School of Computing & Information Technology
University of Wollongong

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

CSCI312 Big Data Management
Singapore 2022-2
Assignment 3

Published on 22 May 2022

Scope

The objectives of Assignment 3 implementation of HBase table, querying and manipulating data in HBase table, simple data processing with Pig, and data processing with Spark.

This assignment is due on **Saturday, 4 June 2022, 9:00pm** (sharp) Singaporean Time (ST).

This assignment is worth **20%** 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 3. A submission procedure is explained at the end of Assignment 3 specification.

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

Only one submission of Assignment 3 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.

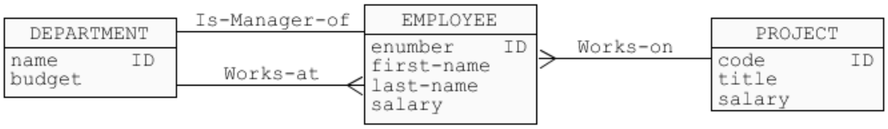
A submission of compressed files (zipped, gzipped, rared, tared, 7-zipped, lhzed, … etc) is not allowed. The compressed files will not be evaluated.

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 (5 marks)

Design and implementation of HBase table

Implement as a single HBase table a database that contains information described by the following conceptual schema.



(1) Create HBase script `solution1.hb` with HBase shell commands that create HBase table and load sample data into the table. Load into the table information about at least one department, three employees such that that one of them is a manager of the others and two projects the employees are working on.

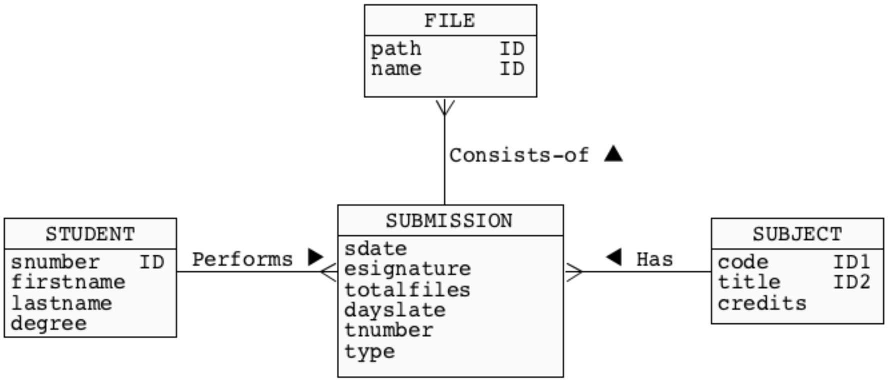
When ready use HBase shell to process a script file `solution1.hb` and to save a report from processing in a file `solution1.rpt`.

Deliverables

A file `solution1.rpt` that contains a report from processing of `solution1.hb` script with the statements that create HBase table and load sample data.

Task 2 (5 marks)
Querying and manipulating data in HBase table

Consider a conceptual schema given below. The schema represents a simple database domain where students submit assignments and each submission consists of several files and it is related to one subject.



Download a file `task2.hb` with HBase shell commands and use HBase shell to process it. Processing of `task2.hb` creates HBase table `task2` and loads some data into it.

Use HBase shell to implement the following queries and data manipulations on the HBase table created in the previous step. Save the queries and data manipulations in a file `solution2.hb`.

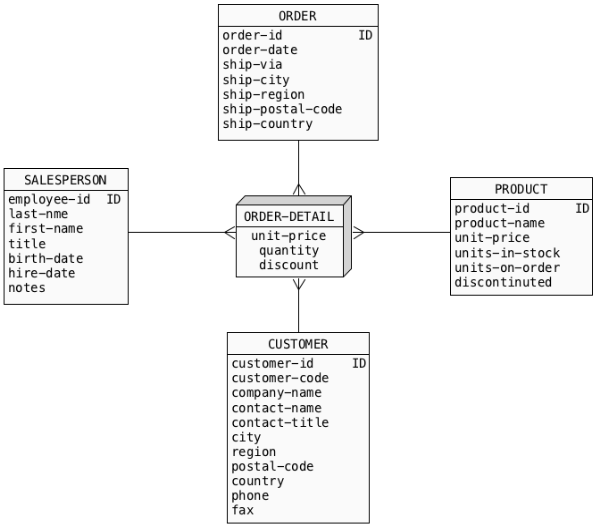
(1) Find all information about a student number `007`, list one version per cell.
(2) Find all information about a submission of assignment `1` performed by a student `007` in a subject `312`, list one version per cell.
(3) Find the first and the last names of all students, list one version per cell.
(4) Find all information about a student whose last name is `Potter`, list one version per cell.
(5) Delete a column family `FILES`.
(6) Add a column family `ENROLMENT` that contains information about dates when the subjects have been enrolled by the students and allow for 2 versions in each cell of the column family.
(7) Insert information about at least two enrolments performed by the students.
(8) List information about all enrolments performed by the students.
(9) Increase the total number of versions in each cell of a column family `ENROLMENT`.
(10) Delete HBase table `task2`.

When ready, start HBase shell and process a script file `solution2.hb` with Hbase command shell. When processing is completed copy the contents of Command window with a listing from processing of the script and paste the results into a file solution2.rpt. Save the file. When ready submit a file solution2.rpt.

Deliverables
A file solution2.rpt with a listing from processing of a script file solution2.hb.

<u>Task 3 (5 marks)</u>

Consider the following conceptual schema of a data warehouse.



Download a file `task3.zip` published on Moodle together with a specification of Assignment 3 and unzip it. You should obtain a folder `TASK3` with the following files: `customer.tbl`, `order_detail.tbl`, `order.tbl`, `product.tbl`, `salesperson.tbl`. The files contain data dumped from a data warehouse whose conceptual schema is given above.

Use editor to examine the contents of `*.tbl` files. Note, that each file has a header with information about the meanings of data in each column. A header is not a data component of each file.

(1) Remove the headers and transfer the files into HDFS.

Create Pig Latin script `solution3.pig` that implements the following queries.

(2) Find the first and the last name (`first-name`, `last-name`) of sales people who handled the orders submitted by a company `Consolidated Holdings`.

(3) Find the total number of products not ordered in `1996`.

(4) Find the summarizations of quantities (`quantity`) per ordered product (`product-id`).

(5) Find the identifiers of orders (`order-id`) that included both `Ikura` and `Tofu.`

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

Deliverables

A file `solution3.rpt` with a report from processing of Pig Latin script `solution3.pig`.

Task 4 (5 marks)

Data processing with Spark

Consider the following sales related information.

```
bolt 45
bolt 5
drill 1
drill 1
screw 1
screw 2
screw 3
...   ...
```

Add more lines to information listed above and load the sales related information into a text file `sales.txt` and later on load the file into HDFS.

An objective of this task is to *find the total sales per part* using three different techniques: Resilient Distributed Datasets, Datasets, and DataFrames with SQL.

Use Spark command line interface to implement the following tasks.

(1) Load the contents of a file `sales.txt` located in HDFS into a Resilient Distributed Dataset (RDD) and use RDD to find the total sales pert part.

When ready copy the contents of `Terminal` screen with a report from implementation of a task (1) and paste it into a file `solution4.rpt`.

(2) Load the contents of a file `sales.txt` located in HDFS into a Dataset and use the Dataset to find the total sales pert part.

When ready copy the contents of `Terminal` screen with a report from implementation of a task (2) and paste/append it at the end of a file `solution4.rpt`.

(3) Load the contents of a file `sales.txt` located in HDFS into a DataFrame and use SQL to find the total sales pert part.

When ready copy the contents of `Terminal` screen with a report from implementation of a task (3) and paste/append it at the end of a file `solution4.rpt`.

Deliverables

A file `solution4.rpt` with a report from of implementation of the tasks (1), (2), and (3) .

<u>Submission of Assignment 3</u>

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.rpt`**, **`solution2.rpt`**, **`solution3.rpt`**, and **`solution4.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 **`ASSESSMENT ITEMS (ASSIGNMENTS)`**
(5) Click at **`In this place you can submit the outcomes of your work on the tasks included in Assignment 3`** 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 **`solution1.rpt`**, **`solution2.rpt`**, **`solution3.rpt`**, and **`solution4.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