Q. Bank (BDA):

Q. What is Data Streaming and explain with example.

→ Data streaming in the context of BDA refers to the continuous, real-time processing and analysis of large volumes of data as they are generated.

Unlike traditional batch processing, where data is collected, stored and then processed at a later time, data streaming processes data on-the-fly as it arrives, enabling immediate insights and responses.

Example:

↳ Data Streaming in Sentiment Analysis on Social Media.

where posts and comments about a brand are monitored in real-time.

Using technologies like Apache Kafka for data ingestion and Apache Flink or Spark Streaming for data processing, companies can analyse sentiments through natural language processing to measure public opinion.

PAGE No.
DATE / /

Q. Explain Content-based Recommendations

→ Content-based recommendation systems are a type of recommendation system that suggest items to users based on the features of the items and a profile of the user's preferences.

Content-based recommendation are widely used in various domains, such as recommending movies, books, news articles and more, tailored to the individual's explicit preferences.

Advantages:

- Personalisation:
  Recommendations are tailored to the individual's specific preferences, based on the content they have engaged with.

- No Cold Start for item:
  Since recommendations are based on item features, new items can be recommended as soon as they are available.

Challenges:

- User-Profile Initialization → New User = No History
  that's why as a beginner user needs to go through direct queries about preferences.

- Over-specialization → User ko same-same content repeat ho sakta jiski wajah se user nayi diversity discover nahi kar payega.

Q. Explain Relational Operation using Map Reduce.

→ MapReduce is a programming model designed to process large volumes of data in a distributed computing environment.

It consists of two main phases: the Map Phase & the Reduce Phase.

Despite being primarily designed for simple aggregation tasks, such as counting occurence of words in a large set of documents, MapReduce can also be adopted to perform more complex operations such as those found in relational databases.

Relational Operations in MapReduce:

I] Filtering (Selection):

Imagine if you have a long list of people with details
III] Aggregation like age & you only want to keep the details of people over 18.
IV] Joining Table In the map step, you go through each record and pick out only those that say "age > 18".
V] Group By. In the reduce step, you just collect all these filtered records together.

II] Projection:

Think of this as choosing only the columns you are interested infrom a table.

If you have a table with columns like Name, Age & City, but you only care about Names, in the Map step, you pick just Name from each record.

The Reduce step might not even be needed unless you want to do something extra

Q. How Bloom filter is useful for big data analytics. Explain with one example.

→ A Bloom filter is a space-efficient probabilistic data structure that is used to test whether an element is a member of a set.

It's particularly useful in situations where ~~saying~~ saving memory is crucial and a small probability of error can be tolerated.

For Big data analytics, where datasets can be extraordinarily large and checking each element against a dataset or database can be time-consuming & resource-intensive, Bloom filters offer a fast & efficient way to reduce the need for expensive lookups or checks.

Advantages :- Space Efficient
Time Efficient

Example : Web Crawling

Let's say you are building a web crawler to ~~&~~ visit different web pages.
You want to avoid revisiting pages you have already seen to save time and resources.
A Bloom filter can quickly tell you if you have already visited a webpage without needing to keep a massive list of every webpage you've ever seen.

PAGE No.

DATE / /

Q. Explore the clustering application

→ Clustering is a fundamental techique in data analysis and machine learning, used to group similar data-points together based on their characteristics or features.

1] Customer Segmentation:

- E-Commerce → Clustering helps businesses understand customer behaviour by grouping customers with similar purchasing patterns. (Unko Ads dikhate hai same!)
- Retail → Retailers use clustering to segment customers based on demographics, purchase history or shopping preferences.

2] Image Segmentation:

- Medical Imaging → MRI, CT Scan me clustering use kiya jaata hai to identify and analyse specific regions of interest.
- Satellite Imagery → Clustering helps identify and classify objects or features in satellite images such as land cover types, vegetation.

3] Recommendation Clustering:

- Content based Filtering → Q2.