**QUESTION 1**

1. What is the difference between highest and lowest level of the team has reached in gameclicks ? (hint: user MAX and MIN operation in postgres).

|  |  |  |
| --- | --- | --- |
|  | A. | 5 |
|  | B. | 6 |
|  | C. | 8 |
|  | D. | 7 |

**0.5 points**

**QUESTION 2**

1. How many user id's (repeats allowed) have reached more than average level of team?

|  |  |  |
| --- | --- | --- |
|  | A. | 755806 |
|  | B. | 0 (ZERO) |
|  | C. | 358220 |
|  | D. | 397586 |

**0.5 points**

**QUESTION 3**

1. How many user id's (repeat allowed) reached the highest level in gameclicks and also clicked the highest costing price in buyclicks?

|  |  |  |
| --- | --- | --- |
|  | A. | 23301 |
|  | B. | 32747 |
|  | C. | 73226 |
|  | D. | 66887 |

**0.5 points**

**QUESTION 4**

1. What the following line of code do in postgres?

SELECT COUNT(\*) FROM (SELECT buyclicks.userID, teamlevel, price FROM buyclicks JOIN gameclicks ON buyclicks.userID = gameclicks.userId) temp WHERE price = 20 and teamlevel = 1;

|  |  |  |
| --- | --- | --- |
|  | A. | count the users who exists between both gameclicks and buyclicks tables |
|  | B. | This is an invalid line of code, the sub-query in not formatted properly |
|  | C. | Display the users who have bought item worth 20 and have had a team with level 1 |
|  | D. | Find the total number of user ids (repeat allowed) in buyclicks and have bought items with price worth 20 and was in a team with level 1 at some point of time |

**0.5 points**

**QUESTION 5**

1. use sample database (db)

in users table how many records are there? (count the records)

|  |  |  |
| --- | --- | --- |
|  | A. | 11188 |
|  | B. | 22288 |
|  | C. | 11177 |
|  | D. | 11199 |

**0.5 points**

**QUESTION 6**

1. In MongoDB data set, what is the username of the twitter account who has a tweet\_followers\_count of axactly 8973882

|  |  |  |
| --- | --- | --- |
|  | A. | Autocenterit |
|  | B. | SasSpear |
|  | C. | FIFAcom |
|  | D. | CreateImage |

**0.5 points**

**QUESTION 7**

1. Fins the tweet\_ID where user\_name is "ActionSportsJax"

|  |  |  |
| --- | --- | --- |
|  | A. | 521531393 |
|  | B. | 757667800 |
|  | C. | 757667800521531393 |
|  | D. | 123456789123456789 |

**0.5 points**

**QUESTION 8**

1. count how many tweets are there with "India" in tweet\_text

|  |  |  |
| --- | --- | --- |
|  | A. | 15 |
|  | B. | 18 |
|  | C. | 17 |
|  | D. | 16 |

**0.5 points**

**QUESTION 9**

1. Count how many tweet are there in db.users with "Cricket" in tweet\_text?

|  |  |  |
| --- | --- | --- |
|  | A. | 5 |
|  | B. | 6 |
|  | C. | 4 |
|  | D. | 3 |

**0.5 points**

**QUESTION 10**

1. how many tweets are there with "Cricket" in tweet\_text and tweet\_menstioned\_count greater than 1

|  |  |  |
| --- | --- | --- |
|  | A. | 0 |
|  | B. | 1 |
|  | C. | 3 |
|  | D. | 2 |

**0.5 points**

**QUESTION 11**

1. You want to ensure that the folder on which a Hive table is based in not deleted when the table is dropped. What should you do?

|  |  |  |
| --- | --- | --- |
|  | A. | Specify LOCATION keyword in the CREATE TABLE statement |
|  | B. | Create the table based on an existing folder |
|  | C. | Specify the INTERNAL keyword in CREATE TABLE statement |
|  | D. | Specify the EXTERNAL keyword in the CREATE TABLE  statement. |

**0.5 points**

**QUESTION 12**

1. The '/data/source' folder in Hadoop distributed file system (hdfs) contains multiple files containing unstructured text data. You run the following CREATE TABLE statement:

CREATE TABLE sourcedata (col1 INT, col2 STRING)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ' '

STORED AS TEXTFILE LOCATION '/data/source';

which of the following outcomes will result from the CREATE TABLE statement?

|  |  |  |
| --- | --- | --- |
|  | A. | The statement will succeed. Rows in the existing files in '/data/source' that do not match the table schema will be deleted. |
|  | B. | The statement will fail because the folder already contains data files |
|  | C. | The statement will succeed. The table sourcedata will be created in the '/data/source' . No existing files will be deleted. |
|  | D. | The statement will succeed. All existing files in '/data/source' will be deleted |

**0.5 points**

**QUESTION 13**

1. You need to create a Hive table for tab-delimited data file that will be stored in the '/data/calls' folder. The data files contains telephone call records in the following format:

2015-08-01-10:03:00       555-123-4567         22.3

2015-08-01:11:22:00       555-098-7654         2.7

2015-08-01:11:53:00       555-111-6543         17.1

The table must be named "calls", and the data files must be deleted if the table is dropped.  
  
Which CREATE TABLE statement should you use to create the table?

|  |  |  |
| --- | --- | --- |
|  | A. | CREATE TABLE calls (calltime STRING, phonenumber STRING, duration DECIMAL); |
|  | B. | CREATE EXTERNAL TABLE calls (calltime STRING, phonenumber STRING, duration DECIMAL)  ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'  STORED AS TEXTFILE LOCATION '/data/calls'; |
|  | C. | CREATE TABLE calls (calltime STRING, phonenumber STRING, duration DECIMAL)  ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'  STORED AS TEXTFILE LOCATION '/data/calls'; |
|  | D. | CREATE TABLE calls (calltime STRING, phonenumber STRING, duration DECIMAL)  ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'; |

**1 points**

**QUESTION 14**

1. You write the following Pig Latin script (line numbers are included for reference only)

01  A = LOAD '/source' USING PigStorage(' ') AS (c1;chararray, c2:long);

02  B = FILTER A BY c2 IS NOT NULL;

03  DUMP B;

04  C = ORDER B BY c1;

05 STORE C INT) '/output';

When you run the script, on which lines will Pig generate mapreduce jobs?

|  |  |  |
| --- | --- | --- |
|  | A. | Line 01 |
|  | B. | Line 04 |
|  | C. | Line 02 |
|  | D. | Line 05 |
|  | E. | Line 03 |

**1 points**

**QUESTION 15**

1. The file '/data/values.txt' contains the following text, which consists of a line containing column headings followed by multiple lines of data:

item     value

1         12.3

2         28.1

3         19.5

you run the following Pig Latin statement to load the contents of the file into a relation named data:

data = LOAD '/data/values.txt' USING PigStorage('\t') AS (item: chararray, value:float);

Which of the following represents the values in the first tuple in the data relation?

|  |  |  |
| --- | --- | --- |
|  | A. | (item, ) |
|  | B. | (item, 12.3) |
|  | C. | (01, 12.3) |
|  | D. | (item, value) |

**1 points**

**QUESTION 16**

1. You upload a Python script named convert.py to the '/scripts' folder in the HDFS

You write the following HiveQL code to use the Python script a a user-defined function when querying as existing table named data. (line numbers are included for reference only)

01

02  SELECT TRANSFORM (col1. col2, col3)

03  USING 'python.exe convert.py' AS

04  (col1 STRING, col2 INT, col3 STRING)

05  FROM data

06  ORDER BY col1;

Which statement should you include in line 01?

|  |  |  |
| --- | --- | --- |
|  | A. | SET hive.execution.engine=tez; |
|  | B. | SHOW TABLES; |
|  | C. | LOAD DATA INPATH '/scriptds/convert.py' INTO TABLE data; |
|  | D. | ADD FILE hdfs://10.0.2.15/scripts/convert.py; |