Review on Big Data

There is no place where Big Data does not exist! The curiosity about what is Big Data has been soaring in the past few years. Let me tell you some mind-boggling facts! Forbes reports that every minute, users watch *4.15 million YouTube videos*, send *456,000 tweets* on Twitter, post *46,740 photos* on Instagram and there are *510,000 comments* posted and *293,000 statuses* updated on Facebook!

Just imagine the huge chunk of data that is produced with such activities. This constant creation of data using social media, business applications, telecom and various other domains is leading to the formation of Big Data.

Evolution of Big Data

Before exploring any further, let me begin by giving some insight into why the this technology has gained so much importance.

When was the last time you guys remember using a floppy or a CD to store your data? Let me guess, had to go way back in the early 21st century right? The use of manual paper records, files, floppy and discs have now become obsolete. The reason for this is the exponential growth of data. People began storing their data in relational database systems but with the hunger for new inventions, technologies, applications with quick response time and with the introduction of the internet, even that is insufficient now. This generation of continuous and massive data can be referred to as Big Data. There are a few other factors that characterize Big Data which I will be explaining later in this blog.

Forbes reports that there are 2.5 quintillion bytes of data created each day at our current pace, but that pace is only accelerating. Internet of Things(IoT) is one such technology which plays a major role in this acceleration. 90% of all data today was generated in the last two years.

What is Big Data?

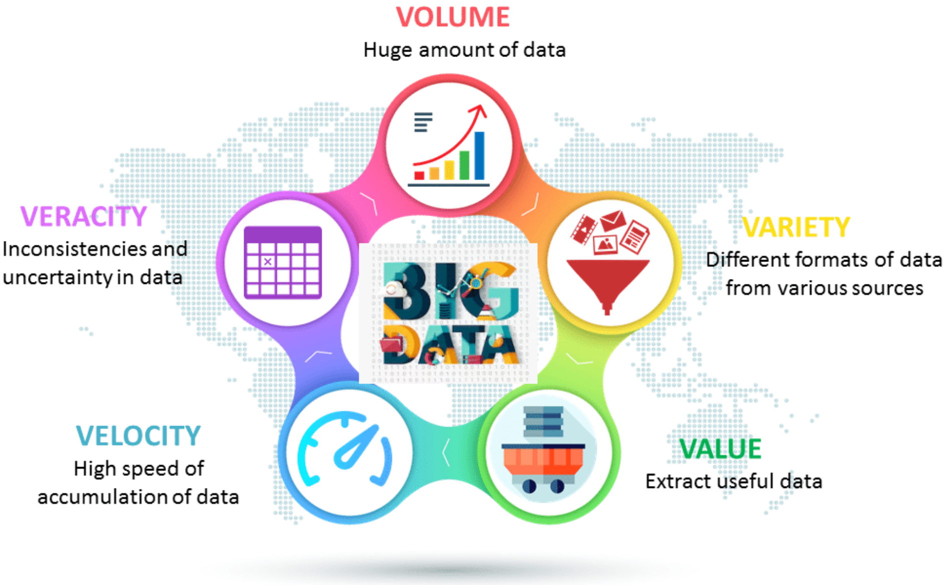
So before I explain what is Big Data, let me also tell you what it is not! The most common myth associated with it is that it is just about the size or volume of data. But actually, it’s not just about the “big” amounts of data being collected. ***Big Data*** refers to the large amounts of data which is pouring in from various data sources and has different formats. Even previously there was huge data which were being stored in databases, but because of the varied nature of this Data, the traditional relational database systems are incapable of handling this Data. Big Data is much more than a collection of datasets with different formats, it is an important asset which can be used to obtain enumerable benefits.

The three different formats of big data are:

1. ***Structured:*** Organised data format with a fixed schema. Ex: RDBMS
2. ***Semi-Structured:*** Partially organised data which does not have a fixed format. Ex: XML, JSON
3. ***Unstructured:*** Unorganised data with an unknown schema. Ex: Audio, video files etc.

Characteristics of Big Data

Following are the characteristics:


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

The above image depicts the five V's of Big Data but as and when the data keeps evolving so will the V's. I am listing five more V's which have developed gradually over time:

- Validity: correctness of data
- Variability: dynamic behaviour
- Volatility: tendency to change in time
- Vulnerability: vulnerable to breach or attacks
- Visualization: visualizing meaningful usage of data

(we covered more than 5 Vs, but these are the most important to remember)

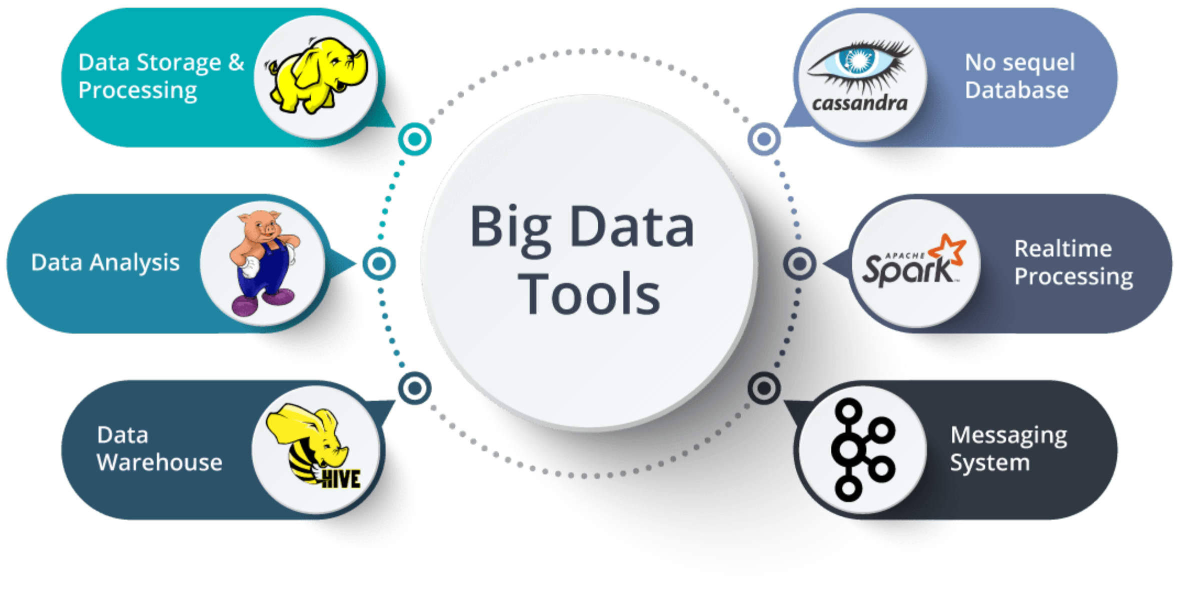
Big Data Analytics

Now that I have told you what is Big Data and how it's being generated exponentially, let me present to you a very interesting example of how *Starbucks*, one of the leading coffeehouse chain is making use of this Big Data.

I came across this article by Forbes which reported how *Starbucks* made use of this technology to analyse the preferences of their customers to enhance and personalize their experience. They analysed their member's coffee buying habits along with their preferred drinks to what time of day they are usually ordering. So, even when people visit a "new" Starbucks location, that store's point-of-sale system is able to identify the customer through their smartphone and give the barista their preferred order. In addition, based on ordering preferences, their app will suggest new products that the customers might be interested in trying. This my friends is what we call Big Data Analytics.

Basically, it is largely used by companies to facilitate their growth and development. This majorly involves applying various data mining algorithms on the given set of data, which will then aid them in better decision making.

There are multiple tools for processing Big Data such as ***Hadoop***, ***Pig***, ***Hive***, ***Cassandra***, ***Spark***, ***Kafka***, etc. depending upon the requirement of the organisation.



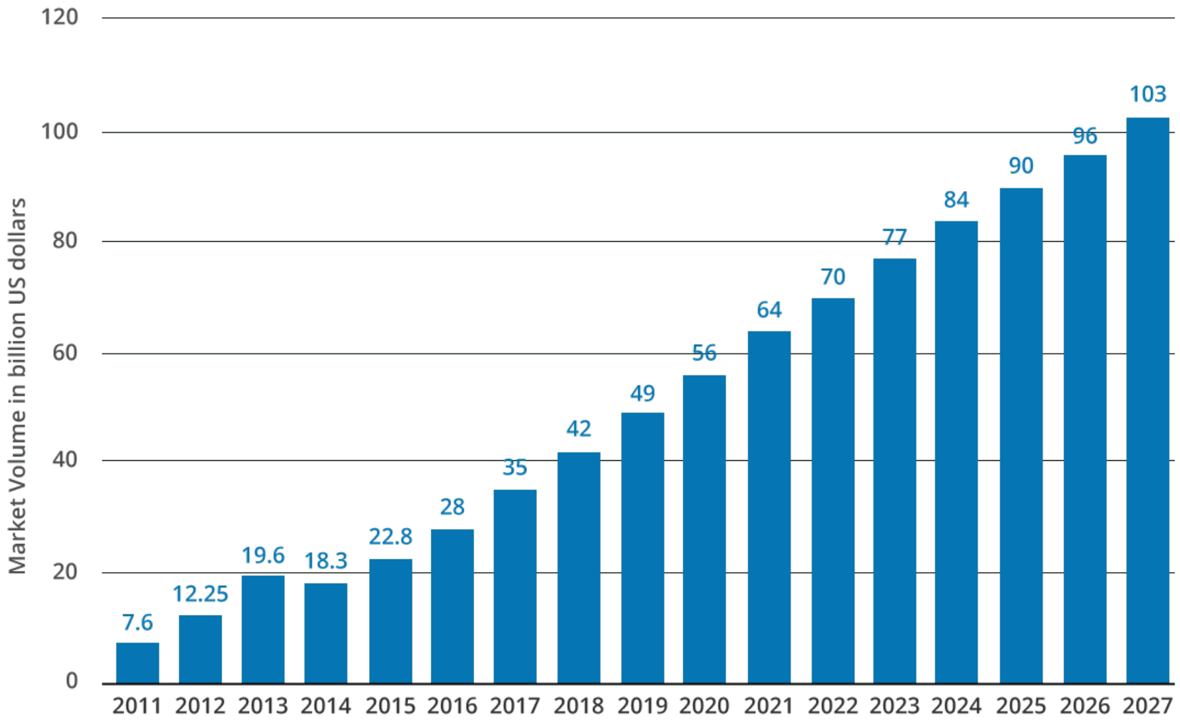
Big Data Applications

These are some of the following domains where ***Big Data Applications*** has been revolutionized:

- **Entertainment:** Netflix and Amazon use it to make shows and movie recommendations to their users.
- **Insurance:** Uses this technology to predict illness, accidents and price their products accordingly.
- **Driver-less Cars:** Google's driver-less cars collect about one gigabyte of data per second. These experiments require more and more data for their successful execution.
- **Education:** Opting for big data powered technology as a learning tool instead of traditional lecture methods, which enhanced the learning of students as well aided the teacher to track their performance better.
- **Automobile:** Rolls Royce has embraced this technology by fitting hundreds of sensors into its engines and propulsion systems, which record every tiny detail about their operation. The changes in data in real-time are reported to engineers who will decide the best course of action such as scheduling maintenance or dispatching engineering teams should the problem require it.
- **Government:** A very interesting use case is in the field of politics to analyse patterns and influence election results. Cambridge Analytica Ltd. is one such organisation which completely drives on data to change audience behaviour and plays a major role in the electoral process.

Scope of Big Data

- **Numerous Job opportunities:** The career opportunities pertaining to the field of Big data include, Big Data Analyst, Big Data Engineer, Big Data solution architect etc. According to IBM, 59% of all Data Science and Analytics (DSA) job demand is in Finance and Insurance, Professional Services, and IT.
- **Rising demand for Analytics Professional:** An article by Forbes reveals that "IBM predicts demand for Data Scientists will soar by 28%". By 2020, the number of jobs for all US data professionals will increase by 364,000 openings to 2,720,000 according to IBM.
- **Salary Aspects:** Forbes reported that employers are willing to pay a premium of $8,736 above median bachelor's and graduate-level salaries, with successful applicants earning a starting salary of $80,265
- **Adoption of Big Data analytics:** Immense growth in the usage of big data analysis across the world.



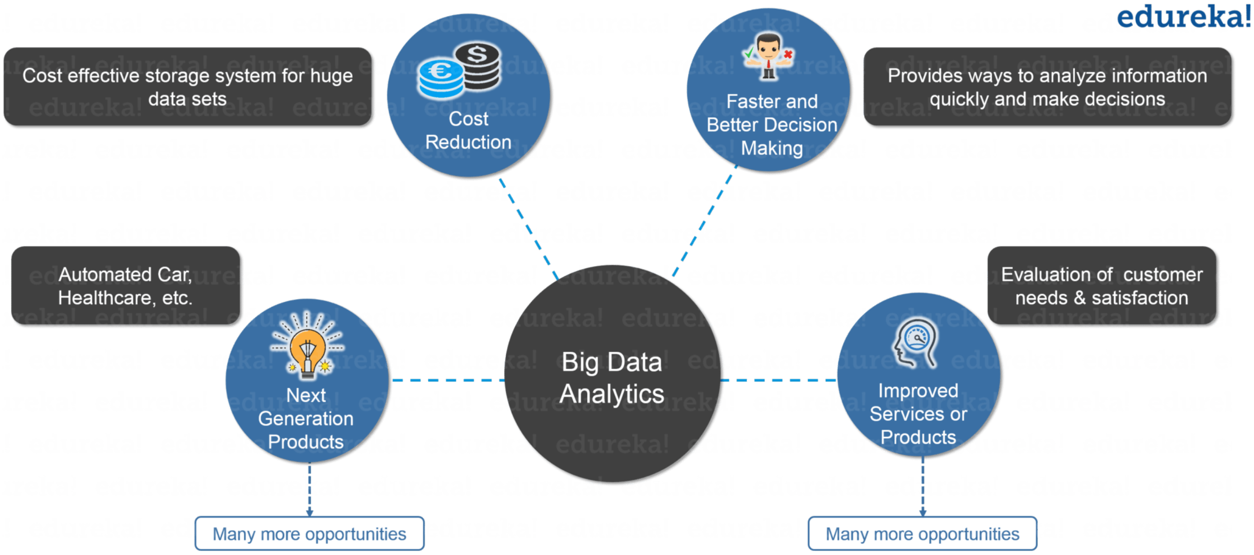
The above image depicts the growing market revenue of Big Data in billion U.S. dollars from the year 2011 to 2027.

Hadoop Review

As to understand what exactly is Hadoop, we have to first understand the issues related to Big Data and the traditional processing system. Advancing ahead, we will discuss what is Hadoop, and how Hadoop is a solution to the problems associated with Big Data. We will also look at the CERN case study to highlight the benefits of using Hadoop.

Big Data 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.

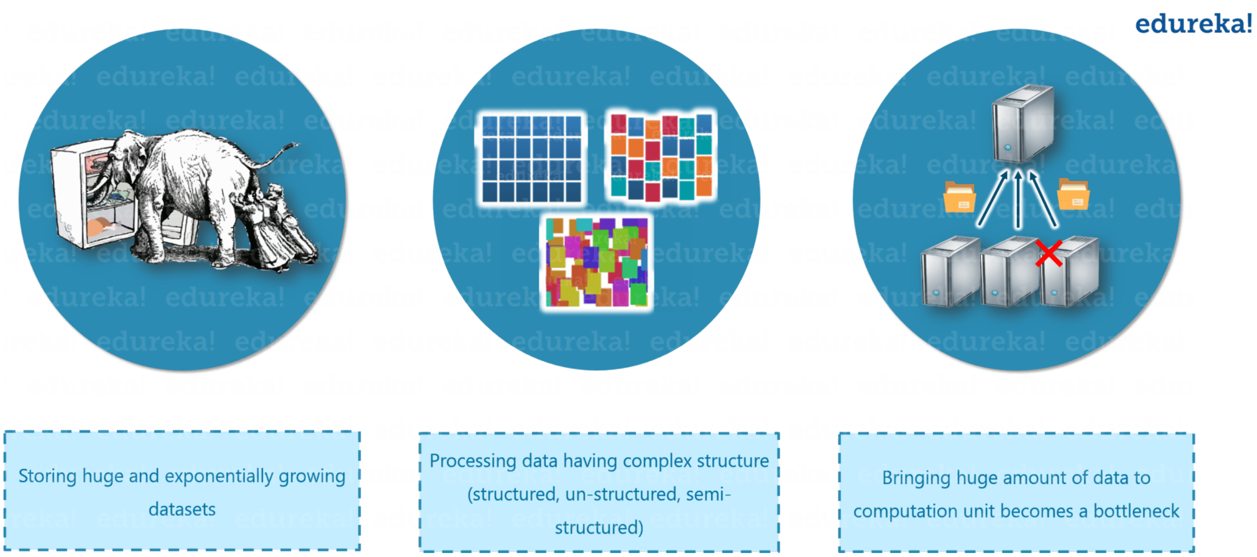

edureka!
Cost effective storage system for huge data sets
Cost Reduction
Faster and Better Decision Making
Provides ways to analyze information quickly and make decisions
Automated Car, Healthcare, etc.
Next Generation Products
Big Data Analytics
Improved Services or Products
Evaluation of customer needs & satisfaction
Many more opportunities
Many more opportunities

So, let us move ahead and know the problems associated with traditional approach in en-cashing Big data opportunities.

Problems with Traditional Approach

In traditional approach, the main issue was handling the heterogeneity of data i.e. structured, semi-structured and unstructured. The RDBMS focuses mostly on structured data like banking transaction, operational data etc. and Hadoop specializes in semi-structured, unstructured data like text, videos, audios, Facebook posts, logs, etc. RDBMS technology is a proven, highly consistent, matured systems supported by many companies. While on the other hand, Hadoop is in demand due to Big Data, which mostly consists of unstructured data in different formats.

Now let us understand what are the major problems associated with Big Data. So that, moving ahead we can understand how Hadoop emerged as a solution.

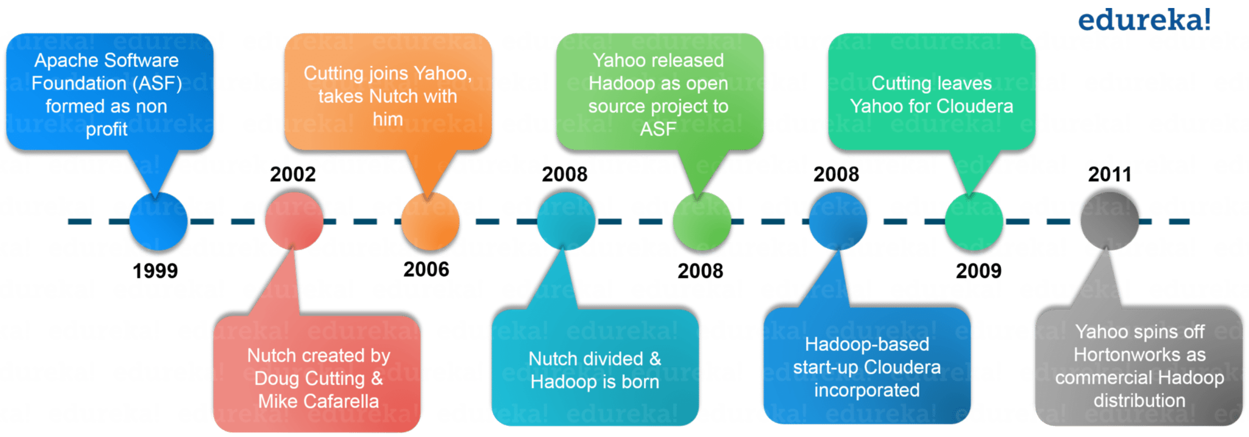

edureka!
Storing huge and exponentially growing datasets
Processing data having complex structure (structured, un-structured, semi-structured)
Bringing huge amount of data to computation unit becomes a bottleneck

The first problem is storing the colossal amount of data.
Storing this huge data in a traditional system is not possible. The reason is obvious, the storage will be limited only to one system and the data is increasing at a tremendous rate.

Second problem is storing heterogeneous data.
Now, we know storing is a problem, but let me tell you, it is just a part of the problem. Since we discussed that the data is not only huge, but it is present in various formats as well like: Unstructured, Semi-structured and Structured. So, you need to make sure that, you have a system to store all these varieties of data, generated from various sources.

*Third problem is accessing and processing speed***.**
The hard disk capacity is increasing but the disk transfer speed or the access speed is not increasing at similar rate. Let me explain you this with an example: If you have only one 100 Mbps I/O channel and you are processing 1TB of data, it will take around 2.91 hours. Now, if you have four machines with one I/O channel, for the same amount of data it will take 43 minutes approx. Thus, accessing and processing speed is the bigger problem than storing Big Data.

Before understanding what is Hadoop, let us first look at the evolution of Hadoop over the period of time.


edureka!
Apache Software Foundation (ASF) formed as non profit
1999
2002
Nutch created by Doug Cutting & Mike Cafarella
Cutting joins Yahoo, takes Nutch with him
2006
2008
Nutch divided & Hadoop is born
Yahoo released Hadoop as open source project to ASF
2008
2008
Hadoop-based start-up Cloudera incorporated
Cutting leaves Yahoo for Cloudera
2009
2011
Yahoo spins off Hortonworks as commercial Hadoop distribution

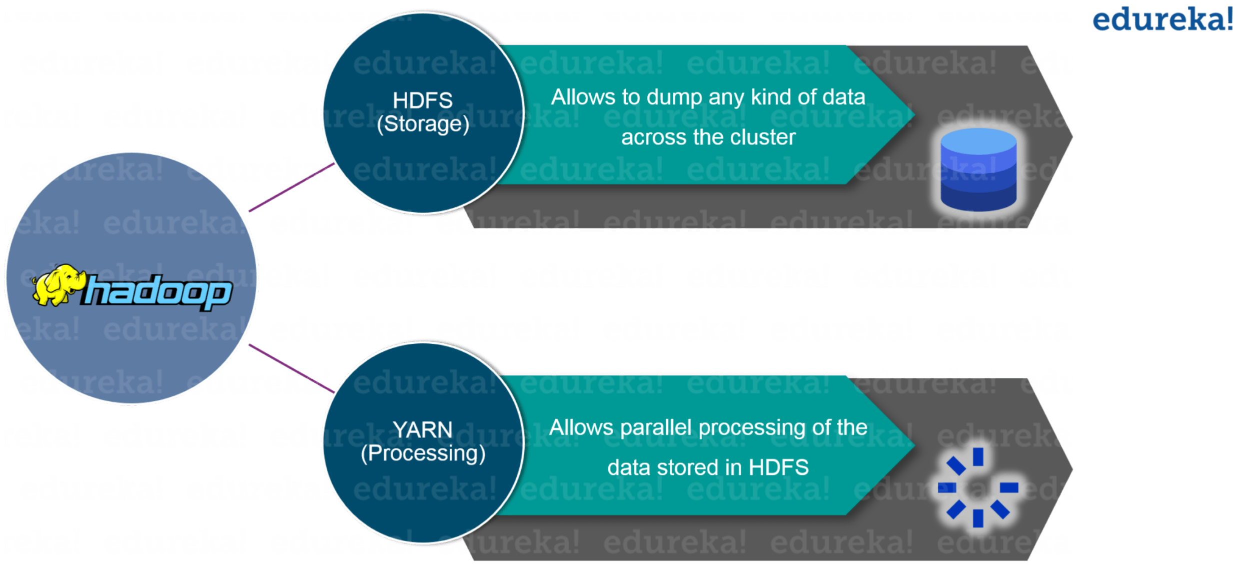
In 2003, Doug Cutting launches project Nutch to handle billions of searches and indexing millions of web pages. Later in Oct 2003 – Google releases papers with GFS (Google File System). In Dec 2004, Google releases papers with MapReduce. In 2005, Nutch used GFS and MapReduce to perform operations. In 2006, Yahoo created Hadoop based on GFS and MapReduce with Doug Cutting and team. You would be surprised if I would tell you that, in 2007 Yahoo started using Hadoop on a 1000 node cluster.

Later in Jan 2008, Yahoo released Hadoop as an open source project to Apache Software Foundation. In Jul 2008, Apache tested a 4000 node cluster with Hadoop successfully. In 2009, Hadoop successfully sorted a petabyte of data in less than 17 hours to handle billions of searches and indexing millions of web pages. Moving ahead in Dec 2011, Apache Hadoop released version 1.0. Later in Aug 2013, Version 2.0.6 was available.

When we were discussing about the problems, we saw that a distributed system can be a solution and Hadoop provides the same. Now, let us understand what is Hadoop.

What is Hadoop?

Hadoop is a framework that allows you to first store Big Data in a distributed environment, so that, you can process it parallelly. There are basically two components in Hadoop:

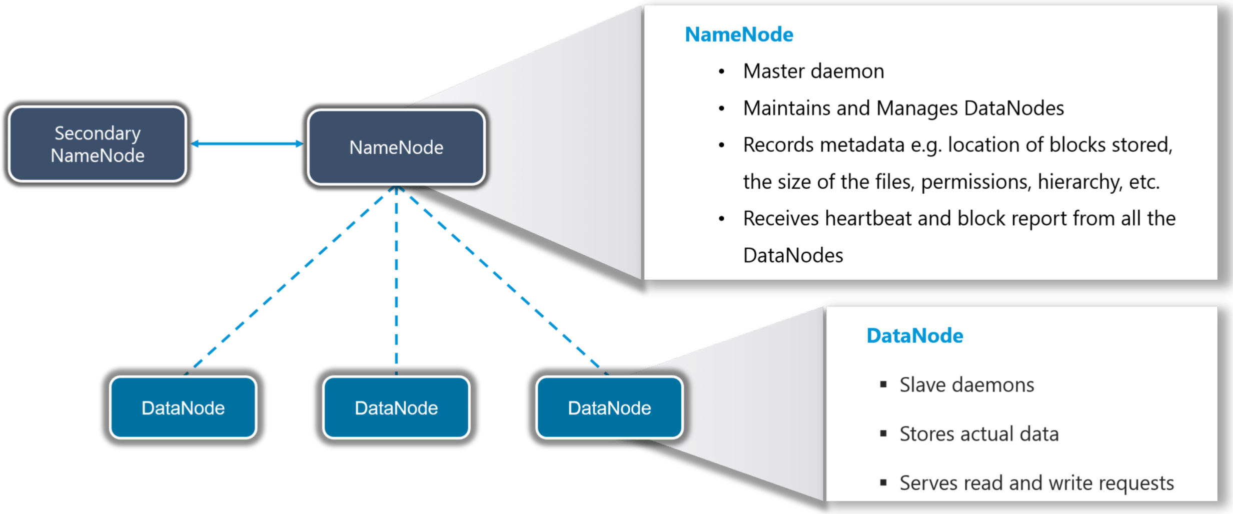

edureka!
hadoop
HDFS
(Storage)
Allows to dump any kind of data across the cluster
YARN
(Processing)
Allows parallel processing of the data stored in HDFS

he first one is ***HDFS*** for storage (Hadoop distributed File System), that allows you to store data of various formats across a cluster. The second one is ***YARN***, for resource management in Hadoop. It allows parallel processing over the data, i.e. stored across HDFS.

Let us first understand HDFS.

HDFS

HDFS creates an abstraction, let me simplify it for you. Similar as virtualization, you can see HDFS logically as a single unit for storing Big Data, but actually you are storing your data across multiple nodes in a distributed fashion. HDFS follows master-slave architecture.



NameNode

- Master daemon
- Maintains and Manages DataNodes
- Records metadata e.g. location of blocks stored, the size of the files, permissions, hierarchy, etc.
- Receives heartbeat and block report from all the DataNodes

DataNode

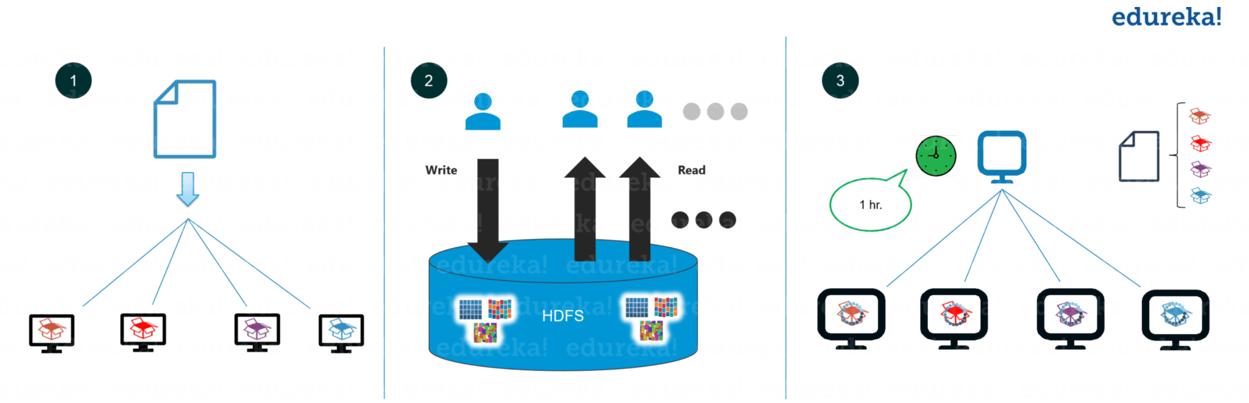
- Slave daemons
- Stores actual data
- Serves read and write requests

In HDFS, Namenode is the master node and Datanodes are the slaves. Namenode contains the metadata about the data stored in Data nodes, such as which data block is stored in which data node, where are the replications of the data block kept etc. The actual data is stored in Data Nodes.

I also want to add, we actually replicate the data blocks present in Data Nodes, and the default replication factor is 3. Since we are using commodity hardware and we know the failure rate of these hardwares are pretty high, so if one of the DataNodes fails, HDFS will still have the copy of those lost data blocks. You can also configure replication factor based on your requirements.

Hadoop-as-a-Solution

Let's understand how Hadoop provided the solution to the Big Data problems that we just discussed.



*The first problem is storing Big data***.**

HDFS provides a distributed way to store Big data. Your data is stored in blocks across the DataNodes and you can specify the size of blocks. Basically, if you have 512MB of data and you have configured HDFS such that, it will create 128 MB of data blocks. So HDFS will divide data into 4 blocks as 512/128=4 and store it across different DataNodes, it will also replicate the data blocks on different DataNodes. Now, as we are using commodity hardware, hence storing is not a challenge.

It also solves the scaling problem. It focuses on ***horizontal scaling*** instead of vertical scaling. You can always add some extra data nodes to HDFS cluster as and when required, instead of scaling up the resources of your DataNodes. Let me summarize it for you basically for storing 1 TB of data, you don't need a 1TB system. You can instead do it on multiple 128GB systems or even less.

*Next problem was storing the variety of data***.**

With HDFS you can store all kinds of data whether it is structured, semi-structured or unstructured. Since in HDFS, there is *no pre-dumping schema validation*. And it also follows write once and read many model. Due to this, you can just write the data once and you can read it multiple times for finding insights.

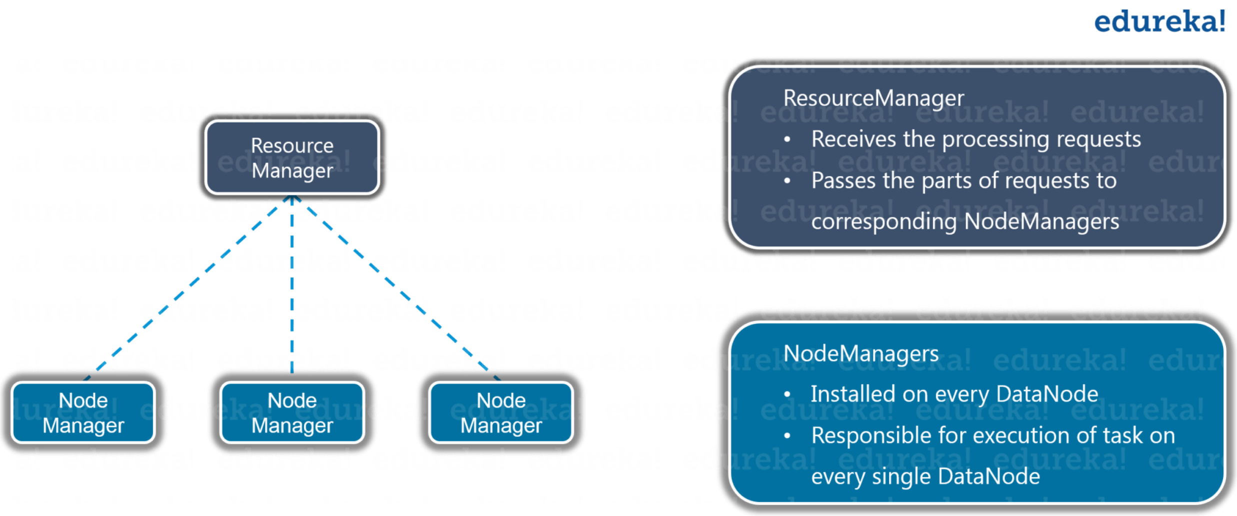
*T*hird challenge was accessing & processing the data faster**.**

Yes, this is one of the major challenges with Big Data. In order to solve it, we move processing to data and not data to processing. What does it mean? Instead of moving data to the master node and then processing it. In MapReduce, the processing logic is sent to the various slave nodes & then data is processed parallely across different slave nodes. Then the processed results are sent to the master node where the results is merged and the response is sent back to the client.

In YARN architecture, we have ResourceManager and NodeManager. ResourceManager might or might not be configured on the same machine as NameNode. But, NodeManagers should be configured on the same machine where DataNodes are present.

YARN

YARN performs all your processing activities by allocating resources and scheduling tasks.



ResourceManager
- Receives the processing requests
- Passes the parts of requests to corresponding NodeManagers

NodeManagers
- Installed on every DataNode
- Responsible for execution of task on every single DataNode

It has two major components, i.e. ResourceManager and NodeManager.

ResourceManager is again a master node. It receives the processing requests and then passes the parts of requests to corresponding NodeManagers accordingly, where the actual processing takes place. NodeManagers are installed on every DataNode. It is responsible for the execution of the task on every single DataNode.

I hope now you are clear with What is Hadoop and its major components. Let us move ahead and understand when to use and when not to use Hadoop.

Where is Hadoop used?

Hadoop is used for:

- *Search* – Yahoo, Amazon, Zvents
- *Log processing* – Facebook, Yahoo
- *Data Warehouse* – Facebook, AOL
- *Video and Image Analysis* – New York Times, Eyealike

Till now, we have seen how Hadoop has made Big Data handling possible. But there are some scenarios where Hadoop implementation is not recommended.