|  |  |
| --- | --- |
| **Assignment Case** |  |
| COMP6579  Big Data Processing |
| **Computer Science** | **E203-COMP6579-AE01-01** |
| ***Valid on*** *Even Semester Year 2019/2020* | **Revision 00** |

## Soal

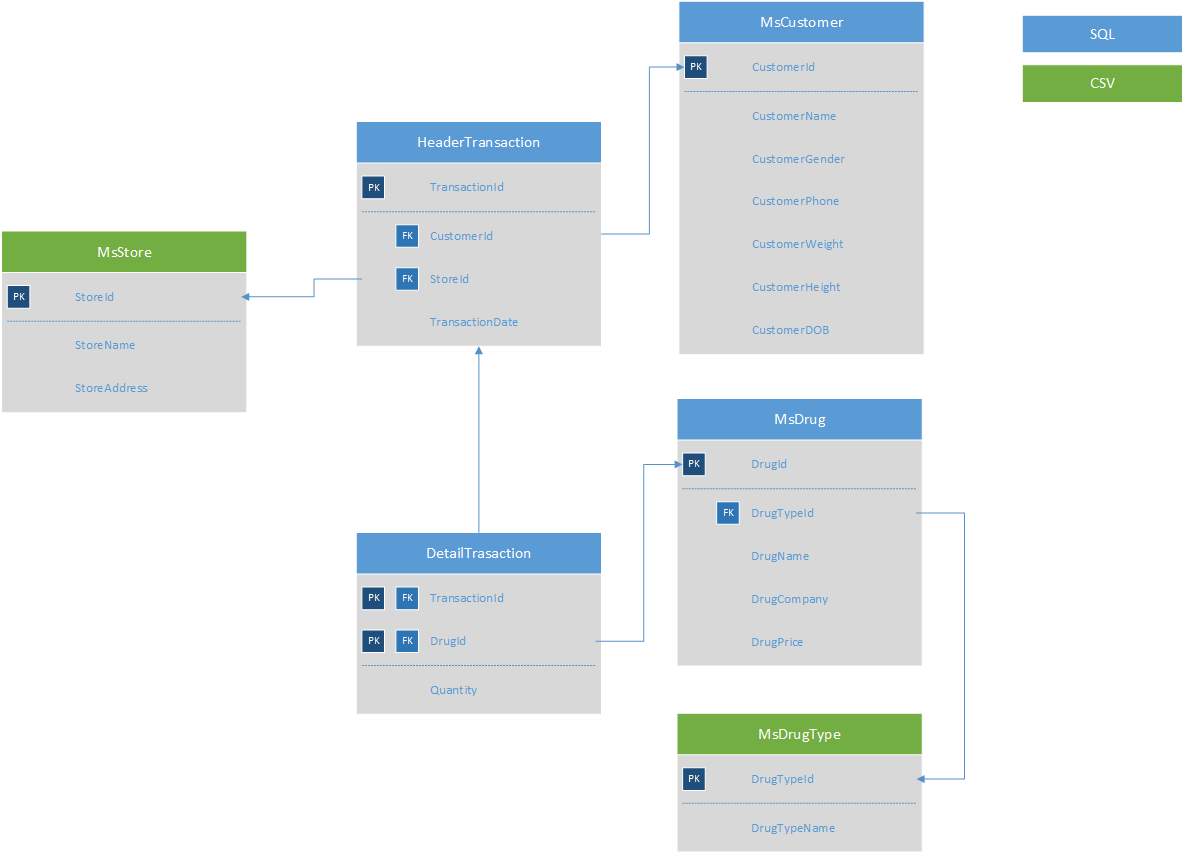
*Case*

**AmorE’s Pharmaceutical**

**AmorE’s Pharmaceutical** is a drug store chain based on Jakarta under **PT**. **Software Laboratory**. Since the sales are going high, they need to improve the store sales more quickly. To do that they intended to do some **analysis** from **different** **kinds** of data they have.

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

Figure 1. AmorE’s Pharmaceutical 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 “**MsDrugType**.**csv**” and “**MsStore**.**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**, **drug**, and **customer**. 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 Pharmaceutical**, below are some statements you need to answer using **Hive** / **Impala** query:

* 1. Show **drugs** which has been bought more than **50 pieces in a month in 2017**
  2. Show **top 10 store** which has the **most transactions in 2018**
  3. Show **top 3 customer** who spent the **most in 2018**
  4. Show **drug type** that has been **bought more than the average** drug type transaction
  5. Show **customer’s voucher** which is determined based on the total drugs they **bought per transaction in 2017 – 2019**. Voucher will only be acquired by customers who **bought more than 5 type of drugs in a transaction**. The value will be determined with terms and condition as below:

|  |  |
| --- | --- |
| Total Drug Bought | Voucher |
| >200 | 1.000.000 |
| 101 - 200 | 500.000 |
| 31 - 100 | 200.000 |
| 21- 30 | 100.000 |
| <21 | 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