1.Write a Map-Reduce Program for Word Count Problem?

https://www.javatpoint.com/mapreduce-word-count-example

2.Explain in detail with a neat diagram about Executing Map Phase – Shuffling and Sorting and Reducing Phase Execution?

https://www.tutorialscampus.com/map-reduce/algorithm.htm

3.Explain the architecture of the pig and its advantages?

ARCHITECTURE

https://www.tutorialspoint.com/apache_pig/apache_pig_architecture.htm

ADVANTAGES

https://data-flair.training/blogs/pig-advantages-and-disadvantages/

4.Explain about pig Data Model and Schema.

https://www.oreilly.com/library/view/programming-pig/9781449317881/ch04.html

5.Explain about pig Relational Operators.

Pig Latin – Relational Operations

The following table describes the relational operators of Pig Latin.

Operator	Description
Loading and Storing	
LOAD	To Load the data from the file system (local/HDFS) into a relation.
STORE	To save a relation to the file system (local/HDFS).
Filtering	
FILTER	To remove unwanted rows from a relation.
DISTINCT	To remove duplicate rows from a relation.
FOREACH, GENERATE	To generate data transformations based on columns of data.
STREAM	To transform a relation using an external program.
Grouping and Joining	
JOIN	To join two or more relations.
COGROUP	To group the data in two or more relations.
GROUP	To group the data in a single relation.
CROSS	To create the cross product of two or more relations.
Sorting	
ORDER	To arrange a relation in a sorted order based on one or more fields (ascending or descending).
LIMIT	To get a limited number of tuples from a relation.
Combining and Splitting	
UNION	To combine two or more relations into a single relation.
SPLIT	To split a single relation into two or more relations.
Diagnostic Operators	
DUMP	To print the contents of a relation on the console.
DESCRIBE	To describe the schema of a relation.
EXPLAIN	To view the logical, physical, or MapReduce execution plans to compute a relation.
ILLUSTRATE	To view the step-by-step execution of a series of statements.

6.Explain about Parameter Substitution with examples?

http://www.hadooplessons.info/2014/09/parameter-substitution-in-pig.html

7.Write about HIVE & its Architecture?

https://www.tutorialspoint.com/hive/hive_introduction.htm

8.Explain HIVE Data Types and Table Creation in hive?

DATA TYPES

https://www.tutorialspoint.com/hive/hive_data_types.htm

TABLE CREATION:

https://www.tutorialspoint.com/hive/hive_create_table.htm

9.Explain the following with examples.

a)Loading data in HIVE Tables

Loading data into a Table

Since Hive has no row-level insert, update, and delete operations, the only way to put data into your table is to use one of the bulk load operations.
Alternatively, we can write the files in directories looked up by Hive. Let's look at an example of bulk load.

- Use LOAD DATA command.

LOAD DATA LOCAL INPATH '/home/hduser/docs/hive/companies/src' INTO TABLE hiveclass.companies;

Here we use the LOAD DATA function for that purpose. You can see that this command is used to load data into a Hive table.
What this command does is that it will just copy the data from the local file path specified and places it inside the directory that was specified, to store that particular table data at the time of the creation of the table. Let's check the contents of the directory where the company's table is defined to store the data.

- Specify a directory as path and not an individual file.

LOAD DATA LOCAL INPATH '/home/hduser/docs/hive/companies' INTO TABLE hiveclass.companies;

Now LOAD DATA, you're going to specify that the data stored in the local file system with key word LOCAL, then using the keyword INPATH you can specify the path of the file that is stored in '*/home/hduser/docs/hive/src' and which is the table that I want to load the data into, which is companies.
LOAD DATA LOCAL INPATH '/home/hduser/docs/hive/src' INTO TABLE companies;

b)Managed Tables

Managed table is also called as Internal table. This is the default table in Hive. When we create a table in Hive without specifying it as external, by default we will get a Managed table.

- Uses default location or explicitly provided location to store data.
- Relevant directory is created.
- Metadata is updated in metastore.
- LOAD DATA statement moves data into the appropriate directory created by Hive.
- When table is dropped, data is removed and metadata is removed from metastore.
- Hive controls the life cycle of the table.

New York Stock Exchange dataset

NYSE_DAILY
exchange
stock symbol
date
open price
high price
low price
close price
volume
adjusted close price

NYSE_DIVIDENDS
exchange
stock symbol
date
dividends

Input Data:

Exchange Symbol Date OpenCloseAdjustedClose HighLow Volume

NYSE AEA 2/8/2010 open*4.42#close*4.24#adj_close*4.24 4.42#4.21 205500

NYSE AEA 2/5/2010 open*4.42#close*4.41#adj_close*4.41 4.54#4.22 194300

NYSE AEA 2/4/2010 open*4.55#close*4.42#adj_close*4.42 4.69#4.39 233800

```
CREATE TABLE daily(
exchangename STRING,
symbol          STRING,
date            STRING,
opencloseadj    MAP<STRING, FLOAT>,
highlow         STRUCT<high:FLOAT, low:FLOAT>,
volume          INT
)
ROW FORMAT DELIMITED
    FIELDS TERMINATED BY '\t'
    COLLECTION ITEMS TERMINATED BY '#'
    MAP KEYS TERMINATED BY '*'
    LINES TERMINATED BY '\n'
```

STORED AS TEXTFILE;

10. Explain the following with examples.

a.External Tables

- Hive does NOT assume that it owns the data.
- Data is NOT moved from its location in HDFS to the table's storage directory. The data remains where it was.
- Metadata for the table gets updated in the metastore.
- When a table is dropped, actual data is NOT removed, only metadata is removed.
- You can copy the schema of an external table to create another table.

CREATE TABLE nyse.dailycopy LIKE nyse.daily;

Input Data:

Exchange	Symbol	Date	dividend
NYSE	AIT	11/12/2009	0.15
NYSE	AIT	8/12/2009	0.15
NYSE	AIT	5/13/2009	0.15

```
CREATE EXTERNAL TABLE dividends(
exchangename STRING,
symbol          STRING,
date            STRING,
dividend        FLOAT
)
```

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t'

LINES TERMINATED BY '\n'

LOCATION '/hive/nyse/dividends/c7v4';

When to use External Tables?

- To query data stored in external systems such as Amazon S3.
- Avoids bringing in that data into HDFS.
- If data to be queried is also used by other tools such as Pig.
- If multiple tables or views are created on the same data.
- Avoids data getting deleted when one table is dropped.

b. Querying HIVE Tables

- Most Hive queries trigger MapReduce jobs.

Does NOT trigger MapReduce jobs:

SELECT * FROM companies;

SELECT * FROM companies LIMIT 1;

Triggers MapReduce jobs:

SELECT * FROM companies WHERE name='Oracle';

SELECT name, marketcap FROM companies;

SELECT * FROM companies;

SELECT * FROM companies LIMIT 1;

- Looks for the entire contents of the file.

- Hive can easily read the contents and provide results.

SELECT * FROM companies WHERE name=’Oracle’;

SELECT name, marketcap FROM companies;

Apply processing on the underlying data:

- Get table metadata
- File format of data stored
- InputFileFormat to read data
- For each row, extract the required columns’ data
- WHERE clause condition satisfied or not

Extract data from Collection Data Types

SELECT name, countries, cxos, hqaddress FROM companies;

countries ARRAY

cxos MAP

hqaddress STRUCT