ISIT312 Big Data Management
SIM S4 2025
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 **20 November 2025 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.
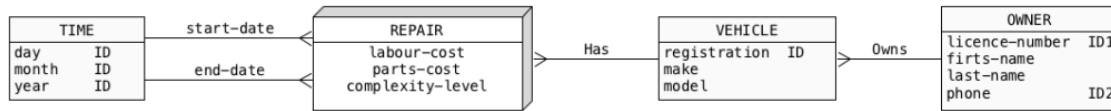
All files left on Moodle in a state **"Draft(not submitted)"** will not be evaluated.

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.

(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 two vehicles such that each one required two repairs. Do not forget about the owners of the vehicles.

When ready use HBase shell to process a script file `solution1.hb` and to save a report from processing in a file `solution1.pdf`.
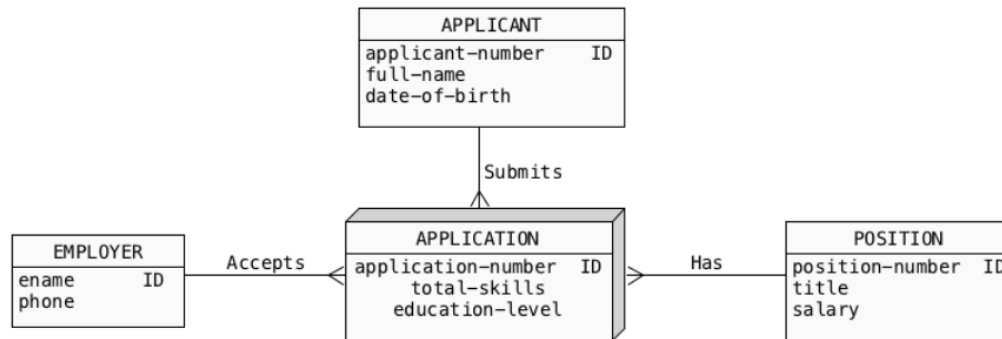
Deliverables
A file `solution1.pdf` that contains a report from processing of `solution1.hb` script with the statements that create HBase table and load sample data.

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

Consider a conceptual schema given below. The schema represents a simple database domain where applicants apply for the positions offered by employers.

Download a file `task2.hb` with HBase shell commands and use HBase shell to process it. Processing of a script `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 position, that has a number `312`, list one version per cell.
(2) Find all information about an application for a position `312` performed by an applicant `007`, list one version per cell.
(3) Delete a column family `EMPLOYER`.
(4) Add to a column family `POSITION` a column that contains information about the total number of applications. Next, update the values in the new column for all positions recorded in a database.
(5) Increase the total number of versions in each cell of a column family `APPLICANT`.

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 the processing of the script and paste the results into a file `solution2.pdf`. Save the file.

Deliverables
A file `solution2.pdf` with a listing from the processing of a script file `solution2.hb`.

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 (SP425) 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