BDI Assignment 1

DATE:

05/05

26/09/24

Kreena Shah
60004210243
C'32

(1) Facebook exemplify the characteristics of Big data in the following manner :

(a) Volume

Facebook generates an enormous volume of data daily, with over 2.7 billion monthly active users generating posts, comments , shares.

(b) Velocity

Date on facebook streams in at a rapid pace, with users constantly uploading photos, videos , status updates as well as engaging with content

This requires real time processing to analyze & respond to user activity effectively.

(c) Variety

Facebook data comes in various forms including text, image , videos , links , interaction.

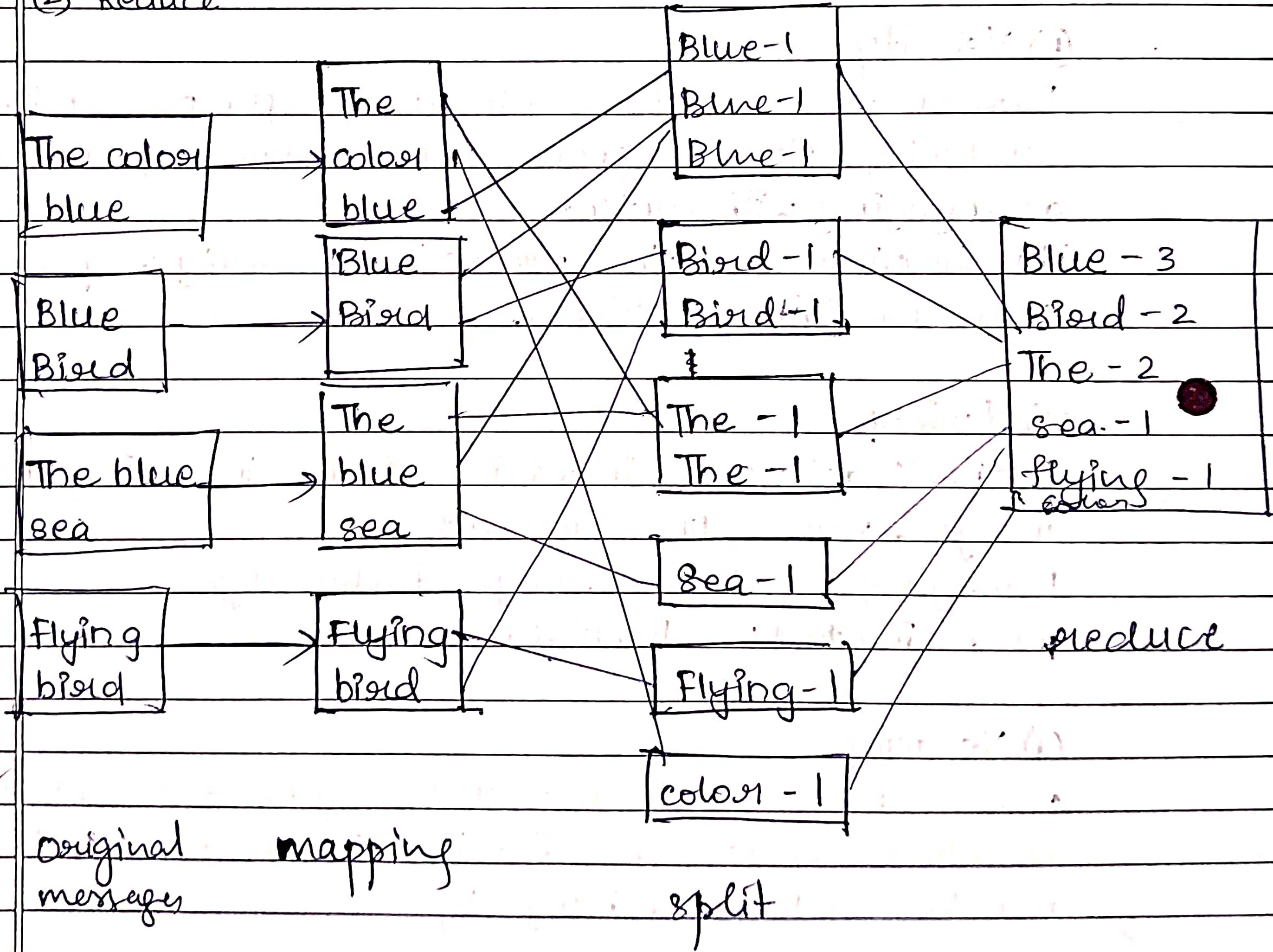
Managing this diverse range of data types requires data processing tools & techniques.

(d) Veracity

With such a vast amount of user generated content, ensuring accuracy & reliability of data is crucial for maintaining trust & integrity

(e) Value
Facebook extracts value from its data by analyzing user behaviour, preferences, & interactions to personalize user experiences

(2) Map Reduce is a programming model & processing techniques used to process & generate large dataset in parallel across distributed computeding clusters. It consits of 2 main phases
(1) Map
(2) Reduce



(3) HBase is a distributed, column oriented database built on top of Hadoop Distributed File System (HDFS) It's schema design includes concept such as

(a) Tables
HBase organizes data into tables

(b) Row Keys
Each row has unique key, used for data retrieval

(c) Column Families
Columns are grouped into column families

(d) Columns
Columns in HBase are not predefined.

BDI Assignment 2

DATE:

05/05

Kreena Shah
60004210243
C'32

(1)

```
from pys$park.sql import SparkSession

spark = SparkSession.builder \
        .appName("sparkQL CRUD operation")
        .getOrCreate()

df = spark.createDataFrame([
        (1, 'Alice', 30),
        (2, 'Bob', 20),
     ])

df.show()

new_row = [(4, 'David', 40)]

df = df.union(spark.createDataFrame(new_row,
      ["id", "name", "age"]))

df.select("name", "age").show()

df = df.withColumn("age", df["age"]+1)

df = df.filter(df.id != 2)

df.show()

spark.stop()
```

(2) Industry Use Cases

(a) Content Management Systems
Storing & managing dynamic content for websites & blogs

(b) Real Time Analytics:
Analysing user behaviour & interactions for personalized recommendations

(c) IoT
Storing sensor data & telemetry information for monitoring & analysis

(d) Mobile Applications
Serving as backend database for mobile apps with offline capabilities

(e) E-commerce Platforms
Managing product catalogs, customer profiles & order data.

(3) Industry Use Cases of Apache Kafka

(a) Real time Stream Processing
Processing & analyzing streaming data from various sources for insights.

DATE:

(b) Log Aggregation
Consolidating log data from distributed systems for monitoring & troubleshooting

(c) Event sourcing
Capturing & storing event data to maintain a full history of changes