Assignment -1

Lambda Architecture

> Explain the factors leads to Big Data, List and explain Major sources of Big Data.

→ 1) Technological Advances:- The exponential growth in computing power & storage capabilities led to the growth of Big Data.

2) Digitalisation:- Business operations, Social interaction research all more has led to the increase in the amount of Big Data

3) Internet & Connectivity:- The wide spread availability of Internet and the proliferation of connected devices

4) Social Media & User Generated Content:- The rise of Social media platforms, blogging and other online platforms.

5) Sensor Networks:- Advances in sensor technologies have enabled the collection of data from various sources.

6) Data Storage & Cloud Computing:-
The adoption of cloud Computing platforms has made it easier & more cost effective

7) Data Analytics tools:-
The Development of Data Analytic tools have made the organisations to store & process large volumes of Data at given time.

huge amount of data storage. Then without loosing the quality is challenge.

→ Fault tolerance:- Computing Fault tolerance is extremely hard.

→ Scalability:- Big Data projects can grow and evolve rapidly. The scalability issue of big Data has lead toward cloud computing.

4) Discuss the problem faced by traditional Database Systems.

→ *Scalability:-

Traditional Database Systems often struggle to handle massive volumes of Data generated by modern applications.

*Performance:-

As Data size and query complexity increase traditional Database may experience performance is compromised. Complex queries, joins and aggregation each can lead to slow response times, making realtime analytics.

* Data Model Rigidity:-

Traditional Database are typically designed with after fixed System. makes it challenging to adapt to change data Requirements.

* Data Consistency:-

Maintaining data Consistency in distributed requirement can be difficult

Traditional Database use Acid transaction which can be limit in distributed System

Lack of Support for unstructured Data:-
The traditional Database are primarily Designed for structured data & may not handle unstructured or semi structured data types.

Security Concerns:-
Data security is critical and traditional databases may face vulnerabilities if not configured and Managed properly.

5) Discuss The required properties of Big Data Systems.

→ - Robustness & fault tolerance:-
Systems need to behave correctly despite machine goes down randomly The complex semantics of consistency in distributed database.

* low latency (read) & update :- The vast majority of applications require reads to be satisfied with very low latency.

* Scalability:-
Scalability is the ability to maintain performance in the face of increasing data or load by adding resources to the System.

Date :
Page :

* Extensibility:-

Extensible Systems allow functionality to be able to be added with a minimal developent cost.

* Adhoc queries:-

Nearly every large dataset has unwanted value within it. Been able to mine a dataset arbitarily gives opportunities optimize and new applications

* Minimal Maintance:-

Maintance is the work required to keep the System Running Smoothly in all environments.

Q) Explain Different layer of Lambda Architecture:-

→ There are 3 layers.

1) Batch layer:-

The layer is responsible for proceeding large volume of Data in a fault tolerant & scalable manner. It stores all the data in its raw form and precomputes batch views on top of it.

2) Speed layer:-

This layer only responsible for processing real time data streams. It stores only the most relevant and recent

data & computs real time views on top of the existing ones.

3) Servis Layer:
This layer is responsible for servis queris on the data. It merges the batch view & real time views to provd a complte view of the data.

7) Differentiate between re-computation Algorithm and incremental algorithm.

→ Incremental Algorithm:
Updts a batch view by usn the new Data and the current state of batch view to perform an updte. This approch is more efficient in terms of resorc usag because it only needs to proces the new data & updte the relevant parts of the batch view.

Re-computation Algorithm:
Updts a batch view by lookin at the entire master dataset, includs the new data. This approach is more genral and can handle any type of updte, but it is less efficint in terms of resorc.

8) List & Explain Reqwirmts & Responsibils of batch layer.

→ Reqwirmts:
→ Proces large volume of Data in a fault tauleran and scalable manner.

→ store all data in its raw form
→ Precompute batch views on top of the raw data
→ Batch views should be immutable & recomputed periodically to reflect the latest Data.

Responsibilities:
* Ingest raw data from various sources
* Validate and clean the Data
* Store the Data in a Distributed file System or Database
* Precompute Batch views on top of the raw Data using Batch processing Algorithms.
* Handle updates to the Data by recomputing the batch views.
* Ensure consistency across the batch views by merging them with the real-time views in the serving layer.

9) Explain the responsibilities of Serving layer:
* low latency access:
The serving layer must provide low latency access to the precomputed batch views.

* Scalability :- The serving layer must be scalable enough to handle the large volume of Data.

Date:
Page:

* Random Reads: The serving layer must be supports random reads with indexes provides direct access to small portion of the views.

* Batch writable:-
The serving layer must be batch writable, means that it can be updated in bulk.

* Error tolerance
The serving layer must be able to correct the errors, This is achieved by recompute the serving layer views.

b) With Example show latency and high throughput can be achieved in serving layer

→ To achieve low latency the Data must be precomputed and indexed in a way that allows for fast queries, This can be achieved by denormalizing the data & tailoring the views to the specific queries they serve.

Let's consider the example of large dataset of customer transactions for an e-commerce website. We want to be able to quickly query the total revenue for a specific product category over a certain time period. To achieve this we can denormalize the data by creating precomputed batch views that aggregate the revenue for each product category over

time. The serving layer can then index the batch views and provide an interface for querying the data. By using a distributed database such as Cassandra we can achieve the given requirements.

11) List the SC requirements & responsibilities of speed layer.

→

Requirements:-
→ low latency updates
→ Incremental Computation
→ Random writes.
→ Fault tolerance.
→ Scalability.

Responsibilities:-
→ Ingest Real time Data.
→ Process Real time Data
→ Store Real time views.
→ Merge batch & real time views.
→ Expiring Data.

12) Differentiate between Batch & Speed layer.

Batch Layer.	Speed Layer.
* The Batch layer Responsible for precomputing batch views from the master Dataset.	* The Speed layer is responsible for providing updates to the real time views.

* Uses Batch Computation to process the entire Dataset and generate batch view	* Incremental computation to process only the new data to update the real time views.
* Designed to handle large volumes of data & is optimized for throughput rather than latency	* Designed to handle real time data is optimized for low latency updates rather than throughput.
* Fault tolerant & can recover from failures by recomputing the batch views	* Recover from failures by replaying the new Data updates to the views.