1. How to load the file and print results in Pig Latin script?

records = LOAD 'Movie\_Log.txt' AS (login\_time:String, logout\_time:String, user\_name:String……);

Dump records;

1. How to get the list of movies grouped by ratings.

Group\_data = GROUP records BY ratings;

1. What is the primitive “DESCRIBE” in Pig Latin?

DESCRIBE is to print the information about the file, The describe operator is used to view the schema of a relation.

1. Extend the previous script to process the clickstream data into user sessions.

ClickStream = LOAD ‘Movie\_Log.txt’ USING org.apache.pig.piggybank.stroage.avro.AvroStroage(‘no\_schema\_check’, ‘schema\_file’, ‘hdfs://bigdatalite.cocaldomain/user/oracle/moviework/schemas/activity.avsv’);

1. How can I use FOREACH statement in Pig Script?

FOREACH operator is used to generate the specified data transformation based on column data.

Such as foreach\_data = FOREACH order\_by\_data GENERATE completed \_movie;

1. How to get the top-rated movies in each group.

order\_by\_data = ORDER records BY rated\_ point ASC;

foreach\_data = FOREACH order\_by\_data GENERATE completed \_movie;

Dump foreach\_data;

1. Select only the clicks, which correspond to starting, browsing, completing, or purchasing movies.

SELECT login\_time, incompleted \_movie, completed \_movie, purchased \_movie FROM records

Exercise 2

1. Describe the structure of students file in Pig Latin script.

records = LOAD ' students.csv' AS (……);

DESCRIBE records;

1. Write a pig script to assign names to the data fields of the students.csv data. The output file should be called “students\_details”.

students\_details = JOIN names BY id, records BY id;

Dump students\_details;

1. Perform any 3 operations on the file “students\_attendance.csv”. The output of the 3rd operation in this case should be called “SA\_details”.

attendance = LOAD ' students\_attendance.csv ' AS (……);

First = FILTER attendance BY hours > 5;

Second = ORDER First BY hours ASC;

SA\_details = FOREACH Second GENERATE names, hours;

1. Write a script to find the sum of hours attended by each student.

Hours\_sum = FOREACH SA\_details GENERATE names, hours, SUM(hours);

1. Write a script to join StudentID, Name, hours attended.

Students\_join = JOIN StudentID BY id, Name BY id, hours\_ attended BY id;

1. Print the results on the screen.

Dump Students\_join;