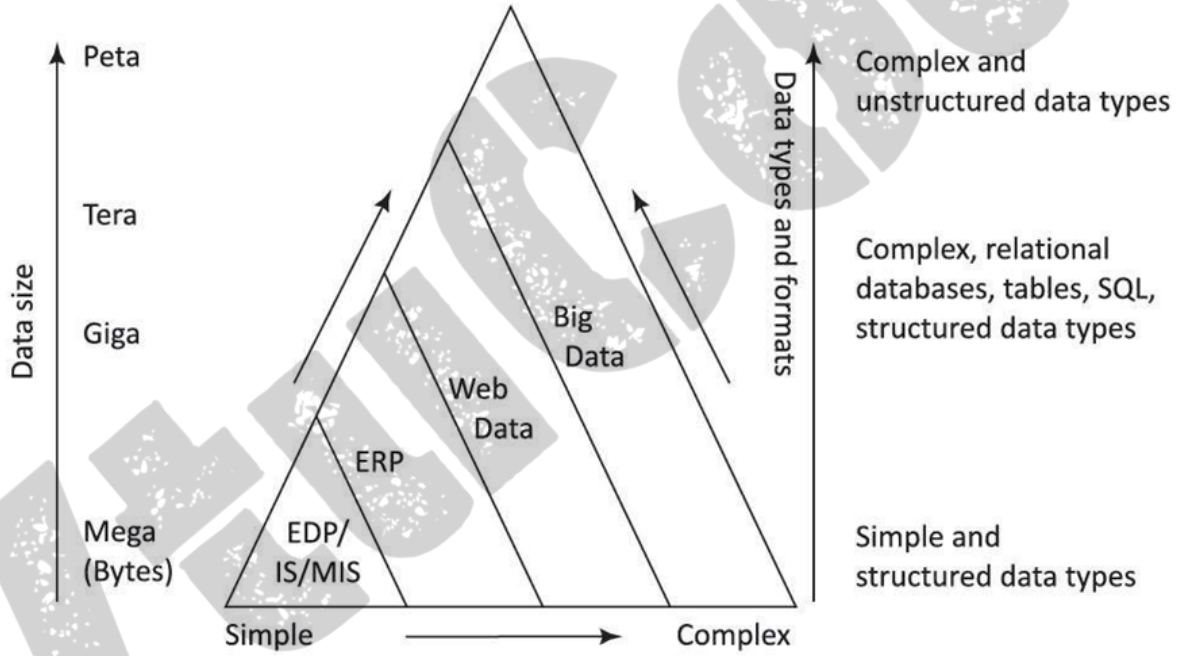
MODULE 1: INTRODUCTION TO BIG DATA ANALYTICS

Need of Big Data

The rise in technology has led to the production and storage of voluminous amounts data. Earlier megabytes were used but nowadays petabytes are used for processing, analysis, discovering new facts and generating new knowledge. Conventional systems for storage, processing and analysis pose challenges in large growth in volume of data, variety of data, various forms and formats, increasing complexity, faster generation of data and need of quickly processing, analyzing and usage.

Figure shows data usage and growth. As size and complexity increase, the proportion of unstructured



Evolution of Big data and their Characteristics

Data

Data has several definitions.Usages can be singular or plural.

- “Data is information, usually in the form of facts or statistics that one can analyze or use for further calculations.”
- "Data is information that can be stored and used by a computer program"
- "Data is information presented in numbers, letters, or other form".

- "Data is information from series of observations, measurements or facts".
- "Data is information from series of behavioural observations, measurements or facts",

Web Data

. Web data is the data present on web servers (or enterprise servers) in the form of text, images, videos,audios and multimedia files for web users. A user (client software) interacts with this data. A client can access (pull) data of responses from a server. The data can also publish (push) or post (after registering subscription) from a server. Internet applications including web sites, web services, web portals, online business applications, emails, chats, tweets and social networks provide and consume the web data.

Classification of Data-Structured, Semi-structured and Unstructured

- Data can be classified as structured, semi-structured, multi-structured and unstructured:
- Structured data conform and associate with data schemas and data models. Structured tables (rows and columns). Nearly 15-20% data are in structured or semi-structured form, Unstructured data
- Unstructured data do not conform and associate with any data models.

Structured Data:

Structured data enables the following:

- Data insert, delete, update and append , Indexing to enable faster data retrieval
- Scalability which enables increasing or decreasing capacities and data processing operations such as, storing, processing and analytics
- encryption and decryption for data security.

Semi-Structured Data

- Examples of semi-structured data are XML and JSON documents.
- Semi-structured data contain tags or other markers, which separate semantic elements and enforce hierarchies of records and fields within the data.
- Semi-structured form of data does not conform and associate with formal data model structures.
- Data do not associate data models, such as the relational database and table models.

Multi-Structured Data

- Multi-structured data refers to data consisting of multiple formats of data, viz. structured, semi-structured and/or unstructured data.
- Multi-structured data sets can have many formats. They are found in non-transactional systems.
- For example, streaming data on customer interactions, data of multiple sensors, data at web or enterprise server or the data warehouse data in multiple formats.
- Multi-or semi-structured data has some semantic meanings and data is in both structured and unstructured formats.

Unstructured Data

- Unstructured data does not possess data feature such as a table or a database.
- Unstructured data are found in file type such as TXT, CSV.
- Data may be have internal structures, such as in e-mails.
- The relationships, schema and features need to be separately established

BIG DATA:

- Big Data is high-volume, high-velocity and/or high-variety information asset that requires new forms of processing for enhanced decision making, insight discovery and process optimization .

- "A collection of data sets so large or complex that traditional data processing applications are inadequate."

- "Data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges."

- "Big Data refers to data sets whose size is beyond the ability of typical database software tool to capture, store, manage and analyze.

Big Data Characteristics

Volume, variety and/or velocity as the key "data management challenges" for enterprises. Analytics also describe the '4Vs', i.e. volume, velocity, variety and veracity.:

- **Volume** The phrase 'Big Data' contains the term big, which is related to size of the data and hence the characteristic. Size defines the amount or quantity of data, which is generated from an application(s).
- **Velocity** The term velocity refers to the speed of generation of data. Velocity is a measure of how fast the data generates and processes. To meet the demands and the challenges of processing Big Data, the velocity of generation of data plays a crucial role.
- **Variety** Big Data comprises of a variety of data. Data is generated from multiple sources in a system. This introduces variety in data and therefore introduces complexity. Data consists of various forms and formats. The variety is due to the availability of a large number of heterogeneous platforms in the industry. This characteristic helps in effective use of data according to their formats.
- **Veracity** is also considered an important characteristics to take into account the quality of data captured, which can vary greatly, affecting its accurate analysis.

Big Data Types

1. Social networks and web data, such as Facebook, Twitter, e-mails, blogs and YouTube.

2 Transactions data and Business Processes (BPs) data, such as credit card transactions, flight bookings, etc. and public agencies data such as medical records, insurance business data etc.

3 Customer master data, such a data for facial recognition and for the name, date of birth, marriage anniversary, gender, location and income category.

4 Machine-generated data, such as machine-to-machine or Internet of Things data, and the data from sensors, trackers, web logs and computer systems log. Computer generated data is also considered as machine generated data from data store. Usage of programs for processing

of data using data repositories, such as database or file, generates data and also machine generated data.

5.Human-generated data such as biometrics data, human-machine interaction data, e-mail records with a mail server and MySQL database of student grades.

Big Data Classification

Basis of classification	**Examples**
Data sources (traditional)	Data storage such as records, RDBMs, distributed databases, row-oriented In- memory data tables, column-oriented In-memory data tables, data warehouse, server, machine-generated data, human-sourced data, Business Process (BP) data, Business Intelligence (BI) data
Data formats (traditional)	Structured and semi-structured
Big Data sources	Data storage, distributed file system, Operational Data Store (ODS), data marts, data warehouse, NoSQL database (MongoDB, Cassandra), sensors data, audit trail of financial transactions, external data such as web, social media, weather data, health records
Big Data formats	Unstructured, semi-structured and multi-structured data
Data Stores structure	Web, enterprise or cloud servers, data warehouse, row-oriented data for OLTP, column-oriented for OLAP, records, graph database, hashed entries for key/value pairs
Processing data rates	

	Batch, near-time, real-time, streaming
Processing Big Data rates	High volume, velocity, variety and veracity, batch, near real-time and streaming data processing,
Analysis types	Batch, scheduled, near real-time datasets analytics
Big Data processing methods	Batch processing (for example, using MapReduce, Hive or Pig), real-time processing (for example, using SparkStreaming, SparkSQL, Apache Drill)
Data analysis methods	Statistical analysis, predictive analysis, regression analysis, Mahout, machine learning algorithms, clustering algorithms, classifiers, text analysis, social network analysis, location-based analysis, diagnostic analysis, cognitive analysis
Data Usuages	Human, business process, knowledge discovery, enterprise applications, Data

SCALABILITY AND PARALLEL PROCESSING

Big Data needs processing of large data volume, and therefore needs intensive computations. Processing complex applications with large datasets (terabyte to petabyte datasets) need hundreds of computing nodes.

Convergence of Data Environments and Analytics

- Big Data processing and analytics requires scaling up and scaling out, both vertical and horizontal computing resources.
- Computing and storage systems when run in parallel, enable scaling out and increase system capacity.
- Scalability enables increase or decrease in the capacity of data storage, processing and analytics. Scalability is the capability of a system to handle the workload as per the magnitude of the work.
- System capability needs increment with the increased workloads. When the workload and complexity exceed the system capacity, scale it up and scale it out.

Analytics Scalability to Big Data

- **Vertical scalability** means scaling up the given system's resources and increasing the system's analytics, reporting and visualization capabilities. Scaling up means designing the algorithm according to the architecture that uses resources efficiently.
- **Horizontal scalability** means increasing the number of systems working incoherence and scaling out the workload. Processing different datasets of a large dataset deploys horizontal scalability.Scaling out means using more resources and distributing the processing and storage tasks in parallel.

The easiest way to scale up and scale out the execution of analytics software is to implement it on a bigger machine with more CPUs for greater volume, velocity, variety and complexity of data. The software will definitely perform better on a bigger machine.

However, buying faster CPUs, bigger and faster RAM modules and hard disks, faster and bigger motherboards will be expensive compared to the extra performance achieved by efficient design of algorithms.If more CPUs add in a computer, but the software does not

exploit the advantage of them, then that will not get any increased performance out of the additional CPUs.

Massively Parallel Processing Platforms

When making software, draw the advantage of multiple computers (or even multiple CPUs within the Scaling uses parallel processing systems. Many programs are so large and/ required to enhance (scale) up the computer system or use massive parallel (MPP) processing (MPPs) platforms.

Parallelization of tasks can be done at several levels:
(i) distributing separate tasks onto separate threads on the same CPU. in distribution
(ii) distributing separate tasks onto separate CPUs on the same computer
(iii) distributing separate tasks onto separate computers

Multiple compute resources are used in parallel processing systems. The computational problem is broken into discrete pieces of sub-tasks that can be processed simultaneously. The system executes multiple program instructions or sub-tasks at any moment in time. Total time taken will be much less than with a single compute resource.

Distributed Computing Model

- A distributed computing model uses cloud, grid or clusters, which process and analyze big and large datasets on distributed computing nodes connected by high-speed networks.
- It gives the requirements of processing and analyzing big, large and small to medium datasets on distributed computing nodes.
- Big Data processing uses a parallel, scalable, and no-sharing program model, such as MapReduce, for computations on it and data is adversely affected.

<u>Cloud Computing</u>

"Cloud computing is a type of Internet-based computing that provides shared processing resources and data to the computers and other devices on demand."

One of the best approaches for data processing is to perform parallel and distributed computing in a cloud computing environment.

Cloud resources can be Amazon Web Service (AWS) Elastic Compute Cloud (EC2), Microsoft Azure or Apache CloudStack. Amazon Simple Storage Service (S3) provides simple web services interface to store and retrieve any amount of data, at any time, from anywhere on the web.

Cloud computing features are:

(i) on-demand service
(ii) resource pooling,
(iii) scalability,
(iv) accountability
(v) broad network access.

Cloud services can be accessed from anywhere and at any time through the Internet. A local private cloud can also be set up on a local cluster of computers. Cloud computing allows availability of computer infrastructure and services "on-demand" basis. The computing infrastructure includes data storage device, development platform, database, computing power or software applications.

Cloud services can be classified into three fundamental types:

1. Infrastructure as a Service (IaaS): Providing access to resources, such as hard disks, network connections, databases storage, data center and virtual server spaces is Infrastructure as a Service (IaaS). Some examples are Tata CloudStack is an open source software for deploying and managing a large network of virtual machines, and offers public cloud services which provide highly scalable Infrastructure as a Service

2. Platform as a Service (PaaS): It implies providing the runtime environment to allow developers to build applications and services, which means cloud Platform as a Service. Software at the clouds support and manage the services, storage, networking, deploying, testing, collaborating, hosting and maintaining applications. Examples are Hadoop Cloud Service (IBM BigInsight, Microsoft Azure HD Insights, Oracle Big Data Cloud Services).

3. Software as a Service (SaaS): Providing software applications as a service to end-users is known as Software as a Service. Software applications are hosted by a service provider and made available to customers over the Internet. Some examples are SQL GoogleSQL, IBM BigSQL, HPE Vertica, Microsoft Polybase and Oracle Big Data SQL.

Grid and Cluster Computing

Grid Computing

Grid Computing refers to distributed computing, in which a group of computers from several locations are connected with each other to achieve a common task.
A group of computers that might spread over remotely comprise a grid. A grid is used for a variety of purposes.
A single grid of course, dedicates at an instance to a particular application only.
Grid computing provides large-scale resource sharing which is flexible, coordinated and secure among its users. The users consist of individuals, organizations and resources.
Grid computing suits data-intensive storage better than storage of small objects of few million of bytes.
To achieve the maximum benefit from data grids, they should be used for a large amount of data that can distribute over grid nodes. Besides data grid, the other variation of the grid.

Grid computing is scalable. Grid computing also forms a distributed network for resource integration.

Drawbacks of Grid Computing Grid computing is the single point, which leads to failure in case of underperformance or failure of any of the participating nodes.
A system's storage capacity varies with the number of users, instances and the amount of data transferred at a given time.
Sharing resources among a large number of users helps in reducing infrastructure costs and raising load capacities.

Cluster Computing

A cluster is a group of computers connected by a network. The group works together to accomplish the same task. Clusters are used mainly for load balancing. They shift processes between nodes to keep an even load on the group of connected computers. Hadoop architecture uses the similar methods.

Volunteer Computing

Volunteers provide computing resources to projects of importance that use resources to do distributed computing and/or storage. Volunteer computing is a distributed computing paradigm which uses computing resources of the volunteers. Volunteers are organizations or members who own personal computers. Projects examples are science-related projects executed by universities or academia in general.

DESIGNING DATA ARCHITECTURE:

Data Architecture Design.

Big Data architecture is the logical and/or physical layout/structure of how Big Data will be stored, accessed and managed within a Big Data or IT environment. Architecture logically defines how Big Data solution will work, the core components used, flow of information, security and more.

Data analytics need a number of sequential steps. Big Data architecture design task simplifies when using the logical layers approach.

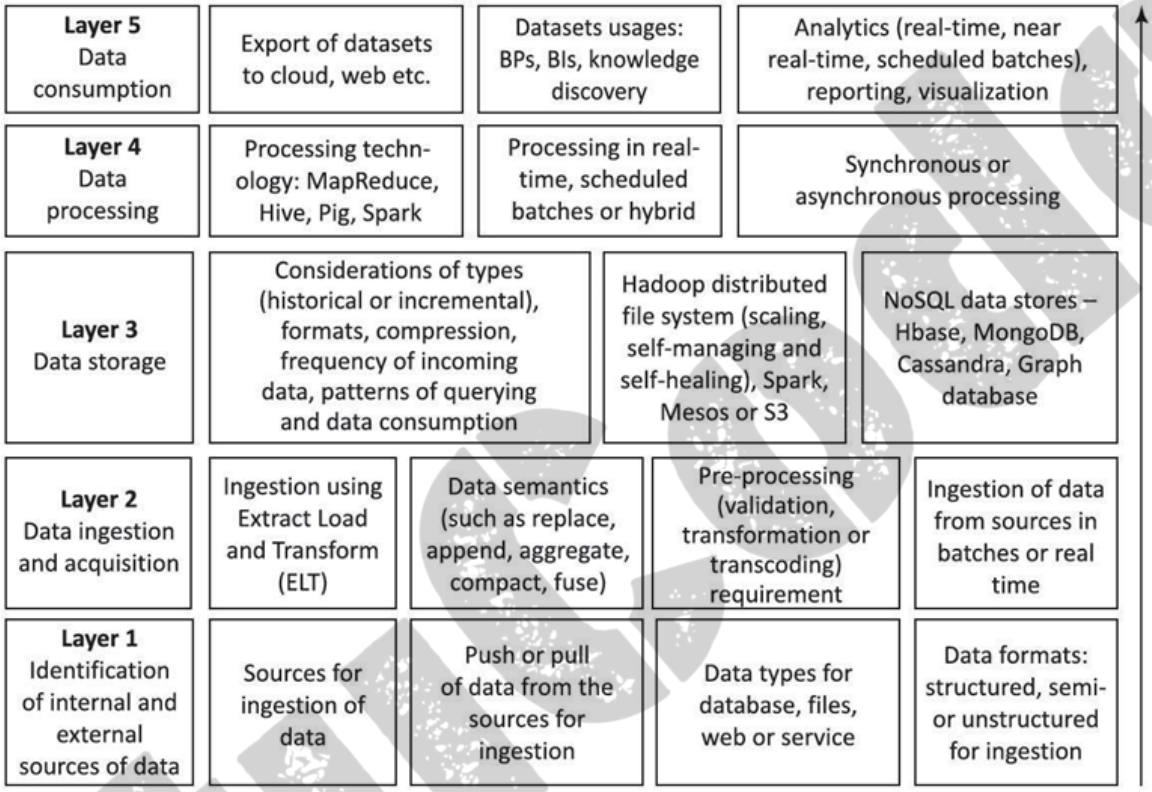


Figure shows the logical layers and the functions which are considered in Big Data architecture. Five vertically aligned textboxes on the left shows the layers. Horizontal textboxes show the functions in each layer.

Data processing architecture consists of five layers:
(i) identification of data sources,
(ii) acquisition, ingestion, extraction, pre-processing, transformation of data,
(iii) data storage at files, servers, cluster or cloud,
(iv) data-processing, and
(v) data consumption in the number of programs and tools.
.

Data consumed for applications, such as business intelligence, data mining, discovering patterns/clusters, artificial intelligence (AI), machine learning (ML), text analytics, descriptive and predictive analytics, and data visualization.

Data ingestion, pre-processing, storage and analytics require special tools and technologies. Logical layer 1 (L1) is for identifying data sources, which are external, internal or both. The layer 2 (L2) is for data-ingestion.Data ingestion means a process of absorbing information, just like the process of absorbing nutrients and medications into the body by eating or drinking them. Ingestion is the process of obtaining and importing data for immediate use transfer. Ingestion may be in batches or in real time using pre-processing or semantics.

The L3 layer is for storage of data from the L2 layer. The L4 is for data processing using software, such as MapReduce, Hive, Pig or Spark. The top layer L5 is for data consumption. Data is used in analytics, visualizations, reporting, export to cloud or web servers.

QUALITY, PRE-PROCESSING

Data Quality:

High quality means data, which enables all the required operations, analysis, decisions, planning and knowledge discovery correctly. A definition for high quality data, especially for artificial intelligence applications, can be data with five R's as follows: Relevancy, recency, range, robustness and reliability. Relevancy is of utmost importance.

Data integrity:

Data integrity refers to the maintenance of consistency and accuracy in data over its usable life. Software, which store, process, or retrieve the data, should maintain the integrity of data. Data should be incorruptible.

Data Noise, Outliers, Missing and Duplicate Values

Noise: One of the factors effecting data quality is noise. Noise in data refers to data giving additional meaningless information besides true (actual/required) information.. Noisy data means data having large additional information. Result of data analysis is adversely affected due to noisy data.

Example: The WRMP organization for weather recording. Consider noise in wind velocity and direction readings due to external turbulences. The velocity at certain instances will appear too high and sometimes too low. The directions at certain instances will appear inclined more towards the north and sometimes more towards the south.

Outliers A factor which effects quality is an outlier. An outlier in data refers to data, which appears to not belong to the dataset. For example, data that is outside an expected range. The

outliers are a result of human data-entry errors, programming bugs, some transition effect or phase lag in stabilizing the data value to the true value.
Example: Consider an outlier in the students' grade-sheets in one subject out of five in the fourth semester result of a student. A result in a semester shows 9.0 out of 10 points in place of 3.0 out of 10. Data 9.0 is an outlier. The student semester grade point average (SGPA) will be erroneously declared and the student may even be declared to have failed in that semester.

Missing Values Another factor effecting data quality is missing values. Missing value implies data not appearing in the data set.
Example: Consider missing values in the sales figures of chocolates. The values not sent for certain dates from an ACVM. This may be due to the failure of power supply at the machine or network problems on specific days in a month. The chocolate sales not added for a day can be added in the next day's sales data. The effect over a month on the average sales per day is not significant. However, if the failure occurred on last day of a month, then the analysis will be erroneous.

Duplicate Values Another factor effecting data quality is duplicate values. Duplicate value implies the same data appearing two or more times in a dataset.
Example: Consider duplicate values in the sales figures of chocolates from an ACVM. This may be due to some problem in the system. The number of duplicates for sales when sent and added, then sales result analysis will get affected. It can even result in false alarms to a service, which maintains the supply chain to the ACVMs.

Data Pre-processing

Pre-processing is a must before data mining and analytics. Pre-processing is also a must before running a Machine Learning (ML) algorithm. Analytics needs prior screening of data quality also. Data when being exported to a cloud service or data store needs pre-processing. Pre-processing needs are:

(i) Dropping out of range, inconsistent and outlier values

(ii) Filtering unreliable, irrelevant, and redundant information.

(iii) Data cleaning, editing,

(iv) reduction and/or wrangling Data validation, transformation or transcoding

(v) ELT processing

Data Cleaning

Data cleaning refers to the process of removing or correcting incomplete, incorrect, inaccurate or irrelevant parts of the data after detecting them.

Data Cleaning Tools Data cleaning is done before mining of data. Incomplete or irrelevant data may result into misleading decisions. It is not always possible to create well structured data.Data cleaning tools help in refining and structuring data into usable data.

Data Enrichment:"Data enrichment refers to operations or processes which refine, enhance or improve the raw data."

Data Editing refers to the process of reviewing and adjusting the acquired datasets. The editing controls the data quality. Editing methods are (i) interactive, (ii) selective, (iii) automatic, (iv) aggregating and (v) distribution.

Data reduction enables the transformation of acquired information into an ordered, correct and simplified form. The reductions enable ingestion of meaningful data in the datasets. The basic concept is the reduction of multitudinous amount of data, and use of the meaningful parts..

Data wrangling refers to the process of transforming and mapping the data. Results from analytics are then appropriate and valuable. For example, mapping enables data into another format, which makes it valuable for analytics and data visualizations.

Big Data Platform

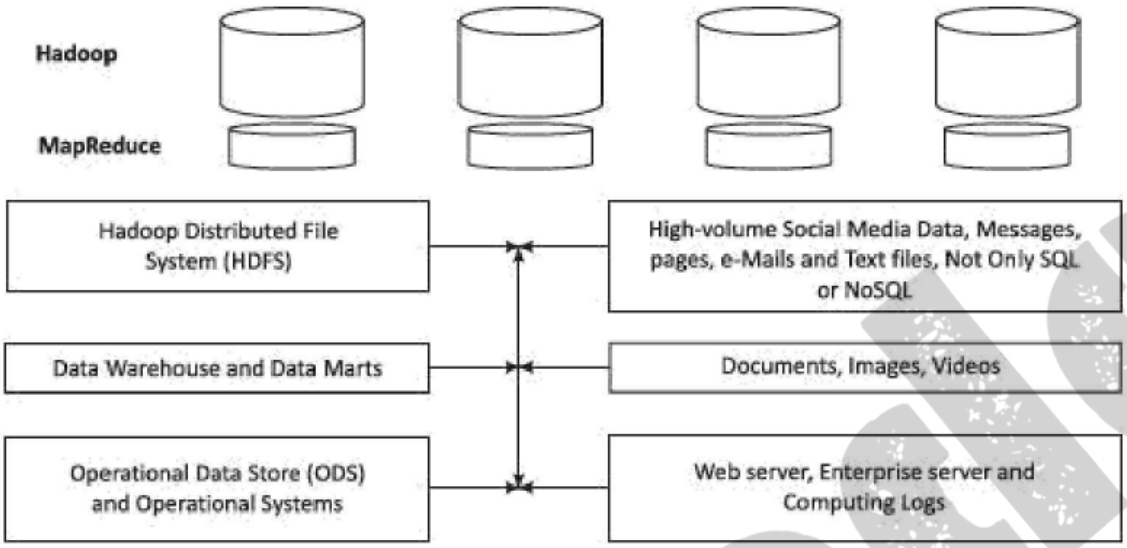
A Big Data platform supports large datasets and volume of data. The data generate at a higher velocity, in more varieties or in higher veracity. Managing Big Data requires large resources of MPPs, cloud, parallel processing and specialized tools. Bigdata platform should provision tools and services for:

1. Storage, processing and analytics,
2 developing, deploying, operating and managing Big Data environment,
3 Reducing the complexity of multiple data sources and integration of applications into one cohesive solution.
4. custom development, querying and integration with other systems, and
5. The traditional as well as Big Data techniques.

Data management, storage and analytics of Big data captured at the companies and services require the following:

- New innovative non-traditional methods of storage, processing and analytics
- Distributed Data Stores
- Creating scalable as well as elastic virtualized platform (cloud computing)
- Huge volume of Data Stores
- Massive parallelism & High speed networks
- High performance processing, optimization and tuning
- Data management model based on Not Only SQL or NoSQL
- In-memory data column-formats transactions processing or dual in-memory data columns as well as row formats for OLAP and OLTP
- Data retrieval, mining, reporting, visualization and analytics
- Graph databases enable analytics with social network messages, pages and data analytics
- Machine learning or other approaches
- Big data sources: Data storages, data warehouse, Oracle Big Data, MongoDB NoSQL, Cassandra NoSQL.
- Data sources: Sensors, Audit trail of Financial transactions data, external data such as Web, Social Media, weather data, health records data.

Hadoop



Hadoop based Big Data environment

Big Data platform consists of Big Data storage(s), server(s) and data management and business intelligence software. Storage can deploy Hadoop Distributed File System (HDFS), NoSQL data stores, such as HBase. MongoDB, Cassandra. HDFS system is a open source storage system. HDFS is a scaling, self-managing and self-healing file system.
The Hadoop system packages application-programming model. Hadoop is a scalable and reliable parallel computing platform. Hadoop manages Big Data distributed databases.

BIG DATA ANALYTICS APPLICATIONS AND CASE STUDIES

Big Data in Marketing and Sales

Data are important for most aspect of marketing, sales and advertising.A definition of marketing is the creation, communication and delivery of value to customers. Customer (desired) value means what a customer desires from a product. Customer (perceived) value means what the customer believes to have received from a product after purchase of the product. Customer value analytics (CVA) means analyzing what a customer really needs. CVA makes it possible for leading marketers, such as Amazon to deliver the consistent customer experiences.
An example of fraud is borrowing money on already mortgage assets. Example of timely compliances means returning the loan and interest installments by the borrowers.

A few examples in service-innovation are as follows: A company develops software and then offers services like Uber. Another example is of a company which develops software for hiring services, and then offers costly construction machinery and equipment.
Big data is providing marketing insights into (i) most effective content at each stage of a sales cycle, (ii) investment in improving the customer relationship management (CRM), (iii) addition to strategies for increasing customer lifetime value (CLTV), (iv) lowering of customer acquisition cost (CAC). Cloud services use Big Data analytics for CAC, CLTV and other metrics, the essentials in any cloud-based business
Contextual marketing means using an online marketing model in which a marketer sends to potential customers the targeted advertisements, which are based on the search terms during latest browsing patterns usage by customers.

Big data Analytics in detection of marketing Fruads:

Fraud detection is vital to prevent financial loss to users. Transferring customer information to third party, falsifying company information to financial institutions, marketing product with compromising quality, marketing product with service level different from the promised, stealing intellectual property, and much more.
Big Data analytics enable fraud detection. Big Data usages has the following features-for enabling detection and prevention of frauds:

1. Fusing of existing data at an enterprise data warehouse with data from sources such as social media, websites, blogs, e-mails, and thus enriching existing data
2. Using multiple sources of data and connecting with many applications
3. Providing greater insights using querying of the multiple source data
4. Analyzing data which enable structured reports and visualization
5. Providing high volume data mining, new innovative applications and thus leading to new business intelligence and knowledge discovery
6. Making it less difficult and faster detection of threats, and predict likely frauds by using various data and information publicly available.

Big Data Risks

Large volume and velocity of Big Data provide greater insights but also associate risks with the data used Data included may be erroneous, less accurate or far from reality. Analytics introduces new errors due to such data.Companies need to take risks of using Big Data and design appropriate risk management procedures. They have to implement robust risk management processes and ensure reliable predictions. Corporate, society and individuals must act with responsibility.

Big Data Credit Risk Management

Financial institutions, such as banks, extend loans to industrial and household sectors. These institutions in many countries face credit risks, mainly risks of (i) loan defaults, (ii) timely return of interests and principal amount.
Financing institutions are keen to get insights into the following:

- Identifying high credit rating business groups and individuals,
- Identifying risk involved before lending money
- Identifying industrial sectors with greater risks
- Identifying types of employees and businesses with greater risks
- Anticipating liquidity issues (availability of money for further issue of credit and rescheduling credit over the years.

Big Data and Healthcare

Big Data analytics in health care use the following data sources:
(i) clinical records,
(ii) pharmacy records,
(3) electronic medical records
(4) diagnosis logs and notes and
(v) additional data, such as deviations from person usual activities, medical leaves from job, social interactions.

Healthcare analytics using Big Data can facilitate the following:

1. Provisioning of value-based and customer-centric healthcare,
2. Utilizing the Internet of Things' for health care
3. Preventing fraud, waste, abuse in the healthcare industry and reduce healthcare costs
4. Improving outcomes
5. Monitoring patients in real time.

Big Data in Medicine

Big Data analytics deploys large volume of data to identify and derive intelligence using predictive models about individuals. Big Data driven approaches help in research in medicine which can help patients. Big Data offers potential to transform medicine and the healthcare system
1 Aggregating large volume and variety of information around from multiple sources the DNAS, proteins, and metabolites to cells, tissues, organs, organisms, and ecosystems, that can enhance the understanding of biology of diseases. Big data creates patterns and models by data mining and help in better understanding and research,

2. Deploying wearable devices data, the devices data records during active as well as inactive periods, provide better understanding of patient health, and better risk profiling the user for certain diseases,

Big Data in Advertising

The impact of Big Data is tremendous on the digital advertising industry. The digital advertising industry sends advertisements using SMS, e-mails, WhatsApp, LinkedIn, Facebook, Twitter and other mediums. Big data real time analytics for faster insights, emerging trends and patterns, and gain actionable insights for facing competitions from similar products in digital advertising and building relationships

Big Data captures data of multiple sources in large volume, velocity and variety of data unstructured and enriches the structured data at the enterprise data warehouse. Big data real time analytics provide emerging trends and patterns, and gain actionable insights for facing competitions from similar products. The data helps digital advertisers to discover new relationships, lesser competitive regions and areas.

Success from advertisements depend on collection, analyzing and mining. The new insights enable the personalization and targeting the online, social media and mobile for advertisements called hyper-localized advertising.