VELAGAPUDI RAMAKRISHNA

SIDDHARTHA ENGINEERING COLLEGE

(AUTONOMOUS)

PART-A

10 x 1 = 10M

1. a. List applications of Big Data.

Any two applications

Log analytics is a common use case for Big Data projects.

The fraud detection pattern

Market Basket Analysis

Weather Forecasting

Retail – Data Processing

Aadhar Project by govt of india

Healthcare Analysis

b. Name the latest versions of Hadoop release.

Any two latest versions

2.7.x 2.7.7 / May 31, 2018

2.8.x 2.8.5 / September 15, 2018

2.9.x 2.9.2 / November 9, 2018

2.10.x 2.10.1 / September 21, 2020

3.1.x 3.1.4 / August 3, 2020

3.2.x 3.2.2 / January 9, 2021

3.3.x 3.3.1 / June 15, 2021

c. Define Big Data?

Ans: The data which can‘t be stored and processed by traditional systems. It may have huge in volume, velocity or veriety.

d. What is the role of job tracker in HDFS?

Ans: *The jobtracker coordinates all the jobs run on the system by* scheduling tasks to run on tasktrackers.

- Tasktrackers run tasks and send progress reports to the jobtracker, which keeps a record of the overall progress of each job. If a task fails, the jobtracker can reschedule it on a different tasktracker.

e. How to copy a file from the HDFS to local file system?

Ans: hadoop fs –copyToLocal source in HDFS> <destination to local filesystem>

f. **Which is the default input formats defined in MapReduce**?

Ans: any two input formats

IntWritable, Text input format

g. What is the key-value pair in Hadoop MapReduce.

Map Reduce works by breaking the processing into two phases:

1. the map phase
2. the reduce phase

Map Phase and Reduce phase receive data in the form of key and value and generates output in the form of key and value.

Map Phase: It is data preparation phase, read input and give required information to reduce in the form of key and value.

Reduce phase: It gets output of map phase followed by sort and shuffle and it will do the aggregate/ required work.

h. List different complex data types in Pig.

Complex Types: tuples, bags, map;

a. A Tuple is an ordered set of fields
b. A Bag is a collection of tuples
c. A Map key, value pair

i. What is metastore in Hive.

Hive chooses respective database servers to store the schema or Metadata of tables, databases, columns in a table, their data types, and HDFS mapping. This is the default metastore for HIVE.

It is Apache Derby Hive Server Process.

j. List any two Pig commands.

Ans: load , Store , dump

PART-B

4 x 15 = 60M

UNIT-I

2. a. Briefly discuss about characteristics of Big Data. 7M

b. Differentiate between Data vs Information vs Big Data. 8M

2 (a) Answer:

Three characteristics of Big Data –

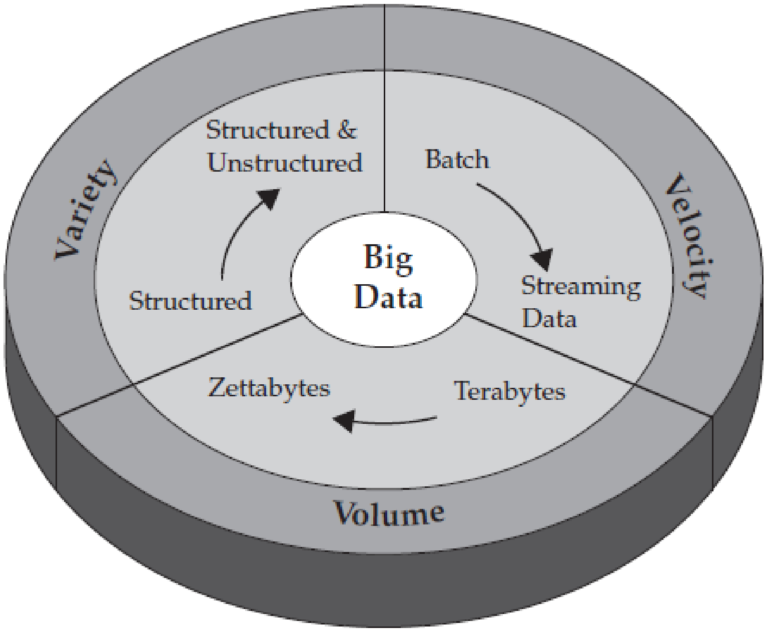
List -2M

Explanation – 3 x 2 = 6M

Volume, Velocity and Variety

Volume of Data(Can there be enough?)

- The sheer volume of data being stored today is exploding.
- In the year 2000, 800000 petabytes of data were stored in the world. Expectation that the data will reach 35zettabytes by 2020.
- Twitter alone generates more than 7 terabytes of data every day, facebook 10 terabytes and some enterprises generates terabytes of data every hour of every day of the year.



Varieties of data

1. Structured Data - arranged in rows and columns – databases, excel sheet
2. Semi structured data – self describing documents,- xml, json,html
3. Unstructured Data- schema less – text, audio, video, image, ppt, pdf

Velocity of Data(How fast is fast?)

- The sheer volume and variety of data we collect and store has changed, so the velocity at which it is generated and needs to be handled.
- How quick the data is arriving and stored, and its associated rates of retrievals.
- Today's enterprises are dealing with PB of data instead of TB, and the increase in RFID sensors and other information streams has led to a constant flow of data at a pace that has made it impossible for traditional systems to handle.
- Big Data scale streams computing is a concept that IBM has been delivering on for some time and serves as a new paradigm for Big Data problem. In traditional processing u can think of running queries against relatively static data.
- Dealing effectively with Big Data requires that u perform analytics against the volume and variety of the data while it is still in motion, not just after it is at rest.

2(b.) Differentiate between Data vs Information vs Big Data. 8M

Ans : Data – 2M

Information – 2M

Big Data – 2M

Comparision – 2M

(OR)

3. a. What is Hadoop? Categorize various tools of Hadoop framework. 8M

b. Explain Hadoop Ecosystem with examples. 7M

a. Answer

Hadoop – 2M

Tools – HDFS – mapReduce – 2M

HDFS – 2M

MapReduce or any tools – 2M

3. **b. Explain Hadoop Ecosystem with examples. 7M**

Hadoop Ecosystem – 3M

Diagram – 2M

Any two examples – 2M

Hadoop Ecosystem

Hadoop Ecosystem:

Common

A set of components and interfaces for distributed filessystems & general I/O(serialization,JavaRPC(remote procedure calls), persistent data structures).

Avro

A serialization system for efficient, cross-language RPC, & persistent data storage.

MapReduce

A distributed data processing model & execution environment that runs on large clusters of commodity machines.

HDFS

A distributed filesystem that runs on large clusters of commodity machines.

Pig

A data flow language and execution environment for exploring very large datasets. Pig runs on HDFS and Mapreduce clusters.

Hive

A distributed data warehouse. Hive manages data stored in HDFS and provides a query language based on SQL for querying the data.

Hbase

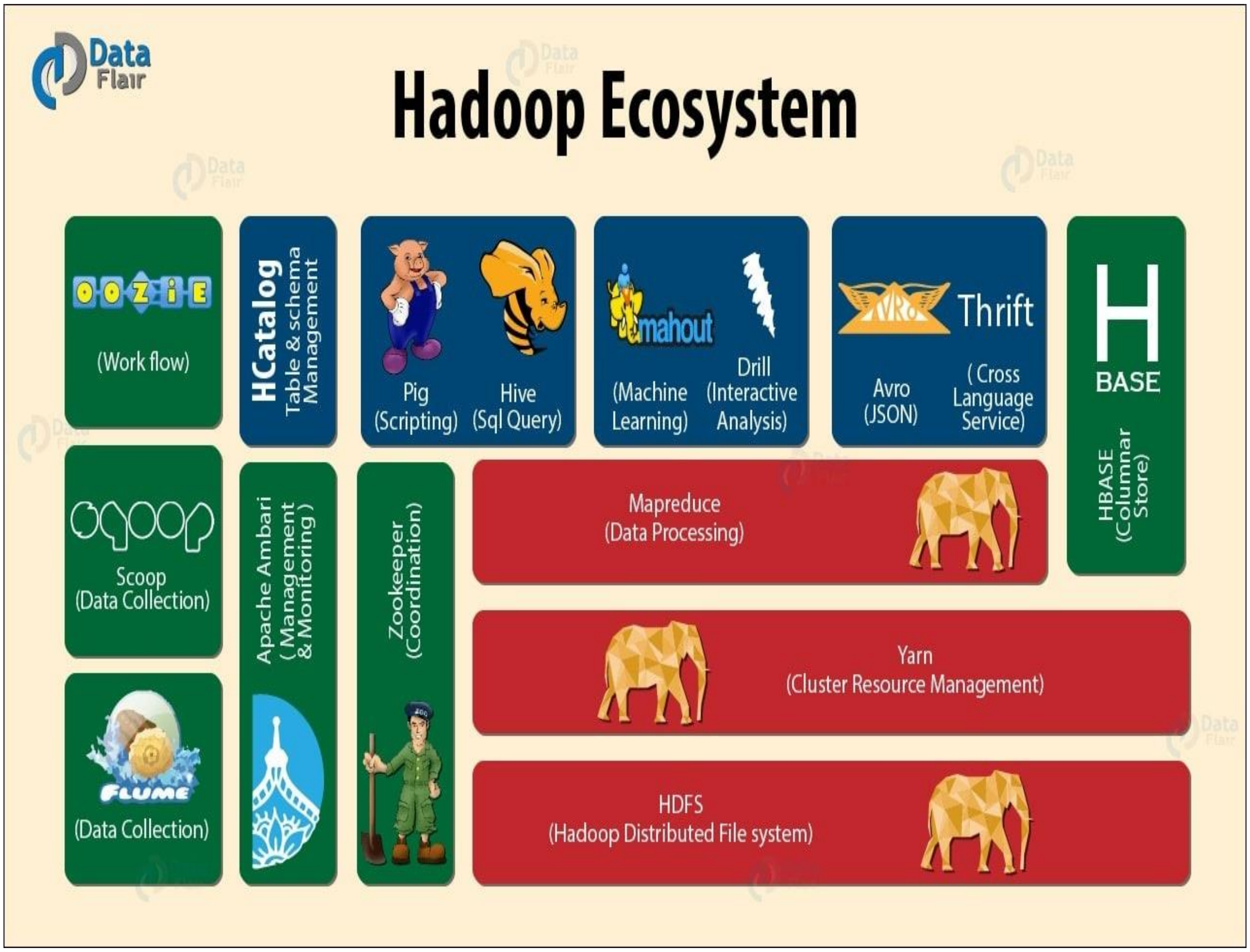
A distributed, column-oriented database. Hbase uses HDFS for its underlying storage, and supports both batch-style computations using MapReduce and point queries(random reads).

ZooKeeper

A distributed, highly available coordination service. ZooKeeper provides primitives such as distributed locks that can be used for building distributed applications.

Sqoop

A tool for efficiently moving data between relational databases and HDFS.



UNIT II

4. (a) Write short notes on name node and data node. 7M

b) Discuss how to read data from a Hadoop URL. 8M

4(a) ANs :

Name node – 3M

Data node – 3M

Diagram – 1M

Namenodes and Datanodes

- **HDFS cluster has two types of nodes operating in a master-worker pattern. A namenode (the master) & a number of datanodes (workers).**
- **The namenode manages the filesystem namespace. It maintains the filesystem tree & the metadata for all the files & directories in the tree. This information is stored on the local disk in the form of two files: the namespace image and their edit log.**
- **The namenode also knows the datanodes on which all the blocks for a given file are located.**
- **Datanodes are the workhorses of the filesystem. They store and retrieve blocks when they are told to and they report back to the namenode periodically with lists of blocks that they are storing.**
- **Without the namenode, the filesystem cannot be used. If the machine running the namenode were failed, all the files on the filesystem would be lost since there would be no way of knowing how to reconstruct the files from the blocks on the datanodes.**

4. b) Discuss how to read data from a Hadoop URL. 8M

Ans :

Diagram – 2M

Explanation – 6M

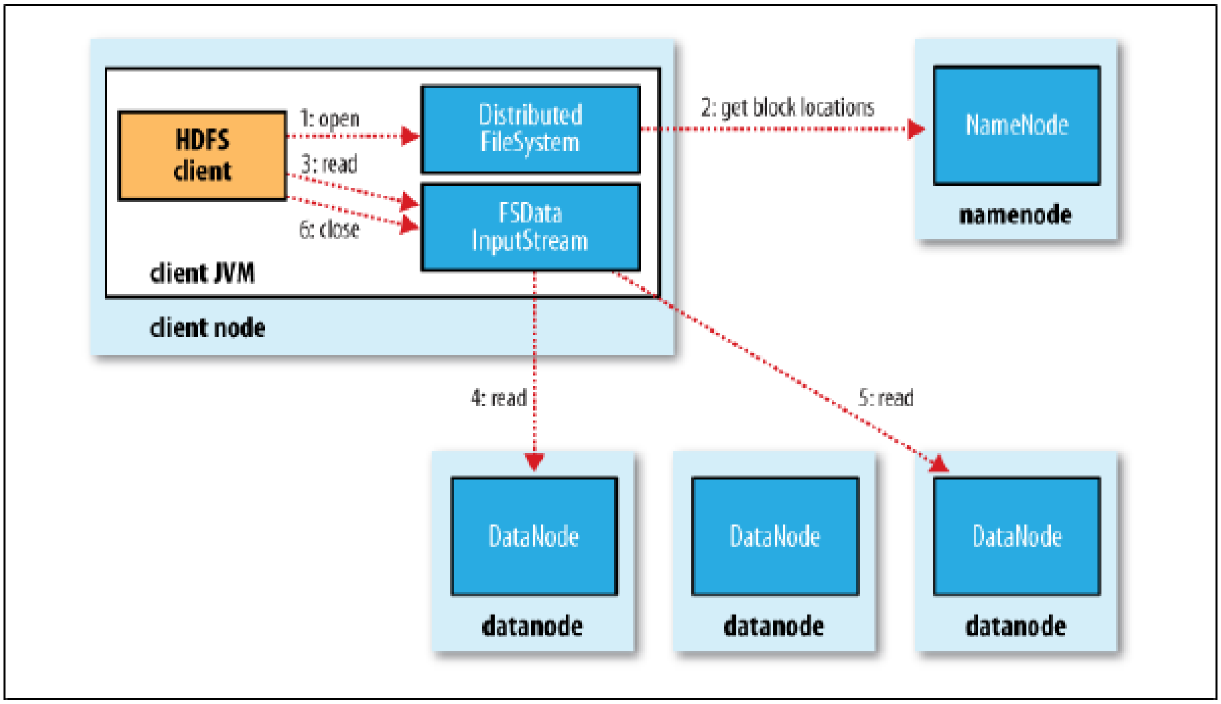
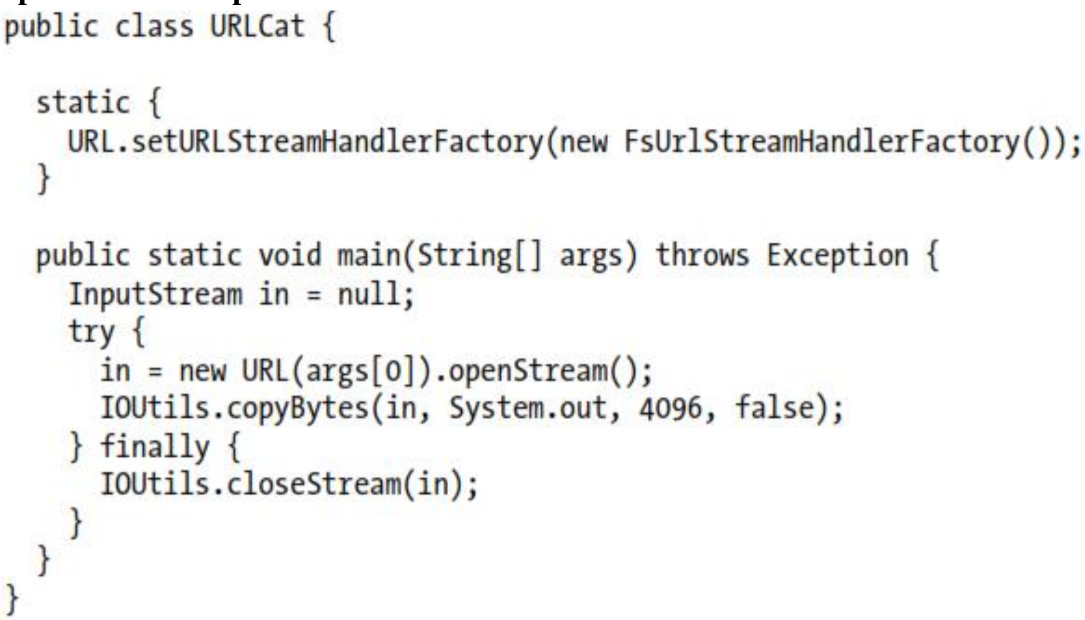


Figure 3-2. A client reading data from HDFS

Read operation Hadoop URL

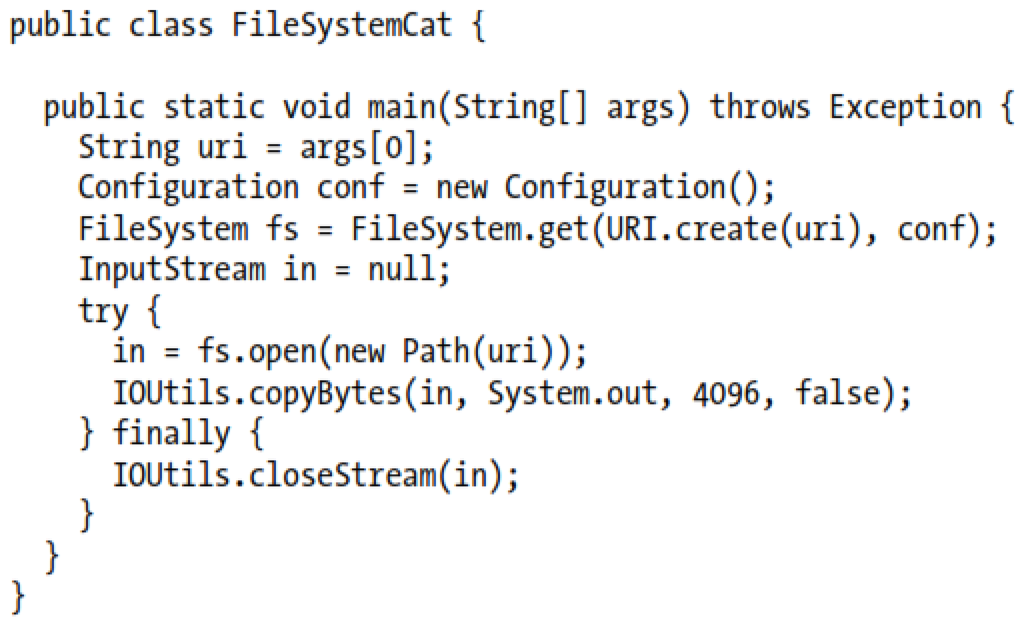


```
public class URLCat {

  static {
    URL.setURLStreamHandlerFactory(new FsUrlStreamHandlerFactory());
  }

  public static void main(String[] args) throws Exception {
    InputStream in = null;
    try {
      in = new URL(args[0]).openStream();
      IOUtils.copyBytes(in, System.out, 4096, false);
    } finally {
      IOUtils.closeStream(in);
    }
  }
}
```

File system API



```
public class FileSystemCat {

  public static void main(String[] args) throws Exception {
    String uri = args[0];
    Configuration conf = new Configuration();
    FileSystem fs = FileSystem.get(URI.create(uri), conf);
    InputStream in = null;
    try {
      in = fs.open(new Path(uri));
      IOUtils.copyBytes(in, System.out, 4096, false);
    } finally {
      IOUtils.closeStream(in);
    }
  }
}
```

(OR)

5. a. Discuss about basic file system operations in HDFS. 7M

b. **Define HDFS? Explain in brief about the basic building blocks of Hadoop.? 8M**

5(A) **a. Discuss about basic file system operations in HDFS. 7M**

Any seven commands – 7M

Basic File System Operations

- The file-system is ready to be used, & we can do all the usual file-system operations such as reading files, creating directories, moving files, deleting data, an listing directories.
- You can type hadoop fs –help to get detailed help on every command.

Copying a file from the local file system to HDFS:

Eg:

% hadoop fs –copyFromLocal input/docs/quangle.txt hdfs://localhost/user/tom/quangle.txt

% hadoop fs –copyFromLocal input/docs/quangle.txt /user/tom/quangle.txt

% hadoop fs –copyFromLocal input/docs/quangle.txt quangle.txt

Copying the file back to the local filesystem and check whether it’s the same:

% hadoop fs –copyToLocal quangle.txt quangle.copy.txt

% md5 input/docs/quangle.txt quangle.copy.txt

% hadoop fs –mkdir books

% hadoop fs –ls

5.(b) **Define HDFS? Explain in brief about the basic building blocks of Hadoop.? 8M**

HDFS – 2M

Namenode – 2M
Datanodes – 2M
Diagram – 2M

HDFS:

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 providing inexpensive solution by using commodity hardware.

Commodity Hardware: The normal machines, no need to purchase a high end machine. We can use existing hardware that is commodity hardware.

- Hadoop doesn't require expensive, highly reliable hardware to run on.
- It's designed to run on clusters of commodity hardware for which the chance of node failure across the cluster is high, at least for large clusters.
- HDFS is designed to carry on working without a noticeable interruption to the user in the face of such failure.

Name Node & Data Node:

- HDFS cluster has two types of nodes operating in a master-worker pattern. A namenode (the master) & a number of datanodes (workers).
- The namenode manages the filesystem namespace. It maintains the filesystem tree & the metadata for all the files & directories in the tree.
- This information is stored on the local disk in the form of two files: the namespace image and their edit log.
- The namenode also knows the datanodes on which all the blocks for a given file are located. Datanodes are the workhorses of the filesystem. They store and retrieve blocks when they are told to and they report back to the namenode periodically with lists of blocks that they are storing.
- Without the namenode, the filesystem cannot be used. If the machine running the namenode were failed, all the files on the filesystem would be lost since there would be no way of knowing how to reconstruct the files from the blocks on the datanodes.

UNIT-III

6. a. Discuss in brief about Mapper and Reducer. 8M
b. Explain Pig's built in types in detail. 7M

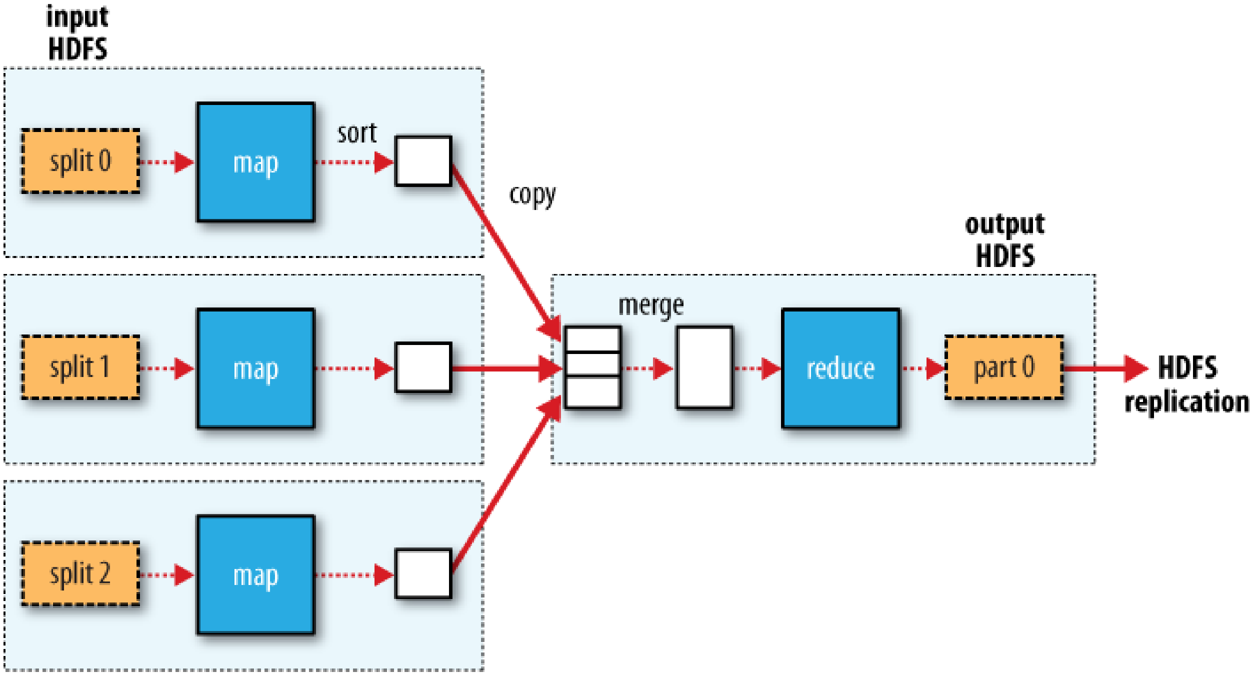
6(a). a. Discuss in brief about Mapper and Reducer. 8M

Diagram – 2M

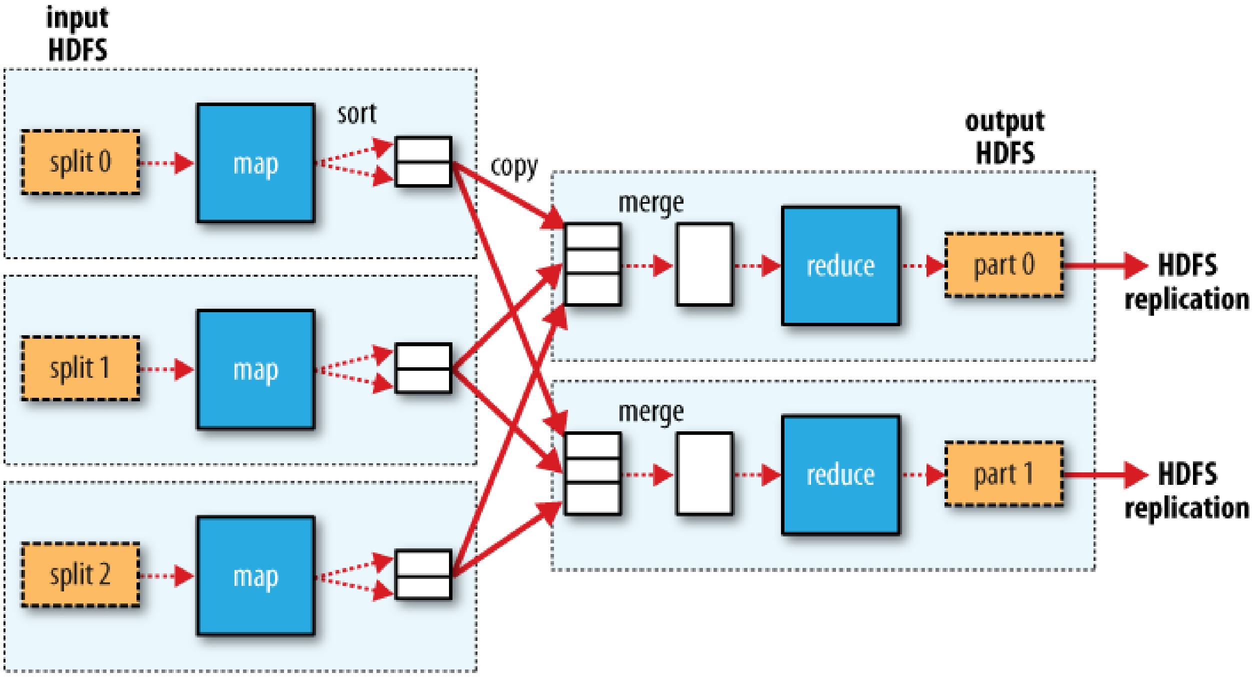
Mapper – 3M

Reducer – 3M

mapreduce data flow – 2M
single reducer - 2M
multiple reducers - 2M
no reducer – 2M



Single reducer



6(b). Explain Pig's built in types in detail. 7M

Ans : any pig seven data types 7 x 1 = 7M

- **Simple Data Types:**
  - **Int, long, float, double, boolean, null, chararray, bytearry;**
- **Complex Types: tuples, bags, map;**
  - **A Tuple is an ordered set of fields**
  - **A Bag is a collection of tuples**
  - **A Map key, value pair**

(OR)

7.(A) List and explain MapReduce input and output formats. 7M

(B) Differentiate between local and distributed modes in pig scripts. 8M

7(a) List and explain MapReduce input and output formats. 7M

Ans: MapReduce – 3M

Any two Input formats – 2M

Any two Output formats- 2M

7(B) Differentiate between local and distributed modes in pig scripts. 8M

Local mode – 4M

Distributed / mapreduce mode – 4M

Local mode

- **Local host and local file system is used**
- **Neither Hadoop nor HDFS is required**
- **Useful for prototyping and debugging**
- **Pig –x local**

- **MapReduce mode**
  - **Run on a Hadoop cluster and HDFS**

Pig –x mapreduce

The default mode is mapreduce

UNIT-IV

8. a. Write hive commands to create a student table with fields: rollnumber, name and address. Also insert two rows into that table. 8M

b) Differentiate HiveQL with traditional SQL. 7M

8 (A) a. Write hive commands to create a student table with fields: rollnumber, name and address. Also insert two rows into that table. 8M

create table – 4M

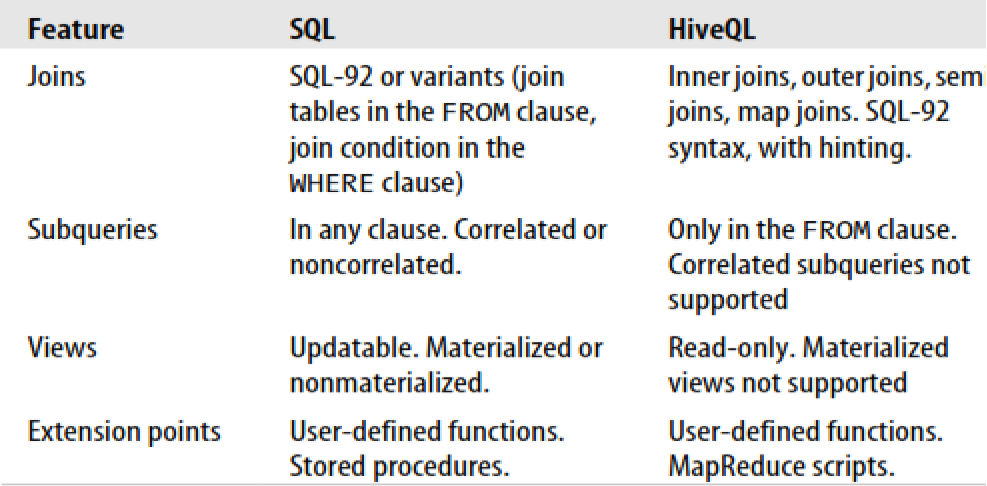
load with example input file – 4M

8(b) b) Differentiate HiveQL with traditional SQL. 7M

HIVEQL – 3M

Any four comparisons 4M

- **It is similar to SQL.**
- **It provides SQL type language for querying called HiveQL or HQL, which is easy to code.**
- **Hive supports rich data types such as structs, lists, and maps.**
- **Hive suppots SQL filters, group-by and order-by clauses.**
- **Custom types, custom functions can be defined.**
-



Feature	SQL	HiveQL
Joins	SQL-92 or variants (join tables in the FROM clause, join condition in the WHERE clause)	Inner joins, outer joins, semi joins, map joins. SQL-92 syntax, with hinting.
Subqueries	In any clause. Correlated or noncorrelated.	Only in the FROM clause. Correlated subqueries not supported
Views	Updatable. Materialized or nonmaterialized.	Read-only. Materialized views not supported
Extension points	User-defined functions. Stored procedures.	User-defined functions. MapReduce scripts.

(OR)

9. (A) Elaborate the procedure to create and manage tables in Hive . 7M

(b) Discuss various types supported by HiveQL with an example. 8M

9(A) Elaborate the procedure to create and manage tables in Hive . 7M

Ans: create tables – 2M

Manage tables – 3M

Example – 2M

- **Hive provides two kinds of tables:**
  - **Managed tables**
  - **External tables**
- **Managed tables**
  - **Hive stores the managed tables under the warehouse folder under Hive.**
  - **The complete life cycle is managed by Hive.**
  - **When the internal table is dropped, it drops the data as well as the metadata.**

- **CREATE TABLE IF NOT EXISTS STUDENT**
  (rollno INT, name STRING, gpa FLOAT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ’\t’;

DESCRIBE STUDENT;

External or self-managed tables

- **When the table is dropped, it retains the data in the underlying location.**
- **Create external table if not exists ext_student (rollno int, name string, gpa float) row format delimited fields terminated by “\t” location ‘/student’**
- **TO load data into table from file**
  - **Load data local inpath ‘/root/file.tsv’ overwrite into table ext_student;**
- **LOCAL keyword is used to load the data from local file system.**
- **To load the data from HDFS remove the local keyword**

9 (b) Discuss various types supported by HiveQL with an example. 8M

Ans: any 8 data types with example – 8M

All the data types in Hive are classified into four types, given as follows:

- **Column Types**
- **Literals**
- **Null Values**
- **Complex Types**
- **Column Types**
- **Column type are used as column data types of Hive. They are as follows:**
- **Integral Types**
- **String Types**
- **Dates**
- **Decimals**
- **Integer type data can be specified using integral data types, INT. When the data range exceeds the range of INT, you need to use BIGINT and if the data range is smaller than the INT, you use SMALLINT. TINYINT is smaller than SMALLINT.**
- **The following table depicts various INT data types:**

Type	**Postfix**	**Example**
TINYINT	**Y**	**10Y**
SMALLINT	**S**	**10S**
INT	**-**	**10**
BIGINT	**L**	**10L**

- **String Types**
- **String type data types can be specified using single quotes (' ') or double quotes (" ").**
- **It contains two data types: VARCHAR and CHAR.**
- **Hive follows C-types escape characters.**
- **The following table depicts various CHAR data types:**

Data Type	**Length**
VARCHAR	**1 to 65355**
CHAR	**255**

- **Timestamp**
- **It supports traditional UNIX timestamp with optional nanosecond precision. It supports java.sql.Timestamp format “YYYY-MM-DD HH:MM:SS.fffffffff” and format “yyyy-mm-dd hh:mm:ss.ffffffffff”.**
- **Dates**
- **DATE values are described in year/month/day format in the form {{YYYY-MM-DD}}.**
- **Decimals**
- **The DECIMAL type in Hive is as same as Big Decimal format of Java. It is used for representing immutable arbitrary precision. The syntax and example is as follows:**
- **DECIMAL(precision, scale) decimal(10,0)**
- **Union Types**
- **Union is a collection of heterogeneous data types. You can create an instance using create union. The syntax and example is as follows:**