SIM GLOBAL EDUCATION

UOW AUSTRALIA

SCIT
School of Computing and Information Technology

Family Name ....

First Name ....

Student Number..

ISIT312
Big Data Management

SAMPLE FINAL EXAMINATION PAPER

Exam value: **40% of the subject assessment.**

Marks available: **60 marks.**

Question 1 (10 marks)

Consider the following fragment of **Filter** application that has the functionality equivalent to the functionality of the following SQL statement.

```
  SELECT key, value
  FROM Sequence-of-key-value-pairs
  WHERE value > given-value;

public class Filter {

    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        String[] otherArgs = new GenericOptionsParser(
                                                  conf, args).getRemainingArgs();
        conf.set("limit", otherArgs[0]);
      Job job = new Job(conf, "Distributed Filter");
      job.setJarByClass(Filter.class);
      job.setMapperClass(FilterMapper.class);
      job.setOutputKeyClass(Text.class);
      job.setOutputValueClass(IntWritable.class);
      job.setNumReduceTasks(0); // Set number of reducers to zero
      FileInputFormat.addInputPath(job, new Path(args[1]));
      FileOutputFormat.setOutputPath(job, new Path(args[2]));
      System.exit(job.waitForCompletion(true) ? 0 : 1);
  }

 public static class FilterMapper
     extends Mapper<Object, Text, Text, IntWritable>{

     private final static IntWritable counter = new IntWritable(0);
     private Text word = new Text();
     private Integer total;
     private Integer limit;
     public void map(Object key, Text value, Context context
                 ) throws IOException, InterruptedException {
       StringTokenizer itr = new StringTokenizer(value.toString());

       limit = Integer.parseInt( context.getConfiguration().get("limit") );

       while (itr.hasMoreTokens()) {
           word.set(itr.nextToken());
             total = Integer.parseInt(itr.nextToken());
           if ( total > limit )
           { counter.set( total );
             context.write(word, counter); }
       }
```

An objective of this task is to use the Java code of **Filter** application listed above to explain how would you implement **IN** application that has the functionality equivalent to the functionality of the following SQL statement.

```
  SELECT key, value
  FROM Sequence-of-key-value-pairs
  WHERE value IN (given-value-1, ... , given value-n);
```

You do not need to re-write entire application. However, if Java code is more convenient for you than written English then you are free to write Java code.

Question 2 (10 marks)

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

A vehicle repair company would like to create a data warehouse to keep information about its past and present activities. The company employs a number of mechanics, junior mechanics and administration people. The company owns a number of vehicle repair facilities in the different cities all over a country. The mechanics and junior mechanics employed in the facilities perform the services and repairs of vehicles owned by the customers.

The company would like to organize a data warehouse such that following information can be retrieved/computed from the warehouse later on.

(1) Find the total number of repairs performed by an employee (mechanic or junior mechanic) per day, per month, per year, per facility, per vehicle serviced or repaired.
(2) Find the total number of services performed by an employee (mechanic or junior mechanic) per day, per month, per year, per facility, per vehicle serviced or repaired.
(3) Find the total time spent on repairs or services per employee (mechanic or junior mechanic), per day, per month, per year.
(4) Find the average costs of repairs or services per vehicle (car, truck, bus, other vehicle).
(5) Find the total number of parts used for services or repairs per day, per month, per year.
(6) Find the average time spent on repairs or services per day, per month, per year, per facility.
(7) Find the total amount fuel consumed on testing during repairs or service per vehicle, per day per month, per year.
(8) Find the total number of repairs or services per facility, per city.
(9) Find the total time spent repairs or services per customer.
(10) Find the average quality evaluation indicator per vehicle, per month, per year.

A vehicle is described by a registration number (unique) fuel consumption, and capacity.

Employees are described be employee number (unique), first name, and last name. Mechanics are described by licence number (unique), junior mechanics are described the titles of courses completed.

Facilities are described by address in a city and cities are described by name (unique).

Customers are described by a customer number (unique), first name and last name.

Your task is to create a conceptual schema of a sample data warehouse domain listed above.

To draw a conceptual schema, use a graphical notation explained to you during the lecture classes in a subject ISIT312 Big Data Management.

Question 3 (10 marks)

The following CREATE TABLE statement implements a fact table in a tabular view of three-dimensional data cube in Hive.

```
CREATE TABLE ORDERS(
ORDERKEY    VARCHAR(20),      /* Orders dimension                        */
PARTKEY     VARCHAR(12),      /* Parts dimension                         */
SUPPKEY     VARCHAR(12),      /* Suppliers dimension                     */
QUANTITY    DECIMAL(7),       /* Quantity measure                        */
DISCOUNT    DECIMAL(4,1) );   /* Discount measure                        */
```

Express the following queries as HQL SELECT statements using window partitioning technique.

(1) For each part list its key (PARTKEY), all its quantities (QUANTITY) and the largest quantity.

(2 marks)

(2) Compute an average applied discount (DISCOUNT) for the parts with the keys (PARTKEY) 1001 and 1002 and list the part keys and the average applied discount.

(2 marks)

(3) For each part list its key (PARTKEY), all its applied discounts, and an average applied discount (DISCOUNT) of the current discount and the previous one in the ascending order of available discounts.

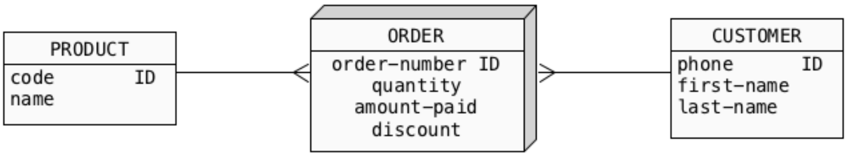
(3 marks)

(4) For each part list its key (PARTKEY), all its applied discounts, and an average applied discount (DISCOUNT) of the current quantity and all previous discounts in the ascending order of applied discounts.

(3 marks)

Question 4 (10 marks)

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



(1) Use HBase shell command language to write the commands that create HBase table implementing a two-dimensional data cube 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 orders submitted by the same customer and including two different products.

(2 marks)

(3) Write the commands of HBase shell command language that retrieve from HBase table created in the previous step the following information:

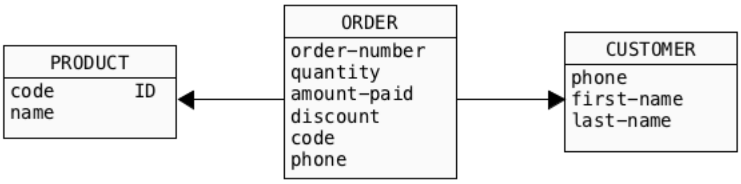
- list all column families for an order number 123,
- list no more than 2 versions of product name used in an order 123,
- list entire table with at most 2 versions per cell,
- list at most 2 versions of orders with timestamps in a range from 1 to 5,
- list all codes of products ordered by a customer whose phone is 345.

(5 marks)

.

Question 5 (10 marks)

Assume that the following logical schema of two-dimensional data cube has been implemented as the data files `customer.txt`, `product.txt`, and `order.txt`.



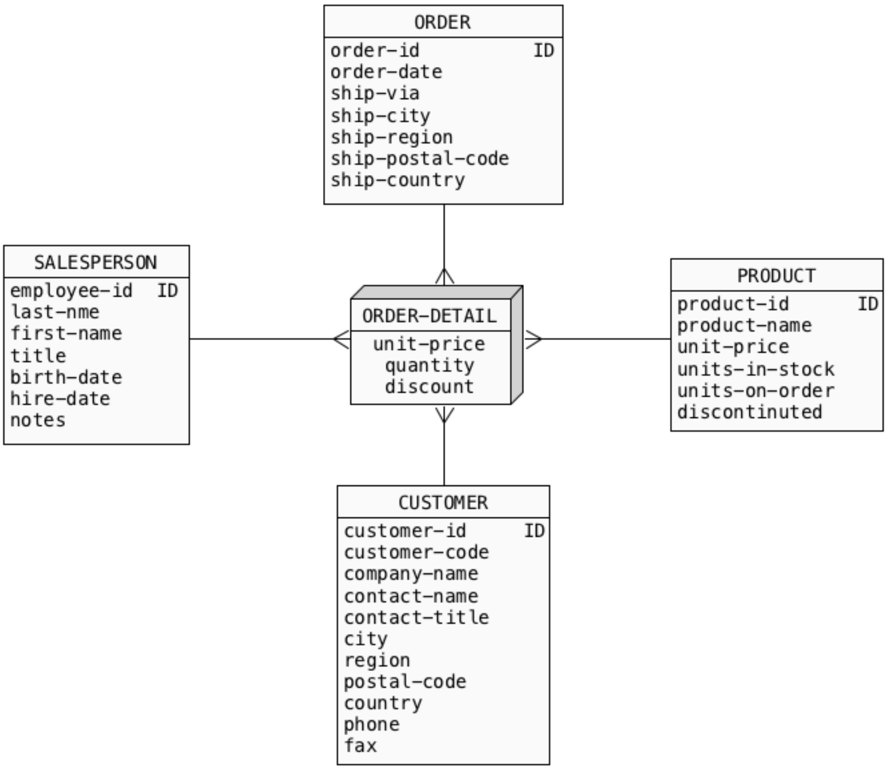
Assume that '|' (vertical bar has been used to separate data items in each row in the data files customer.txt, product.txt, and order.txt.

Assume that the data files `customer.txt`, `product.txt`, and `order.txt` has been uploaded into HDFS.

Write a sequence of Pig-Latin commands that *find the total quantities of products grouped by product name*.

Question 6 (10 marks)

Consider the following conceptual schema of a data warehouse.



The files `customer.tbl`, `order_details.tbl`, `order.tbl`, `product.tbl`, `salesperson.tbl` contain data dumped from a data warehouse whose conceptual schema is given above.

Write the implementations of the following Spark-shell operations.

(1) Create a DataFrame named `orderDetailsDF` that contains information about the details of orders included in a file `order-details.tbl`.
(2) Lists all order details where discount is less than `0.5`.
(3) Find the total number of customers located in `Germany`.
(4) Find the total number of orders per each customer.
(5) Find 5 the most expensive (use attribute `unit-price`) products.

End of Examination