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.

S2-2022 Final Examination

Date: 6 June, 2022

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.

Copyright SCIT, University of Wollongong, 2022.

Question 1 (8 marks)

The World Meteorological Organization created a file `hightemp.txt` with information about the highest temperatures recorded every day in a number of cities all over the world. The file `higtemp.txt` is a text file where information about the highest temperature recorded on a given day, in a given city is stored in a single row. Data items like date, temperature, city name are separated with a single blank. For example, few sample rows from the file are listed below.

```
01-JAN-1991 25 Sydney
01-JAN-1991 30 Brisbane
01-JAN-1991 32 Singapore
     ...    ... ...
02-JAN-1991 25 Sydney
02-JAN-1991 31 Brisbane
02-JAN-1991 35 Singapore
     ...    ... ...
05-JUN-2022 15 Sydney
05-JUN-2022 20 Brisbane
05-JUN-2022 25 Singapore
```

(1) 6 marks

Write an implementation of Map-Reduce application that lists the highest temperature recorded together with a name of city where such temperature was recorded in a given year. Organize your application such that it would be possible to re-use it for any given year.

For example, if we process only 9 rows visible above then the highest temperature recorded in `1991` was `35` in `Singapore`.

Assume that due to the global warming no negative temperatures are recorded.

To write your application use Java programming language or pseudo-code at a level of complexity of Java statements. So, for example, you can write your application in Python or in some other pseudo-code that looks like Java. It is compulsory to include the comments explaining your code.

Please, write only the implementations of Mapper and Reducer.

(2) 2 marks

Assume that source code of your application is written in Java and it is available in a file `highest.java`. List all steps you should follow to get the outcomes from preparation and processing of your application. For example, to list the highest temperatures recorded, together with a city where such temperature was recorded in 2021.

Copyright SCIT, University of Wollongong, 2022.

Question 2 (7 marks)

A large international bank would like to create a data warehouse with information about the loans given to its customers.

A customer is described by a unique account number associated with a loan, full name and address.

The bank records the dates when the customers are provided with the loans and the dates when the loans are fully repaid. A date consists of a day, month and year.

The banks offer the following types of loans: home, investment, personal. Different types of loans are offered at different interest rates.

All loans must be insured at the insurance companies. An insurance company is described by a unique name and address.

The loans are issued by the tellers located at the branches. A description of a teller consists of a unique employee number and full name. A branch is described by a unique name.

The bank plans to use a data warehouse to implement the following classes of analytical applications.

(1) Find the total number of loans issued per day, per month, per year, per branch, per bank teller, per city, per state, per country, per loan type, per customer.

(2) Find the total amount of money loaned to the customers per day, per month, per year, per city, per country, per loan type.

(3) Find the total interest rates on the loans per day, per month, per year, per city, per country, per loan type.

(4) Find an average period of time needed for the loan repayment per loan type, per customer, per city, per country.

(5) Find the total number of different currencies used for the loans.

(6) Find the total amount of money on loans per currency.

Your task is to create a conceptual schema of a data warehouse needed by the bank. To draw a conceptual schema, use a graphical notation explained to you in a presentation 11 Conceptual Data Warehouse Design. Draw a conceptual schema by hand on a piece of paper, take a picture or scan it and submit an image.

Copyright SCIT, University of Wollongong, 2022.

Question 3 (10 marks)

In this question we use the same file `hightemp.txt` as in Question 1 of the examination paper. A file `hightemp.txt` contains information about the highest temperatures recorded every day in a number of cities all over the world. The file `higtemp.txt` is a text file where information about the highest temperature recorded on a given day, in a given city is stored in a single row. Data items like date, temperature and city name are separated with a single blank.

A file `hightemp.txt` has been loaded to HDFS to a location `/bigdata/hightemp`. We must not replicate data already loaded to HDFS.

(1) Write HQL statements that prepare a file `hightemp.txt` to be accessed through `SELECT` statements. Next, write `SELECT` statement that lists the total number of cities where the temperature measurements were recorded.

(2 marks)

(2) Write `SELECT` statement that lists the names of cities, temperature measurements and dates sorted in the ascending order of the temperature measurements per each city. Additionally, list a rank of each temperature measurement in each city.

(2 marks)

(3) Write `SELECT` statement that lists an average temperature per year and per year and city.

(2 marks)

(4) Write `SELECT` statement that lists an average temperature per year and city, per city, and an average temperature from all measurements.

(2 marks)

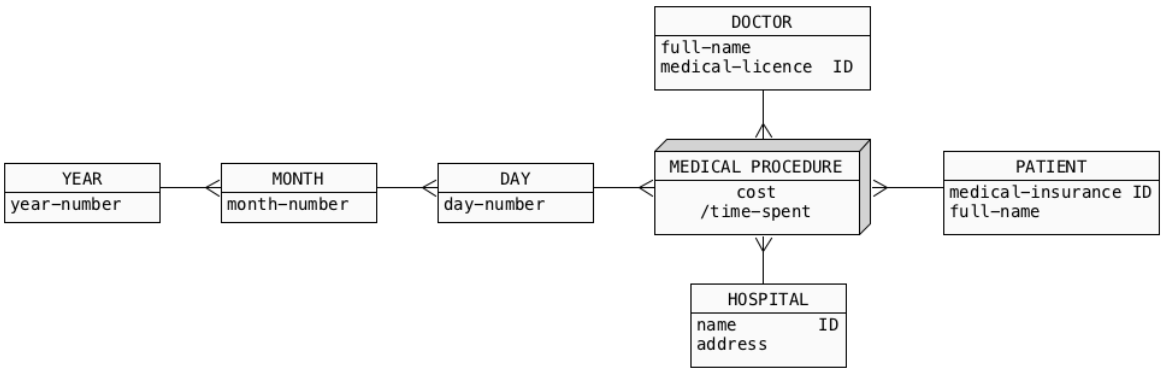
(5) Write `SELECT` statement that lists five month moving average of temperate measurements per each city.

(2 marks)

Copyright SCIT, University of Wollongong, 2022.

Question 4 (8 marks)

Consider the following conceptual schema of a four-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 patients, 2 doctors, 1 hospital, 2 medical procedures.

(3 marks)

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

- Find all information about a patient whose medical insurance number `MI6789`.

(1 mark)

- Find all information about the medical procedures where the costs were higher than `100`.

(1 mark)

(4) Extend Hbase table with information about the consultations between patients and doctors. A description of a consultation consists of a consultation date and consultation topic. Write the commands of HBase shell command language that adds information about two consultations.

(1 mark)

(5) Extend Hbase table with information about the specialisations of doctors. Assume, that a specialisation is described by a name and level. Write the commands of HBase shell command language that adds information about a specialisation of one doctor.

(1 mark)

Copyright SCIT, University of Wollongong, 2022.

Question 5 (4 marks)

In this question we use the same file `hightemp.txt` as in Question 1 of the examination paper. A file `hightemp.txt` contains information about the highest temperatures recorded every day in a number of cities all over the world. The file `higtemp.txt` is a text file where information about the highest temperature recorded on a given day, in a given city is stored in a single row. Data items like date, temperature and city name are separated with a single blank.

A file `hightemp.txt` has been uploaded to HDFS at a location `/bigdata/hightemp`.

Additionally, the World Meteorological Organization uploaded to HDSF at the same location a file `city.txt` that contains information about the cities where the temperature was recorded. The file `city.txt` is a text file with the names of cities and the names countries where the cities are located. Data items are separated with a single blank in each line. Few sample lines from a file `city.txt` are given below.

```
Brisbane Australia
Singapore Singapore
Sydney Australia
...    ...
```

Write Pig-Latin statements that implement the following retrievals.

(1) Find the distinct names of countries that recorded a temperature higher than 50 degrees.
(1 mark)

(2) Find the names of cities to together with the total number of temperature measurements recorded in each city. You can skip the names of cities where temperature was not recorded.
(1 mark)

(3) Find the names of countries, names of cities located in each country and temperatures recorded on `01-JAN-2020`. If a temperature has not been recorded in a city on `01-JAN-2020` then list only a name of city and a name of country.
(1 mark)

(4) Find the names of cities where a temperature has not been recorded on `01-JAN-2020`.
(1 mark)

Copyright SCIT, University of Wollongong, 2022.

Question 6 (3 marks)

In this question we use the same file `hightemp.txt` as in Question 1 of the examination paper. A file `hightemp.txt` contains information about the highest temperatures recorded every day in a number of cities all over the world. The file `higtemp.txt` is a text file where information about the highest temperature recorded on a given day, in a given city is stored in a single row. Data items like date, temperature and city name are separated with a single blank.

A file `hightemp.txt` has been uploaded to HDFS at a location `/bigdata/hightemp`.

(1) Load the contents of a file `hightemp.txt` located in HDFS into a Resilient Distributed Dataset (RDD) and use RDD to find an average temperature in `Sydney` in `2020`.

(1 mark)

(2) Load the contents of a file `hightemp.txt` located in HDFS into a Dataset and use the Dataset to find the total number of temperature measurements per city.

(1 mark)

(3) Load the contents of a file `hightemp.txt` located in HDFS into a DataFrame and use SQL to find an average temperature per city and per year and city.

(1 mark)

End of Examination Paper

Copyright SCIT, University of Wollongong, 2022.