**Big Data Technology**

**The elephant in the room: Hadoop's parallel world**

* There are many Big Data technologies that have been making an impacton the new technology stacks for handling Big Data, but Apache Hadoop is one technology that has been the darling of Big Data talk. Hadoop is an open-source platform for storage and processing of diverse data types that enables data-driven enterprises to rapidly derive the complete value from all their data. He explained the history and overview of Hadoop to us: The original creators of Hadoop are Doug Cutting (used to be at Yahoo! now at Cloudera). Doug and Mike were building a project called “Nutch” with the goal of creating a large Web index. They saw the MapReduce and GFS papers from Google, which were obviously super relevant to the problem. They integrated the concepts from MapReduce and GFS into Nutch; then later these two components were pulled out to form the genesis of the Hadoop project. The name “Hadoop” itself comes from Doug ’s son, he just made the word up for a yellow plush elephant toy that he has.
* Hadoop gives organizations the flexibility to ask questions across their structured and unstructured data that were previously impossible to ask or solve
* T he scale and variety of data have permanently overwhelmed the ability to cost-effectively extract value using traditional platforms.
* The scalability and elasticity of free, open-source Hadoop running on standard hardware **allow organizations to hold onto more data than ever before**.
* Hadoop handles a variety of workloads, including **search, log processing,recommendation systems, data warehousing, and video/image analysis**.
* Apache Hadoop is an open-source project **administered by the Apache Software Foundation**. The software was **originally developed by the world ’s largest Internet companies** to capture and analyze the data that they generate.

**The TWO CRITICAL COMPONENTS OF HADOOP are:**

* **The Hadoop Distributed File System (HDFS):** HDFS is the storage system for a Hadoop cluster.
* **MapReduce:** Because Hadoop stores the entire dataset in small pieces across a collection of servers, analytical jobs can be distributed, in parallel, to each of the servers storing part of the data

**~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~**

**OLD VS NEW APPROACHES TO DATA ANALYTICS**

|  |  |
| --- | --- |
| **Old Approach (Database approach)** | **New Approach (Big data Analytics)** |
| Follows data and analytics technology stack with different layers of cross- communicating data and working on  ―scale-up‖ expensive hardware. | Follows data and analytics platform that does all the data processing and analytics in one layer without moving data back and forth on cheap but scalable (―scale-out‖) commodity hardware. |
| Data is moved to places where they have to be processed. | Data must be processed and converted into usable business intelligence where it sits. |
| Massive parallel processing was not employed due to hardware and storage limitations. | Hardware and storage is affordable and continuing to get cheaper to enable massive parallel processing. |
| Due to technological limitations storing, managing and analyzing massive data sets were difficult. | New proprietary technologies and open source inventions enable different approaches that make it easier and more affordable to store, manage & analyze data. |
| Not able to handle unstructured data. | The variety of data and ability to handle unstructured data is on the rise. Big data approach provides solution to this. |

**~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~**

**Open Source Technology for Big Data Analytics**

* **OPEN- SOURCE SOFTWARE** is computer software that is available in source code form under an open- source license that permits users to study, change, improve and distribute the software. **Hadoop** is a open- source project.
* One of the key attributes of open- source projects is that it is not constrained by someone else’s predetermined ideas or vision which makes it flexible, extensible and low cost.
* **ONE DISADVANTAGE OF OPEN**-

Source is that it has to coexist with the proprietary solution for a long time for many reasons.

For example,

Getting data from hadoop to a database required a hadoop expert in the middle to do the data cleansing and the data type translation .If the data was not 100% (clean which is the case with most circumstances) a developer was needed to get it to a consistent, proper form. Besides wasting the valuable time of that expert, this process meant that business analysts couldn’t directly access and analyze data in hadoop clusters. **SQL-H IS SOFTWARE THAT IS DEVELOPED TO SOLVE THIS PROBLEM.**

**~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~**

### The Cloud and Big Data

* ~~Market economics are demanding that capital- intensive infrastructure costs disappear and business challenges are forcing clients to consider newer models.The cloud-deployment model satisfies such needs. With a cloud model, payment is on subscription basis with no capital expense.Typical 30% maintenance fees are not incurred and all the updates on the platform are automatically available. The traditional cost of value chains is completely eliminated by massively scalable platforms (such as cloud) where marginal cost to deliver an incremental product/service is zero. Whether a private hosted model or a publicly shared one, the true value lies in delivering software, data and/or analytics in an ―as a service’ model.~~

**~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~**

**MOBILE BUSINESS INTELLIGENCE**

* **Simplicity** and **ease of use** had been the **major barriers to BI adoption**. But mobile device have made complicated actions to be performed very easily.
  + For example, a *young child can use an ipad or iphone easily but not a laptop*. This ease of use will drive the wide adoption of mobile BI.
* Multi touch and software oriented devices have brought mobile analytics and intelligence to a much wider audience.
* Ease of mobile application development has also contributed to the wide adoption of mobile BI.

### Three elements that have impacted the viability of mobile BI are-

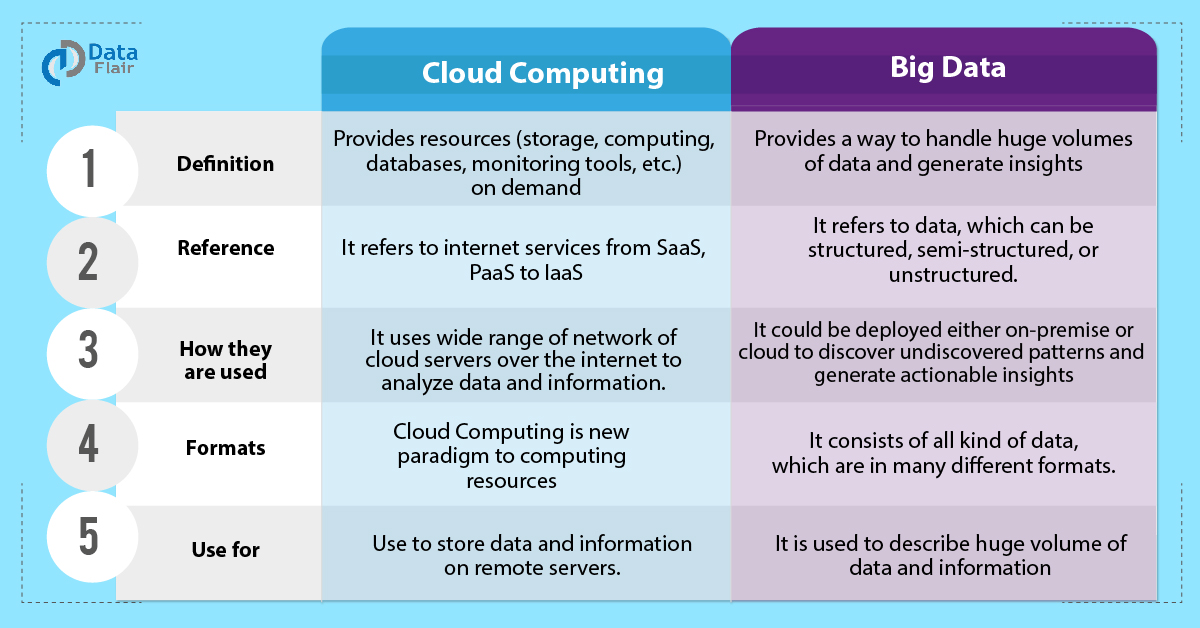
1. Location-GPS component enables finding location easy.
2. Transaction can be done through smart phones.
3. Multimedia functionality allows virtualization.

### Three challenges with mobile BI include-

1. Managing standards for these devices.
2. Managing security (always a big challenge).
3. Managing ―bring your own device‖, where you have devices both owned by the company and devices owned by the individual, both contributing to productivity.

**~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~**

**DIFFERENCE BETWEEN CLOUD COMPUTING AND BIG DATA.**



**~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~**

**BIG DATA CHALLENGES:**

The major challenges associated with big data are as follows:

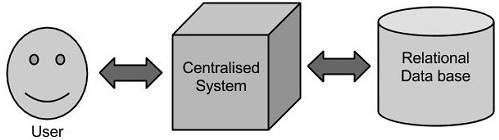
* Capturing data
* Curation
* Storage
* Searching
* Sharing
* Transfer
* Analysis
* Presentation

To fulfill the above challenges, organizations normally take the help of enterprise servers.

**~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~**

# TRADITIONAL APPROACH

In this approach, an enterprise will have a computer to store and process big data. Here data will be stored in an RDBMS like Oracle Database, MS SQL Server or DB2 and sophisticated softwares can be written to interact with the database, process the required data and present it to the users for analysis purpose.



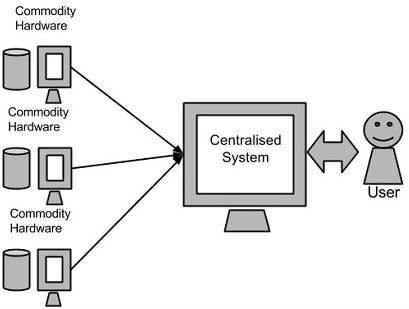
**IT’S LIMITATION**

This approach works well where we have less volume of data that can be accommodated by standard database servers, or up to the limit of the processor which is processing the data. But when it comes to dealing with huge amounts of data, it is really a tedious task to process such data through a traditional database server.

**~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~**

# GOOGLE’S SOLUTION

Google solved this problem using an algorithm called **MAPREDUCE.** This algorithm divides the task into small parts and assigns those parts to many computers connected over the network, and collects the results to form the final result dataset. Above diagram shows various commodity hardwares which could be single CPU machines or servers with higher capacity.

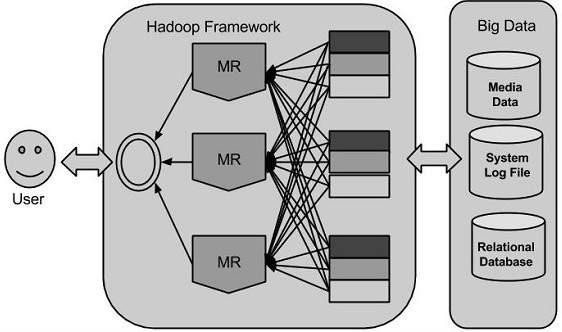
****

**~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~**

# HADOOP

Doug Cutting, Mike Cafarella and team took the solution provided by Google and started an Open Source Project called HADOOP in 2005 and Doug named it after his son's toy elephant.

* + Now Apache Hadoop is a registered trademark of the Apache Software Foundation.
  + **Hadoop runs applications using the MapReduce algorithm**, where the data is processed in **parallel on different CPU nodes**.
  + In short, Hadoop framework is capable enough to develop applications capable of running on clusters of computers and **they could perform complete statistical analysis for a huge amounts of data**.



* + Hadoop is an Apache **open source framework** **WRITTEN IN JAVA** that **allows distributed processing of large datasets across clusters of computers using simple programming models**.
  + A Hadoop frame-worked application **works in an environment that provides distributed storage and computation across clusters of computers**.
  + Hadoop is **designed to scale up from single server to thousands of machines**, each offering local computation and storage.

**~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~**

# HADOOP ARCHITECTURE

**Hadoop framework includes following FOUR MODULES:**

* **Hadoop Common:** These are Java libraries and utilities required by other Hadoop modules. These libraries

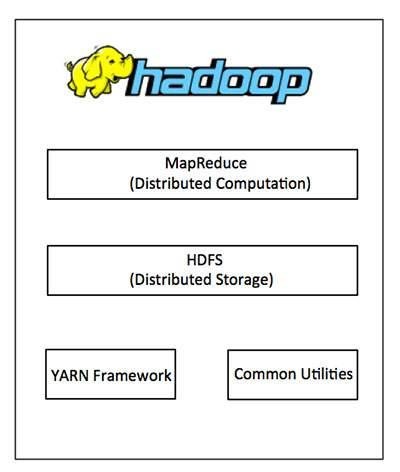
provides filesystem and OS level abstractions and contains the necessary Java files and scripts required to start Hadoop.

* **Hadoop YARN:** This is a framework for job scheduling and cluster resource management.
* **Hadoop Distributed File System (HDFS™):** A distributed file system that provides high-throughput access to

application data.

* **Hadoop MapReduce:** This is YARN-based system for parallel processing of large data sets.

**We can use following diagram to depict these four components available in Hadoop framework.**

Since 2012, the term "Hadoop" often refers not just to the base modules mentioned above but also to the collection of additional software packages that can be installed on top of or alongside Hadoop, such as Apache Pig, Apache Hive, Apache HBase, Apache Spark etc.

**~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~**

# MAPREDUCE

Hadoop **MapReduce** is **a software framework for easily writing applications which process big amounts of data in-parallel on large clusters** (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

**THE TERM MAPREDUCE ACTUALLY REFERS TO THE FOLLOWING TWO DIFFERENT TASKS THAT HADOOP PROGRAMS PERFORM:**

* **The Map Task:** This is the first task, which takes input data and converts it into a set of data, where individual elements are broken down into tuples (key/value pairs).
* **The Reduce Task:** This task takes the output from a map task as input and combines those data tuples into a smaller set of tuples. The reduce task is always performed after the map task.
* Typically both the input and the output are stored in a **file-system**. The framework **takes care of scheduling tasks, monitoring them and re-executes the failed tasks.**
* The MapReduce framework consists of a single master **JobTracker** and one slave **TaskTracker** per cluster-node.
* The master is responsible for resource management, tracking resource consumption/availability and scheduling the jobs component tasks on the slaves, monitoring them and re-executing the failed tasks.
* The slaves **TaskTracker** execute the tasks as directed by the master and provide task-status information to the master periodically.
* The **JobTracker** is a single point of failure for the Hadoop MapReduce service which means if JobTracker goes down, all running jobs are halted.

**~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~**

# HADOOP DISTRIBUTED FILE SYSTEM

* The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on large clusters (thousands of computers) of small computer machines in a reliable, fault-tolerant manner.
* HDFS uses a master/slave architecture where master consists of a single **NameNode** that manages the file system metadata and one or more slave **DataNodes** that store the actual data.
* A file in an HDFS namespace is split into several blocks and those blocks are stored in a set of DataNodes. The NameNode determines the mapping of blocks to the DataNodes. The DataNodes takes care of read and write operation with the file system. They also take care of block creation, deletion and replication based on instruction given by NameNode.
* HDFS provides a shell like any other file system and a list of commands are available to interact with the file system. These shell commands will be covered in a separate chapter along with appropriate examples.