School of Computing & Information Technology
University of Wollongong

Session: **4, 2024**
Lecturer: **Dr Fenghui Ren**

ISIT312 Big Data Management
SIM S4 2024
Assignment 2

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 **3 November 2024 by 9:00 pm** Singaporean Time (ST).

This assignment is worth **20%** of the total evaluation in the subject. The assignment consists of 4 tasks and the specification of each task starts from a new page.

Only electronic submission through Moodle at:
`https://moodle.uowplatform.edu.au/login/index.php`
will be accepted. Email submissions will not be accepted. 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 late submission penalty (25% of the total mark) will be applied for every 24 hours late.

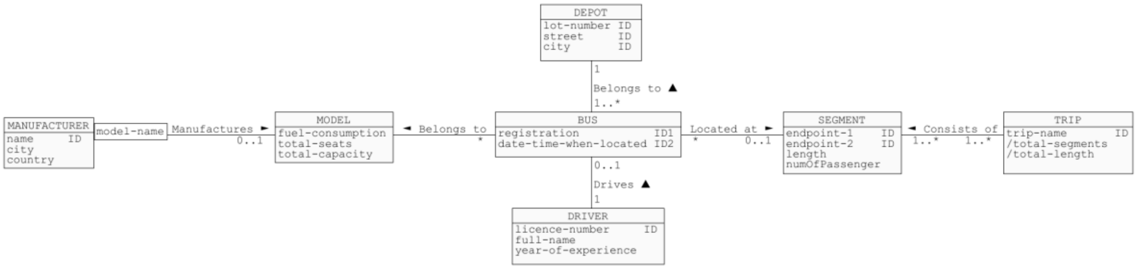
A submission that contains an incorrect file attached is treated as a correct submission with all consequences coming from the evaluation of the file attached.

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

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 (5 marks)
Intuitive design of a data cube from a conceptual schema of an operational database

A bus transportation company maintains an operational database, that contains information about the current locations of the busses owned by the company. A current location of a bus is determined by a trip segment a bus passes through, in a moment. A trip consists of a sequence of trip segments. A bus can traverse a trip in both directions. It is important to note, that the operational database contains only "point-in-time" information. Each time a bus moves to the next segment all information about its past locations (at the previous segments) is removed from a database. The remaining contents of an operational database is consistent with a conceptual schema given below.



The company would like to implement a data warehouse that can be used to implement the following applications.

(i) *find the total number of kilometers travelled by each bus per year, per month per day.*

(ii) *find the total number of trips performed per bus, per driver, per year, per month, per day.*

(iii) *find the total number of drivers per trip.*

(iv) *find the total number of buses travel per trip segment, per trip, per year, per month per day.*

(v) *find an average duration of bus travel per trip segment, per trip, per year, per month per day.*

(vi) *find the total fuel consumption per trip segment, per trip, per bus model, per manufacturer, per year, per month, per day.*

(vii) *find the total number of trips per bus, per depot, per city.*

(viii) *find the total number of passengers per segment, per trip, per year, per month per day.*

(ix) find the largest number of passengers per bus, per trip.

(x) find an average number of passengers per trip.

(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 bus 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. For each of the measurements, provide an explanation how the measurements are obtained.

(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 (5 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 large international network of hotels would like to create a data warehouse to store information about their hotels located in the different cities of different countries, hotel guests visiting the rooms in hotels, and employees working at the hotels. The management of the network would like to store the following information in the data warehouse.

Each hotel is described by its location (country, city, building number), email address and link to a Web page. A hotel offers the rooms to its customers. A room has a unique number within a hotel. A room number consists of a floor number and a unique number at a floor. For example, room 25 at 5th floor has a number 0525.

Each hotel has a number of employees. An employee has a unique employee number, first name, last name, and date of birth. Staff members belong to either administration group or maintenance group. Among the other duties, administration staff members are allowed to perform check-in and check-out of hotel guests. Maintenance staff members perform the maintenance works in the rooms occupied by hotel guests.

Hotel guests stay in hotel rooms. On check-in day a start date of a visit is recorded and on check-out day an end date of a visit is recorded. The data warehouse must contain information about the total number days of each visit and amount of money paid by each hotel guest, total number of facilities used by hotel guests, and the total number of maintenances performed in a room during a visit.

A hotel guest is described by a number of identification document, first name, last name, date of birth and nationality. A hotel guest uses a credit card to pay for his/her stay in a hotel. A credit card number and a name of bank that issued a card is recorded.

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

A management of the hotel network would like to get from a data warehouse information about

- the total number of visits per hotel and per given period of time like day, month, and year,
- about the total number of visits in hotels per city and per country,
- about the total number of check-ins/outs per employee, and
- about the total number of visits paid per credit card used, total number of customers per hotel, per room, per month per year, total profits per hotel, per city where the hotels are located,

- the average length of stay per year, per month, per hotel, average discount applied per hotel, per month per year.

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 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 (5 marks)
Implementation of a table with a complex column type (0NF table) in Hive

Assume that we have a collection of semi-structured data that contains information about the students and their final evaluations (in a scale from 0 to 99) of the enrolled subjects. The first value in each row is a unique student number. Next, we have a list of pairs, that consist of subject code and the final evaluation in a subject.

```
1111,CSIT111:01,CSIT121:23,CSIT101:50,CSIT235:99,ISIT312:02
1112,CSIT101:56,CSIT111:78,CSIT115:10,ISIT312:05
1113,CSIT121:76,CSIT235:87:ISIT312:49
1114,CSIT111:50,ISIT312:45
1115,ISIT115;67,CSCI235:45,CSIT127:56
...
```

(1) Implement HQL script `solution3.hql` that creates an internal table to store information about the student numbers and the evaluations of the subjects. An internal table must be nested (it must be in 0NF).

(2) Include into the script `INSERT` statements that load sample data into the table. You must insert at least 5 records, that have a structure consistent with the records listed above.

(3) Include into the script `SELECT` statements that lists the contents of the table. Assume that no more 5 subjects have been participated by each student.

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.pdf`.

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.pdf` 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 (5 marks)

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

Download the following files: `author.tbl`, `item.tbl`, `dbcreate.sql`.

Use an editor to examine the contents of `*.tbl` files. The files contain synthetic data extracted from the relational tables. A file `dbcreate.sql` contains `CREATE TABLE` statements used to create the relational tables with the synthetic data.

Transfer the files `author.tbl`, `item.tbl` into HDFS.

(1) Implement HQL script `solution4.hql` to create the external tables, that provide a tabular view of synthetic data included in the files `author.tbl`, `item.tbl`. The external tables must overlap on the files transferred to HDFS in the previous step. It is recommended to use `CREATE TABLE` statements included in a file `dbcreate.sql` to create a file `solution4.hql`.

(2) Include into `solution4.hql` script `SELECT` statements, that list the total number of rows in each table, the total number of rows in both tables, and the first 3 rows from each one of the external tables implemented in the previous step.

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

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 `solution4.pdf` with a report from processing of HQL script `solution4.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 absolutely sure that you submit the correct files with the correct contents. Please submit an Academic Consideration in SOLS if an extension (1 week maximally) is required.

Please combine all files into a single zipped file (**`A2-solutions.zip`**). Please submit the zipped file 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 (SP424) Big Data Management`**
(4) Scroll down to a section **`SUBMISSIONS`**
(5) Click at **`Assignment 2`** link.
(6) Click at a button **`Add Submission`**
(7) Move the zipped file **`A2-solutions.zip`** into an area **`You can drag and drop files here to add them`**. You can also use a link **`Add`**...
(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