GOA COLLEGE OF ENGINEERING

“Bhausaheb Bandodkar Technical Education Complex”

ASSIGNMENT 4 **DATE: 03.01.2022**

1) Describe the working of Map reduce with a relevant example.

MapReduce is a processing technique and a program model for distributed computing based on java.
The MapReduce algorithm contains two important tasks, namely Map and Reduce.
Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.
Given below is the data regarding the electrical consumption of an organization. It contains the monthly electrical consumption and the annual average for various years.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Avg
1979	23	23	2	43	24	25	26	26	26	26	25	26	25
1980	26	27	28	28	28	30	31	31	31	30	30	30	29
1981	31	32	32	32	33	34	35	36	36	34	34	34	34
1984	39	38	39	39	39	41	42	43	40	39	38	38	40
1985	38	39	39	39	39	41	41	41	00	40	39	39	45

If the above data is given as input, we have to write applications to process it and produce results such as finding the year of maximum usage, year of minimum usage, and so on. This is a walkover for the programmers with finite number of records. They will simply write the logic to produce the required output, and pass the data to the application written.

But, think of the data representing the electrical consumption of all the largescale industries of a particular state, since its formation.

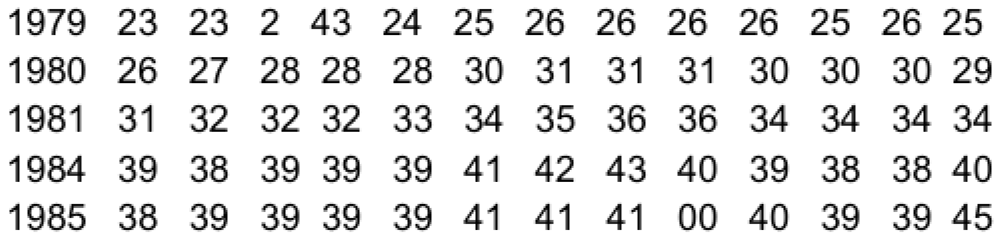
When we write applications to process such bulk data,

- They will take a lot of time to execute.
- There will be a heavy network traffic when we move data from source to network server and so on.

To solve these problems, we have the MapReduce framework.

Input Data

The above data is saved as sample.txt and given as input. The input file looks as shown below.



1979	23	23	2	43	24	25	26	26	26	26	25	26	25
1980	26	27	28	28	28	30	31	31	31	30	30	30	29
1981	31	32	32	32	33	34	35	36	36	34	34	34	34
1984	39	38	39	39	39	41	42	43	40	39	38	38	40
1985	38	39	39	39	39	41	41	41	00	40	39	39	45

2. Differentiate between MapReduce & Apache PIG.

S.No	MapReduce	Pig
1.	It is a Data Processing Language.	It is a Data Flow Language.
2.	It converts the job into map-reduce functions.	It converts the query into map-reduce functions.
3.	It is a Low-level Language.	It is a High-level Language
4.	It is difficult for the user to perform join operations.	Makes it easy for the user to perform Join operations.
5.	The user has to write 10 times more lines of code to perform a similar task than Pig.	The user has to write fewer lines of code because it supports the multi-query approach.
6.	It has several jobs therefore execution time is more.	It is less compilation time as the Pig operator converts it into MapReduce jobs.
7.	It is supported by recent versions of the Hadoop.	It is supported with all versions of Hadoop.

3. Define HDFS? Explain in brief about the basic building blocks of Hadoop?

The Hadoop Distributed File System (HDFS) is a distributed file system for Hadoop. It contains a master/slave architecture.

The Hadoop overall architecture is a distributed master/from architecture consisting of a set of daemons and a set of host programs, and daemons are: Namenode,datanode,secondary namenode,jobtracker,tasktracker

The Namenode,datanode,secondary namenode is divided into stored process classes, while Jobtracker and Tasktracker are divided into computational process classes.

Namenode:

Namenode is the master node of the Hadoop distributed Storage System (HDFS), which itself does not participate in I/O tasks, but instead gives these tasks to the datanode that it manages. Namenode the file system's metadata is stored in memory.

Datanode:

Datanode is a HDFS node from the Hadoop distributed storage System (slave node), which is responsible for the actual task of reading and writing HDFS blocks (a large file is divided into HDFS block) and continuously reporting status to Namenode.

Secondary Namenode:

Secondary Namenode is a worker process used in the cluster to monitor the state of the HDFs cluster. It is also not the same as namenode that it does not accept and record any real-time changes in HDFs. Instead, it deals only with Namenode, and periodically collects snapshots of HDFs states (snapshot), which are used primarily to restore work when the Namenode fails.

Job Tracker:

Job Tracker is our contact for applications and Hadoop, and when we submit code to the Hadoop cluster, it determines the execution plan, including deciding which files to process, assigning different tasks to each node (which is actually assigned to task Tracker, and then forwarding), and monitor all tasks that are running. This process typically runs on the primary node of the cluster.

4. Why is NoSQL essential in handling big data?

NoSQL allows for high-performance, agile processing of information at massive scale. It stores unstructured data across multiple processing nodes, as well as across multiple servers. As such, the NoSQL distributed database infrastructure has been the solution of choice for some of the largest data warehouses.

5. Distinguish between Hadoop 1. X and Hadoop 2.X

S.No	Key	Hadoop1	Hadoop2
1.	New Components and API	As Hadoop 1 introduced prior to Hadoop 2 so has some less components and APIs as compare to that of Hadoop 2.	On other hand Hadoop 2 introduced after Hadoop 1 so has more components and APIs as compare to Hadoop 1 such as YARN API,YARN FRAMEWORK, and enhanced Resource Manager.
2.	Support	Hadoop 1 only supports MapReduce processing model in its architecture and it does not support non MapReduce tools.	On other hand Hadoop 2 allows to work in MapReducer model as well as other distributed computing models like Spark, Hama, Giraph, Message Passing Interface) MPI & HBase coprocessors.
3.	Resource Management	Map reducer in Hadoop 1 is responsible for processing and cluster-resource management.	On other hand in case of Hadoop 2 for cluster resource management YARN is used while processing management is done using different processing models.
4.	Scalability	As Hadoop 1 is prior to Hadoop 2 so comparatively less scalable than Hadoop 2 and in context of scaling of nodes it is limited to 4000 nodes per cluster	On other hand Hadoop 2 has better scalability than Hadoop 1 and is scalable up to 10000 nodes per cluster.
5.	Implementation	Hadoop 1 is implemented as it follows the concepts of lots which can be used to run a Map task or a Reduce task only.	On other hand Hadoop 2 follows concepts of containers that can be used to run generic tasks.
6.	Windows Support	Initially in Hadoop 1 there is no support for Microsoft Windows provided by Apache.	On other hand with an advancement in version of Hadoop Apache provided support for Microsoft windows in Hadoop 2.

6. Write Hive Queries to perform following task

- **Creation of a table**

```
CREATE [TEMPORARY] [EXTERNAL] TABLE [IF NOT EXISTS] [db_name.] table_name

[(col_name data_type [COMMENT col_comment], ...)] [COMMENT table_comment]
[ROW FORMAT row_format] [STORED AS
file_format]
```

- **Joining of table**

```
join_table:

  table_reference JOIN table_factor [join_condition]
  | table_reference {LEFT|RIGHT|FULL} [OUTER] JOIN table_reference join_condition
  | table_reference LEFT SEMI JOIN table_reference join_condition
  | table_reference CROSS JOIN table_reference [join_condition]
```

7. Discuss advantages of MongoDB over RDBMS

Schema less − MongoDB is a document database in which one collection holds different documents. ...
Structure of a single object is clear. No complex joins.
Deep query-ability. ...
Tuning.
Ease of scale-out − MongoDB is easy to scale.

8. Write short notes on visual data analysis techniques.

Data visualization is actually a set of data points and information that are represented graphically to make it easy and quick for user to understand. Data visualization is good if it has a clear meaning, purpose, and is very easy to interpret, without requiring context. Tools of data visualization provide an accessible way to see and understand trends, outliers, and patterns in data by using visual effects or elements such as a chart, graphs, and maps.

Characteristics of Effective Graphical Visual :

It shows or visualizes data very clearly in an understandable manner. It encourages
viewers to compare different pieces of data.
It closely integrates statistical and verbal descriptions of data set.

It grabs our interest, focuses our mind, and keeps our eyes on message as human brain tends to focus on visual data more than written data.
It also helps in identifying area that needs more attention and improvement.

Using graphical representation, a story can be told more efficiently. Also, it requires less time to understand picture than it takes to understand textual data.

9. Discuss functions of

- **Inputsplit**

InputSplit in Hadoop MapReduce is the logical representation of data. It describes a unit of work that contains a single map task in a MapReduce program.

- **Recordreader**

The MapReduce RecordReader in Hadoop takes the byte-oriented view of input, provided by the InputSplit and presents as a record-oriented view for Mapper. It uses the data within the boundaries that were created by the InputSplit and creates Key-value pair.

- **Mapper**

Hadoop Mapper is a function or task which is used to process all input records from a file and generate the output which works as input for Reducer. It produces the output by returning new key-value pairs. The input data has to be converted to key-value pairs as Mapper can not process the raw input records or tuples(key-value pairs).

- **Reducer in case of MapReduce.**

Reducer is a phase in hadoop which comes after Mapper phase. The output of the mapper is given as the input for Reducer which processes and produces a new set of output, which will be stored in the HDFS.

10. Write Hive Queries to perform following task (Use suitable example)

- **Creation a Database company**

```
hive(default)> create database company
             > ;
```

- **Creation of table employee in database company((Expected attributes :Ename, Edoj, Eage, Erole, Edepartment))**

```
use company; hive(company)> create table1
               > (
               • Ename string,
               • Edoj date,
               • Eage int,
               • Erole string,
               • Edepartment string
               >);
```

- **Display list of employee department wise based on date of joining (older first)**

```
>SELECT Edepartment FROM company ORDER BY Edoj DESC;
```

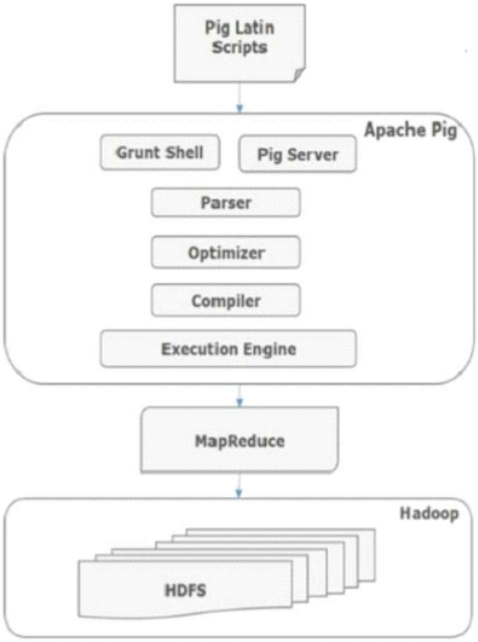
- **Display count of employee department wise**

```
>SELECT COUNT(*) FROM company GROUP BY Edepartment;
```

- **Remove table and database permanently**

```
>DROP TABLE table1;
>DROP DATABASE company;
```

11. With neatly labeled diagram explain Apache PIG framework



Parser

Initially the Pig Scripts are handled by the Parser. It checks the syntax of the script, does type checking, and other miscellaneous checks. The output of the parser will be a DAG (directed acyclic graph), which represents the Pig Latin statements and logical operators.

In the DAG, the logical operators of the script are represented as the nodes and the data flows are represented as edges.

Optimizer

The logical plan (DAG) is passed to the logical optimizer, which carries out the logical optimizations such as projection and pushdown.

Compiler

The compiler compiles the optimized logical plan into a series of MapReduce jobs.

Execution engine

Finally the MapReduce jobs are submitted to Hadoop in a sorted order. Finally, these MapReduce jobs are executed on Hadoop producing the desired results.

12. Explain different types of NoSQL databases.

Key Value Pair Based

Data is stored in key/value pairs. It is designed in such a way to handle lots of data and heavy load.
Key-value pair storage databases store data as a hash table where each key is unique, and the value can be a JSON, BLOB(Binary Large Objects), string, etc.

Column-based

Column-oriented databases work on columns and are based on BigTable paper by Google. Every column is treated separately. Values of single column databases are stored contiguously.

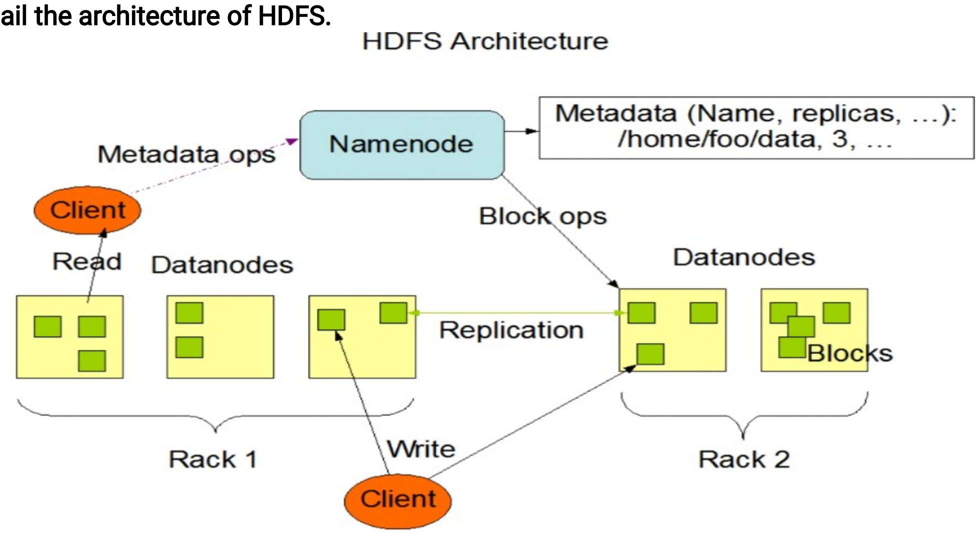
Document-Oriented:

Document-Oriented NoSQL DB stores and retrieves data as a key value pair but the value part is stored as a document. The document is stored in JSON or XML formats. The value is understood by the DB and can be queried.

Graph-Based

A graph type database stores entities as well the relations amongst those entities. The entity is stored as a node with the relationship as edges. An edge gives a relationship between nodes. Every node and edge has a unique identifier.

13. Explain in detail the architecture of HDFS.



The Hadoop Distributed File System (HDFS) is the underlying file system of a Hadoop cluster. It provides scalable, fault-tolerant, rack-aware data storage designed to be deployed on commodity hardware. Several attributes set HDFS apart from other distributed file systems. Among them, some of the key differentiators are that HDFS is:

designed with hardware failure in mind

built for large datasets, with a default block size of 128 MB optimized for
sequential operations

rack-aware

cross-platform and supports heterogeneous clusters

Data in a Hadoop cluster is broken down into smaller units (called blocks) and distributed throughout the cluster. Each block is duplicated twice (for a total of three copies), with the two replicas stored on two nodes in a rack somewhere else in the cluster. Since the data has a default replication factor of three, it is highly available and fault-tolerant. If a copy is lost (because of machine failure, for example), HDFS will automatically re-replicate it elsewhere in the cluster, ensuring that the threefold replication factor is maintained.

HDFS architecture can vary, depending on the Hadoop version and features needed:

Vanilla HDFS

High-Availability HDFS

HDFS is based on a leader/follower architecture. Each cluster is typically composed of a single NameNode, an optional SecondaryNameNode (for data recovery in the event of failure), and an arbitrary number of DataNodes.

14. Explain in detail the Ecosystem of the Hadoop Framework.

People at Google also faced the above-mentioned challenges when they wanted to rank pages on the Internet. They found the Relational Databases to be very expensive and inflexible. So, they came up with their own novel solution. They created the Google File System (GFS).

GFS is a distributed file system that overcomes the drawbacks of the traditional systems. It runs on inexpensive hardware and provides parallelization, scalability, and reliability. This laid the stepping stone for the evolution of Apache Hadoop.

Apache Hadoop is an open-source framework based on Google’s file system that can deal with big data in a distributed environment. This distributed environment is built up of a cluster of machines that work closely together to give an impression of a single working machine.

Here are some of the important properties of Hadoop you should know:

- Hadoop is highly scalable because it handles data in a distributed manner
- Compared to vertical scaling in RDBMS, Hadoop offers horizontal scaling
- It creates and saves replicas of data making it fault-tolerant
- It is economical as all the nodes in the cluster are commodity hardware which is nothing but inexpensive machines
- Hadoop utilizes the data locality concept to process the data on the nodes on which they are stored rather

than moving the data over the network thereby reducing traffic

- It can handle any type of data: structured, semi-structured, and unstructured. This is extremely important in today’s time because most of our data (emails, Instagram,Twitter, IoT devices, etc.) has no defined format

Components:

1. HDFS (Hadoop Distributed File System)
2. MapReduce
3. YARN
4. HBase
5. Pig
6. Hive
7. Sqoop
8. Flume
9. Kafka
10. Zookeeper
11. Spar