NAME : RAJENDRA RAKHA ARYA PRABASWARA

CLASS : 3H

NIM : 1941720080

1. **Examples of cases that require Big Data**

    - **Hospitals** really need Big Data technology. because every day the data stored includes laboratory tests, medical record data, doctor's prescriptions, immunizations, drugs, claims and also payments. Big Data is also used to store data related to claims from doctors and hospitals. The data is used to make decisions to increase revenue and reimbursement. The growth in the number of clients accompanied by a very rapid volume of data makes the infrastructure burdened so that Big Data is the right solution.
    - **Bank** Similarly, bank hospitals require big data technology to store customer data, financial transactions, etc. with big data banks can see the performance of each branch office quickly
    - **Travel apps** Like Traveloka
    - **Online Market Place** Like Shopee

2. **Explanation about Hadoop**

    Hadoop is an apache-based software that is used to store and handle very large amounts of big data and efficiently. Hadoop is designed to keep him reliable at work. The application will continue to run even if one or more servers or clusters fail. Its operation is also efficient because it does not require the transfer of large volumes of data across networks. hadoop can process very large data even terabytes. hadoop is also capable of processing any data, large or small, plain text or binary files (such as images), hadoop also has a MapReduce algorithm, meaning you can parallelize data processing.

3. **Explanation about HDFS**

    HDFS (Hadoop Distributed File System). HDFS is a distributed file system developed by Apache. A distributed file system (distributed file system) is a file system that stores data not on a single hard disk drive (HDD) or other storage media, but the data is broken up and stored scattered in a cluster that consists of several computers, can only be 2 computers, tens or even thousands of computers. HDFS is useful for handling gigantic data.

    HDFS has main components namely namenode and datanode.

    Namenode is a node that acts as the master, while the datanode is the nodes in the cluster that act as the slave. Namenodes are responsible for storing, organizing and controlling the blocks of data stored on nodes spread across the cluster.

Datanode is a datanode which is responsible for storing the data blocks addressed to it, and periodically reporting its condition to the namenode.

So, the namenode is like a manager that manages and controls the cluster. Meanwhile, datanodes are like workers that store data and execute commands from the namenode.

4. **Explanation about MapReduce**

MapReduce is an algorithm program for writing applications that can process Big Data in parallel on multiple nodes. MapReduce provides analytical capabilities for analyzing large volumes of complex data. usually every company has a centralized big data storage system but the centralized system creates too many bottlenecks when processing many files simultaneously. with mapreduce algorithm it will divide the task into small parts and assign it to many computers. Then, the results are collected in one place and integrated to form a result data set.

MapReduce algorithm contains two important tasks, namely Map and Reduce.

- **Map** task takes one data set and transforms it into another data set, where individual elements are split into tuples (key-value pairs).

- **Reduce** task takes the output of the Map as input and combines those data tuples (key-value pairs) into a smaller set of tuples.