SIM GLOBAL EDUCATION

UOW AUSTRALIA

Family Name .....

First Name .....

SCIT

School of Computing and

Information Technology

ISIT312
Big Data Management

This paper is for students studying at the Singapore Institute of Management Pte Ltd.

S4-2021 Final Examination

Date: 30 November, 2021

Time: 2.15 pm – 5.15 pm SGT + 30 minutes submission time

Marks available: **40 marks.**

DIRECTIONS TO CANDIDATES

(1) The answers to the questions included in the final examination must be hand-written with a BLACK or DARK BLUE PEN on the WHITE PIECES of paper. No pencil and no other colour of paper is allowed.

(2) When finished, take the pictures of the hand-written solution, save the pictures in files (`pdf`, `jpeg`, `jpg`, `gif`, `bmp`, `png`, `tiff` formats are all acceptable), and submit the files through Moodle. Using mobile phone cameras is all right. It is possible to take more than one picture per answer to assure the good readability of an answer. The marks will be deducted for submissions in the different formats. No more than 20 files can be submitted and no more than 200Mbytes can be submitted. Please plan well your pictures.

(3) The files should have the names indicating a number of the respective question in the final examination paper like `q1`, `q2`, ... and `q1-1`, `q1-2`, ... when more than one picture is used for an answer of a question. It will help you to avoid a submission of a wrong file or submission of the same file twice.

(4) All answers including the drawings must be hand-written. No printed material will be evaluated. No iPad or other tablet is allowed. The solutions must be hand-written on the pieces of paper. Submission of the typed and/or electronically processed text is a violation of the final/deferred/supplementary examination regulations and it will be considered as a medium level academic misconduct with all consequences coming from such fact.

(5) Marks will be deducted for the late submissions at a rate of 1 mark per 1 minute late.

Question 1 (8 marks)

Assume, that a text file `crime-stories.txt` contain the texts of the large number of crime stories. Assume, that the file is formatted such that one statement is located in one line of the text file.

Assume, that a text file `patterns.txt` contains the text patterns, for example regular expressions. Assume, that the file is formatted, such that one pattern is located in one line of the text file.

To simplify the problem, assume that all text patterns in a file `patterns.txt` are distinct.

Finally, assume that a function `match(text-line, text-pattern)` returns true when `text-line` matches a `pattern text-pattern`. Otherwise, the function returns false.

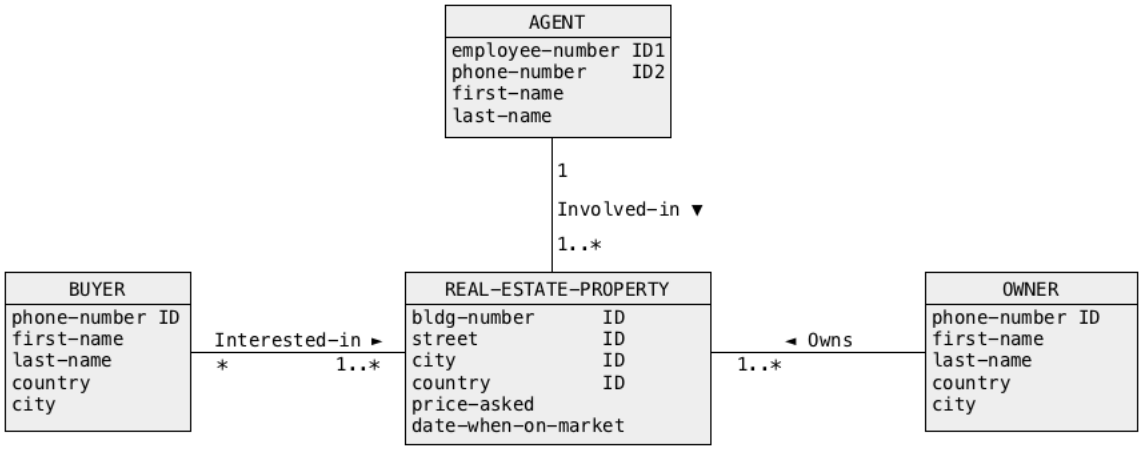
Your task is to explain how to implement a MapReduce application, that for each text pattern in a file `patterns.txt` finds the total number of statements in a file `crime-stories.txt` that match the pattern.

You must specify the parameters (if any) of your application and the key-value data in the input and output of the Map and Reduce stages.

There is no need to write Java code, however, if you like it then it is all right to do so. The precise explanations in plain English or in a pseudocode will do. Please note, that if you decide to use pseudocode then your explanations must precisely explain what happens at each stage of Map-Reduce application.

Question 2 (8 marks)

Consider the following conceptual schema of an operational database owned by a multinational real estate company. The database contains information about the real estate properties offered for sale, owners of the properties, potential buyers who are interested in the properties and real estate agents involved in selling of the properties.



Whenever a property is put on a market by an owner, a description of the property is entered into an operational database. Whenever a property is purchased, its description is removed from an operational database.

The real estate company would like to create a data warehouse to keep information about the finalized real estate transactions, properties involved in the transactions, sellers/owners, and agents involved in the real estate transactions. The real estate company would like to use a data warehouse to implement the following classes of analytical applications.

(1) Find the total number of real estate properties sold per month, year, street, city, country, agent involved.

(2) Find an average asked price of real estate properties sold per month, year, street, city, country, agent involved.

(3) Find an average final price of real estate properties sold per month, year, street, city, country, agent involved.

(4) Find an average period of time on the marked of real estate properties sold per month, year, street, city, country, agent involved.

(5) Find the total number of times each real estate property has been sold in a given period of time.

(6) Find the total number of buyers interested in purchases of real estate properties sold per day, month, year, street, city, country, agent involved.

Note, the operational database does not contain all information necessary to implement the classes of applications listed above. Additional information must be added when data is transferred from an operation database to a data warehouse.

Your task is to create a conceptual schema of the planned data warehouse. To draw a conceptual schema, use a graphical notation explained to you in a presentation 11 Conceptual Data Warehouse Design.

Question 3 (10 marks)

The text files `orders.txt` and `details.txt` contain information about the orders submitted by the customers and about the details of each order, i.e. the items included in each order. Each file consists of a line with a header and a number of lines with data. The sample rows included in the files are the following.

`orders.txt`
```
order,number,customer-number
0001,1234
0002,1234
0003,7890
... ...
```

`details.txt`
```
order-number,line-number,item-name,quantity
0001,01,bolt,25
0001,02,screw,20
0002,01,bolt,100
0002,02,nut,300
0003,01,trolley,1
... ...
```

The headers of the files have been removed and the files have been loaded to HDFS to a location `/bigdata/orders`. Assume, that we would we would like to retrieve from the files, information listed below. Write HQL (Hive) statements that implement the retrievals. An important constraint is such that we must not replicate data already loaded to HDFS.

(1) List the contents of the files `orders.txt` and `details.txt`.

(2 marks)

(2) Find the total quantities summarized per order number and customer number, per customer number, and the total quantity of all orders.

(2 marks)

(3) List the names of items, customer numbers and quantities of items ordered by each customer and sorted in the ascending order of the ordered quantities per each customer. Additionally, list the ranks of all items purchased by each customer.

(2 marks)

(4) Find an average quantity of ordered items per item name, and per order number and customer number.

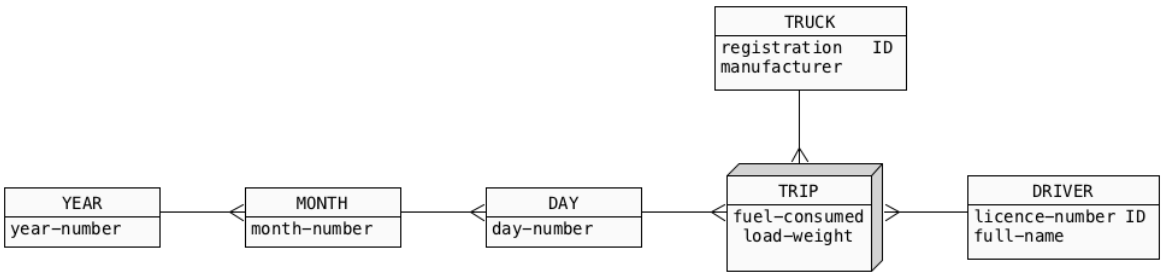
(2 marks)

(5) List the names of items partitioned by item name and the progressing summations of the quantities of items ordered in the ascending order of order numbers within each partition.

(2 marks)

Question 4 (8 marks)

Consider the following conceptual schema of a three-dimensional data cube.



(1) Use HBase shell command language to write the commands that create HBase table implementing a conceptual schema given above.

(1 mark)

(2) Write the commands of HBase shell command language that insert into HBase table created in the previous step information about 2 trucks, 2 drivers, and 2 trips.

(3 marks)

(3) Write the commands of HBase shell command language, that perform the following data retrieval and data manipulation operations on HBASE table.

- Find all information about a truck with a registration number `AL08UK`.

(1 mark)

- Find all information about the trips where a weight of load was higher than `100`.

(1 mark)

- Extend Hbase table with information about a depot of each truck. Record information about a full address of a depo and assume that each truck belongs to only one depot.

(1 mark)

- Extend Hbase table with information about the skills of drivers. Assume, that a skill is described by a name and a driver may have many skills. of course, as kill can be possessed by many drivers.

(1 mark)

Question 5 (4 marks)

The text files `orders.txt` and `details.txt` contain information about the orders submitted by the customers and about the details of each order, i.e. the items included in each order. Each file consists of a line with a header and a number of lines with data. The sample rows included in the files are the following.

`orders.txt`
```
order,number,customer-number
0001,1234
0002,1234
0003,7890
... ...
```

`details.txt`
```
order-number,line-number,item-name,quantity
0001,01,bolt,25
0001,02,screw,20
0002,01,bolt,100
0002,02,nut,300
0003,01,trolley,1
... ...
```

The headers of the files have been removed and the files have been loaded to HDFS to a location `/bigdata/orders`.

Write Pig-Latin statements that implement the following retrievals.

(1) Find the customer numbers (`customer-number`) of the customers who ordered at least one time an item `bolt` and quantity of such order was larger than 100.

(1 mark)

(2) Find the customer numbers (`customer-number`) of the customers who never ordered an item `bolt`.

(1 mark)

(3) Find the customer numbers (`customer-number`) of the customers who ordered an item `bolt` and an item `screw`.

(1 mark)

(4) Find the customer numbers (`customer-number`) together with the total number of submitted orders.

(1 mark)

Question 6 (3 marks)

The text files `orders.txt` and `details.txt` contain information about the orders submitted by the customers and about the details of each order, i.e. the items included in each order. Each file consists of a line with a header and a number of lines with data. The sample rows included in the files are the following.

`orders.txt`

```
order,number,customer-number
0001,1234
0002,1234
0003,7890
... ...
```

`details.txt`

```
order-number,line-number,item-name,quantity
0001,01,bolt,25
0001,02,screw,20
0002,01,bolt,100
0002,02,nut,300
0003,01,trolley,1
... ...
```

The headers of the files have been removed and the files have been loaded to HDFS to a location `/bigdata/orders`.

(1) Load the contents of a file `details.txt` located in HDFS into a Resilient Distributed Dataset (RDD) and use RDD to find the total quantity per each item.

(1 mark)

(2) Load the contents of a file `details.txt` located in HDFS into a Dataset and use the Dataset to find the total quantity per each item.

(1 mark)

(3) Load the contents of a file `details.txt` located in HDFS into a DataFrame and use SQL to find the total quantity per each item.

(1 mark)

End of Examination Paper