EVALUATION OF ANALYTICAL SCLABILITY

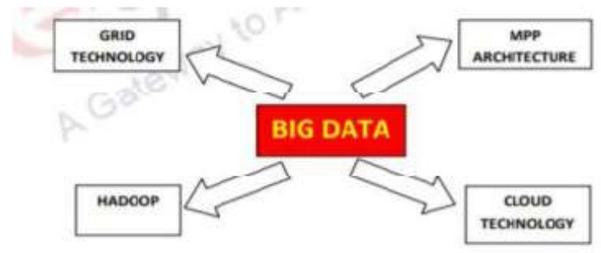


Figure 1. Big data technologies

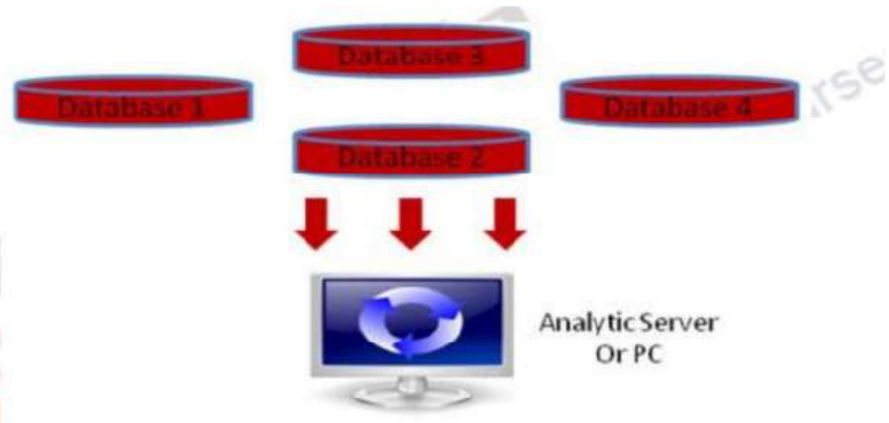


Figure 2. Traditional Analytic Architecture

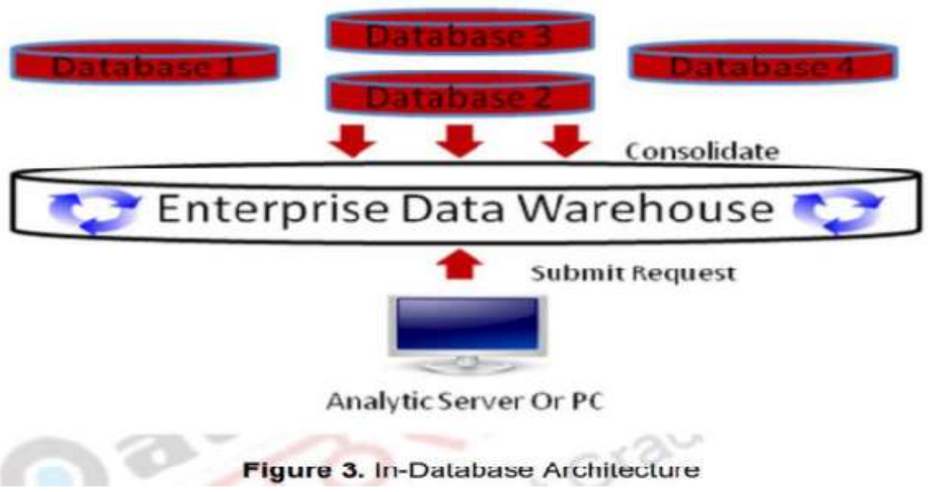
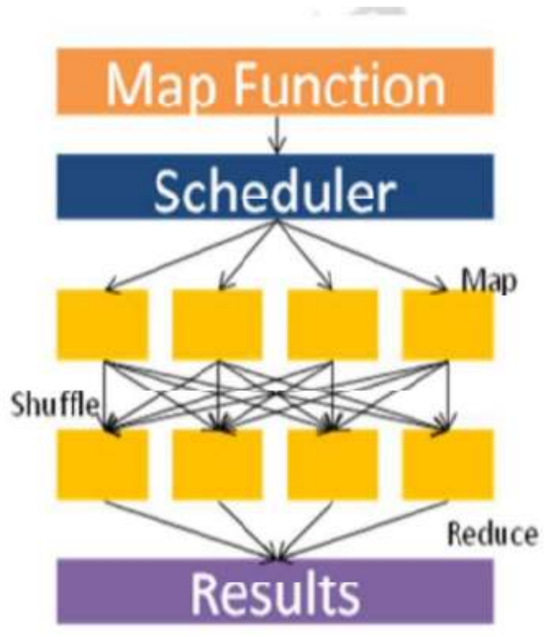


Figure 3. In-Database Architecture



2. Types of Analytics and Types of Bigdata:

i. **Types Of Analytics:**

- Descriptive Analytics
- Diagnostic Analytics
- Predictive Analytics
- Prescriptive Analytics

WEB DATA

Web content= text, image, records etc.

Web structure= hyperlinks, tags etc.

Web usage= http logs, app server logs etc.

TRADITIONAL ANALYTIC ARCHITECTURE

- Traditional analytics collects data from heterogeneous data sources and we had to pull all data together into a separate analytics environment to do analysis which can be an analytical server or a personal computer with more computing capability.
- The heavy processing occurs in the analytic environment, fig 2
- In such environments, shipping of data becomes a must, which might result in issues related with security of data and its confidentiality.

Modern in database architecture

- Data from heterogeneous sources are collected, transformed and loaded into data warehouse for final analysis by decision making.
- The processing stays in the database where data has been consolidated.
- Data is presented in aggregated form for querying.
- Queries from users are submitted to OLAP engines for execution.
- Such in db architecture are tested for their query throughput rather than transaction throughput as in traditional DB environments.
- The data in consolidated form are free from anomalies , since they are pre processed before loading into warehouses which may be used directly for analysis.
- Figure 3

Massive Parallel Processing (MPP)

MPP is the shared nothing approach of parallel computing

It is a type of computing where in the process is being done by many CPU working in parallel to execute a single program. One of the most significant difference between a Symmetric Multi Processing or SMP an MPP is that with MPP, each of the many CPUs has its own memory to assist it in preventing a possible hold up that the user may experience with using SMP when all of the CPU's attempt to access the memory simultaneously.

Features is:

- Loosely couped nodes.
- Nodes linked together by a high speed connection.
- Each node has its own memory.
- Disks are not shared, each being attached to only one node, shared nothing architectures.

Cloud Computing

- Cloud computing is the delivery of computing services over the internet.
- Examples are online file storage, social networking sites, webmail, and online business applications.
- Cloud computing model allows access to information and computer resources from anywhere that a network connection is available.
- It provides a shared pool of resources, including data storage space, networks, computer processing power, and specialized corporate and user applications.
- McKinsey and Company has indicated features of cloud :

1. Mask the underlying infrastructure from the user.
2. Be elastic to scale on demand.
3. On a pay-per-use basis.
4. National Institute of Standards and Technology(NIST)
5. On-demand self services.
6. Broad network access.
7. Resource pooling.
8. Measured service.

Public cloud

- The service and infrastructure are provided off site over the internet
- Less secured and more vulnerable than private cloud

Private cloud

- Infrastructure operated solely for a single organization
- Offer the greatest level of security and control

Grid Computing

- Grid computing is a form of distributed computing where by a “super and virtual computer” is composed of a cluster of networked, loosely coupled computers, acting in concert to perform very large tasks.
- Grid computing (Foster and Kesselman, 1999) is a growing technology that facilitates the executions of large scale resource intensive applications on geographically distributed computing resources.
- Facilitates flexible, secure, coordinated large scale resource sharing among dynamic collections of individual, institution and resources.
- Distributed or grid computing in general is a special type of parallel computing that relies on complete computers connected to a network by a conventional network interface producing commodity hardware, compared to the lower efficiency of designing and constructing a small number of custom super computers.
- Disadvantages:
  The various processors and local storage area do not have high speed connections.

Hadoop

Apache Hadoop is an open source software framework for storage and large scale processing of data sets on clusters of commodity hardware.

2 main building blocks inside this runtime environment are : MapReduce and HDFS

MapReduce

- Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data (multi terabytes data sets) in parallel on large clusters (1000 of node) of commodity hardware in a reliable, fault tolerant manner.
- A MapReduce job usually splits the input data set into independent chunks which are processed by the map tasks in a completely parallel manner.
- The framework sorts the outputs of the map, which are then input to the reduce tasks.
- Typically, both the input and output of job are stored in a file system.
- The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks.

HDFS (Hadoop Distributed File System)

- It is one of the core components of the Hadoop framework and is responsible for the storage aspect.
- Unlike the usual storage available on our computers, HDFS is a distributed File System and parts of a single large file can be stored on different nodes across the cluster.
- HDFS is a distributed, reliable and scalable file system.