**BIG DATA ANALYTICS**

**MODULE I :**

* **Introduction:** Due to the advent of new technologies, devices, and communication means like social networking sites, the amount of data produced by mankind is growing rapidly every year. The amount of data produced by us from the beginning of time till 2003 was 5 billion gigabytes. If you pile up the data in the form of disks it may fill an entire football field. The same amount was created in every two days in 2011, and in every ten minutes in 2013. This rate is still growing enormously. **Though all this information produced is meaningful and can be useful when processed**, it is being neglected.
* 90% of the world’s data was generated in **the last few years**.

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

**Big data and its importance:**

**Big data:** Big data means really a big data , it is a **collection of large datasets that cannot be processed using traditional computing techniques**. Big data is not merely a data, rather it has become a complete subject, which involves various tools, techniques and frameworks.

(or)

Big data refers to datasets whose size is beyond the ability of typical database software tool to capture, store, managed and analyze.

**Here ’s our standard answer in three parts:**

1. **COMPUTING PERFECT STORM:**

**Big Data analytics are the natural result of FOUR MAJOR GLOBAL TRENDS:**

* **Moore's Law** (technology always gets cheaper),
* **Mobile computing** (that smart phone or mobile tablet in your hand),
* **Social networking** (Facebook, Foursquare, Pinterest, etc.), and
* **Cloud computing** (you don't even have to own hardware or software anymore ;).

1. **DATA PERFECT STORM:**

Volumes of transactional data have been around for decades for most big firms, but now:

* Gates have now opened with more **volume**, and the **velocity** and **variety**—the threeVs
* Three Vs makes it extremely **complex** and **cumbersome** with the current data management & analytics technology and practices.

1. **CONVERGENCE PERFECT STORM:**

New alternatives for IT and business executives to address Big Data analytics;

**MERGING OF**

* Traditional DBMS and analytics software and hardware technologies,
* Open-source technology, and
* Commodity hardware

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

Aside from the changes in the actual hardware and software technology, there has also been a -massive change in the actual **EVOLUTION OF DATA SYSTEMS**: (Misha Ghosh - Innovator)

* **DEPENDENT** (Early Days) **=>** Data systems were fairly new and users didn't know quite know what they wanted. IT assumed that “Build it and they shall come.”
* **INDEPENDENT** (Recent Years) **=>** Users understood what an analytical platform was and worked together with IT to define the business needs and approach for deriving insights for their firm.
* **INTERDEPENDENT** (Big Data Era) **=>** Interactional stage between various companies, creating more social collaboration beyond your firm's walls.

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

**A flood of mythic "start up" proportions:**

* ~~People who compare the amount of data produced daily to a deluge of mythic proportions are entirely correct. This flood of data represents something we’ve never seen before. It’s new, it’s powerful, and yes, it’s scary but extremely exciting.~~
* ~~This is something that every entrepreneur takes to heart as they evangelize their start-up’s big idea that they know will impact the world! This is also true with Big Data and the new technology and approaches that have arrived at our doorstep.~~
* ~~Over the past decade companies like Facebook, Google, LinkedIn, and eBay have created treasured firms that rely on the skills of new data scientists, who are breaking the traditional barriers by leveraging new technology and approaches to capture and analyze data that drives their business.~~
* It’s all about finding the right home for the new approaches and making them work for you!

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

**Big data is more than merely big why now?**

* Defines Big Data as “data that becomes large enough that it cannot be processed using conventional methods.”
* Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. This definition is intentionally subjective.
* Technology advances over time, the size of datasets that qualify as big data will also increase. Depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry.
* The real challenge is identifying or developing most cost-effective and reliable methods for extracting value from all the terabytes and peta bytes of data now available. That’s where **BIG DATA ANALYTICS BECOME NECESSARY.**

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

**Why Now?**

**TIMELINE OF RECENT TECHNOLOGY DEVELOPMENTS.**

* If you believe that it’s possible to learn from past mistakes, then one mistake we certainly do not want to repeat is investing in new technologies that didn’t fit into existing business frameworks.
* During the customer relationship management (CRM) era of the 1990s, many companies made substantial investments in customer-facing technologies that subsequently failed to deliver expected value.
* The reason for most of those failures was fairly straightforward: Management either forgot (or just didn’t know) that big projects require a synchronized transformation of people, process, and technology.
* We can avoid those kinds of mistakes if we keep our attention focused on the outcomes we want to achieve. The technology of Big Data is the easy part—the hard part is figuring out what you are going to do with the output generated by your Big Data analytics.

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

**A convergence of key trends, a wider variety of data:**

* The **DIFFERENCE BETWEEN “OLD BIG DATA” AND “NEW BIG DATA” IS ACCESSIBILITY**. Here ’s a brief summary of our interview:
* Companies have always kept large amounts of information. But until recently, they stored most of that information on **tape**. While it’s true that the amount of data in the world keeps growing, the real change has been in the ways that we access that data and use it to create value.
* Today, you have technologies like Hadoop, for example, that make it functionally practical to access a tremendous amount of data, and then extract value from it. The availability of lower-cost hardware makes it easier and more feasible to retrieve and process information, quickly and at lower costs than ever before.
* So it ’s the convergence of several trends—more data and less expensive, faster hardware—that ’s driving this transformation. Today, we ’ve got raw speed at an affordable price. That cost/benefit has

really been a game changer for us.

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

**A Wider Variety of Data**

* The variety of data sources continues to increase. Traditionally, internally focused operational systems, such **as ERP (enterprise resource planning) and CRM applications**, were the major source of data used in analytic processing.

**Wider variety of data sources such as:**

* **Internet data** (i.e., clickstream, social media, social networking links)
* **Primary research** (i.e., surveys, experiments, observations)
* **Secondary research** (i.e., competitive and marketplace data, industry reports, consumer data, business data)
* **Location data** (i.e., mobile device data, geospatial data)
* **Image data** (i.e., video, satellite image, surveillance)
* **Supply chain data** (i.e., EDI, vendor catalogs and pricing, quality information)
* **Device data** (i.e., sensors, PLCs, RF devices, LIMs, telemetry)

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

**The expanding universe of unstructured data**

* **STRUCTURED DATA** (the kind that is easy to define, store, and analyze)
* **UNSTRUCTURED DATA** : The kind that tends to defy easy definition, **takes up lots of storage capacity**, and is typically **more difficult to analyze.**

**INFORMATION THAT EITHER**

* **does not have a predefined data model** and/or
* **does not fit well into a relational database.**
* is **typically text heavy**, but may contain data such as dates, numbers, and facts as well.
* Unstructured data **tends to grow exponentially**, unlike structured data,which tends to grow in a more linear fashion.
* **Most new data is unstructured**
* Unstructured data **represents** almost **95 percent** of **new** data, while **structured** data represents only **5** **percent**.
* **Unstructured data is vastly underutilized**
* The term **SEMI-STRUCTURED DATA**is used to describe structured data that doesn’t fit into a formal structure of data models.
* ~~The amount of data (all data, everywhere) is doubling every two years. Our world is becoming more transparent. We, in turn, are beginning to accept this as we become more comfortable with parting with data that we used to consider sacred and private. Imagine huge deposits of oil or other natural resources that are just sitting there, waiting to be used. That ’s the current state of unstructured data as of today. Tomorrow will be a different story because there ’s a lot of money to be made for smart individuals and companies that can mine unstructured data successfully.~~

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

**Big Data analytics uses a wide variety of advanced analytic, TO PROVIDE:**

* **Deeper insights.** Rather than looking at segments, classifications, regions, groups, or other summary levels you’ll have insights into all the individuals, all the products, all the parts, all the events, all the transactions, etc.
* **Broader insights.** The world is complex. Operating a business in a global, connected economy is very complex given constantly evolving and changing conditions. As humans, we simplify conditions so we can process events and understand what is happening. Big Data analytics takes into account all the data, including new data sources, to understand the complex, evolving, and interrelated conditions to produce more accurate insights.
* **Frictionless actions.** Increased reliability and accuracy that will allow the deeper and broader insights to be automated into systematic actions.

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

**Industry examples of big data: Digital marketing and the online world**

# INDUSTRY EXAMPLES OF BIG DATA

## Digital Marketing

* Introduction
* Database Marketers, Pioneers of Big Data
* Big Data & New School of Marketing
* Cross Channel Life cycle Marketing
* **INTRODUCTION**
* **DIGITAL MARKETING** encompasses using any sort of online media (profit or non-profit) for driving people to a website, a mobile app etc. and retaining them and interacting with them to **understand what consumers really want.**
* Digital marketing is easy when consumers interact with corporate` primary platform (ie. The one the corporate own) because corporate get good information about them. But corporate get very little information once people start interacting with other platforms (eg. *Facebook, Twitter, Google +*).
* One of the **problems** that have to be dealt with in a digital existence, is that **corporate do not have access to data for all consumers** (ie. There is very little visibility about people when they interact in social networking sites). So **corporate lose control of their ability to access the data they need**, in order to make smart, timely decisions.
* Big data on the web will completely transform a company’s ability to understand the effectiveness of its marketing and the ability to understand how its competitors are behaving. Rather than a few people making big decisions, **the organizations will be able to make hundreds and thousands of smart decisions every day.**
* **Database Marketers, Pioneers of Big Data:**

**DATABASE MARKETING** is concerned with building databases containing info about individuals, using that information to better understand those individuals and communicating effectively with some of those individuals to drive business value.

**MARKETING DATABASES ARE TYPICALLY USED FOR**

* **Customer acquisition**
* **Retaining and cross-selling to existing customers which reactivates the cycle**

- As companies grew and systems proliferated, a situation where there was one system for one product and another for another product etc. was landed up (silos).

- Then companies began developing technologies to manage and duplicate data from multiple sources companies started developing software that could eliminate duplicate customer info (de-duping).

- This enable them to extract customer information from silos product systems, manage the info into single database, remove all the duplicates and then send direct mail to subsets of the customers in the database.

- ~~Companies such as~~ **Reader’s Digest** ~~and several other firms were early champions of this new kind of marketing and they used it very effectively. By the 1980’s marketers developed the ability to run reports on the info in their databases which gave them better and deeper insights into buying habits and preferences of customers.~~ **Telemarketing** ~~became popular when marketers figured out how to feed information extracted from customer databases to call centers. In 1990’s email entered the picture and marketers saw opportunities to~~ **reach customers via Internet and** [**WWW.-**](http://WWW/) ~~In the past five years there has been exponential growth in database marketing and the new scale is pushing up against the limits of technology~~

* **Big Data & New School of Marketing:**

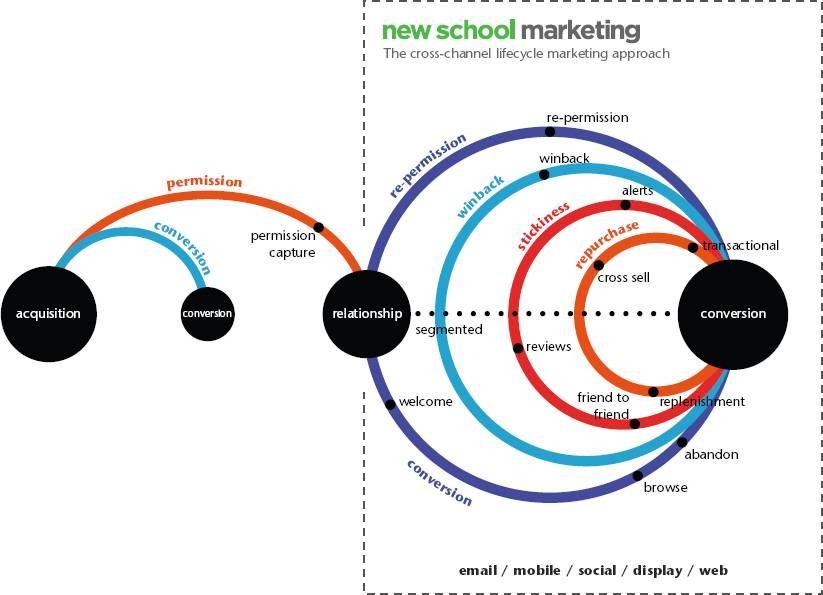
New school marketers **deliver what today’s consumers want** ie. Relevant interactive communication across digital power channels Digital power channels: email, mobile, social display and web. Consumers have changed so must marketers.

* **Right approach – Cross Channel Lifecycle Marketing**
* **CROSS-CHANNEL LIFECYCLE MARKETING** really starts with the capture of customer permission, contact information, and preferences for multiple channels.
* It also requires marketers to have the right integrated marketing and customer information systems, so that

(1) They can have complete understanding of customers through stated preferences and observed behavior at any given time; and

(2) They can automate and optimize their programs and processes throughout the customer lifecycle.

* Once marketers have that, they need a practical framework for planning marketing activities.

The **various loops that guide marketing strategies and tactics in the Cross- Channel Lifecycle Marketing approach**: ***conversion, repurchase, stickiness, win-back, and re-permission are shown in the following figure.***

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

## FINANCIAL SERVICES

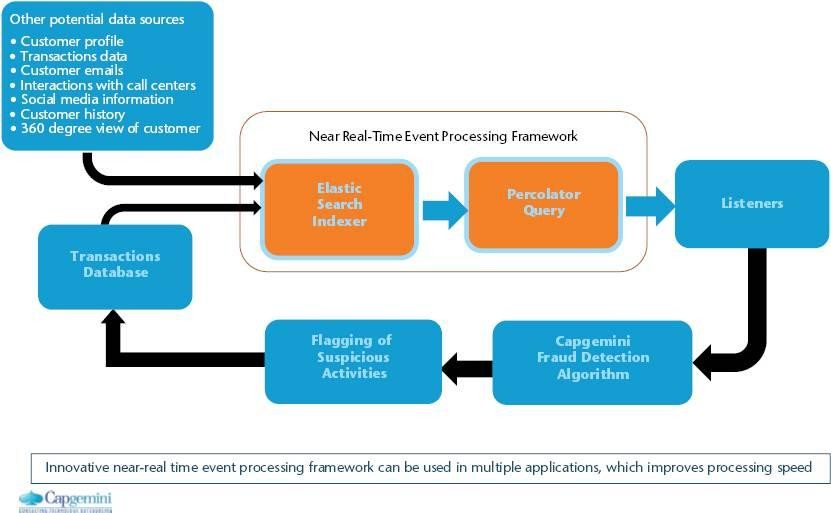
### Fraud & Big Data

* Fraud is intentional deception made for personal gain or to damage another individual.
* One of the most common forms of fraudulent activity is credit card fraud.
* Social media and mobile phones are forming new frontiers fraud.

Capegemini financial services team believes that due to the nature of data streams and processing required BIG Data Technologies provide an optimal technology solution based on the **FOLLOWING THREE VS :**

1. **High volume:** Years of consumer records and transactions (150 billion +records per year).
2. **High velocity:** Dynamic transactions and social media info.
3. **High variety:** Social media plus other unstructured data such as customer E-mails, call center conversations as well as transactional structured data.

## FRAUD DETECTION POWERED BY NEAR REAL-TIME EVENT PROCESSING FRAMEWORK

* The near real-time event processing framework can be used in multiple applications which improves processing speed.
* This fraud detection system uses an open source search server based on Apache Lucene.
* It can be used to search all kind of documents at near real-time. The tool is used to index new transactions which are sourced in real-time, which allows analytics to run in a distributed fashion utilizing the data specific to the index.
* Using this tool, large historical data sets can be used in conjunction with real-time data to identify deviation from typical payment patterns.
* The big data component allows overall historical patterns to be compared and contrasted and allows the number of attributes and characteristics about consumer behavior to be very wide with little impact on overall performance.
* Percolator query performs the function of identifying new transactions that have raised profiles. Percolator query can handle both structured and unstructured data.
* This provides scalability to the event processing framework and allows specific suspicious transactions to be enriched with additional unstructured information (E.g. Phone location/geospatial records, customer travel schedules and so on).
* This ability to enrich the transaction further can reduce false positives and increase the experience of customer while redirecting fraud efforts to actual instances of suspicious activity.
* Capegemini’s fraud Big Data initiative focuses on flagging the suspicious credit card transactions to prevent fraud in near real-time via multi-attribute monitoring.
* Real- time inputs involving transaction data and customers records are monitored via validity checks and detection rules.

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

## Big data and Healthcare

* Big data promises enormous **revolution in healthcare**, with the advancements in everything from the management of chronic disease to the delivery of personalized medicine.
* The healthcare industry now **has huge amount of data**: from biological data such as **gene expression**, **Special Needs Plans** (SNPs), **proteomics**, **metabolomics**, and **next-generation gene sequence data** etc. The exponential growth in data is further accelerated by the digitization of patient level data stored in **Electronic Medical Records** (EMRs) or **Electronic Health Records** (EHRs) and **Health Information Exchanges** (HIEs) enhanced with data from imaging and test results, medical and prescription claims and personal health devices.
* In addition to saving and improving lives, Big Data has the potential to transform the entire health care system by replacing guesswork and intuition with objective, data-driven science (see Figure).

### Figure: Data in the World of Healthcare

The **HEALTHCARE SYSTEM IS FACING SEVERE ECONOMIC, EFFECTIVENESS AND QUALITY CHALLENGES**. These factors are forcing transformation in pharmaceutical business model. Hence the healthcare industry is moving from traditional model built on regulatory approval and settling of claims to medical evidence and proving economic effectiveness through improved analytics derived insights. **The success of this model depends on the creation of robust analytics capability an harnessing integrated real-world patient level data.**

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

## ADVERTISING AND BIG DATA

Big Data is changing the way advertisers address three related needs.

1. How much to spend on advertisements.
2. How to allocate amount across all the marketing communication touch points.

How to optimize advertising effectiveness, given these needs advertisers need to measure their advertising end to end in terms of Reach, Response & Reaction.

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

**LAST YEAR QUESTION PAPER:**

**1.What are REAL-TIME INDUSTRY APPLICATIONS OF HADOOP?**

* Hadoop, well known as Apache Hadoop, is an open-source software platform for scalable and distributed computing of large volumes of data.
* It provides rapid, high performance and cost-effective analysis of structured and unstructured data generated on digital platforms and within the enterprise. It is used in almost all departments and sectors today. I would like to share a interesting video related to hadoop.

**Some of the instances where Hadoop is used:**

* Managing traffic on streets
* Streaming processing
* Content Management and Archiving Emails
* Processing Rat Brain Neuronal Signals using a Hadoop Computing Cluster
* Fraud detection and Prevention
* Advertisements Targeting Platforms are using Hadoop to capture and analyze click stream, transaction, video and social media data
* Managing content, posts, images and videos on social media platforms
* Analyzing customer data in real-time for improving business performance
* Public sector fields such as intelligence, defense, cyber security and scientific research
* Financial agencies are using Big Data Hadoop to reduce risk, analyze fraud patterns, identify rogue traders, more precisely target their marketing campaigns based on customer segmentation, and improve customer satisfaction
* Getting access to unstructured data like output from medical devices, doctor’s notes, lab results, imaging reports, medical correspondence, clinical data, and financial data.

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

**What is THE IMPORTANCE OF UNSTRUCTURED DATA IN REAL WORD? Give Examples?**

**unstructured data**: It might be clear to many, but just so we are all on the same page, unstructured data is images, video, sound and documents.

**PART -2**

**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-**

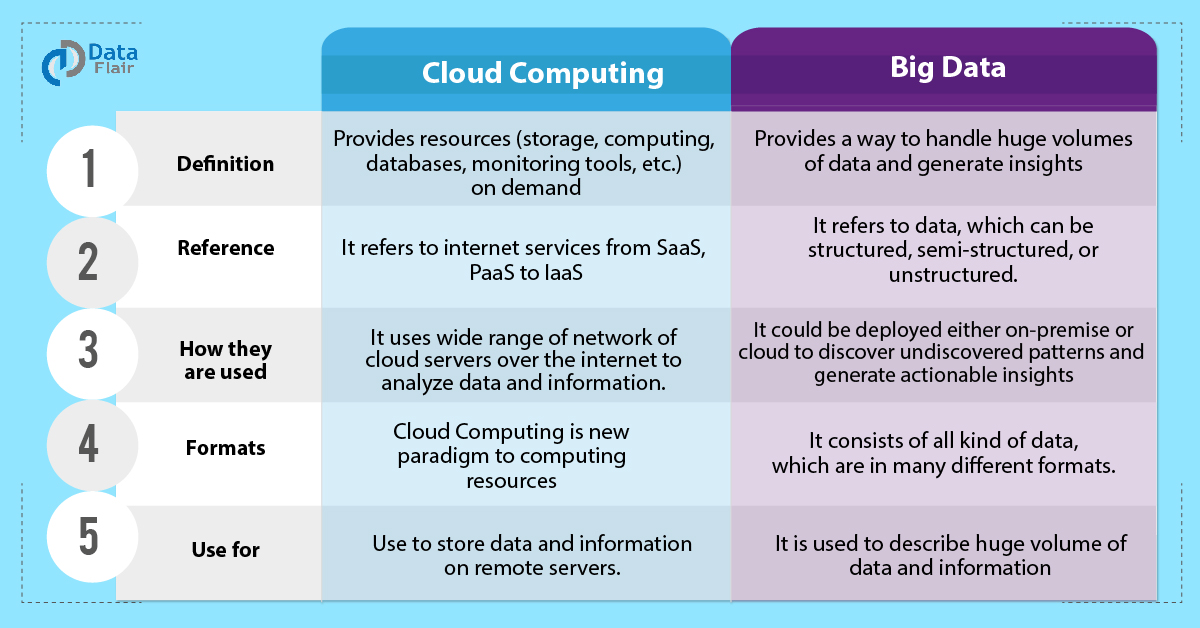
1. Location-GPS component enables finding location easy.
2. Transaction can be done through smart phones.
3. Multimedia functionality allows virtualization.

**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:

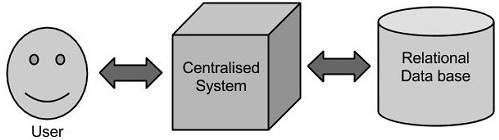
* Capturing data
* Curation
* Storage
* Searching
* Sharing
* Transfer
* Analysis
* Presentation

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.



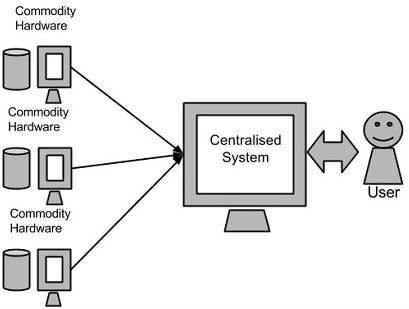
**IT’S LIMITATION**

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.

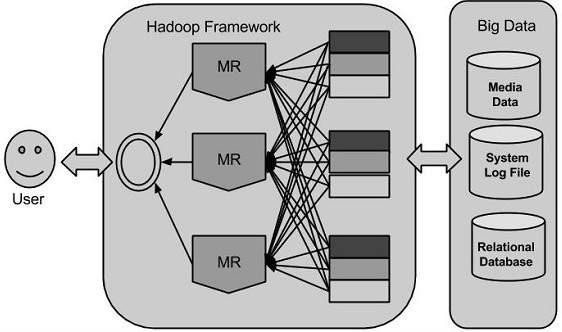
****

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

**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

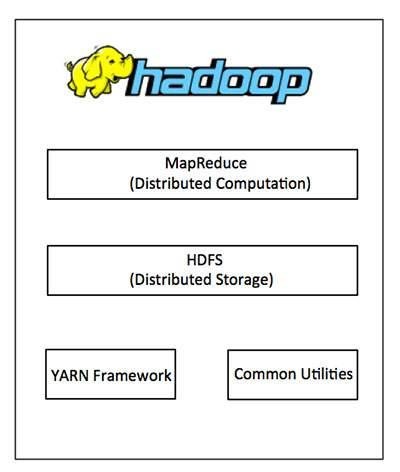
provides filesystem and OS level abstractions and contains the necessary Java files and scripts required to start Hadoop.

* **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.