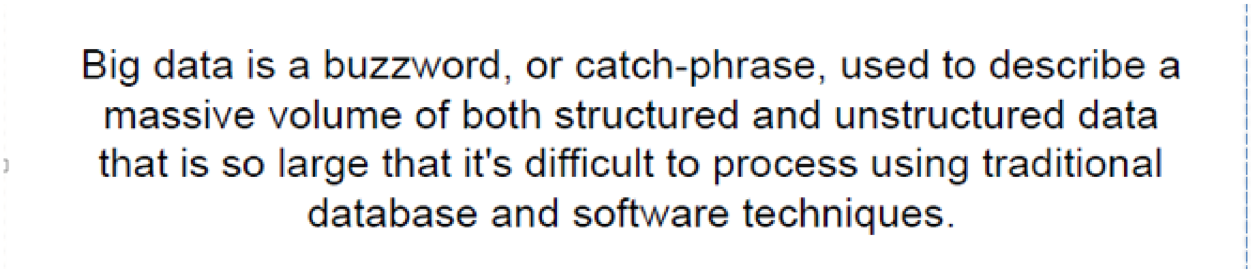
Unit-1

Short Answer Questions

1. **What is Big Data?**

Ans)



Big data is a buzzword, or catch-phrase, used to describe a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques.

2. Define Big Data Analytics?

Ans)

The process of analysis of large volumes of diverse data sets, using advanced analytic techniques is referred to as Big Data Analytics.

3. What are the advantages of Big Data?

Ans)Advantages of Big Data

Big Data Technology has given us multiple advantages.

- Big Data has enabled predictive analysis which can save organizations from operational risks.
- Predictive analysis has helped organizations grow business by analyzing customer needs.
- Big Data has enabled many multimedia platforms to share data Ex: youtube, Instagram
- Medical and Healthcare sectors can keep patients under constant observations.

4. Why is Big Data important?

Ans)
Companies use big data in their systems to improve operations, provide better customer service, create personalized marketing campaigns and take other actions that, ultimately, can increase revenue and profit

- ➔ To be two steps ahead of the competition.
- ➔ Derive insights and drive growth.

5. What are the characteristics of Big Data?

Ans)

Characteristics of Big Data
There are five v's of Big Data that explains the characteristics.
5 V's of Big Data

- Volume
- Veracity
- Variety
- Value
- Velocity

6. What type of data received from the GPS Satellite and the web?

Almanic data, Ephemeris Data, Coordinates.

7. What are the different types of NoSQL databases?

Ans)Here are the four main types of NoSQL databases:

- Document databases
- Key-value stores
- Column-oriented databases
- Graph databases

Unit – 2

Short Answer Questions

1. What is Hadoop?

Ans)

Apache Hadoop is an open source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data. Instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyze massive datasets in parallel more quickly.

2. What are the core components of Hadoop?

Ans)The Core Components of Hadoop are as follows:

- MapReduce
- HDFS
- YARN
- Common Utilities

3. What is commodity hardware?

Ans)

The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on hardware based on open standards is called *commodity hardware*.

4. What is HDFS?

Ans)

HDFS is a distributed file system that handles large data sets running on commodity hardware. It is used to scale a single Apache Hadoop cluster to hundreds (and even thousands) of nodes. HDFS is one of the major components of Apache Hadoop, the others being MapReduce and YARN.

5. What is MapReduce?

Ans)

MapReduce is a programming model for writing applications that can process Big Data in parallel on multiple nodes. MapReduce provides analytical capabilities for analyzing huge volumes of complex data.

6. What are the roles of a Hadoop server?

Ans)

The three major categories of machine roles in a Hadoop deployment are

- Client machines
- Masters nodes
- Slave nodes.

The Master nodes oversee the two key functional pieces that make up Hadoop: storing lots of data (HDFS), and running parallel computations on all that data (Map Reduce).

7. What are the Hadoop daemon processes?

Ans)Daemons mean **Process**. Hadoop Daemons are a set of processes that run on Hadoop.

Apache Hadoop 2 consists of the following Daemons:

- NameNode
- DataNode
- Secondary Name Node
- Resource Manager
- Node Manager

8. What is the replication factor and what is the default replication factor of Hadoop?

Ans)

It is basically the number of times a Hadoop framework replicates each and every Data Block. Block is replicated to provide Fault Tolerance.

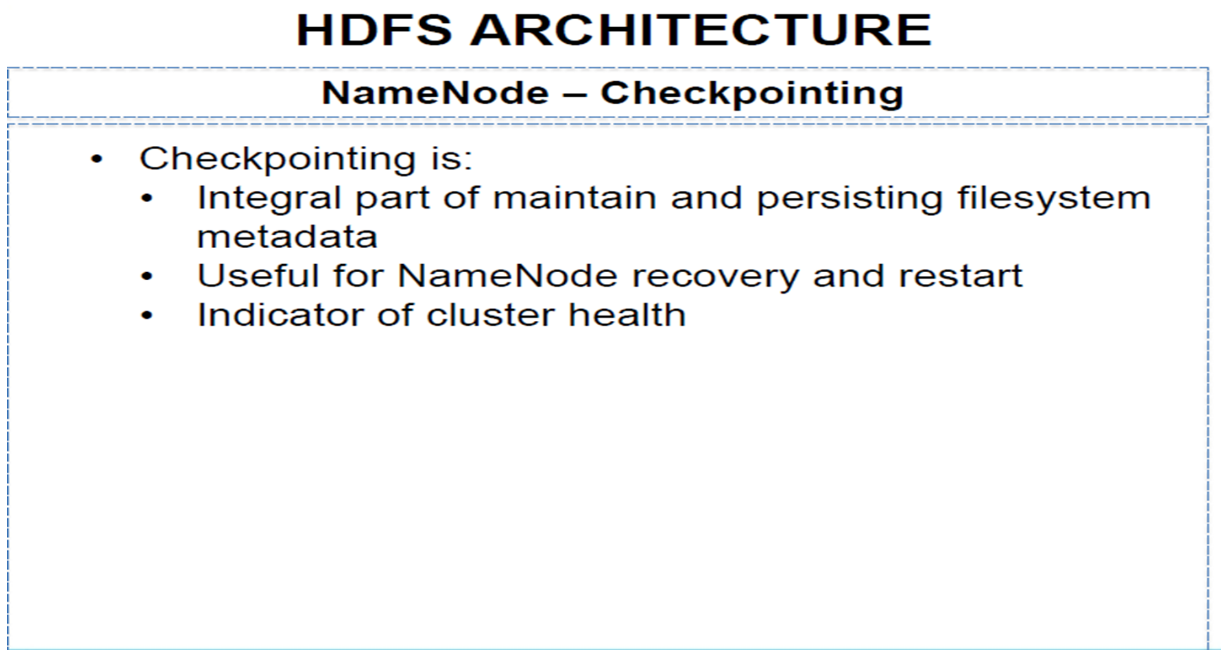
The default replication factor is 3 which can be configured as per the requirement; it can be changed to 2 (less than 3) or can be increased (more than 3.).

9. What is checkpointing?

Ans)

Checkpointing is a process that takes an fsimage and edit log and compacts them into a new fsimage. This way, instead of replaying a potentially unbounded edit log, the NameNode can load the final in-memory state directly from the fsimage. This is a far more efficient operation and reduces NameNode startup time.

Or



HDFS ARCHITECTURE

NameNode – Checkpointing

- Checkpointing is:
  - Integral part of maintain and persisting filesystem metadata
  - Useful for NameNode recovery and restart
  - Indicator of cluster health

Unit -3

Short Answer Questions

1. Write the list of configuration files needs to be edited to setup Hadoop?

ANS)]

2. Write a Hadoop command to copy data from local file system to HDFS and HDFS to local file system?

ANS)

local file system to HDFS:

```
hdfs dfs -put /local-file-path /hdfs-file-path
```

HDFS to local file system:

```
hdfs dfs -copyToLocal /hdfs-file-path /local-file-path
```

3. Write Command line syntax to create SSH?

ANS)

```
putty.exe -ssh  <USERNAME> @ <IP ADDRESS>
```

Set 1

1. What is the Mapper Phase?

ans)Mapper task is the first phase of processing that processes each input record (from RecordReader) and generates an intermediate key-value pair.

2. What is the Reducer Phase?

Ans)

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.

3. What is the InputFormat class?

Ans)

Hadoop InputFormat describes the input-specification for execution of the Map-Reduce job.

InputFormat describes how to split up and read input files. In MapReduce job execution, InputFormat is the first step. It is also responsible for creating the input splits and dividing them into records.

4. What is inputsplit?

Ans)

InputSplit is the logical representation of data in Hadoop MapReduce. It represents the data which individual mapper processes. Thus the number of map tasks is equal to the number of InputSplits. Framework divides split into records, which mapper processes.
MapReduce InputSplit length is measured in bytes.

5. What is Apache Pig?

Ans)

Apache Pig is a high-level data flow platform for executing MapReduce programs of Hadoop. The language used for Pig is Pig Latin.

The Pig scripts get internally converted to Map Reduce jobs and get executed on data stored in HDFS.

6. What scalar data types give examples?

Ans)

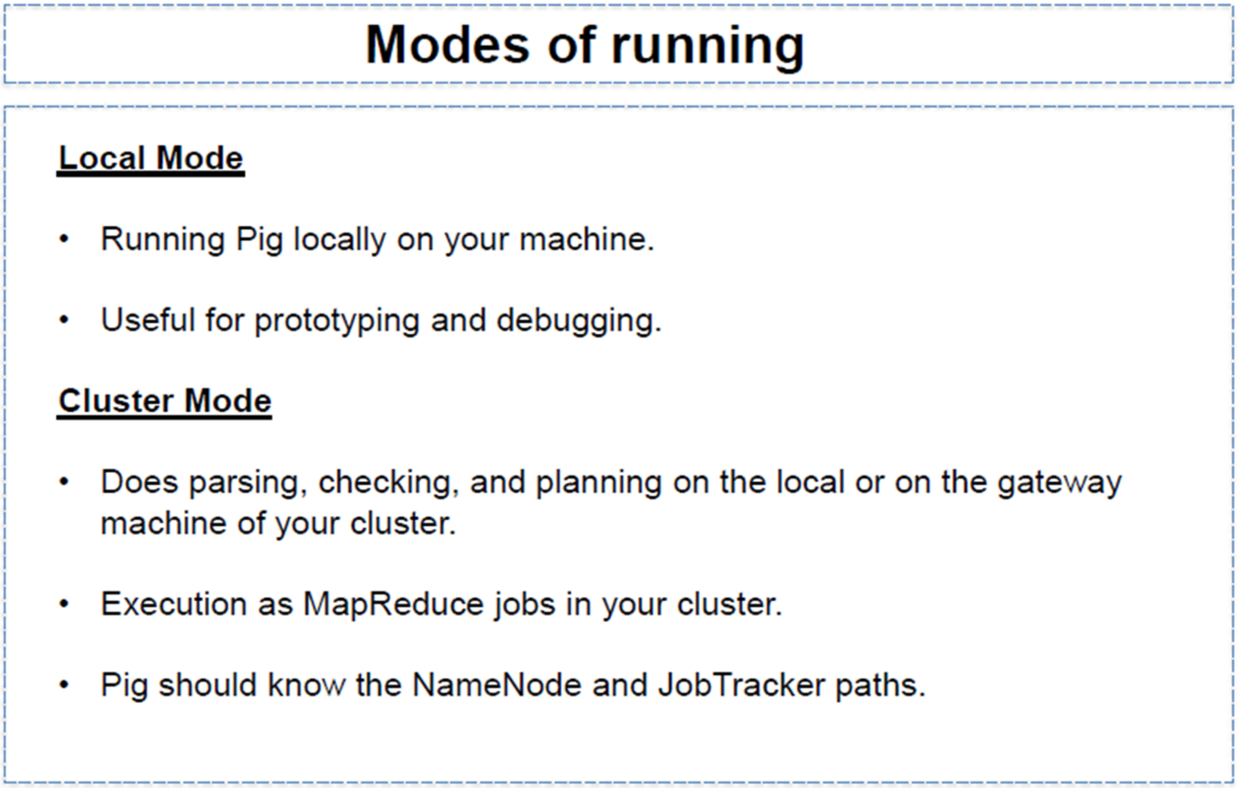
Scalar Data Types

Pig scalar types are simple types that appear in most programming languages.

Data Types	Description	Example
Int	It is a singed 32-bit integer value	2
long	It is a singed 64-bit integer value	15L or 15l
Float	It is a 32-bit floating point value	4.5F or 4.5.5f or 4.5e2f or 4.5E2F
double	It is a 64-bit floating point value	08.5 or 08.5e2 or 08.5E2
Boolean	Boolean represents true or false values	true/false
charArray	It is a Character array	Hello
byteArray	The default datatype in pig is a ByteArray	
bigInteger	It displays the BigInteger	70204096223145
bigDecimal	It displays the BigDecimal	198.789654123133211

7. What are the running modes of apache pig?

Ans)



Modes of running

Local Mode

- Running Pig locally on your machine.
- Useful for prototyping and debugging.

Cluster Mode

- Does parsing, checking, and planning on the local or on the gateway machine of your cluster.
- Execution as MapReduce jobs in your cluster.
- Pig should know the NameNode and JobTracker paths.

8. What is R programming?

Ans)

R is an open-source programming language that is widely used as a statistical software and data analysis tool. R generally comes with the Command-line interface.

9. What is HIVE?

Ans)

Hive is a data warehouse system which is used to analyze structured data. It is built on the top of Hadoop. It was developed by Facebook.

Hive provides the functionality of reading, writing, and managing large datasets residing in distributed storage. It runs SQL-like queries called HQL which get internally converted to MapReduce jobs.

10. What are the advantages of HIVE?

Ans)

Advantages of Hive:

- Keeps queries running fast
- HiveQL is a declarative language like SQL
- Provides the structure on an array of data formats
- Multiple users can query the data with the help of HiveQL
- Very easy to write query including joins in Hive
- Simple to learn and use

Set2

1. What is Sorting and shuffling phase?

Ans)

3. Shuffling in MapReduce

The process of transferring data from the mappers to reducers is known as shuffling i.e. the process by which the system performs the sort and transfers the map output to the reducer as input. So, MapReduce shuffle phase is necessary for the reducers, otherwise, they would not have any input (or input from every mapper).

4. Sorting in MapReduce

The keys generated by the mapper are automatically sorted by MapReduce Framework, i.e. Before starting of reducer, all intermediate key-value pairs in MapReduce that are generated by mapper get sorted by key and not by value. Values passed to each reducer are not sorted; they can be in any order.

2. What are different modes available to set up hadoop?

Ans)

Hadoop can be installed in three different modes.

1. Standalone Mode.
2. Pseudo-Distributed Mode.
3. Fully Distributed Mode.

3. What is OutputFormat class?

Ans)

OutputFormat checks the output specification for execution of the Map-Reduce job. It describes how RecordWriter implementation is used to write output to output files.

4. What are the complex data types?

Ans)
The Hive Complex Data Type are categorized as:
Array
Map
Struct
Union

5. What is a data model in pig?

ANS)
Pig Latin data model allows Pig to handle any kind of data. Pig Latin data model is fully nested and can treat both atomic like integer, float, and non-atomic complex data types such as Map and tuple.

6. What are the relational operators in pig?

Ans)

Pig Latin – Relational Operations

The following table describes the relational operators of Pig Latin.

Operator
Loading and Storing
LOAD
STORE
Filtering
FILTER
DISTINCT
FOREACH, GENERATE
STREAM

Grouping and Joining
JOIN
COGROUP
GROUP
CROSS
Sorting
ORDER
LIMIT
Combining and Splitting
UNION
SPLIT

Diagnostic Operators
DUMP
DESCRIBE
EXPLAIN
ILLUSTRATE

7. What is the procedure for executing a pig program?

Ans)

You can execute the Pig scripts by using one of the methods:

- Grunt Shell
- Script file
- Embedded script
- **Interactive Mode (Grunt shell)** − You can run Apache Pig in interactive mode using the Grunt shell. In this shell, you can enter the Pig Latin statements and get the output (using Dump operator).
- **Batch Mode (Script)** − You can run Apache Pig in Batch mode by writing the Pig Latin script in a single file with .pig extension.

- **Embedded Mode (UDF)** − Apache Pig provides the provision of defining our own functions (User Defined Functions) in programming languages such as Java, and using them in our script.
-

8. Write syntax for creating a table in HIVE?

Ans)

A table in Hive is a set of data that uses a schema to sort the data by given identifiers.

The general syntax for creating a table in Hive is:

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

(col_name data_type [COMMENT 'col_comment'],, ...)

[COMMENT 'table_comment']

[ROW FORMAT row_format]

[FIELDS TERMINATED BY char]

[STORED AS file_format];
```

9. What is a managed table?

Ans)

Managed tables are Hive owned tables where the entire lifecycle of the tables' data are managed and controlled by Hive. All the write operations to the Managed tables are performed using Hive SQL commands.

10. What is an external table?

Ans)External tables are tables where Hive has loose coupling with the data. The writes on External tables can be performed using Hive SQL commands but data files can also be accessed and managed by processes outside of Hive.

Shuffle phase in Hadoop transfers the map output from Mapper to a Reducer in MapReduce. Sort phase in MapReduce covers the merging and sorting of map outputs. Data from the mapper are grouped by the key, split among reducers and sorted by the key. Every reducer obtains all values associated with the same key. Shuffle and sort phase in Hadoop occur simultaneously and are done by the MapReduce framework.

The shuffle and sort phase is done by the framework. Data from all mappers are grouped by the key, split among reducers and sorted by the key. Each reducer obtains all values associated with the same key. The programmer may supply custom compare functions for sorting and a partitioner for data split.

Shuffle phase in Hadoop transfers the map output from Mapper to a Reducer in MapReduce. Sort phase in MapReduce covers the merging and sorting of map outputs.