Information about preparing for ISIT312 final exam

Key contents

- Laboratory and assignment tasks
  - *Review all the tasks and reproduce the codes in the VM*
- Lecture notes
  - *Review all the basic/fundamental concepts*
  - *Implement/reproduce the examples in the VM*

Sample questions:

A. Short-answer questions

Examples:

(1) Explain the three core layers in Apache Hadoop.

(2) What is similarity and difference between Hive and traditional relational database?

(3) What are the advantages of Spark compared with MapReduce?

(4) What is the difference between transformation and action in Spark?

B. Data warehouse design

(1) Build a conceptual model from a provided specification.

Example question:

Read and analyse a specification of data warehouse domain listed below. Your task is to create a conceptual schema for the sample data warehouse domain listed below. To draw a conceptual schema, use a graphical notation explained to you during the lecture classes.

A telephone service provider would like to build a data warehouse to keep information about its past and present activities. The company offers a number of different call programs to its customers.

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

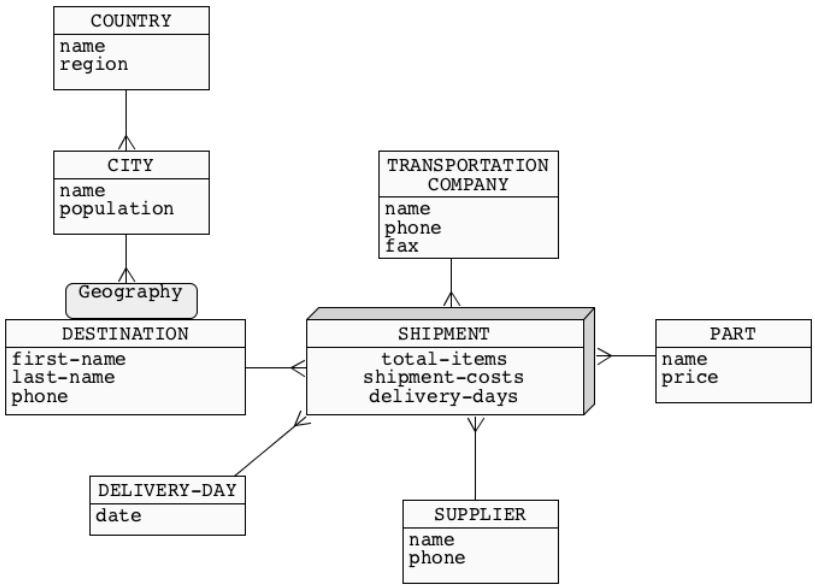
(i) Find, total amount collected per call program, per year and per program and year.
(ii) Find the total duration of calls made by customers per city, per year, per month, per day.
(iii) Find, the total number of weekend calls made from the customers (callers) per city to the customers (callee) per city, per year, per month and year.
(iv) Total duration of international calls started by the customers (callers) per country per year.
(v) Total amount collected from customers per city, per program and per year.

A call program is described by a unique name, price per call, and a short description. A telephone call is described by a phone number of a customer who issued a call (caller), phone number of a customer

who received a call (callee). A call belongs to either a class of local calls or international calls.
A customer is described by a unique phone number (unique), first name and last name, and city.

(2) Translate a conceptual model into a logical model.

Sample conceptual model:



C. HQL statements

Use basic HQL statements to build Hive (internal and external) tables and make data queries.

Example question:

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                    */
```

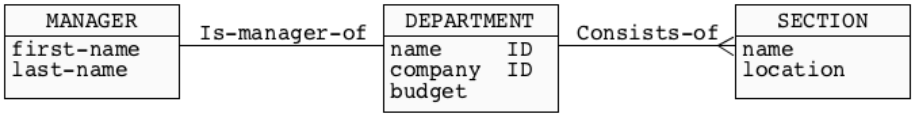
(i) Find the total quantities summarized per part and supplier, per part, and the total quantity of all orders
(ii) Find an average discount per part and supplier, per part, per supplier, and an average discount of all orders.
(iii) Find an average quantity of ordered parts per supplier, per order, and per part.
(iv) Find the total discount per order and part, per supplier, and total discount of all orders

D. HBase implementation and query

Implement HBase tables and queries.

Example question:

Consider the following conceptual schema of a sample database domain.



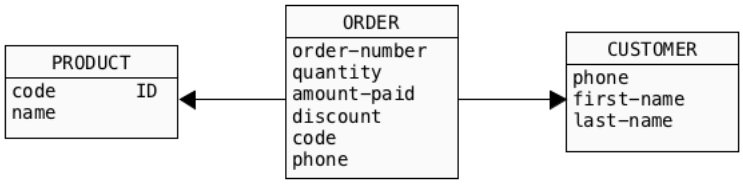
(i) Write the commands of HBase shell command language that create HBase table implementing a sample database domain given above.
(ii) Write the commands of HBase shell command language that insert into HBase table created in the previous step information about at least 2 departments such that each department has one manager and at least 2 sections. All other information is up to you.
(iii) Write the commands of HBase shell command language to retrieve all data in the entire table.

E. Pig dataflow statement

Use Pig-Latin language to implement dataflow statements.

Example question:

Assume that the following logical schema 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` have been uploaded into HDFS.

Write a sequence of Pig-Latin commands that *find the first and the last names of customers who ordered at least one product with a name "bolt".*

F. Spark data processing

Use Spark DataFrame/Dataset and Structured Streaming to process data

Sample questions:

(1) A DataFrame named `flightDF` is defined in Spark-shell. Calling the `show()` method on it returns the following output:

```
flightDF.show(4)
// +-----------------+-------------------+-------+
// |DEST_COUNTRY_NAME|ORIGIN_COUNTRY_NAME|flights|
// +-----------------+-------------------+-------+
// |    United States|            Romania|     15|
// |    United States|            Croatia|      1|
// |    United States|            Ireland|    344|
// |            Egypt|      United States|     15|
```

```
// +-----------------+-------------------+-------+
// only showing top 4 rows
```

You need to compute the average number of flights arriving at each country specified in the `DEST_COUNTRY_NAME` column of `flightDF`.

Explain your operations and write down your code.

(2) Assume that two DataFrames are defined in Spark-shell as follows:

```
val studentsDF = Seq(
  (1023, "James Bond", 001),
  (4102, "Robin Hood", 002),
  (3453, "Harry Potter", 001)
  .toDF("student_id", "name", "course_id")

val coursesDF = Seq(
  (001, "Big Data"),
  (002, "Cyber Security"),
  (003, "Software Engineering"))
  .toDF("course_id", "course_name")
```

Write down your code to perform an inner join for `studentsDF` and `coursesDF` on their `course_id` column.

(3) Complete the following word count application for structured streaming.

```
import spark.implicits._
val lines = spark.readStream
  .format("socket") // socket source
  .option("host", "localhost") // listen to the localhost
  .option("port", 9999) // and port 9999
  .load()
val words = <insert your code here>
val wordCounts = words.groupBy("value").count()
val query = wordCounts.writeStream
  .outputMode("complete") // accumulate counting result of the stream
  .format("console") // use the console as the sink
  .start()
```

(4) Explain how you compile the source code of a self-contained application and submit it to a Spark cluster. Support your answer with Terminal commands.