



# Big Data and Hadoop

NIELIT Chandigarh/Ropar



*"Big Data is at the foundation of all the megatrends that are happening today."* — Chris Lynch



# Introduction to Big Data

1. Today's business enterprises are data-driven and without data no enterprise can have a competitive advantage.
2. ***Data is the new currency and oil of our generation.***
3. “Big Data” is the data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it.\*
4. Big data are extremely large data sets that may be analyzed computationally to reveal pattern, trends and associations especially related to human behavior and interactions
5. With each passing day, Big data is growing bigger, is more difficult to make sense of, is being generated at a much faster rate and this trend is only going to intensify in our data-driven digital world.
6. Today Big Data is so rampant that one has to look which are the companies that are not deploying Big Data.



# Importance of Big Data

- **Improved Decision-Making:** Analyzing Big Data allows businesses to make data-driven decisions, reducing risks and uncovering new opportunities.
- **Enhanced Customer Experiences:** By understanding customer behavior and preferences, businesses can personalize products and services.
- **Operational Efficiency:** Big Data enables process optimization, cost reduction, and better resource allocation.
- **Innovation and Growth:** Data analysis reveals patterns and trends, fostering innovation and driving growth in various industries.
- **Predictive Analytics:** Leveraging machine learning on Big Data helps forecast outcomes and anticipate trends.

# Business Relevance of Big Data

- **Marketing and Sales:** Provides insights into customer preferences, enabling targeted marketing strategies and improved sales performance.
- **Supply Chain Management:** Helps optimize logistics, inventory, and demand forecasting.
- **Healthcare:** Facilitates advancements in diagnostics, personalized medicine, and real-time health monitoring.
- **Finance:** Enhances fraud detection, risk management, and algorithmic trading.
- **Smart Cities:** Big Data plays a critical role in traffic management, energy optimization, and public safety initiatives.
- **E-Commerce:** Drives product recommendations, pricing strategies, and customer retention.





- Big Data is a collection of data that is huge in volume yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.

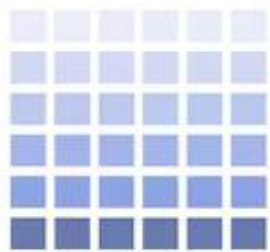


# WHAT IS BIG DATA

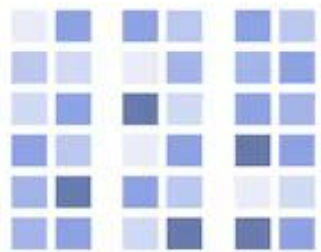
- Big Data are high volume, high velocity, or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery, and process optimization.
- The term 'big data' is self-explanatory – a collection of huge data sets that normal computing techniques cannot process.
- The term not only refers to the data, but also to the various frameworks, tools, and techniques involved.
- Technological advancement and the advent of new channels of communication (like social networking) and new, stronger devices have presented a challenge to industry players in the sense that they have to find other ways to handle the data.
- Big data is an all-inclusive term, representing the enormous volume of complex data sets that companies and governments generate in the present-day digital environment.
- Big data, typically measured in petabytes or terabytes, materializes from three major sources— transactional data, machine data, and social data.

# TYPES OF BIG-DATA

## Types of Big Data



Structured  
Data



Semi-Structured  
Data



Unstructured  
Data

Big Data is generally categorized into three different varieties. They are as shown below:

- Structured Data
- Semi-Structured Data
- Unstructured Data

# TYPES OF BIG-DATA

- Structured Data owns a dedicated data model, It also has a well-defined structure, it follows a consistent order and it is designed in such a way that it can be easily accessed and used by a person or a computer. Structured data is usually stored in well-defined columns and also Databases.

Example: Database Management Systems(DBMS)

- Semi-Structured Data can be considered as another form of Structured Data. It inherits a few properties of Structured Data, but the major part of this kind of data fails to have a definite structure and also, it does not obey the formal structure of data models such as an RDBMS.

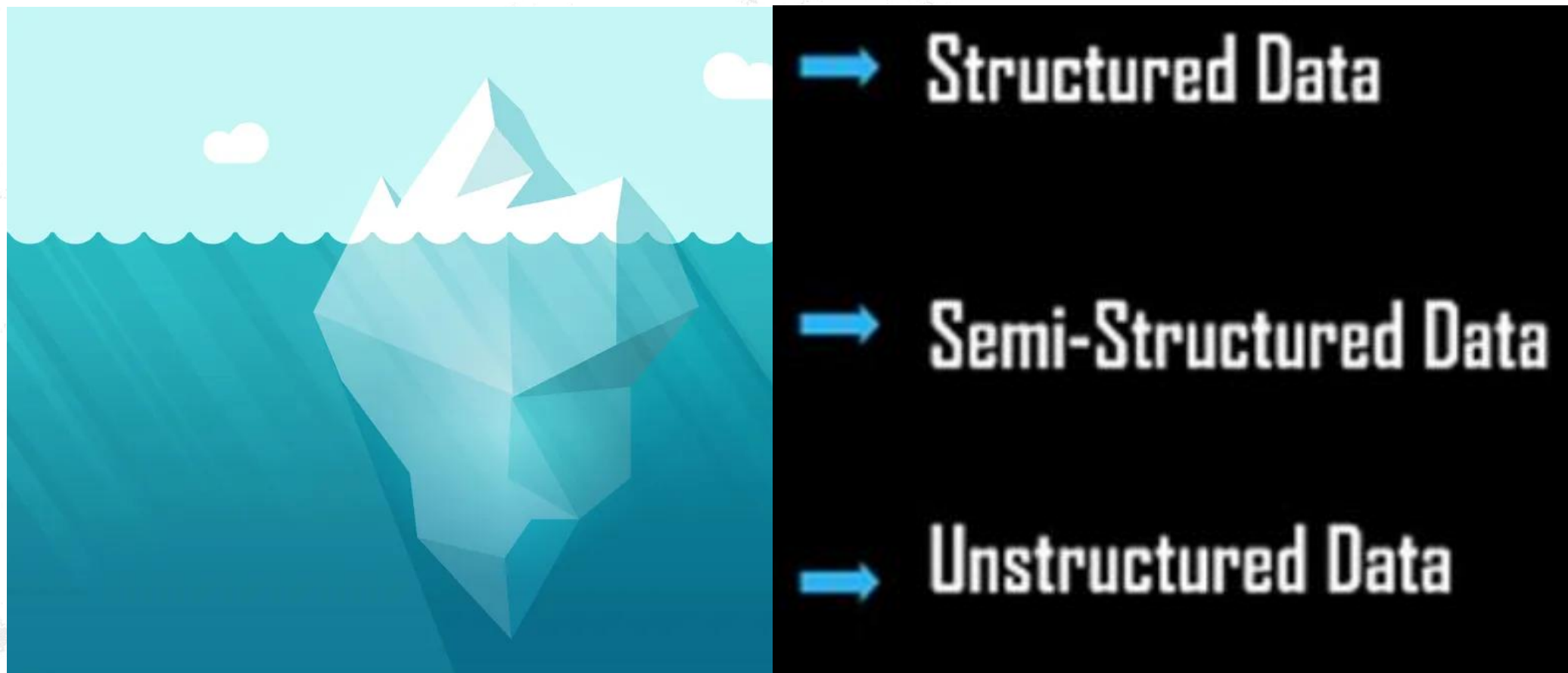
Example: Comma Separated Values(CSV) File.

- Unstructured Data is completely a different type of which neither has a structure nor obeys to follow the formal structural rules of data models. It does not even have a consistent format and it found to be varying all the time. But, rarely it may have information related to data and time.

Example: Audio Files, Images etc



# TYPES OF BIG-DATA



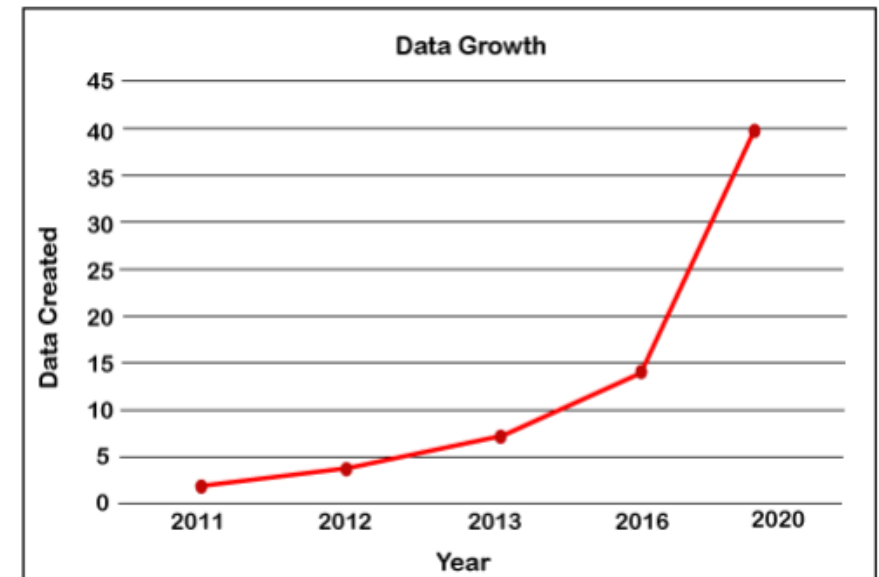
# THE CHARACTERISTICS OF BIG DATA



# THE CHARACTERISTICS OF BIG DATA

## Volume

- Volume refers to the unimaginable amounts of information generated every second from social media, cell phones, cars, credit cards, M2M sensors, images, video, and whatnot. We are currently using distributed systems, to store data in several locations and brought together by a software Framework like Hadoop. Facebook alone can generate about billion messages, 4.5 billion times that the “like” button is recorded, and over 350 million new posts are uploaded each day. Such a huge amount of data can only be handled by Big Data Technologies



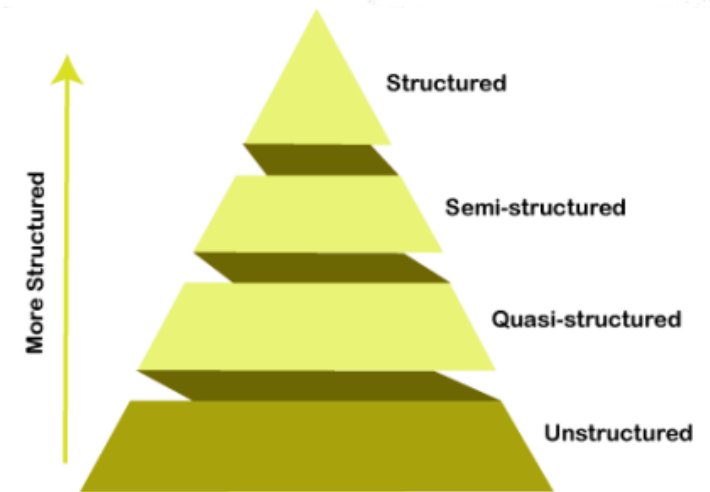
# THE CHARACTERISTICS OF BIG DATA

## Variety

- As Discussed before, Big Data is generated in multiple varieties. Compared to the traditional data like phone numbers and addresses, the latest trend of data is in the form of photos, videos, and audios and many more, making about 80% of the data to be completely unstructured

## Veracity

- Veracity basically means the degree of reliability that the data has to offer. Since a major part of the data is unstructured and irrelevant, Big Data needs to find an alternate way to filter them or to translate them out as the data is crucial in business developments

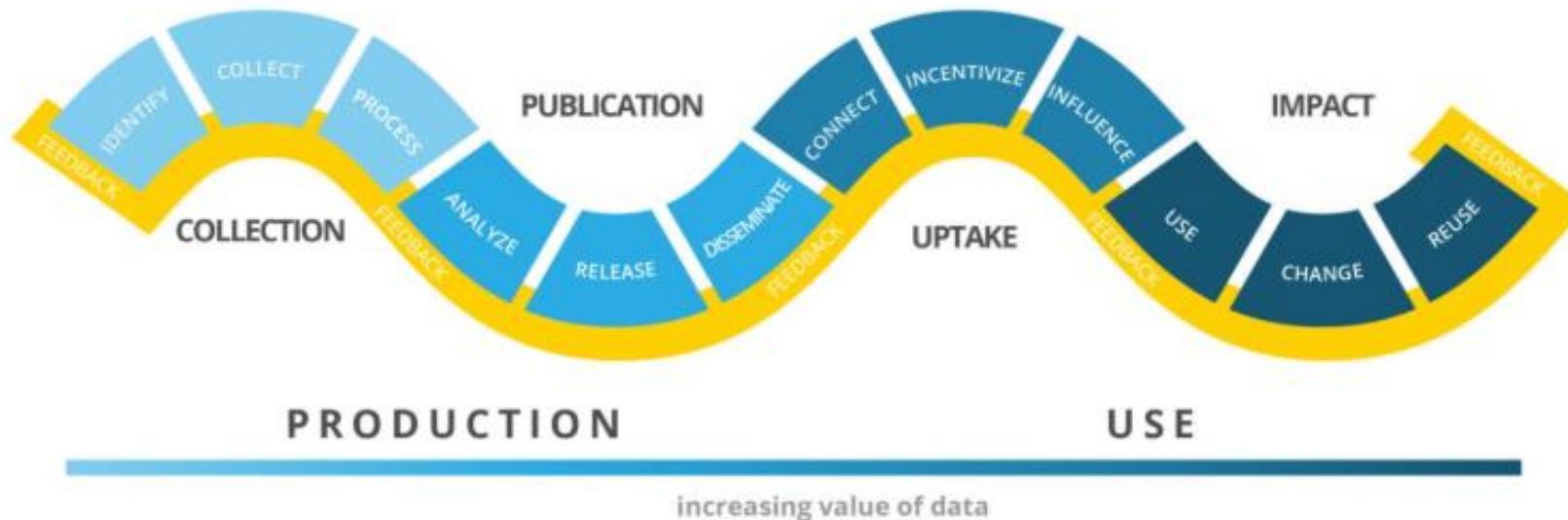




# THE CHARACTERISTICS OF BIG DATA

## Value

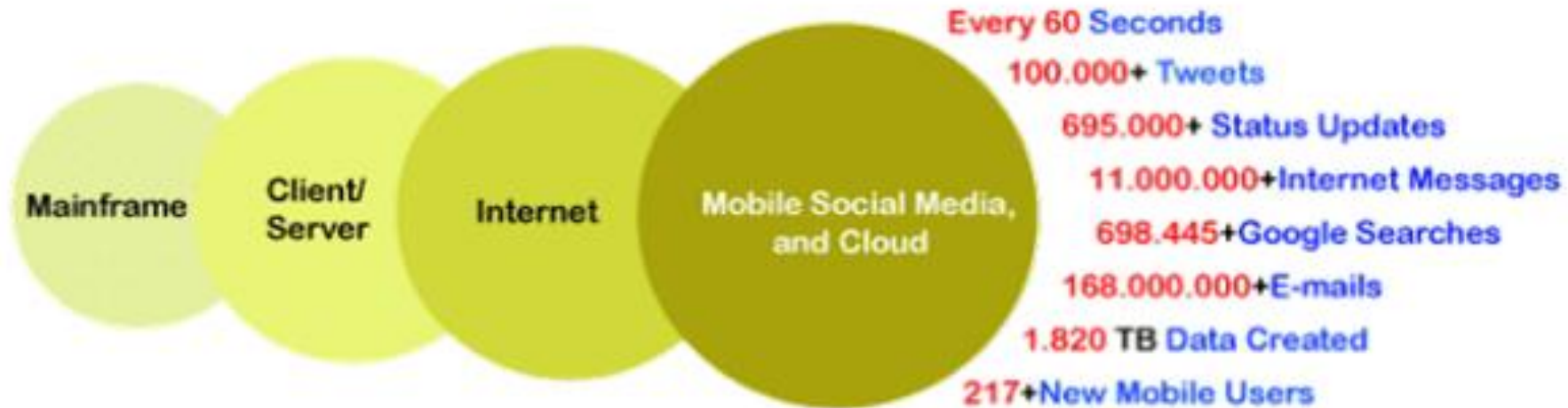
- Value is the major issue that we need to concentrate on. It is not just the amount of data that we store or process. It is actually the amount of valuable, reliable and trustworthy data that needs to be stored, processed, analyzed to find insights.



# THE CHARACTERISTICS OF BIG DATA

## Velocity

- Last but never least, Velocity plays a major role compared to the others, there is no point in investing so much to end up waiting for the data. So, the major aspect of Big Data is to provide data on demand and at a faster pace



# APPLICATIONS OF BIG DATA

## Retail

- Leading online retail platforms are wholeheartedly deploying big data throughout a customer's purchase journey, to predict trends, forecast demands, optimize pricing, and identify customer behavioral patterns. Big data is helping retailers implement clear strategies that minimize risk and maximize profit.

## Healthcare

- Big data is revolutionizing the healthcare industry, especially the way medical professionals in the past diagnosed and treated diseases. In recent times, effective analysis and processing of big data by machine learning algorithms provide significant advantages for the evaluation and assimilation of complex clinical data, which prevent deaths and improve the quality of life by enabling healthcare workers to detect early warning signs and symptoms.

# APPLICATIONS OF BIG DATA

## Financial Services and Insurance

- The increased ability to analyze and process big data is dramatically impacting the financial services, banking, and insurance landscape.
- In addition to using big data for swift detection of fraudulent transactions, lowering risks, and supercharging marketing efforts, few companies are taking the applications to the next levels

## Manufacturing

- Advancements in robotics and automation technologies, modern-day manufacturers are becoming more and more data focused, heavily investing in automated factories that exploit big data to streamline production and lower operational costs.
- Top global manufacturers are also integrating sensors into their products, capturing big data to provide valuable insights on product performance and its usage.



# APPLICATIONS OF BIG DATA

## Energy

- To combat the rising costs of oil extraction and exploration difficulties because of economic and political turmoil, the energy industry is turning toward data-driven solutions to increase profitability.
- Big data is optimizing every process while cutting down energy waste from drilling to exploring new reserves, production, and distribution.

## Logistics & Transportation

- State-of-the-art warehouses use digital cameras to capture stock level data, which, when fed into ML algorithms, facilitates intelligent inventory management with prediction capabilities that indicate when restocking is required.
- In the transportation industry, leading transport companies now promote the collection and analysis of vehicle telematics data, using big data to optimize routes, driving behavior, and maintenance.

# APPLICATIONS OF BIG DATA

## Government

- Cities worldwide are undergoing large-scale transformations to become “smart”, through the use of data collected from various Internet of Things (IoT) sensors.
- Governments are leveraging this big data to ensure good governance via the efficient management of resources and assets, which increases urban mobility, improves solid waste management, and facilitates better delivery of public utility services.

# Some of Companies Generating Huge Data

**1. Facebook :-** Arguably the world's most popular social media network with more than two billion monthly active users worldwide. Every day, we feed Facebook's data beast with mounds of information.

- Every 60 seconds, 136,000 photos are uploaded,
- 510,000 comments are posted
- 293,000 status updates are posted.
- With data like this, Facebook knows who our friends are, what we look like, where we are, what we are doing, our likes, our dislikes, and so much more. Some researchers even say Facebook has enough data to know us better than our therapists!




**2. Twitter, officially known as X since July 2023, :-** is a gold mine of data. Unlike other social platforms, almost every user's tweets are completely public and pullable.

- There are 6,000 tweets per second
- This equates to 500 million tweets every single day
- Handling the huge quantities of data takes skill
- Terabytes worth of data uploaded to the servers every day!

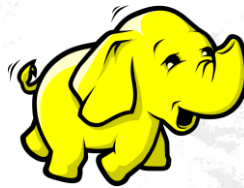


# Some of Companies Generating Huge Data

3. **Instagram** :- Instagram, the social networking app for sharing photos and videos, launched in 2010. Today, it boasts 800 million monthly active users and is owned by Facebook. People interact by showing their love with a heart, commenting and using hashtags. What all of this activity does is create an enormous amount of data..
  - Every 60 seconds, 136,000 photos are uploaded,
  - 510,000 comments are posted
  - 293,000 status updates are posted.
4. In essence starting from technology companies like Google, Apple, Amazon, Microsoft all the way to mining companies like Rio Tinto, retailers like Walmart and hospitality companies like Airbnb are all using big data.
  - **Amazon** – Getting insights on customer data & providing better user experience
  - **Google** – making sense of what the customer is searching for and providing better search results
  - **Walmart** – providing customers what they are looking for in terms of products, discounts, etc.



# What is Hadoop?



Hadoop is an **open-source framework** for storing and processing large datasets in a distributed environment across multiple nodes.

Designed to handle **Big Data** challenges by providing:

- **Scalability**: Easily scales up to accommodate large data.
- **Fault Tolerance**: Ensures data availability even in case of failures.
- **Cost Efficiency**: Runs on commodity hardware.

Hadoop was created by **Doug Cutting** and **Mike Cafarella** in **2006** while working at Yahoo.

- **Why?** : To handle massive amounts of data efficiently.

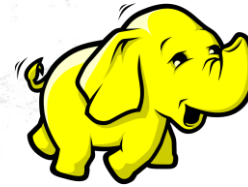
Inspired by Google's papers on the **GFS (Google File System)** and **MapReduce**.

- **Why the Name Hadoop?**

Doug Cutting named it after his son's yellow toy elephant called "Hadoop."

The name has no technical meaning but was catchy and unique

# Who is Doug Cutting ?



- Doug Cutting and his contribution:
- **Hadoop:** Doug Cutting, along with his colleague Mike Cafarella, created **Hadoop** in 2005. Hadoop was inspired by Google's MapReduce programming model and the Google File System (GFS). It was designed to handle large-scale data processing and storage in a way that could scale to massive data volumes across many machines.
- **The Importance of Hadoop:** Hadoop became a foundational technology for big data because it allowed organizations to process huge amounts of unstructured and structured data using inexpensive hardware. Its ability to scale out by adding more machines to handle increasing data loads made it extremely popular in the world of big data.
- **Contributions to Distributed Computing:** Doug Cutting's work on Hadoop helped shape the modern era of distributed computing, particularly in the context of big data storage, processing, and analysis.

# Features of Hadoop

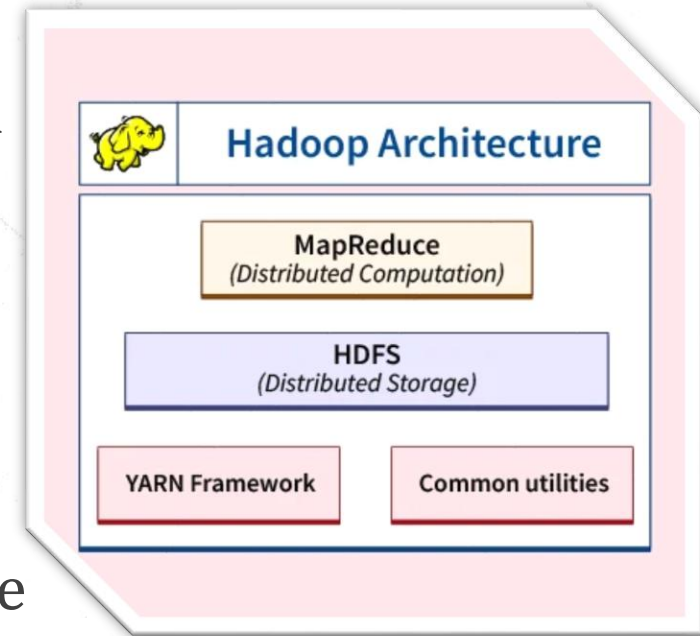
- **Scalability:** Can scale from a single node to thousands of nodes.
- **Fault Tolerance:** Automatically replicates data across nodes to ensure no data is lost in case of hardware failure.
- **Cost-Effective:** Runs on commodity hardware, reducing the overall cost.
- **Flexibility:** Handles structured, semi-structured, and unstructured data.

## Common Use Cases:

- Data warehousing and analytics
- Log processing
- Fraud detection
- Social media analysis
- Recommendation systems

# Core Components of Hadoop Framework

- **Hadoop Distributed File System (HDFS)**
- A **distributed storage system** that splits data into blocks and distributes them across nodes in a cluster.
- **Key Features:**
  - **Block Storage:** Files are divided into 128 MB (default) blocks.
  - **Replication:** Ensures data availability by replicating each block (default replication factor is 3).
  - **Fault Tolerance:** Automatically recovers from hardware failures.
- **Architecture:**
  - **NameNode:** Master node managing metadata.
  - **DataNodes:** Worker nodes storing actual data.





# Benefits of Hadoop Framework

- **Handles Structured, Semi-Structured, and Unstructured Data** : Hadoop can process and store various types of data, including structured, semi-structured, and unstructured data (e.g., text, images, videos).
- **Scalability** : Scales horizontally by adding more nodes to the cluster, allowing Hadoop to handle ever-growing volumes of data efficiently.
- **Fault Tolerance** : Ensures high availability and data integrity through data replication across nodes. If a node fails, Hadoop automatically recovers the data from another replica.
- **Cost Efficiency** : Hadoop runs on commodity hardware, reducing the need for expensive storage systems, making it a cost-effective solution for processing large datasets.
- **Flexibility** : Supports diverse data formats and can integrate with other technologies, making it versatile for various data processing needs.

# Benefits of Hadoop Framework

- **High Throughput and Parallel Processing** : Uses a distributed processing model (MapReduce) to process large datasets in parallel across multiple nodes, improving performance and speed.
- **Security and Data Integrity** : Provides security features such as data encryption and authentication, ensuring secure data access and integrity throughout the processing pipeline.
- **Data Locality** : Processes data where it is stored, minimizing data transfer across the network, reducing latency, and improving overall processing efficiency.
- **Ecosystem and Integration** : The Hadoop ecosystem includes various tools (like Hive, Pig, HBase, Spark) that provide extensive functionality for data processing, querying, and management.
- **Open Source and Community Support** : Being open-source, Hadoop has a large global community that continuously contributes to its development, offering strong support and frequent updates.
- **Batch Processing and Real-Time Processing** : Hadoop supports both batch processing and real-time data processing through frameworks like Apache Spark, allowing for a wide range of use cases.

# Why Apache Hadoop?

1. When the type of data is unstructured, the volume of data is huge, and the results needed are at uncompromisable speeds, then the only platform that can effectively stand up to the challenge is Apache Hadoop.
2. Hadoop was created by Doug Cutting and Mike Cafarella in 2005. It was originally developed to support distribution for **Nutch search engine project**.
3. Doug, who was working at Yahoo! at the time named the project after his **son's toy elephant** who was 2 years old at the time and just began to talk.
4. In 2007, Yahoo successfully tested Hadoop on a 1000 node cluster and start using it. In January of 2008, Yahoo released Hadoop as an **open source project** to **ASF (Apache Software Foundation)**. And in July of 2008, Apache Software Foundation successfully tested a 4000 node cluster with Hadoop.
5. Hadoop owes its runaway success to a processing framework, **MapReduce**, that is central to its existence.
6. MapReduce technology lets ordinary programmers work efficiently without having to worry about intra-cluster complexities, monitoring of tasks, node failure management, and so on.

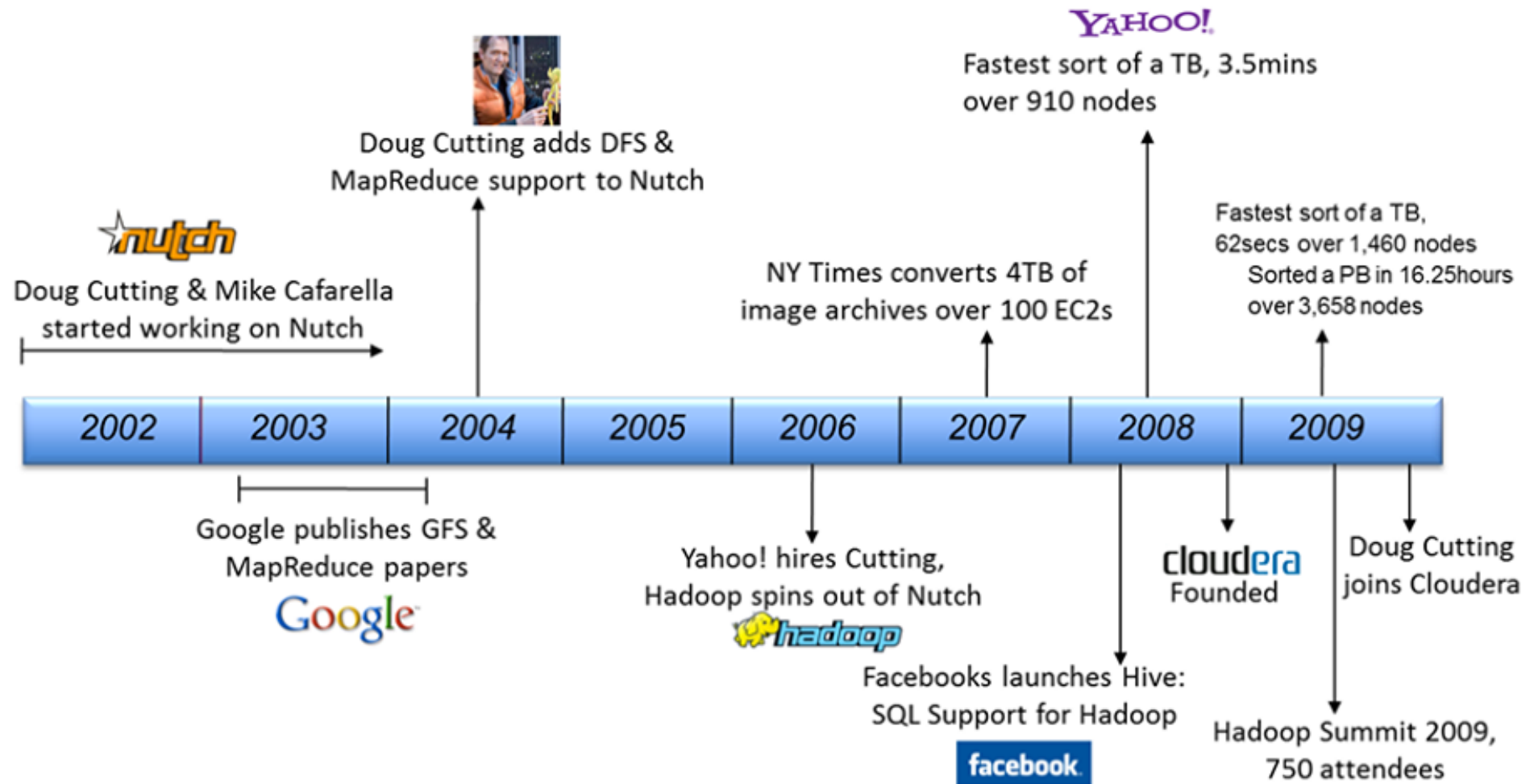
# Why Apache Hadoop?

- Big Data!!
  - Storage
  - Analysis
  - Data management

	Traditional RDBMS	MapReduce
Data size	Gigabytes	Petabytes
Access	Interactive and batch	Batch
Updates	Read and write many times	Write once, read many times
Structure	Static schema	Dynamic schema
Integrity	High	Low
Scaling	Nonlinear	Linear



# History of Hadoop



# Important Skills Required

## Skill 1

1. Linux as the operating system and Ubuntu as server distribution is the preferred choice for Hadoop installations
2. Knowledge of commands and the editor will help during Hadoop installation and file management operations



Linux

## Skill 2

1. Programming knowledge is not required as such, but it depends on the role. For example, a data analyst will know Python and R. A Hadoop developer will know Java or Scala
2. Hadoop is all about handling and processing data



Programming Languages

## Skill 3

1. Knowledge of SQL query and commands are must to go about learning Hadoop
2. Hadoop ecosystem has many software packages like Apache Hive, HBase, and Pig, etc. that extracts data from HDFS using SQL like queries



SQL knowledge

# Introduction to Apache Hadoop

1. Apache Hadoop is the most important framework for working with Big Data. Hadoop biggest strength is scalability. It upgrades from working on a single node to thousands of nodes without any issue in a seamless manner.
2. It is a framework which is based on java programming. It is intended to work upon from a single server to thousands of machines each offering local computation and storage.
3. It supports the large collection of data set in a distributed computing environment.
4. The Apache Hadoop software library based framework that gives permissions to distribute huge amount of data sets processing across clusters of computers using easy programming models.
5. Hadoop helps to execute large amount of processing where the user can connect together multiple commodity computers to a single-CPU, as a single functional distributed system and have the particular set of clustered machines that reads the dataset in parallel and provide intermediate, and after integration gets the desired output.

# Advantages of Apache Hadoop

1. Hadoop framework allows the user to quickly write and test distributed systems. It is efficient, and it automatically distributes the data and work across the machines and in turn, utilizes the underlying parallelism of the CPU cores.
2. 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.
3. Hadoop libraries are developed to find/search and handle the failures at the application layer.
4. Servers can be added or removed from the cluster dynamically at any point of time and Hadoop continues to operate without interruption.
5. Apache Hadoop is an **open source project** based on Java applications and hence compatible on all the platforms.
6. Apache Hadoop is the most popular and powerful big data tool, which provides world's best reliable storage layer –HDFS(Hadoop Distributed File System), a batch Processing engine namely MapReduce and a Resource Management Layer like YARN.



# Advantages of Apache Hadoop

---

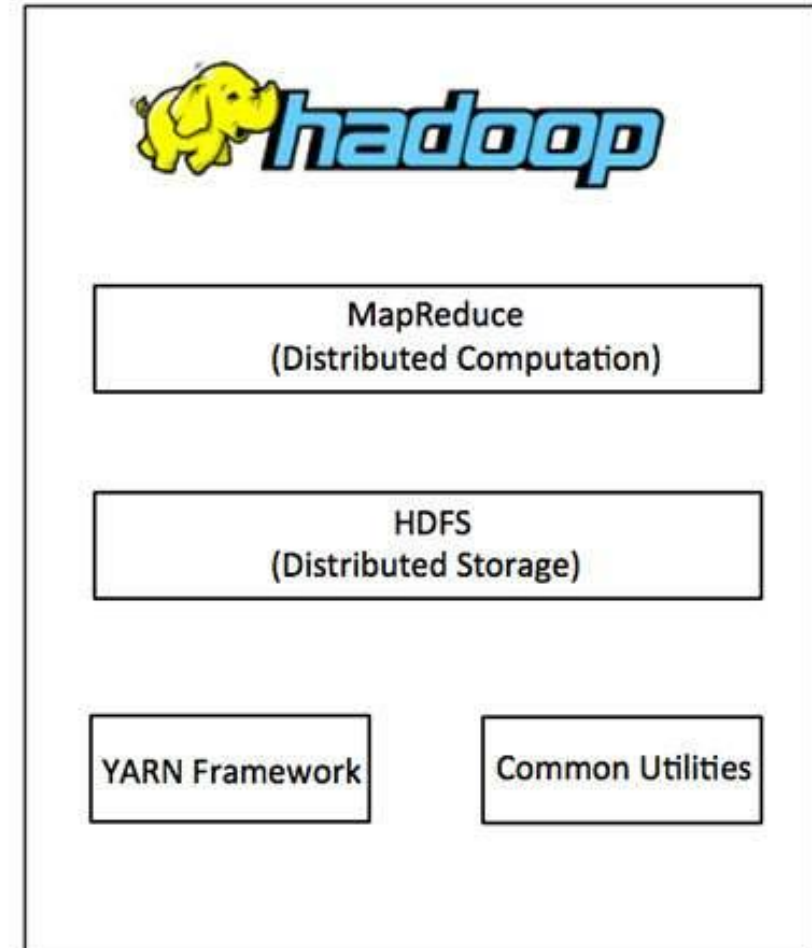
- Vast amounts of data
- Economic
- Efficient
- Scalable
- Reliable

# Applications not for Hadoop

- Low-latency data access
  - HBase is currently a better choice
- Lots of small files
  - All filesystem metadata is in memory
  - The number of files is constrained by the memory size of the name node
- Multiple writers, arbitrary file modifications

# The Core Apache Hadoop Project

- Hadoop Common:
  - Java libraries and utilities required by other Hadoop modules.
- Hadoop YARN:
  - a framework for job scheduling and cluster resource management.
- HDFS:
  - A distributed file system
- Hadoop MapReduce:
  - YARN-based system for parallel processing of large data sets.



# Hadoop Features and Characteristics

1. **Distributed Processing** – The data storage is maintained in a distributed manner in HDFS across the cluster, data is processed in parallel on cluster of nodes.
2. **Fault Tolerance** – By default the three replicas of each block is stored across the cluster in Hadoop and it's changed only when required. Due to replication of data in the cluster, data can be reliable because even if your machine goes down, also then your data will be stored reliably.
3. **Scalability** – Hadoop is highly scalable and in a unique way hardware can be easily added to the nodes. It also provides horizontal scalability which means new nodes can be added on the top without any downtime.
4. **Economic** –We do not require any specialized machine for Hadoop. Hadoop provides huge cost reduction since it is very easy to add more nodes to it. Even if the requirement increases, then there is an increase of nodes, without any downtime and without any much of preplanning.
5. **Easy to use** – No need of client to deal with distributed computing, framework takes care of all the things.
6. **Data Locality** – Hadoop works on data locality principle which states that the movement of computation to data instead of data to computation. When client submits his algorithm, then the algorithm is moved to data in the cluster instead of bringing data to the location where algorithm is submitted and then processing it.



# Hadoop Assumptions

Hadoop is written with huge amount of clusters of computers in mind and is built upon the following assumptions:

1. Hardware may fail due to any external or technical malfunction where instead commodity hardware can be used.
2. Processing will be run in batches and there exits an emphasis on high throughput as opposed to low latency.
3. Applications which run on HDFS have large sets of data. A typical file in HDFS may be of gigabytes to terabytes in size.
4. Applications require a write-once-read-many access model.
5. Moving Computation is cheaper compared to the Moving Data.

# Core Hadoop EcoSystem

2. **YARN :-** Next in the Hadoop ecosystem is YARN (Yet Another Resource Negotiator). It manages the resources on your computing cluster. It is the one that decides who gets to run the tasks, when and what nodes are available for extra work, and which nodes are not available to do so. So, it's like the heartbeat of Hadoop that keeps your cluster going.
3. **MapReduce :-** MapReduce is a High-Performance Parallel Data Processing model that employs the Divide-Conquer principle. Mappers have the ability to transform your data in parallel across your computing cluster in a very efficient manner; whereas, Reducers are responsible for aggregating your data together.
4. **Apache Pig :-** If you are more familiar with a scripting language that has somewhat SQL-style syntax, Pig is for you. In place of writing your code in Java for MapReduce, you can go ahead and write your code in Pig Latin which is similar to SQL. Just writing a Pig Latin code will perform MapReduce functions.
5. **Hive :-** Hive is a way of making the distributed data sitting on your file system somewhere look like a SQL database. It has a language known as Hive SQL. It is just a database in which you can connect to a shell client and ODBC and execute SQL queries on the data that is stored on your Hadoop cluster even though it's not really a relational database under the hood.

# Core Hadoop EcoSystem

6. **Apache Spark** :- It is mainly a real-time data processing engine developed in order to provide faster and easy-to-use analytics than MapReduce. It uses the in-memory processing of data. It can handle SQL queries, do Machine Learning across an entire cluster of information, handle streaming data, etc.
7. **Apache Hbase** :- It is a very fast NoSQL database and a columnar data store, which is meant for achieving large transaction rates. It can expose data stored in your cluster which might be transformed in some way by Spark or MapReduce. It provides a very fast way of exposing those results to other systems.
8. **ZooKeeper** :- ZooKeeper is basically a technology for coordinating everything on your cluster. ZooKeeper can be used for keeping track of which the master node is, which node is up, or which node is down.
9. **Sqoop** :- Sqoop is a tool used for transferring data between relational database servers and Hadoop such as Oracle, MySQL etc.
10. **Kafka** . Kafka aims to provide a unified, low-latency platform to handle real-time data feeds. It is horizontally scalable and fault-tolerant. Asynchronous communication and messages can be established with the help of Kafka.

# Why do you need HDFS File-System?

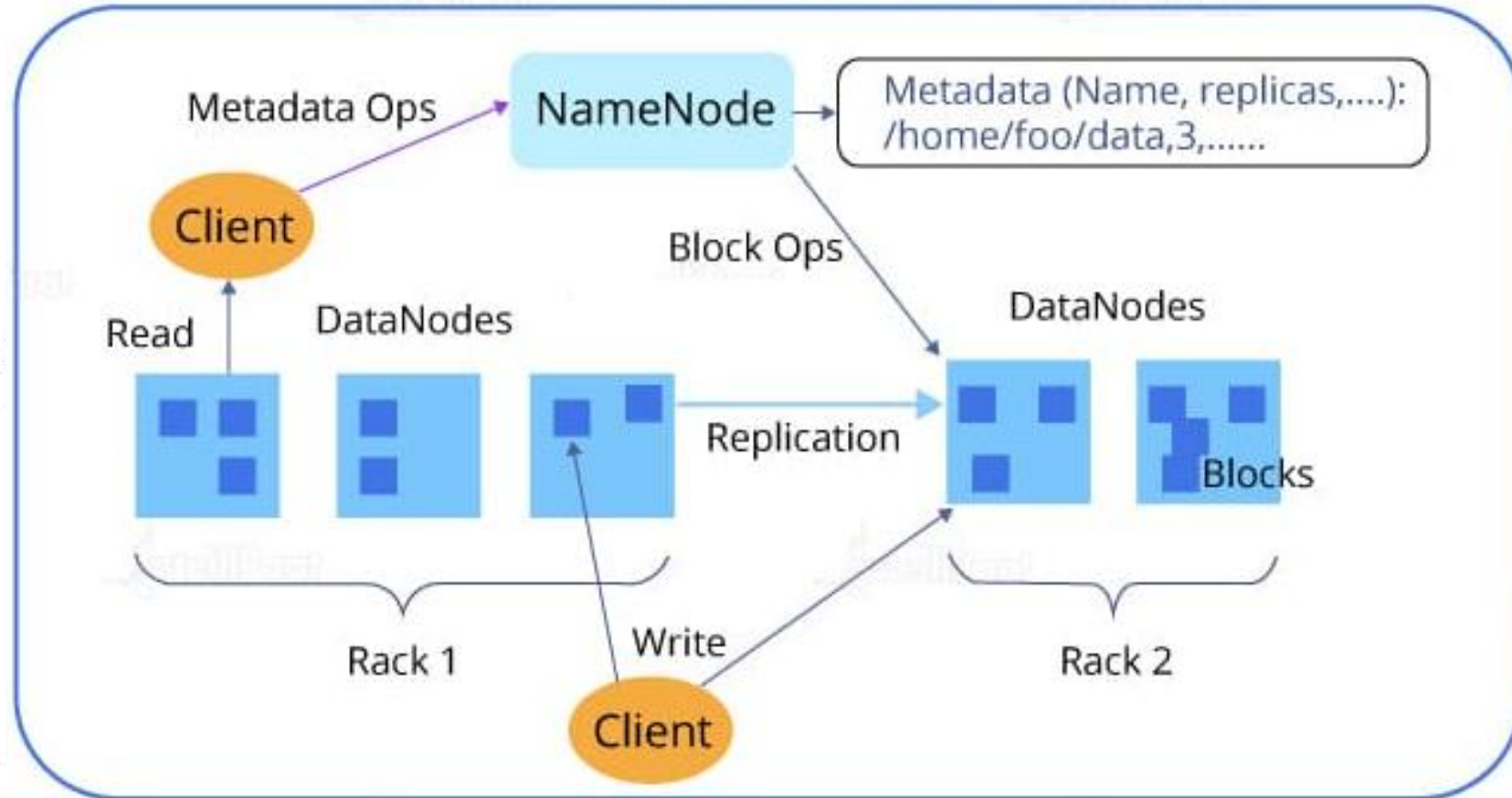
1. Just like a well-organized book, a **file system** manages reading and writing of files on your hard disk and stores metadata regarding your files.
2. With growing data velocity, the data size easily outgrows storage limit of a machine. A solution would be to store the data across a **network of machines** by solving numerous complications of a network. This is where Hadoop comes in & provides one of most reliable filesystems. HDFS for storing data across multiple machines.

## Advantages

1. **Fast recovery from hardware failures:** It is designed to detect failure and automatically recover on its own in case a server in a cluster of HDFS goes down.
2. **Extremely large files:** This file system is designed for storing a very large amount of data in range of petabytes(1000 TB).
3. **Streaming Data Access Pattern:** HDFS is designed on principle of write-once and read-many-times. Once data is written large portions of dataset can be processed any number times.
4. **Commodity hardware:** HDFS uses hardware that is inexpensive and easily available in the market. This is one of feature which specially distinguishes HDFS from other file system.
5. **Portability:** HDFS is portable across hardware platforms and is compatible with many underlying operating systems.



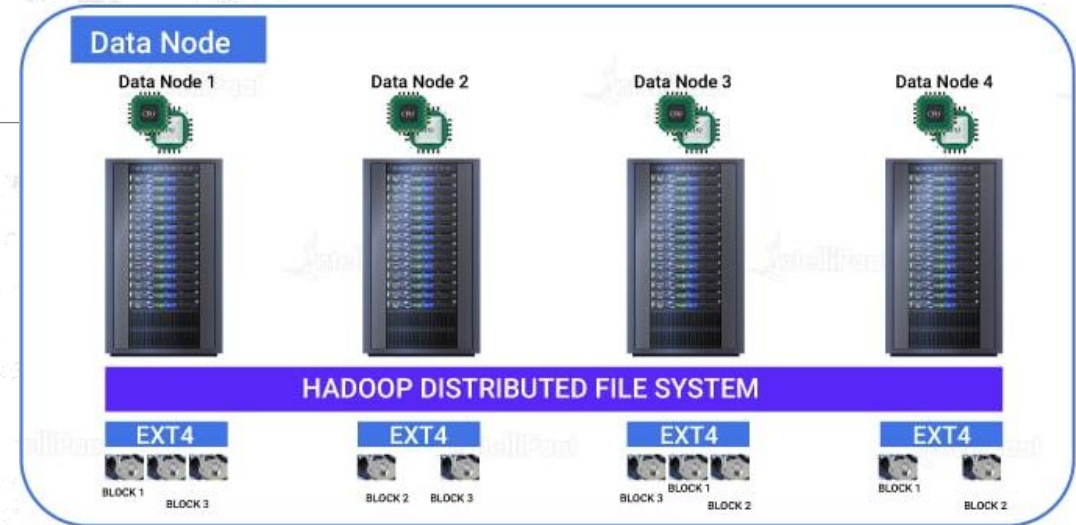
# HDFS Architecture



*The HDFS architecture is a Master-Slave architecture & has several components in it. Two basic nodes in HDFS architecture are DataNode and NameNode*

# DataNode

1. Nodes wherein the blocks are physically stored are known as DataNodes.
2. Every DataNode knows the blocks it is responsible for.
3. Although the DataNode knows about the block it is responsible for, it doesn't care to know about the other blocks and the other DataNodes.
4. As a user because you don't know anything about the blocks other than the file name. Thus, you should be able to work only with the file name in the Hadoop cluster.
5. So the question here is: if the DataNodes do not know which block belongs to which file, then ***who has the key information?*** The key information is maintained by a node called the **NameNode**.



# NameNode

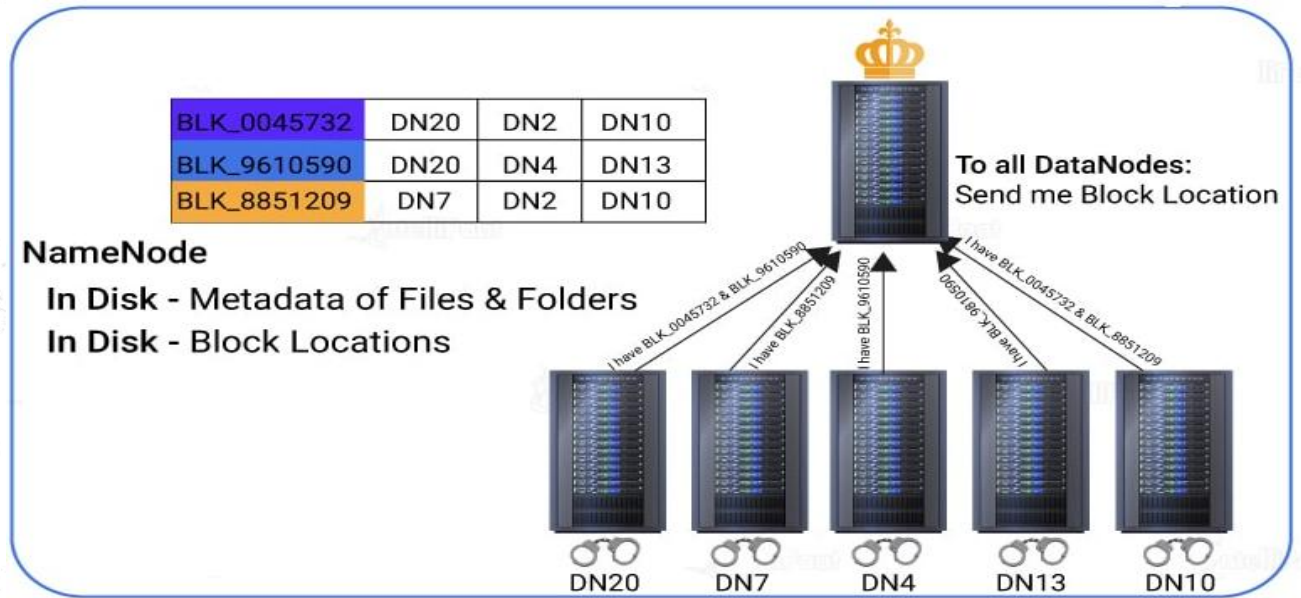
1. A NameNode keeps track of all files or datasets in HDFS. It knows list of file blocks & their locations in HDFS.
2. If a NameNode is down in your Hadoop cluster, there would be no way you could access the files in the cluster.
3. Apart from the block locations, a NameNode also has the metadata of the files and folders in HDFS, which includes information like, the size, replication factor, created by, created on, last modified by, last modified on, etc.
4. Due to the significance of the NameNode, it is also called the **master node** and the DataNodes are called **slave nodes**, and hence the master–slave architecture.
5. NameNode persists all the metadata information about the files and folder and hard disk, except for the block location.





# NameNode

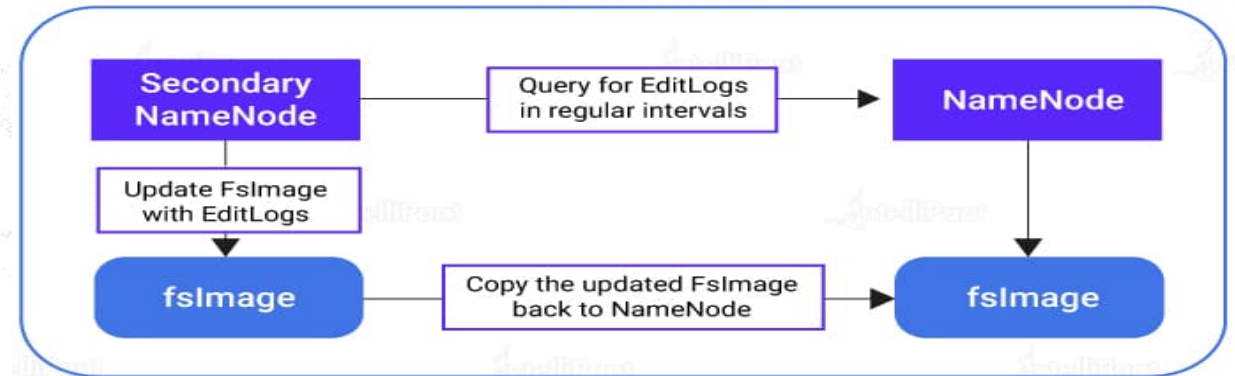
1. When a NameNode starts up, the DataNodes will try to connect with the NameNode and broadcast the list of blocks that each of them is responsible for.
2. NameNode is most powerful node in the cluster in terms of capacity.
3. The NameNode will hold the block locations in memory and never persist the information in the hard disk because in a busy cluster, HDFS is constantly changing with the new data files coming into the cluster and if NameNode has to persist every change by writing the information to a hard disk, it would be a bottleneck.
4. The NameNode thus holds the block locations in memory so that it can give a faster response to the clients.





# Secondary NameNode

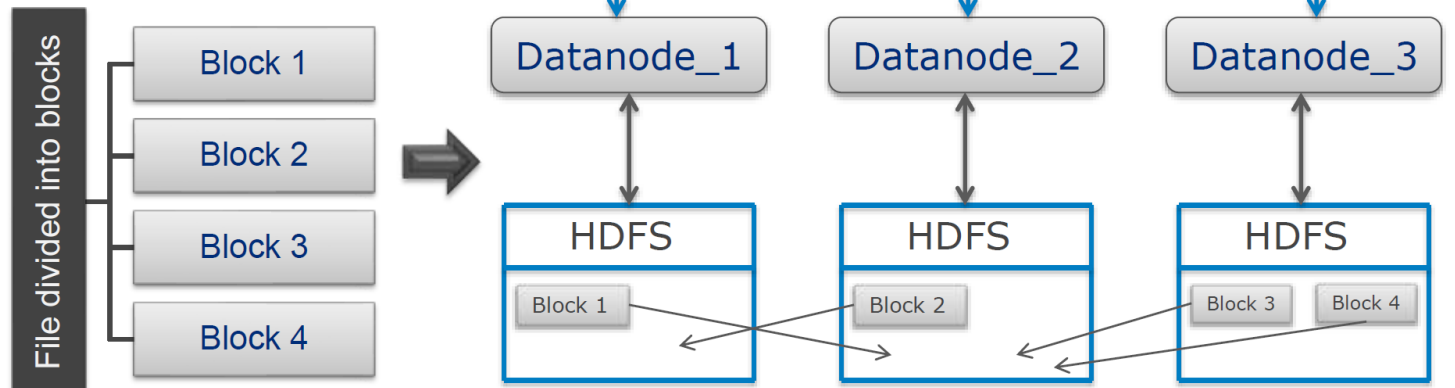
1. Secondary NameNode is a specially dedicated node in HDFS cluster whose main function is to take checkpoints of the file system metadata present on namenode.
2. Hence it is also called the checkpoint node.
3. It is a helper to the primary NameNode but not a backup namenode. The secondary NameNode reads all files, along with the metadata, from the RAM of the NameNode.
4. As the NameNode is the single point of failure in HDFS, if NameNode fails entire HDFS file system is lost. So in order to overcome this, Hadoop implemented Secondary NameNode whose main function is to store a copy of FsImage file and edits log file into the file system or to the hard disk..
5. FsImage is a snapshot of the HDFS file system metadata at a certain point of time and EditLog is a transaction log which contains records for every change that occurs to file system metadata.
6. Whenever a NameNode is restarted, the latest status of FsImage is built by applying edits records on last saved copy of FsImage. Since, NameNode merges FsImage and EditLog files only at start up, if the EditLog is very large, NameNode restart process result in some considerable delay in the availability of FS. Thus, main functions of Secondary NameNode is to keep the edits log as small as possible.
7. It usually runs on a different machine than the primary NameNode since its memory requirements are same as the primary NameNode.



# Storage & Replication of Blocks in HDFS

*When you upload a file into HDFS, it will automatically be split into 128 MB fixed-size blocks. HDFS blocks are huge compared to disk blocks so that the time consumed for transferring data from the disk can be reduced.*

*Default replication factor in HDFS is set to 3 viz one original block and two replicas.*



*A file can be larger than any single disk in the network. HDFS simplifies the storage subsystem & eliminates metadata concerns. It also provides fault tolerance and availability*

