UNIVERSITI
MALAYA

WQD7009

Big Data Applications and Analytics

Group 2

Final Alternative Assessment

Name	Low Boon Kiat
Matric Number	17138399
Submission Date	23rd January 2024

Question 1

(a) Discuss any FIVE (5) data processing and analysis problems associated with traditional smart cities applications.

The first data processing and analysis problems associated with traditional smart cities applications is scalability issues. Traditional systems struggle to handle the increasing volume of data generated by smart city applications. Traditional systems often have fixed resource capacities that are not easily expandable to accommodate the growing data influx. As the number of sensors and connected devices increases, it will lead to performance bottleneck.

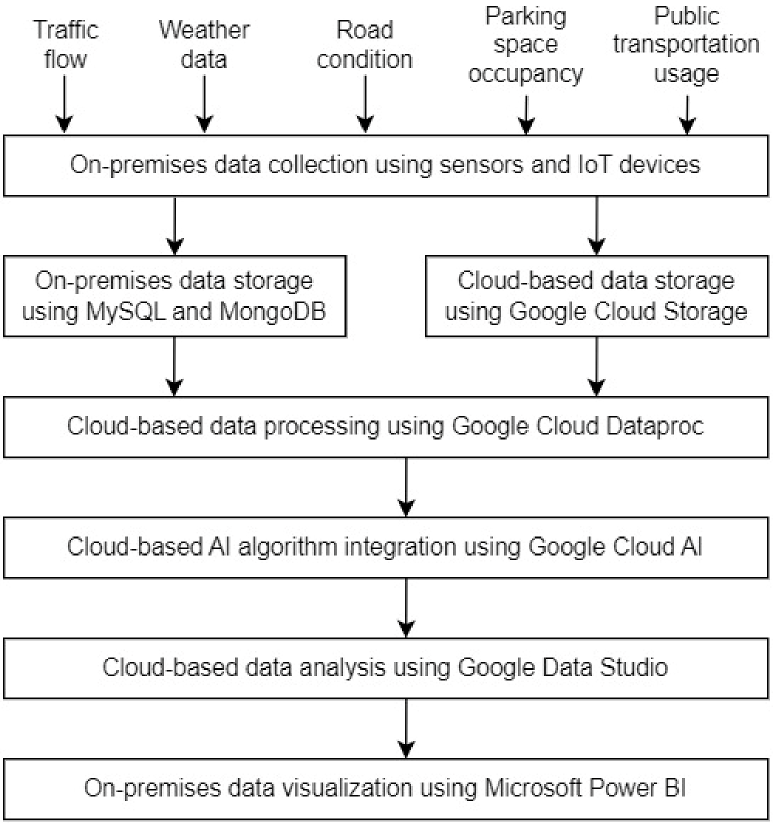
The second data processing and analysis problems associated with traditional smart cities applications is limited real-time processing. Traditional data processing systems are not designed for real-time data processing. They often rely on batch processing or periodic updates. Smart city applications such as traffic management require immediate analysis of data for effective decision-making, making real-time processing an essential requirement.

The third data processing and analysis problems associated with traditional smart cities applications is lack of predictive analytics. Traditional systems often lack advanced analytics capabilities such as predictive analytics. Machine learning algorithms can enhance smart city applications by predicting trends and future developments based on historical data.

The fourth data processing and analysis problems associated with traditional smart cities applications is data silos. Traditional systems often store data in separate databases specific to each city. Siloed data often comes with different formats. This makes it challenging to integrate information from different city systems, leading to incomplete analysis and decision-making.

The fifth data processing and analysis problems associated with traditional smart cities applications is security concerns. Traditional systems may lack robust security measures, exposing sensitive data to potential breaches. With the increasing interconnectedness of smart city devices, it provides more entry points for data breaches.

(b) Based on the above case study, propose an on-premises and cloud-based big data technologies application for smart city data processing, detailing each component of the application in an appropriate diagram. The proposed application must include a generative AI algorithm or relevant AI tools for analysing smart city application data.


Traffic flow
Weather data
Road condition
Parking space occupancy
Public transportation usage
On-premises data collection using sensors and IoT devices
On-premises data storage using MySQL and MongoDB
Cloud-based data storage using Google Cloud Storage
Cloud-based data processing using Google Cloud Dataproc
Cloud-based AI algorithm integration using Google Cloud AI
Cloud-based data analysis using Google Data Studio
On-premises data visualization using Microsoft Power BI

The architecture illustrates the flow of data through various stages, starting with data collection and storage, followed by cloud-based processing, AI integration, analysis, and lastly on-premises visualization. The integration of on-premises and cloud resources provides a scalable and reliable solution for smart city data processing, particularly on traffic management.

- Data Collection (On-Premises):

Sensors and IoT devices are deployed throughout the city to collect various types of data such as traffic flow, weather data, road condition, parking space occupancy, and public transportation usage.

- Data Storage (On-Premises and Cloud):

On-premises infrastructure such as MySQL and MongoDB stores data locally for short-term storage. On-premises database servers provide low-latency access to local data for real-time traffic monitoring where immediate data retrieval is essential.

Cloud-based storage such as Google Cloud Storage is used for scalable and durable long-term storage of large volumes of data. It supports analytics and machine learning applications that require access to historical data.

- Data Processing (Cloud):

Cloud-based distributed computing services such as Google Cloud Dataproc provide scalability for processing large datasets and enhance data quality through data cleaning. This allows the system to handle varying workloads of data cleaning.

- AI Algorithm Integration (Cloud):

Cloud-based AI services such as Google Cloud AI host the generative AI algorithm. By leveraging the scalable computational power of cloud services, generative AI algorithms can analyse historical traffic data and generate predictive models that can provide insights into optimizing the traffic flow.

- Data Analysis (Cloud):

Cloud-based analytics tools such as Google Data Studio utilize the generative AI output as part of their input dataset. City planners and stakeholders can leverage these tools to interactively explore and analyse the generative AI-generated insights.

- Data Visualization (On-Premises):

On-premises data visualization tool such as Microsoft Power BI serves as a local interface for city planners and stakeholders to access the critical information in real-time. With low-latency access to critical data, local dashboards are suitable for traffic management scenarios where real-time insights are essential.

(c) How can the proposed on-premises and cloud-based smart city application ensure cost-effectiveness during the implementation and deployment processes? Include a cost analysis.

The first strategy to ensure cost-effectiveness in the implementation and deployment of a smart city application is to optimize the hybrid infrastructure. For scenarios with low-latency requirements, on-premises infrastructure is used for short-term data storage to ensure quick access to critical data. On the other hand, we embrace the pay-as-you-go model of cloud services to take advantage of scalable resources. Cloud services are used for long-term storage to ensure cost-effective storage of historical data. Cloud services are used for data processing so that we pay only for the resources consumed during actual usage, avoiding upfront infrastructure costs.

The second strategy to ensure cost-effectiveness in the implementation and deployment of a smart city application is to use cloud-based monitoring and optimization tools to monitor resource utilization and identify opportunities for cost savings. To rectify the inefficiencies, we can implement resource optimization strategies such as autoscaling in the cloud. This ensures that computational resources scale up or down based on demand, minimizing unnecessary costs during periods of inactivity.

Cost analysis for the smart city application:

Component	**Cost**	**Optimization**
Cloud-based data storage (Google Cloud Storage)	Calculated based on the amount of data stored over time.	Implement lifecycle policies to automatically transition infrequently accessed data to lower-cost storage classes.

Cloud-based data processing (Google Cloud Dataproc)	Calculated based on the number of virtual machines (VM) and the duration of use.	Implement automatic scaling to adjust the number of VMs based on processing needs.
Cloud-based AI services (Google Cloud AI)	Calculated based on Application Programming Interface (API) usage.	Use efficient model architectures to minimize processing costs. Implement caching strategies to reduce redundant API calls.
On-premises data storage (MySQL, MongoDB), on-premises data visualization (Microsoft Power BI)	Initial setup costs for hardware and software licenses.	Regularly assess hardware requirements and consider hardware upgrades when necessary.

Question 2

(a) Develop and explain a DataOps architecture diagram for data visualization and analysis, considering data from various sources such as Excel files, SAS files, Postgres files, real-time logs, MySQL, or other sources. Include an analysis of the Entrepreneurship use case in the data ops architecture. The diagram should seamlessly integrate these sources into a streaming data pipeline capable of handling volume, variety, and velocity. Ultimately, it should channel the processed data to visualization tools like PowerBI, Tableau, or any other data visualization software.

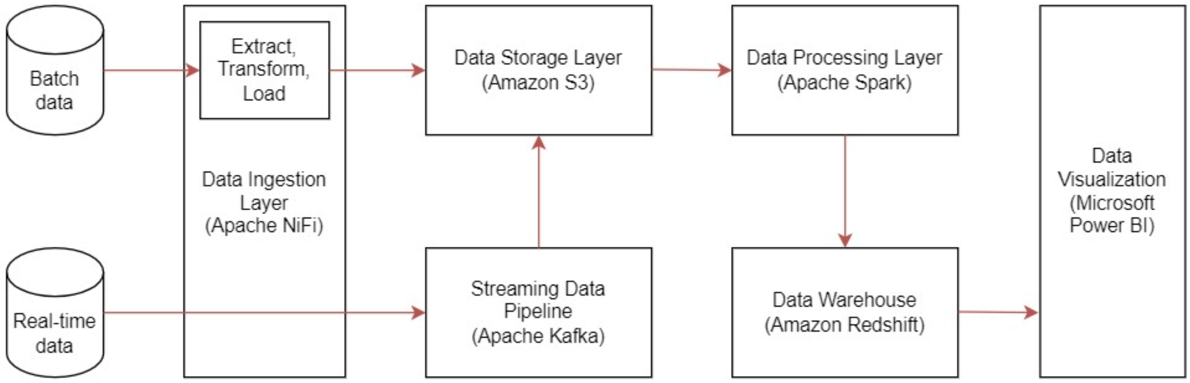

Batch data
Extract, Transform, Load
Data Ingestion Layer (Apache NiFi)
Data Storage Layer (Amazon S3)
Data Processing Layer (Apache Spark)
Data Visualization (Microsoft Power BI)
Real-time data
Streaming Data Pipeline (Apache Kafka)
Data Warehouse (Amazon Redshift)

Figure shows the DataOps architecture diagram for the given use case on Entrepreneurship, which consists of the following components.

- Data Sources:

Input data comes from various data sources including Excel files, SAS files, Postgres databases, real-time logs, and MySQL databases. The entrepreneurship data sources may consist of real-time data such as real-time logs of student interactions with online entrepreneurship platforms, live feedback from entrepreneurship training programs, etc as well as batch data such as historical survey responses, annual reports on entrepreneurship training programs, etc.

- Data Ingestion Layer (Apache NiFi):

Data ingestion layer applies ETL processes (Extract, Transform, Load) onto the batch data using Apache NiFi to extract entrepreneurship data from different sources, transform it into a unified format, and load it into data storage layer. When real-time data enters data ingestion layer, it bypasses traditional ETL processes and flows directly into the streaming data pipeline.

- Streaming Data Pipeline (Apache Kafka):

Apache Kafka serves as a distributed streaming platform that handles real-time data streaming. It decouples producers (data sources) from consumers (data processing layers), allowing for independent scalability of different components within the pipeline. Real-time data are processed in real-time through the streaming data pipeline.

- Data Storage Layer (Amazon S3):

Data storage layer acts as a centralized repository for both batch and real-time data. Amazon S3 is used for scalable and durable long-term storage of the entrepreneurship data. It supports large volumes of data and provides flexibility in accessing historical data.

- Data Processing Layer (Apache Spark):

When analytical processing is needed, both batch and real-time data can be retrieved from the data storage layer and processed using Apache Spark. It performs processing such as aggregations and calculations on the data. In this case, Apache Spark performs analyses on the stored data, identifying trends in student entrepreneurial competencies.

- Data Warehouse (Amazon Redshift):

Data warehouse stores curated and processed data for analytical queries, providing an optimized environment for complex queries. Amazon Redshift is used as a scalable data warehouse for efficient querying and reporting on entrepreneurial attitudes and competencies.

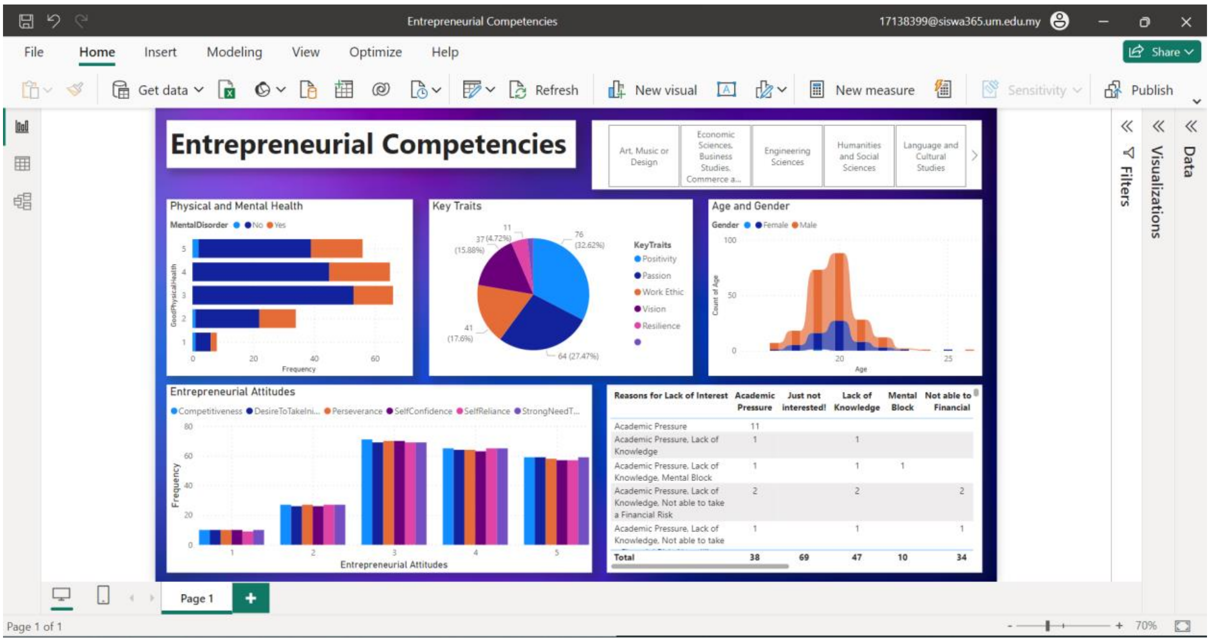
- Data Visualization Tool (Microsoft Power BI):

Visualization tools integrate with the data warehouse for the creation of interactive dashboards. These tools enable stakeholders to visualize key metrics and insights related to entrepreneurial competency among students.

(b) Develop a comprehensive dashboard using Power BI, or Tableau, or a similar data visualization tool for the provided entrepreneurship dataset. The dashboard should incorporate up to five graphs illustrating key parameters. Attach evidence of the dashboard in screenshots and include the published link in the answer document.

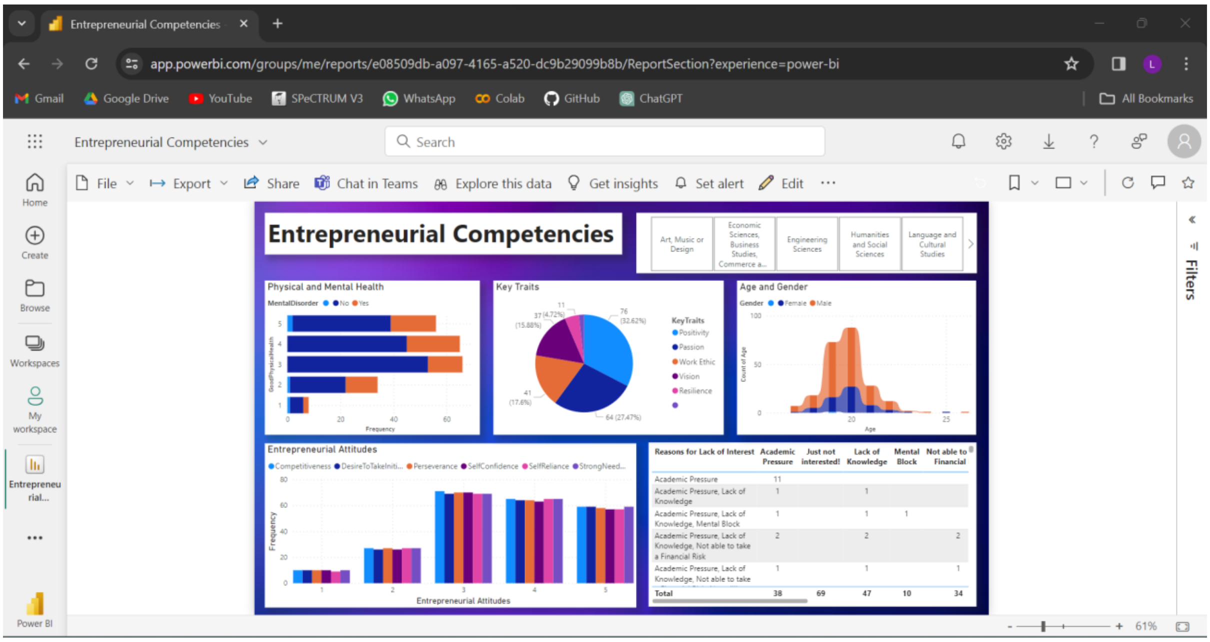
Screenshots:

- Create in Power BI Desktop.



Reasons for Lack of Interest	Academic Pressure	Just not interested!	Lack of Knowledge	Mental Block	Not able to Financial
Academic Pressure	11				
Academic Pressure, Lack of Knowledge	1		1		
Academic Pressure, Lack of Knowledge, Mental Block	1		1	1	
Academic Pressure, Lack of Knowledge, Not able to take a Financial Risk	2		2		2
Academic Pressure, Lack of Knowledge, Not able to take	1		1		1
Total	38	69	47	10	34

- Published to My Workspace.



Reasons for Lack of Interest	Academic Pressure	Just not interested!	Lack of Knowledge	Mental Block	Not able to Financial
Academic Pressure	11				
Academic Pressure, Lack of Knowledge	1		1		
Academic Pressure, Lack of Knowledge, Mental Block	1		1	1	
Academic Pressure, Lack of Knowledge, Not able to take a Financial Risk	2		2		2
Academic Pressure, Lack of Knowledge, Not able to take	1		1		1
Total	38	69	47	10	34

Published Link:
https://app.powerbi.com/groups/me/reports/e08509db-a097-4165-a520-dc9b29099b8b/ReportSection?experience=power-bi

Google Drive:
https://drive.google.com/file/d/1k878mev8pElzvM8lDj_4ZIfRqT0k2pI1/view?usp=sharing

(c) Formulate five questions related to the dataset and graphs from the developed dashboard. Provide answers to these questions using the graphs and narrative, constructing a cohesive and informative story that enhances understanding of the entrepreneurship dataset and its implications.

1. What key traits are prevalent among the students?
   Positivity accounted for 32.62% of the count of key traits, followed by passion at 27.47% and work ethic at 17.6%.

2. How is the physical and mental health of the students?
   Level 3 accounted for 28.82% of the count of good physical health, followed by level 4 at 28.38% and level 5 at 24.45%. 27.47% of the students have mental disorder.

3. How is the age distribution of the students?
   38.1% of students are 20 years of age, followed by 31.6% of students at 19 years of age and 12.12% of students at 21 years of age.

4. What are the primary reasons for lack of entrepreneurial willingness in the students?
   Just not interested and lack of knowledge are the primary reasons for lack of entrepreneurial willingness in students.

5. What are the average entrepreneurial attitudes of the students?
   Between 1 to 5, the average entrepreneurial attitudes of the students are 3.35 for perseverance, 3.62 for desire to take initiative, 3.59 for competitiveness, 3.72 for self-reliance, 3.89 for strong need to achieve, and 3.56 for self-confidence.

The dataset on entrepreneurship reveals that prevailing traits among aspiring entrepreneurs include high levels of positivity (32.62%), passion (27.47%), and work ethic (17.6%). However, concerns arise regarding the mental health of 27.47% of students. The age distribution highlights a concentration of aspiring entrepreneurs around 20 years old (38.1%). Barriers to entrepreneurial willingness are primarily attributed to lack of knowledge and disinterest. Average entrepreneurial attitudes reflect moderate scores across perseverance, initiative, competitiveness, self-reliance, need to achieve, and self-confidence. These insights underscore the need for targeted initiatives addressing mental health, educational pressures, and fostering an environment conducive to entrepreneurship for an innovative future.