Question 1

1  
point

**1. Question 1**

What does it mean for a query language to be declarative?



**The language specifies what data to obtain.**



The language specifies both the process of how to obtain the data and specifies what data to obtain.



A language specific declaration of data types in order to define the method of data retrieval.



The language specifies the process of how to obtain the data.

Question 2

1  
point

**2. Question 2**

Use the following table named "user\_table" to answer the next 2 problems.

|  |  |  |
| --- | --- | --- |
| userId | username | email |
| 1 | admin | admin@corporate.moe |
| 2 | h4xor | 1337@rawr.cte |

How would you go about querying the entire username column (however many)?



SELECT username FROM user\_table WHERE userId=1



SELECT user\_table FROM username



SELECT username FROM userId WHERE \*



**SELECT username FROM user\_table**

Question 3

1  
point

**3. Question 3**

How would you go about querying the entire database table (please refer to question 2's table)?



SELECT user\_table FROM \*



SELECT username, email FROM userId



**SELECT \* FROM user\_table**



SELECT \* FROM \* WHERE user\_table

Question 4

1  
point

**4. Question 4**

What is the global indexing table?



**An index table in order to keep track of a given data type that might exist within multiple machines.**



An index table in order to keep track of data records within one machine.



A global table that uses a specific technique called indexing and the table uses an index as the primary key.



An index table in order to keep track of a given data type that might exist within one machine.

Question 5

1  
point

**5. Question 5**

What are the three computing steps of a semi-join?



Project, Ship, Reduce



**Project, Decompose, Send**



Index, Join, Display



Query, Join, Display



None Applicable

Question 6

1  
point

**6. Question 6**

What is the purpose of a semi-join?



Increase the speed of the join for trade-off of increased data transmission cost.



Another name for join: an operation to combine two tables by column.



**Increase the efficiency of sending data across multiple machines.**

Question 7

1  
point

**7. Question 7**

What is a subquery?



**A query statement within another query.**



A short query than normal.



An alternative query that acts as a substitute for another query.

Question 8

1  
point

**8. Question 8**

What is a correlated subquery?



**A type of query that contains a subquery that requires information from a query one level up.**



A type of query that contains a relationship between a variable attribute x and a variable attribute y. The two variables have a dependent relationship causing a correlation.



A type of query that requires two tables in order to calculate values.

Question 9

1  
point

**9. Question 9**

What is the purpose of GROUP BY queries?



Enables queries within queries.



**Enables calculations based on specific columns of the table.**



Required before you can use functions like AVG, SUM, MIN, MAX, COUNT.

Question 10

1  
point

**10. Question 10**

Consider the following generic statement for questions 10-12:

db.<collection>.find(<query filter>, <projection>).<cursor modifier>

Which part of the statement would reflect that of the FROM statement in SQL as illustrated in the lecture?



<projection>



<query filter>



<cursor modifier>



**<collection>**

Question 11

1  
point

**11. Question 11**

Which part of the statement would reflect that of the SELECT statement in SQL as illustrated in the lecture?



<query filter>



<collection>



<cursor modifier>



**<projection>**

Question 12

1  
point

**12. Question 12**

Which part of the statement would reflect that of the WHERE statement in SQL as illustrated in the lecture?



<collection>



<projection>



<cursor modifier>



**<query filter>**

Question 13

1  
point

**13. Question 13**

A sample part of the data structure is as follows:

{ \_id:1, userIndex: 10, email: “arealeamil@notreallu.asd", retainRate:2}

What would be the most likely statement that we would need to grab email info for user indexes greater than 24?



db.userIndex.find({email:{$gt:24}}, {\_id:0})



db.email.find({userIndex:{$lte:24}}, {email:1, \_id:0})



db.userIndex.find({email:{$lte:24}}, {\_id:0})



**db.email.find({userIndex:{$gt:24}}, {email:1, \_id:0})**

Question 14

1  
point

**14. Question 14**

What does it mean to have a \_id:0 within our query statement?



Grab as many objects as possible.



Grab the first object in the results.



Does not have an effect, simple convention left for compatibility issues.



**Tell MongoDB not to return a document id.**

This quiz encompasses data and content from Week 1 and 2, so we recommend reviewing that material from last week for this quiz as well. What is the highest level that the team has reached in gameclicks? (Hint: use the MAX operation in postgres).



9



6



8

**Correct**



7



10

Question 2

Correct

1 / 1 points

**2. Question 2**

How many user id's (repeats allowed) have reached the highest level as found in the previous question? (Hint: For postgres: you may either use two queries or use a sub-query).



122757



67271



51294

**Correct**



98823



106436

Question 3

Correct

1 / 1 points

**3. Question 3**

How many user id’s (repeats allowed) reached the highest level in game-clicks and also clicked the highest costing price in buy-clicks? Hint: Refer to question 4 for ideas.



73226



32747

**Correct**



66887



23301

Question 4

Correct

1 / 1 points

**4. Question 4**

What does the following line of code do in postgres?

*SELECT count(userid) FROM (SELECT buyclicks.userId, teamLevel, price FROM buyclicks JOIN gameclicks on buyclicks.userId = gameclicks.userId) temp WHERE price=3 and teamLevel=5;*



Counts the users who exists between both gameclicks and buyclicks files.



Finds the total number of user ids (repeats allowed) in buy-clicks that have bought items with prices worth $3 and was in a team with level 5 at some point in time.



Displays the users who have bought items worth $3 and have had a team with level 5.



This is an invalid line of code, the subquery is not formatted properly.

Question 5

Correct

1 / 1 points

**5. Question 5**

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



Autocenterit



FIFAcom



SasSpear



Create

How many tweets have location not null?



6973



6945

**This should not be selected**



No option applicable.



6937



5957

Question 2

Correct

1 / 1 points

**2. Question 2**

How many people have more followers than friends? (Hint : use this.user instead of user).



5809

**Correct**



6673



5590



6238



5206

Question 3

0.60 / 1 points

**3. Question 3**

Perform a query that returns the text of tweets which have the string "http://". Which of the following substrings do NOT occur in the results? (Choose all that apply)



@Infosmessi\_

**Correct**



@Ass0Star

**This should not be selected**



@DundalkFC

**This should be selected**



@espn

**Un-selected is correct**



@TerraceImages

**Un-selected is correct**

Question 4

Incorrect

0 / 1 points

**4. Question 4**

Query: Return all the tweets which contain text "England" but not "UEFA". In these results the string “Euro 2016” appears in...



More than 6 tweets.



2 tweets



0 tweets



3 tweets

**This should not be selected**



5 tweets

Question 5

Incorrect

0 / 1 points

**5. Question 5**

Query: Get all the tweets from the location "Ireland" which also contain the string "UEFA". In this result the user with the highest friends count is...



ProfitwatchInfo



DerekRantsGames

**This should not be selected**



irishexaminer



Insight4News4



Pauldonaghue

How many different countries are mentioned in at least one tweet?



64



44



211



112

Question 2

1  
point

**2. Question 2**

How many times is any country mentioned in a tweet?



397



52



26634



211

Question 3

1  
point

**3. Question 3**

What are the three countries with the highest mentioned count



Thailand, Mexico, Denmark



Nigeria, Slovakia, Germany



Thailand, Iceland, Mexico



Norway, Nigeria, France

Question 4

1  
point

**4. Question 4**

How many times was France mentioned in a tweet?



42



25



8



30

Question 5

1  
point

**5. Question 5**

Which country was mentioned most: Kenya, Wales, or Netherlands?



Netherlands



Kenya



Wales

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*M 2\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**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; |