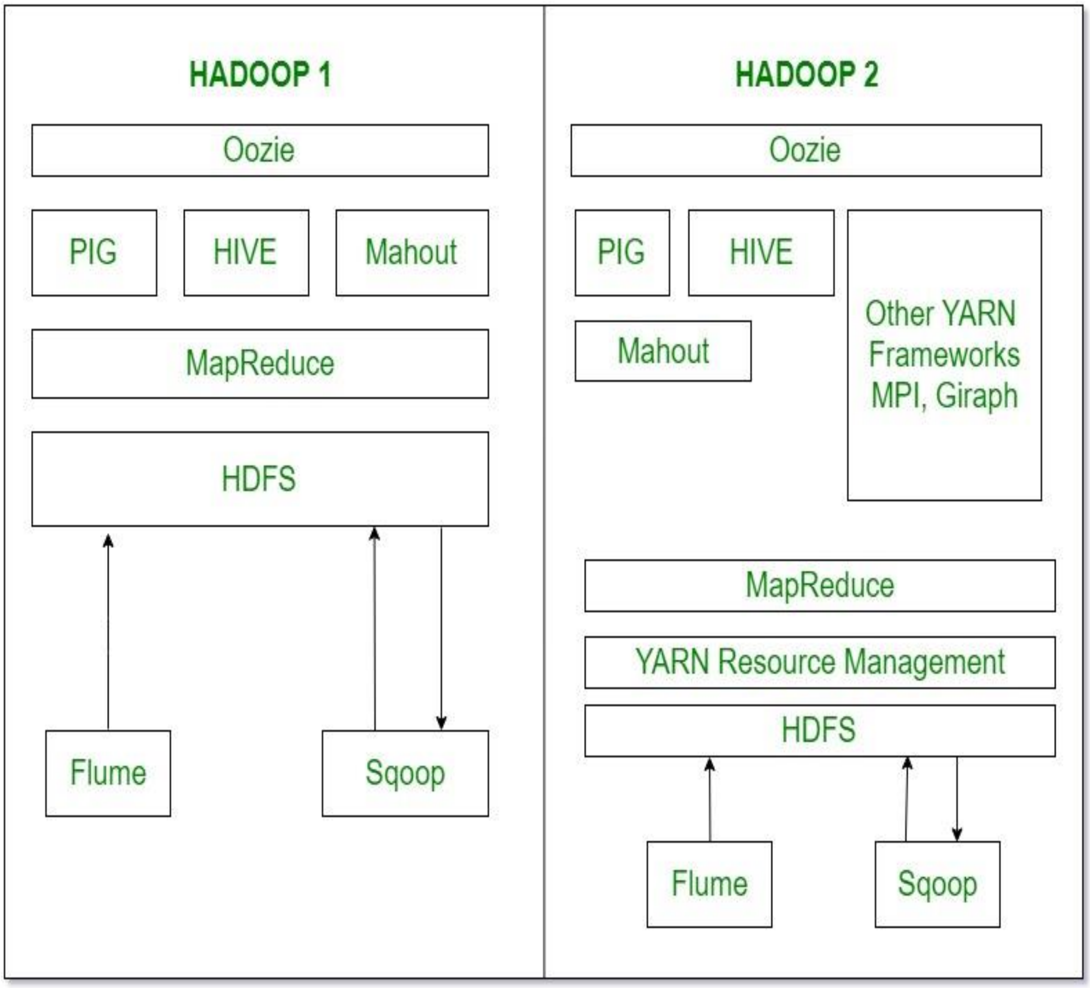


4. Limitations:

Hadoop 1 is a Master-Slave architecture. It consists of a single master and multiple slaves. Suppose if master node got crashed then irrespective of your best slave nodes, your cluster will be destroyed. Again for creating that cluster means copying system files, image files, etc. on another system is too much time consuming which will not be tolerated by organizations in today’s time.

Hadoop 2 is also a Master-Slave architecture. But this consists of multiple masters (i.e active namenodes and standby namenodes) and multiple slaves. If here master node got crashed then standby master node will take over it. You can make multiple combinations of active-standby nodes. Thus Hadoop 2 will eliminate the problem of a single point of failure.

5. Ecosystem



- Oozie is basically Work Flow Scheduler. It decides the particular time of jobs to execute according to their dependency.
- Pig, Hive and Mahout are data processing tools that are working on the top of Hadoop.
- Sqoop is used to import and export structured data. You can directly import and export the data into HDFS using SQL database.
- Flume is used to import and export the unstructured data and streaming data.

Hadoop Architecture

Introduction:

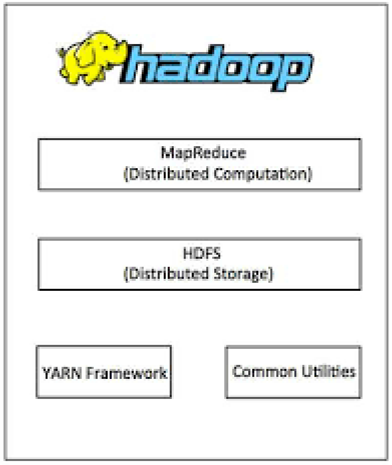
The Hadoop Distributed File System (HDFS) is the primary data storage system used by Hadoop applications. HDFS employs a NameNode and DataNode architecture to implement a distributed file system that provides high-performance access to data across highly scalable Hadoop clusters.

Hadoop itself is an open source distributed processing framework that manages data processing and storage for big data applications. HDFS is a key part of the many Hadoop ecosystem technologies. It provides a reliable means for managing pools of big data and supporting related big data analytics applications.

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. The Hadoop framework application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage. Hadoop Architecture At its core, Hadoop has two major layers namely –

- Processing/Computation layer (MapReduce), and

- Storage layer (Hadoop Distributed File System)



MapReduce MapReduce is a parallel programming model for writing distributed applications devised at Google for efficient processing of large amounts of data (multi-terabyte data-sets), on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. The MapReduce program runs on Hadoop which is an Apache open-source framework. Hadoop Distributed File System The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. It is highly fault-tolerant and is designed to be deployed on low-cost hardware. It provides high throughput access to application data and is suitable for applications having large datasets. Apart from the above-mentioned two core components, Hadoop framework also includes the following two modules –

- Hadoop Common − These are Java libraries and utilities required by other Hadoop modules.

- Hadoop YARN − This is a framework for job scheduling and cluster resource management.

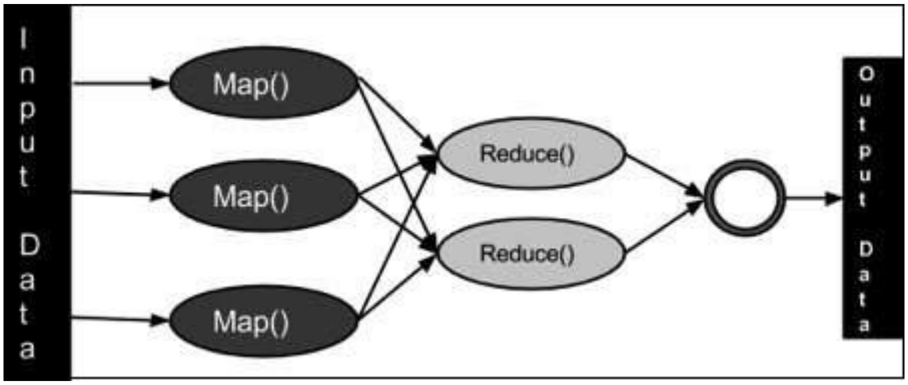
What is MapReduce?

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.

The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into *mappers* and *reducers* is sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the MapReduce model.

The Algorithm

- Generally MapReduce paradigm is based on sending the computer to where the data resides!
- MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.
    - **Map stage** − The map or mapper's job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.
    - **Reduce stage** − This stage is the combination of the **Shuffle** stage and the **Reduce** stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.
- During a MapReduce job, Hadoop sends the Map and Reduce tasks to the appropriate servers in the cluster.
- The framework manages all the details of data-passing such as issuing tasks, verifying task completion, and copying data around the cluster between the nodes.
- Most of the computing takes place on nodes with data on local disks that reduces the network traffic.
- After completion of the given tasks, the cluster collects and reduces the data to form an appropriate result, and sends it back to the Hadoop server.



Hadoop Distributed File system

HDFS is a distributed file system that handles large data sets running on commodity hardware. It is used to scale a single Apache Hadoop cluster to hundreds (and even thousands) of nodes. HDFS is one of the major components of Apache Hadoop, the others being MapReduce and YARN. HDFS should not be confused with or replaced by Apache HBase, which is a column-oriented non-relational database management system that sits on top of HDFS and can better support real-time data needs with its in-memory processing engine.

The goals of HDFS

Fast recovery from hardware failures

Because one HDFS instance may consist of thousands of servers, failure of at least one server is inevitable. HDFS has been built to detect faults and automatically recover quickly.

Access to streaming data

HDFS is intended more for batch processing versus interactive use, so the emphasis in the design is for high data throughput rates, which accommodate streaming access to data sets.

Accommodation of large data sets

HDFS accommodates applications that have data sets typically gigabytes to terabytes in size. HDFS provides high aggregate data bandwidth and can scale to hundreds of nodes in a single cluster.

Portability

To facilitate adoption, HDFS is designed to be portable across multiple hardware platforms and to be compatible with a variety of underlying operating systems.

An example of HDFS

Consider a file that includes the phone numbers for everyone in the United States; the numbers for people with a last name starting with A might be stored on server 1, B on server 2, and so on.

With Hadoop, pieces of this phonebook would be stored across the cluster, and to reconstruct the entire phonebook, your program would need the blocks from every server in the cluster.

To ensure availability if and when a server fails, HDFS replicates these smaller pieces onto two additional servers by default. (The redundancy can be increased or decreased on a per-file basis or for a whole environment; for example, a development Hadoop cluster typically doesn’t need any data redundancy.) This redundancy offers multiple benefits, the most obvious being higher availability.

The redundancy also allows the Hadoop cluster to break up work into smaller chunks and run those jobs on all the servers in the cluster for better scalability. Finally, you gain the benefit of data locality, which is critical when working with large data sets.

How Does Hadoop Work?

It is quite expensive to build bigger servers with heavy configurations that handle large scale processing, but as an alternative, you can tie together many commodity computers with single-CPU, as a single functional distributed system and practically, the clustered machines can read the dataset in parallel and provide a much higher throughput. Moreover, it is cheaper than one high-end server. So this is the first motivational factor behind using Hadoop that it runs across clustered and low-cost machines. Hadoop runs code across a cluster of computers. This process includes the following core tasks that Hadoop performs:

- Data is initially divided into directories and files. Files are divided into uniform sized blocks of 128M and 64M (preferably 128M). • These files are then distributed across various cluster nodes for further processing. • HDFS, being on top of the local file system, supervises the processing.
- Blocks are replicated for handling hardware failure.
- Checking that the code was executed successfully.
- Performing the sort that takes place between the map and reduce stages.
  Sending the sorted data to a certain computer.
- Writing the debugging logs for each job.

Advantages of Hadoop

- Hadoop framework allows the user to quickly write and test distributed systems. It is efficient, and it automatic distributes the data and work across the machines and in turn, utilizes the underlying parallelism of the CPU cores.
- Hadoop does not rely on hardware to provide fault-tolerance and high availability (FTHA), rather Hadoop library itself has been designed to detect and handle failures at the application layer.
- Servers can be added or removed from the cluster dynamically and Hadoop continues to operate without interruption.
- Another big advantage of Hadoop is that apart from being open source, it is compatible on all the platforms since it is Java based.