Welcome to Big Data Analytics Module

Prepared By : Tushar Kakaiya

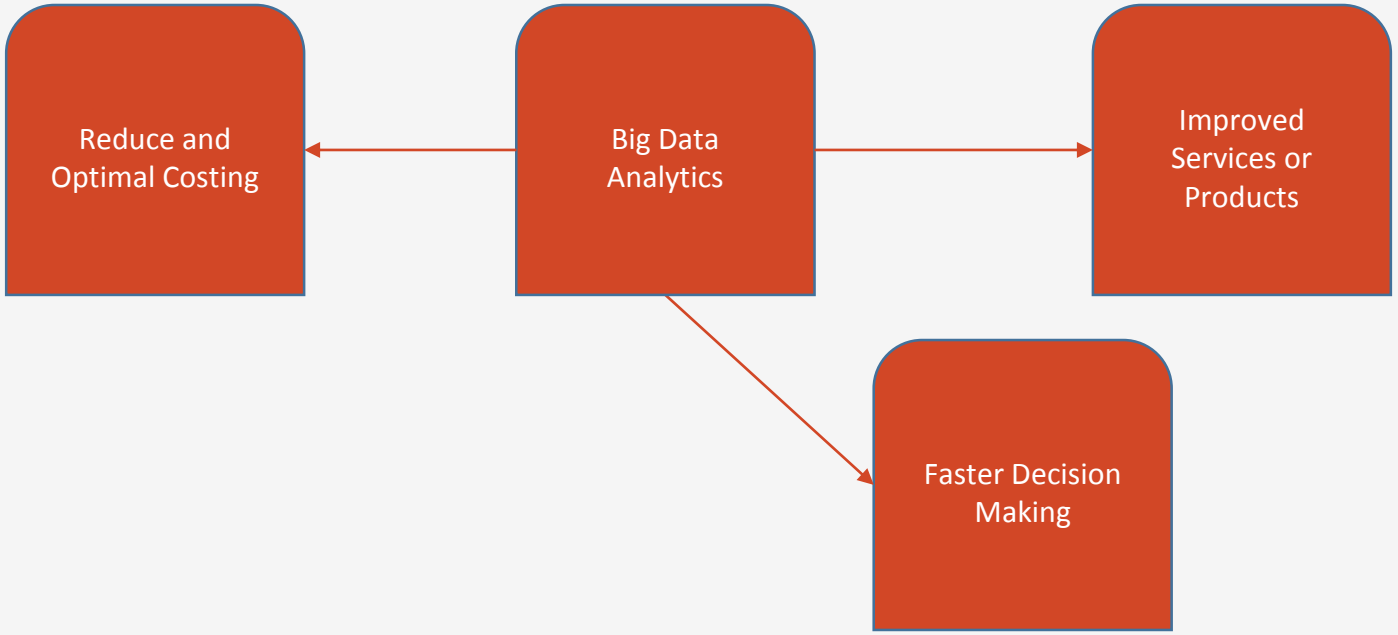
Introducing Hadoop

Chapter - 02

Big Data – A Problem

- Big Data which is a problem for Traditional Database management Systems like RDBMS but actually it is emerging as an opportunity for organizations. Now, organizations have realized that they are getting lots of benefits by Big Data Analytics, as you can see in the below image. They are examining large data sets to uncover all hidden patterns, unknown correlations, market trends, customer preferences and other useful business information.

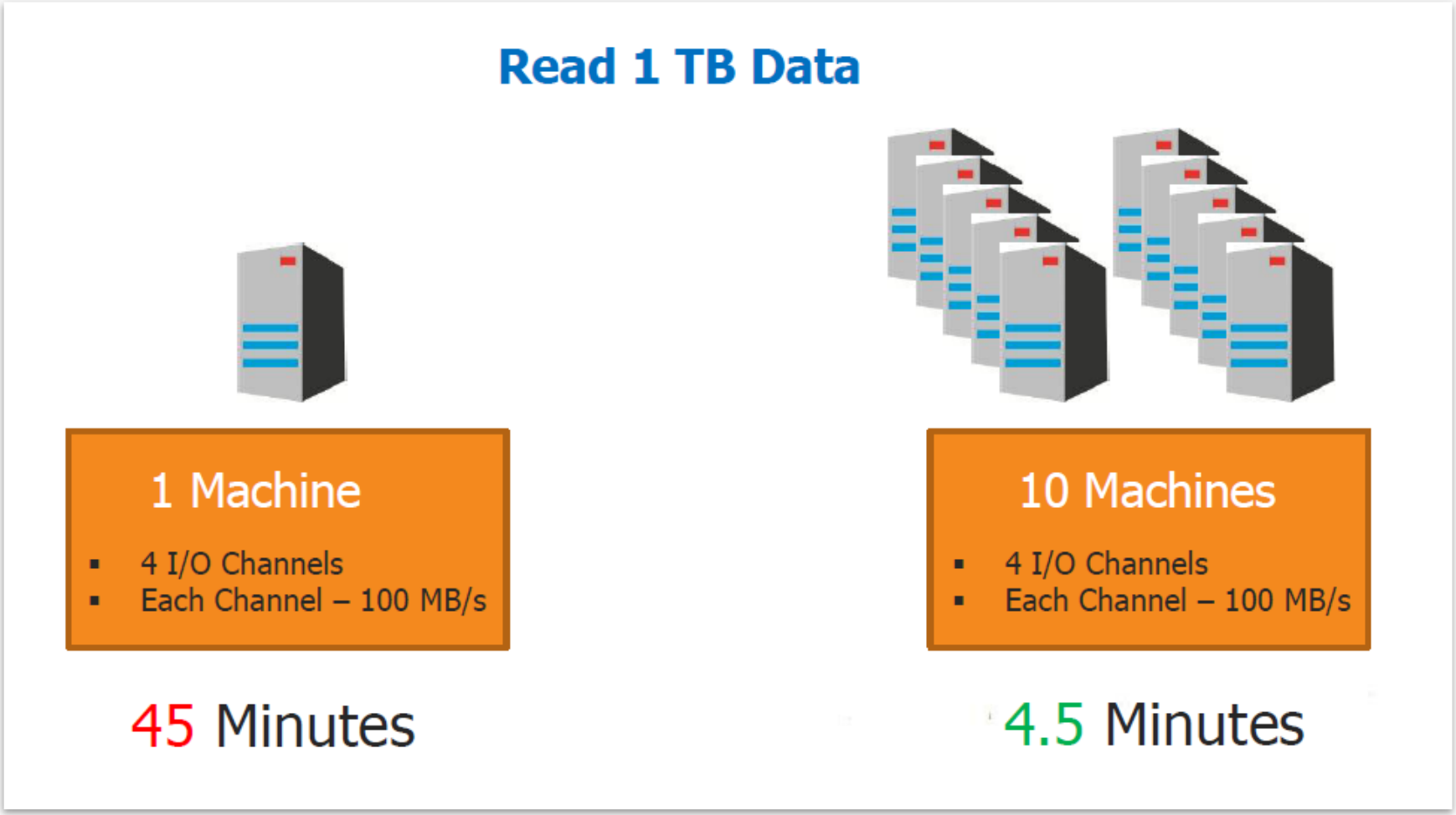
- These analytical findings are helping organizations in more effective marketing, new revenue opportunities, better customer service. They are improving operational efficiency, competitive advantages over rival organizations and other business benefits.



Problems for RDBMS :

- **Storing Large Amount of Data**
- **Storing Heterogeneous Data**
- **Processing data having Complex Structures**
- **Bringing huge amount of data to Computation Unit**

Traditional RDBMS Centralized vs Distributed Systems



Traditional RDBMS Centralized Vs Distributed Systems

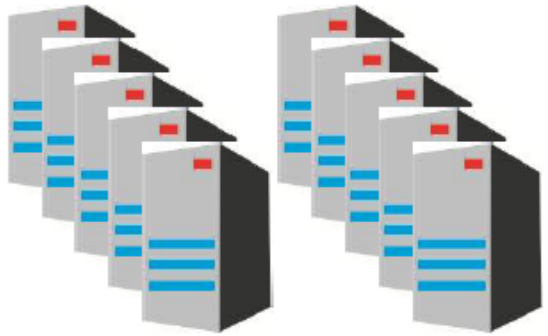
NOT USED

Read 1 TB Data

1 Machine

- 4 I/O Channels
- Each Channel – 100 MB/s

45 Minutes



10 Machines

- 4 I/O Channels
- Each Channel – 100 MB/s

4.5 Minutes

What is Hadoop?

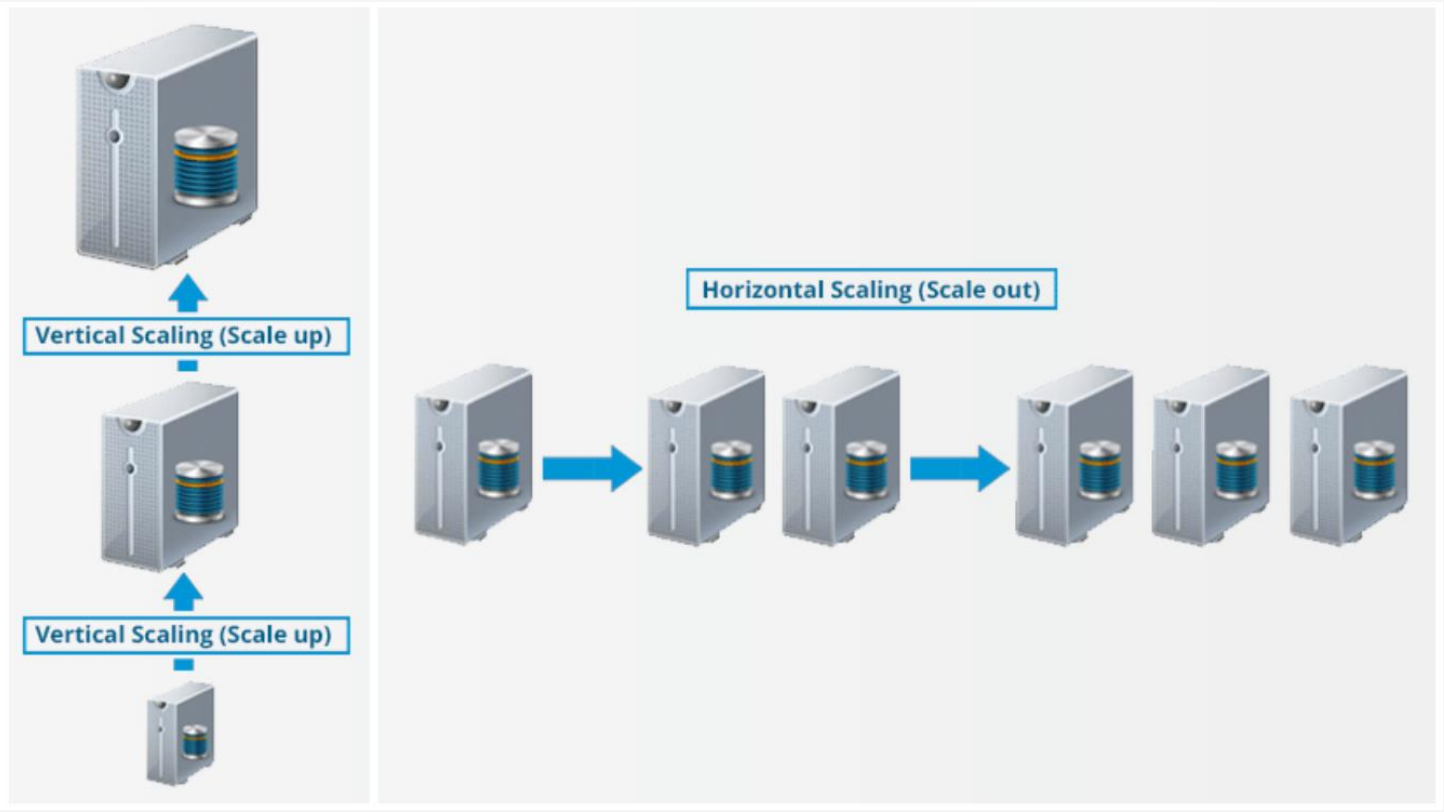
- A framework that is based on Google File System and Google Map Reduce
- Distributed computing platform that handles massive amount of data through parallelism
- Process petabytes of data
- Hadoop is platform independent since it runs on Java
- Applications built using HADOOP are run on large data sets distributed across clusters of commodity computers. Commodity computers are cheap and widely available. These are mainly useful for achieving greater computational power at low cost.
- Scales horizontally to thousands of CPU

What is Hadoop

- **Apache Hadoop** is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity hardware.

- Hadoop is the foundation of most big data architectures.

What is Hadoop



Hadoop-as-a-Solution to Big Data

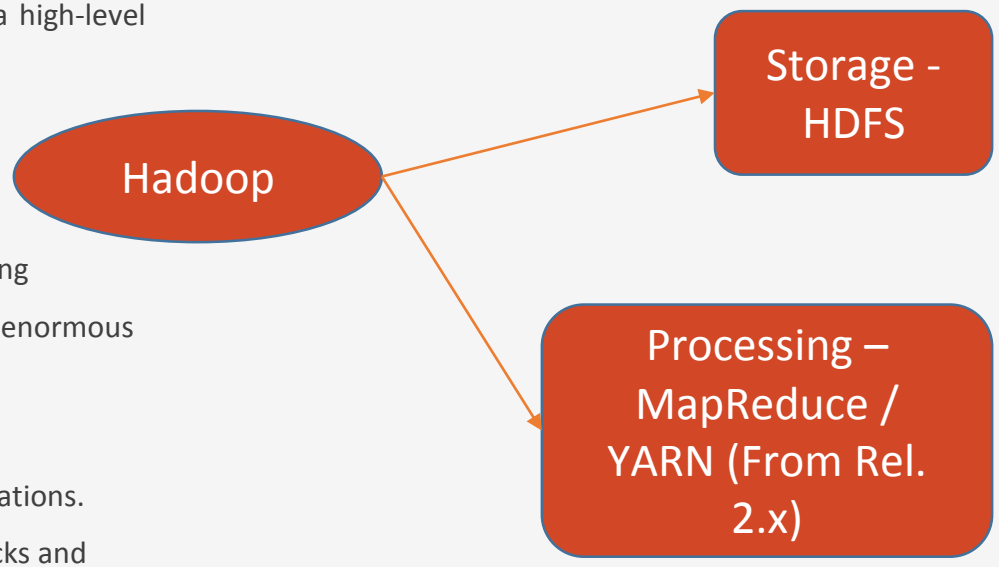
- Similar to data residing in a local file system of a personal computer system, in Hadoop, data resides in a distributed file system which is called as a **Hadoop Distributed File system**. The processing model is based on **'Data Locality'** concept wherein computational logic is sent to cluster nodes(server) containing data. This computational logic is nothing, but a compiled version of a program written in a high-level language such as Java. Such a program, processes data stored in Hadoop HDFS.

- Apache Hadoop consists of two sub-projects –

- **Hadoop MapReduce:** MapReduce is a computational model and software framework for writing applications which are run on Hadoop. These MapReduce programs are capable of processing enormous data in parallel on large clusters of computation nodes.

- **HDFS** (**Hadoop Distributed File System**): HDFS takes care of the storage part of Hadoop applications. MapReduce applications consume data from HDFS. HDFS creates multiple replicas of data blocks and distributes them on compute nodes in a cluster. This distribution enables reliable and extremely rapid computations.

- Although Hadoop is best known for MapReduce and its distributed file system- HDFS, the term is also used for a family of related projects that fall under the umbrella of distributed computing and large-scale data processing. This is known as **Hadoop Eco-System**



History of Hadoop

- Hadoop was started with **Doug Cutting and Mike Cafarella** in the year 2002 when they both started to work on Apache Nutch project. Apache Nutch project was the process of building a search engine system that can index 1 billion pages. After a lot of research on Nutch, they concluded that such a system will cost around half a million dollars in hardware, and along with a monthly running cost of $30, 000 approximately, which is very expensive. So, they realized that their project architecture will not be capable enough to the workaround with billions of pages on the web. So they were looking for a feasible solution which can reduce the implementation cost as well as the problem of storing and processing of large datasets.

LATER

- In **2003**, they came across a paper that described the architecture of Google's distributed file system, called **GFS (Google File System)** which was published by Google, for storing the large data sets. Now they realize that this paper can solve their problem of storing very large files which were being generated because of web crawling and indexing processes. But this paper was just the half solution to their problem.

- In **2004**, Google published one more paper on the technique **MapReduce**, which was the solution of processing those large datasets. Now this paper was another half solution for Doug Cutting and Mike Cafarella for their Nutch project. These both techniques (GFS & MapReduce) were just on white paper at Google. Google didn't implement these two techniques. Doug Cutting knew from his work on Apache Lucene (It is a free and open-source information retrieval software library, originally written in Java by Doug Cutting in 1999) that open-source is a great way to spread the technology to more people. So, together with Mike Cafarella, he started implementing Google's techniques (GFS & MapReduce) as open-source in the Apache Nutch project.

History of Hadoop

- In **2005**, Cutting found that Nutch is limited to only 20-to-40 node clusters. He soon realized two problems:

  (a) Nutch wouldn't achieve its potential until it ran reliably on the larger clusters

  LATER

  (b) And that was looking impossible with just two people (Doug Cutting & Mike Cafarella)

- The engineering task in Nutch project was much bigger than he realized. So he started to find a job with a company who is interested in investing in their efforts. And he found Yahoo!.Yahoo had a large team of engineers that was eager to work on this there project.

- So in **2006**, Doug Cutting joined Yahoo along with Nutch project. He wanted to provide the world with an open-source, reliable, scalable computing framework, with the help of Yahoo. So at Yahoo first, he separates the distributed computing parts from Nutch and formed a new project Hadoop **(He gave name Hadoop it was the name of a yellow toy elephant which was owned by the Doug Cutting's son. and it was easy to pronounce and was the unique word.)** Now he wanted to make Hadoop in such a way that it can work well on thousands of nodes. So with GFS and MapReduce, he started to work on Hadoop.

- In **2007**, Yahoo successfully tested Hadoop on a 1000 node cluster and start using it.

History of Hadoop

- 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.**

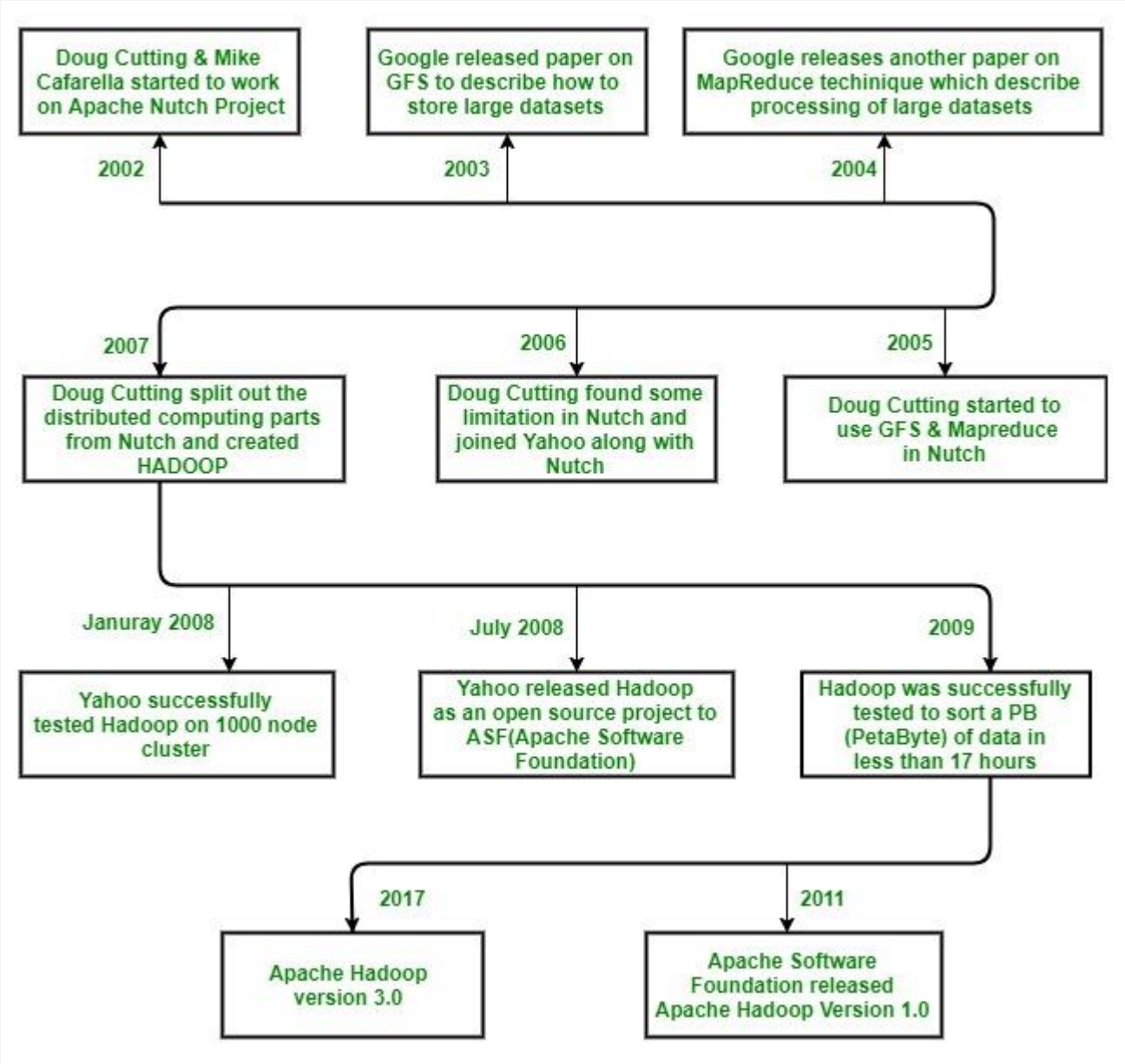
- In **2009,** Hadoop was successfully tested to sort a PB (PetaByte) of data in less than 17 hours for handling billions of searches and indexing millions of web pages. And Doug Cutting left the Yahoo and joined Cloudera to fulfill the challenge of spreading Hadoop to other industries.

- In **December of 2011,** Apache Software Foundation released Apache **Hadoop version 1.0**.

- And later in **Aug 2013, Version 2.0.6** was available.

- And currently, we have Apache Hadoop **Version 3.0** which released in **December 2017.**

Brief History on Hadoop



Milestones of Hadoop

- Jul 2008, Apache tested a 4000 node cluster with Hadoop successfully.
- 2009, Hadoop successfully sorted a petabyte of data in less than 17 hours to handle billions of searches and indexing millions of web pages.
- 2009 – Apache Hadoop completed terabyte sort in 62 seconds using its 1406 nodes cluster

Features Of Hadoop

- **Suitable for Big Data Analysis**
    - As Big Data tends to be distributed and unstructured in nature, HADOOP clusters are best suited for analysis of Big Data. Since it is processing logic (not the actual data) that flows to the computing nodes, less network bandwidth is consumed. This concept is called as data locality concept which helps increase the efficiency of Hadoop based applications.
- **Scalability**
    - HADOOP clusters can easily be scaled to any extent by adding additional cluster nodes and thus allows for the growth of Big Data. Also, scaling does not require modifications to application logic.
- **Cost Effective**
    - As it runs on commodity Hardware
- **Fault Tolerance**
    - HADOOP ecosystem has a provision to replicate the input data on to other cluster nodes. That way, in the event of a cluster node failure, data processing can still proceed by using data stored on another cluster node.

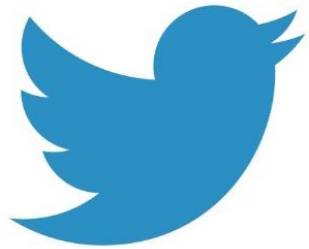
Organizations using Hadoop

In India

- **FLUTURA**
- **HECKYL**
- **Coviam technologies**
- **Cluster Foundry**
- **INC42 Media**

Around the World

- Amazon
- eBay
- Yahoo! Inc.
- Hortonworks
- Facebook
- Apple
- General Dynamics – IT
- EMC Corporation
- Northrop Grumman
- Twitter



Organizations using Hadoop

Hadoop in Facebook

- Messaging in facebook has been one of its popular feature since its inception.
- Another features of facebook such has like button or status updates are done in Mysql database but applications such as facebook messaging system runs on the top of HBASE which is Hadoop's NoSql database framework.
- The data warehousing solution of facebook's lies in HIVE which is built on the top of HDFS.
- The reporting needs of the FACEBOOK is also achieved by using HIVE.
- Post 2011 with increase in the magnitude of data and to improve the efficiency facebook started implementing apache corona which works very much like Yarn framework.
- In apache corona, the a new scheduling framework is used which separates cluster resource management from job coordination.

Organizations using Hadoop

Hadoop in Yahoo!

- When it comes about the size of the hadoop cluster,yahoo beats all by having the 42000 nodes in about 20 YARN (aka MapReduce 2.0)clusters with 600 petabytes of data on HDFS to serve the company's mobile, search, advertising, personalization, media, and communication efforts.
- Yahoo uses hadoop to block around 20.5 billion messages and checks it to enter it into its email server. Yahoo's spam detection abilities has increased to manifolds since it started using Hadoop.
- In the ever growing family of hadoop,yahoo has been one of the major contributor.
- Yahoo has been the pioneer of many new technologies which have already embraced itself into hadoop ecosystem.
- Few notable technologies which yahoo has been using apart from mapreduce and hdfs is Apache tez and spark.
- One of the main vehicle of yahoo's hadoop chariot is pig which started in yahoo and it still tops the chart as 50-60 percent of jobs are processed using pig scripts.

Organizations using Hadoop

Hadoop in Health care companies:

- **Hadoop in Cancer treatment:**

- The response of patients having same type of cancer is different for same cancer medicine and this is because of unique individual genome.

- Each person's genome contains around 1.5 Gigabytes of the data and to understand how a particular drug responds to a particular genome requires the genomic data to be stored and combined with other data like demographics and trial outcomes and finally an analysis to be done to know which medicine is suitable for which kind of genetic spectrum.

- Many top cancer research institutes have applied this hadoop technology to elevate the success rate of their cancer treatments.

- **Hadoop in checking re-occurrence of heart cardiac attack:**

- UC Irvine Health in USA while discharging heart patients is equipping them with a wireless scale so that weight measured by them in home could be transferred automatically and wirelessly to the Hadoop cluster established in the hospital inside which hadoop algorithm running determines a chance for recurrence for heart attack by analyzing the risk factor associated with

Organizations using Hadoop

Hadoop in Telecom industries:

- Telecommunication sector is one of the most data driven industry.
- Apart from processing millions of call per seconds it is also providing services for web browsing,videos,television,streaming music,movies,text messages and email.
- All these sources have flooded the telecom companies with drastic increase in the data due to which storing and process overhead have increased manifolds.
- Some of the case studies related to implementation of Hadoop in telecom sectors has been discussed below:
- **Analyzing call data records**
- To reduce the rate of call drop and improve the sound quality,the call details pouring in to the company's database in real time has to be analyzed to maximum precision.
- Telecom companies have been using tools like Flume to ingest the millions of call records per second into hadoop and then using Apache storm for processing them in real time to identify the troubling patterns.
- **Timely servicing of the equipments**
- Replacing the equipments from transmission tower of telecom companies is very much costlier than the repairing.
- To determine an optimum schedule for maintenance(not too early,not too late),hadoop has been used by the companies for storing unstructured, sensor and streaming data.
- Machine learning algorithms are applied on these data to reduce maintenance cost and to do timely repair of the equipments before it gets any problem.

Organizations using Hadoop

Hadoop in Financial Sectors:

- Companies in the financial sectors have been using hadoop to do deeper analysis on the data to improve operational margins and to detect the malicious activities which gets unnoticed in the normal scenario.
- Some of the case studies which are in practice in financial sectors are as follows:
- **Anti money laundering practice**
- Before Hadoop,finance companies used to follow the approach where selective storing of the data used to take place by discarding historical data due to storage limitations.
- So the sample data available for analytics was not suffice to give a full proof results which could be used to check money laundering.
- But now companies have been using hadoop framework for greater storing and processing abilities and to determine the sources of black money and keep it out of the system.
- Companies are now able to manage millions of customer names and their transactions in real time and the rate of detecting the suspicious transactions have increased drastically after implementing hadoop ecosystem.

Organizations using Hadoop

Hadoop in Banks

- Many banks across the world have been using Hadoop platform to collect and analyze all the data pertaining to their customers like daily transactional data,data coming from interaction from multiple customer touch points like call centers, home value data and merchant records.

- All these data can be analyzed by banks to segregate customers into one or more sections based on their needs in terms of banking product and services,their sales,promotion and marketing accordingly.

- Using Big data Hadoop architecture , many credit card issuing banks has been implementing fraud detection system which detects

- Suspicious activity by analyzing one’s past history with spending patterns and trends and have been disabling the cards of the suspects.

Major Vendors of Hadoop

- Cloudera
- Hortonworks
- MapR