|  |  |
| --- | --- |
| **Assignment Case** |  |
| COMP6579  Big Data Processing |
| **Computer Science** | **O213-COMP6579-AE02-01** |
| ***Valid on*** *Odd Semester Year 2020/2021* | **Revision 00** |

1. Seluruh mahasiswa tidak diperkenankan untuk:

*All students are not allowed to:*

* + - Melihat sebagian atau seluruh jawaban mahasiswa lain,

*Seeing a part or the whole answer from other student*

* + - Menyadur sebagian maupun seluruh jawaban dari buku,

*Adapted a part or the whole answer from the book*

* + - Mendownload sebagian maupun seluruh jawaban dari internet,

*Downloading a part or the whole answer from the internet,*

* + - Mengerjakan soal yang tidak sesuai dengan tema yang ada di soal,

*Working with another theme which is not in accordance with the existing theme in the matter of the case,*

* + - Melakukan tindakan kecurangan lainnya,

*Committing other dishonest actions,*

* + - Secara sengaja maupun tidak sengaja melakukan segala tindakan kelalaian yang menyebabkan hasil karyanya berhasil dicontek oleh orang lain / kelompok lain.

*Accidentally or intentionally conduct any failure action that cause the results of the project was copied by someone else / other groups.*

1. Jika mahasiswa terbukti melakukan tindakan seperti yang dijelaskan butir 1 di atas, maka **nilai mahasiswa** yang melakukan kecurangan (menyontek maupun dicontek) akan di – **NOL** – kan.

*If the student is proved to the actions described in point 1 above, the score of the student which committed dishonest acts (cheating or being cheated) will be “Zero”*

1. Perhatikan jadwal pengumpulan jawaban, segala jenis pengumpulan jawaban di luar jadwal tidak dilayani.

*Pay attention to the submission schedule, all kinds of submission outside the schedule will not be accepted*

1. Persentase penilaiaan untuk matakuliah ini adalah sebagai berikut:

*Marking percentage for this subject is described as follows:*

|  |  |  |
| --- | --- | --- |
| **Tugas Mandiri**  *Assignment* | **Proyek**  *Project* | **UAP**  *Final Exam* |
| 40% | - | 60% |

1. Software yang digunakan pada matakuliah ini adalah sebagai berikut:

*Software will be used in this subject are described as follows:*

|  |
| --- |
| **Software**  *Software* |
| VM Cloudera  Jupyter Notebook |

## Ekstensi file yang harus disertakan dalam pengumpulan tugas mandiri dan proyek untuk matakuliah ini adalah sebagai berikut:

*File extensions should be included in assignment and project collection for this subject are described as follows:*

|  |  |
| --- | --- |
| **Tugas Mandiri**  *Assignment* | **UAP**  *Final Exam* |
| DOCX, SQL | IPYNB |

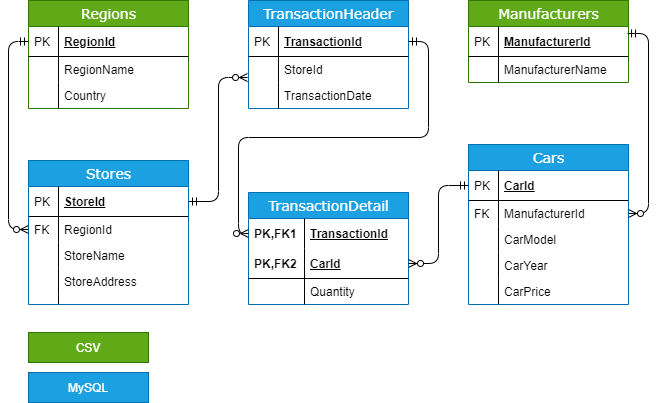
## Soal

*Case*

**AmorE’s Dealership**

**AmorE’s Dealership** is a popular car dealership chain in countries such as America, Canada, and Japan under **PT Software Laboratory.** As **AmorE’s Dealership** has grown rapidly, there is still a need to improve the sales more quickly. To do that, the owner **Amore** wants you to do an **analysis** from **different kinds** of data they have.

From the sales business process, the data can be analyzed to gain sales insight. The data is stored in **Comma-Separated** **Values (CSV)** file and **MySQL** dump file and the data schema is drawn using **Entity Relationship Diagram (ERD)** below:



**Figure 1. AmorE’s Dealership ERD**

You were given the task to gain some insight from the **sales** data using **Hadoop** tools. Below is the task you must do:

# **Load data from CSV to Hive**

Given the file “**Manufacturers**.**csv**” and “**Regions**.**csv**”, you were asked to load the data from **Comma-Separated Values** (**CSV**) file to **Hive** for data integration.

# **Load data from MySQL to Hive**

Given the file “**create+insert.sql**” that consists of the data about **sales**, **cars**, and **stores**. You need to load the data to **MySQL** database, then **ingest** the data from **MySQL** database to **Hive** for data integration.

# **Query Analysis**

From the data in **Hive**, you need to gain some sales insight in **AmorE’s Dealership**, below are some statements you need to answer using **Hive** / **Impala** query:

* 1. Show **cars** which has been bought **more than 30** units **in 2018.**
  2. Show **top 10 region** which has the **most transactions** **in America.**
  3. Show **top 3 stores** which earned the **most gross profit** **in 2019.**
  4. Show **car manufacturers** that have **sell cars more than the average car manufacturer** for cars that are made **after 2010.**
  5. Show **store and total cars sold** for each store who did **more than 5 transactions** within year **2018-2019** and **determine how much investment** will be given based on the total car sold on the **year 2018-2019**:

|  |  |
| --- | --- |
| Total Car Sold | Investment |
| > 200 | 100.000.000 |
| 100 - 200 | 50.000.000 |
| 30 – 99 | 20.000.000 |
| 20 – 29 | 10.000.000 |
| < 20 | 0 |

**Files to be collected**:

[NIM].txt that consist of:

* Command to Load data from CSV to Hive
* Command to Load data from MySQL to Hive
* Hive query for analysis