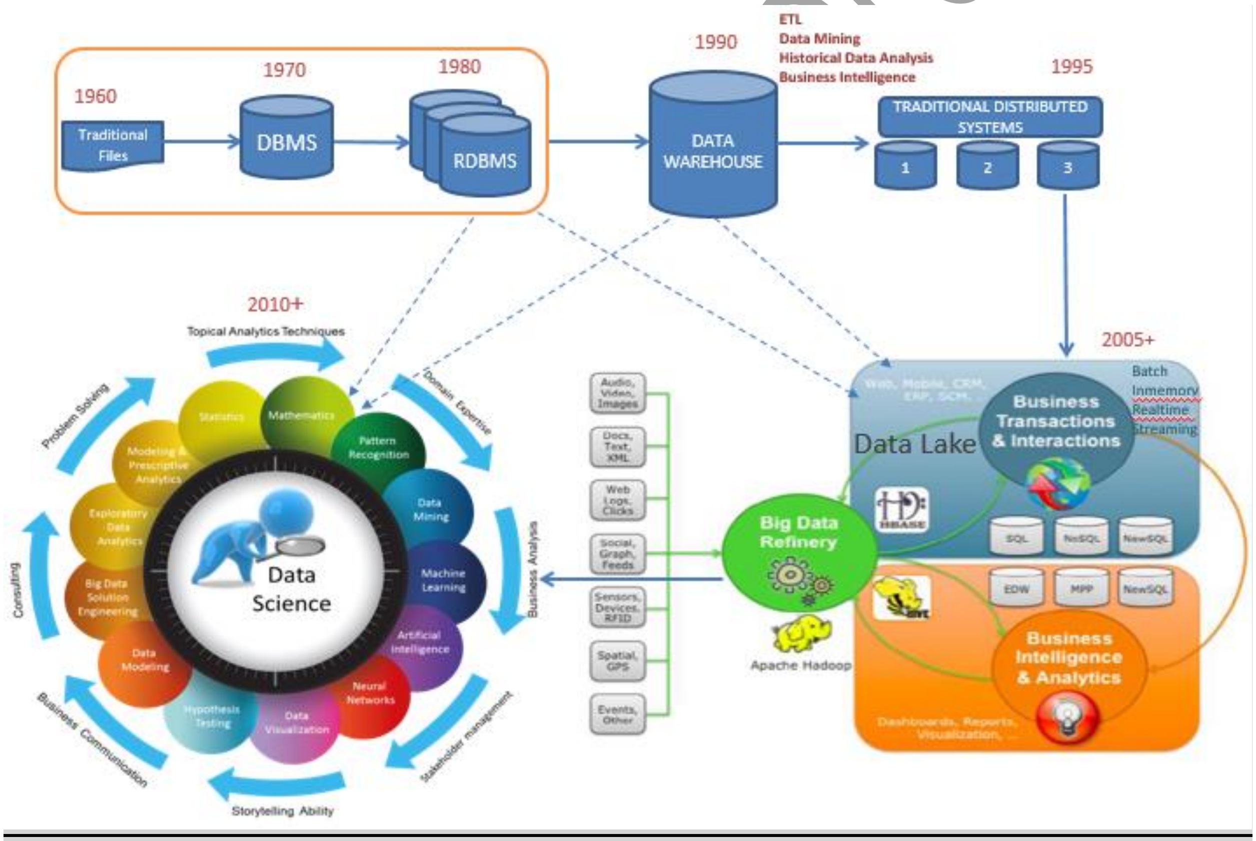
Web: Inceptez.com Mail: info@inceptez.com Call: 7871299810, 7871299817

BigData & Hadoop

Evolution of Data

Application of computation and programming started to sort out the data capture and processing challenges from several decades ago. Let’s see how data management, realization and utilization evolved to produce the value to the systems in the name of ETL/ELT, Data Wrangling, Data Engineering, BigData, Data mining, data Analysis, Analytics, Data Science, Machine learning, AI, Neural networks, Visualization, Dashboarding etc. .



Year 1960+ Free form data captured in the form of scratch files or flat files which stores raw unstructured data collected to store transactions as stories.

Year 1970+ Database Management System (DBMS) is system software for creating and managing databases with wide dataset in rows and column format. The DBMS provides users and programmers with a systematic way to create, retrieve, update and manage data. It is the way of storing data in a denormalized fashion as a fat tables, but we had lot of challenges in terms of maintaining duplicates, not scalable to more columns, non relational, costly to update and maintain.

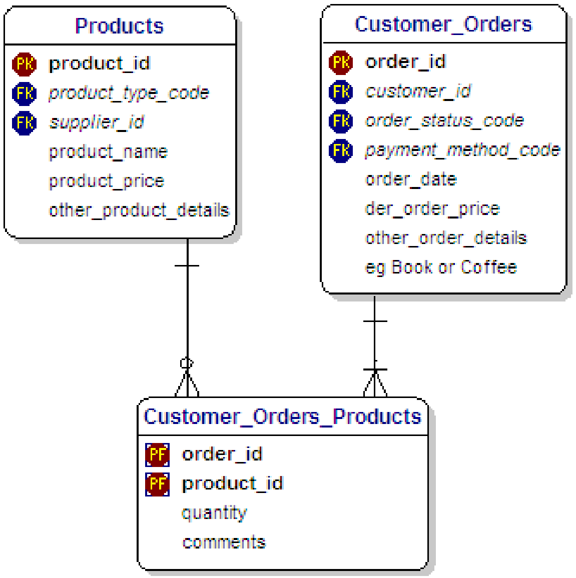
Custname	address	product	product desc	quantity	Amount
Inceptez	27, A, Brahmin street, Velachery	laptop	dell latitude series	2	80000
Inceptez	27, A, Brahmin street, Velachery	table	computer table	2	6000
ABC Infotech	3, bharathi nagar, chrompet	mobile phone	samsung s3	1	20000
Inceptez	27, A, Brahmin street, Velachery	table	Chair	10	12000

Year 1980+ Relational Database Management System (RDBMS) is system software for creating and managing relational databases comprise of tables that can be related with integrity. The RDBMS provides users and programmers with a systematic way to create, retrieve, update and manage relational data by normalizing and relating the tables. Several challenges in terms of maintaining standalone diversified databases for each and every systems, not cost effective, cant handle huge volume of data etc. RDBMS can be used for generating daily or weekly reports and not for generating integrated metrics or historical reports.

custid	custname	address
1	inceptez	27, A, Brahmin street, Velachery
2	ABC Infotech	3, bharathi nagar, chrompet

prodid	product	product desc
1	laptop	dell latitude series
2	table	computer table
3	mobile phone	samsung s3

transid	Prodid	custid	quantity	amount
1	1	1	2	80000
2	2	1	2	6000
3	3	2	1	20000
4	2	1	5	15000

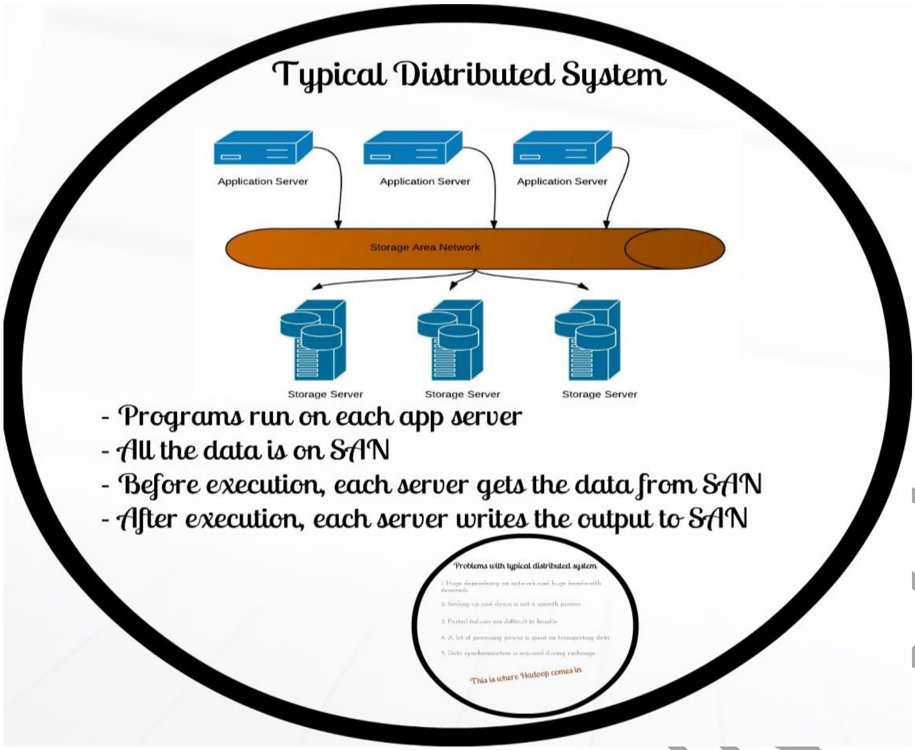


1990+ Datawarehouse - Enterprise level Datawarehouses are formed to maintain a single point of truth for the enterprise wide data for data convergence which uses the model of subject oriented, integrated, non volatile, timevariant and available database, but due limited scalability, not cost effective, only handle structured data and cant used for predicting other than time series data. Datawarehouses started becoming obsolete due to these reasons. It also provides only opportunities for generating historical analysis reports, data mining and root cause analysis to an extend of business intelligence.

Traditional Analysis or Data Mining- An examination of historical data and facts to uncover and understand cause of an issue, thus providing basis for problem solving and decision making by providing reports or spreadsheets.

Year 1995+ Traditional Distributed Systems

1. Huge dependency on network and bandwidth need to port raw data from storage disk to storage to app server and store the processed data back to Storage, 2 way data transfer occurs.
2. Partial failure is not easy to handle.
3. Not easily scalable.
4. Maintaining coordination and synchronisation is not an easy task.



- Programs run on each app server
- All the data is on SAN
- Before execution, each server gets the data from SAN
- After execution, each server writes the output to SAN

Problems with typical distributed system

1. Huge dependency on network and huge bandwidth demands
2. Scaling up and down is not a smooth process
3. Partial failures are difficult to handle
4. A lot of processing power is spent on transporting data
5. Data synchronization is required during exchange

This is where Hadoop comes in

Year 2005+ - Bigdata

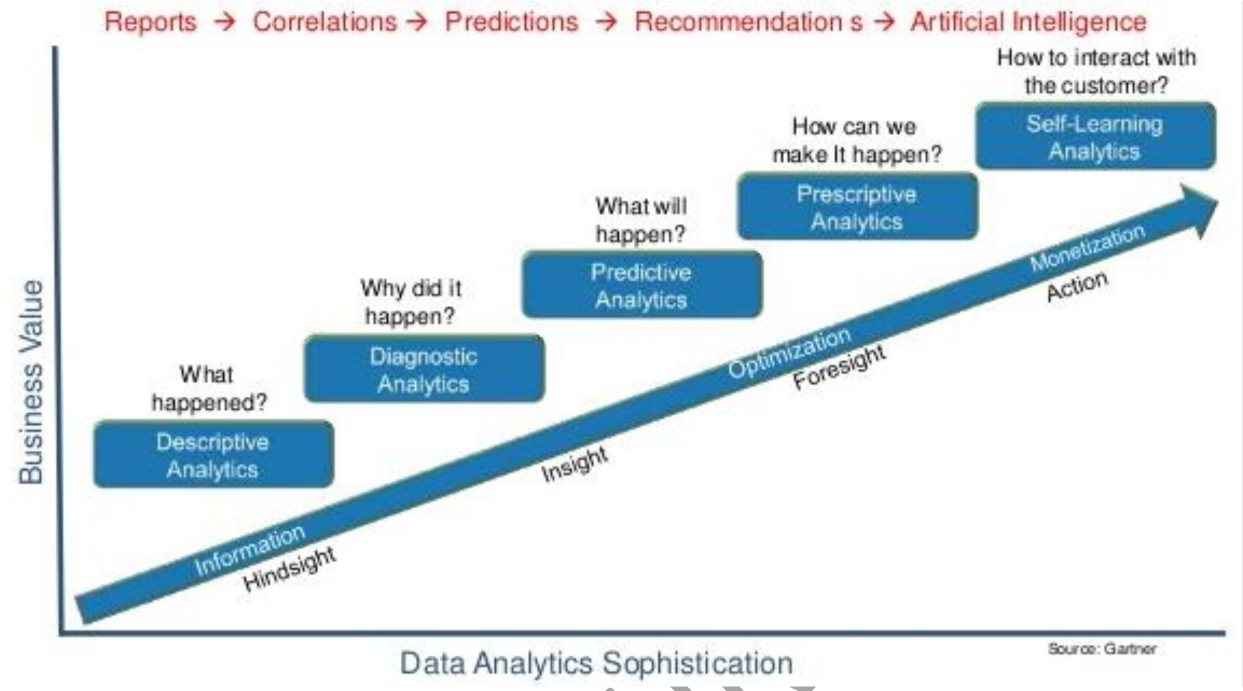
Big data is a phrase or methodology that describes the aquire, cure, process and store the humongous volume of data that is limited to handled by the traditional systems with in the stipulated period of time.

Big data is simply the large sets of data that businesses and other parties put together to serve specific goals and operations. Big data can include many different kinds of data in many different kinds of formats. As a rule, it’s raw and unsorted until it is put through various kinds of tools and handlers.

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time, extremely handles large data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions.

Data Analytics - Analytics is the process of deriving insights from patterns found in data (historical or live) to inform decision-making and improved outcomes.is an encompassing and multidimensional field that uses mathematics, statistics, predictive modeling and machine-learning techniques to find meaningful patterns and knowledge in recorded data.

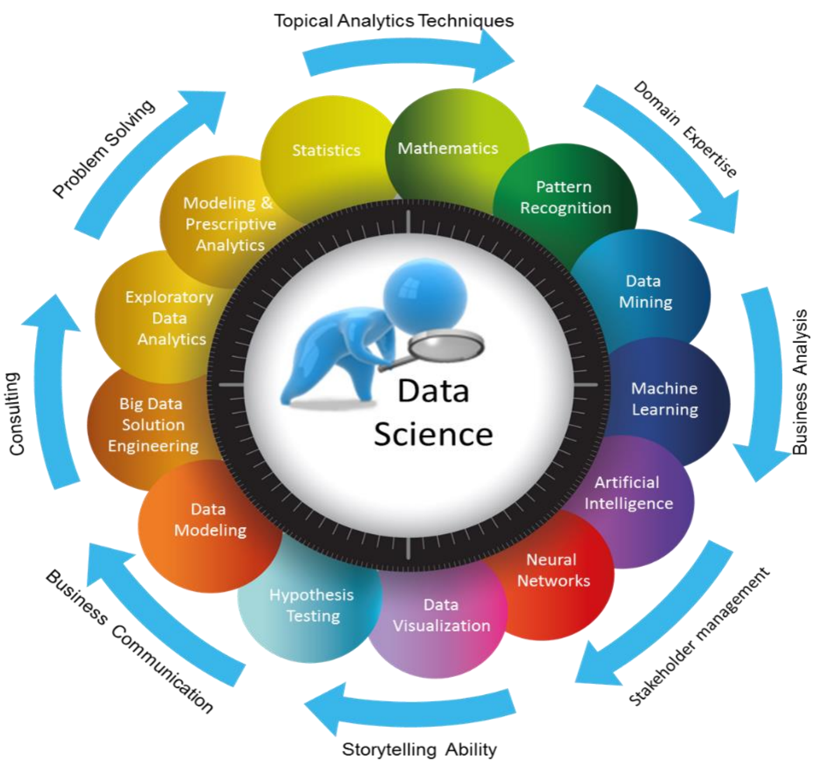
Evolution of Analytics



DataMining - Data mining is a traditional process of extracting larger set of any raw data and analysing data patterns using one or more software techniques to identify trends and root cause of the business problem. Techniques such as SQL queries, iterative processing, visualization, explanatory analysis etc are used to perform data mining.

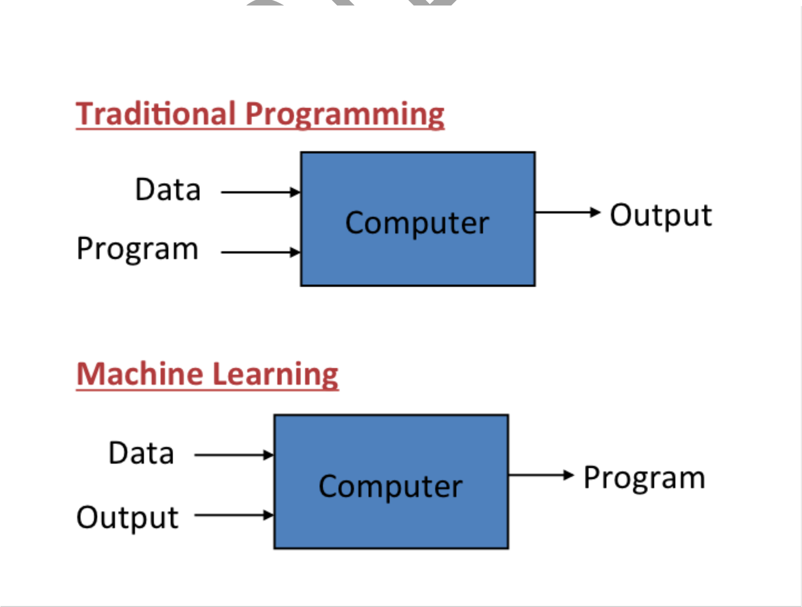
DataScience

Data science, is the advancement of Analytics which is also known as data-driven science, is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining using modern computing techniques including classification of data, clustering, applying algorithms, models, regression, augumented analytics, Deep learning, natural language processing etc,. that provides the complete disciplinary of applying analysis, analytics, learning of trends of data and mining of data using the power of science.



Machine Learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.



All about Bigdata

Types Of Big Data:

Broadly classified into man made and machine made data such as click stream event logs, device logs, photos, videos, historical datawarehouse data, geo satellite data etc. Shared nothing architecture.

Big data sourced from web data, social media, click stream data, man-made data, machine/device data etc.

Web log click stream data eg. Online banking transactions comparison between traditional and bigdata.

Man made data eg. Sentimental analysis through twitter, fb etc. Product campaign analysis.

Machine/device data such as geospacial data from gps devices such as Telematics, truck roll for efficient technian productivity measurement.

Why Bigdata draws market attention:

Every industry has its own particular big data challenges. Banks need to analyze streaming transactions in real time to quickly identify potential fraud. Utility companies need to analyse energy usage data to gain control over demand. Retailers need to understand the social sentiment around their products and markets to develop more effective campaigns and promotions.

Size hierarchy

GB Gigabyte

TB Terabyte (1024 GB)

PB Petabyte (1024 TB)

EB Exabyte (1024 PB)

ZB Zetabyte (1024 EB)

YB Yotabyte (1024 ZB)

BigData Statistics

Common data sets

- 40 Zetabytes of data exist in the digital universe today.
- Data is doubling every 2 years, 90% of world's data volume is created in past 2 years.

- Facebook stores, accesses, and analyzes 30+ Petabytes of user generated data.

- Walmart handles more than 1 million customer transactions every hour, which is imported into databases estimated to contain more than 2.5 petabytes of data.

- More than 5 billion people are calling, texting, tweeting and browsing on mobile phones worldwide.

The Rapid Growth of Unstructured Data

- YouTube users upload 48 hours of new video every minute of the day.

- 571 new websites are created every minute of the day.

Big Data & Real Business Issues

- According to estimates, the volume of business data worldwide, across all companies, doubles every 1.2 years.

- Poor data can cost businesses 20%–35% of their operating revenue.

- Bad data or poor data quality costs US businesses $600 billion annually.

- According to execs, the influx of data is putting a strain on IT infrastructure. 55 percent of respondents reporting a slowdown of IT systems and 47 percent citing data security problems, according to a global survey from Avanade.

- In that same survey, by a small but noticeable margin, executives at small companies (fewer than 1,000 employees) are nearly 10 percent more likely to view data as a strategic differentiator than their counterparts at large enterprises.

- Three-quarters of decision-makers (76 per cent) surveyed anticipate significant impacts in the domain of storage systems as a result of the “Big Data” phenomenon.

Bigdata is the center of attraction for all the technologies storage, processing and business value derivation needs. With the current/traditional systems industry was concentrating of implementing digital provisions and solutions for their business starting from application development programming to help clients interact with the business providing better UI experience, datawarehouse solution to convert, store, process and produce business indicators to promote the business, data storage solutions provided by db companies to store and retrieve RDBMS data efficiently. Now bigdata is in high demand due to all traditional systems are incapable of handling humongous volume of data being sent from different mediums of different format that has to be processed in a stipulated period of time and brings a real insight and business output from those data including the search of new business oppurtunities such as customer retention, scaling the business etc..

- **Bigdata & Analytics** is one of the top trending technology, Bigdata, Cloud computing, Enterprise mobility, DevOps, IOT.
- **Enterprise mobility -** The term refers not only to mobile workers and mobile devices, but also to the mobility of corporate data. An employee may upload a corporate presentation from his or her desktop PC to a cloud storage service, then access it from a personal iPad to show at a client site, for example. Enterprise mobility can improve employee productivity, but it also creates security risks.Enterprise mobility management products, such as data loss prevention technologies, are available to help IT departments address these risks. A strong acceptable use policy for employees can also contribute to a successful enterprise mobility strategy.
- **Internet of Things** : Monitor kids, old parents, motion sensor to track any unwanted movement around or inside house etc.
- **Cloud Computing**
- **Analytics :** The application of a sequence of steps (algorithms) or transformations to generate insights from processed datasets. It is an encompassing and multidimensional field that uses mathematics, statistics, predictive modeling and machine-learning techniques to find meaningful patterns and knowledge in recorded data.

Characteristics: 3vs Defined by IBM

- Volume – huge volume of machine and man made data such as logs, click events, media files and historical warehouse data.

  Eg . FB, Historical data (Bank example, Airlines example) etc.

- Velocity – Speed at which data is generated, captured, processed and stored.
  Eg. Tweets, click stream events, GPS Geospatial data etc.

- Variety – not only structured, it can accommodate un structured, semi structured, media, logs etc.
  Eg. STB logs, logs generated from different devices.

- Veracity – Trustworthy of the data we receive, how much true data we receive. Data accuracy is based on the veracity of source data that we receive at the very lower granular level. **Eg.** Website click steam logs.

- Value – ROI can be derived base on utilizing the data stored and bringing business insights from it.

Eg. Click stream logs ROI derived based on the interest shown by the customer in the webpage (positive and negative scores).

- Variable – data varies from time to time wrt size, type that is unpredictable.
  Eg. A company started with only deals (Snapdeal) then enter into retail marketing and business grown in all media.

Big data tools (other than hadoop):

Spark – In memory computing provided with Low latency SQL solution standalone and Hadoop, provides machine learning libraries and graph database solution in a single place.

MongoDB – Document oriented DB, stores JSON (Java script object notation data).

TD Aster – Teradata uses mapreduce for the MPP store and process.

Cloud MapReduce - Initially developed at Accenture Technology Labs, Cloud Mapreduce can be broadly defined as Mapreduce Implementation on Amazon Cloud OS. When compared with other open source implementation, it is completely different architecture than others.

Amazon EMR

Microsoft HD Insight

IBM Big Insight

What is hadoop:

Hadoop is a open source Apache software stack that runs on cluster of commodity hardware that provides distributed storage and processing of large data sets.

It is highly scalable, fault tolerant, data locality, highly distributed coordination, distributed MPP for heterogeneous data and heterogeneous software frameworks, low cost commodity hardware, high availability, highly secured, open source, resilient etc.

Evolution of Hadoop (spider to elephant)

Year	Evolution
1997	Doug Cutting working in lucene search engine.
	Doug Lucene to ASF and Mike Caferella 'web crawler' to index www , at that

2001	time 1.7 million web sites was there. Doug and Caferella joined and dubbed web crawler and named as NUTCH and tested in 1GB/1TB server, indexed about 100 webs /sec, and capable of maximum indexing of 100million. Scalability is an issue, tried with 4 nodes, but didn't achieved the expected performance.
2003	GFS white paper released by google, that exactly addressed the issues faced by Doug and Caferella.
2004	Google released another paper on MR for processing solutions. for processing solutions. Doug and Caferella made NDFS build based on GFS.
2005	Doug and Caferella wrote MR on top of NDFS.
2006	Doug made a subproject of lucene using MR and HDFS and named it as Hadoop – the name of his son's yellow elephant toy. Yahoo faced the same processing and scalability issue, employed Doug to implement hadoop in yahoo.
2007	Budding social media giants started their own projects on top of hadoop, such as Facebook released (hive, cassantra), linkedin – kafka, twitter – storm etc.
2008	Employees from google, facebook, yahoo and Berkleydb started 1st hadoop distribution called Cloudera.
2009	Yahoo sorts 1 PB in 16 hours using 3600 nodes cluster.
2011	Yahoo sponsored to its own hadoop company namely Hortonworks.
2012	Arun murthy and other open source community users proposed YARN.
2013	Hadoop 1.1.2 and Hadoop 2.0.3 alpha. Ambari, Cassandra, Mahout have been added

Hadoop - Features

Handle Huge volume of data

If the Data is too large to store and process in one computer, Hadoop can solve the problem by dividing the data and store it into cluster of several nodes and it is ready to process in a shorter period of time. Data is visible to the client in a unified way.Distributed parallel read and processing enhance performance

Highly Scalable:

Hadoop scales linearly. Due to linear scale, a Hadoop Cluster can be added with tens, hundreds, or even thousands of servers at any time with minimal effort.

Flexible:

Hadoop enables businesses to easily access new data sources and tap into different types of data (both structured and unstructured) to generate value from that data. This means businesses can use Hadoop to derive valuable business insights from data sources such as social media, email conversations or clickstream data. In addition, Hadoop can be used for a wide variety of purposes, such as log processing, recommendation systems, data warehousing, market campaign analysis and fraud detection.

Data locality - Move computation rather than data

Data blocks reside on each nodes locally. In hadoop, there is no central system to hold the data as like traditional systems that stores data in SAN that bring the data through network to the program running node, process and store back the result to the SAN for client's access, hadoop runs the program on the node where data resides and coordinate all nodes to produce the output and produce to the client.

Highly Reliable

Data is replicated across multiple nodes (replication factor is configurable) and if a node goes down, the required data can be read from another node which has the copy of that data. And it also ensures that the replication factor is maintained, even if a node goes down, by replicating the data to other available nodes.

Integerated

Data integrity will be maintained by comparing the checksum value of the data at the time to data movement between data nodes and clients.

Cost Effectiveness

Commodity hardware, no need of high end servers. Open source, no licence or reneval fee.

Fault Tolerant

Hadoop has built-in fault tolerance. If program fails in one node can be handled automatically by executing the same piece of program in some other node has the same replica data, hence hadoop is highly reliable.

When to Use Hadoop

Hadoop can be used in various scenarios including some of the following:

- Your data may expect to grow exponentially in huge volume.
- If you don't want to spend much on licensing cost and hardware, since hadoop is open source in commodity hardware.
- If you wanted to store raw data of its own format ie schema less storage system.

- You need to store heterogeneous historical and live data that is available to process at the same time.
- Need to bring all features such as acquisition, ETL, retention, data mining, analytics and machine learning under single roof.
- Need more granular data for the true analytics for business intensive applications such as banking, scientific domains.

When Not to Use Hadoop

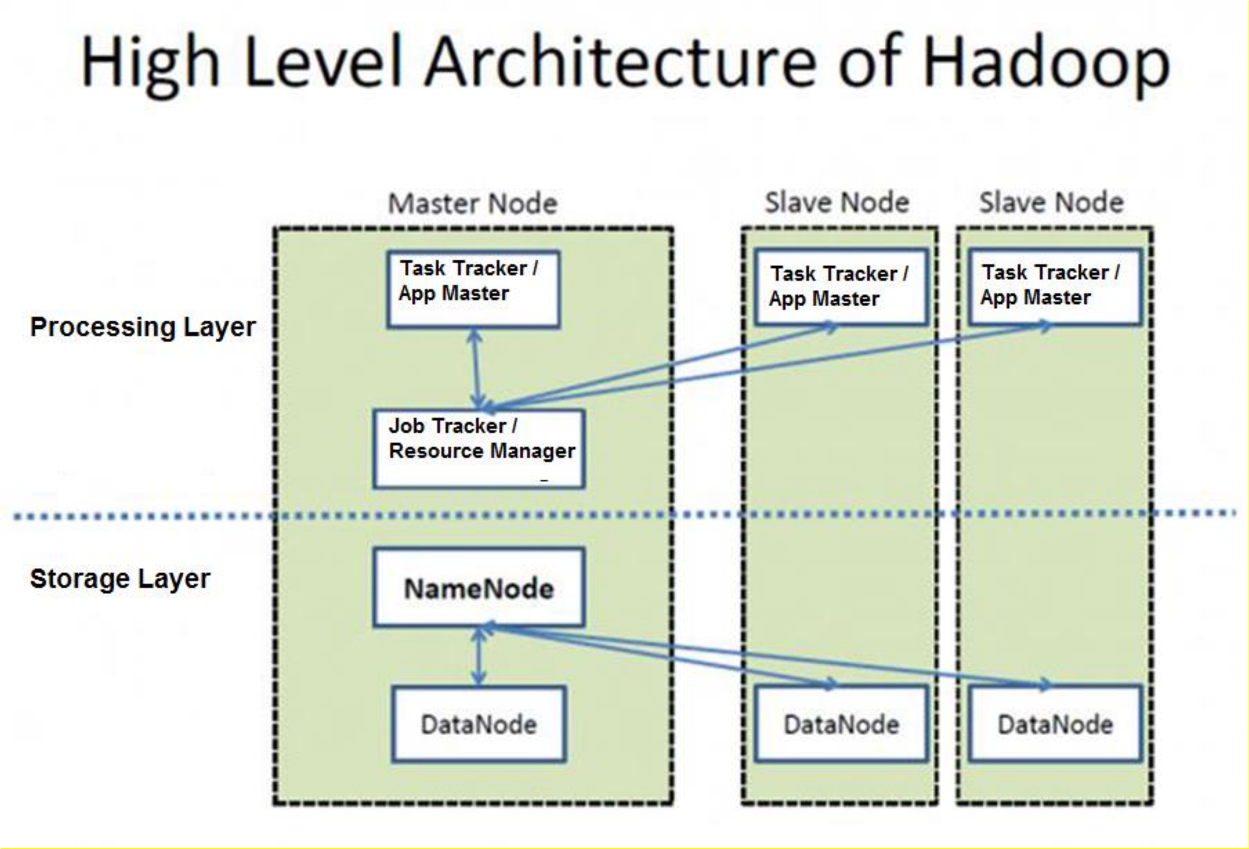
There are few scenarios in which Hadoop is not the right fit currently.

- POS systems with low latency data processing with huge number of small files and perform frequent transactions.
- If needed to replace the complete traditional datawarehouse - Hadoop is complementary approach to a traditional data warehouse, not a replacement for it..
- For the mission critical business intensive systems where frequent upgradation or frequent change in application is a risky process to perform trial runs.

Components of Hadoop

1. **Hadoop Common:** contains libraries and utilities needed by other Hadoop modules.
2. **Hadoop Distributed File System (HDFS):** a distributed file-system that stores data on the commodity machines, providing very high aggregate bandwidth across the cluster.
3. **Hadoop MapReduce:** a programming model for large scale data processing.
4. **Hadoop YARN:** a resource-management platform responsible for managing compute resources in clusters and using them for scheduling of users' applications.

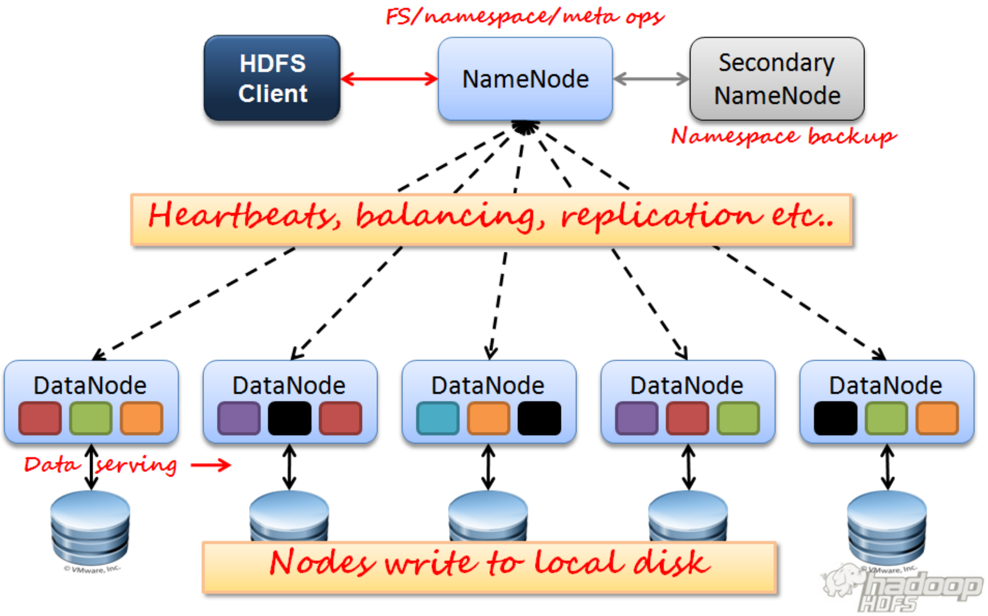
Architecture

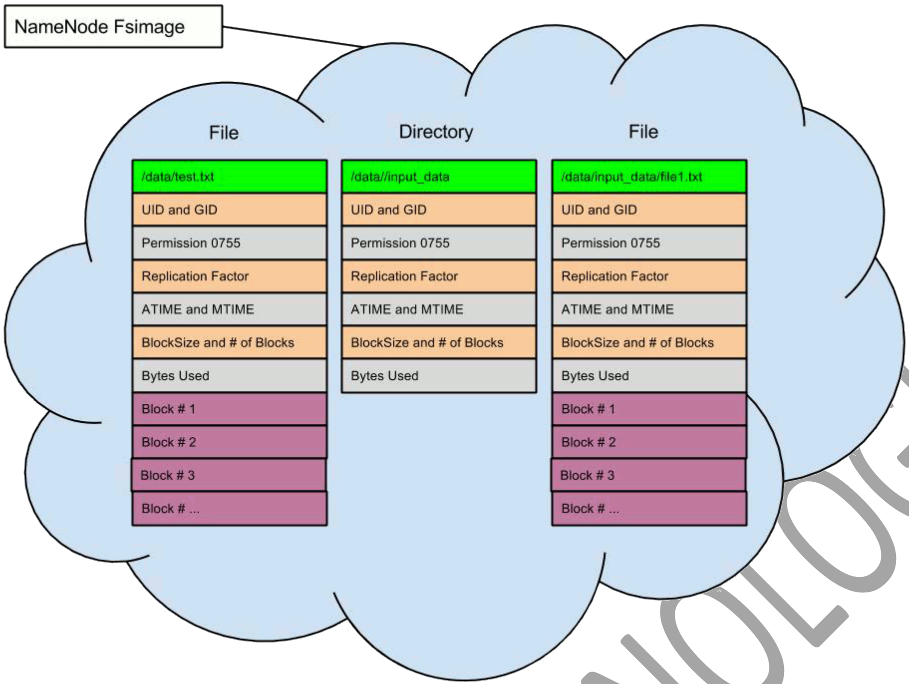


High Level Architecture of Hadoop

Master Node
Slave Node
Slave Node
Processing Layer
Task Tracker / App Master
Task Tracker / App Master
Task Tracker / App Master
Job Tracker / Resource Manager
Storage Layer
NameNode
DataNode
DataNode
DataNode

HDFS Architecture





NameNode is the master node or DFS Master of the system. It maintains the name system (directories and files) and manages the blocks which are present on the DataNodes.

Role of NameNode:

1. Name node holds the metadata for HDFS in memory and in disk.
2. Metadata in Memory serves the purpose of instant access to the metadata.
3. Files (fsimage and editlog) serves only when the cluster is restarted due to failure which will be referencing these files.
4. Controls read/write access to files, check the the existence of the directory/files before read/write.
5. Manages blocks, replication and re replication.

Metadata Components:

fsimage – Stores the inode details like modification time, access time, access permission, replication. **editlogs** – This keeps tracking of each and every change that is being done on HDFS. (Like adding a new file, deleting a file, moving it between folders..etc).

Any change that we do to HDFS will be tracked in edit logs, not in the fsimage. So the edit logs file will keep on growing whereas the fsimage size remains the same. This won't have any impact unless or until we restart the cluster. When we restart the cluster, the fsimage needs to get loaded in to the main memory. Since all changes are present in the edilogs not in the fsimage, hadoop will try to write editlogs changes to fsimage (this takes some time depends on the size of the editlogs file).

The namenode maintains the entire metadata in RAM, which helps clients receive quick responses to read requests. Therefore, it is important to run namenode from a machine that has lots of RAM at its disposal. The higher the number of files in HDFS, the higher the consumption of RAM. The namenode daemon also maintains a persistent checkpoint of the metadata in a file stored on the disk called thefsimage file.

Whenever a file is placed/deleted/updated in the cluster, an entry of this action is updated in a file called the edits logfile. After updating the edits log, the metadata present in-memory is also updated accordingly.

It is important to note that the fsimage file is not updated for every write operation.

Secondary Name Node:

1. Secondary name node act as a backup node in the case of name node crashed and the fsimage/editlog is lost completely in namenode. With the Admin interaction the FSImage from Secondary name node can be copied to a new name node and bring the name node up and running using the fsimage copied from SNN.
2. The secondary name node is responsible for performing periodic housekeeping functions for the *NameNode*. It only creates checkpoints of the filesystem present in the *NameNode*.

DataNode

1. Data Nodes are the slaves which are deployed on each machine and provide the actual storage.
2. They are responsible for serving read and write requests for the clients.
3. All datanodes send a heartbeat message to the namenode every 3 seconds to say that they are alive. If the namenode does not receive a heartbeat from a particular data node for 10 minutes, then it considers that data node to be dead/out of service and initiates replication of blocks which were hosted on that data node to be hosted on some other data node.
4. The data nodes can talk to each other to rebalance by move and copy data around and keep the replication high.
5. When the datanode stores a block of information, it maintains a checksum for it as well. The data nodes update the namenode with the block information periodically and before updating verify the checksums. If the checksum is incorrect for a particular block i.e. there is a ***disk level corruption for that block,*** it skips that block while reporting the block information to the namenode. In this way, namenode is aware of the disk level corruption on that datanode and takes steps accordingly.
6. **Blockreport** - The DataNode stores HDFS data in files in its local file system. The DataNode has no knowledge about HDFS files. It stores each block of HDFS data in a separate file in its local file

system. The DataNode does not create all files in the same directory. Instead, it uses a heuristic to determine the optimal number of files per directory and creates subdirectories appropriately. It is not optimal to create all local files in the same directory because the local file system might not be able to efficiently support a huge number of files in a single directory. When a DataNode starts up, it scans through its local file system, generates a list of all HDFS data blocks that correspond to each of these local files and sends this report to the NameNode.

Terminologies

Cluster: A cluster is a group of servers and other resources that act like a single system and enable high availability and, in some cases, load balancing and parallel processing.
Node: A single machine in cluster
Block: A block is the smallest unit of data that can be stored or retrieved from the disk. File systems deal with the data stored in blocks.