|  |  |
| --- | --- |
| **Assignment Case** |  |
| COMP6549001  Big Data Processing |
| **Computer Science** | **O223-COMP6579-BR03-01** |
| ***Valid on*** *Odd Semester Year 2021/2022* | **Revision 00** |

1. Seluruh mahasiswa tidak diperkenankan untuk:

*All students are not allowed to:*

* + - Berdiskusi dan/atau bekerja sama dengan mahasiswa lainnya

*Discuss and/or work together with other student participants*

* + - Melihat sebagian atau seluruh jawaban mahasiswa lain

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

* + - Membuka dan menyalin dari **BUKU** atau **CATATAN**, **VIDEO** dari pengajar (recording kelas, VBL, Youtube, dsb) dan **REFERENSI** lainnya

*Open and copy from any resources such as notes, videos (class recording, VBL, Youtube, etc) and other references*

* + - Membuka dan menyalin jawaban dari internet (google, stackoverflow, dsb)

*Open and copy answer from the internet (google, stackoverflow, etc)*

* + - 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. Bila Anda tidak membaca peraturan ini, maka Anda dianggap telah membaca dan menyetujuinya

*If you have missed to read these regulations, so you are considered to have read and agreed on it*

1. Persentase penilaiaan untuk matakuliah ini adalah sebagai berikut:

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

|  |  |
| --- | --- |
| **Tugas Mandiri**  *Assignment* | **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 uap untuk matakuliah ini adalah sebagai berikut:

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

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

## Soal

*Case*

**BR Box**

**BR Box** is a music album store selling music from 1900s era. Since the sales are going high, they need to improve the store sales more quickly. To do that, **BR Box** need to do **analysis** from **different kinds** of data they have.

From the sales business process, some data can be analyzed to gain sales insight in **BR Box**. 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:

Diagram

Description automatically generated

Figure . BR Box 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 “**albums**.**csv**” and “**genres**.**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 **users** **and sales**. 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 **BR Box**, below are some statements you need to answer using **Hive** / **Impala** query:

* 1. Show **albums** and **total quantity that have been sold** which the **album** was **released in 2001**.
  2. Show **users** and their **total spending** in **year 2021**.
  3. Show **users** and their **total variant of** **purchased album genres,** which the **total variant of purchased genres** is **more than equal to 5**.
  4. Show **albums** and **total profit** which the **genre of the album is the most sold genre** **of all time**, **transaction date occurred in year of 2021,** and **the total profit is more than equal to 15000**.
  5. Show **albums**, **total quantity sold**, and **album rating label** which **album** **critic score and user score are not null**. The **album rating label** is determined with **following conditions**:

|  |  |
| --- | --- |
| Album rating label | Condition |
| Recommended | * Total quantity sold is more than average of all album’s total quantity sold. * Critic score >= 70 * User score >= 7 |
| Popular | Total quantity sold is more than average of all album’s total quantity sold. |
| Worth to try | - |

**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

**References:**

* https://www.kaggle.com/kauvinlucas/30000-albums-aggregated-review-ratings