WIE3007 – Data Mining and Warehousing

Individual Assignment

IoT-based Energy Consumption Monitoring in Retail Store

Amir Firdaus Bin Abdul Hadi

17204620/2

Contents

1. Introduction……………………………………………………………………………..3

2. Requirement Analysis
    a. Business Problems………………………………………………………………4
    b. Project Management…………………………………………………………….5

3. Design Model
    a. Data Schema Model……………………………………………………….….8
    b. Data and Sensors Needed………………………………………………...……9

4. ETL Architecture
    a. Architecture Diagram……………………………………………………...…11
    b. Design Details and Process…………………………………………………11

5. References…………………………………………………………………………….13

1.0 Introduction

The IoT or Internet of Things is a game-changing technology approach that enhances the communication between electronic devices and sensors to further assist our lifestyle. IoT utilises smart devices and internet to promote innovative and new solutions to problems related to various field specifically businesses. Smart city, smart homes, pollution control, smart transportation, smart industries are some of the result transformations due to IoT.

One of the things IoT can provide in retail is the ability to analyse and offer real-time monitoring and tracking the energy consumption information of equipment and buildings. For US retailers, after labour, rent and marketing costs, energy is in top 4 biggest in-store operating cost. This calls for efforts for cost reduction. Moreover, cost is not the only reason retailers should be focusing in reducing energy usage. Many companies and business should have the obligations to do their responsibility in the effort against climate change and carbon emissions.

In a retail store, a convenience store in particular, a lot of data can be generated from multiple sources. Sources such as freezer, lights, etc. In addition to that, as time goes by, many data about energy consumption will accumulate in the data storage to even the terabyte level and more if IoT is implemented with the added sensors throughout the premises. Traditional data processing techniques and architecture will be struggling to meet the system demands. It is a challenging task to process and monitor data in a near real-time scenario. It is vital to develop a cost-effective high performance data pipeline and architecture to serve the demands of the storage and manage the data from the sensors, legacy data, and analyse energy consumption. One of it is by implementing the correct data mining and data warehousing approach.

2.0 Requirement Analysis

2.1 Business Problems

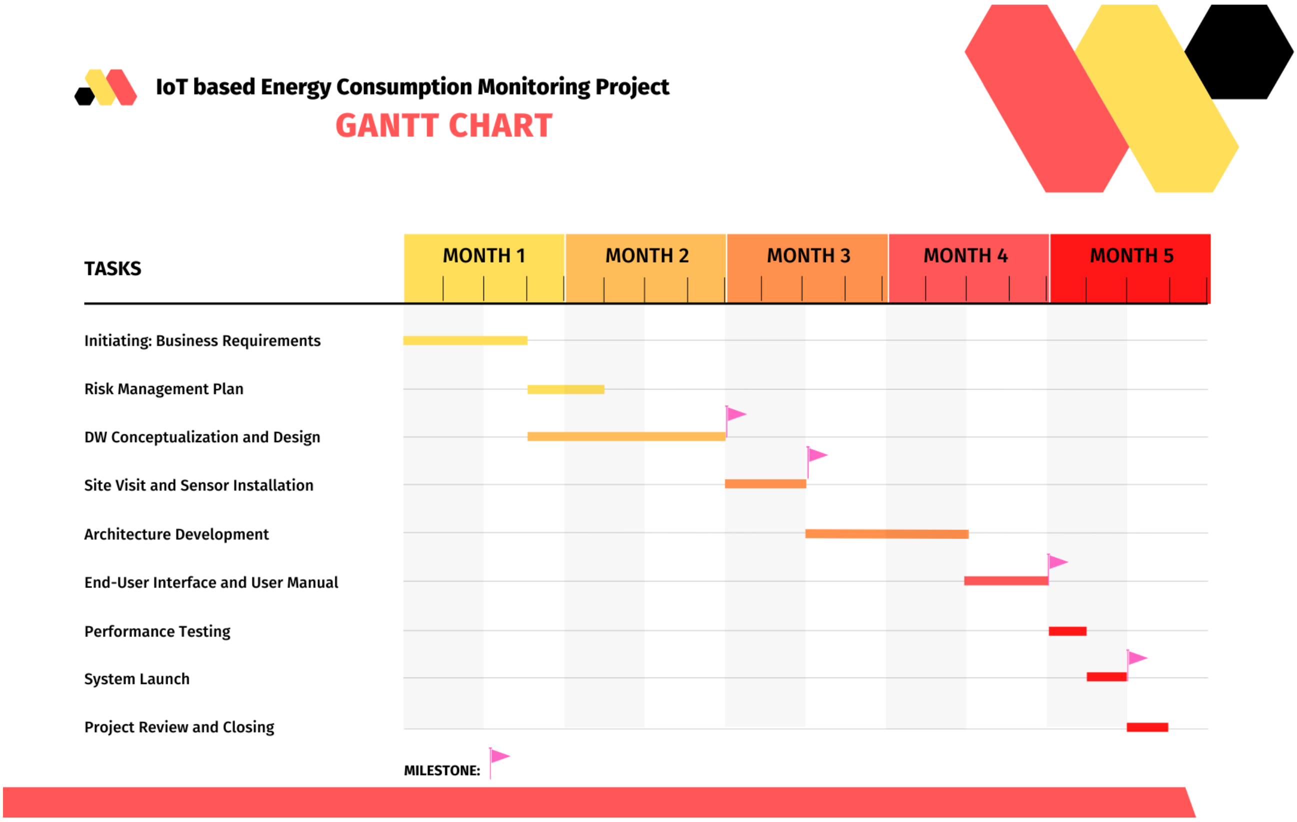
Reducing energy consumption is a vital aspect and a goal in retail stores in minimizing cost and managing resources efficiently. As a retail business grows and branching to multiple stores, tracking energy uses poses difficulty to managers and stakeholders. The business needed a proper way of monitoring energy consumption to provide information and optimization steps in order to achieve the goal.

2.2 Key Issues

- What kind of sensors needed to monitor energy consumption?
- What kind of technologies suitable to be used for data warehousing and big data analysis?
- From the monitoring, how the stakeholders or managers should approach to optimize energy usage?

2.3 Project Management

2.3.1 Project Timeline and Milestone



Phase	Task	Estimated Days	Responsibility
1 Initiating	1.1 Discussion with Stakeholders	7	Project manager
	1.2 Identify Business Objectives and Scopes	3	Project manager
	1.3 Identify Problems and Requirements	4	Project manager
	1.4 Preliminary Research	14	All
2 Planning	2.1 Conceptualization and Platform Selection	14	Business analyst
	2.2 Develop Risk Management Plan	3	Project manager
	2.3 Design Data Warehouse Architecture	21	DW system analyst, DW solution architect
	3.1 Equipment Check and Install Sensor	14	IoT engineer

3 Implementation	3.2 Develop Data Pipelining and Warehouse	14	Data engineer, DevOps engineer
	3.3 Develop Data Mining Algorithm	7	DW solution architect
	3.4 Develop Data Policies	7	DW solution architect
	3.5 Develop End-user Interface	14	Data Scientist, DW Administrator
	3.6 Data Warehouse Performance Testing	3	QA engineer
	3.7 System Launch	2	DW Administrator
	3.8 Create User Manual and Conduct Training	2	DW Administrator
4 Closure	4.1 Finalize Project and Review	3	All
	4.2 Project Closure	1	Project manager
	Total Days	133	

2.3.2 Roles and Responsibility

Roles	Responsibility
Project Manager	• Maintain project scheduling, updates project to stakeholders • Defining project scope such as objectives and deliverables
Business Analyst	• Documenting the details of the Data Warehouse (DW) solution • Outlining functional and non-functional requirements and limitations of data warehouse
DW System Analyst	• Identify specification for data warehouse system requirements • Analysing data sources, processes and analytics tools
DW Solution Architect	• Develop a data warehouse design solution architecture • Develop data governance strategy • Determining data warehouse tech stack
Data Engineer	• Developing the ETL/ELT process • Designing data models, schema and structures • Developing and maintaining a data pipeline from multiple sources into data warehouse
QA Engineer	• Preparing test and performance strategy • Review DW tech design documents • Evaluate the developed DW solution

DevOps Engineer	• Setting up the DW software infrastructure • Introducing continuous integration/continuous deployment (CI/CD) pipelines to automate and streamline data warehouse development processes
Data Warehouse Administrator	• Performs technical administration duties for the development and maintenance of data warehouse.
IoT Engineer	• Develop and setting up sensors. • Install sensors and provide technician supports to sensor deployment.
Data Scientist	• Develop end-user interface from data warehouse to reporting tools • Creates tracking and monitoring dashboards

2.3.3 Major Outcome

I. Able to design data warehousing architecture that enables energy consumption tracking and optimization.

II. Able to generate data required from multiple sensors type.

III. Able to use the data acquired for monitoring, analysing and optimising energy consumption.

IV. Able to reduce energy consumption and costs upon implementing the proposed IoT solution.

3.0 Design Model

3.1 Data Schema – Snowflake Model

Snowflake schema is used for this data warehouse instead of star schema as its dimension tables are normalized. This schema uses less storage to store dimension tables although becomes more complex. Snowflake schemas reduces the occurrence of redundant data, which are easier to manage. Snowflake are a better option for data warehouses while star schemas are good for data marts due to it simple relationships. In addition to that, this schema is more compatible with many OLAP database modelling tools. The multidimensional data model includes:

SHOP_SENSOR_FACT_TABLE, SENSOR_DIM_TABLE, EQUIPMENT_DIM_TABLE, SHOP_DIM_TABLE, ENERGY_SENSOR, DOOR_SENSOR, TEMPERATURE_SENSOR, MANAGER_TABLE, MANUFACTURER_TABLE

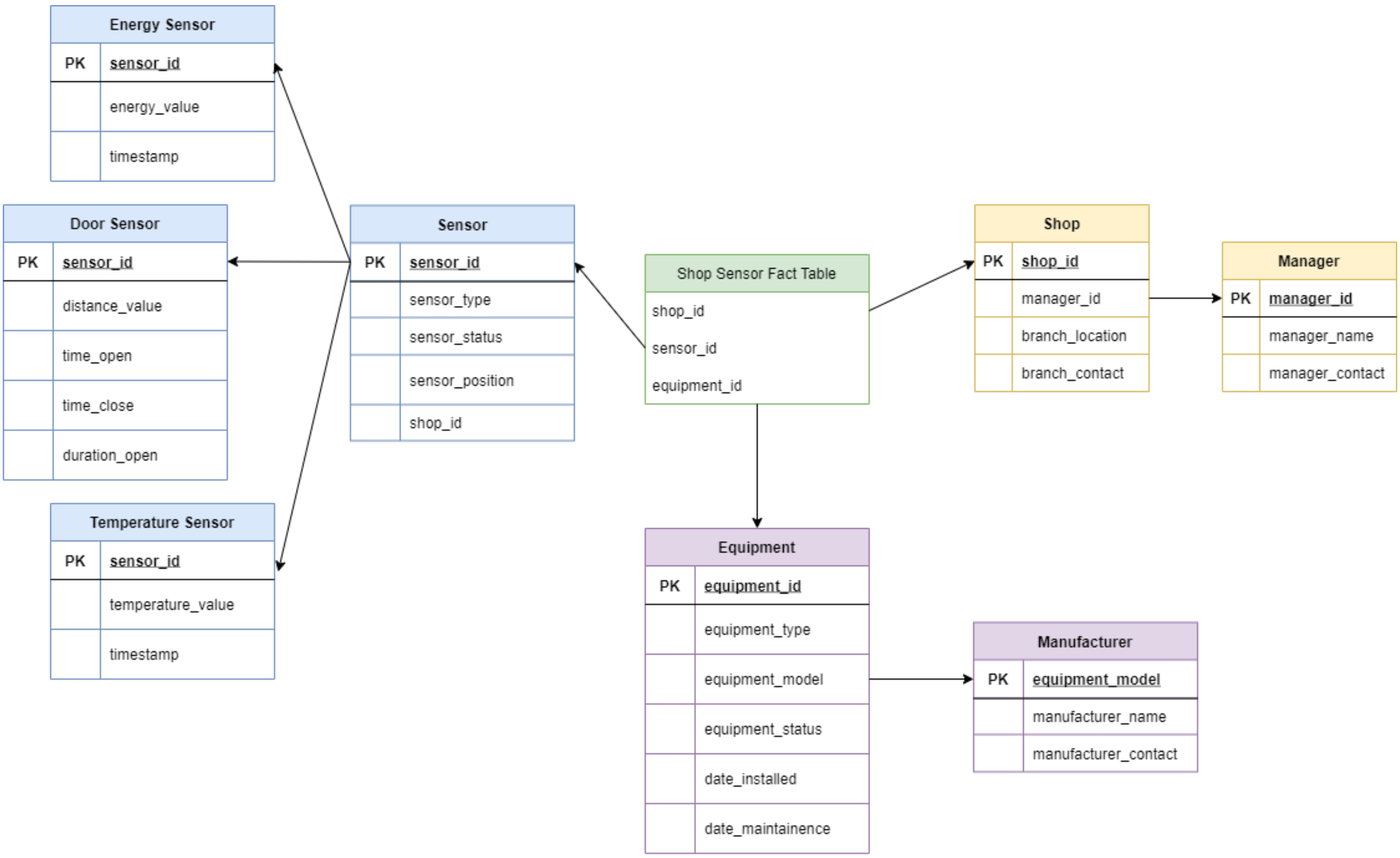


Figure: Snowflake schema for retail store energy consumption monitoring data warehouse

3.2 Data and Sensors Needed

a) **Shop**

To store shop branch locations and person-in-charge. To keep track which sensors or equipment in which shop branch.

a. **Manager**

Information about the shop branch's person-in-charge as they are the one who will be using the system to monitor the energy consumption of their own respective branch.

b) **Sensor**

To store information about individual sensors and their type; energy, door and temperature sensors. The sensor status tells whether the sensor is active or offline or under maintenance. The sensor position shows where the sensor is placed within the shop premise (in refrigerator, front door, kitchen's light, etc)

a. **Energy Sensor**

Track energy consumptions of electrical equipment (Ex: refrigerator, air-conditioners, lighting) at time intervals. In case of some equipment showed different energy consumption patterns, it might be indicating there is a fault in the equipment. This will notify manager to send technician to check and repair the equipment.

b. **Door Sensor**

Measure door/window movements, determining whether they are open or closed (Ex: freezer). Freeze that left open used more energy to maintain its cold temperature, hence use more energy. This will alert the manager to make sure every equipment that requires closing or sealing to be properly closed.

c. **Temperature Sensor**

Measure temperature at time intervals to make sure the temperature at the premise at optimum temperature to maintain efficiency.

c) **Equipment**

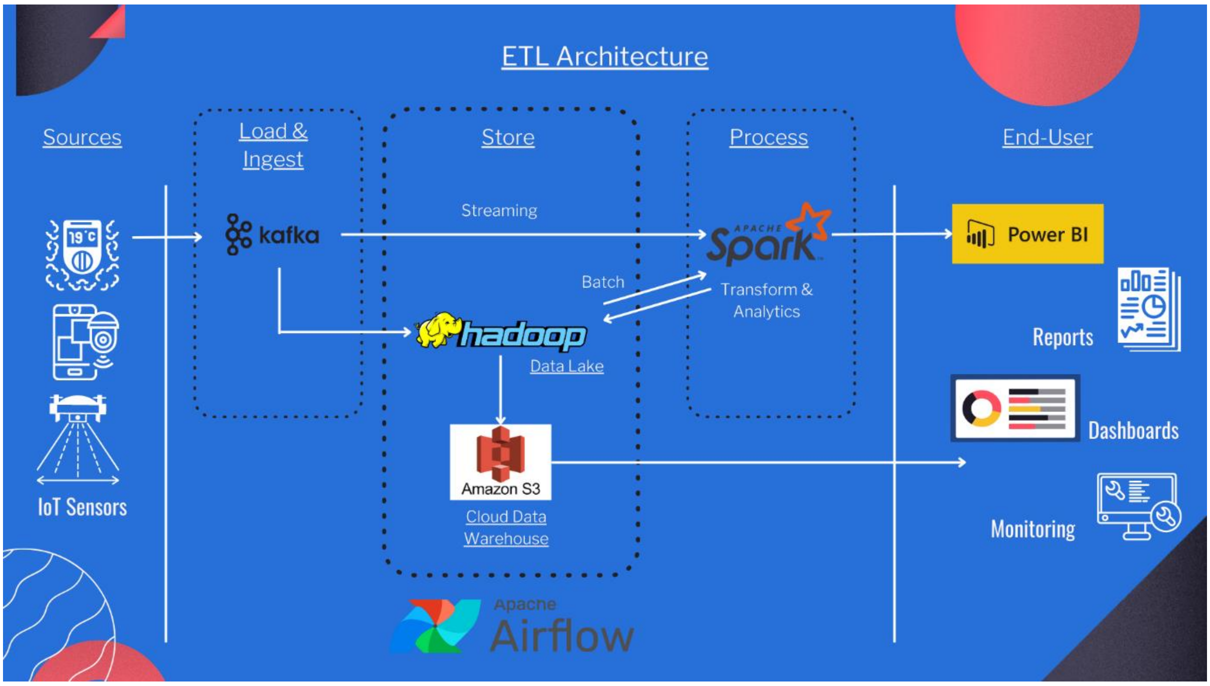
To store information about electrical equipment in the store such as type (cooling, lighting), model (specific model/brand from manufacturer), status (working, under maintenance), date installed to know the age of the equipment, and date of maintenance service. This information is important to know the condition of each equipment in case there is anomalies in sensor reading for easy check-ups.

a. **Manufacturer**

   The manufacturer or supplier of the equipment used. Easy to refer when needed to send for repair.

4.0 ETL Architecture

4.1 Architecture Diagram



4.2 Design Details and Process

a) Ingest

We are using Kafka to ingest and stream data created by IoT devices and electricity equipment. Kafka will store preserving the sequence of events received. Kafka being an open-source, contains Distributed Data Streaming tools that provides real-time event and also batch processing driven applications. This is needed as energy consumption monitoring is conducted to show in real time to be effective for the retail stores. The data is taken from the IoT sensors directly to Kafka and processed in real-time fashion and also loaded into the data lake and data warehouse respectively in an incremental way.

b) Store

The data flowed from Kafka will be stored first in a data lake based on Hadoop. Hadoop runs on a cluster of commodity servers and can scale up to hundreds and thousands of nodes. In that context, the amount of data being handled can be large, and a lot of data sources can be process at the same time which is expected from numerous sensors installed in every store branch for a certain retail business. Other than that, the data is most usually stored in a Hadoop Distributed File System (HDFS). This system allows for parallel processing of data.

From the data lake, the data is then stored on a cloud-based data warehouse, Amazon Simple Storage Service (Amazon S3) which is an object storage service that offers security, scalability, data availability, and performance. Amazon S3 can be used to store and retrieve any amount of data at any time and supply it for monitoring and dashboard purposes. It can provide cloud storage needs at scale for small and large retail.

c) Process

Apache Spark is used for the data processing and analytics. It is an open-source data processing framework for implementing big data analytics on distributed cluster of computers. Other than that, it is a unified tools for working with Big Data or data with massive amounts. Spark can support a variety of data analytics tasks and jobs, ranging from simple data loading and queries to machine learning and streaming computation, over the same computing engine and with a consistent set of APIs. It is because of multiple functional libraries such as machine learning (MLlib), stream processing (Spark Streaming and the newer Structured Streaming), libraries for SQL and structured data (Spark SQL), and graph analytics (GraphX).

d) Data Flow

To monitor the data flow, scheduling and pipeline management, Apache Airflow is utilised as a way to keep everything in order and neat. Airflow is an open-source platform to flexibly author, schedule, and monitor workflows. These workflows will help data engineers to move data, filter datasets, design data policies, manipulation, monitoring and even call microservices to trigger database management tasks. The benefits of using Airflow are that it allows us to schedule and monitor workflows, not just author them. Moreover, workflows assist us on different actions in a process contribute to a valuable business outcome.

e) End-User

At the end of the pipeline, the processed data is pulled to create meaningful reports, dashboards and also monitoring every sensor or vital parameter for energy tracking. One of the tools can be used is Power BI, a collection of business intelligence (BI), reporting, and data visualization products and services for individuals and teams. It stands out with streamlined publication and distribution capabilities, as well as integration with other Microsoft products and services. Power BI also can connect to Apache ecosystem and Amazon S3 to ease data flow from storage to end-user access.

5.0 References

Apache airflow: Overview, use cases, and benefits. (n.d.). Retrieved from https://www.contino.io/insights/apache-airflow

Chang, C., Jiang, F., Yang, C., & Chou, S. (2019). On construction of a big data warehouse accessing platform for campus power usages. *Journal of Parallel and Distributed Computing*, *133*, 40-50. doi:10.1016/j.jpdc.2019.05.011

FabragaMS. (n.d.). Analytics end-to-end with Azure synapse. Retrieved from https://learn.microsoft.com/en-us/azure/architecture/example-scenario/dataplate2e/data-platform-end-to-end?tabs=portal

Introduction to Apache spark tutorial. (n.d.). Retrieved from https://www.projectpro.io/apache-spark-tutorial/tutorial-introduction-to-apache-spark

Kumar, S., Tiwari, P., & Zymbler, M. (2019). Internet of things is a revolutionary approach for future technology enhancement: A review. *Journal of Big Data*, *6*(1). doi:10.1186/s40537-019-0268-2

Kunnathuvalappil Hariharan, N. (2021). Trends in data warehousing techniques. doi:10.31219/osf.io/6cyq4

Marinakis, V., & Doukas, H. (2018). An advanced IoT-based system for intelligent energy management in buildings. *Sensors*, *18*(2), 610. doi:10.3390/s18020610

Modern real-time ETL with Kafka example. (n.d.). Retrieved from https://etl-tools.info/en/examples/kafka-real-time-etl.htm

Pointer, I. (n.d.). What is Apache spark? The big data platform that crushed Hadoop. Retrieved from https://www.infoworld.com/article/3236869/what-is-apache-spark-the-big-data-platform-that-crushed-hadoop.html

Ramírez-Faz, J., Fernández-Ahumada, L. M., Fernández-Ahumada, E., & López-Luque, R. (2020). Monitoring of temperature in retail refrigerated cabinets applying IoT over open-source hardware and software. *Sensors*, *20*(3), 846. doi:10.3390/s20030846

Vaz, D. (2021, July 6). Data lake & Hadoop : How can they power your analytics? Retrieved from https://www.cuelogic.com/blog/data-lake-and-hadoop

Waehner, K. (2022, July 21). Data warehouse and data lake modernization: From legacy on-premise to cloud-native infrastructure. Retrieved from https://www.kai-waehner.de/blog/2022/07/15/data-warehouse-data-lake-modernization-from-legacy-on-premise-to-cloud-native-saas-with-data-streaming/

Waehner, K. (2022, July 21). Data streaming for data ingestion into the data warehouse and data lake. Retrieved from https://www.kai-waehner.de/blog/2022/07/05/data-streaming-for-data-ingestion-into-data-warehouse-and-data-lake/

Wei, M., Hong, S. H., & Alam, M. (2016). An IoT-based energy-management platform for industrial facilities. *Applied Energy*, *164*, 607-619. doi:10.1016/j.apenergy.2015.11.107