FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY

WQD7007
BIG DATA MANAGEMENT

SEMESTER 1, 2023/2024

CASE STUDY (GROUP 1)

Name	**Matric Number**
Kar Hong Sam	S2191926

CONTENT

1.0 Big Data Resource

Big data analytics integration is changing the tourism business in a dynamic way. Considered a sophisticated toolkit, it enables companies to analyze large datasets, exploring details beyond numbers to understand visitor behavior, preferences, and trends. This analytical method makes it easier to create experiences that are customized to each person's interests. But the significance goes beyond customization. Big data improves operational effectiveness by allowing companies to more accurately predict demand, streamline logistics, and effectively manage resources. It also plays a crucial part in risk management by offering perspectives on how to deal with unforeseen difficulties. Big data also improves marketing tactics by interpreting web platforms, social media, and reviews. Lastly, it advances sustainable practices by directing the sector toward more environmentally responsible choices.

According to information provided by (Airbnb, 2023), the website has significantly contributed to the expansion of the travel and tourism sector. This is explained by the fact that it has low entry barriers for new hosts and is popular with domestic tourists. Hence, the Singapore Airbnb dataset from Kaggle was chosen as the big data resource for tourism industry. Table 1 is the metadata of the dataset which contains the field name and its description.

Table 1: Metadata of Singapore Airbnb.

Field Name	**Description**
id	Unique identifier for each listing on Airbnb.
name	Descriptive name or title of the Airbnb listing.
host_id	Unique identifier for the host of the Airbnb listing.
host_name	Name of the host associated with the Airbnb listing.
neighbourhood_group	Geographical grouping of neighborhoods (if applicable).
neighbourhood	Specific neighborhood where the listing is located.
latitude	The geographic latitude of the listing's location.
longitude	The geographic longitude of the listing's location.
room_type	Type of room or accommodation offered (e.g., entire home, private room).
price	Cost per night for booking the Airbnb listing.
minimum_nights	Minimum number of nights required for booking.
number_of_reviews	Total number of reviews received for the Airbnb listing.
last_review	Date of the most recent review for the listing.
reviews_per_month	Average reviews received per month.
calculated_host_listings_count	The total number of listings managed by the host.
availability_365	Number of days the listing is available for booking within a year.

id	name	host_id	host_nam	neighbourl	neighbourl	latitude	longitude	room_typ	price	minimum_	number_o	last_review	reviews_p	calculated	availability_365
49091	COZICOMI	266763	Francesca	North Regi	Woodland	1.44255	103.7958	Private roo	83	180	1	10/21/2013	0.01	2	365
50646	Pleasant R	227796	Sujatha	Central Re	Bukit Tima	1.33235	103.7852	Private roo	81	90	18	12/26/2014	0.28	1	365
56334	COZICOMI	266763	Francesca	North Regi	Woodland	1.44246	103.7967	Private roo	69	6	20	10/1/2015	0.2	2	365
71609	Ensuite Ro	367042	Belinda	East Regio	Tampines	1.34541	103.9571	Private roo	206	1	14	8/11/2019	0.15	9	353
71896	B&B Roor	367042	Belinda	East Regio	Tampines	1.34567	103.9596	Private roo	94	1	22	7/28/2019	0.22	9	355
71903	Room 2-ne	367042	Belinda	East Regio	Tampines	1.34702	103.961	Private roo	104	1	39	8/15/2019	0.38	9	346
71907	3rd level Ju	367042	Belinda	East Regio	Tampines	1.34348	103.9634	Private roo	208	1	25	7/25/2019	0.25	9	172
241503	Long stay a	1017645	Bianca	East Regio	Bedok	1.32304	103.9136	Private roo	50	90	174	5/31/2019	1.88	4	59
241508	Long stay a	1017645	Bianca	East Regio	Bedok	1.32458	103.9116	Private roo	54	90	198	4/28/2019	2.08	4	133
241510	Long stay a	1017645	Bianca	East Regio	Bedok	1.32461	103.9119	Private roo	42	90	236	7/31/2019	2.53	4	147
275343	Convenien	1439258	K2 Guesth	Central Re	Bukit Mera	1.28875	103.8081	Private roo	44	15	18	4/21/2019	0.23	32	331
275344	15 mins to	1439258	K2 Guesth	Central Re	Bukit Mera	1.28837	103.811	Private roo	40	30	10	9/13/2018	0.11	32	276
289234	Booking fo	367042	Belinda	East Regio	Tampines	1.34561	103.9598	Private roo	417	2	12	1/1/2019	0.14	9	239
294281	5 mins wal	1521514	Elizabeth	Central Re	Newton	1.31125	103.8382	Private roo	65	2	125	8/22/2019	1.35	6	336
324945	20 Mins to	1439258	K2 Guesth	Central Re	Bukit Mera	1.28976	103.809	Private roo	44	30	13	2/2/2019	0.15	32	340
330089	Accomo@	1439258	K2 Guesth	Central Re	Bukit Mera	1.28677	103.8124	Private roo	40	30	10	4/27/2019	0.14	32	331
330095	10 mins to	1439258	K2 Guesth	Central Re	Bukit Mera	1.28537	103.8109	Private roo	31	90	3	8/22/2016	0.04	32	361
344803	Budget sho	367042	Belinda	East Regio	Tampines	1.34943	103.9595	Private roo	49	2	45	8/11/2019	0.5	9	357

Figure 1: Sample data from Singapore Airbnb.

The Airbnb dataset serves as a crucial big data resource, embodying the quintessential characteristics of big data - volume, velocity, veracity, visualization, and value. Boasting a daily data production of 20 TB and an accumulated dataset reaching 1.4 PB according to ProjectPro (2023), its sheer volume is both impressive and indicative of its continuous growth. This substantial data influx necessitates high velocity, ensuring swift processing through various stages, from generation to storage in a data lake. Veracity takes center stage in maintaining the reliability of the dataset. The imperative lies in meticulous validation, guaranteeing the accuracy of crucial details such as listing descriptions, prices, and availability. This, in turn, establishes a trustworthy foundation for analytical processes and decision-making related to Airbnb listings. Incorporating visualization techniques becomes paramount to extract meaningful insights from the dataset. Geographical distribution maps, generated using latitude and longitude data, provide an intuitive understanding of property locations. Visualization extends to depicting trends in price distribution, room type popularity, and review counts, enhancing comprehension through charts, graphs, and maps. Beyond its utility for individual listings, the dataset's value is expansive. Hosts and property managers benefit from strategic decision-making, encompassing pricing optimization, effective marketing strategies, and enhanced customer satisfaction. Furthermore, its impact transcends individual stakeholders, influencing the broader tourism industry. Identified trends have the potential to inform policy decisions, guide investment strategies, and shape marketing campaigns.

In essence, the Airbnb dataset transcends being merely a repository of raw data; it functions as a transformative force, converting vast amounts of information into actionable knowledge. Its comprehensive representation of the 5V's of big data—volume, velocity, veracity, visualization, and value—positions it as an invaluable resource with the potential to significantly benefit the tourism industry in Malaysia. By leveraging this dataset, stakeholders within the industry can make well-informed decisions, drawing insights into trends, pricing strategies, and customer preferences. Hosts and property managers can optimize pricing and marketing strategies, leading to increased customer satisfaction and enhanced competitiveness. Additionally, policymakers can utilize identified trends to inform tourism-related policies, and investors can make strategic decisions based on valuable insights derived from this rich source of data. The Airbnb dataset, with its diverse and dynamic information, thus becomes a key instrument in shaping the trajectory of the tourism sector in Malaysia.

2.0 Big Data Storage

In the context of managing Airbnb's structured CSV dataset within the tourism industry, prioritizing MySQL as a storage solution offers several advantages. MySQL, renowned for its role as a reliable relational database management system (RDBMS), excels in handling well-structured data, aligning seamlessly with the tabular nature of the Airbnb dataset. Its support for SQL enables efficient execution of complex queries and analytics, providing a familiar environment for developers and analysts. MySQL's ACID compliance ensures data integrity, a crucial aspect in the tourism industry where accurate and consistent information is paramount. Additionally, MySQL's normalization capabilities mitigate redundancy and enhance data integrity, particularly advantageous when dealing with datasets that may have overlapped information. The maturity, stability, and extensive community support associated with MySQL further contribute to its suitability for managing the Airbnb dataset. While other big data tools such as Hive, HBase, MongoDB, and Spark are potent solutions for large-scale, unstructured datasets, the structured nature of the Airbnb dataset makes MySQL a pragmatic choice, ensuring efficiency and reliability in data storage and retrieval within the tourism industry. While other big data tools such as Hive, HBase, MongoDB, and Spark are potent solutions for large-scale, unstructured datasets, the structured nature of the Airbnb dataset makes MySQL a pragmatic choice, ensuring efficiency and reliability in data storage and retrieval within the tourism industry.

3.0 Demonstration of Storing and Accessing Big Data Resources

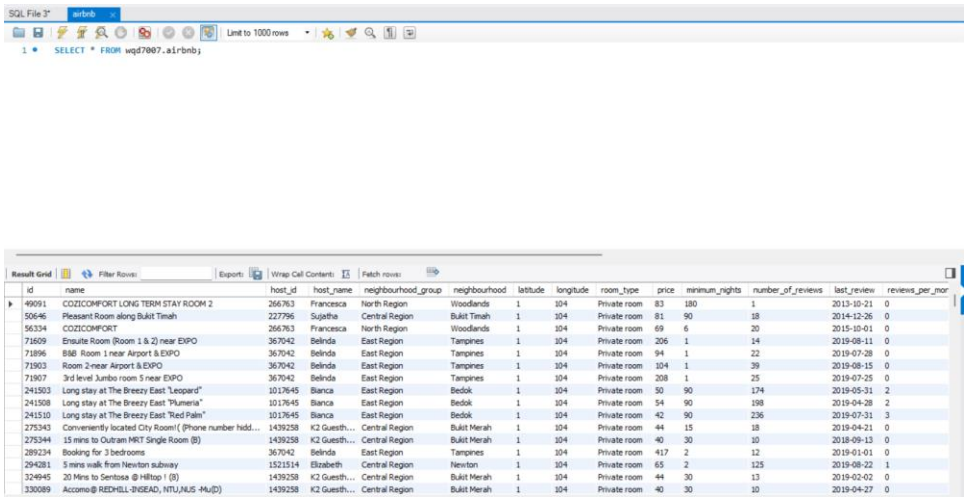
To set up the Airbnb dataset in MySQL Workbench, a SQL script was meticulously crafted. The initial step involved downloading the Singapore Airbnb dataset from Kaggle (Singapore Airbnb, 2019). Subsequently, within MySQL Workbench, a database named 'wqd7007' was established. A script was then devised to create a table labeled 'airbnb,' encompassing requisite columns with specified data types. Finally, the downloaded CSV file was loaded into the 'airbnb' table, facilitating seamless integration of the dataset into the MySQL Workbench environment. This comprehensive process ensures that the dataset is not only acquired but also effectively structured and integrated for further analysis within the MySQL framework.

```sql
DROP TABLE IF EXISTS airbnb;

-- Create a new table
CREATE TABLE airbnb (
    id INT,
    name VARCHAR(255),
    host_id INT,
    host_name VARCHAR(255),
    neighbourhood_group VARCHAR(255),
    neighbourhood VARCHAR(255),
    latitude INT,
    longitude INT,
    room_type VARCHAR(255),
    price INT,
    minimum_nights INT,
    number_of_reviews INT,
    last_review DATE NULL,
    reviews_per_month INT NULL,
    calculated_host_listings_count INT,
    availability_365 INT
);

LOAD DATA INFILE 'C:/Users/kh/Desktop/wqd7007/finals/listings.csv'
INTO TABLE airbnb
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 ROWS;
```

Figure 2: MySQL script to create table and load CSV dataset



```sql
SELECT * FROM wqd7007.airbnb;
```

id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_mon
49091	COZICOMFORT LONG TERM STAY ROOM 2	266763	Francesca	North Region	Woodlands	1	104	Private room	83	180	1	2013-10-21	0
50646	Pleasant Room along Bukit Timah	227796	Sujatha	Central Region	Bukit Timah	1	104	Private room	81	90	18	2014-12-26	0
56334	COZICOMFORT	266763	Francesca	North Region	Woodlands	1	104	Private room	69	6	20	2015-10-01	0
71609	Ensuite Room (Room 1 & 2) near EXPO	367042	Belinda	East Region	Tampines	1	104	Private room	206	1	14	2019-08-11	0
71896	B&B Room 1 near Airport & EXPO	367042	Belinda	East Region	Tampines	1	104	Private room	94	1	22	2019-07-28	0
71903	Room 2-near Airport & EXPO	367042	Belinda	East Region	Tampines	1	104	Private room	104	1	39	2019-08-15	0
71907	3rd level Jumbo room 5 near EXPO	367042	Belinda	East Region	Tampines	1	104	Private room	208	1	25	2019-07-25	0
241503	Long stay at The Breezy East "Leopard"	1017645	Bianca	East Region	Bedok	1	104	Private room	50	90	174	2019-05-31	2
241508	Long stay at The Breezy East "Plumeria"	1017645	Bianca	East Region	Bedok	1	104	Private room	54	90	198	2019-04-28	2
241510	Long stay at The Breezy East "Red Palm"	1017645	Bianca	East Region	Bedok	1	104	Private room	42	90	236	2019-07-31	3
275343	Conveniently located City Room!((Phone number hidd...	1439258	K2 Guesth...	Central Region	Bukit Merah	1	104	Private room	44	15	18	2019-04-21	0
275344	15 mins to Outram MRT Single Room (B)	1439258	K2 Guesth...	Central Region	Bukit Merah	1	104	Private room	40	30	10	2018-09-13	0
289234	Booking for 3 bedrooms	367042	Belinda	East Region	Tampines	1	104	Private room	417	2	12	2019-01-01	0
294281	5 mins walk from Newton subway	1521514	Elizabeth	Central Region	Newton	1	104	Private room	65	2	125	2019-08-22	1
324945	20 Mins to Sentosa @ Hilltop ! (B)	1439258	K2 Guesth...	Central Region	Bukit Merah	1	104	Private room	44	30	13	2019-02-02	0
330089	Accomo@ REDHILL-INSEAD, NTU,NUS -Mu(D)	1439258	K2 Guesth...	Central Region	Bukit Merah	1	104	Private room	40	30	10	2019-04-27	0

Figure 3: Sample data queried from MySQL Workbench

Identifying meaningful insights from Airbnb data is crucial for business owners to comprehend traveler preferences, enhance customer experiences, and, consequently, maintain competitiveness in the market. Here are some analyses that can help business owners to extract the information by using SQL.

Queries and Results:

1. Discovering the Top 10 Highest-Priced Listings

```
SELECT name, host_name, neighbourhood_group, neighbourhood, room_type, price
FROM airbnb
ORDER BY price DESC
LIMIT 10;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

name	host_name	neighbourhood_group	neighbourhood	room_type	price
Comfortable & Quiet Master Bedroom	Yolivia	West Region	Bukit Panjang	Private room	10000
YOUR entire PRIVATE LUXURY PENTHOUSE condo unit	Jj	West Region	Tuas	Entire home/apt	10000
Testing	David	Central Region	Kallang	Private room	10000
The Club Residences - Contemporary Manor	Darren	Central Region	Southern Islands	Entire home/apt	8900
The Club Residences - Contemporary Manor (A)	Darren	Central Region	Southern Islands	Entire home/apt	8900
P	Yin	Central Region	Outram	Private room	7000
Lakeside Master room of condo 裕廊湖畔公寓主人房	X-Roy	West Region	Jurong East	Private room	7000
Love	Lily	West Region	Bukit Batok	Entire home/apt	6944
Hotel style master bedroom	Jo	Central Region	Kallang	Private room	6000
旅行者	Xia	East Region	Bedok	Private room	5000

Figure 4: Query 1 and the results

Based on the result, it turns out that Central Region is the highest price region in Singapore mainly because it has the highest population density compared to another region and the economic hub also within that region.

2. Discovering the Average Price per Room Type

```
SELECT room_type, AVG(price) AS avg_price
FROM airbnb
GROUP BY room_type;
```

Result Grid | Filter Rows: | Export:

room_type	avg_price
Entire home/apt	226.9983
Private room	110.9385
Shared room	65.6751

Figure 5: Query 2 and the results

Upon examining the query results, it becomes evident that entire homes or apartments tend to cost more than private and shared rooms. This correlation makes sense as the square footage of an entire house is typically larger.

3. Identifying the Top 10 Hosts with the Most Listings

```
SELECT host_id, host_name, neighbourhood_group, COUNT(*) AS total_listings
FROM airbnb
GROUP BY host_id, host_name, neighbourhood_group
ORDER BY total_listings DESC
LIMIT 10;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

host_id	host_name	neighbourhood_group	total_listings
66406177	Jay	Central Region	237
8492007	Alvin	Central Region	202
209913841	Richards	Central Region	156
29420853	Aaron	Central Region	140
2413412	Kaurus	Central Region	112
31464513	Darcy	Central Region	112
219550151	Rain	Central Region	111
23722617	Alex	Central Region	84
8948251	Joey	Central Region	83
159804766	Xiaoyu	Central Region	79

Figure 6: Query 3 and the results

The results indicate that most of the listings belong to property owners situated in the central region of Singapore.

4. Identifying the Top 10 Listings with the Highest Availability

```
SELECT name, host_id, host_name, neighbourhood_group, neighbourhood, room_type, price, availability_365
FROM airbnb
ORDER BY availability_365 DESC, price
LIMIT 10;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

name	host_id	host_name	neighbourhood_group	neighbourhood	room_type	price	availability_365
Walking Distance to MRT	23994004	Orion	North Region	Sembawang	Private room	19	365
New bto common room for rent at canberra crescent	24214608	Akmal	North Region	Sembawang	Private room	19	365
Cosy single bedroom in Yishun SGP	63207270	Jiaquan	North Region	Yishun	Private room	21	365
Jurong West Spacious Common Room (near NTU)	44165891	Vikneswaran	West Region	Jurong West	Private room	22	365
SINGLE BED CAPSULE IN HOSTEL RM 2	46545593	Meadows	Central Region	Kallang	Shared room	22	365
SINGLE BED CAPSULE IN HOSTEL RM 3	46545593	Meadows	Central Region	Kallang	Shared room	22	365
SINGLE BED CAPSULE IN HOSTEL RM4	46545593	Meadows	Central Region	Kallang	Shared room	22	365
SINGLE BED CAPSULE IN HOSTEL RM5	46545593	Meadows	Central Region	Kallang	Shared room	22	365
COSY SINGLE BED CAPSULE IN HOSTEL 7	46545593	Meadows	Central Region	Kallang	Shared room	22	365
Mu	168666642	Moe	North Region	Woodlands	Private room	22	365

Figure 7: Query 4 and the results

Private rooms and shared rooms are always available throughout the year in Singapore due to the high accommodation demand from tourists.

5. Analyzing the Average Reviews and Price by Different Room Type

```
SELECT room_type,
       AVG(reviews_per_month) AS avg_reviews_per_month,
       AVG(number_of_reviews) AS avg_number_of_reviews,
       AVG(price) AS avg_price
FROM airbnb
GROUP BY room_type;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

room_type	avg_reviews_per_month	avg_number_of_reviews	avg_price
Entire home/apt	1.0333	12.1837	226.9983
Private room	0.9830	13.6099	110.9385
Shared room	0.7184	12.4619	65.6751

Figure 8: Query 5 and the results

Based on the findings, the average monthly review count barely surpasses 1, which could be attributed to either a less effective review system provided by Airbnb or the infrequent practice of leaving reviews among tourists. Additionally, the average prices shed light on the market rates for various room types, revealing that shared rooms emerge as the most economical option. However, it's important to note that opting for a shared room does come at the expense of sacrificing personal privacy.

4.0 Big Data Pipeline

A data pipeline plays a pivotal role in managing the entire lifecycle of data, facilitating a seamless flow from data generation to decision-making. By orchestrating various stages, including data generation, acquisition, storage, analysis, and visualization, a well-constructed data pipeline ensures the efficient movement of data across these phases. It streamlines the process of collecting and transforming raw data into meaningful insights, providing a structured framework for analysis. This systematic flow enhances the accessibility and quality of data, enabling decision-makers to derive valuable insights more effectively. Moreover, the automated nature of a data pipeline reduces manual intervention, minimizing errors and improving the data's overall reliability. Ultimately, a well-designed data pipeline empowers organizations to make informed decisions by delivering timely, accurate, and actionable insights derived from their data sources.

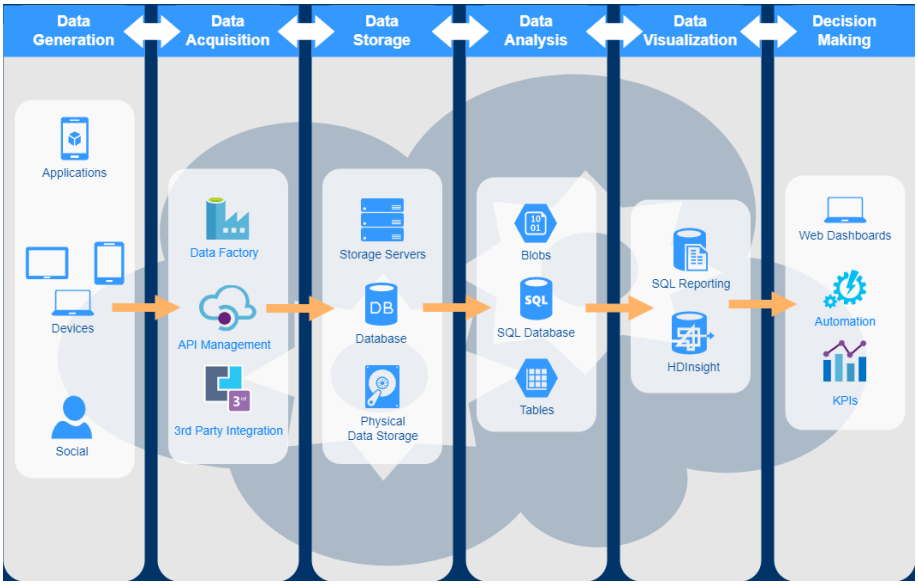


Figure 9: A proposed big data pipeline for Airbnb data

The proposed big data pipeline includes the six phases of big data as below:

1. Data Generation: This marks the initial steps in the big data process, where information is generated from various sources. Airbnb data, for instance, can originate from mobile applications, websites, and various social media platforms.
2. Data Acquisition: This phase involves collecting and storing data in a centralized platform, whether cloud-based or on-premises databases, for subsequent transformation or analysis. In this data pipeline, the information can be gathered through APIs or third-party integrations, facilitating the extraction and transfer of Airbnb data to the preferred data storage.
3. Data Storage: This step involves storing a variety of data, be it structured or unstructured, in a suitable data storage platform. MySQL was chosen as the data storage solution due to the tabular format of the collected data, making it easy to retrieve and manage.
4. Data Analysis: This entails understanding the data to uncover trends through the application of data analytics methods such as descriptive and predictive analytics.
5. Data Visualization: To enhance comprehension, data is visualized, making information more accessible than traditional tabular views. Insights are quickly gleaned, with charts highlighting trends in a meaningful way.
6. Decision Making: The final step involves management making decisions based on data-driven results, completing the data pipeline's journey from generation to informed decision-making.

References:

Airbnb. (2023, March 21). New report highlights Airbnb's contribution to inclusive growth of tourism. Airbnb Newsroom. https://news.airbnb.com/new-report-highlights-airbnbs-contribution-to-inclusive-growth-of-tourism/

ProjectPro. (2023, November 17). How data science increased Airbnb's valuation to $25.5 bn? https://www.projectpro.io/article/how-data-science-increased-airbnbs-valuation-to-25-5-bn/199#:~:text=propel%20its%20growth%3F-

Singapore Airbnb. (2019, August). Singapore Airbnb [Dataset]. Kaggle. https://www.kaggle.com/datasets/jojoker/singapore-airbnb/data