School of Computing & Information Technology
University of Wollongong

Session: **2, April 2021**
Lecturer: **Janusz R. Getta**

CSCI312 Big Data Management
Singapore 2021-2
Assignment 2

Published on 26 April 2021

Scope

The objectives of Assignment 2 include conceptual modelling of a data warehouse, logical design of a data warehouse, implementation of a data warehouse as a collection of external tables in Hive, and querying a data cube.

This assignment is due on **Friday, 14 May 2021, 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 functional specification of operational database

A train company has an operational database with information about the daily train trips between the cities located in the same or different countries. 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 made by trains in a given year, departing from the stations locating in a given country and arriving at the stations located in a given country.*

(ii) *find the total duration of international trips in a given year, that is, trips departing from a station located in a country and arriving at a station located in another country,*

(iii) *find the total number of trips that departed from or arrived at a given city in a given month of a given year,*

(iv) *find and average duration of train trips in a given country in a given year,*

(v) *for all trips in a given year, find an average number of passengers on a trip.*

(vi) *find an average number of passengers all trips between two given city.*

(vii) *find total number of trips per each driver.*

(viii) *find the total number of trips that used a given train type in a given year.*

(1) Use the specifications of applications listed above to find a data cube, that should be implemented by the train company to create a data warehouse. In your specification of a data cube, list the names of dimensions, hierarchies, and measures.

(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 names of dimensions, list of hierarchies, list of measures and a list of attributes as a result of task (1),
(2) a perspective drawing of three-dimensional data cube as a result of task (2).

Task 2 (6 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 total number of visits in hotels per city and per country, about 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, 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.

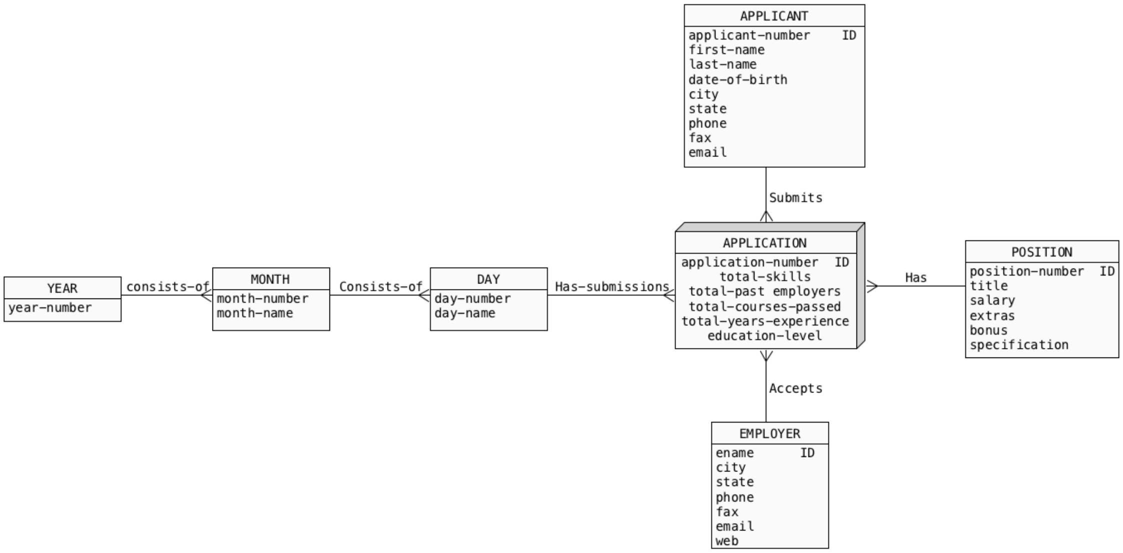
To draw your diagram, you can use UMLet diagram drawing tool and apply a "Conceptual modelling" notation, Selection of a drawing notation is available in the right upper corner of the main menu of UMLet diagram drawing tool. UMLet 14.3 software is can be downloaded from the subject's Moodle Web site 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 (4 marks)
Logical modelling of a data warehouse

Consider the following conceptual schema of a data warehouse.

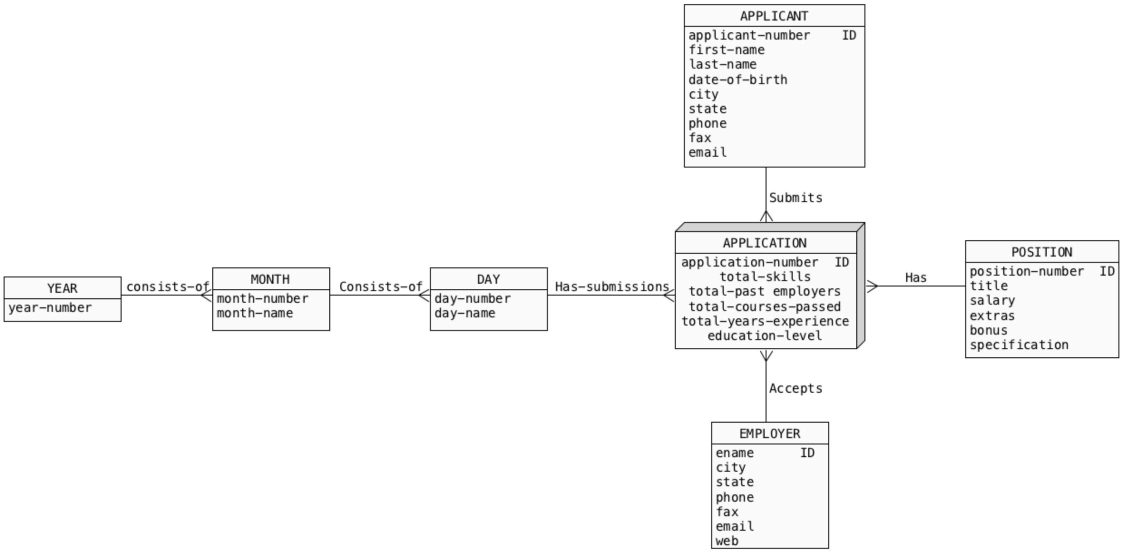


Perform a step of logical design to transform a conceptual schema given above into a logical schema (star schema). Use UMLet diagram drawing tool and apply a "Logical modelling" notation to draw a logical schema. Selection of a drawing notation is available in the right upper corner of the main menu of UMLet. Save a diagram of logical schema in a file `solution3.uxf` and export it to a file `solution3.pdf`.

<u>Deliverables</u>
A file `solution3.pdf` with a drawing of a logical schema.

Task 4 (6 marks)
Implementation of a data warehouse as a collection of external tables in Hive

Consider the following conceptual schema of a data warehouse.



Download a file `task4.zip` and unzip it. You should obtain a folder `task4` with the following files: `applicant.tbl`, `position.tbl`, `employer.tbl`, and `application.tbl`.

Use text editor to examine the contents of `*.tbl` files. The order of columns with values is usually consistent with the order of properties in the entity types of a conceptual schema above. In the case of a file `application.tbl` an order of columns with values is a bit different. It is your task to discover the most appropriate order. Note, that you may have to "clean" the files. It means that you may have to remove small mistakes in the files. It is called Extract, Transform, and Load (ETL).

When ready, 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 you can re-use the outcomes of a logical design performed in Task 3 above.

Include into `solution4.hql` script `SELECT` statements that lists any 3 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 `solution4.rpt`.

Deliverables

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

Task 5 (6 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) Implement the following query using GROUP BY clause with CUBE operator.

For the order clerks (O_CLERK) Clerk#000000522, Clerk#000000154, find the total number of ordered parts per customer (O_CUSTKEY), per supplier (L_SUPPKEY), per customer and supplier (O_CUSTKEY, L_SUPPKEY), and the total number of ordered parts.

(2) Implement the following query using GROUP BY clause with ROLLUP operator.

For the parts with the keys (L_PARTKEY) 7, 8,9 find the largest discount applied (L_DISCOUNT) per part key (L_PARTKEY) and per part key and supplier key (L_PARTKEY, L_SUPPKEY) and the largest discount applied at all.

(3) Implement the following query using GROUP BY clause with GROUPING SETS operator.

Find the smallest price (L_EXTENDEDPRICE) per order year (O_ORDERDATE), and order clerk (O_CLERK).

Implement the following SQL queries as SELECT statements using window partitioning technique.

(4) 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.

(5) 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().

(6) 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.

Deliverables

A file solution5.rpt that contains a report from processing of SELECT statements.

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.pdf`**, **`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 (SP221) 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.pdf`**, **`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