Question 1

Based on your own experience of using Apache Pig and Apache Spark in this subject, briefly explain (i) the similarity of data processing between Apache Pig and Apache Spark, and (ii) the advantages of Apache Spark over Apache Pig.

Question 2

A table named X contains a column of keys and a column of values:

key	value
k1	1
k1	2
k2	3
k2	4

Note that the first row of X contains the column names. Explain how to implement the following SQL-like query in the MapReduce model:

SELECT key, SUM(value) FROM X GROUPBY key

You need to specify the key-value data in the input and output of the Map and Reduce stages.

Question 3

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

A university administration would like to record time spent by the students in the lecture classes. To achieve that, the administration installed at the entries to the lecture halls the devices that can read and recognise student cards. Now, each time a student enters and leaves a lecture hall she/he swaps a card against a device, and the total time spent by a student in a lecture hall is computed and recorded.

The administration has the operational databases that contain information about students, subjects, timetables and enrolments. The administration would like to use these databases and information obtained from scanning of student cards to create a data warehouse. It should be possible to get from the warehouse information about the total time spent by each student in lecture classes per given period of time, such as per day, per week, per month, per session, and per year. It should be possible to get information about the total number of times each student attended the lectures per day, per week, per month etc. It should be possible to get information about time spent by each student in the lecture classes per subject, per degree enrolled, and per school that offers the subjects. It should be also possible to get information about the total number of students who attended lectures per subject, per degree enrolled, and per school that offers the subjects.

In your solution, describe each level of dimension with at least 1 and at most 3 attributes. The names of attributes that you choose must relate to the domain described above. Use either ID tags or underscores to denote the identifiers. The notations that you use must conform with the one taught in ISIT312 or ISIT912.

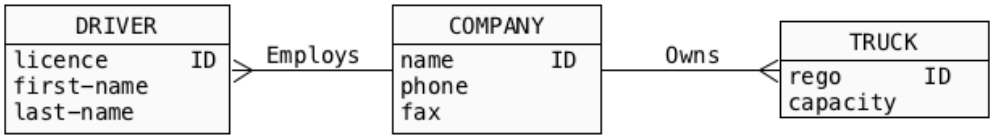
(3.1) Create a conceptual schema for the above specification of a sample data warehouse domain. The graphical notation that you use must conform with the one taught in ISIT312 or ISIT912.

(3.2) Use the relational-algebraic notation to specific the following OLAP query:

- “Find the total time spent by each student in lecture classes per session and lecture”

Question 4

Consider the following conceptual schema of some data cube:



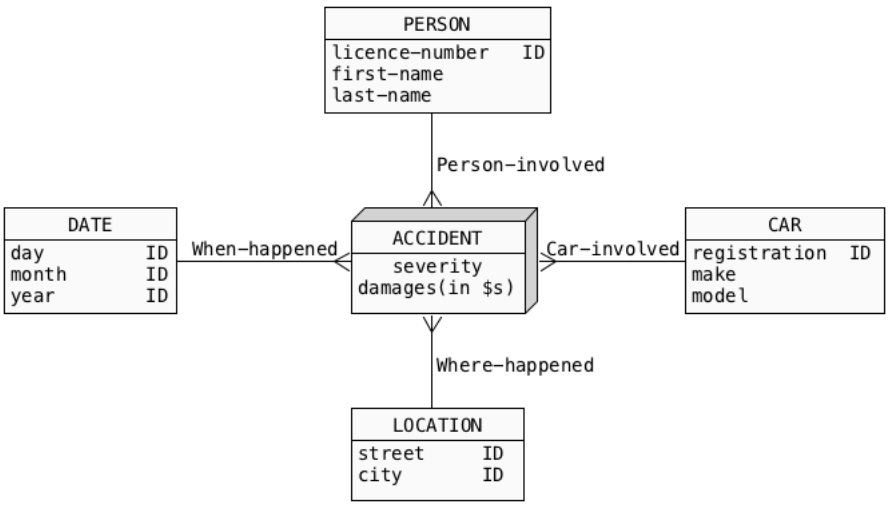
(4.1) Present a drawing of a *logical schema* based on the above conceptual schema. The notation that you use must conform with the one taught in ISIT312.

(4.2) Write the HQL statements that implement the logical schema resulted in the previous step as *internal* tables in Hive. (You should make reasonable assumption about the row, column and file formats of the physical data.)

(4.3) Write an HQL query for finding "the average number of drivers which are associated with a truck".

Question 5

Implement as a single HBase table a database that contains information described by the following conceptual schema.



(5.1) Write HBase shell commands that create the HBase table and load some sample data into the table. The sample data includes information about *at least* two accidents and two cars and one person involved in *both* accidents. All other information is up to you.

(5.2) Assume that this HBase table has been in use and a lot of data has been populated in it. Write the HBase shell commands that implement the following queries.

(a) Find all information about the accidents having damages higher than 1000.
(b) List the first and last names of people involved in accidents in Sydney in 2019.

Question 6

(6.1) Assume that the file `people.txt` is in the HDFS directory `hdfs://localhost:8080/people/`. The contents in `people.txt` are as follows:

```
Michael, 29, software engineer
Andy, 30, data scientist
Justin, 19, business analyst
...
```

Explain how to load `people.txt` into a Spark dataframe named `peopleDF`, where the first and third columns should have a string field and the second column should have an integer field. Also present the Scala source code of your operations.

(6.2) Assume that a dataframe named `FlightsDF` of flight statistics is defined in Spark, with the following code processed:

```
FlightsDF.printSchema()
Out:
root
 |-- DEST_CITY: string (nullable = true)
 |-- DEST_COUNTRY_NAME: string (nullable = true)
 |-- ORIGIN_CITY: string (nullable = true)
 |-- ORIGIN_COUNTRY_NAME: string (nullable = true)

DF.show(2)
Out:
+---------------------+-----------+--------------+
|DEST_CITY|DEST_COUNTRY|ORIGIN_CITY|ORIGIN_COUNTRY|
+---------+-----------+-----------+--------------+
|Sydney   |Australia   |Melbourne  |Australia     |
|Auckland |New Zealand |Singapore  |Singapore     |
+---------+-----------+-----------+--------------+
only showing top 2 rows
```

Based on `FlightsDF`, write down the Scala code in Spark to implement the operation

- "Find the country or countries with most international flights"

Note. An international flight has different original and destination countries.