1) What is VMware?

→. VMware is a virtualization and cloud computing software.

↳ VMware software allows you to run a virtual computer on your physical computer.

↳ It increases the efficiency of computer system.

2) What is iso file related to OS?

→ ISO file are like digital copy of a whole operating system CD or DVD.

↳ An ISO file often represents a complete image of an operating system installation disk.

[ISO files support archiving of file folders such as system and installation files of operating system, software program, file folders, movies, videos, or even computer games]

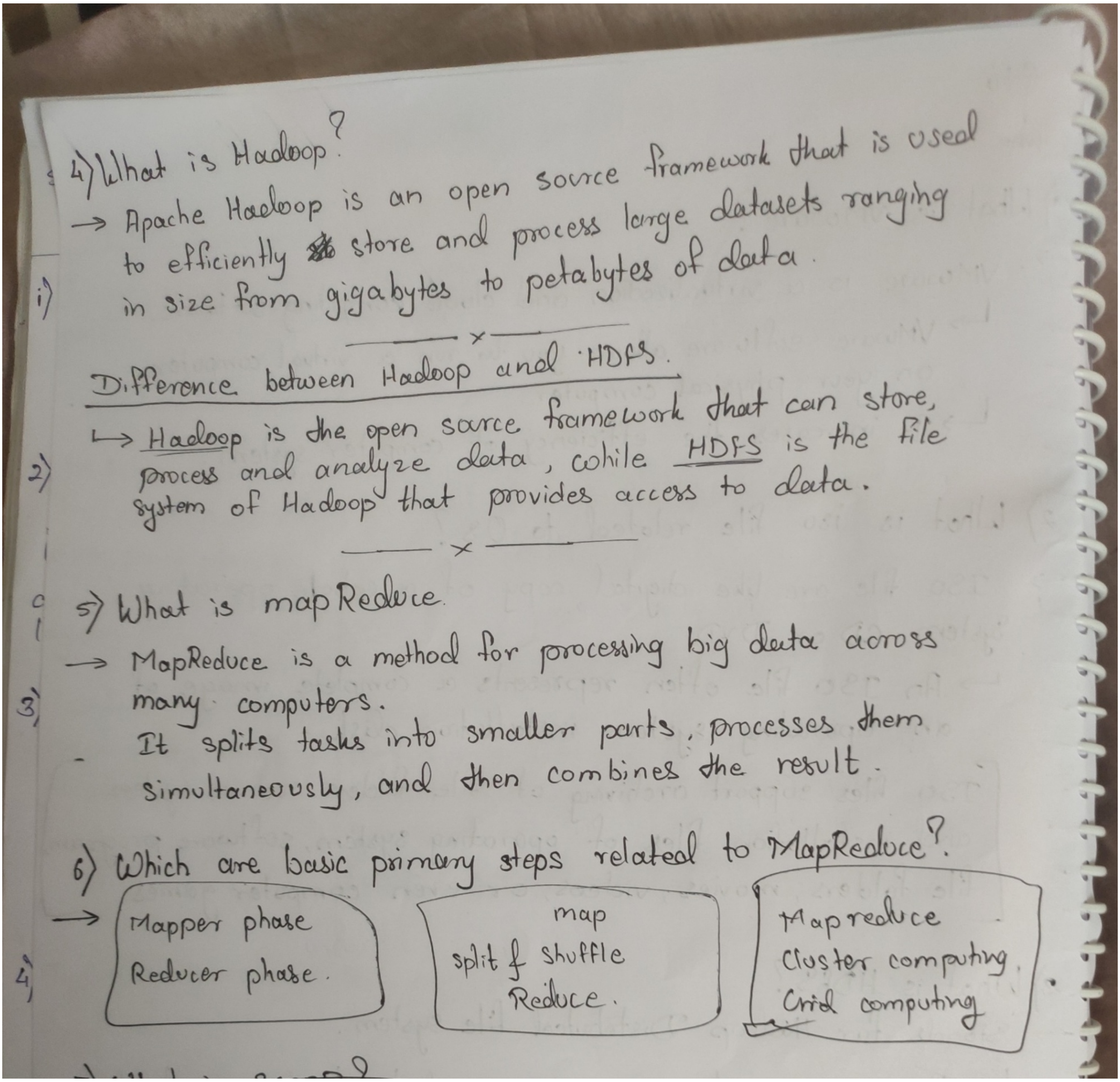
3) What is HDFS?

→ Stands for Hadoop Distributed file system.

↳ The primary storage system used by hadoop applications.

↳ HDFS operates as a distributed file system designed to run on commodity hardware

HDFS is a module of Hadoop



4) What is Hadoop?

→ Apache Hadoop is an open source framework that is used to efficiently ~~[illegible]~~ store and process large datasets ranging in size from gigabytes to petabytes of data.

—— × ——

Difference between Hadoop and HDFS.

↳ Hadoop is the open source framework that can store, process and analyze data, while HDFS is the file system of Hadoop that provides access to data.

—— × ——

5) What is mapReduce.

→ MapReduce is a method for processing big data across many computers.
It splits tasks into smaller parts, processes them simultaneously, and then combines the result.

6) Which are basic primary steps related to MapReduce?

→ Mapper phase
Reducer phase.

map
split & shuffle
Reduce.

Mapreduce
Cluster computing
Grid computing

7) What is Scoop?

→ SCOOP stands for "SQL to Hadoop and Hadoop to SQL".

↳ Scoop is a tool designed to transfer data between Hadoop and relational database servers

8) What is FLUME?

→ Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data.

9) What is Spark? application.

→ Apache Spark is an open-source, distributed processing system used for big data workloads.

↳ It utilizes in-memory caching, and optimized query execution for fast analytic queries against data of any size

Application.

i) Financial Services: Spark is used in banking to predick customer churn, and recommend new financial products.

ii) Health care: Spark is used to build comprehensive patient. care, by making data available to front-line health workers for every patient interaction.

iii). Retail: Spark is used to attrack, and keep customers through personalized services and offers.

10) What is Hive?

→ Hive is a data warehouse system which is used to analyze structured data.

Features

↳ Hive is fast & scalable.

↳ It is capable of analyzing large datasets stored in HDFS

11) What is Pig?

→ Pig is a high-level data flow platform for executing MapReduce programs of Hadoop.

↳ Language used for Pig is Pig latin

12) What is 'R'?

↳ R programming provides comprehensive sets of tools such as in-built functions and a wide range of packages to perform data analytis, represent data and build visualization.

13) Use & advantages of R?

14) What is tableau?

→ Tableau is a powerful tool used for data analysis & visualization.

15) Which type of dataset is accepted by tableau?

→ Tableau accepts structured data, cloud-based data sources, ~~geograp~~ geospatial data and real-time data stream

16) What is Big Data ?

→ Big data refers to extremely large and diverse collection of structured, unstructured, and semi-structured data

↳ used in machine learning, predictive modeling.

17) Which are 5 V's related to big data.

→ velocity, volume, value, variety, veracity.

18) Which function is required to present graph in 'R' language ?

→ plot() → primary function used to create graph.

hist() → Histogram.
barplot() → bar charts.
scatterplot() → scatter plots.

ggplot2
Package is used for creating highly complea visualizations

19) Differentiate Between Pig, Hive & Mapreduce.

Pig	Hive	Mapreduce
① scripting language	SQL like quering language	compiled language.
② Higher level of abstraction	Higher level of abstraction	lower level of abstraction.
③ ~~More lines of code than maprduce & less than~~ comparatively less lines of code than Mapreduce	~~com~~ comparatively less lines of code than MapReduce & Pig.	More lines of code.

20) How to copy file from one directory to other?

→ primary command for copying files in HDFS.

" hadoop fs -cp "

?

21) Which command need to provide to check Hadoop is installed properly.

→ hadoop version

22) What is difference between single & distributed system.

→

23) Which are pre-requiste for Hadoop installation.

→

check Java version &.

Hardware requirement : 4GB RAM & minimum 60 GB hard disk for better performance.