W / T / F / S / SU /

Big data modeling and management

Date

Week 4 :-

quiz

Data formats and streaming.

1. What is true bt^n data modeling and formatting of the data?

→ The data doesn't necessarily need to be formatted in way that represents the data model. Just so long as it can be extrapolated.

2. What is streaming?

→ Utilizing real time data to compute and change the state of an app. continuously.

video 1:-

Data model vs Data format

serialized representation of data and data model.

CSV doesn't mean Relational.

video 2:-

What is data stream.

Big data challenges ↓ velocity

→ key characteristics of a data stream

→ requirements of streaming data systems.

→ Recognize the data streams you use in your life.

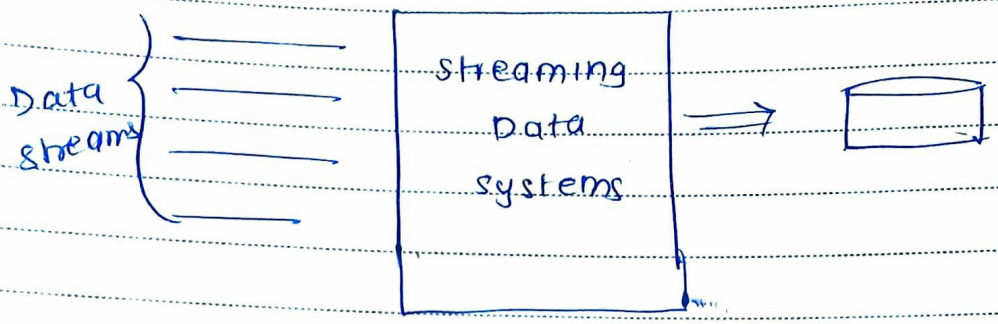
⇒ social media :- sales Trends + sales Distribution ↓ Data-driven Marketing

monitoring and Fault Detection.

fn. A possibly unbounded sequence of data record ↓ may or may not be related

→ Time-stapped

→ Geo-tagged



Streaming data systems → To overcome conventional data system

- → Manage one record or small time window
- → near-real-time
- → independent computations.
- → Non-interactive.

} Many challenges.

Some streaming data systems :-

- → amazon kinesis
- → Apache storm
- → Flink
- → spark streaming

M / T / W / T / F / S / SU/ Date

video 3:-

Why is streaming Data different

→ compare and contrast "data-in-motion" and "data-at-rest"

→ differentiate bt^n streaming and batch data processing

→ List management and processing challenges for streaming data

Data-at-Rest :- Mostly static data from one or more sources

- collected prior to analysis.

Data-in-Motion :- Analyzed as it is generated

Ex. sensor data from self-driving vehicles.

- Stream processing.

i) static/Batch processing :- size determines time and space.

ii) streaming processing :- Unbounded sizes but finite time and space

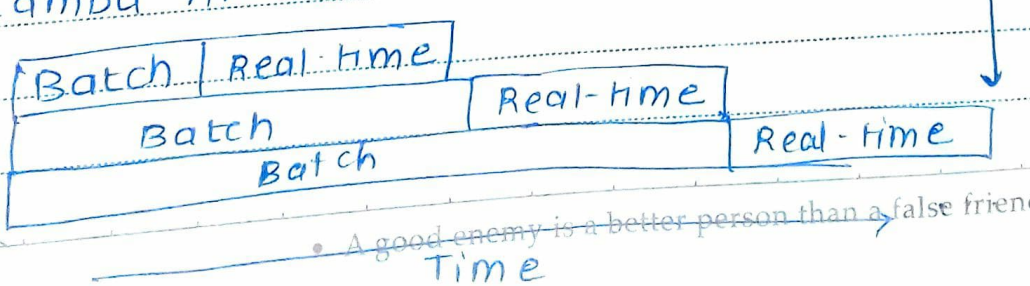
Streaming Data management and Processing

→ compute one data element or a small window of data elements at a time.

→ Relatively fast and simple computations.

→ No interaction with the data source

Lamba Architecture



Priority

• A good enemy is a better person than a false friend. •

streaming Data changes over size + Frequency

M/T/W/T/F/S/SU/ Date

changes can be periodic or sporadic

i) Periodic : evenings, weekends

ii) Sporadic : major events (Breaking News)

Example of extreme changes :-

World Records for Tweets

- size → unbounded
- size and frequency → unpredictable
- processing → Fast and simple

video 4: Understanding Data Lakes

→ Describe how data lakes enable batch processing of streaming data

→ schema-on-write vs schema-on-read

→ Organizes data stream, data lakes and data warehou on a spectrum of big data management ad storage

What is Data Lake ? → click-streams, social-media, sensor data, sales transaction, Geo location

How Data Lake Works :-

Load data from source

↓

store raw data

↓

Add data model on read

M / T / W / T / F / S / SU /

Date .....

Schema - On - Read Approach

schema - on - Read

⇓

Data Warehouse

⇓

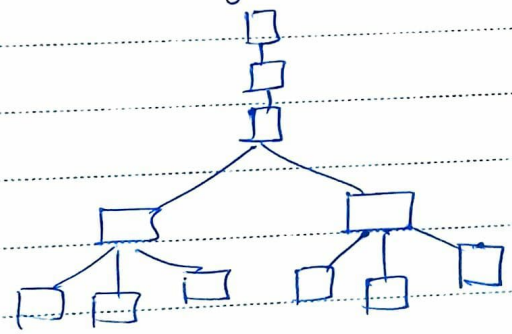
schema - on - write

⇓

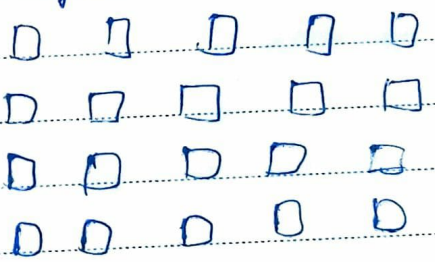
Transform all structu before load

Data lake vs. Data Warehouse

Data Warehouse:
Hierarchical File system



Data lake:
object storage



Data Lake object storage

Application

API