Scalable and Emerging Information System Technique

Techniques for Voluminous Data

Voluminous Data

Voluminous data or big data is the application of specialized techniques and technologies used to process very large sets of data. Such data are so large and complex that it becomes difficult to process using traditional data mining techniques and database management tools.

Techniques to analyze Voluminous Data:

Association Rule
- Association rule learning is a method for discovering interesting correlations between variables in large databases.
- It helps to understand closeness of products with each other so as to increase sales.
- It can be used to monitor system logs to detect intruders and malicious activity.
- It helps to extract information about visitors to websites from web server logs.

Classification Tree Analysis
- Statistical classification is the method of identifying categories that a new observation belongs to.
- It requires training set of correctly identified historical data.
- It mainly helps in assigning objects to categories and groups automatically.

Machine Learning
- Machine learning is the method of inducing human like sense to real world inside the machine.
- It provides ability to the computers to learn without being explicitly programmed.
- It helps in predictions based on known properties learned from sets of training data.

Regression Analysis
- Regression analysis is the method that involves manipulating some independent variable to see how it influences a dependent variable.
- It describes how the value of dependent variable changes when the independent variable is varied.
- It is used to understand customer satisfaction against loyalty.

Sentiment Analysis
- Sentiment analysis is the method that determine the sentiments (view) of speakers or writers with respect to a topic.
- It is used in improving services by analyzing guest comments and customers demands.

Characteristics of Voluminous Data:

Volume:
- Volume indicates the quantity of generated and stored data.
- The size of data determines the potential value and insight.
- It also determines whether it is considered to be big data or not.
- The volume of data in the world is increasing exponentially .

Variety:
- Variety indicates the different types and nature of data.
- All the data present in big data analysis may not be of same type.
- Even a single application may be generating variety of data.
- This increases complexity in big data analysis and knowledge extraction.
- The data may be web data, relational data, XML, structured data, streaming data, graph data and so on.
- For efficient extraction of information or patterns, all these variety of data must be linked together and analyzed together.

Velocity:
- Velocity indicates the speed at which data is generated and processed to meet the demands.
- The data obtained is of dynamic nature, so they must be analyzed very fast to provide efficient and effective knowledge.

Cloud Computing and their types

Cloud

- Cloud is the network basically which is present at remote location.
- It can provide services over public and private networks.

Cloud Computing

- Cloud computing is the method of manipulating, configuring and accessing the hardware and software resources remotely through online data storage, infrastructure and application.
- It provides platform independent services.

Benefits of Cloud Computing

1. One can access applications over the Internet. It reduces the necessity of installation of software in the system.
2. One can manipulate and configure applications online at any time using any devices.

3. It provides tools for online development and deployment.
4. It provides platform independent resources.
5. It operates at high efficiency with optimal utilization.
6. It provides load balancing services.

Risks of Cloud Computing

1. Cloud computing is provided by third party, this may cause risk to handover the sensitive information to cloud service providers.
2. It is difficult to switch from one cloud service provider to another.
3. The services are accessible by any one from the Internet. So, there may be compromise if necessary security system is not applied.
4. In some cases, data deletion may be insecure or incomplete.

Characteristics of Cloud Computing

1. On demand self service
2. Broad network access
3. Resource pooling
4. Rapid elasticity

Classification based on Access Type

1. Public Cloud
- It allows systems and services to be easily accessible to the general public.
- It is insecure due to openness characteristics.

2. Private Cloud
- It allows systems and services to be accessible within an organization.
- It is secure because of its private nature.

3. Community Cloud
- It allows systems and services to be accessible by a group of organizations.

4. Hybrid Cloud
- It is the mixture of public and private cloud.
- The sensitive activities are hosted using private cloud.
- The general activities are hosted using public cloud.

Classification based on Service Type

1. Anything as a Service (XaaS)
- It is the cloud service that provides services related to network, business, identity, database and strategy.

2. Infrastructure as a Service (IaaS)
- It provides infrastructure services to the users.
- The services include virtual machines, servers, storage, networks and so on.

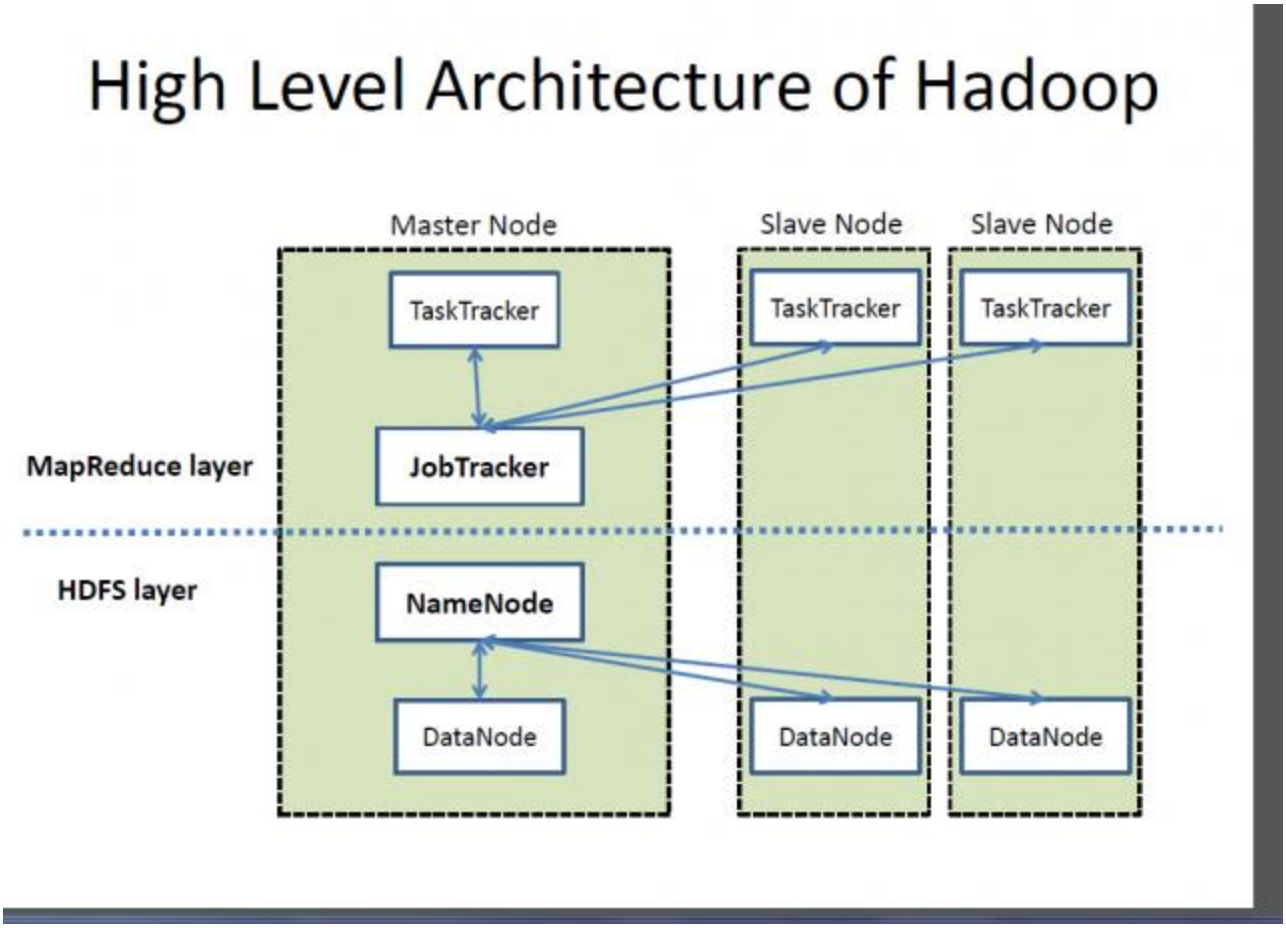
3. Platform as a Service (PaaS)
- It provides the runtime environment for applications, development and deployment tools and so on.
- It includes database, web server, deployment tools and so on.

4. Software as a Service (SaaS)
- It provides users to access software applications as a service to the end users.
- It includes CRM, email, games, virtual desktop and so on.

Map Reduce and Hadoop System

Hadoop System

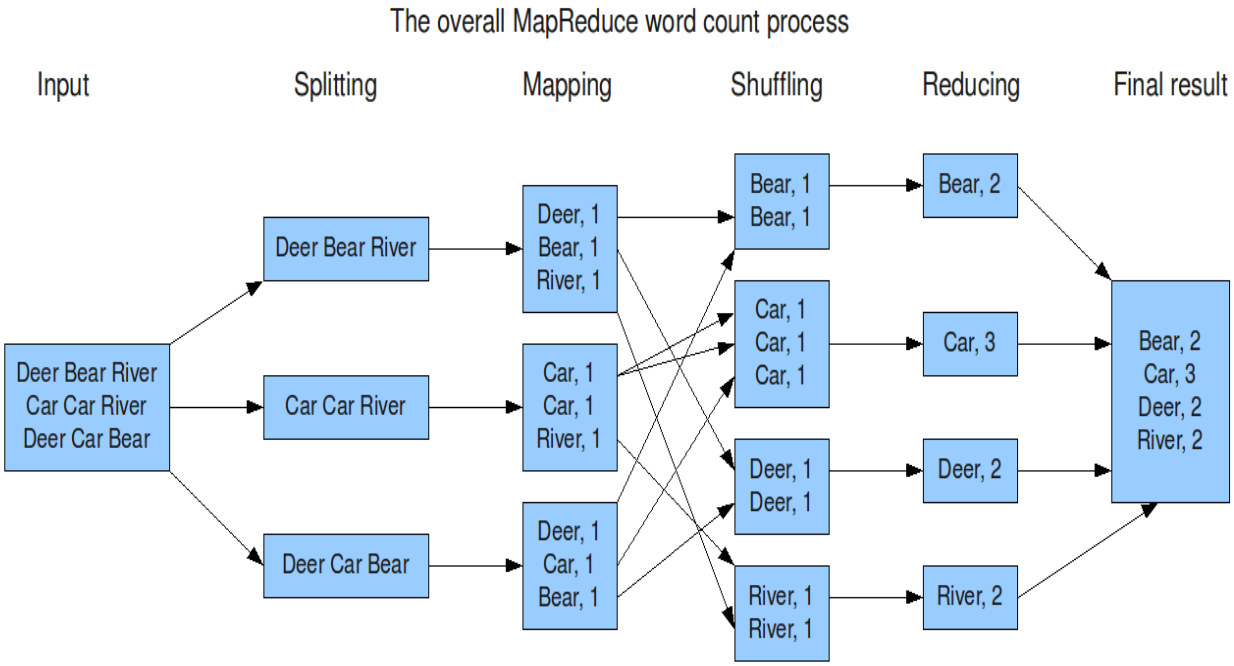
- Hadoop is a framework that allows to process and store huge data sets.
- It is a batch oriented data processing system that works by storing and tracking data across multiple machines and can scale to thousands of servers.
- It is generally used to process huge data sets that are unstructured in nature.
- The data loaded to Hadoop system is split into pieces and spreads across different servers.
- It keeps track of where the data is.
- The complex queries can be performed with faster performance as all the processors are working in parallel.
- For executing such distributed queries, it uses MapReduce.
- It can be divided into two parts: processing and storage.



MapReduce

- MapReduce is the programming model that allows user to process huge data stored in distributed Hadoop system.
- It provides platform to perform distributed and parallel processing on large data sets in a distributed environment.
- It consists of tow distinct tasks: Map and Reduce.
- Map task reads and processes a block of data to produce key-value pairs as intermediate output.
- The output of a Mapper is the input to the Reducer.
- Reduce task receives the key-value pair from multiple map jobs and then aggregates them into a single result set.
- The single result set is the final output of the system.

The working of MapReduce for word count problem is shown in figure below:



Data Management in the Cloud

Transactional Data Management

- Transactional data are those data that needs ACID property to be guaranteed.
- Such data in the cloud is not a perfect match because of following reasons:
1. Cloud provides shared-nothing architecture but transactional data can not be implemented in such architecture.
2. Since the data are replicated over large geographic distances, it is difficult to maintain the ACID properties.
3. Storing transactional data on an untrusted cloud host arise a lot of risk of data compromization.

Analytical Data Management

- Analytical data means those data that are queried up on for use in business planning, decision support and problem solving.
- The scale of such data is larger as it contains all historical data too.
- Such data are well suited to run in cloud environment due to following reasons:

1. It uses shared-nothing architecture.
2. It does not require ACID properties to be guaranteed.
3. It generally do not contain sensitive data. So, there is no risk of data compromise.