MODULE 1

Introduction to Big Data Analytics

Need of Big Data

The rise in technology has led to the production and storage of voluminous amounts of data. Earlier megabytes were used but nowadays petabytes are used for processing, analysis, discovering new facts and generating new knowledge. Figure 1.1 shows data usage and growth. As size and complexity increase, the proportion of unstructured data types also increase

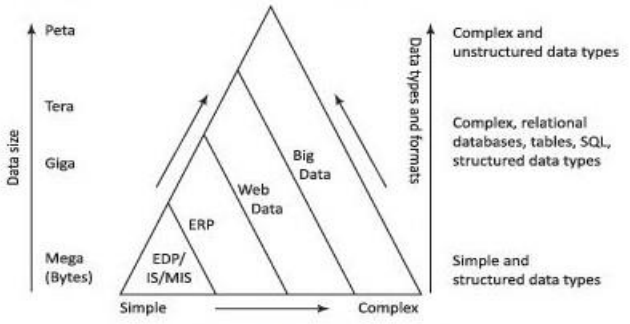

Figure 1.1 Evolution of Big Data and their characteristics

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. When integrated with the Internet of Things, sensors and machines data, the veracity of data is an additional V. Big Data requires new tools for processing and analysis of a large volume of data. For example, unstructured, NoSQL (not only SQL) data or Hadoop compatible system data.

1.1 Big Data

Definitions of Data : Data has several definitions. Usages can be singular or plural.

"Data is information, usually in the form of facts or statistics that one can analyze or use for further calculations." [Collins English Dictionary] "Data is information that can be stored and used by a computer program.". [Computing]

Definition of Web Data: Web is large scale 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.

Some examples of web data are :

1. Wikipedia is a web-based, free-content encyclopaedia project supported by the Wikimedia Foundation.
2. Google Maps is a provider of real-time navigation, traffic, public transport and nearby places by Google Inc.
3. McGraw-Hill Connect is a targeted digital teaching and learning environment that saves students' and instructors' time by improving student performance for a variety of critical outcomes.
4. Oxford Bookstore is an online book store where people can find any book that they wish to buy from millions of titles. They can order their books online at www.oxfordbookstore.com
5. YouTube allows billions of people to discover, watch and share originally-created videos by Google Inc.

Classification of Data— Structured, Semi-structured and Unstructured

Data can be classified as structured, semi-structured, multi-structured and unstructured.

Structured data conform and associate with data schemas and data models. Structured data are found in tables (rows and columns). Nearly 15– 20% data are in structured or semi-structured form.

Unstructured data do not conform and associate with any data models. Applications produce continuously increasing volumes of both unstructured and structured data.

Data sources generate data in three forms, viz. structured, semi-structured and unstructured.

Using Structured Data :

Structured data enables the following:

• data insert, delete, update and append

• Indexing to enable faster data retrieval

• Scalability which enables increasing or decreasing capacities and data processing operations such as, storing, processing and analytics

• Transactions processing which follows ACID rules (Atomicity, Consistency, Isolation and Durability)

• encryption and decryption for data security.

Using 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.

Using 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.

Using Unstructured Data:

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 or object-oriented features, such as extendibility.

Following are some 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 etc.

Big Data Definitions

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.

or

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

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 defines the amount or quantity of data, which is generated from an application(s). The size determines the processing considerations needed for handling that data.
- **Velocity:** The term velocity refers to the speed of generation of data. Velocity is a measure of how fast the data generates and processes. To meet the demands and the challenges of processing Big Data, the velocity of generation of data plays a crucial role.
- **Variety**: Big Data comprises of a variety of data. Data is generated from multiple sources in a system. This introduces variety in data and therefore introduces 'complexity'. Data consists of various forms and formats. The variety is due to the availability of a large number of heterogeneous platforms in the industry.
- **Veracity**: is also considered an important characteristic to take into account the 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.

Big Data Types

Following are the suggested types:

1. Social networks and web data, such as Facebook, Twitter, e-mails, blogs and YouTube.
2. Transactions data and Business Processes (BPs) data, such as credit card transactions, flight bookings, etc. and public agencies data such as medical records, insurance business data etc.
3. Customer master data, such as data for facial recognition and for the name, date of birth, marriage anniversary, gender, location and income category,
4. Machine-generated data, such as machine-to-machine or Internet of Things data, and the data from sensors, trackers, web logs and computer systems log.
5. Human-generated data such as biometrics data, human– machine interaction data, e-mail records with a mail server and MySQL database of student grades. Humans also records their experiences in ways such as writing these in notebooks or diaries, taking photographs or audio and video clips.

Example

Think of a manufacturing and retail marketing company, such as LEGO toys. How does such a toy company optimize the services offered, products and schedules, devise ways and use Big Data processing and storing for predictions using analytics?

Solution : Assume that a retail and marketing company of toys uses several Big Data sources, such as (i) machine-generated data from sensors (RFID readers) at the toy packaging, (ii) transactions data of the sales stored as web data for automated reordering by the retail stores and (iii) tweets, Facebook posts, e-mails, messages, and web data for messages and reports. The company uses Big Data for understanding the toys and themes in present days that are popularly demanded by children, predicting the future types and demands. The company using such predictive analytics, optimizes the product mix and manufacturing processes of toys. The company optimizes the services to retailers by maintaining toy supply schedules. The company sends messages to retailers and children using social media on the arrival of new and popular toys.

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

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.

1.2 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

Big Data processing and analytics requires scaling up and scaling out, both vertical and horizontal computing resources. Computing and storage systems when run in parallel, enable scaling out and increase system capacity.

Scalability enables increase or decrease in the capacity of data storage, processing and analytics. Scalability is the capability of a system to handle the workload as per the magnitude of the work. System capability needs increment with the increased workloads. When the workload and complexity exceed the system capacity, scale it up and scale it out. The following subsection describes the concept of analytics scalability.

Analytics Scalability to Big Data :

Vertical scalability means scaling up the given system's resources and increasing the system's analytics, reporting and visualization capabilities. Scaling up means designing the algorithm according to the architecture that uses resources efficiently. For example, x terabyte of data take 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. If r resources in a system process x terabyte of data in time t, then the (p × 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 Inter Processing nodes Communication (IPC).

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. However, buying faster CPUs, bigger and faster RAM modules and hard disks, faster and bigger motherboards will be expensive compared to the extra performance achieved by efficient design of algorithms. Also, if more CPUs add in a computer, but the software does not exploit the advantage of them, then that will not get any increased performance out of the additional CPUs.

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. Here, it is required to enhance (scale) up the computer system or use massive parallel processing (MPPs) platforms. Parallelization of tasks can be done at several levels: (i) distributing separate tasks onto separate threads on the same CPU, (ii) distributing separate tasks onto separate CPUs on the same computer and (iii) distributing separate tasks onto separate computers.

When making software, draw the advantage of multiple computers (or even multiple CPUs within the same computer) and 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**

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. Table 1.2 gives the requirements of processing and analyzing big, large and small to medium datasets on distributed computing nodes. Big Data processing uses a parallel, scalable and no-sharing program model, such as MapReduce, for computations on it. (Chapter 2)

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**

Wikipedia defines cloud computing as, "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.
It offers high data security compared to other distributed technologies. Cloud resources can be Amazon Web Service (AWS) Elastic Compute Cloud (EC2), Microsoft Azure or Apache CloudStack. Amazon Simple Storage Service (S3) provides simple web services interface to store and retrieve any amount of data, at any time, from anywhere on the web. [Amazon EC2 name possibly drives from the feature that EC2 has a simple web service interface, which provides and configures the storage and computing capacity with minimal friction].
Cloud computing features are: (i) on-demand service (ii) resource pooling, (iii) scalability, (iv) accountability, and (v) broad network access. Cloud services can be accessed from anywhere and at any time through the Internet. A local private cloud can also be set up on a local cluster of computers. Cloud computing allows availability of computer infrastructure and services "on-demand" basis. The computing infrastructure includes data storage device, development platform, database, computing power or software applications.
Cloud services can be classified into three fundamental types:
1. **Infrastructure as a Service (IaaS):** Providing access to resources, such as hard disks, network connections, databases storage, data center and virtual server spaces is Infrastructure as a Service (IaaS). Some examples are Tata 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).
2. **Platform as a Service (PaaS):** It implies providing the runtime environment to allow developers to build applications and services, which means cloud Platform as a Service. Software at the clouds support and manage the services, storage, networking, deploying, testing, collaborating, hosting and maintaining applications. Examples are Hadoop Cloud Service (IBM BigInsight, Microsoft Azure HD Insights, Oracle Big Data Cloud Services).
3. **Software as a Service (SaaS):** Providing software applications as a service to end-users is known as Software as a Service. Software applications are hosted by a service provider and made available to customers over the Internet. Some examples are SQL GoogleSQL, IBM BigSQL, HPE Vertica, Microsoft Polybase and Oracle Big Data SQL.

- **Grid and Cluster Computing**

Grid Computing 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 disperse.

A grid is used for a variety of purposes. A single grid of course, dedicates at an instance to a particular application only. Grid computing provides large-scale resource sharing which is flexible, coordinated and secure among its users. The users consist of individuals, organizations and resources

Features of Grid Computing 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.

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

- Cluster Computing

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

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

1.3 Designing Data Architecture

Data Architecture Design

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

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

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

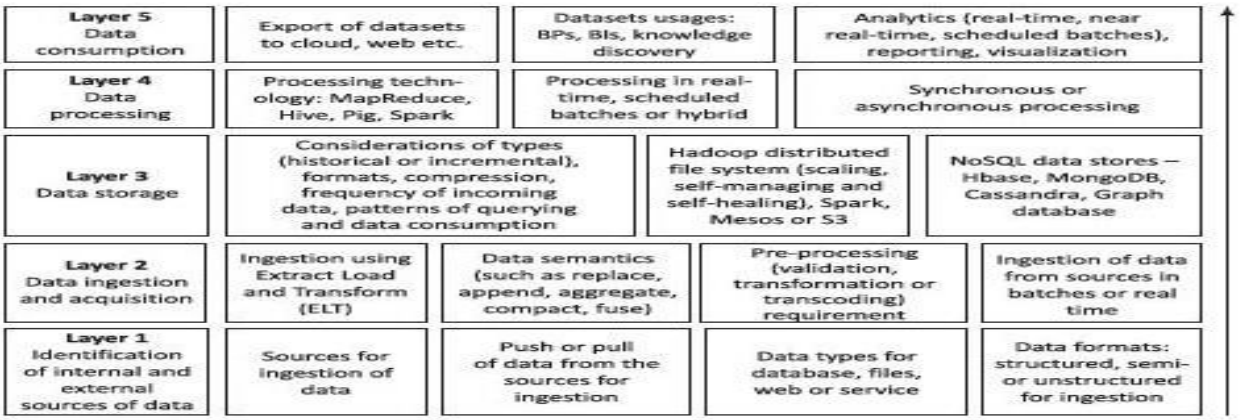

Figure 1.2 Design of logical layers in a data processing architecture, and functions in the layers

Data ingestion, pre-processing, storage and analytics require special tools and technologies.
Logical layer 1 (L1) is for identifying data sources, which are external, internal or both.
The layer 2 (L2) is for data-ingestion. Data ingestion means a process of absorbing information. 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.
The L3 layer is for storage of data from the L2 layer.
The L4 is for data processing using software, such as MapReduce, Hive, Pig or Spark.
The top **layer L5** is for data consumption. Data is used in analytics, visualizations, reporting, export to cloud or web servers.

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 data-types: Database, files, web or service
• Source formats, i.e., semi-structured, unstructured or structured.

L2 considers the following aspects:
• 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.

L3 considers the followings aspects:
• 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.

L4 considers the followings aspects:
• 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.

L5 considers the consumption of data for the following:
• 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.

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.
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.

1.4 Data Sources, Quality, Pre-Processing, And Storing

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 (L4 in Figure 1.2).

- **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. Examples of 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.

A data source name implies a defined name, which a process uses to identify the source. The name needs to be a meaningful name. For example, 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 UniversityStudents_DataPlusGrades. A master-directory server can also be called NameNode.

Microsoft applications consider two types of sources for processing: (i) machine sources and (ii) file sources. (i) Machine sources are present on computing nodes, such as servers. A machine identifies a source by the user-defined name, driver-manager name and source-driver name. (ii) File sources are stored files. An application executing the data, first connects to a driver manager of the source. A user, client or application does not register with the source, but connects to the manager when required.

Oracle applications consider two types of data sources: (i) database, which identifies the database information that the software needs to connect to database, and (ii) logic-machine, which identifies the machine which runs batches of applications and master business functions. Source definition identifies the machine. The source can be on a network. The definition in that case also includes network information, such as the name of the server, which hosts the machine functions.

The applications consider data sources as the ones where the database tables reside and where the software runs logic objects for an enterprise. Data sources can point to:

1. A database in a specific location or in a data library of OS

2. A specific machine in the enterprise that processes logic
3. 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.

Example

(i) How would you name the data sources of the student grade-sheets? (ii) How does an analytics application (Analysis_APP) access student grade-sheet data source, using the Data Dictionary or data-source master-table for the grade-sheets of students? (iii) How does the application connect and access the data source of students' grade-sheets? Assume each student can have a grade-sheet for each of the six semesters in UG Computer Science programme.

Solution

(i) Assume SemID is distinct key for a semester. StudID is a key assigned to a student, whether in CS or another subject, and whether in UG or PG. A StudID is unique. Data source can be file data source named 'UG_CS_Sem_StudID_Grades' for all UG CS student grades. UG_CS_Sem_StudID_Grades database consists of maximum six grade sheets UG_CS_SemID_StudID_Grades, i.e., one for each semester. Assume that Analysis_APP does not connect or directly links to the data source UG_CS_Sem_StudID_Grades database. Then, the Analysis_APP links to a Data Dictionary or data source master table, which is data repository for the pointers of all six semesters of UG Computer Science program and other subject programs.

(ii) (ii) Assume that Analysis_APP associates to Oracle data-source master-table. The table stores the data-source definitions for all UG and PG, and all subjects and semester grades of the students. The data-source master-table stores the pointers of all semester grades. The table thus points to UG_CS_Sem_StudID_Grades DB for the student identified by StudID.

(iii) Assume that application deploys Microsoft DB. Then, first Analysis_APP links to a Driver Manager. The Driver Manager then calls the ODBC functions in the Driver Manager. The application identifies the target driver for the UG_CS_Sem_StudID_Grades data source with a connection handle. When the Driver Manager loads the driver, the Driver Manager builds a table of pointers to the functions in that driver. It uses the connection handle passed by the application to find the address of the function in the target driver and calls that function by address.

- **Unstructured Data Sources**

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 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**

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 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.

Data Quality

Data 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, especially for artificial intelligence applications, can be data with five R's as follows: Relevancy, recency, range, robustness and reliability. Relevancy is of utmost importance. A uniform definition of data quality 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** 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. For example, in Example 1.7 the grades of students should remain unaffected upon processing.
- **Data Noise, Outliers, Missing and Duplicate Values**

Noise: One of the factors effecting data quality is noise. Noise in data refers to data giving additional meaningless information besides true (actual/ required) information. Noise refers to difference in the value measured from true value due to additional influences. The values show nearly equal positive and negative deviations. A statistical analysis of deviation can select the noise in data and true values can be retrieved.

Outliers: A factor which effects quality is an outlier. An outlier in data refers to data, which appears to not belong to the dataset. For example, data that is outside an expected range. The outliers are a result of human data-entry errors, programming bugs, some transition effect or phase lag in stabilizing the data value to the true value.

Missing Values: Another factor effecting data quality is missing values. Missing value implies data not appearing in the data set.

Duplicate Values : Another factor effecting data quality is duplicate values. Duplicate value implies the same data appearing two or more times in a dataset. The following example explains noise, outliers, missing values and duplicate data.

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 (iii) Data cleaning, editing, reduction and/ or wrangling (iv) Data validation, transformation or transcoding (v) ELT processing.

- **Data Cleaning -** Data cleaning refers to the process of removing or correcting incomplete, incorrect, inaccurate or irrelevant parts of the data after detecting them.
- **Data Cleaning Tools -** Data cleaning is done before mining of data. Incomplete or irrelevant data may result into misleading decisions. It is not always possible to create well-structured data. Data 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. Examples of such tools are OpenRefine and DataCleaner.
- **Data Enrichment** - Techopedia definition is as follows: "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, (iii) 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 basic concept is the reduction of multitudinous amount of data, and use of the meaningful parts. The reduction uses editing, scaling, coding, sorting, collating, smoothening, interpolating and preparing tabular summaries.
- **Data Wrangling -** Data wrangling refers to the process of transforming and mapping the data. Results from analytics are then appropriate and valuable. For example, mapping enables data into another format, which makes it valuable for analytics and data visualizations.
- **Data Format used during Pre-Processing**
  Examples of formats for data transfer from (a) data storage, (b) analytics application, (b) service or (d) cloud can be:
  (i) Comma-separated values CSV
  (ii) Java Script Object Notation (JSON) as batches of object arrays or resource arrays
  (iii) Tag Length Value (TLV)
  (iv) Key-value pairs
  (v) Hash-key-value pairs.

CSV Format - An example is a table or Microsoft Excel file which needs conversion to CSV format. A student_record.xlsx converts to student_record.csv file. Comma-separated values (CSV) file refers to a plain text file which stores the table data of numbers and text. When processing for data visualization of Excel format file, the data conversion will be done from csv to xlsx format. Each CSV file line is a data record. Each record consists of one or more fields, separated from each other by commas. RFC 4180 standard specifies the various specifications. A CSV file may also use space, tab or delimiter tab-separated formats for the values in the fields. This is a loose terminology. The following example explains the conversion process.

Example

Consider the example of a table in a grade sheet.

Subject Code	Subject Name	Grade
CS101	"Theory of Computations"	7.8
CS102	"Computer Architecture"	7.2
-	-	-

Solution The first and second lines in the CSV file are:

Subject Code, Subject Name, Grade

CS101, "" Theory of Computations"", 7.8.

CS102, "" Computer Architecture"", 7.8.

The two consecutive double-quotes mean that one of the double quotes is retained in the text "Theory of Computations". That one specifies that characters are inside the double quotes and represent a string.

Data Format Conversions

Transferring the data may need pre-processing for data-format conversions. Data sources store need portability and usability. A number of different applications, services and tools need a specific format of data only. Pre-processing before their usages or storage on cloud services is a must.

Data Store 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.

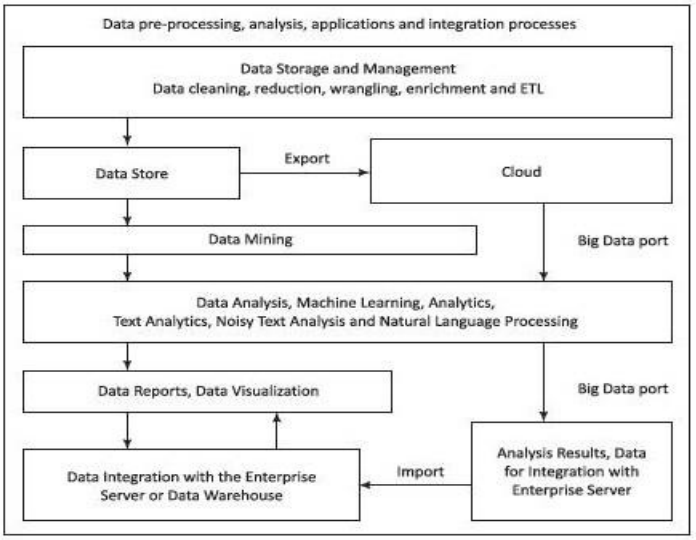

Figure 1.3 Data pre-processing, analysis, visualization, data store export

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.

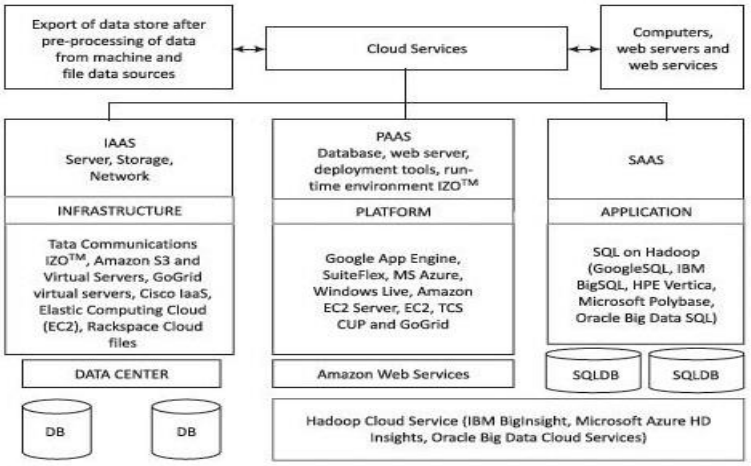

Figure 1.4 Data store export from machines, files, computers, web servers and web services

Example: Export of Data to AWS and Rackspace Clouds

The following example explains the export processes to Amazon and Rackspace clouds.

(a) How do the rows in MySQL database table export to Amazon AWS? (b) How do the rows in MySQL database table export to Rackspace?

Solution (a) Following are the steps for export to an EC2 instance: (i) A process pre-processes the data from data-rows at table in MySQL database and creates a CSV file. (ii) An EC2 instance provides an AWS data pipeline. (iii) The CSV file exports to Amazon S3 using pipeline. The CSV file then copies into an S3 bucket. Coping action deploys an EC2 instance. (iv) AWS notification service (SNS) sends notification on completion.

(b) Following are the steps for export to Rackspace:
(i) An instance name has maximum 255 characters. One or more databases create a database instance. The process of creation can be configured to create an instance now or later. Each database can have a number of users.
(i) Default port number for binding of MySQL is port 3306.
(ii) A command mysqldump – u root – p database_name > database_name.sql exports to Rackspace cloud.
(iv)When a database is at a remote host then a command mysqldump – h host_name -u user_name – p database_name > database_name.sql exports to the cloud database.

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.

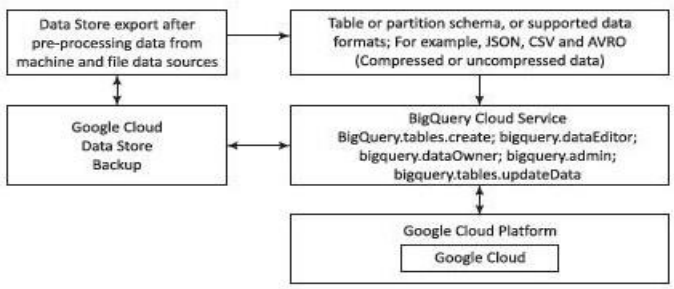

Figure 1.5 BigQuery cloud service at Google cloud platform

Data Store 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. (Avro is a data serialization system in Hadoop related tools for Big Data.) 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.

1.5 Data Storage and Analysis

The following subsections describe data storage and analysis, and comparison between Big Data management and analysis with traditional database management systems.

Data Storage and Management: Traditional Systems

- **Data Store with Structured or Semi-Structured Data** Traditional systems use structured or semi-structured data. The following example explains the sources and data store of structured data.

Example 1.1
What are the sources of structured data store? Solution The sources of structured data store are: • Traditional relational database-management system (RDBMS) data, such as MySQL DB2, enterprise server and data warehouse • Business process data which stores business events, such as registering a customer, taking an order, generating an invoice, and managing products in pre-defined formats. The data falls in the category of highly structured data. The data consists of transaction records, tables,

relationships and metadata that build the information about the business data. • Commercial transactions • Banking/ stock records • E-commerce transactions data.

Example 1.2

Give examples of sources of data store of semi-structured data. Solution Examples of semi-structured data are: • XML and JSON semi-structured documents7,8 • A comma-separated values (CSV) file. The CSV stores tabular data in plain text. Each line is a data record. A record can have several fields, each filed separated by a comma. Structured data, such as database include multiple relations but CSV does not consider the relations in a single CSV file. CSV cannot represent object-oriented databases or hierarchical data records. A CSV file is as follows: Preeti, 1995, MCA, Object Oriented Programming, 8.75 Kirti, 2010, M.Tech., Mobile Operating System, 8.5 Data represent the data records for columns and rows of a table. Each row has names, year of passing, degree name, course name and grade point out of 10. Rows are separated by a new line and the columns by a comma. JSON Object Data Formats: CSV does not represent object-oriented records, databases or hierarchical data records. JSON and XML represent semi-structured data and represent object-oriented and hierarchical data records. Example 3.5 explains CSV and JSON objects and the hierarchical data records in the JSON file format.

- **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 does the following:

  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.

  Relational database examples are MySQL PostGreSQL Oracle database, Informix, IBM DB2 and Microsoft SQL server.

- **Large Data Storage using RDBMS**

  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).

- **Distributed Database Management System**

  A distributed DBMS (DDBMS) is a collection of logically interrelated databases at multiple system over a computer network. The features of a distributed database system are:

  1. A collection of logically related databases.
  2. 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.

3. Should 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. Data in a column are kept together in-memory in columnar format. A single memory access, therefore, loads many values at the column. An 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

  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 obtaining online summarized information and automated reports for a large database. Metadata describes the data. Pre-storing of calculated values provide consistently fast response. Result formats from the queries are based on Metadata.

- **In-Memory Row Format Databases**

  A row format in-memory allows much faster data processing during OLTP (online transaction processing). For example, the 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 (Figure 1.3). Enterprise data integration may also include integration with application(s), such as analytics, visualization, reporting, business intelligence and knowledge discovery.

  Enterprise data warehouse store the databases, and data stores after integration, using tools from number of sources.

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

  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

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

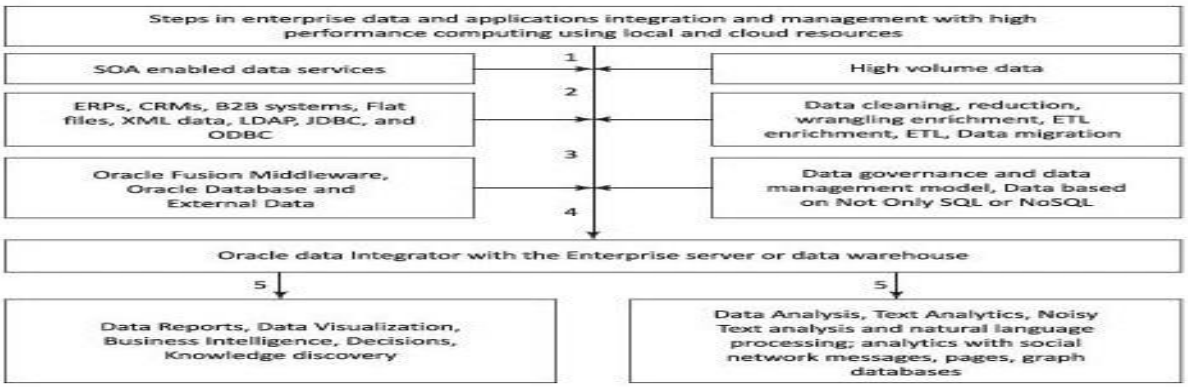

Figure 1.6 Steps 1 to 5 in Enterprise data integration and management with Big-Data for high performance computing using local and cloud resources for the analytics, applications and services

- **Big Data Storage**

Following subsections describe Big Data storage concepts:

Big Data NoSQL or Not Only SQL

NoSQL databases are considered as semi-structured data. Big Data Store uses NoSQL. The stores do not integrate with applications using SQL. NoSQL is also used in cloud data store.

Features of NoSQL are as follows:

1. 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 [BigTable, 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) Consistency means all copies have the same value like in traditional DBs. Availability means at least one copy is available in case a partition becomes inactive or fails. For example in web applications, the other copy in other partition is available. Partition means parts which are active but may not cooperate as in the distributed DBs.

Coexistence of Big Data, NoSQL and Traditional Data Stores

Figure 1.7 shows co-existence of data at server, SQL, RDBMS with NoSQL and Big Data at Hadoop, Spark, Mesos, S3 or compatible Clusters. Table 1.4 gives various data sources for Big Data along with its examples of usages and the tools used.

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

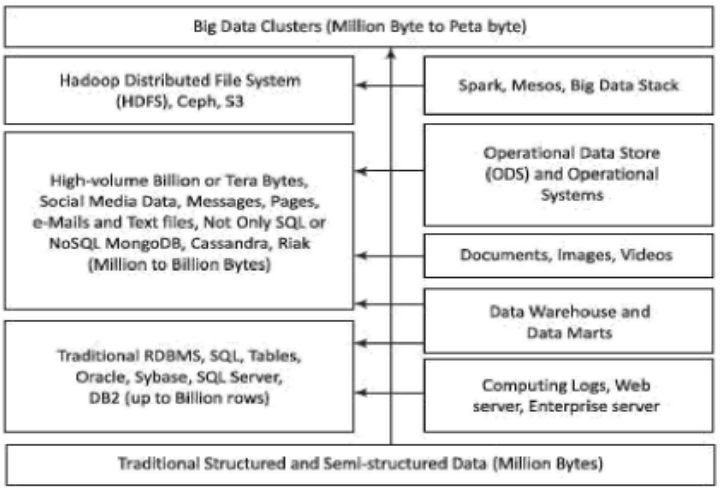

Figure 1.7 Coexistence of RDBMS for traditional server data, NoSQL and Hadoop, Spark and compatible Big Data Clusters

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. 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.

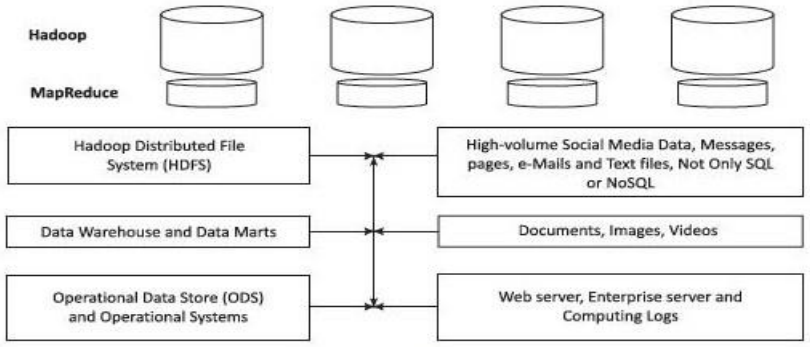

Figure 1.8 Hadoop based Big Data environment

- **Mesos** - Mesos v0.9 is a resources management platform which enables sharing of cluster of nodes by multiple frameworks and which has 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. The stack uses cluster of high performance machines. Table 1.5 gives Big Data management, storage and processing tools.

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 EC2, S3, GoogleQuery, Google App Engine, AWS Elastic Beanstalk, Salesforce Heroku
Storage	Hadoop Distributed File System, Amazon S3, Mesos

Big Data Analytics

Data is collected and analyzed to answer questions, test the hypotheses or disprove theories.

Definition: 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. Data analysis helps in finding business intelligence and helps in decision making.

Data analysis can be defined as, "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." (Wikipedia)

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 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.

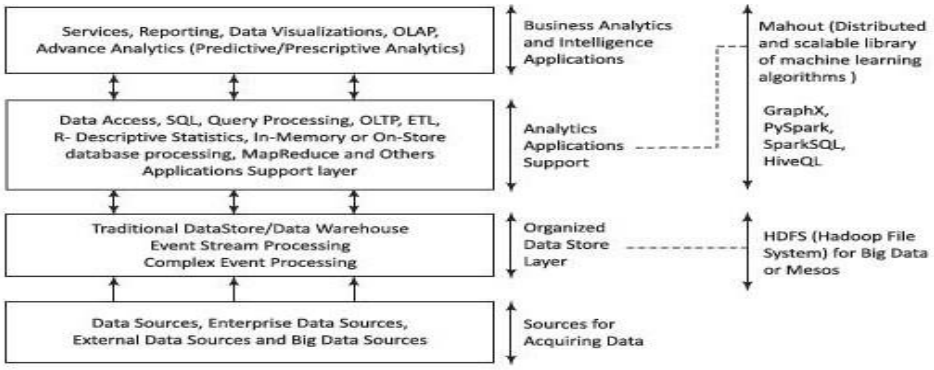

Figure 1.9 Traditional and Big Data analytics architecture reference model

The captured or stored data require a well-proven strategy to calculate, plan or analyze. When Big Data combine with high-powered data analysis, enterprise achieve valued business-related tasks.

Examples are:

• Determine root causes of defects, faults and failures in minimum time.

• Deliver advertisements on mobiles or web, based on customer's location and buying habits.

• Detect offender before that affects the organization or society.

Berkeley Data Analytics Stack (BDAS)

The 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:

1) cost reduction, 2) time reduction, 3) new product planning and development, 4) smart decision making using predictive analytics and 5) knowledge discovery.

Big Data analytics need innovative as well as cost effective techniques. BDAS is an open-source data analytics stack for complex computations on Big Data. It supports efficient, large-scale in-memory data processing, and thus enables user applications achieving three fundamental processing requirements; accuracy, time and cost.

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.

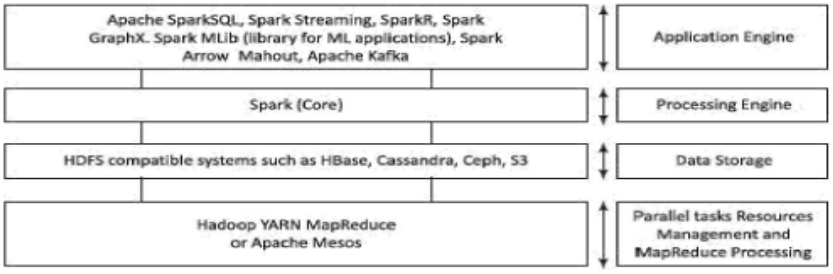

Figure 1.10 Four layers architecture for Big Data Stack consisting of Hadoop, MapReduce, Spark core and SparkSQL, Streaming, R, GraphX, MLib, Mahout, Arrow and Kafka

1.6 Big Data Analytics, Applications and Case Studies

Many 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. Some of the popular ones :

- **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.

  Following are the five application areas in order of the popularity of Big Data use cases:

  1. CVA using the inputs of evaluated purchase patterns, preferences, quality, price and post sales servicing requirements
  2. Operational analytics for optimizing company operations
  3. Detection of frauds and compliances
  4. New products and innovations in service
  5. Enterprise data warehouse optimization.

  An example of fraud is borrowing money on already mortgage assets. Example of timely compliances means returning the loan and interest installments by the borrowers. A few examples in service-innovation are as follows: A company develops software and then offers services like Uber.

  Big data is providing marketing insights into (i) most effective content at each stage of a sales cycle, (ii) investment in improving the customer relationship management (CRM), (iii) addition to strategies for increasing customer lifetime value (CLTV), (iv) lowering of customer acquisition cost (CAC).

  Contextual marketing means using an online marketing model in which a marketer sends to potential customers the targeted advertisements, which are based on the search terms during latest browsing patterns usage by customers. For example, if a customer is searching an airline for flights on a specific date from Delhi to Bangalore, then a smart travel agency targeting that customer through advertisements will show him/ her, at specific intervals, better options for another airline or different but cheap dates for travel or options in which price reduction occurs gradually. The following example explains the use of search engine optimization.

Big Data Analytics in Detection of Marketing Frauds

Fraud detection is vital to prevent financial loses to users. Fraud means someone deceiving deliberately. For example, 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, and much more. Big Data analytics enable fraud detection.

Big Data usages has the following features for enabling detection and prevention of frauds:

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

Big Data Risks

Large volume and velocity of Big Data provide greater insights but also associate risks with the data used. Data included may be erroneous, less accurate or far from reality. Analytics introduces new errors due to such data. Big Data can cause potential harm to individuals. For example, 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, described by Bernard Marr are data security, data privacy breach, costs affecting profits, bad analytics and bad data. Companies need to take risks of using Big Data and design appropriate risk management procedures. They have to implement robust risk management processes and ensure reliable predictions. Corporate, society and individuals must act with responsibility.

Big Data Credit Risk Management

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

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.

The insight using Big Data decreases the default rates in returning of loan, greater accuracy in issuing credit and faster identification of the non-payment or fraud issues of the loan receiving entities.

- ## Big Data and Healthcare

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

Healthcare analytics using Big Data can facilitate the following:

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

Value-based and customer-centric healthcare means cost effective patient care by improving healthcare quality using latest knowledge, usages of electronic health and medical records and 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 such as overutilization of services in short intervals, different hospitals in different locations simultaneously, or identical prescriptions for the same patient filed from multiple locations. Improving outcomes is possible by accurately diagnosing patient conditions, early diagnosis, predicting problems such as congestive heart failure, anticipating and avoiding complications, matching treatments with outcomes and predicting patients at risk for disease or readmission.

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 system— Dr. Eric Schadt and Sastry Chilukuri.

Following are some findings: building the health profiles of individual patients and predicting models for diagnosing better and offer better treatment,

1. Aggregating large volume and variety of information around from multiple sources the DNAs, proteins, and metabolites to cells, tissues, organs, organisms, and ecosystems, that can enhance the understanding of biology of diseases. Big data creates patterns and models by data mining and help in better understanding and research,
2. Deploying wearable devices data, the devices data records during active as well as inactive periods, provide better understanding of patient health, and better risk profiling the user for certain diseases,

- ## Big Data in advertisements

The impact of big data is tremendous on the digital advertisements industry. The digital advertisement industry sends advertisements using SMS, e-mails, 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. Big data help in identify management and can provide an advertising mix for building better branding exercising. 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.