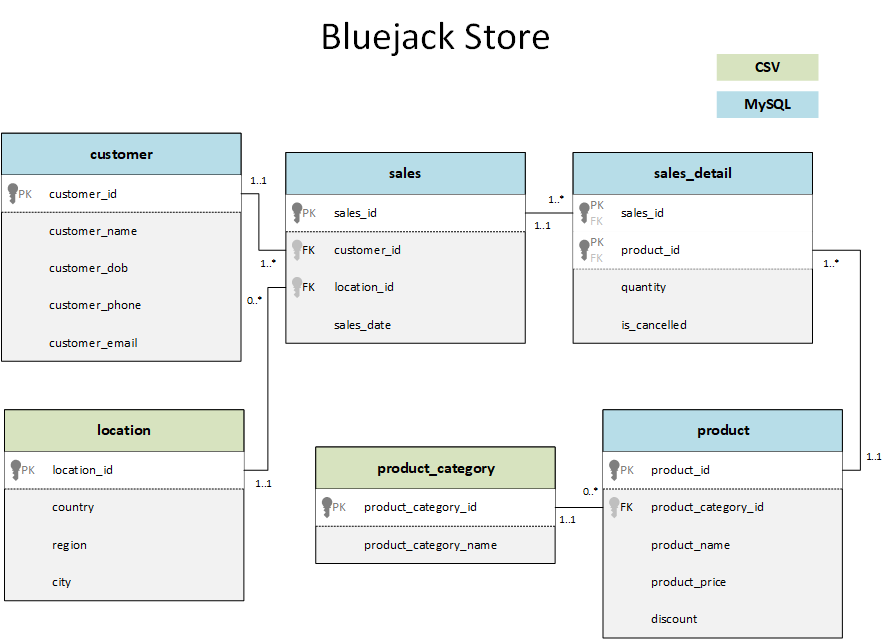
|  |  |
| --- | --- |
| **Soal Kualifikasi**  *Qualification Case* |  |
| COMP6579  Big Data Processing |
| **Post-Training**  *Case created on Odd 2019/2020* | **COMP6579-WL** |

**Bluejack Store**

**Bluejack Store** is a retail store based on Jakarta under **PT**. **Software Laboratory**. As **Bluejack Store** has grown rapidly and planned to “go public” there is a 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 **Bluejack Store**. 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**. **Bluejack 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:

# **Load data from CSV to Hive**

Given the file “**location**.**csv**” and “**product\_category**.**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 “**BluejackStore**.**sql**” that consists of the data about **sales**, **product**, 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 **Bluejack Store**, below are some statements you need to answer using **Hive** / **Impala** query:

* 1. Show the most profitable product category in 2019.
  2. Show the location where the most transactions occurred in the most recent year that transactions occurred.
  3. Show the most profitable product by calculating its revenue. When there is a discount cut the product price by discount. Show data with profit more than 15.000.000.
  4. Show customer who handled transactions more than the average of the total transactions by each customer.
  5. Show customers spending for last Christmas (December 2019). All customer who spends more than 10.000.000 will get a voucher for the next transaction with terms and condition as below:

|  |  |
| --- | --- |
| Spending | Voucher |
| 10.000.000 – 24.999.999 | 1.000.000 |
| 25.000.000 – 49.999.999 | 5.000.000 |
| > 50.000.000 | 10.000.000 |

**Files to be collected**:

Answer.docx that consist of:

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