SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

An Autonomous Institution

Accredited by NAAC – UGC with 'A' Grade

Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

COURSE NAME : 19CS508 - BIG DATA ANALYTICS

III YEAR / V SEMESTER

Unit 4- STREAM MEMORY

Topic 1 : Introduction to Streams Concepts and Stream Data Model and Architecture

Streams Concepts

- Data stream refers to the continuous flow of data generated by various sources in real-time.
- It plays a crucial role in modern technology, enabling applications to process and analyze information as it arrives, leading to timely insights and actions.
- A data stream is an existing, continuous, ordered (implicitly by entrance time or explicitly by timestamp) chain of items. It is unfeasible to control the order in which units arrive, nor it is feasible to locally capture stream in its entirety.

  It is enormous volumes of data, items arrive at a high rate.

Types of Data Streams

- **Data stream –**
- A data stream is a(possibly unchained) sequence of tuples. Each tuple comprised of a set of attributes, similar to a row in a database table.
- **Transactional data stream –**
- It is a log interconnection between entities
- Credit card – purchases by consumers from producer
- Telecommunications – phone calls by callers to the dialed parties
- Web – accesses by clients of information at servers
- **Measurement data streams –**
- Sensor Networks – a physical natural phenomenon, road traffic
- IP Network – traffic at router interfaces
- Earth climate – temperature, humidity level at weather stations

Characteristics of Data Streams

1.Large volumes of continuous data, possibly infinite.

2.Steady changing and requires a fast, real-time response.

3.Data stream captures nicely our data processing needs of today.

4.Random access is expensive and a single scan algorithm

5.Store only the summary of the data seen so far.

6.Maximum stream data are at a pretty low level or multidimensional in creation, needs multilevel and multidimensional treatment.

Applications of Streams

1.Fraud perception

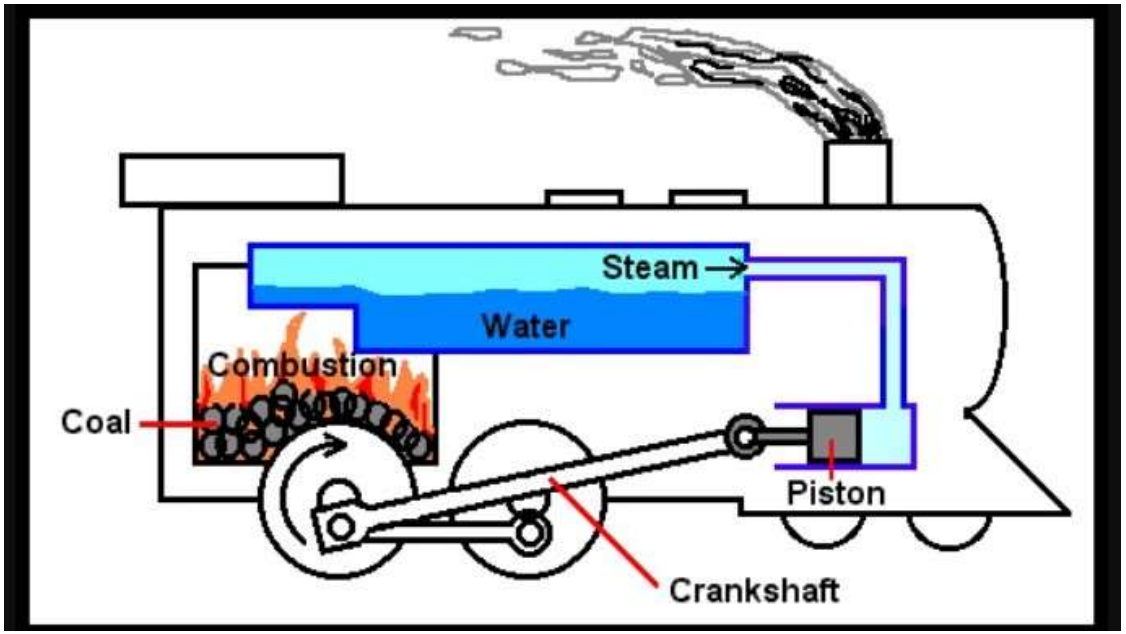
2.Real-time goods dealing

3.Consumer enterprise

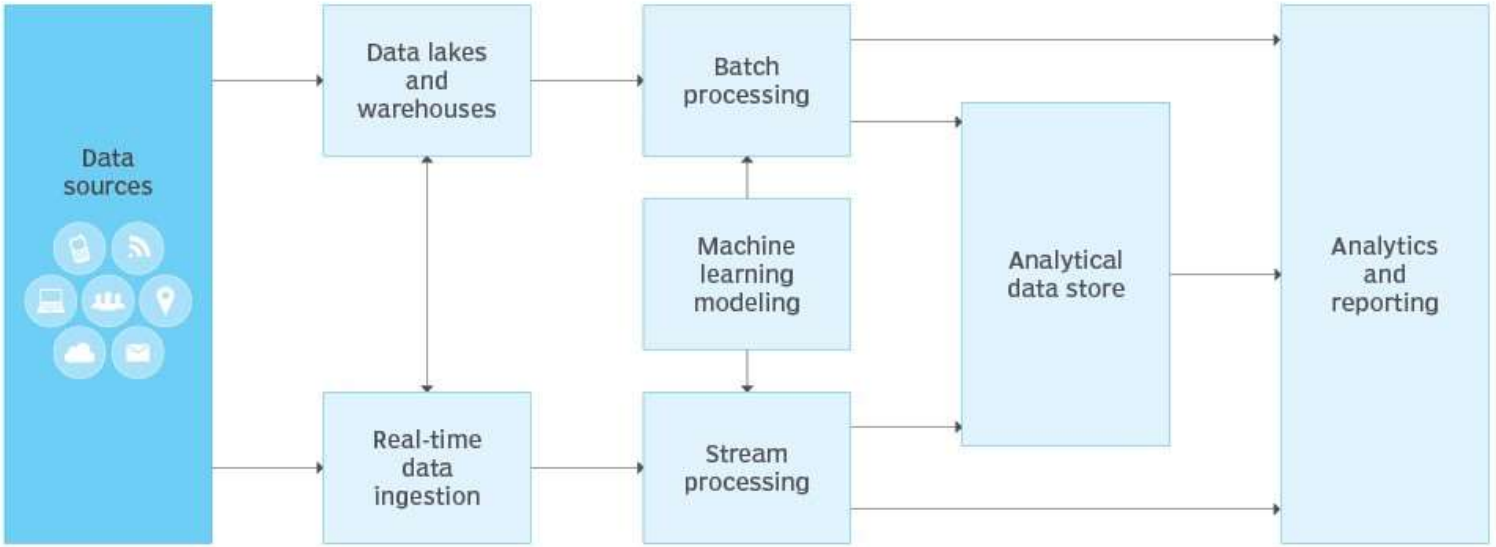
4.Observing and describing on inside IT systems

What are the key components of a Stream Data Model and Architecture ?

- **A data streaming architecture typically consists of the following key components:**
- Data Sources
- Stream Ingestion
- Stream Storage
- Stream Processing Engine
- Data Analytics
- Data Sink / Destination
- Visualization and Reporting



Stream Data Model and Architecture



What are the tools for data streaming?

- Key tools for data streaming include:

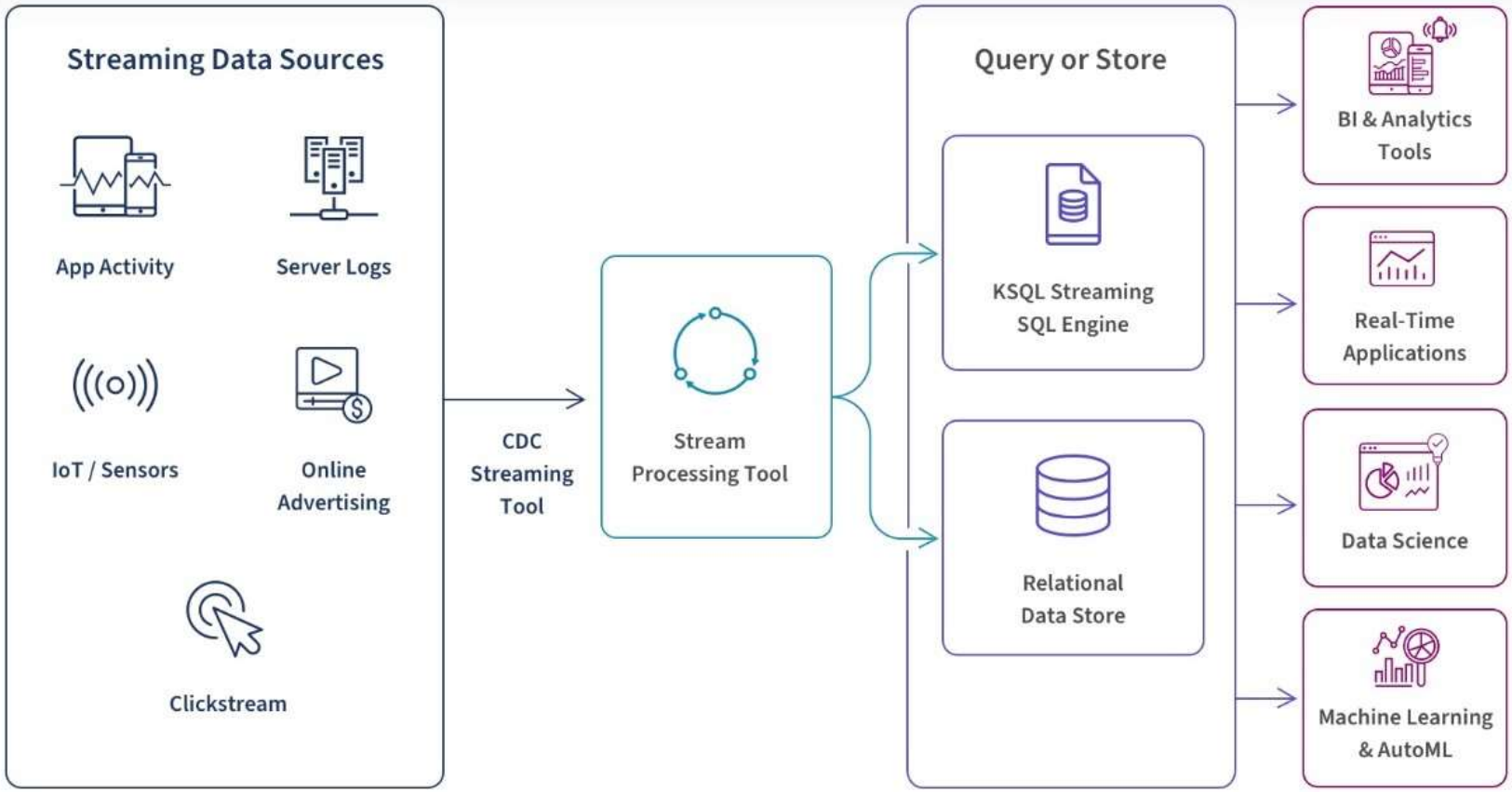
•**Message Brokers**: Apache Kafka, Amazon Kinesis.

•**ETL Tools**: Apache Flink, Apache Storm, Spark Streaming.

•**Data Storage**: Amazon S3, Google BigQuery, Apache Hadoop.

•**Analytics Engines**: Elasticsearch, Apache Druid.

Stream Data Model and Architecture - I



What are the challenges in working with streaming data?

Streaming data architecture requires special considerations due to the nature and volume of data.

- **Availability**
- Streaming data applications require consistency, low latency, and high availability. Consumers are constantly taking new data from the stream to process it. Delays from the producer could back up the system and cause errors.
- **Scalability**
- Raw data streams can surge rapidly and unexpectedly. For example, social media posts spike during a big sporting event. Therefore, the system should prioritize proper data sequencing, availability, and consistency—even during peak loads.
- **Durability**
- Because of the time sensitivity of data, the stream processing system has to be fault tolerant. Otherwise, the data will be lost forever in an outage or failure.

Details	Batch processing	Stream processing
Data scope	Queries or processing over all or most of the data in the dataset.	Queries or processing over data within a rolling time window, or on just the most recent data record.
Data size	Large batches of data.	Individual records or micro batches consisting of a few records.
Performance	Latencies in minutes to hours.	Requires latency in the order of seconds or milliseconds.
Analysis	Complex analytics.	Simple response functions, aggregates, and rolling metrics.

Activity

Advantages of Streams

- This data is helpful in upgrading sales
- Help in recognizing the fallacy
- Helps in minimizing costs
- It provides details to react swiftly to risk

- **Disadvantages of Streams**
- Lack of security of data in the cloud
- Hold cloud donor subordination
- Off-premises warehouse of details introduces the probable for disconnection

Assessment 1

1. List out the advantages of Stream processing
   a)_____
   b)_____
   c)_____
   d)_____

2. Identify the disadvantages of Stream processing
   a)_____
   b)_____
   c)_____
   d)_____

REFERENCES

1. Anand Rajaraman and Jeffrey David Ullman, "Mining of Massive Datasets", Cambridge University Press, 2012.
2. David Loshin, "Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph", Morgan Kaufmann/Elsevier Publishers, 2013
3. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data", Wiley publishers, 2015.

THANK YOU