BIG DATA ANALYTICS
NOTES

CHAPTER-1

INTRODUCTION:

Dawn of the Big Data Era:

The term of big data was coined under the explosive increase of global data and was mainly used to describe these enormous datasets. Compared with traditional datasets, big data generally includes masses of unstructured data that need more real-time analysis.

At present, big data has attracted considerable interest from industry, academia, and government agencies. For example, issues on big data are often covered in public media, including The Economist , New York Times, and National Public Radio etc. The era of big data is coming beyond all doubt. The rapid growth of cloud computing and the Internet of Things (IoT) further promote the sharp growth of data.

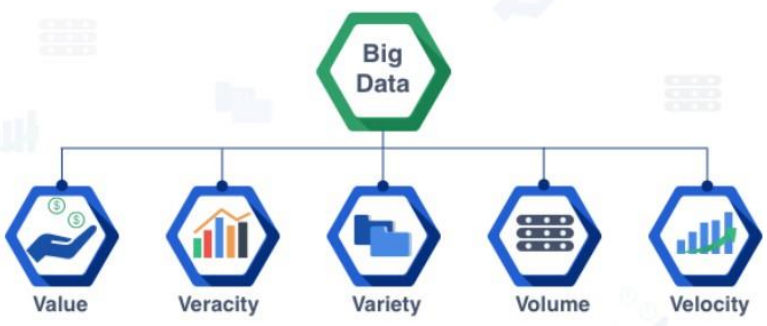
Definition and Features of Big Data:

Big data is an abstract concept. Apart from masses of data, it also has some other features, which determine the difference between itself and "massive data" or "very big data."

In 2010, Apache Hadoop defined big data as "datasets which could not be captured, managed, and processed by general computers within an acceptable scope." On the basis of this definition, in May 2011, McKinsey & Company, a global consulting agency announced Big Data as "the Next Frontier for Innovation, Competition, and Productivity." Big data shall mean such datasets which could not be acquired, stored, and managed by classic database software. This definition includes two connotations: First, the dataset volumes that conform to the standard of big data are changing, and may grow over time or with technological advances; Second, the dataset volumes that conform to the standard of big data in different applications differ from each other.

Features:

Big data is a collection of data from many different sources and is often described by five characteristics: volume, value, variety, velocity, and veracity.



- Volume: the size and amounts of big data that companies manage and analyze.

- Value: the most important “V” from the perspective of the business, the value of big data usually comes from insight discovery and pattern recognition that lead to more effective operations, stronger customer relationships and other clear and quantifiable business benefits
- Variety: the diversity and range of different data types, including unstructured data, semi-structured data and raw data
- Velocity: the speed at which companies receive, store and manage data – e.g., the specific number of social media posts or search queries received within a day, hour or other unit of time
- Veracity: the “truth” or accuracy of data and information assets, which often determines executive-level confidence.

The additional characteristic of variability can also be considered:

- Variability: the changing nature of the data companies seek to capture, manage and analyze – e.g., in sentiment or text analytics, changes in the meaning of key words or phrases.

The Development of Big Data:

Let's look at the development of big data. A project called Hadoop was born in 2005. Hadoop is a very important technology in the field of big data. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model. Many countries around the world and some research institutes have conducted some pilot projects on Hadoop, and have achieved a series of results.

In 2011, EMC held a global summit on Cloud Meets Big Data, and in May of the same year, McKinsey published a related research report. They expected that the so-called digital universe will contain 35 zettabytes of information within the next decade. EMC introduced what it is calling “The EMC Big Data Stack” defining their view of how to store, manage, and act on the big data coming downstream.

In December of the same year, China's Ministry of Industry and Information Technology issued the 12th Five-Year Development Plan for the Internet of Things. China will increase financial support for the smart industry, smart agriculture, smart logistics, smart transportation, smart grid, smart environmental protection, smart security, smart medical care, and smart home in the future. It represents the initial application of big data.

Between 2012 and 2015, many governments and companies around the world, including the United Nations, published a series of related ideas or outlines of action to promote the development of big data. After that, big data has entered a high-speed developing phase, and the 13th Five-Year Development Plan of the big data industry was born in China in 2017, which means that big data has begun to be widely used and developed at a high speed worldwide.

Challenges of Big Data:

1) **Data Representation:** many datasets have certain levels of heterogeneity in type, structure, semantics, organization, granularity, and accessibility. Data representation aims to make data more meaningful for computer analysis and user interpretation. Nevertheless, an improper data representation will reduce the value of the original data and may even obstruct effective data analysis. Efficient data representation shall reflect data structure, class, and type, as well as integrated technologies, so as to enable efficient operations on different datasets.

2) **Redundancy Reduction and Data Compression:** generally, there is a high level of redundancy in datasets. Redundancy reduction and data compression is effective to reduce the indirect cost of the entire system on the premise that the potential values of the data are not affected. For example, most data generated by sensor networks are highly redundant, which may be filtered and compressed at orders of magnitude.

3) **Data Life Cycle Management:** compared with the relatively slow advances of storage systems, pervasive sensors and computing are generating data at unprecedented rates and scales. We are confronted with a lot of pressing challenges, one of which is that the current storage system could not support such massive data.

4) **Analytical Mechanism:** the analytical system of big data shall process masses of heterogeneous data within a limited time. However, traditional RDBMSs are strictly designed with a lack of scalability and expandability, which could not meet the performance requirements. Non-relational databases have shown their unique advantages in the processing of unstructured data and started to become mainstream in big data analysis.

5) **Data Confidentiality:** most big data service providers or owners at present could not effectively maintain and analyze such huge datasets because of their limited capacity. They must rely on professionals or tools to analyze the data, which increase the potential safety risks. For example, the transactional dataset generally includes a set of complete operating data to drive key business processes. Such data contains details of the lowest granularity and some sensitive information such as credit card numbers. Therefore, analysis of big data may be delivered to a third party for processing only when proper preventive measures are taken to protect the sensitive data, to ensure its safety.

6) **Energy Management:** the energy consumption of mainframe computing systems has drawn much attention from both economy and environment perspectives. With the increase of data volume and analytical demands, the processing, storage, and transmission of big data will inevitably consume more and more electric energy. Therefore, system-level power consumption control and management mechanisms shall be established for big data while expandability and accessibility are both ensured.

7) **Expendability and Scalability:** the analytical system of big data must support present and future datasets. The analytical algorithm must be able to process increasingly expanding and more complex datasets.

8) **Cooperation:** analysis of big data is an interdisciplinary research, which requires experts in different fields to cooperate to harvest the potential of big data. A comprehensive big data network architecture must be established to help scientists and engineers in various fields access different kinds of data and fully utilize their expertise, so as to cooperate to complete the analytical objectives.

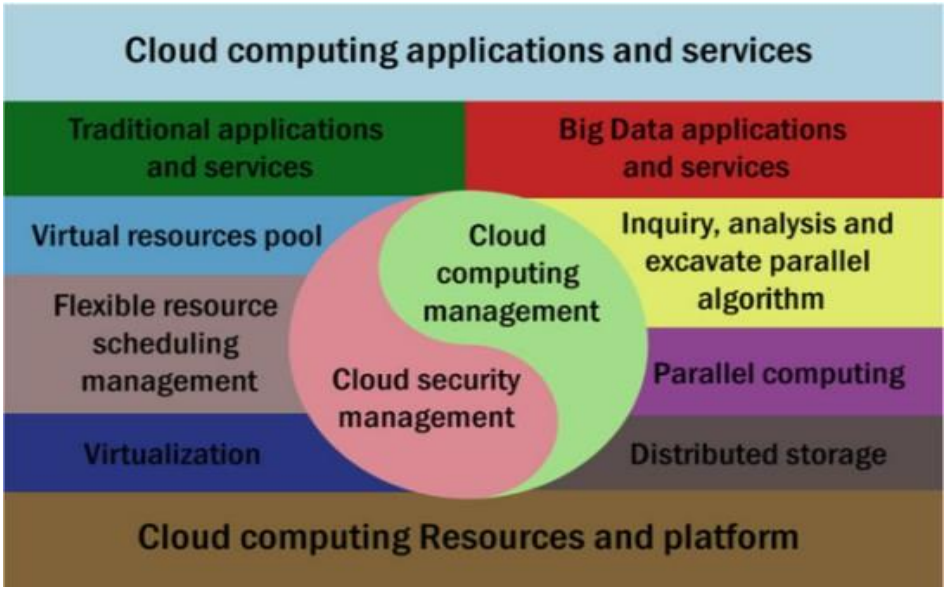
BIG DATA RELATED TECHNOLOGIES:

Cloud Computing Preliminaries:

Cloud Computing is evolved from Distributed Computing, Parallel Computing, and Grid Computing, or a commercial realization of the computer-scientific concept. In a narrow sense, cloud computing means the delivery and use mode of IT infrastructure, i.e., acquiring necessary resources through the Internet on-demand or in an expandable way. In a general sense, cloud computing means the delivery and use mode of services, i.e., acquiring necessary services through the Internet on-demand or in an expandable way.

Relationship Between Cloud Computing and Big Data:

Cloud computing is closely related to big data. Big data is the object of the computation operation and stresses the storage capacity and computing capacity of a cloud server. The main objective of cloud computing is to use huge computing resources and computing capacities under concentrated management, so as to provide applications with resource sharing at a granularity and provide big data applications with computing capacity. On the other hand, the emergence of big data also accelerates the development of cloud computing. The distributed storage technology based on cloud computing allows effective management of big data; the parallel computing capacity by virtue of cloud computing can improve the efficiency of acquiring and analyzing big data.



IoT Preliminaries:

The basic idea of IoT is to connect different objects in the real world, such as RFID, bar code readers, sensors, and mobile phones, etc., to realize information exchange and to make them cooperate with each other to complete a common task. IoT is deemed as the extension of the Internet and is an important part of the future Internet. IoT is mainly characterized with that it accesses every object in the physical world such that the objects can be addressed, controlled, and communicated with.

Compared with the Internet, IoT has the following main features.

• Various terminal equipments

• Automatic data acquisition

• Intelligent terminals

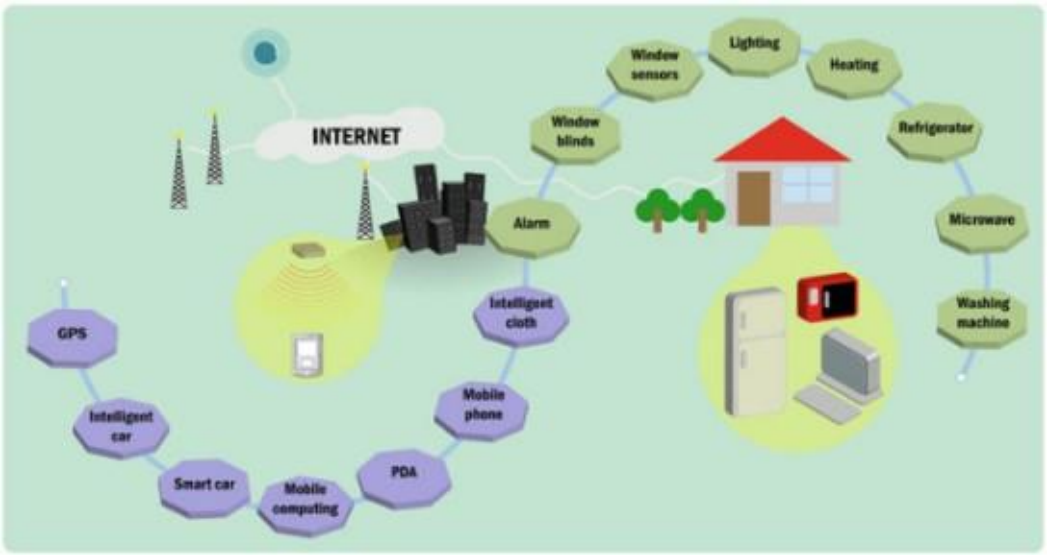


Illustration of the IoT architecture

Relationship Between IoT and Big Data:

The big data generated by IoT has different characteristics compared with general big data because of the different types of data collected, of which the most classical characteristics include heterogeneity, variety, unstructured feature, noise, and rapid growth.A report from Intel pointed out that big data in IoT has three features that conform to the big

data paradigm: (a) abundant terminals generating masses of data; (b) data generated by IoT is usually semi-structured or unstructured; (c) data of IoT is useful only when it is analyzed At present, the data processing capacity of IoT has fallen behind the collected data and it is extremely urgent to accelerate the introduction of big data technologies to catch up with the development of IoT. On one hand, the widespread deployment of IoT drives the high growth of data both in quantity and category, thus providing the opportunity for the application and development of big data. On the other hand, the application of big data technology.

Data Center:

In the big data paradigm, a data center is not only an organization for concentrated storage of data, but also undertakes more responsibilities, such as acquiring data, managing data, organizing data, and leveraging the data values and functions. Data centers are mainly concerned with "data" other than "center."

- Enterprises must take the development of data centers into consideration to improve the capacity of rapidly and effectively processing of big data under limited price/performance ratio. The data center shall provide the infrastructure with a large number of nodes, build a high-speed internal network, effectively dissipate heat, and effectively backup data. Only when a highly energy-efficient, stable, safe, expandable, and redundant data center is built, the normal operation of big data applications may be ensured.

- Many big data applications have developed their unique architectures and directly promote the development of storage, network, and computing technologies related to data centers. As the scale of data centers is increasingly expanding, it is also an important issue on how to reduce the operational cost for the development of data centers.

- In the big data paradigm, a data center shall not only be concerned with hardware facilities but also strengthen soft capacities, i.e., the capacities of acquisition, processing, organization, analysis, and application of big data. The data center may help business personnel analyze the existing data, discover problems in business operations, and develop solutions from big data.

Hadoop:

Hadoop is a technology closely related to big data, which forms a powerful big data systematic solution through data storage, data processing, system management, and integration of other modules. Such technology has become indispensable to cope with the challenges of big data. Hadoop is a set of large-scale software infrastructures for Internet applications similar to Google's FileSystem and MapReduce.

Relationship between Hadoop and Big Data:

Presently, Hadoop is widely used in big data applications in the industry, e.g., spam filtering, network searching, clickstream analysis, and social recommendation. In addition, considerable academic research is now based on Hadoop. Big Data is high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. More than 80% of data captured today is unstructured, from sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals, etc. All of this unstructured data is Big Data.

CHAPTER-2

INTRODUCTION TO HADOOP:

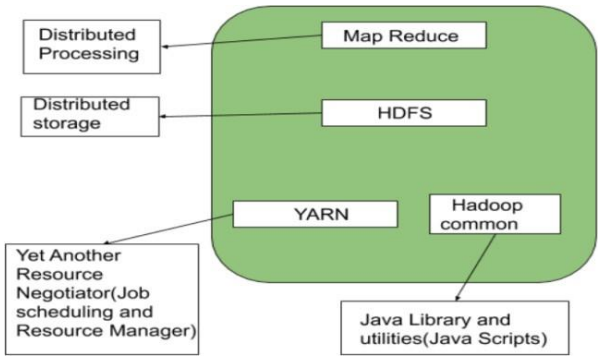
Hadoop, as a Big Data framework, provides businesses with the ability to distribute data storage, parallel processing, and process data at higher volume, higher velocity, variety, value, and veracity. HDFS, MapReduce, and YARN are the three major components for this Hadoop.

Hadoop HDFS uses name nodes and data nodes to store extensive data. MapReduce manages these nodes for processing, and YARN acts as an Operating system for Hadoop in managing cluster resources.

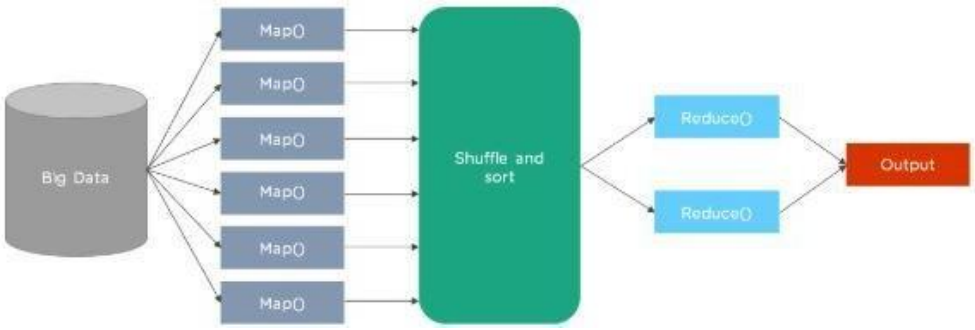
Hadoop Architecture:

The Hadoop Architecture Mainly consists of 4 components they are:-

1. MapReduce
2. HDFS(Hadoop distributed File System)
3. YARN(Yet Another Resource Framework)
4. Hadoop Common

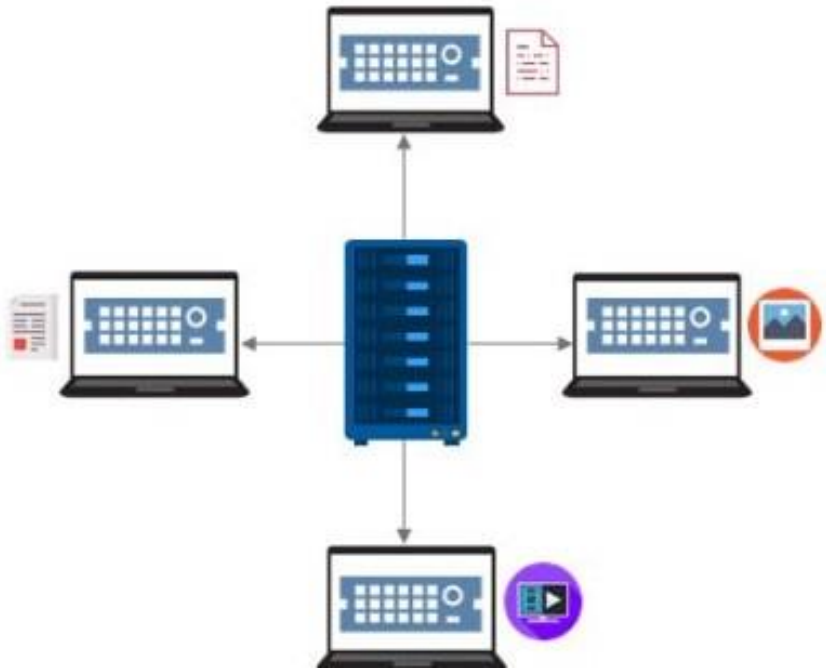


1. **Map Reduce:** Hadoop data processing is built on MapReduce, which processes large volumes of data in a parallelly distributed manner. With the help of the figure below, we can understand how MapReduce works:



As we see, we have our big data that needs to be processed, with the intent of eventually arriving at an output. So in the beginning, input data is divided up to form the input splits. The first phase is the Map phase, where data in each split is passed to produce output values. In the shuffle and sort phase, the mapping phase's output is taken and grouped into blocks of similar data. Finally, the output values from the shuffling phase are aggregated. It then returns a single output value.

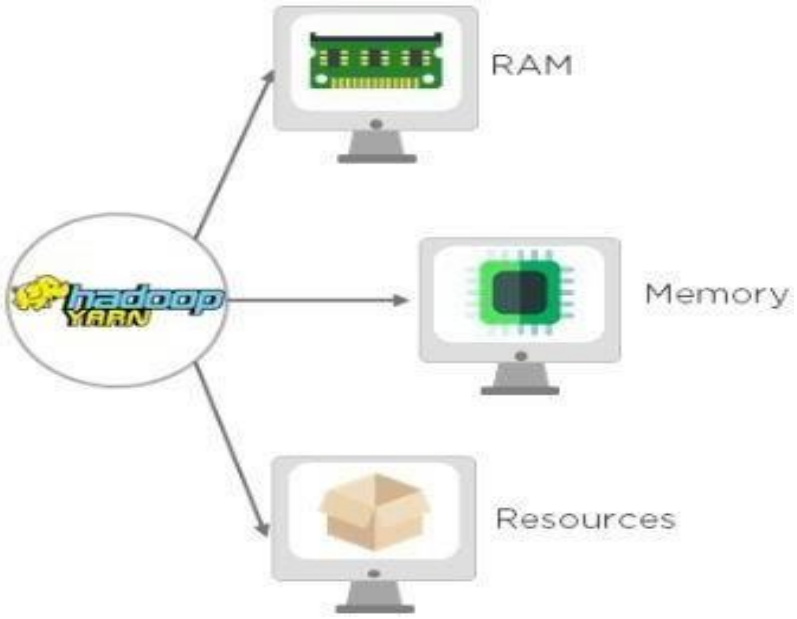
2. **HDFS:** In the traditional approach, all data was stored in a single central database. With the rise of big data, a single database was not enough to handle the task. The solution was to use a distributed approach to store the massive volume of information. Data was divided up and allocated to many individual databases. HDFS is a specially designed file system for storing huge datasets in commodity hardware, storing information in different formats on various machines.



There are two components in HDFS:

1. NameNode - NameNode is the master daemon. There is only one active NameNode. It manages the DataNodes and stores all the metadata.
2. DataNode - DataNode is the slave daemon. There can be multiple DataNodes. It stores the actual data.

3. **YARN(Yet Another Resource Negotiator):** YARN is an acronym for Yet Another Resource Negotiator. It handles the cluster of nodes and acts as Hadoop's resource management unit. YARN allocates RAM, memory, and other resources to different applications.



YARN has two components :

1. ResourceManager (Master) - This is the master daemon. It manages the assignment of resources such as CPU, memory, and network bandwidth.
2. NodeManager (Slave) - This is the slave daemon, and it reports the resource usage to the Resource Manager.

4. **Common Utilities:** Hadoop common or Common utilities are nothing but our java library and java files or we can say the java scripts that we need for all the other components present in a Hadoop cluster. these utilities are used by HDFS, YARN, and MapReduce for running the cluster. Hadoop Common verifies that Hardware failure in a Hadoop cluster is common so it needs to be solved automatically in software by Hadoop Framework.

Common Hadoop Shell commands:

1. **Version Check:** To check the version of Hadoop.
2. **list Command**: List all the files/directories for the given Hdfs destination path.
3. **mkdir Command:** HDFS Command to create the directory in HDFS.
4. **put Command** : Copy file from single src, or multiple Srcs from local file system to the destination file system.
5. **get Command :**HDFS Command to copy files from Hdfs to the local file system.
6. **cat Command:** HDFS Command that copies source paths to stdout.

Anatomy of File Read:

Step 1: First the Client will open the file by giving a call to open() method on FileSystem object
Step 2: DistributedFileSystem calls the Namenode, using RPC (Remote Procedure Call), to determine the locations of the blocks for the first few blocks of the file.
Step 3: The client then calls read() on the stream. DFSInputStream, which has stored the DataNode addresses for the first few blocks in the file, then connects to the first closest DataNode for the first block in the file.
Step 4: Data is streamed from the DataNode back to the client, which calls read() repeatedly on the stream.
Step 5: When the end of the block is reached, DFSInputStream will close the connection to the DataNode
Step 6: Blocks are read in order, with the DFSInputStream opening new connections to datanodes as the client reads through the stream.

Anatomy of File Write:

Step 1: The client creates the file by calling create() method on DistributedFileSystem.
Step 2: DistributedFileSystem makes an RPC call to the namenode to create a new file in the filesystem's namespace, with no blocks associated with it.
Step 3: As the client writes data, DFSOutputStream splits it into packets, which it writes to an internal queue, called the data queue.

HADOOP ECOSYSTEM:

Hadoop Ecosystem is a platform or a suite which provides various services to solve the big data problems. It includes Apache projects and various commercial tools and solutions.

Following are the components that collectively form a Hadoop ecosystem:

1. HDFS: Hadoop Distributed File System
2. YARN: Yet Another Resource Negotiator
3. MapReduce: Programming based Data Processing
4. Spark: In-Memory data processing
5. PIG, HIVE: Query based processing of data services
6. HBase: NoSQL Database
7. Mahout, Spark MLLib: Machine Learning algorithm libraries
8. Solar, Lucene: Searching and Indexing
9. Zookeeper: Managing cluster
10. Oozie: Job Scheduling

PIG: Pig was basically developed by Yahoo which works on a pig Latin language, which is Query based language similar to SQL.It is a platform for structuring the data flow, processing and analyzing huge data sets.Pig does the work of executing commands and in the background, all the activities of MapReduce are taken care of. After the processing, pig stores the result in HDFS.

HIVE: With the help of SQL methodology and interface, HIVE performs reading and writing of large data sets. However, its query language is called as HQL (Hive Query Language).It is highly scalable as it allows real-time processing and batch processing both. Also, all the SQL data types are supported by Hive thus, making the query processing easier.

Apache Spark: It's a platform that handles all the process consumptive tasks like batch processing, interactive or iterative real-time processing, graph conversions, and visualization, etc.It consumes in memory resources hence, thus being faster than the prior in terms of optimization.

Apache HBase: It's a NoSQL database which supports all kinds of data and thus capable of handling anything in the Hadoop Database. It provides capabilities of Google's BigTable, thus able to work on Big Data sets effectively.At times where we need to search or retrieve the occurrences of something small in a huge database, the request must be processed within a short quick span of time.

Flume: Flume is another data collection and ingestion tool, a distributed service for collecting, aggregating, and moving large amounts of log data. It ingests online streaming data from social media, logs files, web server into HDFS.

Schedulers - Fair and Capacity:

- **Fair Scheduler:** Fair scheduling is a method of assigning resources to jobs such that all jobs get, on average, an equal share of resources over time. When there is a single job running, that job uses the entire cluster. When other jobs are submitted, tasks slots that free up are assigned to the new jobs, so that each job gets roughly the same amount of CPU time.it assigns

equal amount of resource to all running jobs. When the job completes, a free slot is assigned to a new job with equal amount of resources. Here, the resource is shared between queues.

- **Capacity Scheduler:** The Capacity Scheduler is designed to allow sharing a large cluster while giving each organization a minimum capacity guarantee. The central idea is that the available resources in the Hadoop cluster are partitioned among multiple organizations who collectively fund the cluster based on computing needs. There is an added benefit that an organization can access any excess capacity not being used by others. This provides elasticity for the organizations in a cost-effective manner. On the other hand, it assigns resources based on the capacity required by the organization. This is set up by queues for each organization with specified amount of capacity. The queue is based on FIFO scheduling.

SPARK:

Apache Spark is a lightning-fast cluster computing technology, designed for fast computation. It is based on Hadoop MapReduce and it extends the MapReduce model to efficiently use it for more types of computations, which includes interactive queries and stream processing. The main feature of Spark is its in-memory cluster computing that increases the processing speed of an application. Spark is designed to cover a wide range of workloads such as batch applications, iterative algorithms, interactive queries and streaming. Apart from supporting all these workload in a respective system, it reduces the management burden of maintaining separate tools.

Apache Spark: It's a platform that handles all the process consumptive tasks like batch processing, interactive or iterative real-time processing, graph conversions, and visualization, etc.It consumes in memory resources hence, thus being faster than the prior in terms of optimization.

Apache HBase: It's a NoSQL database which supports all kinds of data and thus capable of handling anything of Hadoop Database. It provides capabilities of Google's BigTable, thus able to work on Big Data sets effectively.At times where we need to search or retrieve the occurrences of something small in a huge database, the request must be processed within a short quick span of time.

Features of Apache Spark:

• **Speed** − Spark helps to run an application in Hadoop cluster, up to 100 times faster in memory, and 10 times faster when running on disk. This is possible by reducing number of read/write operations to disk. It stores the intermediate processing data in memory.

• **Supports multiple languages** − Spark provides built-in APIs in Java, Scala, or Python. Therefore, you can write applications in different languages. Spark comes up with 80 high-level operators for interactive querying.

• **Advanced Analytics** − Spark not only supports 'Map' and 'reduce'. It also supports SQL queries, Streaming data, Machine learning (ML), and Graph algorithms.

Resilient Distributed Datasets (RDDs):

RDDs are the main logical data units in Spark. There are a distributed collection of objects, which are stored in memory or on disks of different machines of a cluster. A single RDD can be divided into multiple logical partitions so that these partitions can be stored and processed on different machines of a cluster. RDDs are immutable (read-only) in nature. You cannot change an original RDD, but you can create new RDDs byperforming coarse-grain operations, like transformations, on an existing RDD.

Spark RDD Operations:

Two types of Apache Spark RDD operations are- Transformations and Actions. A Transformation is a function that produces new RDD from the existing RDDs but when we want to work with the actual dataset, at that point Action is performed. When the action is triggered after the result, the new RDD is not formed like a transformation. In this Apache Spark RDD operations tutorial we will get the detailed view of what is Spark RDD, what is the transformation in Spark RDD, various RDD transformation operations in SparkTransformations: These are functions that accept the existing RDDs as input and output one or more RDDs. However, the data in the existing RDD in Spark does not change as it is immutable. Some of the transformation operations are provided in the table below

Function	**Description**
map()	Returns a new RDD by applying the function on each data element
filter()	Returns a new RDD formed by selecting those elements of the source on which the function returns true
reduceByKey()	Aggregates the values of a key using a function
groupByKey()	Converts a (key, value) pair into a (key, <iterable value>) pair
union()	Returns a new RDD that contains all elements and arguments from the source RDD
intersection()	Returns a new RDD that contains an intersection of the elements in the datasets

Actions: Actions in Spark are functions that return the end result of RDD computations. It uses a lineage graph to load data onto the RDD in a particular order. After all of the transformations are done, actions return the final result to the Spark Driver. Actions are operations that provide non-RDD values. Some of the common actions used in Spark are given below:

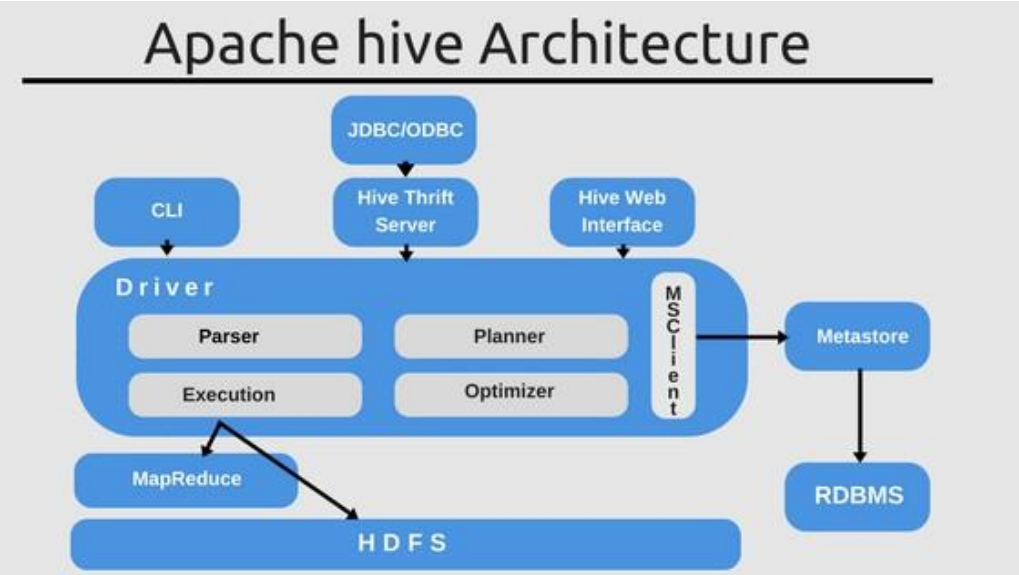
Function	**Description**
count()	Gets the number of data elements in an RDD
collect()	Gets all the data elements in an RDD as an array
reduce()	Aggregates data elements into an RDD by taking two arguments and returning one
take(n)	Fetches the first n elements of an RDD
foreach(operation)	Executes the operation for each data element in an RDD
first()	Retrieves the first data element of an RDD

RDD Lineage and RDD Persistence: Basically, evaluation of RDD is lazy in nature. It means a series of transformations are performed on an RDD, which is not even evaluated immediately. While we create a new RDD from an existing Spark RDD, that new RDD also carries a pointer to the parent RDD in Spark. That is the same as all the dependencies between the RDDs those are logged in a graph, rather than the actual data. It is what we call a lineage graph.

RDD lineage is nothing but the graph of all the parent RDDs of an RDD. We also call it an RDD operator graph or RDD dependency graph. To be very specific, it is an output of applying transformations to the spark. Then, it creates a logical execution plan. Also, physical execution plan or execution DAG is known as DAG of stages. Let's start with one example of Spark RDD lineage by using Cartesian or zip to understand well. However, we can also use other operators to build an RDD graph in Spark.

HIVE:

Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data query and analysis. Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop. Traditional SQL queries must be implemented in the MapReduce Java API to execute SQL applications and queries over distributed



Hive-Meta store: All Hive implementations need a meta store service, where it stores metadata. It is implemented using tables in a relational database. By default, Hive uses a built-in Derby SQL server. It provides single process storage, so when we use Derby, we cannot run instances of Hive CLI. Whenever we want to run Hive on a personal machine or for some developer task, then it is good, but when we want to use it in a cluster, then MySQL or any other similar relational database is required.

CHAPTER-3

BIG DATA GENERATION AND ACQUISITION :

Big data generation:

Data generation is the first step of big data. Specifically, it is large-scale, highly diverse, and complex datasets generated through longitudinal and distributed data sources. Such data sources include sensors, videos, click streams, and/or all other available data sources. At present, main sources of big data are the operation and trading information in enterprises, logistic and sensing information in the IoT, human interaction information and position information in the Internet world, and data generated in scientific research, etc.

1. Enterprise Data:

In 2013, IBM issued a report titled "Analytics: The Real-world Use of Big Data," which indicates that the internal data of enterprises are the main sources of big data. The internal data of enterprises mainly consists of online trading data and online analysis data, most of which are historically static data and are managed by RDBMSs in a structured manner. In addition, production data, inventory data, sales data, and financial data, etc., also constitute enterprise internal data, which aims to capture informationized and data-driven activities in enterprises, so as to record all activities of enterprises in the form of internal data.

2. IoT Data:

IoT is an important source of big data. Among smart cities constructed based on IoT, big data may come from industry, agriculture, traffic and transportation, medical care, public departments, and households, etc. According to the processes of data acquisition and transmission in IoT, its network architecture may be divided into three layers:

- **The sensing layer** - This layer is responsible for data acquisition and mainly consists of sensor networks.
- **The network layer** - The layer is responsible for information transmission and processing, where close transmission may rely on sensor networks, and remote transmission shall depend on the Internet.
- **The application layer** - This layer supports specific applications of IoT. According to the characteristics of IoT, the data generated from IoT has the following features:
- **Large-Scale Data:** In IoT, masses of data acquisition equipment are distributedly deployed, which may acquire simple numeric data or complex multimedia data. In order to meet the demands of analysis and processing, not only the currently acquired data, but also the historical data within a certain time frame should be stored.
- **Heterogeneity:** Because of the variety of data acquisition devices, the acquired data is also different and such data features heterogeneity.

● **Strong Time and Space Correlation:** In IoT, every data acquisition device is placed at a specific geographic location and every piece of data has a timestamp. The time and space correlations are important properties of data from IoT. During data analysis and processing, time and space are also important dimensions for statistical analysis.

● **Effective Data Accounts for Only a Small Portion of Big Data:** A great quantity of noises may occur during the acquisition and transmission of data in IoT. Among datasets acquired by acquisition devices, only a small amount of abnormal data is valuable.

3. Internet data:

Internet data consists of searching entries, Internet forum posts, chatting records, and microblog messages, among others, which have similar features, such as high value and low density. Such Internet data may be valueless individually, but through exploitation of accumulated big data, useful information such as habits and hobbies of users can be identified, and it is even possible to forecast users' behavior and emotional moods.

4. Bio-medical data:

As a series of high-throughput bio-measurement technologies are innovatively developed in the beginning of the twenty-first century, the frontier of research in the bio-medicine field also enters the era of big data. By constructing smart, efficient, and accurate analytical models and theoretical systems for bio-medicine applications, the essential governing mechanism behind complex biological phenomena may be revealed.

5.Data generation from other fields:

As scientific applications are increasing, the scale of datasets is gradually expand- ing, and the development of some disciplines greatly relies on the analysis of masses of data.In addition, pervasive sensing and computing among nature, commercial, Inter- net, government, and social environments are generating heterogeneous data with unprecedented complexity. These datasets have their unique data characteristics in scale, time dimension, and data category. For example, mobile data were recorded with respect to positions, movement, approximation degrees, communications, multimedia, use of applications, and audio environment.

Big Data Acquisition:

As the second phase of the big data system, big data acquisition includes data collection, data transmission, and data pre-processing. During big data acquisition, once the raw data is collected, an efficient transmission mechanism should be used to send it to a proper storage management system to support different analytical applications. The collected datasets may sometimes include much redundant or useless data, which unnecessarily increases storage space and affects the subsequent data analysis. Data compression techniques can be applied to reduce the redundancy. Therefore, data pre-processing operations are indispensable to ensure efficient data storage and exploitation.

1. Data Collection:

Data collection is to utilize special data collection techniques to acquire raw data from a specific data generation environment. Four common data collection methods are shown as follows.

- **Log files:** As one widely used data collection method, log files are record files automatically generated by the data source system, so as to record activities in designated file formats for subsequent analysis. Log files are typically used in nearly all digital devices.

To capture activities of users at the web sites, web servers mainly include the following three log file formats: public log file format (NCSA), expanded log format (W3C), and IIS log format (Microsoft). All the three types of log files are in the ASCII text format. Databases other than text files may sometimes be used to store log information to improve the query efficiency of the massive log store. There are also some other log files based on data collection, including stock indicators in financial applications and determination of operating states in network monitoring and traffic management.

- **Sensors:** Sensors are common in daily life to measure physical quantities and transform physical quantities into readable digital signals for subsequent processing (and storage). Sensory data may be classified as sound wave, voice, vibration etc. Sensed information is transferred to a data collection point through wired or wireless networks. Sometimes the accurate position of a specific phenomenon is unknown, and sometimes the monitored environment does not have the energy or communication infrastructures. The wireless communication must be used to enable data transmission among sensor nodes under limited energy and communication capability.

- **Methods for acquiring network data:** At present, network data acquisition is accomplished using a combination of web crawler, word segmentation system, task system, and index system, etc. Web crawler is a program used by search engines for downloading and storing web pages . Generally speaking, a web crawler starts from the uniform resource locator (URL) of an initial web page to access other linked web pages, during which it stores and sequences all the retrieved URLs. The current network data acquisition technologies mainly include traditional Libpcap-based packet capture technology, zero-copy packet capture technology, as well as some specialized network monitoring software such as Wireshark,SmartSniff, and WinNetCap.

- **Libpcap-Based Packet Capture Technology:**

Libpcap (packet capture library) is a widely used network data packet capture function library. It is a general tool that does not depend on any specific system and is mainly used to capture data in the data link layer. It features simplicity, easy-to-use, and portability, but has a relatively low efficiency. Therefore, under a high-speed network environment, considerable packet losses may occur when Libpcap is used.

● **Zero-Copy Packet Capture Technology:**

The so-called zero-copy (ZC) means that no copies between any internal memories occur during packet receiving and sending at a node. In sending, the data packets directly start from the user buffer of applications, pass through the network interfaces, and arrive at an external network. On receiving, the network interfaces directly send data packets to the user buffer. The basic idea of zero-copy is to reduce data copy times, reduce system calls, and reduce CPU load while datagrams are passed from network equipment to user program space.

● **Mobile Equipments:**

As mobile device functions become increasingly stronger, they feature more complex and multiple means of data acquisition as well as more variety of data.Mobile devices may acquire geographical location information through positioning systems; acquire audio information through microphones; acquire pictures,videos, streetscapes, two-dimensional barcodes, and other multimedia information through cameras; acquire user gestures and other body language information through touch screens and gravity sensors.

2. Data transportation:

Upon the completion of raw data collection, data will be transferred to a data storage infrastructure for processing and analysis Those data is mainly stored in a data center.Therefore, data transmission consists of two phases:

● **Inter-DCN transmissions** does the transmission from data source to data center through physical network infrastructure. The bandwidth of the electronic bottleneck is the limited key factor of the traditional optical transmission technologies. IP-based wavelength division multiplexing (WDM) network architecture, orthogonal frequency-division multiplexing (OFDM) are developed to ignore that key factor.

● **Intra-DCN transmissions** are the data communication flows within data centers.Intra-DCN transmissions depend on the communication mechanism within the data center (i.e., on physical connection plates, chips, internal memories of data servers, network architectures of data centers, and communication protocols).A data center consists of multiple integrated server racks interconnected with its internal connection networks.

3. Data preprocessing:

Because of the wide variety of data sources, the collected datasets vary with respect to noise, redundancy, and consistency, etc., and it is undoubtedly a waste to store meaningless data. In addition, some analytical methods have stringent requirements on data quality. Pre-processing data not only reduces storage expense, but also improves analysis accuracy. Some relational data pre-processing techniques are discussed in the following.

● **Integration:** Data integration is the cornerstone of modern commercial informatics, which involves the combination of data from different sources and provides users with a uniform view of data. Two methods have been widely recognized: data warehouse and data federation. Data warehousing includes a process named ETL (Extract, Transformand Load). Extraction involves connecting source systems, selecting, collecting,analyzing, and processing necessary data.

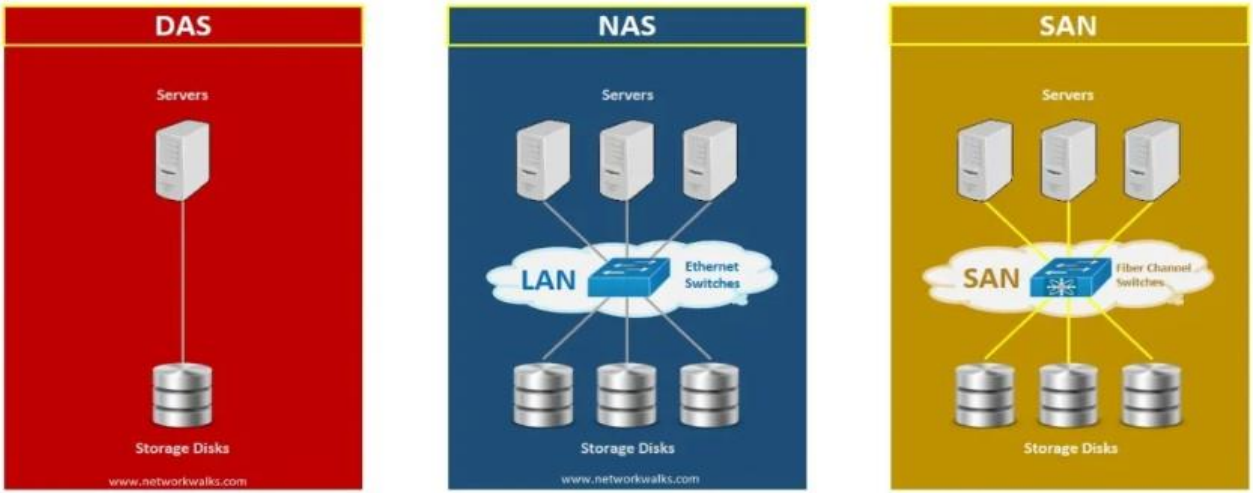
Transformation is the execution of a series of rules to transform the extracted data into standard formats. Loading means importing extracted and transformed data into the target storage infrastructure.

- **Cleaning:** Data cleaning is a process to identify inaccurate, incomplete, or unreasonable data, and then modify or delete such data to improve data quality. Generally, data cleaning includes five complementary procedures: defining and determining error types, searching and identifying errors, correcting errors, documenting error examples and error types, and modifying data entry procedures to reduce future errors.Classic data quality problems mainly come from software defects, customized errors, or system mis-configuration.Data cleaning is of vital importance to keep the data consistency.
- **Redundancy Elimination:** Data redundancy refers to data repetitions or surplus, which usually occurs in many Datasets. Data redundancy can increase the unnecessary data transmission expense and cause defect on storage systems, e.g., waste of storage space, leading to data inconsistency, reduction of data reliability, and data damage. Therefore, various redundancy reduction methods have been proposed, such as redundancy detection,data filtering, and data compression. Such methods may apply to different datasets. However, redundancy reduction may also bring about certain negative effects.

BIG DATA STORAGE:

1. Storage system for massive data:

Data storage refers to the storage and management of large-scale datasets, while achieving reliability and availability. A data storage system consists of two parts: infrastructure and data storage methods or mechanisms. The hardware infrastructure includes massive shared Information Communication Technology (ICT) resources utilized to feedback instant demands of tasks, and such ICT resources are organized in an elastic manner. Data storage equipment is becoming increasingly more important, and storage cost becomes the main expense of many Internet companies. A large number of storage systems emerge to meet the demands of big data. Existing storage technologies can be classified as DAS ,NAS and SA



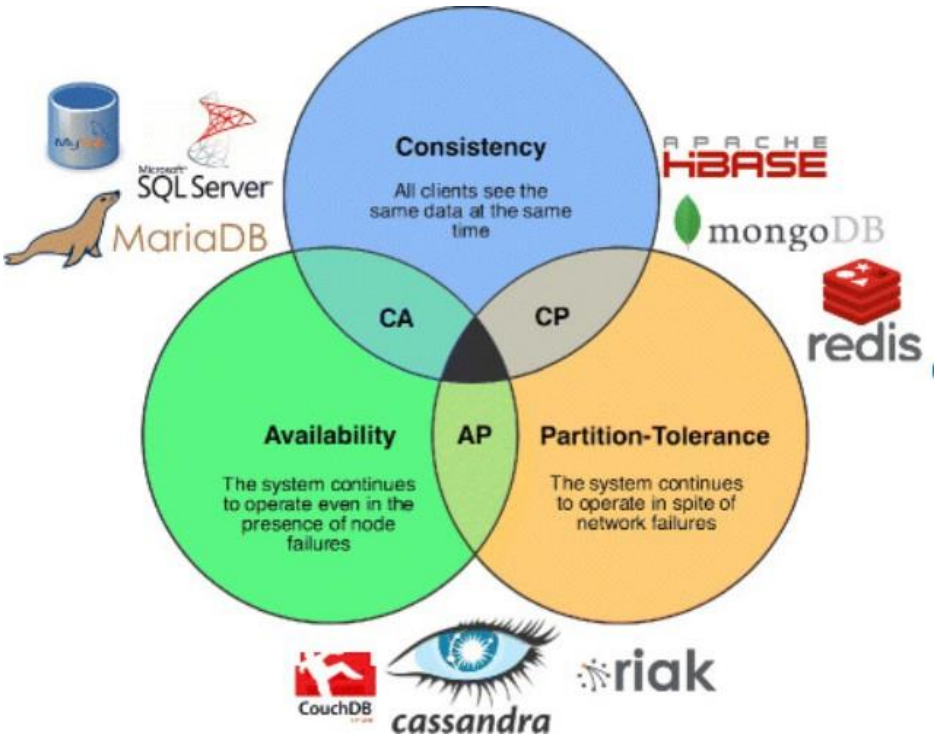
● **DAS(Direct Attached Storage):** In DAS, disc drives are directly connected with servers. DAS applies to a few server environments but, when the storage capacity is increased, the efficiency of storage supply will be quite low and the upgradeability and expandability will be greatly limited. DAS is mainly used in personal computers and small-sized servers, which only support such applications requiring low storage capacities and does not directly support multi-computer shared storage. Tap drivers and RAID (redundant array of independent disks) are classic DAS equipment.

● **NAS(Network Attached Storage):** NAS is actually an auxiliary storage equipment of a network. It is directly connected to a network through a hub or switch, communicating with the TCP/IP protocol. NAS is geared to message passing, and transmits data in the form of files. NAS has two prominent features. First, on physical connection, it directly connects the storage equipment to a network and then hangs the storage at the rear end of a server, thus avoiding the I/O burden at the server.

● **SAN(Storage Area Network):** SAN focuses on data storage with a flexible network topology and high-speed optical fiber connections. It allows multipath data switching among any internal nodes. Data storage management is located in a relatively independent storage local area network, so as to achieve a maximum degree of data sharing and data management, as well as seamless extension of the system.

2.Distributed storage system:

To use a distributed system to store massive data, the following factors should be taken into Consideration:



● **Consistency:** A distributed storage system requires multiple servers to cooperatively store data. As there are more servers, the probability of server failures will be larger. Usually data is divided into multiple pieces to be stored at different servers to ensure availability in case of server failure.Consistency refers to assuring that multiple copies of the same data are identical.
● **Availability:** A distributed storage system operates in multiple sets of servers. As more servers are used, server failures are inevitable. It would be desirable if the entire system is not serious affected with respect to serving the reading and writing requests from customer terminals. This property is called availability.
● **Partition tolerance:** Multiple servers in a distributed storage system are connected by a network. The network could have link/node failures or temporary congestion. The distributed system should have a certain level of tolerance to problems caused by network failures. It would be desirable that the distributed storage still works well when the network is partitioned.

Eric Brewer proposed a CAP [1, 2] theory in 2000, which indicated that a distributed system could not simultaneously meet the requirements on consistency, availability, and partition tolerance; at most two of the three requirements can be satisfied simultaneously.

CA system by ignoring partition tolerance,i.e,they could not handle network failures. Therefore, CA systems are generally deemed as storage systems with a single server, such as the traditional small-scale relational databases.Such systems feature single copy of data, such that consistency is easily ensured. Availability is guaranteed by the excellent design of relational databases. However, since CA systems could not handle network failures, they could not be expanded to use many servers.

CP systems by ignoring availability, i.e, could not ensure sound availability CP systems generally maintain several copies of the same data in order to ensure a level of fault Tolerance . CP systems also ensure data consistency, i.e., multiple copies of the same data are guaranteed to be completely identical. Eg: BigTable and Hbase.

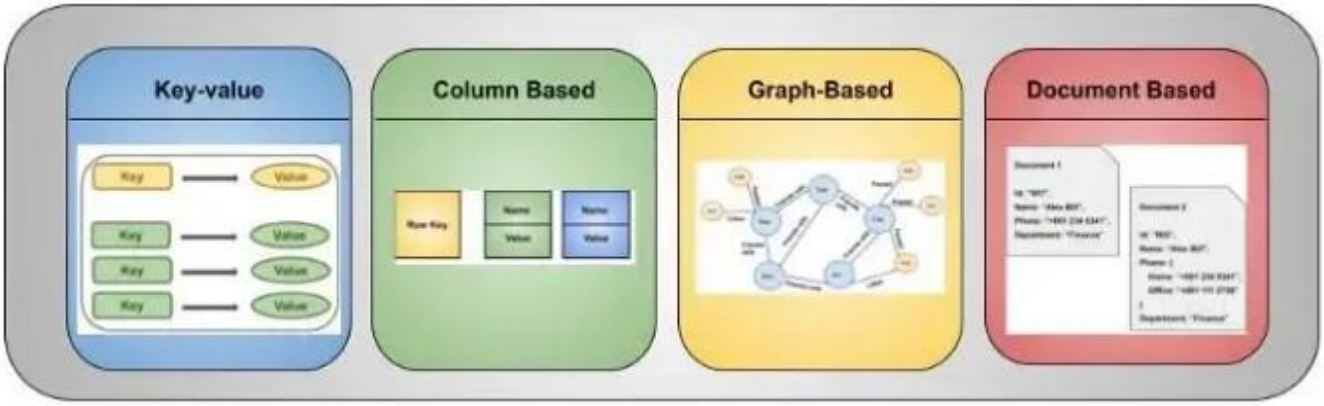
AP systems that ignore consistency, AP systems only ensure eventual consistency rather than strong consistency, accurate data can still be obtained after a certain amount of delay. Eg: Dynamo and Cassandra.

3.Storage management for big data:

Considerable research on big data promotes the development of storage mechanisms for big data. Existing storage mechanisms of big data may be classified into three bottom-up levels:
● File systems
● Databases
● Programming models

Database technology:

Various database systems are developed to handle datasets at different scales and support various applications. NoSQL databases (i.e., non traditional relational databases) are becoming more popular for big data storage. NoSQL databases feature flexible modes, support for simple and easy copy, simple API, eventual consistency, and support of large volume data. NoSQL databases are becoming the core technology for big data.



● Key-value databases:

Key-value Databases are constituted by a simple data model and data is stored corresponding to key-values. Every key is unique and customers may input queried values according to the keys. Such databases feature a simple structure and the modern key-value databases are characterized with high expandability and smaller query response time higher than those of relational databases.

1. **Dynamo:** Dynamo is a highly available and expandable distributed key-value data storage system. It is used to manage store status of some core services in the Amazon e-Commerce Platform. Amazon e-Commerce Platform provides multiple services and data storage that can be realized with key access. The Dynamo interface is constituted by simple reading and writing of data items. Dynamo achieves elasticity and availability through the data partition, data copy, and object edition mechanisms. The Dynamo partition plan relies on Consistent Hashing to divide load for multiple main storage machines.

2. **Voldemort:** Voldemort is also a key-value storage system, which was initially developed for and is still used by LinkedIn. Key words and values in Voldemort are composite objects constituted by tables and images. The voldemort interface includes three simple operations: reading, writing, and deletion, all of which are confirmed by key words. Voldemort provides asynchronous updating concurrent control of multiple editions but does not ensure data consistency. Voldemort may store data in RAM but allows data to be inserted into a storage engine. It is worth noting that Voldemort supports two storage engines including Berkeley DB and Random Access Files.

● Column-oriented databases:

The column-oriented databases store and process data according to columns other than rows. Columns and rows are segmented in multiple nodes to realize expandability.

1. **BigTable:** BigTable is a distributed, structured data storage system, which is designed to process the large-scale (PB class) data among thousands of commercial servers. The basic data structure of BigTable is a multidimensional sequenced mapping with sparse, distributed, and persistent storage. The BigTable API features the creation and deletion of Tablets and column families as well as modification of metadata of clusters, tables, and column families, and access control rights.

2. **Cassandra:** Cassandra is a distributed storage system to manage the huge amount of structured data distributed among multiple commercial servers . The system was developed by Facebook and became an open source tool in 2008. It adopts the ideas and concepts of both Amazon Dynamo and Google BigTable, especially integrating the distributed system technology of Dynamo with the BigTable data model.

● Document databases:

Compared with key-value storage, document storage can support more complex data forms.

1. **MongoDB:** MongoDB is an open-source document-oriented database. MongoDB stores documents as Binary JSON (BSON) objects, which is similar to objects. Every document has an ID field as the main keyword. Query in MongoDB is expressed with syntax similar to JSON. A database driver sends the query as a BSON object to MongoDB. The system allows querying on all documents, including embedded objects and arrays. Indexes may be created for queryable fields in documents to enable rapid query.

2. **SimpleDB:** SimpleDB is a distributed database and a web service of Amazon. Data in SimpleDB is organized into various domains in which data may be stored, acquired, and queried. Domains include different properties and name/value pair sets of projects.Date is copied to different machines at different data centers in order to ensure data safety and improve performance. This system does not support automatic partition and thus could not be expanded with the change of data volume.

3. **CouchDB:** Apache CouchDB is a document-oriented database written in Erlang . Data in CouchDB is organized into documents that consist of fields named by keywords/names and values, and are stored and accessed as JSON objects. Every document is provided with a unique identifier.

● Platform for Nimble universal table storage:

Platform for Nimble Universal Table Storage (PNUTS) is a large-scale parallel geographically-distributed system for Yahoo!'s web applications. It relies on a simple relational data model in which data is organized into a property record table. In addition to the classic data types, blob (binary large object or basic large object) is also an effective data type that allows any structures within records.

Design factors:

Of the various database systems, there is not a single system that can achieve the optimal performance under all workload circumstances. In each database system, some performance goals have to be compromised to achieve optimized operation for specific applications the design factors are:

- **Data Model:**

This section examined three core data models, i.e. key-value, column, and document models. In particular, PNUTS uses a row-oriented data Model.

- **Data Storage:**

In some systems data are designed to be stored in RAM and their snapshots or copies are stored in discs. Other systems store data in discs, with the cache stored in RAM. A few systems have pluggable background programs that are allowed to use different data storage media, or standardized underlying document systems are required.

- **Concurrency Control:**

There are three concurrency control mechanisms used in the existing systems: lock, MVCC, and non-concurrency control. The lock mechanism only allows a user to read or modify a real object (i.e., object, document, or row) at any time. The MVCC mechanism ensures the reading consistency.

- **Consistency:**

According to the CAP theorem, strict consistency could not be simultaneously achieved along with availability and partition tolerance. The weak consistency, eventual consistency, and time axis consistency of both types should be generally compromised for each other.

- **CAP Option:**

The CAP theorem indicates that a shared data system may achieve at most two properties, among consistency, availability, and partition tolerance. Databases based on cloud computing need to copy data from different servers in order to handle system failure in some regions, which basically requires consistency and availability. This way, the trade-off between consistency and availability can bedetermined.

Database Programming Model:

The massive datasets of big data are generally stored in hundreds and even thousands of commercial servers. Apparently, the traditional parallel models (e.gMessage Passing Interface (MPI) and Open Multi-Processing (OpenMP)) may not be adequate to support such large-scale parallel programs. Some parallel programming modes have been proposed:

- **MapReduce:** MapReduce is a simple but powerful programming model for large-scale computing using a large number of clusters of commercial PCs to achieve automatic parallel processing and distribution. In MapReduce, the computational workload is caused by inputting key-value pair sets and generating key-value pair sets. The computing model only has two functions, i.e., Map and Reduce, both of which are programmed by users. The Map function processes input and generates intermediate key-value pairs. Then, MapReduce will combine all the intermediate values related to the same key and transmit them to the Reduce function. Next,

the Reduce function receives the intermediate key and its value set, merges them, and generates a smaller value set.

- **Dryad:** Dryad is a general-purpose distributed execution engine for processing parallel applications of coarse-grained data. The operational structure of Dryad is a directed acyclic graph, in which vertices represent programs and edges represent data channels. Dryad executes operations on the vertices in computer clusters and transmits data via data channels, including documents, TCP connections, and shared-memory FIFO. During operation, resources in a logic operation graph are automatically map to physical resources.
- **All-Pairs:** All-Pairs [33] is a system specially designed for biometrics, bioinformatics, and data mining applications. It focuses on comparing element pairs in two datasets by a given function. The All-Pairs problem may be expressed as a three-tuples (Set A,Set B, and Function F), in which Function F is utilized to compare all elements in Set A and Set B. All-Pairs is implemented in four phases: system modeling, input data distribution, batch job management, and result collection.
- **Pregel:** The Pregel system of Google facilitates the processing of large-sized graphs, e.g., analysis of network graphs and social networking services. A computational task is expressed by a directed graph constituted by vertices and directed edges, in which every vertex is related to a modifiable and user-defined value. Directed edges are related to their source vertices and every edge is constituted by a modifiable and user-defined value and an identifier of a target vertex. After the graph is built, the program conducts iterative calculations, which are called supersteps among which global synchronization points are set until algorithm completion and output completion.

CHAPTER-4

BIG DATA ANALYSIS:

Traditional Data Analysis: Traditional data analysis means to use proper statistical methods to analyze massive first-hand data and second-hand data Big data analysis can be deemed as the analysis of a special kind of data. Therefore, many traditional data analysis methods may still be utilized for big data analysis. Several representative traditional data analysis methods are examined in the following

• **Cluster Analysis:** cluster analysis is a statistical method for grouping objects, and specifically, classifying objects according to some features. Cluster analysis is an unsupervised study method without the use of training data.

• **Factor Analysis:** grouping several closely related variables and then every group of variables becomes a factor (called a factor because it is unobservable, i.e., not a specific variable), and the few factors are then used to reveal the most valuable information of the original data.

• **Correlation Analysis:** correlation analysis is an analytical method for determining the law of correlations among observed phenomena and accordingly conducting forecast and control

• **Regression Analysis:** regression analysis is a mathematical tool for revealing correlations between one variable and several other variables. Based on a group of experiments or observed data, regression analysis identifies dependence relationships among variables hidden by randomness

• **A/B Testing**: also called bucket testing. It is a technology for determining plans to improve target variables by comparing the tested group.

• **Statistical Analysis:** Statistical analysis is based on the statistical theory, a branch of applied mathematics. In statistical theory, randomness and uncertainty are modeled with Probability Theory. Statistical analysis can provide description and inference for large-scale datasets. Descriptive statistical analysis can summarize and describe datasets and inferential statistical analysis draws conclusions from data subject to random variations.

• **Data Mining:** Data mining is a process for extracting hidden, unknown, but potentially useful information and knowledge from massive, incomplete, noisy, fuzzy, and random data. Big Data Analytic Methods: At present, the main processing methods of big data are shown as follows.

• **Bloom Filter:** Bloom Filter is actually a bit array and a series of Hash functions. The principle of Bloom Filter is to store Hash values of data other than data itself by utilizing a bit array, which is in essence a bitmap index that uses Hash functions to conduct loss compression storage of data

• **Hashing:** it is a method that essentially transforms data into shorter fixed-length numerical values or index values.

• **Index:** index is always an effective method to reduce the expense of disc reading and writing, and improve insertion, deletion, modification, and query speeds in both traditional relational

databases that manage structured data, and technologies that manage semi-structured and unstructured data

• **Triel:** also called trie tree, a variant of Hash Tree. It is mainly applied to rapid retrieval and word frequency statistics. The main idea of Trial is to utilize common prefixes of character strings to reduce comparison on character strings to the greatest extent, so as to improve query efficiency

• **Parallel Computing:** compared to traditional serial computing, parallel computing refers to utilizing several computing resources to complete a computation task. Its basic idea is to decompose a problem and assign them to several independent processes to be independently completed, so as to achieve coprocessing.

Architecture for Big Data Analysis:

Due to the wide range of sources and variety, different structures, and the broad application fields of big data, different analytical architectures shall be considered for big data with different application requirements.

• **Real-Time vs. Offline Analysis:** Big data analysis can be classified into real-time analysis and off-line analysis according to the real-time requirement. Real-time analysis is mainly used in Ecommerce and finance. . The main existing architectures of real time analysis include

(a) parallel processing clusters using traditional relational databases.

(b) memory-based computing platforms.

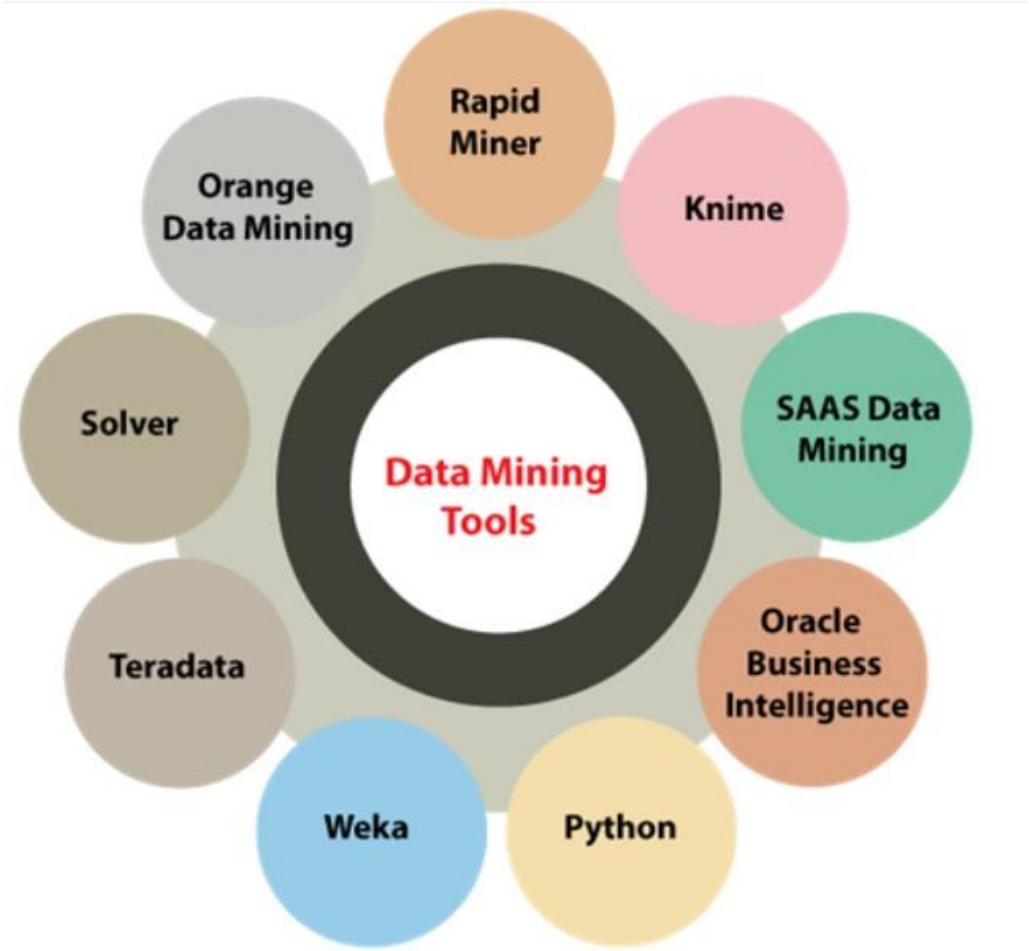
• **Analysis at Different Levels:** Big data analysis can also be classified into memory level analysis, Business Intelligence (BI) level analysis, and massive level analysis, which are examined in the following.

1. **Memory-Level:** Memory-level analysis is for the case when the total data volume is within the maximum level of the memory of a cluster. The memory of the current server cluster surpasses hundreds of GB while even the TB level is common. Memory-level analysis is extremely suitable for real-time analysis. MongoDB is a representative memory-level analytical architecture. With the development of SSD (Solid-State Drive), the capacity and performance of memory-level data analysis has been further improved and widely applied.
2. **BI:** BI analysis is for the case when the data scale surpasses the memory level but may be imported into the BI analysis environment. Currently, mainstream BI products are provided with data analysis plans supporting the level over TB.
3. **Massive:** Massive analysis for the case when the data scale has completely surpassed the capacities of BI products and traditional relational databases. At present, most massive analyses utilize HDFS or Hadoop to store data and use MapReduce for data analysis. Most massive analysis belongs to the offline analysis category.

• **Analysis with Different Complexity**: The time and space complexity of data analysis algorithms differ greatly from each other according to different kinds of data and application

demands. For example, for applications that are amenable to parallel processing, a distributed algorithm may be designed and a parallel processing model may be used for data analysis.

Tools for Big Data Mining and Analysis: Many tools for big data mining and analysis are available, including professional and free open source software. review the top five widely used software, according to a survey of "What Analytics, Data mining, Big Data software you used in the past 12 months for a real project" of 798 professionals made by KDNuggets in 2012 .



• **R (30.7 %):** R, an open source programming language and software environment, is designed for data mining/analysis and visualization. In addition, skilled users may directly call R objects in C. R is a realization of the S language. S is an interpreted language S was mainly implemented in S-PLUS, but S-PLUS is a commercial software. Compared to S, R is more popular since it is open source.

• **Excel (29.8 %):** Excel, a core component of Microsoft Office, provides powerful data processing and statistical analysis capability, and aids decision making. When Excel is installed, some advanced plug-ins, such as Analysis ToolPak and Solver Add-in, with powerful functions for data analysis are also integrated but such plug-ins can be used only if users enable them. Excel is also the only commercial software among the top five.

• **Rapid-I Rapidminer (26.7 %):** Rapidminer is an open source software used for data mining, machine learning, and predictive analysis. In an investigation of KDnuggets in 2011, it was more

frequently used than R (ranked Top 1). Data mining and machine learning programs provided by RapidMiner include Extract, Transform and Load (ETL), data pre-processing and visualization, modeling, evaluation, and deployment. RapidMiner is written in Java

• **KNIME (21.8 %):** KNIME (Konstanz Information Miner) is a user-friendly, intelligent, and open-source-rich data integration, data processing, data analysis, and data mining platform [4]. KNIME was written in Java and, based on Eclipse, provides more functions as plug-ins. Through plugin files, users can insert processing modules to files,pictures, and time series, and integrate them into various open source projects, e.g., R and Weka.

• **Weka/Pentaho (14.8 %):** Weka, abbreviated from Waikato Environment for Knowledge Analysis, is a free and open-source machine learning and data mining software written in Java. Weka provides such functions as data processing, feature selection, classification, regression, clustering, association rule, and visualization, etc. Pentaho is one of the most popular open-source commercial intelligent software. It is a BI kit based on the Java platform Weka's data processing algorithms are also integrated in Pentaho and can be directly called.

CHAPTER-5

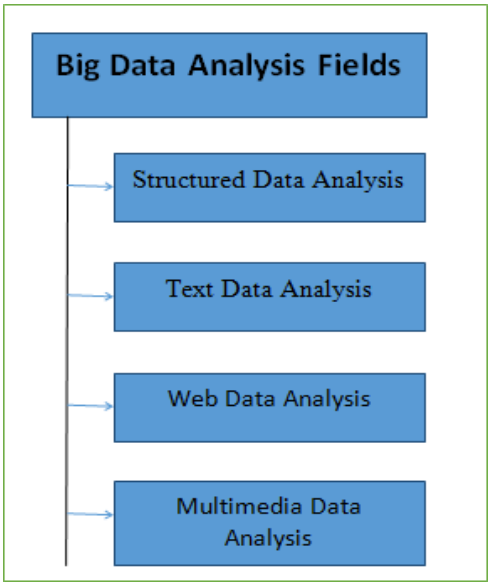
BIG DATA APPLICATIONS:

Application Evolution:

Recently, big data and big data analysis has been proposed for describing datasets and as analytical technologies in large-scale complex programs, which need to be analyzed with advanced analytical methods. Some potential and influential applications from different fields and their data and analysis characteristics are discussed as follows.

- **Evolution of Commercial Applications:** The beginning of the twenty-first century, networks and websites provided a unique opportunity for organizations to have online displays and directly interact with customers. Abundant products and customer information, including clickstream data logs and user behavior, etc., can be acquired from the websites.
- **Evolution of Network Applications:** Network data accounts for a major percentage of the global data volume. The Web has become a common platform for interconnected pages, full of various kinds of data, such as text, images, videos, pictures, and interactive contents, etc. Therefore, plentiful advanced technologies used for semi-structured or unstructured data emerged at the right moment.
- **Evolution of Scientific Applications:** The U.S. The National Science Foundation (NSF) has recently announced the BIG DATA Research Initiative to promote research efforts to extract knowledge and insights from large and complex collections of digital data. Some scientific research disciplines have developed massive data platforms and obtained useful outcomes.IPlant dataset have high varieties in form, including specification or reference data, experimental data, analog or model data, observation data, and other derived data.

Big Data Analysis Fields:



Data analysis research can be divided into six key technical fields, i.e., structured data analysis, text data analysis, website data analysis, multimedia data analysis, network data analysis, and mobile data analysis.

- **Structured Data Analysis :**

Business applications and scientific research may generate massive structured data, of which the management and analysis rely on mature commercialized technologies, such as RDBMS, data warehouse, OLAP, and BPM (Business Process Management). Data analysis is mainly based on data mining and statistical analysis. Exploiting data characteristics, time and space mining may extract knowledge structures hidden in high-speed data flows and sensor data models and modes.

- **Text Data Analysis** :

The most common format of information storage is text, e.g., email communication, business documents, web pages, and social media. Therefore, text analysis is deemed to feature more business-based potential than structured data mining. Generally, tax analysis, also called text mining, is a process to extract useful information and knowledge from unstructured text. Most text mining systems are based on natural language processing (NLP), NLP can enable computers to analyze, interpret, and even generate text. Some NLP-based technologies have been applied to text mining, including information extraction, topic models, text summarization, classification, clustering, question answering, and opinion mining.

- **Web Data Analysis:**

Web analysis has emerged as an active research field. Web analysis aims to automatically retrieve, extract, and evaluate information from Web documents and services so as to discover useful knowledge. Web content mining is the process to discover useful knowledge in Web pages, which generally involve several types of data, such as text, image, audio, video, code, metadata, and hyperlink. The research on image, audio, and video mining has recently been called multimedia analysis. Since most Web content data is unstructured text data, the research on Web data analysis mainly centers around text and hypertext. Hypertext mining involves mining semi-structured HTML files that contain hyperlinks Web structure mining involves models for discovering Web link structures. Here, the structure refers to the schematic diagrams linked in a website or among multiple websites. Models are built based on topological structures provided with hyperlinks with or without link description. Such models reveal the similarities and correlations among different websites and are used to classify website pages. Page Rank and CLEVER make full use of the models to look up related website pages. Topic oriented crawler is another successful case by utilizing the models.

- **Multimedia Data Analysis:**

Multimedia data (mainly including images, audios, and videos) have been growing at an amazing speed. multimedia data is heterogeneous and most of such data contains richer information than simple structured data and text data, Research on multimedia analysis covers many disciplines. Audio summarization can be accomplished by simply extracting the prominent

words or phrases from metadata or synthesizing a new representation. Video summarization is to interpret the most important or representative video content sequence, and it can be static or dynamic.

- **Network Data Analysis:**

Many prevailing online social networking services include Twitter, Facebook, and LinkedIn, etc. have been increasingly popular over the years. Such online social networking services generally include massive linked data and content data. include Twitter, Facebook, and LinkedIn, etc. have been increasingly popular over the years. Such online social networking services generally include massive linked data and content data. Social networking service contexts can be classified into two categories: link-based structural analysis and content-based analysis link-based structural analysis has always been committed on link prediction Link prediction is to predict the possibility of future connection between two vertices. Content-based analysis in SNS is also known as social media analysis.

- **Mobile Traffic Analysis:**

By April 2013, Android Apps had provided more than 650,000 applications, covering nearly all categories. By the end of 2012, the monthly mobile data flow had reached 885 PB As a whole, mobile data has unique characteristics, e.g., mobile sensing, moving flexibility, noise, and a large amount of redundancy. Recently, new research on mobile analysis has been started in different fields. Because of the immaturity of the research on mobile analysis, we will only introduce some recent and representative analysis applications in this section RFID labels are used to identify, locate, track, and supervise physical objects in a cost-effective manner. RFID is widely applied to inventory management and logistics. However, RFID brings about many challenges to data analysis:

(a) RFID data is very noisy and redundant

(b) RFID data is instant and streaming data with a huge volume and limited processing time.

We can track objects and monitor system status by deducing some original events through mining the semantics of RFID data, including location, cluster, and time, etc. In addition, we may design the application logic as complex events and then detect such complex events, so as to realize more advanced business applications .

Key Applications:

- **Application of Big Data in Enterprises:**

At present, big data mainly comes from and is used in enterprises, while BI and OLAP can be regarded as the predecessors of big data application In particular, in marketing, with correlation analysis of big data, enterprises can more accurately predict the behavior of consumers and mine new business models. In particular, in marketing, with correlation analysis of big data, enterprises can more accurately predict the behavior of consumers and mine new business models. On the supply chain, using big data, enterprises may conduct inventory optimization, logistic optimization, and supplier coordination, etc., In finance, the application of

big data in enterprises has been rapidly developed. For example, China Merchants Bank (CMB) utilizes data analysis to recognize that such activities as "Multi-times score accumulation" and "score exchange in shops," are effective for attracting quality customers. By analyzing customers' transaction records, potential small and micro corporate customers can be effectively identifiedData Cube of Taobao is a big data application on the Taobao platform, through which, merchants can be ware of the macroscopic industrial status of the Taobao platform, market conditions of their brands, and consumers' behaviors, etc.,

- **Application of IoT Based Big Data:**

The Internet of Things is not only an important source of big data, but also the main market of application of big data. In the Internet of Things, every object in the real world may be both the producer and consumer of data. Logistic enterprises may have profoundly experienced the application of big data to the Internet of Things. Trucks of UPS are installed with sensors, wireless adapters, and GPS, so the Headquarter can track truck positions and prevent engine failures. Smart city is a hot research area based on the application of Internet of Things data. The U.S. Miami-Dade County is a sample of smart cities. The smart city project cooperation between Miami-Dade County in Florida and IBM closely connects 35 types of key county government departments and Miami City, and helps government leaders obtain better information support in decision making for managing water resources, reducing traffic jam, and improving public safety.

- **Application of Online Social Network-Oriented Big Data:** Online SNS is a social structure constituted by social individuals and connections among individuals based on an information network. Big data of online SNS mainly comes from instant messages, online social, micro blog, and shared space, etc. Classic applications of big data of online SNS are introduced in the following, which mainly mine and analyze content information and structural information to acquire values.
    1. **Content-Based Applications:** Language and text are two most important forms of representation in SNS. Through the analysis of language and text, user preferences, emotions, interests, and demands, etc. may be revealed.
    2. **Structure-Based Applications:** On SNS with users as nodes, social relation, interest, and hobbies, etc. aggregate relations among users into a clustered structure. Such a structure with close relations among internal individuals but loose externally relations is also called a community. The community-based analysis is of vital importance to improve information propagation and for the research on interpersonal relation analysis.

- **Applications of Healthcare and Medical Big Data:**

Medical data is continuously and rapidly growing containing abundant and various information values. Big data has unlimited potential for effectively storing, processing, querying, and analyzing medical data. For example, Aetna Life Insurance Company selected 102 patients from a pool of 1,000 patients to complete an experiment in order to help predict the recovery of patients with metabolic syndrome. In an independent experiment, it scanned 600,000 laboratory test results and 180,000 claims through a series of detection test results of metabolic

syndrome of patients in three consecutive years .The goal is to manage individual health information in individual and family medical equipment.

• Collective Intelligence:

With the rapid development of wireless communication and sensor technologies, mobile phones and tablet computers have integrated more and more sensors, with increasingly stronger computing and sensing capacities. The goal is to complete large-scale and complex social sensing tasks. In crowd sensing, participants who complete complex sensing tasks do not need to have professional skills. Crowd sensing modes represented by Crowdsourcing has been successfully applied Crowdsourcing, a new approach for problem solving, takes a large number of general users as the foundation and distributes tasks in a free and voluntary way. The main idea of Crowdsourcing is to distribute tasks to general users and to complete tasks that users could not individually complete or do not anticipate to complete. The operation framework of Spatial Crowdsourcing is shown as follows. A user may request the service and resources related to a specified location. Then the mobile users who are willing to participate in the task will move to the specified location to acquire related data (such as video, audio, or pictures). Finally, the acquired data will be send to the service requester.

• Smart Grid:

Smart Grid is the next generation power grid constituted by traditional energy networks integrated with computation, communications and control for optimized generation, supply, and consumption of electric energy.

Smart Grid related big data are generated from various sources, such as

(a) power utilization habits of users

(b) phasor measurement data, which are measured by phasor measurement unit (PMU) deployed national-wide

(c) energy consumption data measured by the smart meters in the Advanced Metering Infrastructure (AMI)

(d) energy market pricing and bidding data

(e) management, control and maintenance data for the devices and equipment in the power generation, transmission and distribution networks (such as Circuit Breaker Monitors and transformers).

Smart Grid brings about the following challenges on exploiting big data.

1.Grid Planning:

By analyzing data in Smart Grid, the regions can be identified that have excessive high electrical load or power outage frequencies. Even the transmission lines with high failure possibility can be predicted. Such analytical results may contribute to grid upgrading, transformation, and maintenance, etc.

2. Interaction Between Power Generation and Power Consumption:

An ideal power grid shall balance power generation and power consumption. The traditional power grid is constructed based on a one-directional approach of

transmission-transformation-distribution-consumption, which could not adjust the generation capacity according to the demand of power consumption, thus leading to electric energy redundancy and waste. Smart electric meters are developed to enable the interaction between power consumption and power generation, and to improve power supply efficiency.

3. **Access of Intermittent Renewable Energy:**

At present, many new energy resources, such as wind energy and solar energy, are also accessed to power grids. However, since the power generation capacities of such new energy resources are closely related to climate conditions that feature randomness and intermittency, it is challenging to access them to power grids.