UNIVERSITY OF MALAYA

Semester 1 (2023/2024)

Faculty of Computer Science & Information Technology

WQD7007

Big Data Management

Case Study Report

Name: Muhammad Asyraff bin Ponan

Matric number: S2128251

Lecturer: Dr. Hoo Wai Lam

Table of Contents

Question 1 (a)

Kaggle is chosen as the big resource for this case study. Generally, Kaggle is a platform where data science and machine learning enthusiasts obtain, explore and collaborate as a community. Kaggle allows users to collaborate and compete with one another to solve real-world problems. Kaggle is also suitable for a wide range of users, from beginners interested in data science and artificial intelligence to the world's most experienced data scientists. (Majhi, 2023).

Kaggle can be considered as big data resource based on the 7 Vs of big data. The first one is volume. Kaggle deals with large amounts of data, ranging from small to massive datasets used in a variety of competitions and projects. Users can access and analyse datasets of various sizes, allowing them to work on projects involving varying data volumes. According to the article *Kaggle: All You Need to Know About This Platform* (2023), GPUs, as well as a large amount of community-published data and code, are available for free. More than 50,000 public datasets and 400,000 public notebooks are available for everyone to use.

Furthermore, in terms of velocity, Kaggle supports rapid data generation and analysis. Kaggle has over 536,000 active members, receives nearly 150,000 submissions per month. Kaggle also organizes several data science competitions. Kaggle include a data set and a problem, and participants must create and submit a model that solves the problem or predicts the target variable with the highest accuracy while meeting tight deadlines. As a result, it will generate high-speed data (Coursera Staff, 2023).

In terms of veracity, we can observe from Kaggle itself where tons of datasets are from various users and sources in the community where the accuracy and reliability of data is not always trustable and trustworthy. Also, data quality can also be questioned because of the reality of data is mostly dirty. To deal with that, Kaggle encourages good data practices via community discussions, forums, and documentation (Coursera Staff, 2023).

In terms of variety, Kaggle datasets can come in various forms such as comma-separated values (CSV), JSON, ZIP archives and BigQuery. In addition to that, Kaggle notebooks can be in various forms too like scripts, RMakrkdown, Jupyter Notebook that use either R and Python (Majhi, 2023).

In terms of variability, Kaggle datasets and competitions can vary in terms of data distribution, format, and attributes. This variability requires users to adapt their approaches and models to different scenarios, which improves their ability to deal with a wide range of data challenges (Coursera Staff, 2023). Kaggle covers a wide range of topics, from attempting to predict the onset of cancer by examining patient records to analysing the movie reviews sentiment. The platform provides interesting and challenging projects for contributors to learn and practice (Kaggle: All You Need to Know About This Platform, 2023).

Next, data visualization is made possible in the Kaggle platform due to usage of Python package and also R that can demonstrate numerous interactive data exploration using graphical approach and also convincing data storytelling through Kaggle notebook (Majhi, 2023).

Lastly, the main aim of Kaggle datasets, notebook and competitions is to offer value through significant and actionable insights where data scientists and machine learning engineers compete to create the best models for solving specific problems or analysing certain data sets (Coursera Staff, 2023).

Question 1(b)

Kaggle offers a lot of notebook and datasets across all sectors and areas. By searching tourism keywords in the Kaggle website, we can explore numerous notebooks, discussions and datasets covering tourism. One of the interesting datasets found is about Thailand domestic tourism statistics from January 2019 to February 2023 (Thaweewat, 2023). The dataset consists of several attributes which includes province and region in Thailand, the number of tourists divided into 2 categories which are local Thailand citizens and foreign tourists. It also included some interesting statistics such as occupancy rate and profit generated in all province.

This dataset can be useful to tourism industry through data exploration to gain interesting and actionable insights

1. **Infrastructure Planning:** Use occupancy rate data to plan and optimize tourism infrastructure like hotels, transportation, and attractions.

2. **Regional variations:** Identify regional differences in tourist activity, net profit, and other metrics. This insight can assist policymakers and businesses in addressing gaps, allocating resources effectively, and promoting balanced development across multiple areas.

3. **Occupancy Rate Insights:** Examine occupancy rates across provinces and regions to identify popular destinations and areas that may require extra promotional efforts. High occupancy rates can indicate areas in high demand, whereas low rates might require action to promote tourism.

4. **Demand Forecasting:** Use historical data to forecast tourism demand in various provinces. This can help with resource allocation, infrastructure planning, and peak season planning.

Question 2

Since the Kaggle dataset found are defined by rows and columns, it is better to use a relational database and also structured query language (SQL) databases such as MySQL. The advantages and disadvantages of both databases are as follows:

1. Relational database

Advantages:

Simple and easy: In comparison to other database models, the relational database model is significantly simpler. Users can quickly access the information they need without having to cope with the database's complexity. The Structured Query Language (SQL) is used to run complex queries (Akhtar, 2021).

Flexibility: It is effortless to add, update, or delete tables, relationships, and other data changes as needed without affecting the overall database structure or existing applications (What Is a Relational Database (RDBMS)? | Google Cloud, n.d.).

Disadvantages:

Cost: The relational database system is expensive to set up and maintain. The initial cost of the software alone can be prohibitively expensive for smaller businesses (Akhtar, 2021)

Maintenance Problem: The maintenance of the relational database becomes more challenging over time as the volume of data increases. Database maintenance requires a significant amount of time from developers and programmers (Akhtar, 2021).

2. MySQL

Advantages:

Free and open source: MySQL is the preferred choice for start-ups and developers due to its free and open-source nature. MySQL is free, making it ideal for organization that prioritize cost-cutting. MySQL provides almost all of the features desired in a database server (MySQL Advantages and Disadvantages - Blue Claw Database Developer Resource, 2021).

Platform Support: MySQL supports any platform because it is a cross-platform database server. It is compatible with all operating systems, including Windows, Linux, MacOS, Linux Server, and Windows Server (Roomi, 2022).

Disadvantages:

Poor performance under high loads: MySQL is ideal for many use cases, but it is not appropriate for large enterprises with millions of records and transactions. MySQL does not support read and write operations at such high volumes, which is why this occurs (MySQL Advantages and Disadvantages - Blue Claw Database Developer Resource, 2021).

Stability: MySQL corruption is common due to issues such as lack of stability. This problem is most commonly encountered when performing auditing or transaction-related tasks (Roomi, 2022)

Question 3

Hive will be used to perform data access and analysis. First of all, the dataset downloaded from Kaggle is uploaded to the Ubuntu Terminal from Windows. The file is named as 'thailand_tourism_2019_2023_v2.csv'.

```
asyraff@User:~$ cp /mnt/c/Users/asyra/Downloads/thailand_tourism_2019_2023_v2.csv /home/asyraff
asyraff@User:~$ ls
 Batting.csv                  churn_reduced.csv      hive              pig-0.16.0.tar.gz       sample.txt
 Set02.csv                    derby.log              input             pig_1704003309131.log   sample1.txt
 Set05.csv                    grep_example           metastore_db      pig_1704003811399.log   sqoop
 apache-hive-2.3.9-bin.tar.gz hadoop                 metastore_db.tmp  pig_1704505784808.log   sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz
 churn.java                   hadoop-2.10.2.tar.gz   pig               'principal[.]*'         thailand_tourism_2019_2023_v2.csv
```

Then, upload the dataset to HDFS file directory '/case_study'.

```
asyraff@User:~$ hdfs dfs -put /home/asyraff/thailand_tourism_2019_2023_v2.csv /user/hdfs/case_study
```

The file is successfully uploaded to HDFS

```
asyraff@User:~$ hdfs dfs -ls /user/hdfs/case_study
Found 1 items
-rw-r--r--   1 asyraff supergroup    1634448 2024-01-21 20:53 /user/hdfs/case_study/thailand_tourism_2019_2023_v2.csv
```

Next, Apache Hive is launched on the Ubuntu terminal and Hive table named 'tourism' is created by defining the attributes in the dataset and specifying the data types as well. The Hive table is successfully created.

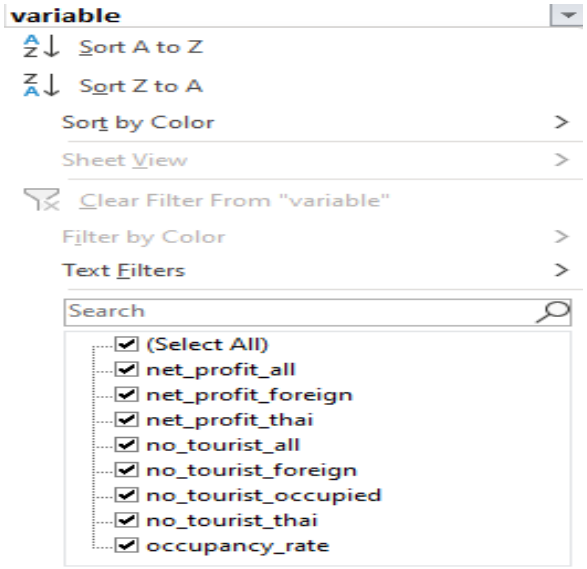
```
asyraff@User:~$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/asyraff/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/asyraff/hadoop/share/hadoop/common/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in jar:file:/home/asyraff/hive/lib/hive-common-2.3.9.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or
 using Hive 1.X releases.
hive> CREATE EXTERNAL TABLE IF NOT EXISTS tourism (
    > `date` STRING, province_eng STRING, region_eng STRING, variable STRING, value int
    > )
    > COMMENT 'Thailand domestic tourism stats'
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE
    > LOCATION '/user/hdfs/case_study'
    > ;
OK
Time taken: 4.74 seconds
```

All rows are loaded properly as the number of rows in Hive table matched the number of row in Excel data.

```
Time taken: 1.483 seconds, Fetched: 30801 row(s)
```

30797	1/12/2022	Ubon Ratchathani	east_northeast	net_profit_foreign	2.48
30798	1/12/2022	Sakon Nakhon	east_northeast	net_profit_foreign	3.52
30799	1/12/2022	Yasothon	east_northeast	net_profit_foreign	0.21
30800	1/12/2022	Amnat Charoen	east_northeast	net_profit_foreign	0.66
30801	1/12/2022	Nong Bua Lamphu	east_northeast	net_profit_foreign	0.2

The challenge when dealing in this dataset is that the variable such as 'occupancy_rate', 'net_profit_all' and others are defined as row output instead column attributes as demonstrated in the screenshot below.



Thus, separate Hive table need to be created for the desired variable so that we able to analyse the value of the variables for each province and region.

```
hive> CREATE TABLE occupancy_rate AS SELECT * FROM tourism WHERE variable = 'occupancy_rate';
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark
, tez) or using Hive 1.X releases.
Query ID = asyraff_20240121211841_d9ccabdf-2cd6-45b1-9fe5-e57e3be35de2
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1705837570589_0002, Tracking URL = http://User.localdomain:8088/proxy/application_1705837570589_0002/
Kill Command = /home/asyraff/hadoop/bin/hadoop job  -kill job_1705837570589_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2024-01-21 21:18:48,408 Stage-1 map = 0%,  reduce = 0%
2024-01-21 21:18:54,744 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.75 sec
MapReduce Total cumulative CPU time: 3 seconds 750 msec
Ended Job = job_1705837570589_0002
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://localhost:9000/user/hive/warehouse/.hive-staging_hive_2024-01-21_21-18-41_088_41944760607926623-1/-ext-10002
Moving data to directory hdfs://localhost:9000/user/hive/warehouse/occupancy_rate
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 3.75 sec   HDFS Read: 1639000 HDFS Write: 181866 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 750 msec
OK
Time taken: 15.488 seconds
```

The Hive table ‘occupancy_rate’ is successfully created to show the value of occupancy rate for each province. To access the top province with high occupancy rate, the latest date, 2022 is selected. New Hive table named occupancy_rate_2022 is created.

```
hive> CREATE TABLE occupancy_rate_2022 AS SELECT * FROM occupancy_rate WHERE `date` LIKE '%/2022';
```

Next, occupancy_rate_2022 is sorted to view the top 20 rows from the table.

```
hive> SELECT * FROM occupancy_rate_2022 ORDER BY value DESC LIMIT 20;
```

```
1/12/2022       Chiang Mai      north   occupancy_rate  95
1/12/2022       Chiang Rai      north   occupancy_rate  90
1/12/2022       Nan     north   occupancy_rate  87
1/10/2022       Buriram         east_northeast  occupancy_rate  86
1/12/2022       Chonburi        east    occupancy_rate  83
1/11/2022       Chiang Mai      north   occupancy_rate  82
1/12/2022       Lampang         north   occupancy_rate  82
1/12/2022       Phuket  south   occupancy_rate  82
1/12/2022       Nakhon Ratchasima       east_northeast  occupancy_rate  80
1/11/2022       Chiang Rai      north   occupancy_rate  79
1/11/2022       Phuket  south   occupancy_rate  79
1/2/2022        Chiang Rai      north   occupancy_rate  79
1/12/2022       Prachuap Khiri Khan     central occupancy_rate  79
1/12/2022       Rayong  east    occupancy_rate  78
1/12/2022       Tak     north   occupancy_rate  76
1/12/2022       Phayao  north   occupancy_rate  76
1/11/2022       Rayong  east    occupancy_rate  76
1/12/2022       Yala    south   occupancy_rate  76
1/9/2022        Buriram         east_northeast  occupancy_rate  75
1/12/2022       Lamphun         north   occupancy_rate  75
Time taken: 18.504 seconds, Fetched: 20 row(s)
```

From the screenshot, it is observed that Chiang Mai, Chiang Rai and Nan in the north region in the last month of the year 2022 recorded a very high occupancy rate. High occupancy rates indicate that there is a high demand for accommodation in a specific location at a given time. This can be useful knowledge for businesses in the hospitality sector in order to increase the accommodation service such as hotels, resorts and homestay to deal with high traffic of tourists during certain period.

Next, the data will be sorted again to analyse the total profit generated from each province throughout the year 2019 to 2023. Thus, tourism Hive table is sorted using net_profit_all values to show the net profit from all type of tourists from each month of the year included.

```
hive> CREATE TABLE net_profit_all AS SELECT * FROM tourism WHERE variable = 'net_profit_all';
```

The challenge now is to sum the profit gained from each year for the same province. To implement that, the sum function is used to total up the revenue gained in each province.

```
hive> CREATE TABLE `total_revenue` AS SELECT `province_eng`, SUM(`value`) AS total_revenue FROM `net_profit_all` GROUP BY `province_eng`;
```

Then, top 10 rows are generated.

```
Bangkok 1846943
Phuket  830753
Chonburi        483633
Chiang Mai      260018
Krabi   165777
Surat Thani     155867
Songkhla        112339
Prachuap Khiri Khan     111161
Chiang Rai      95533
Phetchaburi     83749
Time taken: 16.598 seconds, Fetched: 10 row(s)
```

From the screenshot, it is observed that Bangkok has the highest total revenue among others from the year 2019 to 2023. Thus, from this information, government, businesses and other tourism agency can focus more on this province in order to maintain or increase the net profit generated.

On the other hand, the bottom 10 province based on total revenue generated is processed by sorting the table in ascending order.

```
hive> SELECT * FROM total_revenue ORDER BY `total_revenue` ASC LIMIT 10;
```

```
Nong Bua Lamphu         873
Amnat Charoen   1051
Yasothon        1813
Pattani         2315
Sing Buri       2399
Ang Thong       2530
Maha Sarakham   2724
Kalasin         2802
Bueng Kan       2969
Chainat         3436
Time taken: 15.119 seconds, Fetched: 10 row(s)
```

From all the 10 regions shown from the screenshot above, involved organization and also governments can implement tourism promotion campaign to increase the revenue generated by promoting the landmarks, culture and other aspects that can attract local and foreign tourists to visit those provinces more frequently.

Question 4

Big data pipeline:

1. **Data source:**
   The tourism dataset is obtained from Kaggle website which is about Thailand domestic statistics. The dataset is saved to the device.

2. **Data pre-processing:**
   There are couple of columns that are loaded wrongly which the province and region in Thailand language. Using Microsoft Excel, those two columns are deleted because there are similar columns that represents the province and region which in English language. This is to ease the analysis part later on.

3. **Data storage:**
   The processed dataset is then uploaded to Hadoop HDFS via Ubuntu terminal.

4. **Data access and analysis:**
   Hive is very useful and easy for data access and analysis because it is SQL-like language. Every variable is analysed to discover interesting knowledge and statistics in order to make decision

5. **Data insights:**
   After data is carefully analysed, useful and fruitful insights can be gained for certain organization to act on in order to improve the tourism industry in Thailand such as policy making, promotional campaigns and others.

References

Akhtar, Z. (2021, August 2). Relational Database Benefits and Limitations (Advantages & Disadvantages) - DatabaseTown. *DatabaseTown*. https://databasetown.com/relational-database-benefits-and-limitations/#:~:text=The%20main%20benefits%20of%20using,issue%20of%20speed%20can%20arise.

Coding Ninjas. (n.d.). *Coding Ninjas Studio*. https://www.codingninjas.com/studio/library/kaggle

Coursera Staff. (2023, November 29). *What is kaggle and what is it used for?* Coursera. https://www.coursera.org/articles/kaggle

Kaggle : All you need to know about this platform. (2023, October 30). Data Science Courses | DataScientest. https://datascientest.com/en/kaggle-all-about-this-platform

Majhi, A. (2023, October 17). *Coding Ninjas Studio*. Coding Ninjas. https://www.codingninjas.com/studio/library/kaggle

MySQL Advantages and Disadvantages - Blue Claw Database Developer Resource. (2021, January 4). Blue Claw Database Developer Resource. https://blueclawdb.com/mysql/advantages-disadvantages-mysql/

Roomi, M. (2022, October 29). *5 Advantages and Disadvantages of MySQL | Limitations & Benefits of MySQL*. HitechWhizz - the Ultimate Tech Experience. https://www.hitechwhizz.com/2022/10/5-advantages-and-disadvantages-limitations-benefits-of-mysql1.html

What is a relational database (RDBMS)? | Google Cloud. (n.d.). Google Cloud. https://cloud.google.com/learn/what-is-a-relational-database

Link to the Kaggle dataset

https://www.kaggle.com/datasets/thaweewatboy/thailand-domestic-tourism-statistics