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
Supplementary & Deferred Spring 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.

Accordingly to Wikipedia "Greasy pole or grease pole refers to a pole that has been made slippery and thus difficult to grip. Greasy pole climbing is a name of events that involve staying on, climbing up, walking over or otherwise traversing such a pole."

The competitions organized by World Tour Championships in Greasy Pole Climbing (WTC-GPC) include only climbing of vertical greasy poles. At the beginning of each yearly cycle WTC-GPC prepares a schedule of the competitions. A description of a competition includes a period of time when it is planned, location determined by country, city, and address of a venue, and entry fee to be paid after registration.

During a competition, the referees measure the time taken by a climber to climb from earth level to a platform at the top of the greasy pole. The climber with the shortest time wins the competition. Each competitor has 3 climbing attempts. A competitor is not required to complete in all three attempts. All competitors who fail to reach a platform in any attempt are disqualified and score no points no matter what were the results of the previous attempts. A very good result in the first attempt might mean a competitor would choose to skip the remaining attempts.

The final results of a competition include a description of each competitor, the outcomes of each one of the up to three climbing attempts made by each competitor, the total number of points scored in the competition by each competitor and prize money won by each competitor. The competitors are described by the first name, last name, date of birth and nationality. It is all right to assume that a pair of attributes full name and date of birth uniquely identifies each competitor.

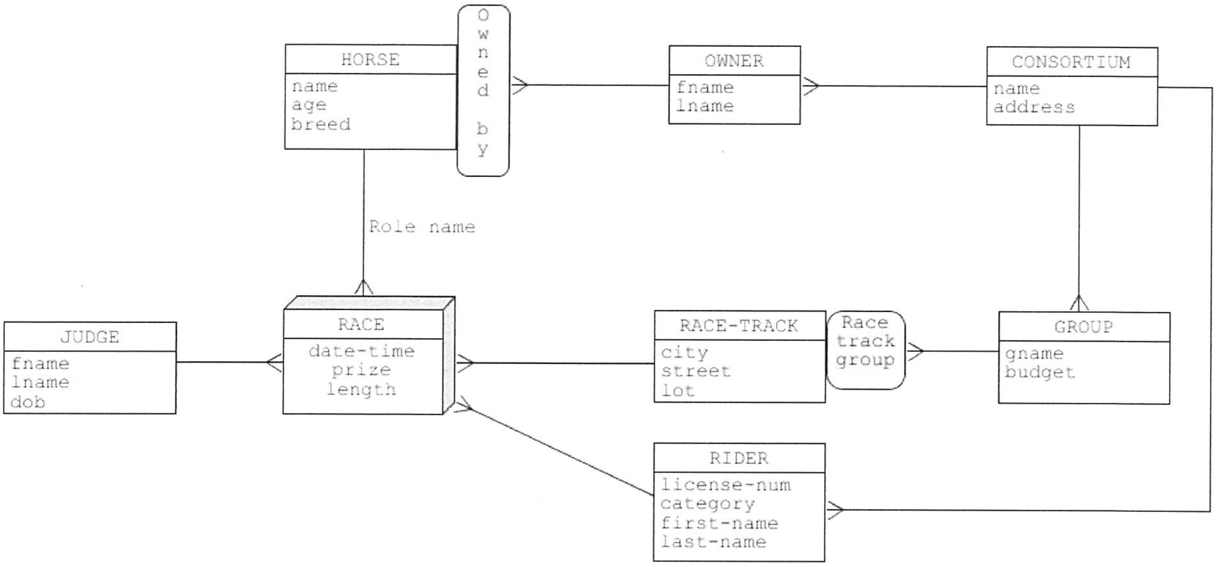
The points scored in a competition are added to the present yearly standings. WTC-GPC keeps information about the total number of points scored by each competitor in the present yearly cycle of the competitions. A competitor who scores the largest total number of points across a yearly cycle wins the prize of the Golden Greasy Pole at the end of the cycle.

WTC-GPC maintains the personal details of all professional climbers, along with a lifetime ranking, and total prize money earned. A lifetime ranking of a professional climber is computed as the total number of points scored in all individual competitions from the beginning of his/her career up to now.

The competitions are organized into a few categories of competitors such as junior, senior, and over 60 competitions organized separately for males and females.

Question 2 (9 marks)

Consider the following conceptual schema of a sample data warehouse.



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 (11 marks)

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

(4 marks)

(2) Explain how the file reading is performed on HDFS. You should describe the main steps.

(4 marks)

(3) What is difference between the client-mode and the cluster-mode in Spark?

(3 marks)

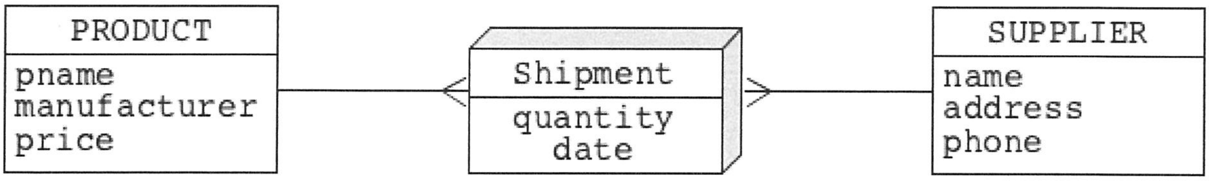
Question 4 (5 marks)

Suppose there is text file on HDFS.

Please explain how to use the MapReduce model to produce two files, one of which contains wordcount output for words beginning with letters from a to m, and the other of which contains wordcount output for words beginning with letters from n to z.

You should describe the main components of the MapReduce model.

Question 5 (7 marks)
Consider the following conceptual schema of a sample data warehouse.



(1) Write the commands of HBase shell command language that create HBase table implementing a sample data warehouse 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 shipments performed by 2 different suppliers. All other details are up to you.

(4 marks)

Question 6 (9 marks)

```
--students.txt
--schema: studentid, name, majorid
1,Henry,100
2,Karen,100
3,Paul,101
4,Jimmy,102
5,Janice,102

--majors.txt
--schema: majorid, majorname
100,SoftwareEngineering
101,InformationManagement
102,BigData
103,CyberSecurity

-- GPAs.txt
--schema: studentid, marjorid, GPA
1,100,3.0
2,100,4.0
3,101,2.5
4,102,4.5
```

Suppose the above files `students.txt`, `majors.txt` and `GPAs.txt` have been uploaded to the root directory of 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.

(1) Load 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 `GPAs_grouped` that groups `GPAs` by `majorid`. Then dump `GPA_grouped`.

(3 marks)

(3) Define a relation `majors_students_outerjoin` that is the full outer join of `majors` and `students` by `majorid`. Then dump `majors_students_outerjoin`.

(3 marks)

Question 7 (5 marks)

This question is related to Apache Hive component of Big Data Technologies.

(1) Explain the concepts of *internal* and *external Hive tables*. In your explanations elaborate on what are the differences and similarities between *Hive tables* and *relational tables* and what are the differences and similarities between *internal* and *external Hive tables*

(2 mark)

(2) Describe a situation where it is necessary to use *internal Hive tables* and describe another situation where it is necessary to use *external Hive tables*.

(3 marks)

In you answers to the questions above you are allowed to refer to your experiences with Hive tables acquired during the laboratory class in the subject.

Question 8 (5 marks)

Use an example to illustrate the difference between the `map()` transformation and the `flatMap()` transformation in Spark.

You should present the data of the two RDDs that you create (e.g., by using the `collect()` action).