UOW AUSTRALIA

School of Computing and Information Technology

Student to complete:

Family name	
Other names	
Student number	
Table number	

ISIT312
Big Data Management
Wollongong Campus

Examination Paper
Spring Session 2017

Exam duration	3 hours
Items permitted by examiner	None
Aids supplied	None
Directions to students	8 questions to be answered.

This examination is worth 60% of the total marks for the subject

This exam paper must not be removed from the exam venue

Question 1 (9 marks)

Read and analyse a specification of data warehouse domain listed below.

Create a conceptual schema for a sample data warehouse domain listed below. Use a graphical notation explained to you during the lecture classes in ISIT312 Big Data Management to draw the conceptual schema.

A group of real estate agencies would like to keep historical information about the real estate properties offered for sale, owners of the properties, buyers, agents employed by the agencies, prices asked and prices paid, and the details of all finalized transactions.

There are three types of real estate properties offered for sale: houses, flats and blocks of land. A house is described by an address that consists of city name, street name, and house number, and by the total number of bedrooms. A flat is described an address that consists of city name, street name, house number, and flat number, and the total number of bedrooms. A block of land is described by an address and area.

Sellers put their real estate properties for sale. Both, sellers and potential buyers are identified by a mobile phone number and described by the first and last name and present address.

Real estate agents are described by a name of agency they work for and a unique employee number.

A real estate transaction includes information about a real estate property sold, seller, buyer, agents involved, price asked, price paid, date when real estate properties have been put on a market, date when a transaction has been finalized, and the total number days a property has been on a market.

Your data warehouse must be designed such that it is possible to implement the following classes of applications.

(1) Find the total number of real estate transactions sold per year, then per month in a given year and then per week in a given month of a given year.

(2) Find the average price paid for the real estate properties per city and then per street of a given city.

(3) Find, the total number of real estate transactions per agent involved the transactions and later on per real estate agency involved in the transactions.

(4) Find the total amount of money paid for the real estate properties purchased per buyer, and per city a buyer is located at.

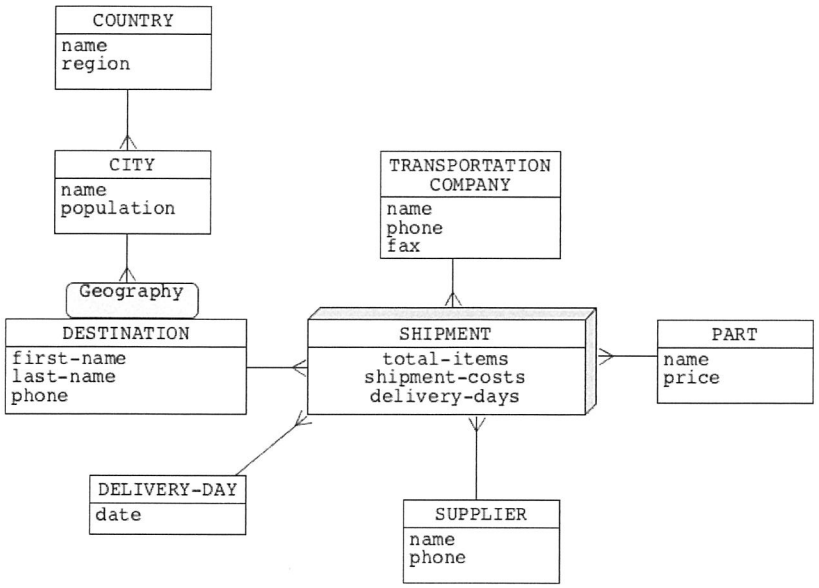
(5) Find the total amount of money earned for the sold real estate properties sold per seller, and per city a buyer is located at.

(6) Find the total number of real estate properties sold per type of real estate property (house, flat, block of land).

(7) Find the average time a real estate property was on a market per city where a real estate property is located at.

Question 2 (9 marks)

Consider the following conceptual schema of a sample data warehouse that supposed to contain historical information about shipments, transportation companies involved in shipments, parts shipped, shipment destinations, suppliers, and delivery days.



Your task is to implement the sample data warehouse as a collection of relational tables that can be used to store information modelled in a diagram above. To implement the data warehouse, perform a stage of logical modelling that transforms a conceptual schema given above into a collection of relational tables. List the names of relational tables and the attributes included in each table. For each relational table found, list the primary keys, candidate keys (if any) and foreign keys (if any).

There is no need to write `CREATE TABLE` statements of SQL! The names of relational tables together with the names of attributes and specifications of key constraints valid in each table are completely sufficient.

Question 3 (8 marks)

(1) Explain the main functions (or services) of the key HDFS components: `NameNode` and `DataNode`. For each component, you should provide at least three of the main functions (or services).

(4 marks)

(2) Explain the main components of the MapReduce data processing model.

(4 marks)

Question 4 (8 marks)

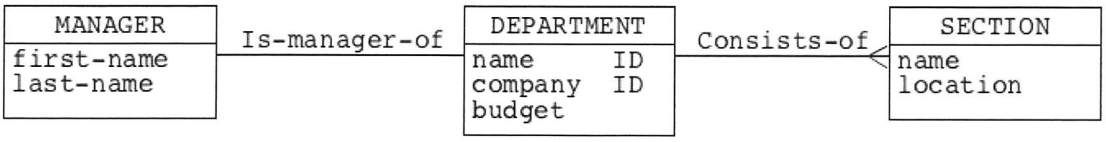
(1) Use one hypothetical example to explain the function of the combiner in the MapReduce model

(4 marks)

(2) Use one hypothetical example to explain why the combiner cannot be the same as the reducer in the MapReduce model for some application.

(4 marks)

Question 5 (7 marks)
Consider the following conceptual schema of a sample database domain.



(1) Write the commands of HBase shell command language that create HBase table implementing a sample database domain given above.

(3 marks)

(2) Write the commands of Hbase shell command language that insert into HBase table created in the previous step information about at least 2 departments such that each department has one manager and at least 2 sections. All other information is up to you.

(4 marks)

Question 6 (9 marks)

Consider the following datasets:

```
--salespeople.txt
--schema: salespersonid, name, storeid
1,Henry,100
2,Karen,100
3,Paul,101
4,Jimmy,102
5,Janice,

--stores.txt
--schema: storeid, name
100,Hayward
101,Baumholder
102,Alexandria
103,Melbourne

--sales.txt
--schema: salespersonid, storeid, salesamt
1,100,38
2,100,75
3,101,55
4,102,12
```

Suppose the above files `salespeople.txt`, `stores.txt`, and `sales.txt` have been uploaded to HDFS.

Write down the Pig-Latin commands that perform operations specified in (1), (2) and (3) below. For (2) and (3), also write down the output data.

(1) Load the three datasets by using the provided relation names and field names. The fields of each relation must have the suitable types.

(3 marks)

(2) Define a relation `salespeople_grouped` that groups `salespeople` by the `storeid`. Then dump `salespeople_grouped`.

(3 marks)

(3) Define a relation `stores_sales_innerjoin` that is the inner join of `stores` and `sales` by the `storeid`. Then dump `stores_sales_innerjoin`.

(3 marks)

Question 7 (5 marks)

Consider the following datasets (the same as in **Question 6** above):

```
--salespeople.txt
--schema: salespersonid, name, storeid
1,Henry,100
2,Karen,100
3,Paul,101
4,Jimmy,102
5,Janice,

--stores.txt
--schema: storeid, name
100,Hayward
101,Baumholder
102,Alexandria
103,Melbourne

--sales.txt
--schema: salespersonid, storeid, salesamt
1,100,38
2,100,75
3,101,55
4,102,12
```

(1) Draw a conceptual schema of a simple data warehouse that might contain data given above.

(1 mark)

(2) Explain how would you implement the data warehouse designed in step (1) using Apache Hive technology.

(4 marks)

Question 8 (5 marks)

Suppose there is a file `text.txt` in the root directory of HDFS.

The words in each line in `text.txt` are separated by space.

You are required to develop a Spark application in the Spark-shell to count the average number of words appearing in one line in `text.txt`.

The first line of the Scala program is provided for you as below. Please complete the program.

```
val lines = sc.textFile("text.txt")
...
```

Write down the codes as far as possible. Use annotations to explain the operations (i.e., transformations and actions) that you use if helpful.

Note. Minor typos in the operation names are acceptable, but wrong operations result in reduced marks.