|  |  |
| --- | --- |
| **Assignment Case** |  |
| COMP6579  Big Data Processing |
| **Computer Science** | **O213-COMP6579-YT01-01** |
| ***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 |

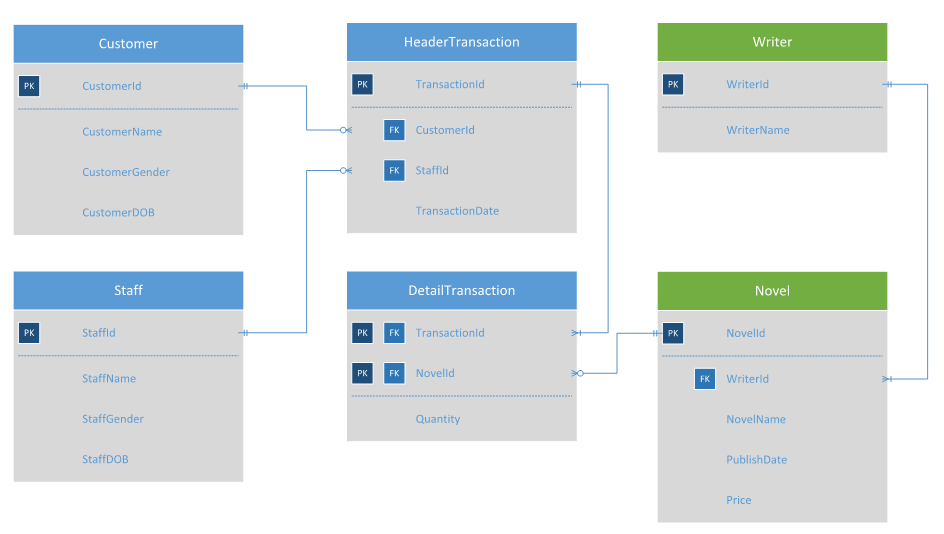
## Soal

*Case*

**Roman Store**

**Roman Store** is a new novel store developed by **Software Laboratory**. As Roman Store become more famous in the society, they wanted to do some analytics on their data.

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. Roman Store ERD***

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

1. **Load data from CSV to Hive**

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

1. **Load data from MySQL to Hive**

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

1. **Query Analysis**

From the data in **Hive**, you need to gain some sales insight in **Roman Store**, below are some statements you need to answer using **Hive** / **Impala** query:

* 1. Showthe **number of books sold** for **five oldest novels** writtenby **Raymund Shmyr.**
  2. Show **top 3 writers** who wrote the most books and was published **5 years ago.**
  3. Show **top 3 female customers** that do the **most transaction**.
  4. Show **male staff** who have **sold** **books** **more than the average**.
  5. Show **novel** and its **total price after discount per transaction** for novel that is published **after 1999** and the novel name must contain **at least 3 words**. The **discount per novel** will be based from its **total price** with the following condition:

|  |  |
| --- | --- |
| **Total Price Per Novel** | **Discount Per Novel** |
| More than 10000000 | 80% |
| More than 1000000 and below or equals 10000000 | 70% |
| More than 500000 and below or equals 1000000 | 50% |
| Others | 20% |

**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