Experiment No:

SAP ID: 60004200107

Q1. Consider any 1 case study of big data solution. State its characteristics in terms of 10 Vs of Bi

JP Morgan can use big data in several ways. The objectives are

- optimizing the sales of foreclosed properties
- Develop new marketing initiatives
- Credit assessment.
- Identify possible opportunities to make money

(i) Volume

Hadoop is designed to handle large volumes of data and JPMC uses Hadoop to store and process vast amounts of financial data, including market data and customer data.

(ii) Velocity

Hadoop has several tools such as Apache Storm Spark that can perform real time data processing and analytics

(iii) Variety

Hadoop is a flexible platform that can handle various types of data. It is utilised by JPMC to store and interpret data from several sources

(iv) Veracity

Financial data used by JPMC must be precise and can use Apache Nifi and Apache Atlas

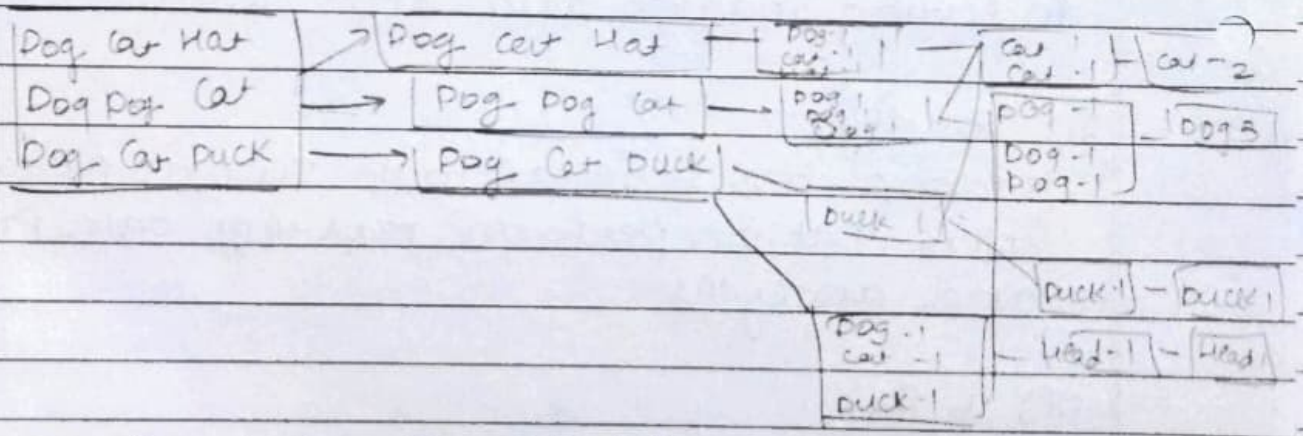
(v) Value

JPMC may get insight into consumer behavior, market trends and potential hazards by keeping and processing vast volumes of ~~data~~ data.

Q2 Explain concept of MapReduce to find frequent words

MapReduce is maploop framework and programming model for processing big data using automatic parallelization & distribution.

The output of map tasks is used as input in reduce tasks and the data is shuffled and reduced.



The input data is divided into multiple segments, the processed. In parallel to reduce processing. Data will be divided into three input splits so that

work can be distributed over map tasks
- The mapper counts the number of times each word occurs from input split the form of key value pairs where the key is the word and value is frequency.
- The shuffler phase in which the values are grouped by keys in the form of key value pairs.
- Next reducer is used for pairs with same key.
- Final output is displayed as frequency of words.

Q3 Advance Indexing in HBASE

- In HBASE the row key provides the same data retrieval benefits as a primary index. So when you create a secondary index, use elements that are different from row key.
- Secondary indexes allow you to have a secondary way to read an HBASE table. They provide a way to efficiently access records by means of some piece of information other than primary key.
- They require additional cluster space and processing because the act of creating a secondary index requires both space and processing.
- A method of index maintenance called diff index can help. Eg. SQL create secondary index for Hbase, maintain those indexes are speed up queries.