School of Computing & Information Technology
University of Wollongong

Session: **2, April 2022**
Lecturer: **Janusz R. Getta**

CSCI312 Big Data Management
Singapore 2022-2
Assignment 2

Published on 24 April 2022

Scope

The objectives of Assignment 2 include conceptual modelling of a data warehouse, implementation of 0NF tables in HQL, implementation of external tables in HQL, and querying a data cube.

This assignment is due on **Saturday, 21 May 2022, 9:00pm** (sharp) Singaporean Time (ST).

This assignment is worth **30%** of the total evaluation in the subject.

Only electronic submission through Moodle at:
`https://moodle.uowplatform.edu.au/login/index.php`
will be accepted. All email submissions will be deleted and mark 0 ("zero") will be immediately granted for Assignment 2. A submission procedure is explained at the end of Assignment 2 specification.

A policy regarding late submissions is included in the subject outline.

Only one submission of Assignment 2 is allowed and only one submission per student is accepted.

A submission marked by Moodle as "late" is always treated as a late submission no matter how many seconds it is late.

A submission that contains an incorrect file attached is treated as a correct submission with all consequences coming from the evaluation of the file attached.

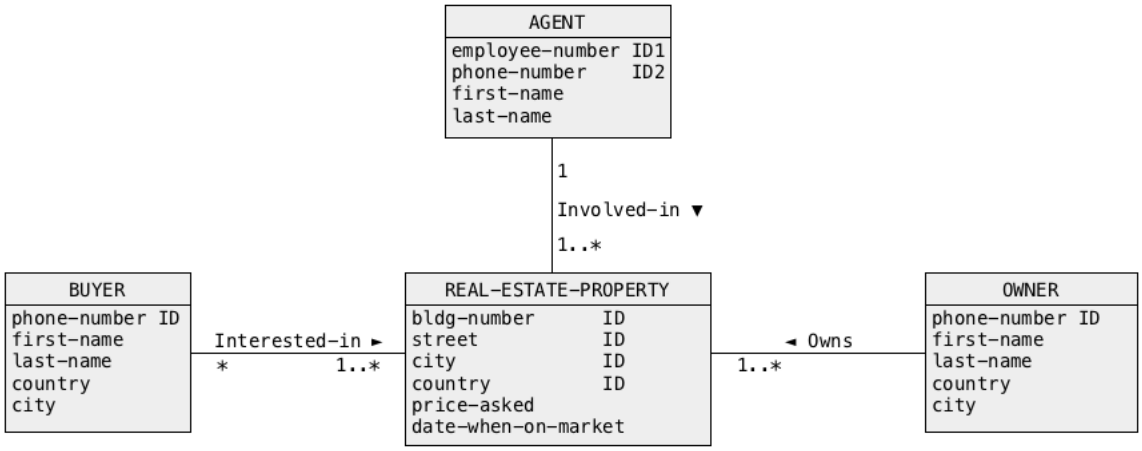
All files left on Moodle in a state **"**`Draft(not submitted)`**"** will not be evaluated.

A submission of compressed files (zipped, gzipped, rared, tared, 7-zipped, lhzed, … etc) is not allowed. The compressed files will not be evaluated.

The second assignment is an **individual assignment** and it is expected that all its tasks will be solved **individually without any cooperation** with the other students. However, it is allowed to declare in the submission comments that a particular component or task of this assignment has been implemented in cooperation with another student. In such a case evaluation of a task or component may be shared with another student. In all other cases plagiarism will result in a **FAIL** grade being recorded for entire assignment. If you have any doubts, questions, etc. please consult your lecturer or tutor during laboratory/tutorial classes or over e-mail.

Task 1 (6 marks)
Intuitive design of a data cube from a conceptual schema of an operational database

Consider the following conceptual schema of an operational database owned by a multinational real estate company. The database contains information about the real estate properties offered for sale, owners of the properties, potential buyers who are interested in the properties and real estate agents involved in selling of the properties.



Whenever a property is put on a market by an owner, a description of the property is entered into an operational database. Whenever a property is purchased, its description is removed from an operational database.

The real estate company would like to create a data warehouse to keep information about the finalized real estate transactions, properties involved in the transactions, sellers/owners, and agents involved in the real estate transactions. The real estate company would like to use a data warehouse to implement the following classes of analytical applications.

(1) Find the total number of real estate properties sold per month, year, street, city, country, agent involved.
(2) Find an average asked price of real estate properties sold per month, year, street, city, country, agent involved.
(3) Find an average final price of real estate properties sold per month, year, street, city, country, agent involved.
(4) Find an average period of time on the marked of real estate properties sold per month, year, street, city, country, agent involved.
(5) Find the total number of times each real estate property has been sold in a given period of time.
(6) Find the total number of buyers interested in purchases of real estate properties sold per day, month, year, street, city, country, agent involved.

Note, the operational database does not contain all information necessary to implement the classes of applications listed above. Additional information must be added when data is transferred from an operation database to a data warehouse.

(1) Use a short explanation of a database domain and a conceptual schema given above, to find a data cube, that should be implemented by the multinational real estate company to create a data warehouse. In your specification of a data cube, list the facts, the measures, the names of dimensions and the hierarchies.

(2) Pick any three dimensions from a data cube found in the previous step and at least 4 values in each dimension and one measure to draw a sample three-dimensional data cube in a perspective view similar to a view included in a presentation `09 Data Warehouse Concepts`, slide 6.

Deliverables

A file `solution1.pdf` that contains

(1) a specification of data cube as a list of facts, measures, dimensions, and hierarchies obtained as a result of task (1),

(2) a perspective drawing of three-dimensional data cube as a result of task (2).

Task 2 (7 marks)
Conceptual modelling of a data warehouse

An objective of this task is to create a conceptual schema of a sample data warehouse domain described below. Read and analyse the following specification of a data warehouse domain.

A person is represented as either a patient or a medical worker or an administration worker. Medical and administration workers work in the medical facilities that have a name, address, possibly (not obligatory) specialization. Each medical worker is described a unique staff number at a facility, name, address, and phone number.

A patient visits a medical facility for service of a health problem. Each service involves a patient, a medical worker, and administration worker. The service can be diagnosis, treatment, or checkup. A description and date of each service is recorded. Time spent on service and the costs are recorded as well.

A patient is eligible for his or her company health care benefits. Patient data includes name, id number (social security number), address (street, city, state, zip), and phone.

A medical worker must hold one or more credentials that are granted to work in a particular medical facility. Doctors are allowed to deliver diagnosis and give treatment based on their specialization Paramedics are allowed to deliver only emergency diagnosis and treatment for any type of life-threatening problems. Nurses do not deliver diagnosis, but they do participate in treatment, particularly if the patient must be prepared for surgery or remain at the facility overnight.

The administration workers are concerned with personnel needs and assignments. Each medical worker must have at most one assignment at a facility. Several administration workers can be assigned to one assignment.

Medical facilities are located in different suburbs of different cities. A medical facility is uniquely identified by an address.

A data warehouse must be designed such it should be possible to easily implement the following classes of applications.

A management of the medical facilities would like to get from a data warehouse information about

- *the total number of medical services performed per medical facility, per year, per month per day, per city and per medical worker,*
- *total length of medical services per medical facility, per year, per month per day, per city and per medical worker,*
- *average length of medical services per medical facility, per year, per month per day, per city and per medical worker,*

- *total number of doctors/ paramedics/nurses involved in medical services, per year, per month per day, per medical facility, per city,*
- *average time spend on medical services per year, month, day,*
- *total costs of medical services per year, month, day, per medical facility, per city.*

To draw a conceptual schema, use a graphical notation explained to you in a presentation 11 Conceptual Data Warehouse Design.

To create a conceptual schema of a sample data warehouse domain, follow the steps listed below.

Step 1 Find a fact entity, find the measures describing a fact entity.
Step 2 Find the dimensions.
Step 3 Find the hierarchies over the dimensions.
Step 4 Find the descriptions (attributes) of all entity types.
Step 5 Draw a conceptual schema.

To draw a conceptual schema, you must use a graphical notation explained to you in a presentation 11 Conceptual Data Warehouse Design.

To draw your diagram, you can use UMLet diagram drawing tool and apply a "Conceptual modelling" notation, Selection of a drawing notation is available in the right upper corner of the main menu of UMLet diagram drawing tool. UMLet 14.3 software is can be downloaded from the subject's Moodle Web site in a section WEB LINKS. A neat hand drawing is still all right.

Deliverables

A file `solution2.pdf` with a drawing of a conceptual schema of a sample data warehouse domain.

Task 3 (6 marks)
Implementation of a table with a complex column type (0NF table) in Hive

Assume that we have a collection of semi-structured data with information about the employees (unique employee number and full name) the projects they are assigned to (project name and percentage of involvement) and their programming skills (the names of known programming languages). Some of the employee are on leave and they are not involved in any project. Also, some of the employee do not know any programming languages. Few sample records from the collection are listed below.

```
007|James Bond|DB/3:30,Oracle:25,SQL-2022:100|Java,C,C++
008,Harry Potter|DB/3:70,Oracle:75|
010,Robin Banks| |C,Rust
009,Robin Hood| |
...
```

(1) Implement HQL script `solution3.hql` that creates an internal relational table to store information about the employees, the projects they are assigned to (project name and percentage of involvement) and their programming skills.

(2) Include into the script `INSERT` statements that load sample data into the table. Insert at least 5 rows into the relational table created in the previous step. Two employees must participate in few projects and must know few programming languages. One employee must participate in few projects and must not know any programming languages. One employee must know few programming languages and must not participate in any projects. One employee must not know programming languages and must not participate in the projects.

(3) Include into the script `SELECT` statements that lists the contents of the table.

When ready, use a command line interface `beeline` to process a script `solution3.hql` and to save a report from processing in a file `solution3.rpt`.

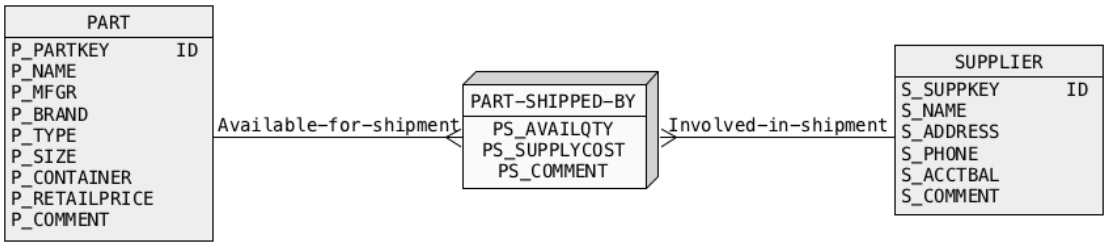
If the processing of the file returns the errors then you must eliminate the errors! Processing of your script must return NO ERRORS! **A solution with errors is worth no marks!**

Deliverables
A file `solution3.rpt` with a report from processing of HQL script `solution3.hql`. The report **MUST NOT include any errors**, and the report **must list all SQL statements processed**.

Task 4 (6 marks)

Implementation of a data warehouse as a collection of external tables in Hive

Consider the following two-dimensional data cube.



The data cube contains information about that parts that can be shipped by the suppliers.

Download and unzip a file `task4.zip`. You should obtain a folder `task4` with the following files: `part.tbl`, `supplier.tbl`, `partsupp.tbl`.

Use an editor to examine the contents of `*.tbl` files. Note, that the contents of the files can be loaded into the relational tables obtained from the transformation of the two-dimensional data cube given above into the relational table `PART`, `SUPPLIER`, and `PARTSUPP`.

Transfer the files into HDFS.

Implement HQL script `solution4.hql` that creates the external tables obtained from a step of logical design performed earlier. The external tables must overlap on the files transferred to HDFS in the previous step. Note, that a header in each `*.tbl` file must be removed before creating the external tables.

Include into `solution4.hql` script `SELECT` statements that any 5 rows from each one of the external tables implemented in the previous step and the total number of rows included in each table.

When ready, use a command line interface `beeline` to process a script `solution4.hql` and to save a report from processing in a file `solution3.rpt`.

Deliverables

A file `solution4.rpt` with a report from processing of HQL script `solution4.hql`.

Task 5 (5 marks)
Querying a data cube

Download a file `task5.zip` and unzip the file. You should obtain a folder `task5` with the following files: `dbcreate.hql`, `dbdrop.hql`, `partsupp.tbl`, `lineitem.tbl`, and `orders.tbl`.

A file `orders.tbl` contains information about the orders submitted by the customers. A file `lineitem.tbl` contains information about the items included in the orders. A file `partsupp.tbl` contains information about the items and suppliers of items included in the orders.

Open Terminal window and use `cd` command to navigate to a folder with the just unzipped files. Start Hive Server 2 in the terminal window (remember to start Hadoop first). When ready process a script file `dbcreate.hql` to create the internal relational tables and to load data into the tables. You can use either `beeline` or SQL Developer. A script `dbdrop.hql` can be used to drop the tables.

The relational tables `PARTSUPP`, `LINEITEM`, `ORDERS` implement a simple two-dimensional data cube. The relational tables `PARTSUPP` and `ORDERS` implement the dimensions of parts supplied by suppliers and orders. A relational table `LINEITEM` implements a fact entity of a data cube.

(1) For each part list its key (`PS_PARTKEY`), all its available quantities (`PS_AVAILQTY`), the smallest available quantity, and the average available quantity. Consider only the parts with the keys `5` and `15`.

(2) For each part list its key (`PS_PARTKEY`), all its available quantities (`PS_AVAILQTY`), and summarized all available quantities. List the results in the order of increasing results of summarization. Consider only the parts with the keys `5`, `10`, `15`, and `25`. Cube/Rollup

(3) For each part list its key (`PS_PARTKEY`), all its available quantities (`PS_AVAILQTY`), and summarized all available quantities. List the results in the order of increasing results of summarization and supply costs (`PS_SUPPLYCOST`). Consider only the parts with the keys `5`, `10`, `15`, and `25`. Window partitioning

(4) For each part list its key (`PS_PARTKEY`) and all its available quantities (`PS_AVAILQTY`) sorted in descending order and a rank (position number in an ascending order) of each quantity. Consider only the parts with the keys `10` and `20`. Use an analytic function `ROW_NUMBER()`.

(5) For each part list its key (`PS_PARTKEY`), its available quantity, and an average available quantity (`PS_AVAILQTY`) of the current quantity and all previous quantities in the ascending order of available quantities. Consider only the parts with

the keys `15` and `25`. Use `ROWS UNBOUNDED PRECEEDING` sub-clause within `PARTITION BY` clause.

When ready, save your `SELECT` statements in a file `solution5.hql`. Then, process a script file `solution5.hql` and save the results in a report `solution5.rpt`.

If the processing of the file returns the errors then you must eliminate the errors! Processing of your script must return NO ERRORS! **A solution with errors is worth no marks!**

Deliverables

A file `solution5.rpt` with a report from processing of HQL script `solution5.hql`. The report **MUST NOT include any errors**, and the report **must list all SQL statements processed**.

Submission of Assignment 2

Note, that you have only one submission. So, make it absolutely sure that you submit the correct files with the correct contents. No other submission is possible !

Submit the files **solution1.pdf**, **solution2.pdf**, **solution3.rpt**, **solution4.rpt**, and **solution5.rpt** through Moodle in the following way:

(1) Access Moodle at **http://moodle.uowplatform.edu.au/**
(2) To login use a **Login** link located in the right upper corner the Web page or in the middle of the bottom of the Web page
(3) When logged select a site **ISIT312 (SP222) Big Data Management**
(4) Scroll down to a section **SUBMISSIONS**
(5) Click at **In this place you can submit the outcomes of your work on the tasks included in Assignment 2** link.
(6) Click at a button **Add Submission**
(7) Move a file **solution1.pdf** into an area **You can drag and drop files here to add them**. You can also use a link **Add**...
(8) Repeat step (7) for the remaining files **solution2.pdf**, **solution3.rpt**, **solution4.rpt**, and **solution5.rpt**
(9) Click at a button **Save changes**
(10) Click at a button **Submit assignment**
(11) Click at the checkbox with a text attached: **By checking this box, I confirm that this submission is my own work,** ... in order to confirm authorship of your submission.
(12) Click at a button **Continue**

End of specification