Anay Shah
60004210073
TE Comps C12

5 Incomplete

BDI Assignment-I

Q] Facebook is a great example of bigdata, demonstrating the characteristics of volume, velocity, variety and value in various ways:

i) Volume- Facebook generates an enormous amount of data daily due to its large user base. Example of high volume data include posts, comments, likes, shares, messages, photos, videos and other user interactions on the platform. Each piece of ~~conten~~ content and interaction contributes to the overall volume, creating a massive pool of data that Facebook collects and analyzes.

ii) Velocity - Facebook processes data at an incredibly high speed to provide real-time updates and responses to user actions. When users post or interact with content, Facebook's algorithms process these actions instantly updating news feeds, sending notifications, and providing personalized recommendations.

iii) Variety - Facebook deals with a wide variety of data types, ranging from text-based posts and comments to multimedia content like photos and videos. This variety of data allows Facebook to create detailed user profiles and target advertisement effectively based on users interests & behaviour.

iv) Value - Facebook leverages big data analytics to extract value from the vast amount of data it collects. By analyzing user behaviour Facebook can improve its platform's usability, enhance user engagement and personalize content recommendations.

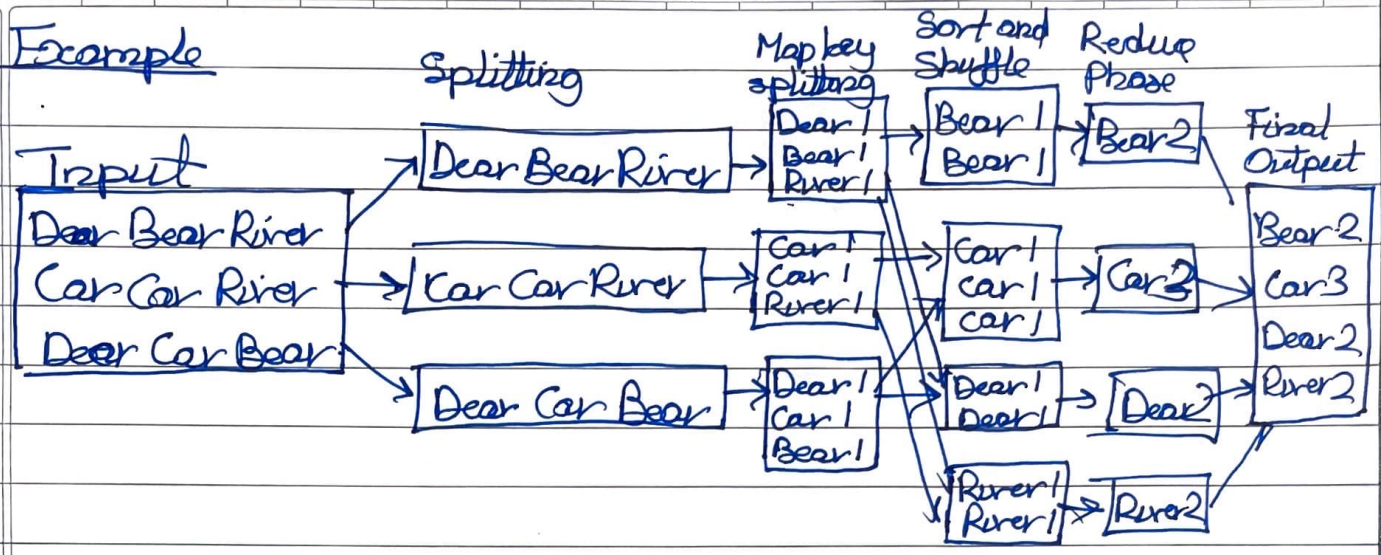
v) Veracity - Veracity is crucial for Facebook to maintain trust and credibility among its users, especially amidst the proliferation of misinformation. Through content modification, fact-checking partnerships and algorithmic detection, Facebook strives to ensure the accuracy and reliability of the content shared on its platform.

Q2] MapReduce is a programming model and processing technique used to handle large-scale data processing tasks in a distributed computing environment, typically on a cluster of computers. The concept of MapReduce is inspired by functional programming principles and is particularly effective for tasks like word-frequency counting.

MapReduce process can be broken down into two main ~~process~~ phases:

i) Map Phase - In the map phase, the input data is divided into smaller chunks and each chunk is processed independently by multiple worker nodes. For word frequency counting each worker node reads a portion of the input text and tokenizes it into individual words. The worker nodes then emit key-value pairs, where the key is a word and the value is the number '1' indicating the occurrence of that word.

ii) Reduce Phase - In the reduce phase the intermediate key-value pairs generated by the Map ~~h~~phase are shuffled and sorted based on their keys. Each unique word key is then passed to a reducer node, which aggregates the values associated with that key.



Q3] Table :- HBase organises data into tables. Table names are strings & composed of characters that are safe for use in a file system path.

Row :- Within a table, data is stored according to its row. Rows are identified uniquely by their row key.

Column Family :- Data within a column family is addressed via its column qualifier or simply column. Column qualifiers need not be specified in advance. Need not be consistent between rows.

Cell :- A combination of row key, column family & column qualifier uniquely.

Timestamp :- Values within a cell are versioned. Versions are identified by their version number, which by default is the timestamp of when the cell was written. Default is 3.