### QUESTION 1

1. A Hive table named **staging** contains supplier data. An extract of the table is shown here:

|  |  |  |
| --- | --- | --- |
| supplierid | suppliername | country\_region |
| 1 | Contoso, Ltd. | United States |
| 2 | Northwind Traders | Canada |
| 3 | Adventure Works Cycles | United States |

1. You create a partitioned table for the supplier data by running the following CREATE TABLE statement:
2. CREATE TABLE suppliers
3. (supplierid INT,
4. suppliername STRING)
5. PARTITIONED BY (country\_region STRING);
6. You need to insert only the rows in the **staging** table with a **country\_region** value of 'United States' into the **suppliers** table.
7. Which code should you use?
8. Which code should you use?

|  |  |  |
| --- | --- | --- |
|  | 1. | -- Option 3  INSERT INTO TABLE suppliers   PARTITION(country\_region='United States')  SELECT supplierid, suppliername  FROM staging  WHERE country\_region = 'United States'; |
|  | 2. | -- Option 1  INSERT INTO TABLE suppliers  SELECT supplierid, suppliername, country\_region  FROM staging  WHERE country\_region = 'United States'; |
|  | 3. | -- Option 4  SET hive.exec.dynamic.partition = true;  SET hive.exec.dynamic.partition.mode=nonstrict;  INSERT INTO TABLE suppliers   PARTITION(country\_region)  SELECT supplierid, suppliername  FROM staging; |
|  | 4. | - Option 2  INSERT INTO TABLE suppliers   PARTITION(country\_region='United States')  SELECT supplierid, suppliername  FROM staging; |

**2 points**

### QUESTION 2

1. Refer Question 21 doc file which is attached

Refer graph (with node c as a spider trap) in file which is uploaded doc file as Question 21. There are links from page A to pages B, C, and D and so on and so forth. Calculate Page rank for each of the pages A, B, C and D using Column adjacency matrix using power iteration method. [use excel or R for calculation]

|  |  |
| --- | --- |
| Selected Answer: | [Page\_Ranking\_Algorithm.zip](https://blackboard.svkm.ac.in/courses/1/NMMPSMU_1718_CL_FMTEDSC_02_0D_01/attempt/_2838472_1/s/20344745c87541799ce5acbe38d7f8c2_Page_Ranking_Algorithm.zip?uploaded_filename=20344745c87541799ce5acbe38d7f8c2_Page_Ranking_Algorithm.zip)  [Remove](https://blackboard.svkm.ac.in/webapps/assessment/take/launch.jsp?course_assessment_id=_53072_1&course_id=_40648_1&content_id=_510582_1&isPasswordAttempted=true&password=cb7c6c9fabd8989d4a579fe25e381853&step=) |

* 1. Attach File

**5 points**

### QUESTION 3

1. You upload a Python script named convert.py to the /scripts folder in the HDFS container hosted in Microsoft Azure Storage.

You write the following HiveQL code to use the Python script as a user-defined function when querying an 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?

|  |  |  |
| --- | --- | --- |
|  | 1. | SHOW TABLES; |
|  | 2. | LOAD DATA INPATH '/scripts/convert.py' INTO TABLE data; |
|  | 3. | ADD FILE hdfs://10.0.2.15/scripts/convert.py; |
|  | 4. | SET hive.execution.engine=tez; |

**2 points**

### QUESTION 4

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?

|  |  |  |
| --- | --- | --- |
|  | 1. | (01, 12.3) |
|  | 2. | (item, value) |
|  | 3. | (item,) |
|  | 4. | (item, 12.3) |

**2 points**

### QUESTION 5

1. The '/data/source' folder 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?

|  |  |  |
| --- | --- | --- |
|  | 1. | The statement will succeed. Rows in the existing files in '**/data/source**' that do not match the table schema will be deleted. |
|  | 2. | The statement will succeed. All existing files in '**/data/source**' will be deleted. |
|  | 3. | The statement will succeed. The table will be created on the '**/data/source**' table. No existing files will be deleted. |
|  | 4. | The statement will fail because the folder already contains data files. |

**2 points**

### QUESTION 6

1. A Hive table named **call\_log** is based on the **/call\_log** folder, whch contains telehone call log files from a call center.

You regularly run the following HiveQL query, which returns the daily total call duration for each telephone extension.

SELECT call\_date, extension, SUM(duration) AS total\_minutes

FROM call\_log

GROUP BY call\_date, extension

ORDER BY call\_date, extension;

You want to enable business users to retrieve this data into Excel for analysis, but you would prefer to avoid them needing to enter the query each time. Instead, you would like users to be able to enter the following query:

SELECT \* FROM daily\_durations;

Which three HiveQL statements could you use to accomplish this goal?

|  |  |  |
| --- | --- | --- |
|  | 1. | -- Option 2  CREATE TABLE daily\_durations  (call\_date DATE,   extension INT   total\_minutes FLOAT)  STORED AS TEXTFILE LOCATION '/call\_log'; |
|  | 2. | -- Option 3  CREATE TABLE daily\_durations  AS  SELECT call\_date, extension, SUM(duration) AS total\_minutes  FROM call\_log  GROUP BY call\_date, extension  ORDER BY call\_date, extension; |
|  | 3. | -- Option 5  CREATE TABLE daily\_durations  (call\_date DATE,   extension INT   total\_minutes FLOAT);  INSERT INTO TABLE daily\_durations  SELECT call\_date, extension, SUM(duration)  FROM call\_log  GROUP BY call\_date, extension  ORDER BY call\_date, extension; |
|  | 4. | -- Option 1  CREATE VIEW daily\_durations  AS  SELECT call\_date, extension, SUM(duration) AS total\_minutes  FROM call\_log  GROUP BY call\_date, extension  ORDER BY call\_date, extension; |
|  | 5. | -- Option 4  CREATE TABLE daily\_durations  (call\_date DATE,   extension INT   total\_minutes FLOAT)  STORED AS TEXTFILE LOCATION '/daily\_durations';  LOAD DATA INPATH '/call\_log' INTO TABLE daily\_durations; |

**6 points**

### QUESTION 7

1. You use the following HiveQL code to create a Hive table:

CREATE TABLE sensor

(logdate DATE,

 DeviceID STRING,

 Reading INT)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ' '

STORED AS TEXTFILE LOCATION '/sensor';

You upload multiple text files containing sensor data to a folder named /staging in the hdfs storage container for your cluster. Then you run the following HiveQL statement:

LOAD DATA INPATH '/staging' INTO TABLE sensor;

What will the **/staging** folder contain after the LOAD statement runs?

|  |  |  |
| --- | --- | --- |
|  | 1. | Nothing. |
|  | 2. | Sensor data files that contain rows that do not match the sensor table schema. |
|  | 3. | All of the sensor data files. |
|  | 4. | A single file named **part-r-00000** that contains all of the sensor data as space-delimited text. |

**2 points**

### QUESTION 8

1. A folder named **/data/traffic** contains comma-delimited text files containing road traffic data. An extract from one of these files is shown here:

2015-08-01,09:01:27,A103,car,SW,33.2

2015-08-01,09:04:16,A103,bus,NE,21.6

2015-08-01,09:04:17,A103,car,NE,42.1

You need to use Pig to process the road traffic data. You must start by loading the data into a relation named **Traffic** in which tuples contain the following fields:

* + obs\_date
  + obs\_time
  + road
  + vehicle
  + direction
  + speed

Which Pig Latin statement should you use to load the data?

Which Pig Latin statement should you use to load the data?

|  |  |  |
| --- | --- | --- |
|  | 1. | -- Option 2  Traffic = LOAD '/data/traffic' USING PigStorage() AS  (obs\_date:chararray, obs\_time:chararray, road:chararray, vehicle:chararray, direction:chararray, speed:float); |
|  | 2. | -- Option 4  Traffic = LOAD '/data/traffic' USING PigStorage('\t') AS  (obs\_date:chararray, obs\_time:chararray, road:chararray, vehicle:chararray, direction:chararray, speed:float); |
|  | 3. | -- Option 3  Traffic = LOAD '/data/traffic' USING PigStorage(',') AS  (obs\_date:chararray, obs\_time:chararray, road:chararray, vehicle:chararray, direction:chararray, speed:float); |
|  | 4. | -- Option 1  Traffic = LOAD '/data/traffic'; |

**2 points**

### QUESTION 9

1. A Hive table named **staging** is based on the **/tables/staging** folder and contains staged comma-delimited text data in two columns named **itemid** and **itemvalue**. Some **itemvalue** values are null. You need to load rows in the **staging** table for which the **itemvalue** column has a non-null value into a new table named **items**. The new table should be based on the **/tables/items** folder and have the same schema as the **staging**table. The **items** table has not been created.

You need to perform the job with a single HiveQL statement.

Which HiveQL statement should you use?

Which HiveQL statement should you use?

|  |  |  |
| --- | --- | --- |
|  | 1. | -- Option 4  CREATE TABLE items  ROW FORMAT DELIMITED FIELDS TERMINATED BY ','  STORED AS TEXTFILE LOCATION '/tables/items'  AS  SELECT itemid, itemvalue  FROM staging  WHERE itemvalue IS NOT NULL; |
|  | 2. | -- Option 2  CREATE TABLE items  (itemid INT,   itemvalue INT)  ROW FORMAT DELIMITED FIELDS TERMINATED BY ','  STORED AS TEXTFILE LOCATION '/tables/staging'; |
|  | 3. | -- Option 1  LOAD DATA INPATH '/tables/staging' INTO TABLE items |
|  | 4. | -- Option 3  INSERT OVERWRITE TABLE items  SELECT itemid, itemvalue  FROM staging  WHERE itemvalue IS NOT NULL; |

**2 points**

### QUESTION 10

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 INTO '/output';

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

|  |  |  |
| --- | --- | --- |
|  | 1. | Line 04 |
|  | 2. | Line 05 |
|  | 3. | Line 02 |
|  | 4. | Line 03 |
|  | 5. | Line 01 |

**4 points**

### QUESTION 11

1. A folder named **/data/stock** contains tab-delimited text files of stock prices. An extract from one of these files is shown here:

|  |  |  |
| --- | --- | --- |
| quotetime | stock | price |
| 09:17:01 | XYZ | 48.67 |
| 09:17:03 | ABC | 14.21 |
| 09:17:07 | XYZ | 49.39 |

1. You need to create a Pig Latin script to process the stock price data. You start your script with the following statement:
2. Stock = LOAD '/data/stock' USING PigStorage('\t') AS
3. (time:chararray, stock:chararray, price:float);
4. You need to continue the script by creating a relation named **Data** in which the header row values, such as “quotetime”, have been removed from the **Stock** relation and only tuples containing actual time, stock, and price values remain.
5. Which statement should you add to the script?
6. Which statement should you add to the script?

|  |  |  |
| --- | --- | --- |
|  | 1. | -- Option 4  Data = FOREACH Stock GENERATE    FLATTEN(time), FLATTEN(stock), FLATTEN(price); |
|  | 2. | -- Option 2  Data = FOREACH Stock GENERATE    REPLACE(time, 'quotetime', ''), REPLACE(stock,'stock', ''), REPLACE(price, 'price', ''); |
|  | 3. | -- Option 3  Data = ORDER Stock BY time; |
|  | 4. | -- Option 1  Data = FILTER Stock BY time != 'quotetime'; |

**2 points**

### QUESTION 12

1. You write the following Pig Latin script:

East = LOAD 'data/east' USING PigStorage(',') AS (store:int, sales:float);

West = LOAD 'data/west' USING PigStorage(',') AS (store:int, sales:float);

Next, you need to write a statement to create a relation named **All** that has a store field and a sales field and contains all of the store and sales values in the **East** and **West** relations.

Which Pig Latin operator should you use?

Which Pig Latin operator should you use?

|  |  |  |
| --- | --- | --- |
|  | 1. | JOIN |
|  | 2. | UNION |
|  | 3. | SPLIT |
|  | 4. | GROUP |

**2 points**

### QUESTION 13

1. Your company stores archived web server log files in a cloud storage and plans to process the files on the last Saturday of each month to summarize the data they contain and load the results into a database for analysis. You plan to use Hadoop technologies in cloud to accomplish this. There are no other Hadoop processing jobs that need to be performed, and you need to process the log data every month while minimizing cost.

What should you do?

What should you do?

|  |  |  |
| --- | --- | --- |
|  | 1. | Provision a cloud cluster with a single data node, leave it running permanently. Dynamically scale up the number of data nodes on the last Saturday of the month to process the log data. |
|  | 2. | Provision an Cloud cluster with the maximum number of data nodes possible, leave it running permanently, and process the log data on the last Saturday of the month. |
|  | 3. | Provision multiple cloud clusters and divide the data processing across them, leave them running permanently, and process the log data on the last Saturday of the month. |
|  | 4. | Provision a cloud cluster on the last Saturday of the month, process the log files, and delete the cluster as soon as the data processing work is complete. |

**2 points**

### QUESTION 14

1. You need to create a Hive table for product data. Product records have a **category** field containing one of 25 possible category names.

You need to store the data files in a separate folder for each category.

Which CREATE TABLE statement should you use?

Which CREATE TABLE statement should you use?

|  |  |  |
| --- | --- | --- |
|  | 1. | -- Option 3  CREATE TABLE products  (id INT,   name STRING,   price DECIMAL,   category STRING)  SKEWED BY (category) ON ('name') STORED AS DIRECTORIES  ROW FORMAT DELIMITED FIELDS TERMINATED BY ','; |
|  | 2. | -- Option 2  CREATE TABLE products  (id INT,   name STRING,   price DECIMAL,   category STRING)  CLUSTERED BY (category) INTO 25 BUCKETS  ROW FORMAT DELIMITED FIELDS TERMINATED BY ','; |
|  | 3. | -- Option 4  CREATE TABLE products  (id INT,   name STRING,   price DECIMAL)  PARTITIONED BY (category STRING)  ROW FORMAT DELIMITED FIELDS TERMINATED BY ','; |
|  | 4. | -- Option 1  CREATE TABLE products  (id INT,   name STRING,   price DECIMAL,   category STRING)  ROW FORMAT DELIMITED FIELDS TERMINATED BY ','; |

**2 points**

### QUESTION 15

1. Find attached files:

1. Schema for traffic data

2. Actual traffic data in csv

Load the data and clean them up (by removing header column). Summarize (group) the data by year and then find numbers of passenger vehicles (car + buses and coaches) and distance cover in kilometer per year. (12 marks)

Write an UDF to convert yearly - kilometer to miles (5 marks)

submit intermediate result (befoer conversion) and final report (yearly data: year, milescovered in that year, NoOfPassengerVehicles  in that year ).

submit all python and pig files! (integration 2 mark)

* 1. Attach File

**19 points**

### QUESTION 16

1. If you create a hive table called table\_name without specifying any LOCATION, what happens?

|  |  |  |
| --- | --- | --- |
|  | 1. | Error message that location is not mentioned |
|  | 2. | Table is created at last location used in previous queries |
|  | 3. | nothing happens |
|  | 4. | Table is create at default location/table\_name |

**2 points**

### QUESTION 17

1. A Hive table named **temps** contains daily temperature readings (in Celsius) for each day between January 1st 2000 and December 31st 2010. An extract of the table is shown here:

|  |  |  |  |
| --- | --- | --- | --- |
| year | month | day | temp |
| 2000 | 1 | 1 | 7.1 |
| 2000 | 1 | 2 | 7.3 |
| ... |  |  |  |
| 2010 | 12 | 30 | 9.3 |
| 2010 | 12 | 31 | 10.1 |

1. You need to write a HiveQL query that summarizes the data to show the average temperature for each year and month combination, as shown in this example:

|  |  |  |
| --- | --- | --- |
| year | month | avgtemp |
| 2000 | 1 | 7.0 |
| 2000 | 2 | 7.6 |
| ... |  |  |
| 2010 | 11 | 12.3 |
| 2010 | 12 | 11.7 |

1. Which HiveQL SELECT statement should you use?

|  |  |  |
| --- | --- | --- |
|  | 1. | -- Option 2  SELECT year, month, AVG(temp) AS avgtemp  FROM temps  GROUP BY year, month  ORDER BY year, month; |
|  | 2. | -- Option 4  SELECT year, month, temp AS avgtemp  FROM temps  GROUP BY year, month; |
|  | 3. | -- Option 3  SELECT year, month, temp AS avgtemp  FROM temps  ORDER BY year, month; |
|  | 4. | -- Option 1  SELECT year, month, AVG(temp) AS avgtemp  FROM temps  ORDER BY year, month; |

**2 points**

### QUESTION 18

1. A Hive table named **staging** contains patients’ medical data. An extract of the table is shown here:

|  |  |  |
| --- | --- | --- |
| patient\_id | Hospital\_id | physician\_id |
| 1034 | 1209-123 | 512462 |
| 2129 | 1209-123 | 872537 |
| 1732 | 1612-098 | 915267 |

1. You create a partitioned table for the patient data by running the following CREATE TABLE statement:
2. CREATE TABLE patients
3. (patient\_id INT,
4. physician\_id INT)
5. PARTITIONED BY (hospital\_id INT);
6. You need to insert all rows in the **staging** table into the **patients** table.
7. Which code should you use?
8. Which code should you use?

|  |  |  |
| --- | --- | --- |
|  | 1. | -- Option 3  INSERT INTO TABLE patients  PARTITION(hospital\_id)  SELECT patient\_id, physician\_id  FROM staging; |
|  | 2. | -- Option 1  INSERT INTO TABLE patients  SELECT patient\_id, physician\_id, hospital\_id  FROM staging; |
|  | 3. | -- Option 4  SET hive.exec.dynamic.partition = true;  SET hive.exec.dynamic.partition.mode=nonstrict;  INSERT INTO TABLE patients  PARTITION(hospital\_id)  SELECT patient\_id, physician\_id, hospital\_id  FROM staging; |
|  | 4. | -- Option 2  INSERT INTO TABLE patients  PARTITION(hospital\_id)  SELECT patient\_id, physician\_id, hospital\_id  FROM staging;  incorrect |

**2 points**

### QUESTION 19

1. You need to create a Hive table for tab-delimited data files that will be stored in the **/data/calls** folder. The data files contain 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?

|  |  |  |
| --- | --- | --- |
|  | 1. | CREATE TABLE calls  (calltime STRING,   phonenumber STRING,   duration DECIMAL)  ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'  STORED AS TEXTFILE LOCATION '/data/calls'; |
|  | 2. | CREATE TABLE calls  (calltime STRING,   phonenumber STRING,   duration DECIMAL); |
|  | 3. | CREATE EXTERNAL TABLE calls  (calltime STRING,   phonenumber STRING,   duration DECIMAL)  ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'  STORED AS TEXTFILE LOCATION '/data/calls'; |
|  | 4. | CREATE TABLE calls  (calltime STRING,   phonenumber STRING,   duration DECIMAL)  ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'; |

**2 points**

### QUESTION 20

1. You need to create a Hive table on a folder that will be used to store comma-delimited text files. If the table is dropped, the folder and its contents must not be deleted.

Which CREATE TABLE statement should you use?

Which CREATE TABLE statement should you use?

|  |  |  |
| --- | --- | --- |
|  | 1. | -- Option 1  CREATE TABLE sourcedata  (col1 STRING,   Col2 INT)  ROW FORMAT DELIMITED FIELDS TERMINATED BY ','; |
|  | 2. | -- Option 3  CREATE INTERNAL TABLE sourcedata  (col1 STRING,   Col2 INT)  ROW FORMAT DELIMITED FIELDS TERMINATED BY ','  STORED AS TEXTFILE LOCATION '/data'; |
|  | 3. | -- Option 4  CREATE EXTERNAL TABLE sourcedata  (col1 STRING,   Col2 INT)  ROW FORMAT DELIMITED FIELDS TERMINATED BY ','  STORED AS TEXTFILE LOCATION '/data'; |
|  | 4. | -- Option 2  CREATE TABLE sourcedata  (col1 STRING,   Col2 INT)  ROW FORMAT DELIMITED FIELDS TERMINATED BY ','  STORED AS TEXTFILE LOCATION '/data'; |

**2 points**

### QUESTION 21

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

|  |  |  |
| --- | --- | --- |
|  | 1. | Specify the LOCATION keyword in the CREATE TABLE statement. |
|  | 2. | Create the table based on an existing folder. |
|  | 3. | Specify the EXTERNAL keyword in the CREATE TABLE statement. |
|  | 4. | Specify the INTERNAL keyword in the CREATE TABLE statement. |

**2 points**

### QUESTION 22

1. Is there any default file format in HIVE?

|  |  |  |
| --- | --- | --- |
|  | 1. | the hive.default.fileformat  configuration parameter determines the default file format |
|  | 2. | TAB separated file |
|  | 3. | CSV (comma separated files) |
|  | 4. | no default file |