School of Computing & Information Technology
University of Wollongong

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

ISIT312 Big Data Management
SIM S4 2024
Assignment 3

Scope

The objectives of Assignment 3 are the 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 **17 November 2024 by 9:00 pm** (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. Email submissions will not be accepted. 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 late submission penalty (25% of the total mark) will be applied for every 24 hours late.

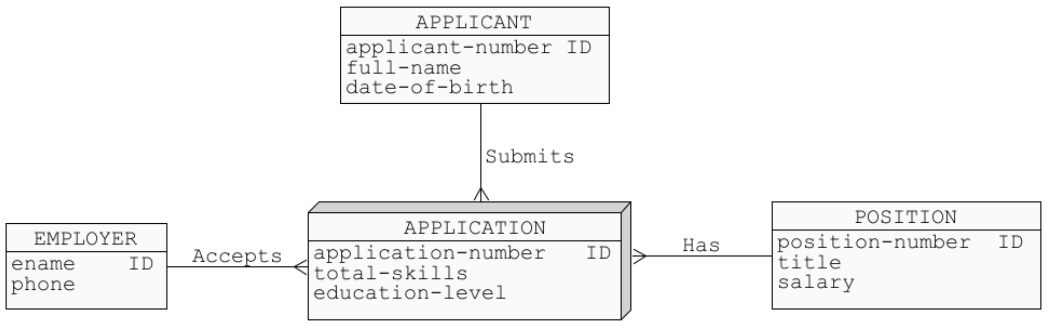
A submission that contains an incorrect file attached is treated as a correct submission with all consequences coming from the evaluation of the file attached.

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

The third 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)
Design and implementation of HBase table

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



Create an HBase script, `solution1.hb`, containing HBase shell commands that perform the following tasks:

1. Create an HBase table.
2. Load sample data into the table with details for:
    - At least **three applications**,
    - **Two employers**,
    - **Two applicants**, and
    - **Two positions**.

The data should include the following relationships:

- One applicant submits an application that involves **one employer** and **one position**.
- Another applicant submits **two applications**, each involving **a different employer** and **a different position**.

The specific values for the data are up to you but should be sensible and consistent with the relationships described.

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

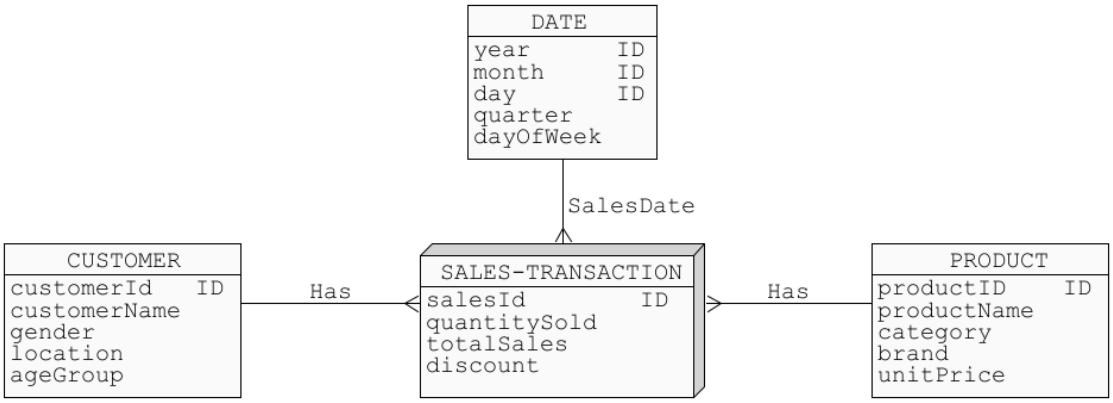
Deliverables

Submit the following file:

- `solution1.pdf`, containing the output report from the processing of the `solution1.hb` script, with the statements that create the HBase table and load the sample data.

Task 2 (6 marks)
Querying and manipulating data in HBase table

This assignment focuses on implementing HBase queries and data manipulations for a database schema that models customer involvement in sales transactions.



Instructions

1. **Set Up the HBase Table**
   - Download the task2.hb file, which contains HBase shell commands to create the HBase table task2 and load initial data into it.
   - Use the HBase shell to process task2.hb and confirm that the table task2 is created and populated with data.
2. **Implement Queries and Data Manipulations**
   - Use the HBase shell to write and save the following queries and data manipulations in a new file named solution2.hb:
     - **Query 1**: Retrieve all available information for the product with product number 101, listing each version in a separate cell.
     - **Query 2**: Retrieve all available information about sales of product 101 made by customer 007, listing each version in a separate cell.
     - **Query 3**: Retrieve the name, location (address), and age group of all customers, listing each customer in a separate cell.
     - **Query 4**: Retrieve all information about products under the brand 'Samsung', listing each product in a separate cell.
     - **Task 1**: Add a new column family named SELLER that will store seller information, including the seller's name, address, and contact person. Insert data for at least two sellers with reasonable values.
     - **Task 2**: Add information for two sales transactions, each by a different seller, with reasonable values.
     - **Task 3**: Increase the total number of versions in each cell of the PRODUCT column family to 4.

3. **Run and Capture Output**
    - After writing solution2.hb, use the HBase shell to process it.
    - Copy the output displayed in the Command window from the processing of solution2.hb and paste it into a file named solution2.pdf.

Deliverables

- Submit a file named solution2.pdf containing the Command window output from running solution2.hb, which includes a listing of the queries and data manipulations performed.

Task 3 (8 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
```

Load the sales related information listed above 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 `solution3.pdf`.

(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 `solution3.pdf`.

(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.pdf`.

Deliverables
A file `solution3.pdf` with a report from of implementation of the tasks (1), (2), and (3) .

Submission of Assignment 3

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 (**`A3-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 3`** link.
(6) Click at a button **`Add Submission`**
(7) Move the zipped file **`A3-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