Big Data Analytics

[18CS72]

Introduction

- A field to analyze and to extract information about the big data involved in the business or the data world so that proper conclusions can be made is called big data Analytics.
- These conclusions can be used to predict the future or to forecast the business.
- Also, this helps in creating a trend about the past. Skilled professionals in statistics and engineering with domain knowledge are needed in the analysis of big data as the data is huge, and analysis needs proper determination and skillset.
- This data is more complex that it cannot be dealt with traditional methods of analysis.

How does Data Analytics work?

- Data Collection- collect data from different sources- Internet, web server logs, cloud applications, mobile application, social media ,email,mobile phone records, sensors
- Data Processing- Organize, Configure and Partition data for queries
- Data Cleansing- Scrub data using scripting or enterprise s/w for errors and inconsistencies
- Data Analyzed: Data is analyzed using tools
- Data Mining
- Predictive Analysis
- Machine Learning Algorithm
- Deep Learning
- Text mining statistical analysis
- Artificial Intelligence
- Data Visualization tools

Big Data Analytic Technologies and Tools

Hadoop- Open source framework for storing and processing structured and unstructured data

Predictive analysis- Machine Learning and statistical algorithms used in Fraud detection, risk mgt, marketing

Stream analytic tools- filters, aggregate and analyze big data in different formats and platforms

Distributed storage-Data is replicated on a non RDBMS

NoSQL- Non RDBMS for raw and unstructured data

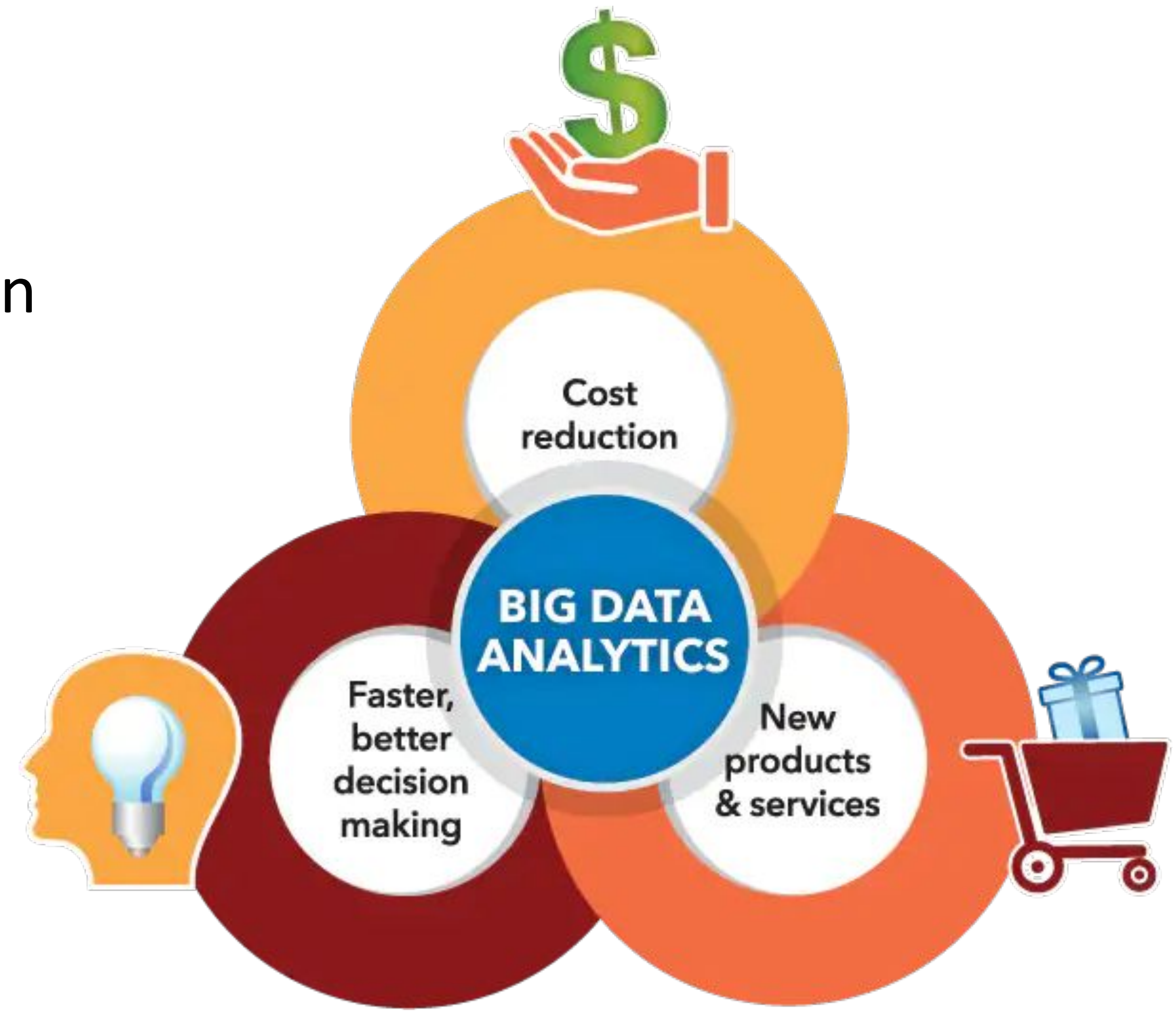
Data Integration s/w- Data to be streamlined across different platforms: Amazon EMR , MongoDB, Hadoop, Apache

Uses of BDA

- Customer acquisition & retention

E.g. Amazon, Netflix, Spotify etc.

- Targeted ads
- Product development
- Price Optimisation
- Risk management
- Improved decision making



Course Outcomes

Cos	Description	BL
C402.1	Identify the features and applications of Big Data.	L3
C402.2	Analyze Hadoop framework and Hadoop Distributed file system.	L4
C402.3	Explain the concepts of NoSQL using MongoDB and Cassandra.	L5
C402.4	Apply MapReduce programming model to process big data with Hadoop.	L3
C402.5	Make use of machine learning algorithms for big data and web contents, Social Networks to provide analytics with relevant visualization tools.	L3

Outline

Module-1:Introduction to Big Data Analytics

Module-2:Introduction to Hadoop

Module-3:NoSQL Big Data Management, MongoDB and Cassandra

Module-4:MapReduce, Hive and Pig

Module-5:Machine Learning Algorithms for Big Data Analytics

TEXT BOOKS :

1. Raj Kamal and Preeti Saxena, “Big Data Analytics Introduction to Hadoop, Spark, and Machine-Learning”, McGraw Hill Education, 2018 ISBN: 9789353164966, 9353164966
2. Douglas Eadline, "Hadoop 2 Quick-Start Guide: Learn the Essentials of Big Data Computing in the Apache Hadoop 2 Ecosystem", 1stEdition, Pearson Education, 2016. ISBN13: 978-9332570351

REFERENCE BOOKS :

1. Tom White, “Hadoop: The Definitive Guide”, 4th Edition, O’Reilly Media, 2015.ISBN-13: 978- 9352130672
2. Boris Lublinsky, Kevin T Smith, Alexey Yakubovich, "Professional Hadoop Solutions", 1st Edition, Wrox Press, 2014ISBN-13: 978-8126551071
3. Eric Sammer, "Hadoop Operations: A Guide for Developers and Administrators",1st Edition, O'Reilly Media, 2012.ISBN-13: 978-9350239261
4. Arshdeep Bahga, Vijay Madisetti, "Big Data Analytics: A Hands-On Approach", 1st Edition, VPT Publications, 2018. ISBN-13: 978-0996025577

Module-1: Introduction to Big Data Analytics

Topics Covered:

- Big Data: Classification of Data, Definitions, Characteristics, Types.
- Big Data classification, data handling techniques.
- Scalability and Parallel processing: Analytics scalability to Big Data, massively parallel processing platforms.
- Cloud Computing, Grid and Cluster Computing, Volunteer computing.
- Designing Data Architecture: Data architecture design, managing data for analysis.
- Data Sources, Quality, Pre-Processing and Storing
- Data Storage and Analysis: Data storage and management: Traditional systems, big data storage.
- Big data platforms, big data analytics.
- Big Data Analytics Applications and Case Studies: Big data in Marketing and sales, Health care
- Big data in medicine, Advertising.

Introduction to Big Data

What is Data?

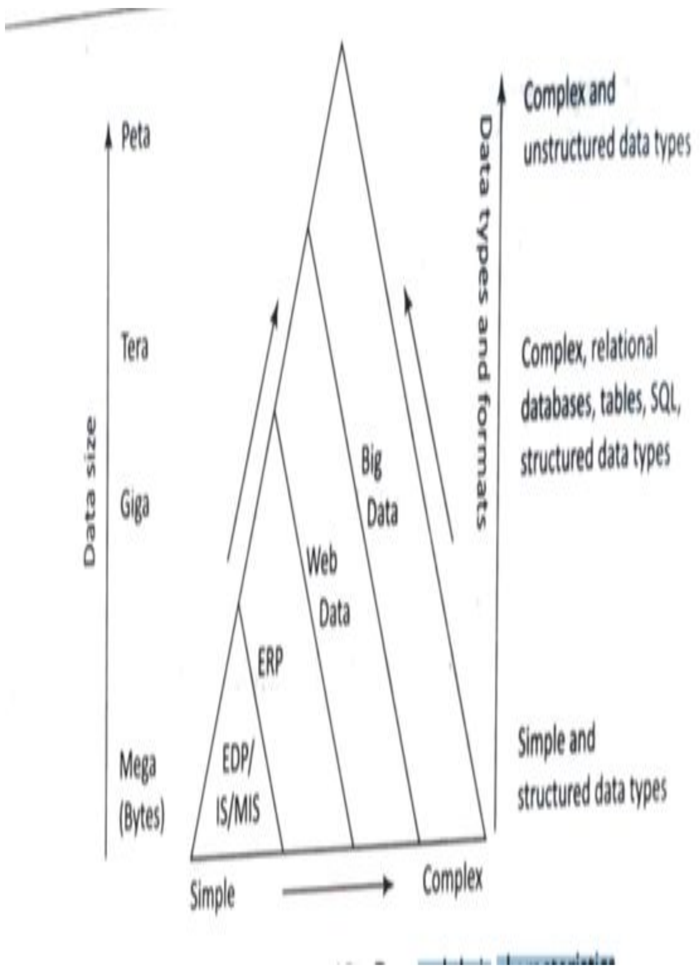
- Anything that is recorded is data.

What is Big Data?

- Bigdata is a term used to describe a collection of data that is huge in size and yet growing exponentially with time.
- Big data is an umbrella term for a collection of data sets so large and complex that it becomes difficult to process them using traditional data management tools.
- In 2000, there were 800,000 Petabytes of data in the world. It is already grown to 35 zettabytes in theyear 2020.
- Big data means that the data is unable to be handled and processed by most current information system or methods
- Challenges in big data include the entire range of operations from capture, curation, storage, search, sharing, analysis, and visualization.
- Big data is more valuable when analyzed as a whole.

Need of Big Data

- The rise in technology has led to the production and storage of voluminous amounts of data.
- Earlier megabytes (10° B) were used but nowadays petabytes (10" B) are used for processing, analyse new facts and generating new knowledge.
- Conventional systems for storage, processing and the rise in technology has led to the production and storage of voluminous amounts of data.
- Discovering analysis and formats, increasing pose challenges in large growth in volume of data, variety of data, various forms and formats: complexity, faster generation of data and need of quickly processing, analyzing and usage.



Need of Big Data(Contd..)

- As size and complexity increase, the proportion of unstructured data types also increase.
- Example of a traditional tool for structured data storage and querying is RDBMS.
- Volume, velocity and variety (3Vs) of data need the usage of number of programs and tools for analyzing and processing at a very high speed.
- When integrated with the Internet of Things, sensors and machines data, the veracity of data is an additional V.
- Application Programming Interface (API) refers to a software component which enables a user to a an application, service or software that runs on a local or remote computing platform.
- Data Model refers to a map or schema, which represents the inherent properties showing groupings of data elements.
- Data Repository refers to a collection of data. A data-seeking program relies upon the data repository for reporting.
- Ex: Repositories are database, flat file and spreadsheet.
- Data Store refers to a data repository of a set of objects.

- **Data store** is a general concept for data such as database, repositories, relational database, flat file, spreadsheet, mail server, web server and directory services. The objects in data store model are instances of the classes which the database schemas define. A data store may consist of multiple schemas or may consist of data in only one schema.

  Ex: of only one scheme for a data store is a relational database.

- **Distributed Data Store** refers to a data store distributed over multiple nodes. Apache Cassandra is one example of a distributed data store. Database (DB) refers to a grouping of tables for the collection of data. A table ensures a systematic way for accessing, updating and managing data. A database pertains to the applications, which access them.

- **A database** is a repository for querying the required information for analytics, processes, intelligence and knowledge discovery. The databases can be distributed across a network consisting of servers and data warehouses.

- **Flat File** means a file in which data cannot be picked from in between and must be read from the beginning to be interpreted. A file consisting of a single-table file is called a flat file.

  Ex: CSV (comma-separated value) file. A flat file is also a data repository.

- Flat File Database refers to a database in which each record is in a separate row unrelated to each other. CSV File refers to a file with comma-separated values

Big Data -Definition

- 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 -- (Gartner' 2012)
- A collection of data sets is so large or complex that traditional data processing applications are inadequate. - Wikipedia .
- Data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges. - (traditional database of authoritative definitions)
- Big Data refers to data sets whose size is beyond the ability of typical database software tool to capture, store, manage and analyse.

Characteristics Of Big Data

A well known definition of Big Data has 3Vs

- Volume (Data is Huge)
- Variety (Data is coming from multiple sources in multiple forms)
- Velocity (Data is changing with time and coming with a velocity)

Big Data Characteristics

Characteristics of Big Data, called 3Vs (and 4Vs also used) are: Volume The phrase "Big Data' contains the term big, which is related to size of the data and hence the characteristic.

- **Size :** Amount or quantity of data, which is generated from an application(s). The size determines the processing considerations needed for handling that data.
- **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 means that the type to which Big Data belongs to is also an important characteristic that needs to be known for proper processing of data. This characteristic helps in effective use of data according to their formats, thus maintaining the importance of Big Data.
- **Veracity :**Quality of data captured, which can vary greatly, affecting its accurate analysis.
- The 4Vs (i.e. volume, velocity, variety and veracity) data need tools for mining, discovering patterns, business intelligence, artificial intelligence (AI), machine learning (ML), text analytics, descriptive and predictive analytics, and the data visualization tools.

Example of features of 3Vs in Big Data and application.

Consider satellite images of the Earth's atmosphere and its regions. The Volume of data from the satellites is large. A number of Indian satellites, such as KALPANA, INSAT-1A and INSAT-3D generate this data. Foreign satellites also generate voluminous data continuously. Satellites record the images of full disk and sectors, such as east and west Asia sectors and regions.

- Velocity is also large. A number of satellites collect this data round the clock. Big Data analytics helps in drawing of maps of wind velocities, temperatures and other weather parameters.
- Variety of images can be in visible range, such as IR-I (infrared range -1), IR-2(infrared range .2),shortwave infrared (SWIR), MIR (medium range R) and color composite.
- Data Veracity: arises due to poor resolutions used for recording or noise in images due to signal impairments
- Data processing needs increased speed of computations due to higher volumes. Need of data management, storage and increased analytics requires new innovative non-traditional methods.
- Big Data of satellites helps in predicting weather, and mapping of different crops and from that estimating the expected crop yield.

Types of Big Data

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

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

Master data - data for facial recognition and for the name, date of birth, marriage anniversary, gender, location and income category 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 - 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.

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

Humans also records their experiences in ways such as writing these in notebooks or diaries, taking photographs or audio and video clips.

Human-sourced information is now almost entirely digitized and stored everywhere from personal computers to social networks. Such data are loosely structured and often ungoverned.

Examples of machine-generated data are:

1. Data from computer systems: Logs, web logs, security/surveillance systems, videos/images etc.

2. Data from fixed sensors: Home automation, weather sensors, pollution sensors, traffic sensors etc.

3. Mobile sensors (tracking) and location data.

Big Data Classification

- Big Data can be classified on the basis of its characteristics that are used for designing data architecture for processing and analytics.

Table 1.1 Various classification methods for data and Big Data

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 usages	Human, business process, knowledge discovery, enterprise applications, Data Stores

Definitions of Data

- 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 - Computing
- Data is information presented in numbers, letters, or other form. [Electrical Engineering. Circuits. Computing and Control]
- Data is information from series of observations, measurements or facts" -Science
- Data is large scale of integration and presence of data on web servers.

- Web is a part of the Internet that stores web data in the form of documents and other web resources.
- URLs enable the access to web data resources.
- 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.

Examples of Web Data:

- Wikipedia is a web-based, free-content encyclopedia project supported by the Wikimedia Foundation.
- Google Maps is a provider of real-time navigation, traffic, public transport and nearby places by Google Inc.
- McGraw-Hill Connect is a targeted digital teaching and learning environment that save students and instructors' time by improving student performance for a variety of critical outcomes.
- Oxford Bookstore is an online book store where people can find any book that they wish to buy from millions of titles

Classification of Data

Structured:

- Structured data conform and associate with data schemas and data models.
- 15-20% data are in structured or semi-structured form.
- An 'Employee' table in a database is an example of Structured Data

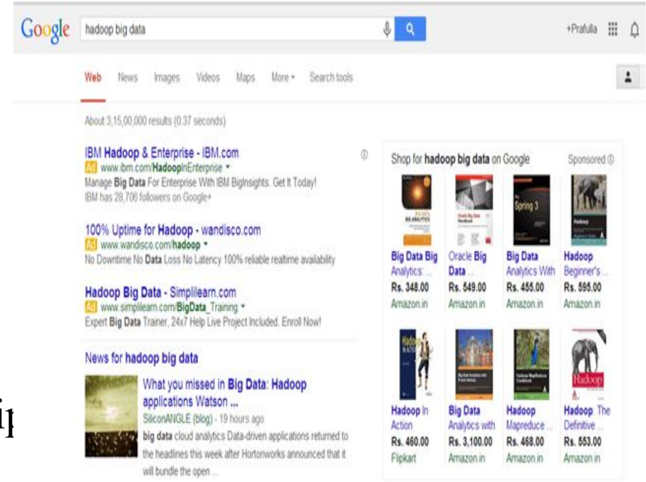
EMPNO	ENAME	JOB	HIREDATE	MGR	SAL	COMM	DEPTNO
7369	SMITH	CLERK	17-DEC-80	7902	800		20
7499	ALLEN	SALESMAN	20-FEB-81	7698	1600	300	30
7521	WARD	SALESMAN	22-FEB-81	7698	1250	500	30
7566	JONES	MANAGER	02-APR-81	7839	2975		20
7654	MARTIN	SALESMAN	28-SEP-81	7698	1250	1400	30
7698	BLAKE	MANAGER	01-MAY-81	7839	2850		30
7782	CLARK	MANAGER	09-JUN-81	7839	2450		10
7788	SCOTT	ANALYST	19-APR-87	7566	3000		20
7839	KING	PRESIDENT	17-NOV-81		5000		10
7844	TURNER	SALESMAN	08-SEP-81	7698	1500	0	30
7876	ADAMS	CLERK	23-MAY-87	7788	1100		20
7900	JAMES	CLERK	03-DEC-81	7698	950		30
7902	FORD	ANALYST	03-DEC-81	7566	3000		20
7934	MILLER	CLERK	23-JAN-82	7782	1300		10

Structured data enables the following:

- data insert, delete, update and append
- Indexing to enable faster data retrieval
- Data mining and analytics, data retrieval, data reporting, data visualization and machine-learning Big Data tools.
- Eg: names, addresses, credit card numbers, geolocation, and so on.

Unstructured:

- Do not conform and associate with any data models.
- Eg: media, text, social media activity, surveillance imagery, and so on.



Semi-structured:

- does not conforms to a data model but has some structure.
- Cannot be stored in the form of rows and columns as in Databases
- contains tags and elements (Metadata) which is used to group data
- Eg: E-mails, XML and other markup languages, Binary executables, TCP/IP packets, Zi files, Integration of data from different sources
- Web pages

```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>
```

Features	Structured Data	Semi Structured Data	Unstructured Data
Transaction management	Matured transaction and various concurrency techniques	Transaction is adapted from DBMS not matured	No transaction management and no concurrency
Version management	Versioning over tuples,row,tables	Versioning over tuples or graph is possible	Versioned as a whole
Flexibility	It is schema dependent and less flexible	It is more flexible than structured data but less flexible than unstructured data	It is more flexible and there is absence of schema
Scalability	It is very difficult to scale DB schema	It's scaling is simpler than structured data	It is more scalable.
Robustness	Very robust	New technology, not very spread	–
Query performance	Structured query allow complex joining	Queries over anonymous nodes are possible	Only textual queries are possible

Big Data Handling Techniques

Techniques deployed for Big Data storage, applications, data management and mining and analytics:

- Huge data volumes storage, data distribution, high-speed networks and high-performance computing
- Applications scheduling using open source, reliable, scalable, distributed file system. Distributed database.
- Open-source tools which are scalable, elastic and provide virtualized environment, clusters of data nodes, task and thread management
- Data management using NoSQL, document database, column-oriented database. graph database
- Other form of databases used as per needs of the applications and in-memory data management using columnar or Parquet formats during program execution.

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.
- Processing of this much distributed data within a short time and at minimum cost is problematic.
- Scalability, scaling up, scaling out in distributed computing, Massively Parallel Processing (MPP), cloud, grid, volunteering computing systems .

Convergence of Data Environments and Analytics

- Big data can co-exist with traditional data store.
- Traditional data stores use RDBMS tables or data warehouse.
- 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 workload and complexity exceed the system capacity, scale it up and scale it out.

Analytics Scalability to Big Data

- Vertical scalability is scaling up the given system's resources and increasing the system's analytics, reporting and visualization capabilities.
- Additional way to solve problems of greater complexities.
- Scaling up means designing the algorithm according to the architecture that uses resources efficiently.
- Example, x terabyte of data take time for processing, code size with increasing complexity increase by factor n, then scaling up means that processing takes equal, less or much less than (n x t).
- Horizontal scalability means increasing the number of systems working in coherence 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.
- If r resources in a system process x terabyte of data in time t, then the (p x x) terabytes process on p parallel distributed nodes such that the time taken up remains t or is slightly more than t (due to the additional time required for IPC).
- Implement it on a bigger machine with more CPUs for greater volume, velocity, variety and complexity of data.

Analytics Scalability to Big Data

- Software will perform better on a bigger machine.
- Faster CPUs, RAM ,hard disks, motherboards will be expensive compared to the extra performance achieved by efficient design of algorithms.
- 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
- Alternative ways for scaling up and out processing of analytics software and Big Data analytics deploy the Massively Parallel Processing Platforms (MPPs), cloud, grid, clusters, and distributed computing software.

Massively Parallel Processing Platforms

- Scaling uses parallel processing systems.
- Many programs are so large and complex to execute them on a single computer system (limited memory)
- Required to enhance (scale) up the computer system or use massive parallel processing (MPPs) platforms.
- Parallelization of tasks can be done at several levels:

 Distributing separate tasks onto separate threads on the same CPU

 Distributing separate tasks onto separate CPUs on the same computer

 Distributing separate tasks onto separate computers

- A solution for Big data processing is to perform parallel and distributed computing in a cloud computing environment

Massively Parallel Processing Platforms Contd..

- When making software, draw the advantage of multiple computers (or even multiple CPUs within the same computer)
- Software which need to be able to parallelize tasks.
- 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

- This Model uses cloud, grid or clusters, which process and analyse big and large datasets on distributed computing nodes connected by high-speed networks.
- Big Data processing uses a parallel, scalable and no-sharing program model, such as MapReduce, for computations on it.

Table 1.2 Distributed computing paradigms

Distributed computing on multiple processing nodes/clusters	Big Data > 10 M	Large datasets below 10 M	Small to medium datasets up to 1 M
Distributed computing	Yes	Yes	No
Parallel computing	Yes	Yes	No
Scalable computing	Yes	Yes	No
Shared nothing (No in-between data sharing and inter-processor communication)	Yes	Limited sharing	No
Shared in-between between the distributed nodes/clusters	No	Limited sharing	Yes

Cloud Computing

- "Cloud computing is a type of Internet-based computing that provides shared processing resources and data to the computers and other devices on demand."
- Best approach for data processing is to perform parallel and distributed computing in a cloud computing environment.
- Cloud usages circumvent the single point failure due to failing of one node.
- Multiple nodes perform automatically and interchangeably.
- It offers high data security compared to other distributed technologies.
- Cloud resources - Amazon Web Service (AWS) Elastic Compute Cloud (Ec2), Microsoft Azure or Apache CloudStack.
- Amazon Simple Storage Service ($3) 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:

- On-demand service
- Resource pooling
- Scalability
- Accountability.
- Broad network access.

- Cloud services can be accessed from anywhere and at any time through 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.
- Computing infrastructure - data storage device, development platform, database, computing power or software applications.

Cloud Services

3 Types

Infrastructure as a Service (IaaS):

- Access to resources like hard disks, network connections, databases storage, data center and virtual server spaces .
- Ex: Tata Communications, Amazon data centers and virtual servers.
- Apache Cloud Stack - Open source software for deploying and managing a large network of virtual machines and offers public cloud services which provide highly scalable Infrastructure

Platform as a Service (PaaS):

- Provides the runtime environment to allow developers to build applications and services
- Software at the clouds support and manage the services, storage, networking, deploying, testing, collaborating, hosting and maintaining applications.
- Ex: Hadoop Cloud Service (IBM BigInsight, Microsoft Azure HD Insights, Oracle Big Data Cloud Service.

Software as a Service (SaaS):

- Provides software applications as a service to end-users
- Software applications are hosted by a service provider and made available to customers over the Internet.
- Ex: SQL GoogleSQL, IBM BigSQL, HPE Vertica. Microsoft Polybase and Oracle Big Data SQL

Grid and Cluster Computing

Grid Computing

- Distributed computing where group of computers from several locations are connected with each other to achieve a common task.
- Resources are heterogeneously and geographically disperse.
- Group of computers spread over remotely comprise a grid.
- Single grid 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(individuals, organizations and resources)
- Suits data-intensive storage better than storage of small objects of few millions of bytes.
- To achieve maximum benefit , should be used for a large amount of data which can distribute over grid nodes.
- Besides data grid, the other variation of grid, i.e., computational grid focuses on computationally intensive operations.
- Features of Grid Computing Grid computing, similar to cloud computing, is scalable.

Limitations:

- Single point failure or failure of nodes
- Storage capacity varies with no of users, instances and amount of data transferred.

Advantages:

- Reducing infrastructure costs and raising load capabilities.

Cluster Computing

- Cluster is a group of computers connected by a network.
- Group works together to accomplish the same task.
- Used for load balancing.
- Shift processes between nodes to keep an even load on the group of connected computers.
- Hadoop architecture uses the similar methods

Table 1.3 Grid computing and related paradigms

Distributed computing	Cluster computing	Grid computing
Loosely coupled Heterogeneous Single administration	• Tightly coupled • Homogeneous • Cooperative working	• Large scale • Cross organizational • Geographical distribution • Distributed management

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
- Ex: Science-related projects executed by universities or academia in general

Some issues with volunteer computing systems are:

- Volunteered computers heterogeneity
- Drop outs from the network over time
- Sporadic availability
- Incorrect results at volunteers are unaccountable as they are essentially from anonymous volunteers,

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
- Defines how Big Data solution will work, the core components (hardware, database, software, storage) used, flow of information, security and more.
- Characteristics of Big Data make architecture a complex process.
- Faster additions of new technological innovations increase the complexity in design.
- The requirements for offering competing products at lower costs in the market make the designing task more challenging for a Big Data architect.

Layer 5 Data consumption	Export of datasets to cloud, web etc.	Datasets usages: BPs, BIs, knowledge discovery	Analytics (real-time, near real-time, scheduled batches), reporting, visualization	
Layer 4 Data processing	Processing techn-ology: MapReduce, Hive, Pig, Spark	Processing in real-time, scheduled batches or hybrid	Synchronous or asynchronous processing	
Layer 3 Data storage	Considerations of types (historical or incremental), formats, compression, frequency of incoming data, patterns of querying and data consumption	Hadoop distributed file system (scaling, self-managing and self-healing), Spark, Mesos or S3	NoSQL data stores – Hbase, MongoDB, Cassandra, Graph database	
Layer 2 Data ingestion and acquisition	Ingestion using Extract Load and Transform (ELT)	Data semantics (such as replace, append, aggregate, compact, fuse)	Pre-processing (validation, transformation or transcoding) requirement	Ingestion of data from sources in batches or real time
Layer 1 Identification of internal and external sources of data	Sources for ingestion of data	Push or pull of data from the sources for ingestion	Data types for database, files, web or service	Data formats: structured, semi- or unstructured for ingestion

Figure 1.2 Design of logical layers in a data processing architecture, and functions in the layers

Data Architecture Design

- **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

(v) Data consumption in the number of programs and tools.

Layer 1: Logical layer 1 (L1) is for identifying data sources, which are extrenal, internal or both.

Layer 2: The layer 2 (L2) is for data-ingestion.

- Data ingestion means a process of absorbing information, just like the process of absorbing nutrients medications into the body by eating or drinking them .
- Ingestion is the process of obtaining and importing data for immediate use or transfer.
- Ingestion may be in batches or in real time using pre-processing or semantics.

Layer 3: L3 is for storage of data from the L2 layer.

Layer 4: L4 is for data processing using software, such as MapReduce, Hive, Pig or Spark.

Layer 5: L5 is for data consumption. Data is used in analytics, Visualizations, reporting, export to cloud or web servers.

Layers considers the following aspects in a design

Layer 1 :

- Amount of data needed at ingestion layer 2 (L2)
- Push from LI or pull by L2 as per the mechanism for the usages
- Source data-types: Database, files, web or service
- Source formats, i.e., semi-structured, unstructured or structured.

Layer 2:

- Ingestion and ELT processes in real time

Layer 3 :

- Data storage type(historical or incremental),format compression, frequency and consumption requirements
- Data storage using Hadoop DFS or NoSQL

Layer 4 :

- Data Processing software: MapReduce, Hive, Pig, Spark, Mahout
- Processing in scheduled batches or real time or hybrid
- Synchronous or asynchronous

Layer 5 :

- Data Integration
- Datasets usages for reporting and visualization
- Analytics(real time, near real time, scheduled batches0
- Export of datasets to cloud, web or other systems

Managing Data for Analysis

- Data managing means enabling, controlling, protecting, delivering and enhancing the value of data and information asset.

- Reports, analysis and visualizations need well-defined data. Data management also enables data usage in applications.

- The process for managing needs to be well defined for fulfilling requirements of the applications Data management functions include:

- Data assets creation, maintenance and protection
- Data governance, which includes establishing the processes for ensuring the availability, usability integrity, security and high-quality of data. The processes enable trustworthy data availability for analytics, followed by the decision making at the enterprise.
- Data architecture creation, modelling and analysis
- Database maintenance, administration and management system. For example, RDBMS (relational database management system), NoSQL
- Managing data security, data access control, deletion, privacy and security
- Managing the data quality
- Data collection using the ETL process

Managing Data for Analysis Contd..

- Managing documents, records and contents
- Creation of reference and master data, and data control and supervision
- Data and application integration
- Integrated data management, enterprise-ready data creation, fast access and analysis, automation and simplification of operations on the data
- Data warehouse management
- Maintenance of business intelligence
- Data mining and analytics algorithms.

Data Sources

- Applications, programs and tools use data.
- Sources can be external, such as sensors, trackers, web logs, computer systems logs and feeds.
- Sources can be machines, which source data from data-creating programs.
- Data sources can be structured, semi-structured, multi-structured or unstructured.
- Data sources can be social media.
- A source can be internal. Sources can be data repositories, such as database, relational database, flat file, spreadsheet, mail server, web server, directory services, even text or files such as comma-separated values (CSV) files.
- Source may be a data store for applications.

Structured Data Sources

- Data source for ingestion, storage and processing can be a file, database or streaming data. The source may be on the same computer running a program or a networked computer.
- Ex: SQL Server, MySQL, Microsoft Access database,Oracle DBMS, IBM DB2, Informix,
- Amazon SimpleDB or a file-collection directory at a server.
- A data source name implies a defined name, which a process uses to identify the source. The name needs to be a meaningful name.
- Ex: a name which identifies the stored data in student grades during processing.
- The data source name could be StudentName_Data_Grades.
- A data dictionary enables references for accesses to data.
- The dictionary consists of a set of master lookup tables.
- The dictionary stores at a central location.
- The central location enables easier access as well as administration of changes in sources. The name of the dictionary can be University Students_DataPlusGrades.
- A master-directory server can also be called Name Node.

Microsoft applications consider two types of sources for processing:

- Machine sources on computing nodes(servers)

 A machine identifies a source by the user-defined name, driver- manager name and source-driver name.

- File sources are stored files

Oracle applications consider two types of data sources:

- Database, identifies the database information that the software needs to connect to Database data sources and logic-machine data sources in Oracle applications database, and
- Logic-machine, identifies the machine which runs batches of applications and master business functions." Source definition identifies the machine. The source can be on a network.

Data sources can point to:

- A database in a specific location or in a data library of OS
- A specific machine in the enterprise that processes logic
- A data source master table which stores data source definitions. The table may be at a centralized source (enterprise server) or at server-map for the source.

Unstructured Data Sources

- Data sources are distributed over high-speed networks.
- Data need high velocity processing. Sources are from distributed file systems.
- Sources are of file types, such as .txt (text file), .csv (comma separated values file). Data may be as key-value pairs, such as hash key-values pairs.
- Data may have internal structures, such as in e-mail, Facebook pages, twitter messages etc.
- The data do not model, reveal relationships, hierarchy relationships or object-oriented features, such as extensibility.

Data Sources-Sensors, Signals and GPS

- Data sources can be sensors, sensor networks, signals from machines, devices, controllers and intelligent edge nodes of different types in the industry M2M communication and the GPS systems.
- Sensors are electronic devices that sense the physical environment.
- Sensors are devices which are used for measuring temperature, pressure, humidity, light intensity, traffic in proximity, acceleration, locations, object(s) proximity, orientations and magnetic intensity, and other physical states and parameters.
- Sensors play an active role in the automotive industry.
- RFIDs and their sensors play an active role in RFID based supply chain management, and tracking parcels, goods and delivery.
- Sensors embedded in processors, which include machine-learning instructions, and wireless communication capabilities are innovations.

Data Quality

- Quality is high when it represents the real-world construct to which references are taken.
- High quality means data, which enables all the required operations, analysis, decisions, planning and knowledge discovery correctly.
- A definition for high quality data for artificial intelligence applications, can be data with five R's as follows: Relevancy, recency, range, robustness and reliability.
- A uniform definition is difficult.
- A reference can be made to a set of values of quantitative or qualitative conditions, which must be specified to say that data quality is high or low.

Data Integrity

- Maintenance of consistency and accuracy in data over its usable life.
- Data should be incorruptable

Data Pre-processing

- Data pre-processing is an important step at the ingestion layer .
- 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

(ii) Data cleaning, editing, reduction and/or wrangling

(iv) Data validation, transformation or transcoding

(v).ELT processing.

Data Cleaning

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 can generate in a system in many formats when it is obtained from the web.
- Data cleaning tools help in refining and structuring data into usable data.

EX: Open Refine and Data Cleaner.

Data Enrichment

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

Data Editing

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, (ii) automatic, (iv) aggregating and (v) distribution.

Data Reduction

- 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 reduction uses editing, scaling, coding, sorting, collating, smoothening, interpolating and preparing tabular summaries.

Data Wrangling

- Process of transforming and mapping the data.
- Results from analytics are then appropriate and valuable.

Ex: mapping enables data into another format, which makes it valuable for analytics and data visualizations.

Data Format used during Pre-Processing

Examples of formats for data transfer from

- data storage
- analytics application
- service
- cloud

Need of data format conversion of data CSV, JSON, key-value pairs or other data from Data Store; for example, in the form of tables

- Comma-separated values CSV
- Java Script Object Notation (JSON) as batches of object arrays or resource
- Tag Length Value (TLV)
- Key-value pairs
- Hash-key-value pairs

CSV Format

- Table or MS excel file can be converted to CSV format
- CSV is a plain text file which stores table data of no’s and text

Subject Code	Subject Name	Grade
CS101	“Theory of Computation”	7.8
CS102	“Computer Architecture”	7.2

CSV file:

Subject Code, Subject Name, Grade

CS101,””Theory of Computation””,7.8


```
{
    "status": "0, Good, the operation completed successfully.",
    "info": "Position Loading wagon, mm",
    "calvints": 1499990404234,
    "sourcets": "2017-07-14 00:00:04.064000",
    "value": "-10.0036686659",
    "id": "ns=2;s=/Nck/MachineAxis/aaIm [3]",
    "tag": "aaIm_Z",
    "serverts": "2017-07-14 00:00:04.124000",
    "type": "Double",
    "origin": "GU-1"
}
```

Data storage and Analysis

Data Store with Structured or Semi-Structured Data

- Traditional systems use structured or semi-structured data.

Sources of Structured data store

- Traditional RDBMS – MySQL ,DB2,enterprise server and data warehouse.
- Business process data stores business events- registering, taking order, generating invoice and managing products in pre defined formats- Transaction records, tables, relationship and metadata to build business data
- Commercial transactions
- Banking/stock records
- E- Commerce transaction data

Sources of Data Store of semi- structured data

- XML and JSON semi- structured documents
- CSV file

SQL

- An RDBMS uses SQL (Structured Query Language).
- SQL is a language for viewing or changing (update, insert or append or delete) databases. It is a language for data access control, schema creation and data modifications.
- SQL was originally based on the tuple relational calculus and relational algebra. SQL can embed within other languages using SQL modules, libraries and pre-compilers.

SQL does the following:

- SQL is language for data querying, updating, inserting, appending and deleting the databases.

1. Create schema, which is a structure which contains description of objects (base tables, views, constraints) created by a user. The user can describe the data and define the data in the database.

2. Create catalog which consists of a set of schemas which describe the database.

3. Data Definition Language (DDL) for the commands which depicts a database, that include creating, altering and dropping of tables and establishing the constraints. A user can create and drop databases and tables, establish foreign keys, create view, stored procedure, functions in the database etc.

4. Data Manipulation Language (DML) for commands that maintain and query the database. A user can manipulate (INSERT/UPDATE) and access (SELECT) the data.

5.Data Control Language (DCL) for commands that control a database, and include administering of privileges and committing. A user can set (grant, add or revoke) permissions on tables, procedures and views.

- A relational DB is a collection of data in multiple tables relate to each other through special fields, called keys (primary key, foreign key and unique key)
- Relational databases provide flexibilities.
- Ex: MysQL PostGreSQL, Oracle database. Informix. IBM DB2 and Microsoft SQL server.

Large Data Storage using RDBMS

- Relational database is collection of data into multiple tables which relates to each other through special fields, called keys.
- RDBMS tables store data in a structured form. The tables have rows and columns.
- Data management of Data Store includes the provisions for privacy and security, data integration, compaction and fusion.
- The systems use machine generated data, human-sourced data, and data from business processes (BP) and business intelligence (BI).
- A set of keys and relational keys access the fields at tables, and retrieve data using queries (insert, modify, append. join or delete).
- RDBMSs use software for data administration also.

Distributed Database Management System

- A distributed DBMS (DDBMS) is a collection of logically interrelated Distributed DB is a collection database at multiple system over a computer network.

Features of a distributed database system are:

- A collection of logically related databases.
- Cooperation between databases in a transparent manner.
- Transparent means that each user within the system may access all of the data within all of the databases as if they were a single database.
- 'location independent' which means the user is unaware of where the data is located, and it is possible to move the data from one physical location to another without affecting the user.

In-Memory Column Formats Data

- A columnar format in-memory allows faster data retrieval when only a few columns in a table need to be selected during query processing or aggregation.
- Data in a column are kept together in-memory in columnar format.
- A single memory access, therefore, loads many values at the column.
- Address increment to a next memory address for the next value is fast when compared to first computing the address of the next value, which is not the immediate next address.
- Online Analytical Processing (OLAP) in real-time transaction processing is fast when using in-memory column format tables. OLAP enables real-time analytics.
- The CPU accesses all columns in a single instance of access to the memory in columnar format in-memory data-storage.
- OLAP enables online viewing of analyzed data and visualization up to the desired granularity enables view by rolling up or drilling
- OLAP enables obtaining online summarized information and automated reports for a large database.

In-Memory Row Format Databases

- A row format in-memory allows much faster data processing during OLTP .
- Each row record has corresponding values in multiple columns and the on-line value store at the consecutive memory addresses in row format.
- A specific day's sale of five different chocolate flavours is stored in consecutive columns c to c+5 at memory.
- A single instance of memory accesses loads values of all five flavours at successive columns during online processing.

Ex: Total number of chocolates sold computes online.

- Data is in-memory row-formats in stream and event analytics. The stream analytics method does continuous computation that happens as data is flowing through the system.
- Event analytics does computation on event and use event data for tracking and reporting events.

Enterprise Data-Store Server and Data Warehouse

- Enterprise data, after data cleaning process, integrate with the server data at warehouse.
- Enterprise data server use data from several distributed sources which store data using various technologies.
- All data merge using an integration tool.
- Integration enables collective viewing of the datasets at the data warehouse Enterprise data integration may also include integration with applications), such as analytics, visualization, reporting, business intelligence and knowledge discovery.
- Heterogeneous systems execute complex integration processes when integrating at an enterprise server or data warehouse.

Some standardized business processes defined in the Oracle application-

1.Integrating and enhancing the existing systems and processes
2.Business intelligence
3. Data security and integrity
4.New business services/products (Web services)
5. Collaboration/knowledge management
6. Enterprise architecture/SOA
7. e-commerce
8. External customer services
9. Supply chain automation/visualization
10. Data Centre optimization

Big Data Storage

Big Data NosQL (Not Only SQL)

- NoSQL databases are considered as semi-structured data.
- Big Data Store uses NoSQL.
- NoSQL stands for No SQL or Not Only SQL.
- The stores do not integrate with applications using SQL.
- NoSQL is also used in cloud data store.

Features of NoSQL are as follows:

- It is a class of non-relational data storage systems, and the flexible data models and multiple schema:

i) Class consisting of uninterrupted key/value or big hash table [Dynamo (Amazon S3)]

ii) Class consisting of unordered keys and using JSON (PNUTS)

iii) Class consisting of ordered keys and semi-structured data storage systems [Big Table, Cassandra (used in Facebook/Apache) and HBase]

iv) Class consisting of JSON (MongoDB)

v) Class consisting of name/value in the text (CouchDB)

vi) May not use fixed table schema

vii) Do not use the JOINS

viii) Data written at one node can replicate at multiple nodes, therefore Data storage is fault tolerant,

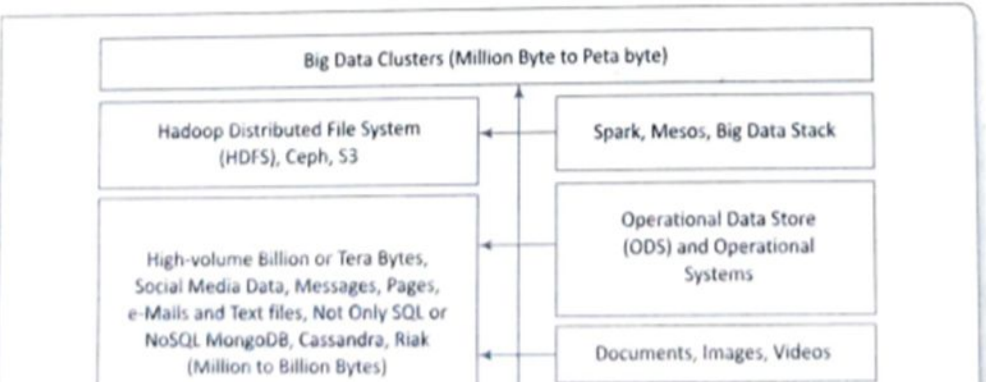
ix) May relax the ACID rules during the Data Store transactions.

x) Data Store can be partitioned and follows CAP theorem (out of three properties, consistency, availability and partitions, at least two must be there during the transactions)

Coexistence of Big Data, NoSQL and Traditional Data Stores

Table 1.4 Various data sources and examples of usages and tools

Data Source	Examples of Usages	Example of Tools
Relational databases	Managing business applications involving structured data	Microsoft Access, Oracle, IBM DB2, SQL Server, MySQL, PostgreSQL Composite, SQL on Hadoop [HPE (Hewlett Packard Enterprise) Vertica, IBM BigSQL, Microsoft Polybase,
Analysis databases (MPP, columnar, In-memory)	High performance queries and analytics	Sybase IQ, Kognitio, Terradata, Netezza, Vertica, ParAccel, ParStream, Infobright, Vectorwise
NoSQL databases (Key-value pairs, Columnar format, documents, Objects, graph)	Key-value pairs, fast read/write using collections of name-value pairs for storing any type of data; Columnar format, documents, objects, graph DBs and DSs	Key-value pair databases: Riak DS (Data Store), OrientDB, Column format databases (HBase, Cassandra), Document oriented databases: CouchDB, MongoDB; Graph databases (Neo4j, Tetan)
Hadoop clusters	Ability to process large data sets across a distributed computing environment	Cloudera, Apache HDFS
Web applications	Access to data generated from web applications	Google Analytics, Twitter
Cloud data	Elastic scalable outsourced databases, and data administration services	Amazon Web Services, Rackspace, GoogleSQL
Individual data	Individual productivity	MS Excel, CSV, TLV, JSON, MIME type
Multidimensional	Well-defined bounded exploration especially popular for financial applications	Microsoft SQL Server Analysis Services
Social media data	Text data, images, videos	Twitter, LinkedIn



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, developing.

2. 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 the

5. traditional as well as Big Data techniques.

Data management, storage and analytics of Big data captured at the companies and services requires

- 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 to 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

- Big Data platform consists of Big Data storage(s), server($) 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 an opensource 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.

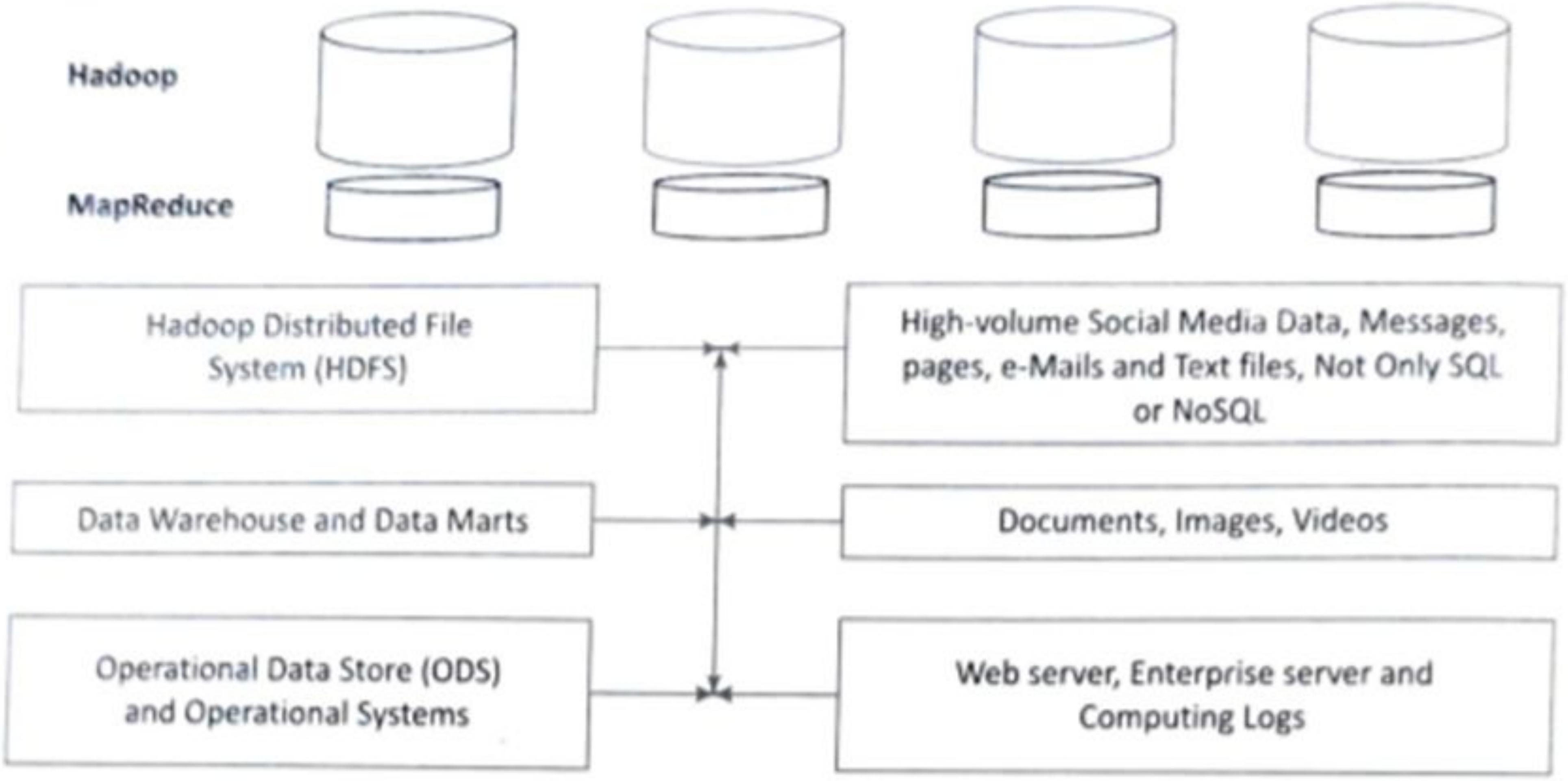


Figure 1.8 Hadoop based Big Data environment

Mesos

- Mesos vo.9 is a resources management platform which enables sharing of cluster of nodes by multiple frameworks

- Compatibility with an open analytics stack [data processing (Hive, Hadoop, HBase, Storm), data management (HDFS)

Big Data Stack

- A stack consists of a set of software components and data store units.

- Applications, machine-learning algorithms, analytics and visualization tools use Big Data Stack (BDS) at a cloud service, such as Amazon EC2, Azure or private cloud.

- Stack uses cluster of high-performance machines.

Types	Examples
MapReduce	Hadoop, Apache Hive, Apache Pig, Cascading, Cascalog, mrjob (Python MapReduce library), Apache S4, MapR, Apple Acunu, Apache Flume, Apache Kafka
NoSQL Databases	MongoDB, Apache CouchDB, Apache Cassandra, Aerospike, Apache HBase, Hypertable
Processing	Spark, IBM BigSheets, PySpark, R, Yahoo! Pipes, Amazon Mechanical Turk, Datameer, Apache Solr/Lucene, ElasticSearch
Servers	Amazon EC2, S3, GoogleQuery, Google App Engine, AWS Elastic Beanstalk, Salesforce Heroku
Storage	Hadoop Distributed File System, Amazon S3, Mesos

Big Data Analytics

- Data analysis need pre-processing of raw data and gives information useful for decision making.
- Analysis brings order, structure and meaning to the collection of data.
- Data is collected and analyzed to answer questions, test the hypotheses or disprove theories.

Definition

- It is statistical and mathematical data analysis that clusters, segments, ranks and predicts future possibilities.
- Feature of data analytics is its predictive, forecasting and prescriptive capability.
- Analytics uses historical data and forecasts new values or results.
- Analytics suggests techniques which will provide the most efficient and beneficial results for an enterprise. Data analysis helps in finding business intelligence and helps in decision making.

Phases in Analytics

- Analytics has the following phases before deriving the new facts, providing business intelligence and generating new knowledge.

1. Descriptive analytics enables deriving the additional value from visualizations and reports
2. Predictive analytics is advanced analytics which enables extraction of new facts and knowledge, and then predicts/forecasts
3. Prescriptive analytics enable derivation of the additional value and undertake better decisions for new option(s) to maximize the profits .
4. Cognitive analytics enables derivation of the additional value and undertake better decisions. Analytics integrates with the enterprise server or data warehouse.

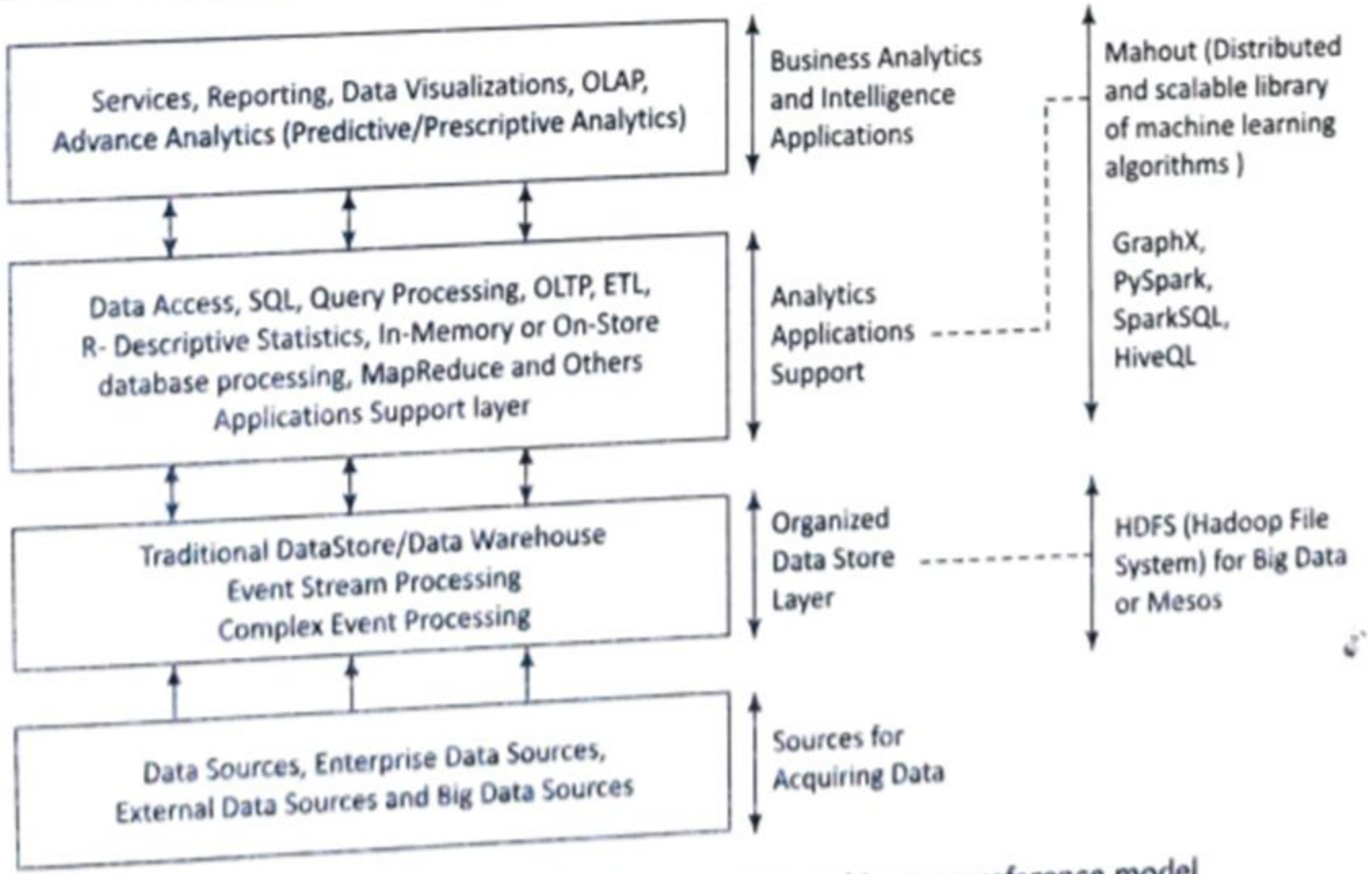


Figure 1.9 Traditional and Big Data analytics architecture reference model

Berkeley Data Analytics Stack (BDAS)

Importance of Big Data lies in the fact that what one does with it rather than how big or large it is.

Identify whether the gathered data is able to help in obtaining the following findings:

- cost reduction
- time reduction
- new product planning and development
- smart decision making using predictive analytics
- knowledge discovery.
- Big Data analytics need innovative as well as cost effective techniques.

Berkeley Data Analytics Stack (BDAS)

- BDAS is an open-source data analytics stack for complex computations on Big Data.
- It supports efficient, large-scale in-memory data processing
- enables user applications achieving three fundamental processing requirements; accuracy, time and cost.
- Berkeley Data Analytics Stack consists of :

 data processing

 data management

 resource management layers

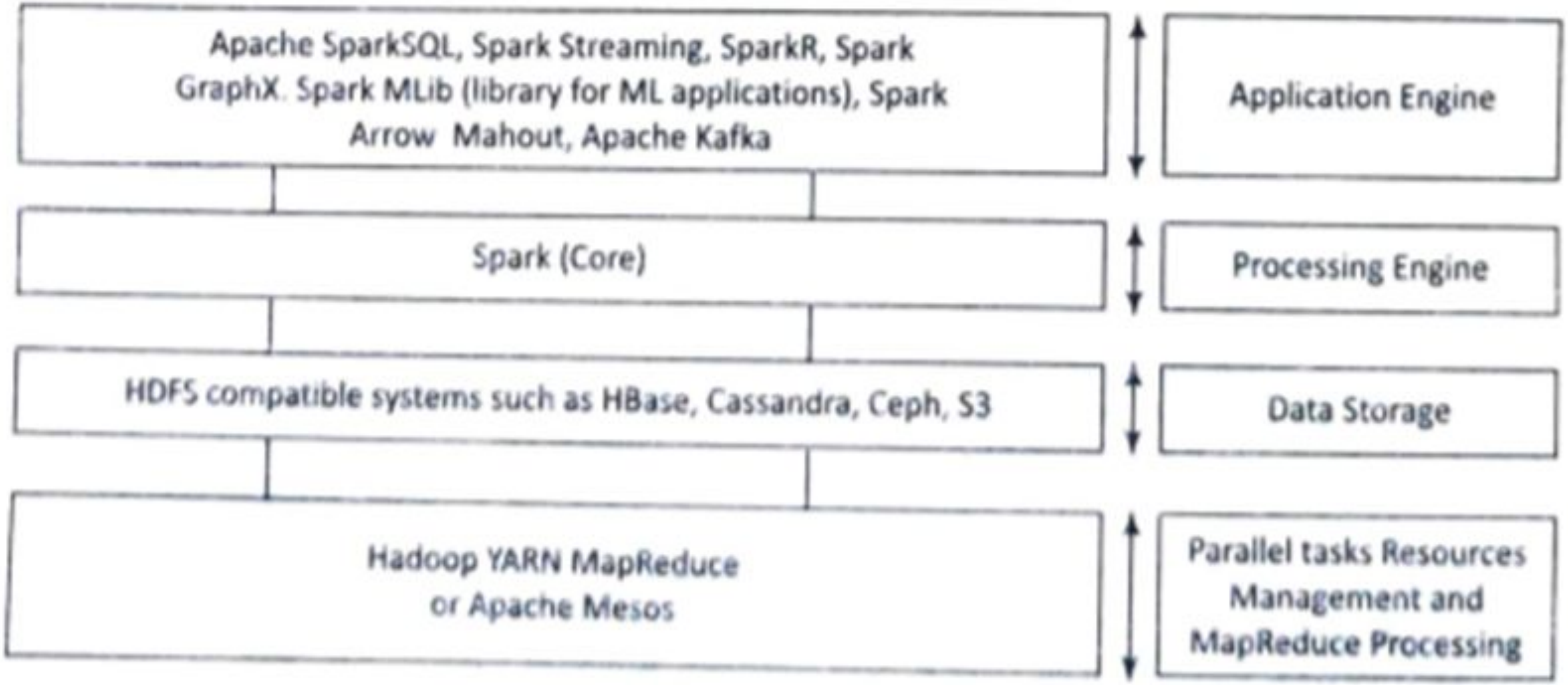


Figure 1.10 Four layers architecture for Big Data Stack consisting of Hadoop, MapReduce, Spark core and SparkSQL, Streaming, R, GraphX, MLib, Mahout, Arrow and Kafka

Big Data Analytics Applications and Case Studies

- Applications such as social network and social media, cloud applications, public and commercial web sites, scientific experiments, simulators and e-government services generate Big Data.
- Big Data analytics find applications in many areas.
- Popular ones are marketing, sales, health care, medicines, advertising etc.

Big Data in Marketing and Sales

- Data are important for most aspect of marketing, sales and advertising.
- Customer Value (CV) depends on three factors- quality, service and price.
- Big data analytics deploy large volume of data to identify and derive intelligence using predictive models about the individuals.
- The facts enable marketing companies to decide what products to sell.
- 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.

Five application areas in order of the popularity of Big Data use cases:

- CVA using the inputs of evaluated purchase patterns, preferences, quality, price and post sales servicing requirements
- Operational analytics for optimizing company operations
- Detection of frauds and compliances
- New products and innovations in service Enterprise data warehouse optimization.
- An example of fraud is borrowing money on already motgage assets.
- Example of timely compliances means returning the loan and interest instalments by the borrowers.

Ex: service-innovation

- A company develops software and then offers services like Uber.
- Company which develops software for hiring services, and then offers costly construction machinery and equipment.
- That service company might be rendering the services by hiring themselves from the multiple sources and locations of big construction companies.

Big data is providing marketing insights into

- Most effective content at each stage of a sales cycle
- investment in improving the customer relationship management (CRM)
- addition to strategies for increasing customer lifetime value (CLTV)
- lowering of customer acquisition cost (CAC).

Ex: Use of search engine optimization.

Why does the search engine at a company product website of a travel agency need optimization?

Solution:

- Consider a travel agency website offers search results for flights between two destinations A and C, which do not connect directly.
- The search shows the results in order of increasing travel cost through stopover at an intermediate airport B.
- Assume that search results show up just mechanically, without embedding intelligence and optimization.
- The customers find uncomfortable solutions with such searches.
- The searches show the cheaper options but sometimes show results such as the customer would reach C through stopover at B after 8 hours or even sometimes on the next day.

- The searches at that travel agency do not consider stopover options at different Bs, options available in different airlines to cut short travel time from B to C at cheaper costs, or newly introduced flights.
- The searches therefore need optimization for parameters of travel cost, multiple intermediate stopovers and airlines that will provide maximum customer convenience as well as cost.

Big Data Analytics in Detection of Marketing Frauds

- Fraud detection is vital to prevent financial loses to users.
- Fraud means someone deceiving deliberately.

Ex:

- Mortgaging the same assets to multiple financial institutions.
- Compromising customer data and 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.
- Big Data analytics enable fraud detection.

Big Data usages provides features-for enabling detection and prevention of frauds:

- Using 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
- Using multiple sources of data and connecting with many applications
- Providing greater insights using querying of the multiple source data
- Analysing data which enable structured reports and visualization
- Providing high volume data mining, new innovative applications and thus leading to new business intelligence and knowledge discovery
- 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. Big Data can cause potential harm to individuals.

Ex:

when someone puts false or distorted data about an individual in a blog, Facebook post, WhatsApp groups or tweets, the individual may suffer loss of educational opportunity, job or credit for his/her urgent needs. A company may suffer financial losses.

Five data risks - Bermard Marr are :

- data security
- data privacy breach
- costs affecting profits.
- bad analytics
- bad data

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:

1. Identifying high credit rating business groups and individuals
2. Identifying risk involved before lending money
3. Identifying industrial sectors with greater risks
4. Identifying types of employees (such as daily wage earners in construction sites) and businesses (such as oil exploration) with greater risks
5. Anticipating liquidity issues (availability of money for further issue of credit and rescheduling credit instalments) over the years.

- The insight using Big Data decreases the default rates in returning of loan, greater accuracy in issuing credit.
- One innovative way to manage credit risks and liquidity risks is use of available data and Big Data, High volume of data analysis gives greater insight into the default patterns, emerging patterns and thus credit risk
- Big Data analytics monitors social media, interactions data, contact addresses, mobile numbers, website financial status, activities or job changes to find the emerging credit risk that may affect a customer loan returning capacity.

- Digital footprints across social media provide a valuable alternative data source for credit risk analysis.
- The data companies assist in rating the customer in application processing and also during the period of repayment of a loan.
- Friends on Facebook and their credit rating, comments and assets posted also help in determining the risks.
- The data insights from the analytics lead to credit and liquidity risk management and faster reactions
- 3 benefits are :
  - Minimize the non-payments and frauds
  - Identifying new credit opportunities, new customers and revenue streams
  - Marketing to low risk businesses and households.

Big Data and Healthcare

Big Data analytics in health care use the following data sources:

- Clinical records,
- Pharmacy records,
- Electronic medical records
- Diagnosis logs and notes and
- Additional data, such as deviations from person usual activities, medical leaves from job, social interactions.

Healthcare analytics using Big Data can facilitate the following

- Provisioning of value-based and customer-centric healthcare,
- Utilizing the 'Internet of Things' for health care
- Preventing fraud, waste, abuse in the healthcare industry and reduce healthcare costs

  Ex: excessive or duplicate claims for clinical and hospital treatment, unnecessary tests, unnecessary use of medicines, such as tonics and testing facilities.
- Improving outcomes
- Monitoring patients in real time.

Value-based and customer-centric healthcare :

- Cost effective patient care by improving healthcare quality using latest knowledge, usages of electronic health and medical records
- Improving coordination among the healthcare providing agencies, which reduce avoidable overuse and healthcare costs.

Healthcare Internet of Things :

- Create unstructured data.
- The data enables the monitoring of the devices data for patient parameters, such as glucose, BP, ECGs and necessities of visiting physicians.
- Prevention of fraud, waste, and abuse uses Big Data predictive analytics and help resolve excessive or duplicate claims in a systematic manner.

- The analytics of patient records and billing help in detecting anomalies

Ex:

- Overutilization of services in short intervals
- Different hospitals in different locations simultaneously
- Identical prescriptions for the same patient filed from multiple locations.

Patient real-time monitoring

- uses machine learning algorithms which process real-time events.
- They provide physicians the insights to help them make life-saving decisions and allow for effective interventions.
- The process automation sends the alerts to care providers and informs them instantly about changes in the condition of a patient.

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 .

Some finding are:

- Building the health profiles of individual patients and predicting models for diagnosing better and offer better treatment

- 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,.

- 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

- Impact of Big Data is tremendous on the digital advertising industry.
- Digital advertising industry sends advertisements using SMS, e-mails, WhatsApp, LinkedIn, Facebook, Twitter and other mediums.
- Big Data technology and analytics provide insights, patterns and models, which relate the media exposure of all consumers to the purchase activity of all consumers using multiple digital channels.
- Identity management and can provide an advertising mix for building better branding exercises.
- Captures data of multiple sources in large volume, velocity and variety of data unstructured and enriches the structured data at the enterprise data warehouse.
- Real time analytics provide emerging trends and patterns, and gain actionable insights for facing competitions from similar products.

Big Data in Advertising

- The data helps digital advertisers to discover new relationships, lesser competitive regions and areas.
- Success from advertisements depend on collection, analysing and mining.
- The new insights enable the personalization and targeting the online, social media and mobile for advertisements called hyper-localized advertising.