Module I

Introduction to Big Data

1.1 Need of Big Data

The rise in technology has led to the production and storage of voluminous amounts of data. Earlier megabytes (10^6 B) were used but nowadays petabytes (10^{15} B) 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 1.1 shows data usage and growth. As size and complexity increase, the proportion of unstructured data types also increase.

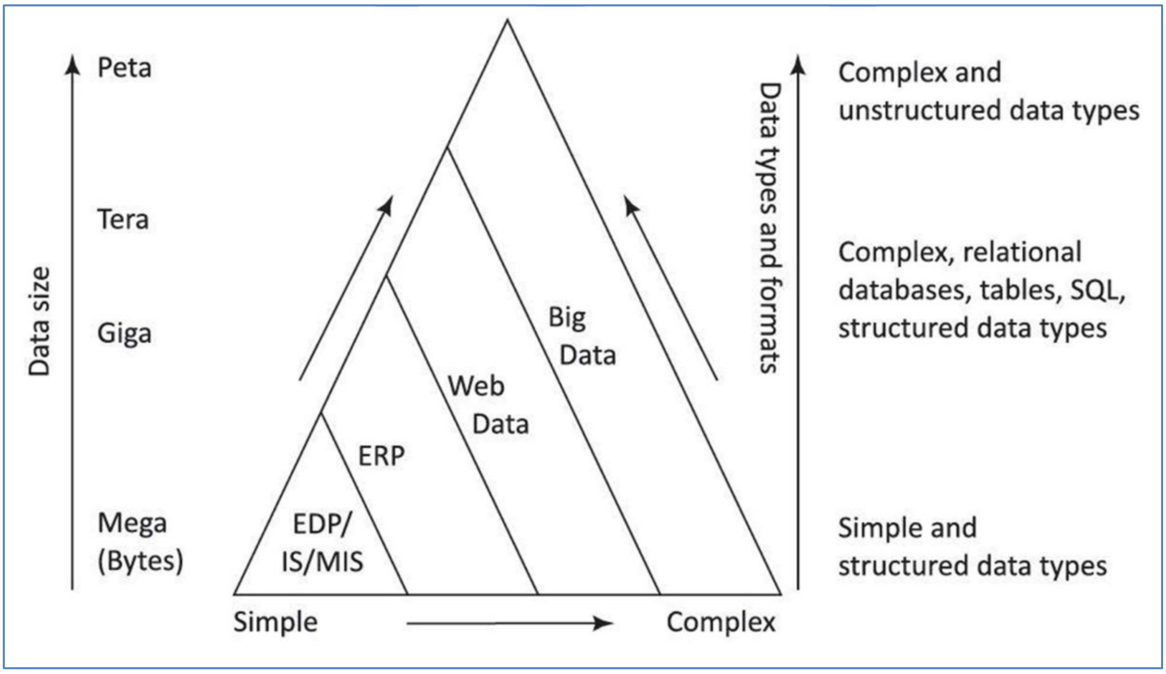


Figure 1.1 Evolution of Big Data and their characteristics

An 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.

1.2 BIG 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. 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 behavioral observations, measurements, or facts.

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**
- **Google Maps**
- **YouTube**
- **Facebook**

1.2.1 Classification of Data

Data can be classified as

a. **Structured**
b. **Semi-structured**
c. **Multi-structured**
d. **Unstructured**.

a. Structured Data

Structured data conform and associate with data schemas and data models. Structured data are found in tables (rows and columns). Nearly 15-20% data are in structured or semi-structured form. 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
- Transaction processing which follows ACID rules (Atomicity, Consistency, Isolation and Durability)
- *Encryption* and *decryption* for data security.

b. 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.

c. 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.

d. Unstructured Data

- Data does not possess data features such as a table or a database.
- Unstructured data are found in file types such as .TXT, .CSV.
- Data may be as key-value pairs, such as hash key-value pairs.
- Data may have internal structures, such as in e- mails.
- The data do not reveal relationships, hierarchy relationships.
- The relationships, schema and features need to be separately established.

Examples of unstructured Data

- **Mobile data**: Text messages, chat messages, tweets, blogs and comments
- **Website content data**: YouTube videos, browsing data, e-payments, web store data, user-generated maps
- **Social media data**: For exchanging data in various forms
- **Texts and documents**
- **Personal documents and e-mails**
- **Text internal to an organization**: Text within documents, logs, survey results
- Satellite images, atmospheric data, surveillance, traffic videos, images from Instagram, Flickr (upload, access, organize, edit and share photos from any device from anywhere in the world).

1.2.2 Big Data Definitions

- Big Data is high-volume, high-velocity and/or high-variety information 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.” -Wikipedia
- Data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges-oxford English dictionary.
- Big Data refers to data sets whose size is beyond the ability of typical database software tool to capture, store, manage and analyze

1.2.3 Big Data Characteristics

- **Volume**: is related to size of the data
- **Velocity**: refers to the speed of generation of data.
- **Variety**: comprises of a variety of data
- **Veracity**: quality of data captured, which can vary greatly, affecting its accurate analysis

1.2.4 Big Data Types

- Social networks and web data, such as Facebook, Twitter, e-mails, blogs and YouTube.
- 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.
- Customer master data such as 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 is also considered as machine generated data
- Human-generated data such as 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 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

Examples of Big Data

- Chocolate Marketing Company with large number of installed Automatic Chocolate Vending Machines (ACVMs).
- Automotive Components and Predictive Automotive Maintenance Services
- (ACPAMS) rendering customer services for maintenance and servicing of (Internet) connected cars and its components
- Weather data Recording, Monitoring and Prediction (WRMP) Organization.

Method	Type of Data
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 Usage	Human, business process, knowledge discovery, enterprise applications, Data

1.2.5 Big Data Classification

Big Data can be classified based on its characteristics that are used for designing data architecture for processing and analytics.

1.2.6 Big Data Handling Techniques

Following are the 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, parallel and distributed computing systems, such as Hadoop or Spark
- 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 and other form of databases used as per needs of the applications and in- memory data management using columnar or Parquet formats during program execution
- Data mining and analytics, data retrieval, data reporting, data visualization and machine- learning Big Data tools.

1.3 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 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.

- Scalability enables increase or decrease in the capacity of data storage, processing& analytics.

1.3.1 Analytical Scalability

Vertical scalability means scaling up the given system's resources and increasing the system's analytics, reporting and visualization capabilities. This is an additional way to solve problems of greater complexities. Scaling up means designing the algorithm according to the architecture that uses resources efficiently.

x terabyte of data takes time t 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 * 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. The easiest way to scale up and scale out 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.

1.3.2 Massive Parallel Processing 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.

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. Big Data processing uses a parallel, scalable and no-sharing program model, such as MapReduce, for computations on it.

Distributed Computing on multiple nodes	Big data	Large data	Small to Medium data
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

1.3.3 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."
- One of the best approach 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.

Features of Cloud Computing

- on-demand service
- resource pooling,
- scalability,
- accountability,
- broad network access.
- Cloud services can be accessed from anywhere and at any time through the Internet.

Cloud Services

There are three types of Cloud Services

- Infrastructure as a Service (IaaS):
- Platform as a Service (PaaS):
- Software as a Service (SaaS):

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 Communications, Amazon data centers and virtual servers.
- Apache 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 (IaaS).

Platform as a Service

- 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).

Software as a service

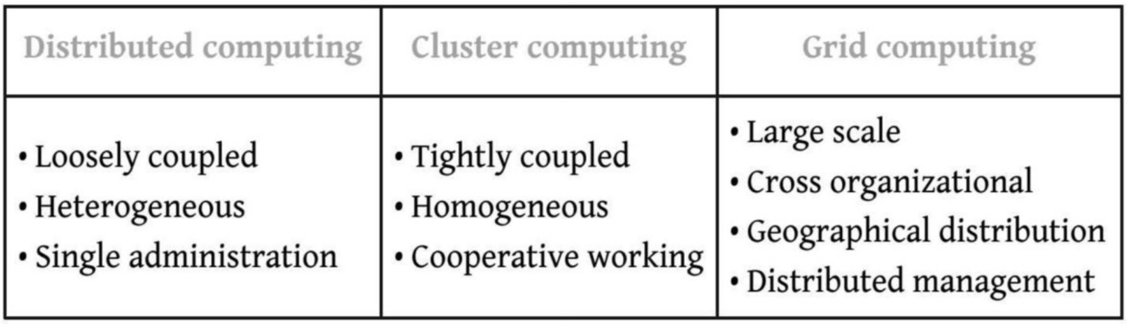
- Providing software applications as a service to end- users is known as Software as a Service.
- Software applications are hosted by a service provider and made available to customers over the Internet.
- Some examples are SQL Google SQL, IBM BigSQL, Microsoft Polybase and Oracle Big Data SQL.

1.3.4 Grid and Cluster 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.

- The computer resources are heterogeneously and geographically dispersing. A group of computers that might spread over remotely comprise a grid.
- A single grid of course, dedicates at an instance to a particular application only.
- Grid computing, similar to cloud computing, is scalable.
- Cloud computing depends on sharing of resources (for example, networks, servers, storage, applications and services) to attain coordination and coherence among resources similar to grid computing.
- Similarly, grid also forms a distributed network for resource integration.

Cluster Computing is 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.



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

1.3.5 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.

Some issues with volunteer computing systems are:

- Volunteered computers heterogeneity
- Dropouts from the network over time
- Their sporadic availability

1.4 Designing the Data Architecture

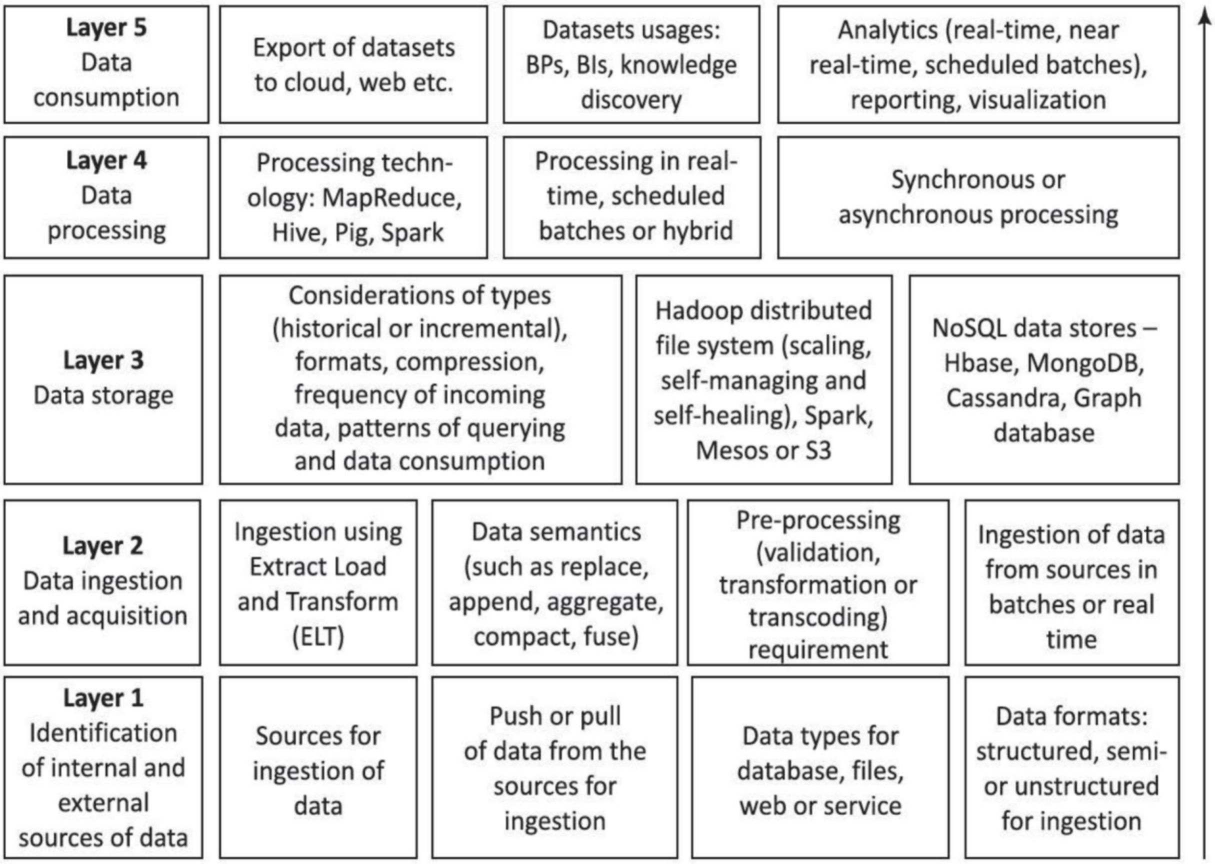
Big Data architecture is the logical and/or physical layout/structure of how Big Data will be stored, accessed and managed within a Big Data or IT environment. Architecture logically defines how Big Data solution will work, the core components (hardware, database, software, storage) used, flow of information,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

Data analytics need the number of sequential steps. Big Data architecture design task simplifies when using the logical layers approach. Figure 1.2 shows the logical layers and the functions which are considered in Big Data architecture

1.4.1 Data processing architecture consists of five layers:

- identification of data sources,
- acquisition, ingestion, extraction, pre-processing, transformation of data,
- Data storage at files, servers, cluster or cloud,
- data-processing,

- data consumption in the number of programs and tools.



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

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 or transfer. Ingestion may be in batches or in real time using pre- processing or semantics.

Layer 1

- L1 considers the following aspects in a design:
  - ➢ Amount of data needed at ingestion layer 2 (L2)
  - ➢ Push from L1 or pull by L2 as per the mechanism for the usages
  - ➢ Source datatypes: Database, files, web or service
- Source formats, i.e., semi-structured, unstructured or structured.

Layer 2

- Ingestion and ETL processes either in real time, which means store and use the data as generated, or in batches.

- Batch processing is using discrete datasets at scheduled or periodic intervals of time.

Layer 3

- Data storage type (historical or incremental), format, compression, incoming data
- frequency, querying patterns and consumption requirements for L4 or L5
- Data storage using Hadoop distributed file system or NoSQL data stores—HBase, Cassandra, MongoDB.

Layer 4

- Data processing software such as MapReduce, Hive, Pig, Spark, Spark Mahout, Spark Streaming
- Processing in scheduled batches or real time or hybrid
- Processing as per synchronous or asynchronous processing requirements at L5.

Layer 5

- Data integration
- Datasets usages for reporting and visualization
- Analytics (real time, near real time, scheduled batches), BPs, BIs, knowledge discovery
- Export of datasets to cloud, web or other systems

1.4.2 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 functions include:

1. Data assets creation, maintenance and protection
2. 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.
3. Data architecture creation, modelling and analysis
4. Database maintenance, administration and management system. For example, RDBMS (relational database management system), NoSQL
5. Managing data security, data access control, deletion, privacy and security

6. Managing the data quality
7. Data collection using the ETL process
8. Managing documents, records and contents
9. Creation of reference and master data, and data control and supervision
10. Data and application integration
11. Integrated data management, enterprise-ready data creation, fast access and analysis, automation and simplification of operations on the data,
12. Data warehouse management
13. Maintenance of business intelligence
14. Data mining and analytics algorithms.

1.5 Data Source

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.

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

1.5.1 Types of Data Source

- structured
- semi-structured
- multi-structured or unstructured

Structured Data Source

- 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
- Structured data sources are SQL Server, MySQL, Microsoft Access database, Oracle DBMS, IBM DB2, Informix, Amazon SimpleDB or a file-collection directory at a server.

Unstructured Data Source

- Unstructured data sources are distributed over high-speed networks.
- The data need high velocity processing. Sources are from distributed file systems.
- The 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 Sources - Sensors, Signals and GPS

The 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. They are sources in IoT applications.

1.5.2 Data Quality

High quality means data, which enables all the required operations, analysis, decisions, planning and knowledge discovery correctly. Five R's as follows:

- **Relevancy,**
- **Recency,**
- **Range,**
- **Robustness**
- **Reliability.**

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

Factors Affecting Data Quality

- Data Noise
- Outlier
- Missing Value
- Duplicate value

Data Noise

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. Noise is random in character, which means frequency with which it occurs is variable over time**.**

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. Actual outliers need to be removed from the dataset, else the result will be effected by a small or large amount.

Missing Value, duplicate Value

Missing Values is another factor effecting data quality. Missing value implies data not appearing in the data set. Another factor effecting data quality is duplicate values. Duplicate value implies the same data appearing two or more times in a dataset.

1.5.3 Data Preprocessing

Data preprocessing is an important step at the ingestion layer. Preprocessing is a must before data mining and analytics. Preprocessing is also a must before running a Machine Learning (ML) algorithm. Preprocessing needs are:

- Dropping out of range, inconsistent and outlier values
- Filtering unreliable, irrelevant and redundant information
- Data cleaning, editing, reduction and/or wrangling
- Data validation, transformation, or transcoding
- ELT processing (Extract, Load, Transform)

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 is done before mining of data. Incomplete or irrelevant data may result into misleading decisions.

- Data cleaning tools help in refining and structuring data into usable data. Examples of such tools are OpenRefine and DataCleaner.

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

- Data reduction enables the transformation of acquired information into an ordered, correct and simplified form.
- Data wrangling refers to the process of transforming and mapping the data. Results from analytics are then appropriate and valuable.
- Mapping enables data into another format, which makes it valuable for analytics and data visualizations

Data format using preprocessing

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

1.5.4 Data Export to Cloud

Figure 1.3 shows resulting data pre-processing, data mining, analysis, visualization and data store. The data exports to cloud services. The results integrate at the enterprise server or data warehouse.

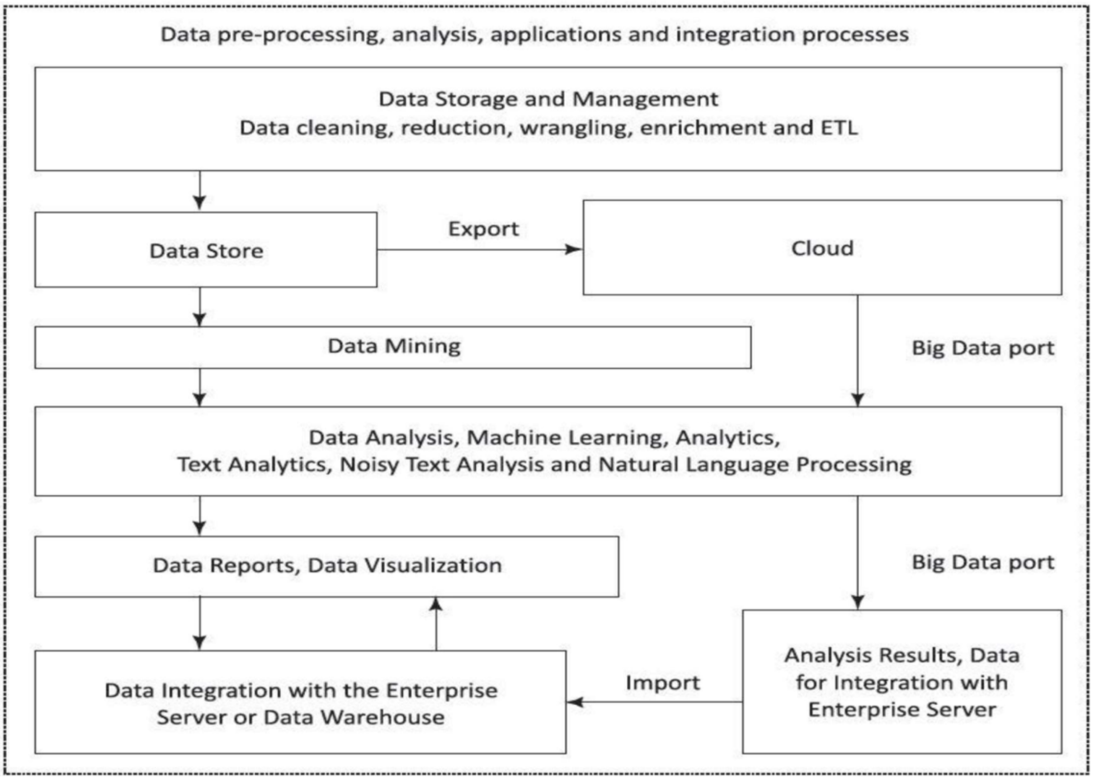

Data pre-processing, analysis, applications and integration processes
Data Storage and Management
Data cleaning, reduction, wrangling, enrichment and ETL
Data Store
Export
Cloud
Data Mining
Big Data port
Data Analysis, Machine Learning, Analytics,
Text Analytics, Noisy Text Analysis and Natural Language Processing
Data Reports, Data Visualization
Big Data port
Data Integration with the Enterprise Server or Data Warehouse
Import
Analysis Results, Data for Integration with Enterprise Server

Figure 1.3 Data preprocessing, analysis, visualization, data store export

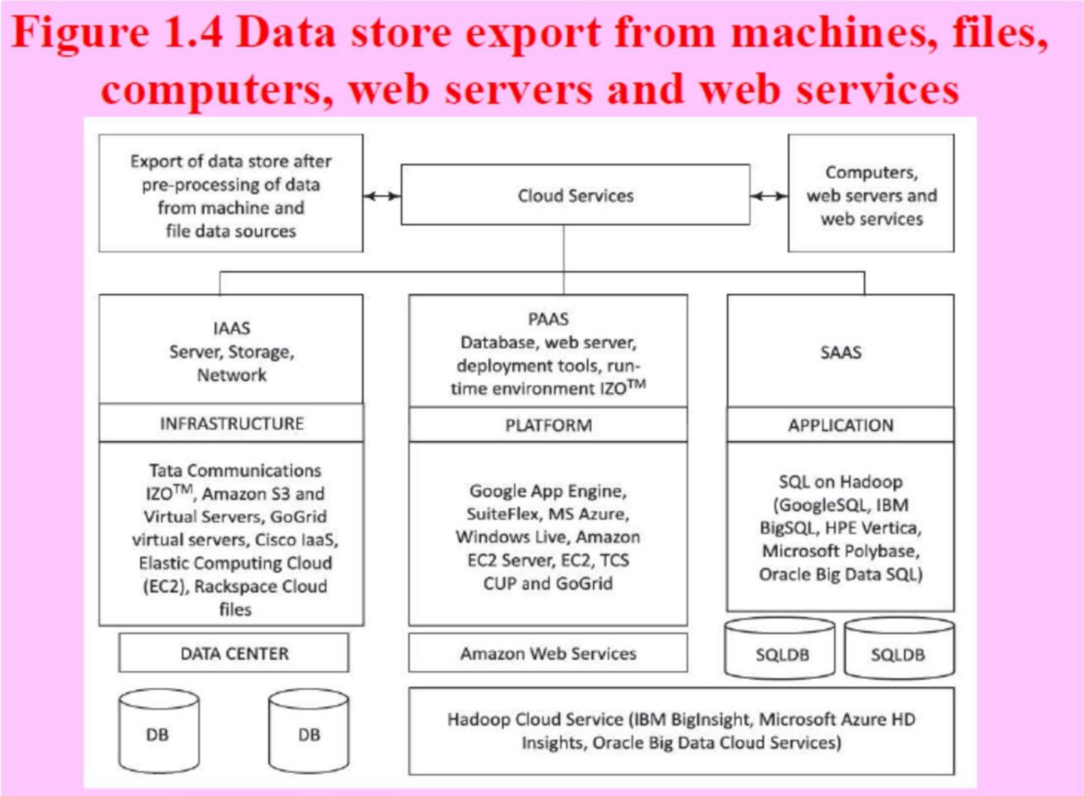


Figure 1.4 Data store export from machines, files, computers, web servers and web services
Export of data store after pre-processing of data from machine and file data sources
Cloud Services
Computers, web servers and web services
IAAS
Server, Storage, Network
INFRASTRUCTURE
Tata Communications IZO™, Amazon S3 and Virtual Servers, GoGrid virtual servers, Cisco IaaS, Elastic Computing Cloud (EC2), Rackspace Cloud files
DATA CENTER
DB
DB
PAAS
Database, web server, deployment tools, run-time environment IZO™
PLATFORM
Google App Engine, SuiteFlex, MS Azure, Windows Live, Amazon EC2 Server, EC2, TCS CUP and GoGrid
Amazon Web Services
SAAS
APPLICATION
SQL on Hadoop (GoogleSQL, IBM BigSQL, HPE Vertica, Microsoft Polybase, Oracle Big Data SQL)
SQLDB
SQLDB
Hadoop Cloud Service (IBM BigInsight, Microsoft Azure HD Insights, Oracle Big Data Cloud Services)

Figure 1.4 Data store export from machines, files, computers, web servers and web services

Cloud Services

Cloud offers various services. These services can be accessed through a cloud client (client application), such as a web browser, SQL or other client. Figure 1.4 shows data-store export from machines, files, computers, web servers and web services. The data exports to clouds, such as IBM, Microsoft, Oracle, Amazon, Rackspace, TCS, Tata Communications or Hadoop cloud services.

Export of Data to AWS and Rackspace Clouds

Google cloud platform provides a cloud service called BigQuery Figure 1.5 shows BigQuery cloud service at Google cloud platform. The data exports from a table or partition schema, JSON, CSV or AVRO files from data sources after the pre-processing.

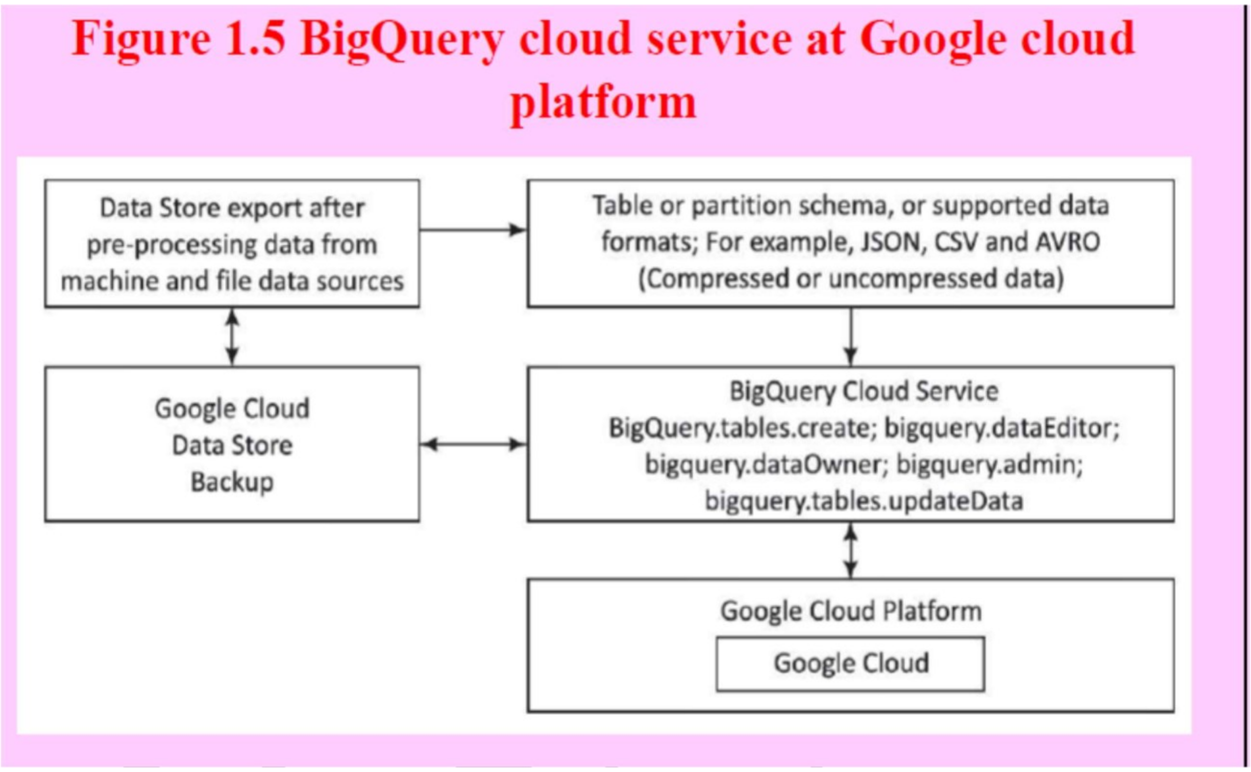


Figure 1.5 BigQuery cloud service at Google cloud platform
Data Store export after pre-processing data from machine and file data sources
Table or partition schema, or supported data formats; For example, JSON, CSV and AVRO (Compressed or uncompressed data)
Google Cloud Data Store Backup
BigQuery Cloud Service BigQuery.tables.create; bigquery.dataEditor; bigquery.dataOwner; bigquery.admin; bigquery.tables.updateData
Google Cloud Platform
Google Cloud

Figure 1.5: BigQuery cloud service at Google cloud platform

Data Store first pre-processes from machine and file data sources. Pre-processing transforms the data in table or partition schema or supported data formats. For example, JSON, CSV and AVRO. Data then exports in compressed or uncompressed data formats.

Cloud service BigQuery consists of bigquery.tables.create; bigquery.dataEditor; bigquery.dataOwner; bigquery.admin; bigquery.tables.updateData and other service functions. Analytics uses Google Analytics 360. BigQuery cloud exports data to a Google cloud or cloud backup only.

1.6 Data Storage and Analysis

This section describes data storage and analysis, and comparison between Big Data management and analysis with traditional database management systems.

1.6.1 Data Storage and Management: Traditional Systems

- Traditional systems use structured or semi-structured data
- The sources of structured data store are:
- Traditional relational database-management system (RDBMS) data, such as MySQL DB2, enterprise server and data warehouse

SQL

An RDBMS uses SQL (Structured Query Language). SQL is a language for viewing or changing (update, insert or append or delete) databases.

1. *Create schema, 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.

Distributed Database Management System

- A distributed DBMS (DDBMS) is a collection of logically interrelated databases at multiple system over a computer network.
- A collection of logically related databases.
- Cooperation between databases in a transparent manner.
- Be '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.
- *Online Analytical Processing* (OLAP) in real-time transaction processing is fast when using in-memory column format tables.
- *The CPU accesses all columns in a single instance of access to the memory in columnar format in memory data-storage.*

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 values store at the consecutive memory addresses in row format.

Enterprise Data-Store Server and Data 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 application(s), such as analytics, visualization, reporting, business intelligence and knowledge discovery

Following are some standardized business processes, as defined in the Oracle application-integration architecture:

- Integrating and enhancing the existing systems and processes
- Business intelligence
- Data security and integrity
- New business services/products (Web services)
- Collaboration/knowledge management
- Enterprise architecture/SOA
- e-commerce
- External customer services
- Supply chain automation/visualization
- Data centre optimization

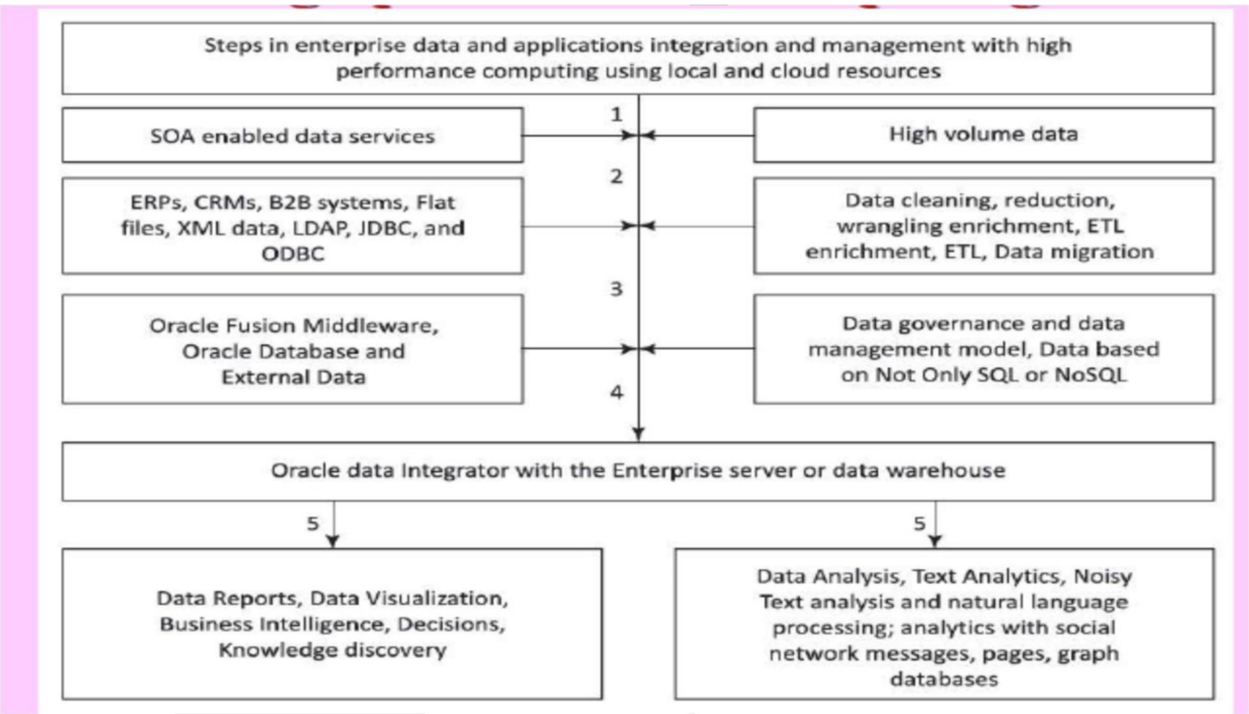

Steps in enterprise data and applications integration and management with high performance computing using local and cloud resources
SOA enabled data services
High volume data
ERPs, CRMs, B2B systems, Flat files, XML data, LDAP, JDBC, and ODBC
Data cleaning, reduction, wrangling enrichment, ETL enrichment, ETL, Data migration
Oracle Fusion Middleware, Oracle Database and External Data
Data governance and data management model, Data based on Not Only SQL or NoSQL
1
2
3
4
Oracle data Integrator with the Enterprise server or data warehouse
5
5
Data Reports, Data Visualization, Business Intelligence, Decisions, Knowledge discovery
Data Analysis, Text Analytics, Noisy Text analysis and natural language processing; analytics with social network messages, pages, graph databases

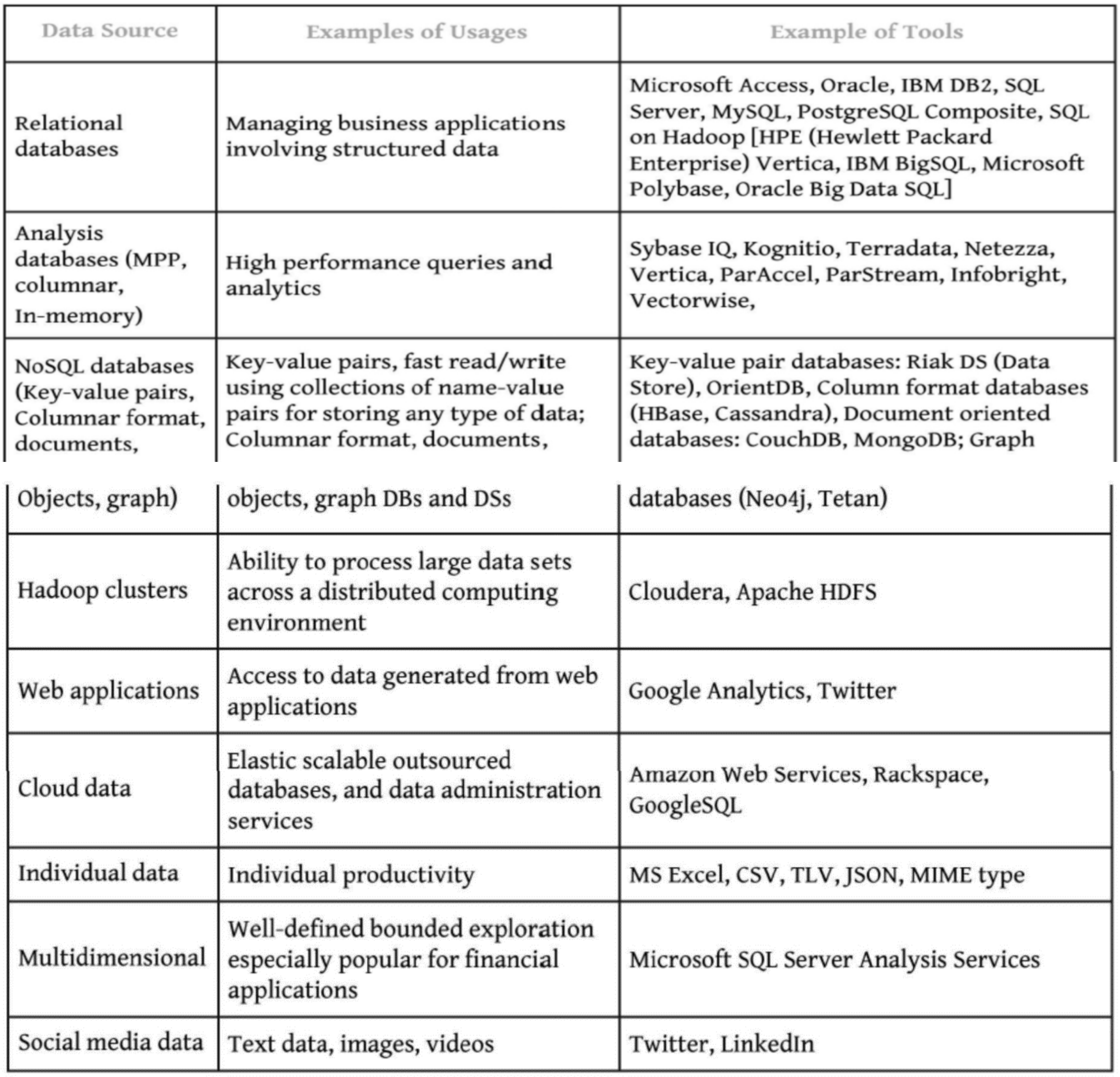
Figure 1.6 Steps 1 to 5 in Enterprise data integration and management with Big Data for high performance computing using local and cloud resources for the analytics, applications and services

1.6.2 Big Data Storage

NO 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 datastore.
- Features of NoSQL are as follows:
- It is a class of non-relational data storage systems, and the flexible data models and multiple schema:
- Class consisting of uninterrupted key/value or big hash table
- Class consisting of unordered keys and using JSON (PNUTS)
- Class consisting of ordered keys and semi-structured data storage systems [BigTable, Cassandra (used in Facebook/Apache) and HBase]
- Do not use the JOINS
- Data written at one node can replicate at multiple nodes, therefore Data storage is fault-tolerant,
- May relax the ACID rules during the Data Store transactions.

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, Oracle Big Data SQL]
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

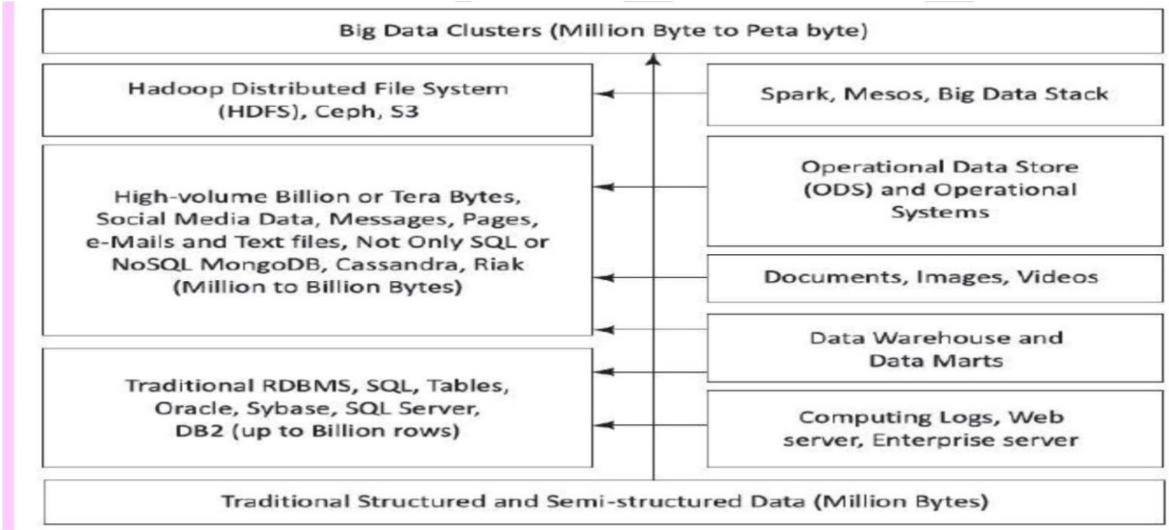

Big Data Clusters (Million Byte to Peta byte)
Hadoop Distributed File System (HDFS), Ceph, S3
Spark, Mesos, Big Data Stack
High-volume Billion or Tera Bytes, Social Media Data, Messages, Pages, e-Mails and Text files, Not Only SQL or NoSQL MongoDB, Cassandra, Riak (Million to Billion Bytes)
Operational Data Store (ODS) and Operational Systems
Documents, Images, Videos
Data Warehouse and Data Marts
Traditional RDBMS, SQL, Tables, Oracle, Sybase, SQL Server, DB2 (up to Billion rows)
Computing Logs, Web server, Enterprise server
Traditional Structured and Semi-structured Data (Million Bytes)

Figure 1.7: Coexistence of RDBMS for traditional server data, NoSQL and Hadoop, Spark and compatible Big Data Clusters

1.6.3 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:

1. New innovative non-traditional methods of storage, processing and analytics
2. Distributed Data Stores
3. Creating scalable as well as elastic virtualized platform (cloud computing)
4. Huge volume of Data Stores
5. Massive parallelism
6. High speed networks
7. High performance processing, optimization and tuning
8. Data management model based on Not Only SQL or NoSQL
9. In-memory data column-formats transactions processing or *dual in-memory data* columns as well as row formats for OLAP and OLTP
10. Data retrieval, mining, reporting, visualization and analytics
11. Graph databases to enable analytics with social network messages, pages and data analytics
12. Machine learning or other approaches
13. Big data sources: Data storages, data warehouse, Oracle Big Data, MongoDB NoSQL, Cassandra NoSQL
14. 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(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 an open-source storage system. HDFS is a scaling, self-managing and self-healing file system.

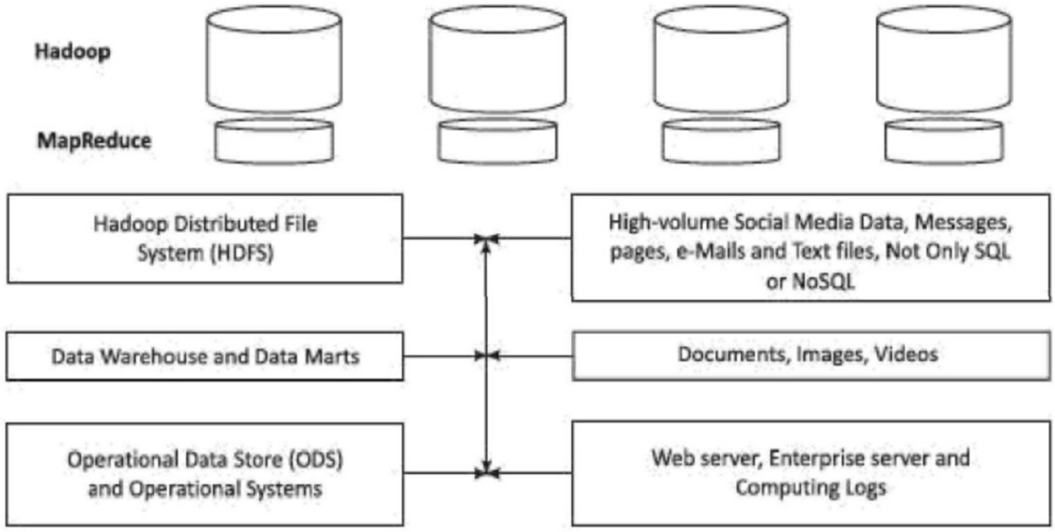

Hadoop
MapReduce
Hadoop Distributed File System (HDFS)
Data Warehouse and Data Marts
Operational Data Store (ODS) and Operational Systems
High-volume Social Media Data, Messages, pages, e-Mails and Text files, Not Only SQL or NoSQL
Documents, Images, Videos
Web server, Enterprise server and Computing Logs

Figure 1.8 Hadoop based Big Data environment

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 shows Hadoop based Big Data environment. Small height cylinders represent MapReduce and big ones represent the Hadoop.

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. The stack uses cluster of high-performance machines.

Table 1.5 Tools for Big Data environment

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 ECZ, S3, GoogleQuery, Google App Engine, AWS Elastic Beanstalk, Salesforce Heroku
Storage	Hadoop Distributed File System, Amazon S3, Mesos

1.6.2 Big Data Analytics

Data Analytics can be formally defined as the statistical and mathematical data analysis that clusters, segments, ranks and predicts future possibilities. An important 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

Analysis of data is a process of inspecting, cleaning, transforming and modeling data with the goal of discovering useful information, suggesting conclusions and supporting 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 decision.

Figure 1.9 shows an overview of a reference model for analytics architecture. The figure also shows on the right-hand side the Big Data file systems, machine learning algorithms and query languages and usage of the Hadoop ecosystem

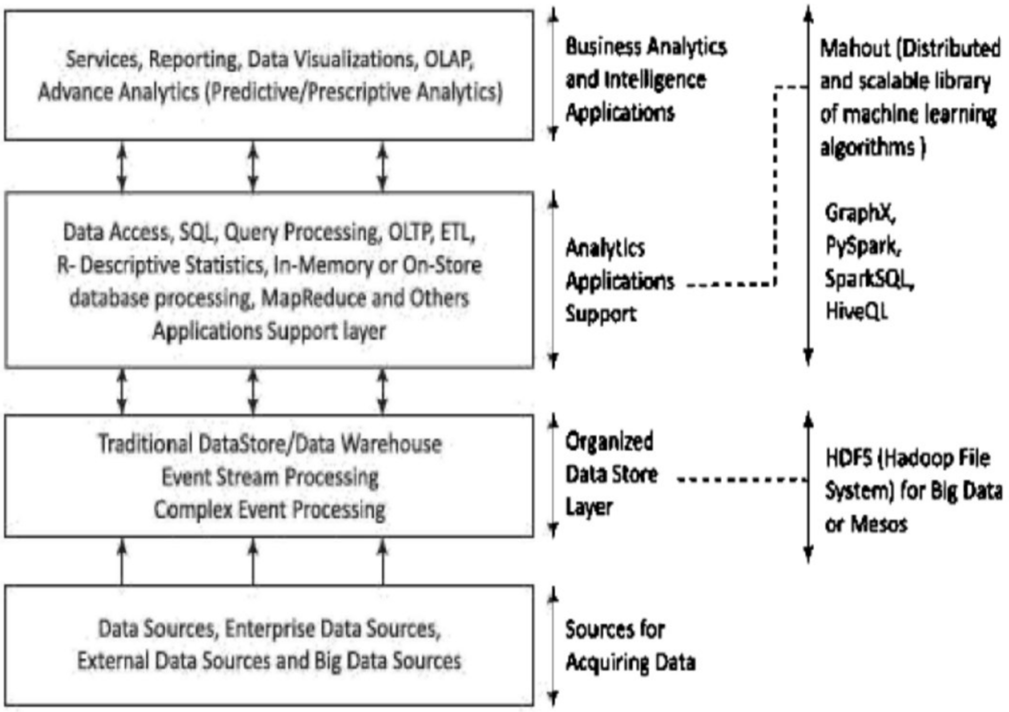

Services, Reporting, Data Visualizations, OLAP, Advance Analytics (Predictive/Prescriptive Analytics)
Business Analytics and Intelligence Applications
Mahout (Distributed and scalable library of machine learning algorithms)
Data Access, SQL, Query Processing, OLTP, ETL, R- Descriptive Statistics, In-Memory or On-Store database processing, MapReduce and Others Applications Support layer
Analytics Applications Support
GraphX, PySpark, SparkSQL, HiveQL
Traditional DataStore/Data Warehouse Event Stream Processing Complex Event Processing
Organized Data Store Layer
HDFS (Hadoop File System) for Big Data or Mesos
Data Sources, Enterprise Data Sources, External Data Sources and Big Data Sources
Sources for Acquiring Data

Figure 1.9 Traditional and Big Data analytics architecture reference model

Berkely Data Analysis Stack (BDAS)

Berkeley Data Analytics Stack (BDAS) consists of data processing, data management and resource management layers. Following list these:

1. Applications, AMP-Genomics and Carat run at the BDAS. Data processing software component provides in-memory processing which processes the data efficiently across the frameworks. AMP stands for Berkeley's Algorithms, Machines and Peoples Laboratory.
2. Data processing combines *batch, streaming* and *interactive* computations.
3. Resource management software component provides for sharing the infrastructure across various frameworks.

Figure 1.10 shows a four layers architecture for Big Data Stack that consists of Hadoop, MapReduce, Spark core and SparkSQL, Streaming, R, GraphX, MLib, Mahout, Arrow and Kafka

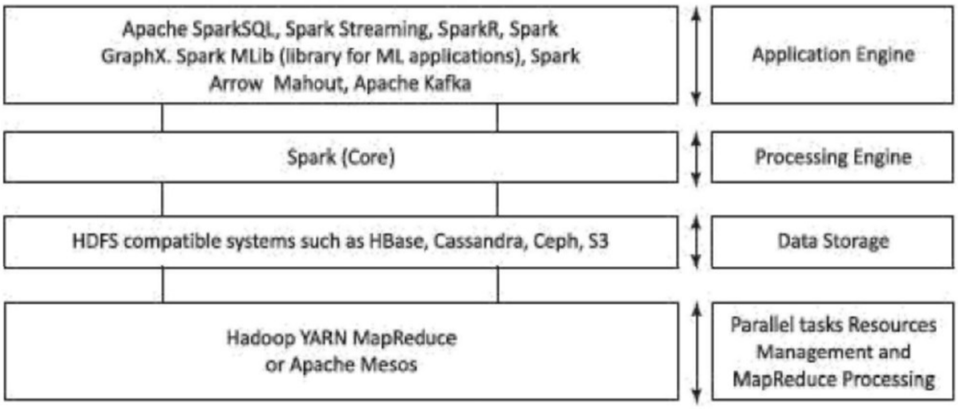

Apache SparkSQL, Spark Streaming, SparkR, Spark GraphX. Spark MLib (library for ML applications), Spark Arrow Mahout, Apache Kafka
Application Engine
Spark (Core)
Processing Engine
HDFS compatible systems such as HBase, Cassandra, Ceph, S3
Data Storage
Hadoop YARN MapReduce or Apache Mesos
Parallel tasks Resources Management and MapReduce Processing

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

1.7 Big Data Applications

1.7.1 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.

1.7.2 Big Data Analytics in Detection of Marketing Frauds

Big Data analytics enable fraud detection. Big Data usages has the following features-for enabling detection and prevention of frauds:

- Fusing of existing data at an enterprise data warehouse with data from sources such as social media, websites, biogs, 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
- Analyzing 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

1.7.3 **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.

Five data risks, described by Bernard Marr are data security, data privacy breach, costs affecting profits, bad analytics and bad data

1.7.4 Big Data Credit Card 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 installments) over the years.

1.7.5 **Big Data in Healthcare**

Big Data analytics in health care use the following data sources: (1) clinical records, (2) pharmacy records, (3) electronic medical records (4) diagnosis logs and notes and (5) 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

(Examples of frauds are excessive or duplicate claims for clinical and hospital treatments. Example of waste is unnecessary tests. Abuse means unnecessary use of medicines, such as tonics and testing facilities.)

4. Improving outcomes
5. Monitoring patients in real time.

1.7.6 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. Following are some findings: 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.

1.7.7 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 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. Advertising on digital medium needs optimization. Too much usage can also affect negatively. Phone calls, SMSs, e-mail-based advertisements can be nuisance if sent without appropriate researching on the potential targets. The analytics help in this direction. The usage of Big Data after appropriate filtering and elimination is crucial enabler of Big Data Analytics with appropriate data, data forms and data handling in the right manner.