EXECUTIVE SUMMARY

INTRODUCTION

This executive summary seeks to evaluate and synthesise the proposal for the completed design and build of the logical database tailored to the client's needs. The proposal covered an explanation of the current client's situation and proposed a solution to the client's issue. In this document, we will elaborate on what was discussed in the initial proposal in addition to focusing on the strengths and weaknesses of the proposed solution. Further, we will address the client's legal and compliance requirements when it comes to the processing, handling, and storing of the data, paying particular importance to how to handle PII (Personally Identifiable Information), which is considered sensitive.

CLIENT BACKGROUND

The client, a successful barbershop in the UK, needs an efficient way to measure its turnover to manage their budget, taxes, and human resources. Employee performance must also be evaluated regularly to ensure the business's success. Currently, the client relies on a time-consuming manual process that involves arranging, processing, and storing data from four separate sources in spreadsheets to create daily reports. This process has the potential to be more efficient due to the disconnected data sources and the manual nature of the process.

The client has two main types of data flow:

- One is mainly static and has slowly changing dimensions. This first type relates to information about employees and customers.
- The second type is strictly transactional.

Looking further at the client's data and after having assessed the past performance of the client, we learned that there are around 200 customers being served each month. There are eight hairdressers working full-time and no data specialists employed. Thus, this provides a rough idea of the volume of data; this information has been taken into consideration for the proposal.

UNDERLYING CONCEPT

To address the client's issues, implementing a data warehouse solution is apt. This is achieved in this case by using an e-commerce web application that sends the data through an API to Google Cloud Platform (GCP) and into the data warehouse (DWH). An open-source data transformation tool (DBT) is then used to automate the ETL (extract, transform, load) process, transforming the raw data into clean data. This is demonstrated in figure 1 below.

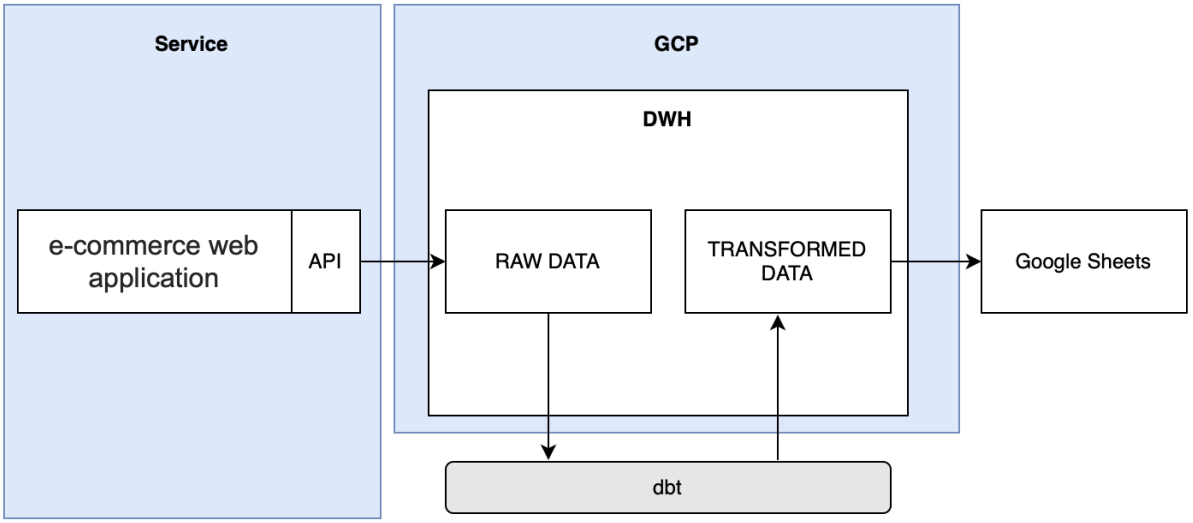


Figure 1: A diagram of the proposed solution, showing data ingestion from the web application into the DWH hosted in GCP.

A user interface will also be created to give the client a simple and easy way to access and analyse the data. This solution stores the data in one place and automatically transforms and structures the data. Furthermore, automated reports can be generated via Google Sheets from this data. Reports in Google Sheets are cloud-based and will automatically update for all of the accessing users. Whenever a report is accessed, all users will see the same, up-to-date information.

Utilising Google Sheets in place of a dedicated BI tool was a conscious decision aimed at decreasing the overall cost of implementation and maintenance; however, it has its drawbacks. The access rights and controls on Google Sheets can be easily changed even by a non-technical user. There is also a relatively high risk that the owner of the sheet will allow other users to make updates that may disrupt the structure of the report and, in turn, break the whole pipeline. This would then need the intervention of a data professional in order to be fixed. However, due to its simplicity and the automated process, Google Sheets is an adequate solution for the client's reporting purposes.

Furthermore, using GCP provides one of the highest levels of security available among cloud providers. It also allows for implementing data compliance checks in the pipeline.

To summarise, the above has been chosen as a solution because the database design and technologies used for implementation are cheap or free in usage and maintenance. Relatively low data volumes processed by the barber shop make it possible to limit the costs related to the implementation of the database.

CONCEPT'S STRENGTHS AND WEAKNESSES

The underlying concept of the proposed database model is a star schema; it consists of a fact table and is directly connected to dimensional tables (ThoughtSpot, 2022).

Although the star schema has its drawbacks, we have opted to use it in our solution as its benefits outweigh its disadvantages for our application. For example, one such disadvantage is that the star schema has redundant data, which increases the storage on disk. This should not be an issue currently, but if the business scales, it may become a cost worth considering. This is caused by the denormalization of the data. Another disadvantage is that there is also a higher potential for errors in data as well as limited flexibility for non-dimensional data. Keboola (2022) states that, whilst the star schema provides benefits, it also has the limitations mentioned above and can be hard to maintain.

Additionally, although the pipeline setup is relatively easy for a data professional, it is not for a non-data-savvy person. Therefore, the maintenance of the pipeline (in case errors occur) may require the client to consult with or hire a professional each time the pipeline breaks, which generates additional costs for the business. This complexity (for non-technical users) also makes it hard and potentially costly to scale and extend existing reporting.

However, the advantages are that it simplifies data analysis, improves query performance, and enables reporting and scalability (Kimball & Ross, 2013: 16-18). The simplification of data analysis played a significant role in our decision to choose the star schema.

There are other advantages to the concept. For example, the data pipeline implements five layers in the processing architecture. This choice allows for improved readability of the code and implementation. It also allows for better control over the quality of the data. Each of the layers is responsible for a different, single operation (e.g., cleaning, transforming, and reporting), which is in line with software engineering principles (Singh, 2022).

This pipeline setup allows for full automation of the final reports, which are integrated with the data warehouse. This allows to save time on manually updating the reporting as well as allow for improved data governance. By the time the data reaches the final report, it has already gone through 4 layers on which quality testing and transformations are applied, ensuring that the final report contains valid and quality information.

DEEP DIVE INTO THE TABLES' STRUCTURE

The following section describes and explains the makeup of the tables and references to the appendix throughout:

The fact table is named fct_transactions (Appendix F) with all transactions; it stores all business process events and contains numerical data (Kimball & Ross, 2013: 41).

The dimensional tables hold information about explanatory attributes, and the dimensional tables' primary keys point to the fact tables' foreign keys (ThoughtSpot, 2022). The dimensional tables used in the database model are dim_service

(Appendix A), dim_customer (Appendix B), dim_employee (Appendix D), and dim_payment (Appendix E).

All services and products provided by our client are listed in the dim_service table, and an e-commerce web application is used to manage the payments for the provided products and services. The payment information is stored in the table dim_payment (Appendix E).

Additionally, customer-related data are stored in the table dim_customer (Appendix B)**,** and the country_code_glossary (Appendix C) contains all countries and their respective country code. These tables are linked together with the primary key country_code in the country_code_glossary (Appendix C) and the foreign key phone_country_code in the table country_code_glossary (Appendix C).

Our client needs the ability to evaluate its employees, whose data are stored in the dim_employee table (Appendix D).

The dimensional tables linked to the fact table, as discussed above, result in a star schema (Appendix G). In order to provide automated reports about employee performance and sales_report, the tables employee_monthly_performance_report (Appendix H) and sales_report (Appendix I) are needed. These two tables are on top of the star schema and reflect the monthly performance of their employees and sales in an aggregated form (Appendix J).

LEGAL AND COMPLIANCE REQUIREMENT

To ensure legal compliance pertaining to data processing, the security of the data is paramount, as this has been proven to improve regulatory compliance (Kwon & Johnson, 2011). This does not mean that security equates to compliance; however, with strong security, compliance is easier to achieve (Comptia.org, 2016).

Some of the primary methods of security include encryption, backup and recovery, and access control (Rodrigues, 2023). Since the data is being stored and processed primarily using GCP, encryption is handled by Google (*Default encryption at rest | documentation | google cloud*). Encryption is vital since personal, sensitive data for employees and customers, such as email addresses and phone numbers, is being collected and processed in the database. Encryption ensures greater security for data. Backup and recovery are also available via a service in GCP. Additionally, access will be limited to the employees that require access in order to do their job (Byun & Li, 2006). This will also help ensure compliance, as only authorised users will be able to access the data.

By practising the above security methods, compliance with regulations such as GDPR becomes easier (McQuillian, 2022). For example, Art. 5 (f) of GDPR states that personal data shall be processed securely. Art. 6 details the lawfulness of processing the data; any customer or employee must provide consent for the processing of their data. Due to the nature of the business, agreements must be in place within the employee contract as well as either consent provided by any customer for storing their personal data or the data may fall into Art. 6 (b) i.e., if it is

necessary for the performance of the contract (Regulation (EU) 2016/679 (General Data Protection Regulation), 2016).

Additionally, the rights of the data subjects must be considered, which are covered in Art. 12–23. In this case, the database must allow for the rectification and erasure of personal data for both employees and customers. Since this is stored in only a single table, it is relatively simple to remove and amend data upon request.

CONCLUSION

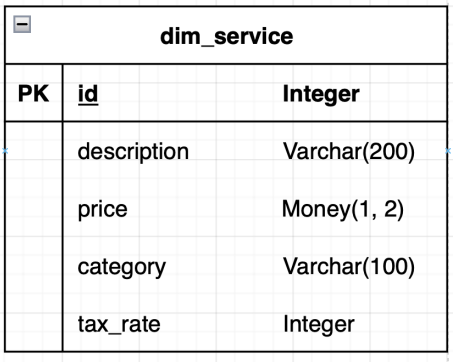
The proposed solution leverages GCP in conjunction with processing data through an open-source software, dbt, to transform the data before providing reporting which is available in Google Sheets.

After weighing up the advantages and disadvantages of the star schema data model, it has been chosen as it ensures that the data is easy to understand and analyse; this model is suitable for the current size of the business as well as the volume of data.

Additionally, by using GCP, data security and compliance is made easier and the necessary steps have been taken to ensure that any PII is handled correctly. Only the management has access to the database and the reports currently, but if the business grows then there is a likelihood that the need would arise for more organised access controls. Further, the current design may reflect the current business requirements, but the business growing may necessitate more fields to be included and thus will be missing from the database and performance issues could arise if the database grows. Thus, the chosen solution may need to be adapted and upgraded in this event. This has been accounted for in relation to the scalability of both the GCP and the star schema.

APPENDICES

APPENDIX A: Table dim_service



dim_service		
PK	id	Integer
	description	Varchar(200)
	price	Money(1, 2)
	category	Varchar(100)
	tax_rate	Integer

Columns:

- **id**: A unique identifier for all services and products.
- **description**: This field contains a product or service description.
- **price**: The price for the product or service.
- **category**: Each product or service is assigned to a category.
- **tax_rate**: Each product or service is allocated to the appropriate tax_rate.

APPENDIX B: Table dim_customer

dim_customer		
PK	id	Integer
	first name	Varchar(50)
	last name	Varchar(50)
	email_address	Varchar(100)
	phone_country_code	Varchar(20)
	phone_number	Varchar(20)

Columns:

- **id**: A unique identifier for all customers
- **first name**: Contains our customer's first name.
- **last name**: Contains our customer's last name.
- **email_address**: The customer's email address is stored to inform him about appointment changes or promotions.
- **phone_country_code**: The phone_country code is a foreign key that links the country_code_glossary table to each other.
- **phone_number**: To contact our customers about appointment changes on short notice.

APPENDIX C: Table country_code_glossary

country_code_glossary		
PK	country_code	Varchar(3)
	country	Varchar(100)

Columns:

- **country_code**: A unique identifier for all countries
- **country**: This field contains the country name for each country code.

APPENDIX D: Table dim_employee

dim_employee		
PK	id	Integer
	first name	Varchar(50)
	last name	Varchar(50)
	date_of_birth	Date
	date_joined	Date
	date_left	Integer
	email_address	Varchar(100)
	phone	Varchar(20)
	annual_salary	Decimal(1, 2)

Columns:

- **id**: A unique identifier for all employees.
- **first name**: This field contains the employee's first name.

- **last name**: This field contains the employee's last name.
- **date_of_birth**: It contains the employee's birthday.
- **date_joined**: The date an employee joins the company can be used to evaluate the level of work experience of an employee in the company.
- **email_address and phone**: This is needed to contact the employee.
- **annual_salary**: This field shows the employee's annual salary

APPENDIX E: Table dim_payment

dim_payment		
PK	**id** **Integer**	
FK	product_id	Integer
	amount_paid	Integer
	original_amount	Varchar(100)
	is_discounted	Boolean
	payment_type	Varchar(25)
	discount_rate	Integer
	customer_satisfaction	Integer
	tax_rate	Integer
	net_amount_after_tax	Integer

Columns:

- **id**: A unique identifier for all payments
- **amount_paid**: The amount of money that has been paid for the service or product.
- **Original_amount**: The original amount of cost for the product or service.
- **Is_discounted**: This flag indicates if the service or product was discounted.
- **Payment_type**: The type of payment that has been used for paying for the product or service.
- **Discount_rate**: It shows the rate of a discount.
- **Customer_satisfaction**: This reflects how satisfied a customer was with the service.
- **Tax_rate**: This field shows the tax rate for the service or product.
- **Net_amount_after_tax**: Shows the amount without taxes.

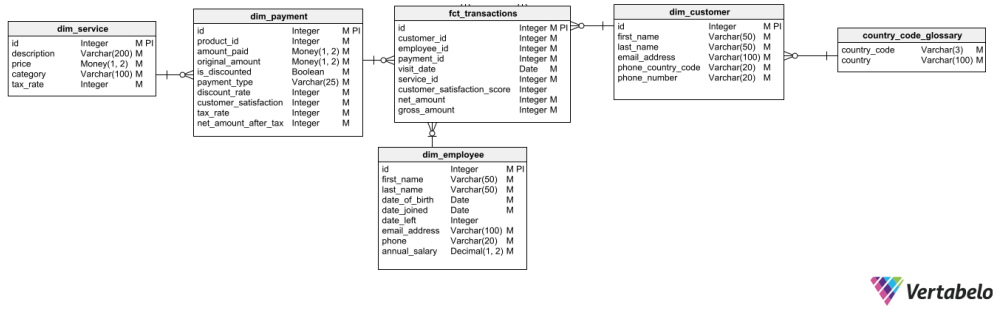
APPENDIX F: Table fct_transaction

fac_transaction		
PK	**id**	**Integer**
FK	customer_id	Integer
FK	employee_id	Integer
FK	payment_id	Integer
	visit_date	Date
FK	service_id	Integer
	customer_satisfaction_score	Integer
	net_amount	Integer
	gross_amount	Integer

Columns:

- **id**: A unique identifier for all business events.
- **customer_id**: This foreign key links to the dim_customer table.
- **payment_id**: This foreign key links to the dim_payment table.
- **visit_date**: This date is the date of the transaction.
- **Service_id**: A foreign key links to the dim_service table.
- **customer_satisfaction_score**: It shows the customers' satisfaction level.
- **net_amount**: It shows the money earned with this transaction with tax deducted.
- **gross_amount**: It shows the money earned with taxes.

APPENDIX G: Star Schema



APPENDIX H: Table employee_monthly_performance_report

employee_monthly_performance_report		
PK	**primary_key**	**Integer**
	date_generated	Date
FK	employee_id	Integer
	total_sales	Decimal(1,2)
	avg_satisfaction_score	Decimal(1,2)
	total_clients_served	Decimal(1,2)

Columns:

- **primary key**: A unique identifier for all monthly performance reports.
- **data_generated**: The date is automatically generated every time a report is called.
- **employee_id**: This foreign key links to the dim_employee table.

- **total_sales**: This field shows the total sales for every employee.
- **avg_satisfaction_score**: This score shows how satisfied the employee's customers were with the provided products and services.
- **total_clients_served**: The total_clients_served information shows how many customers have been served by an employee.

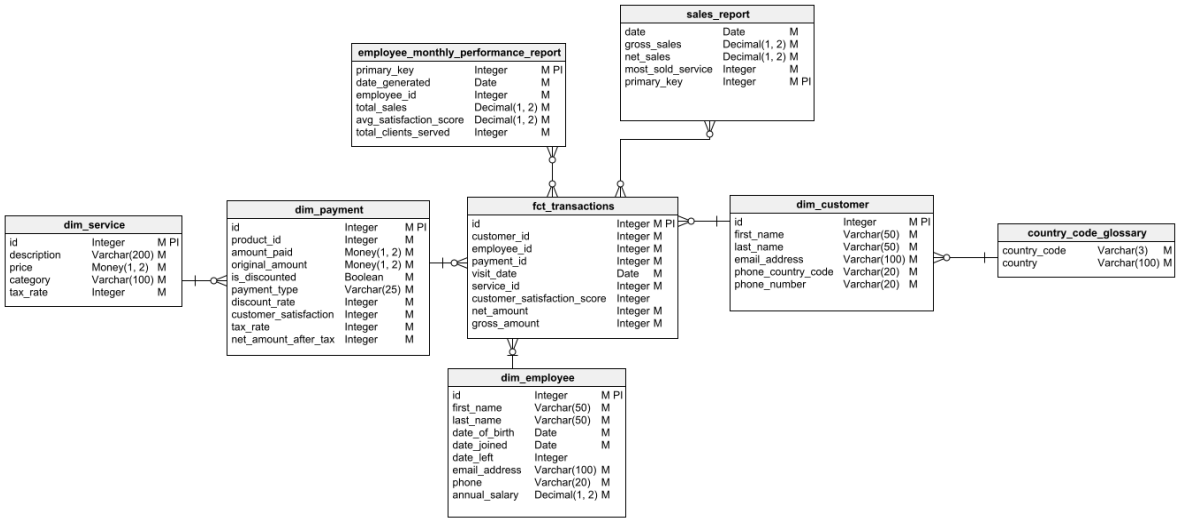
APPENDIX I: Table sales_report

	sales_report	
PK	**primary_key**	**Integer**
	date	Date
	gross_sales	Decimal(1,2)
	net_sales	Decimal(1,2)
	most_sold_services	Integer

Columns:

- **primary_key**: A unique identifier for all sales report events.
- **date**: It shows the date from the report.
- **gross_sales**: It shows the overall sales including taxes.
- **net_sales**: It shows the overall sales without taxes.
- **most_sold_services**: This field shows the most sold products and services.

APPENDIX J: Final Star Schema



The full database schema can be viewed here.

REFERENCES

Byun, J.-W. and Li, N. (2006) “Purpose based access control for privacy protection in Relational Database Systems,” *The VLDB Journal*, 17(4), pp. 603–619. Available at: https://doi.org/10.1007/s00778-006-0023-0.

Default encryption at rest | documentation | google cloud (no date) *Google*. Google. Available at: https://cloud.google.com/docs/security/encryption/default-encryption (Accessed: April 1, 2023).

Kimball, R. and Ross, M. (2013) *The Data Warehouse toolkit the Definitive Guide to Dimensional Modeling*. 3rd edn. Indianapolis, IN: Wiley.

Kwon, J. and Johnson, E. (2011) “The Impact of Security Practices on Regulatory Compliance and Security Performance,” *International Conference on Information Systems* [Preprint].

McQuillan, R. (2022) *What is Data Access Control: In-depth guide*, *Budibase*. Available at: https://budibase.com/blog/app-building/data-access-control/ (Accessed: April 1, 2023).

Quick start guide to security compliance: Cybersecurity: Comptia (no date) *Compitia*. Available at: https://connect.comptia.org/content/guides/quick-start-guide-to-security-compliance (Accessed: April 1, 2023).

Rodrigues, J. (2023) *Top 5 methods of protecting data*, *TitanFile*. Available at: https://www.titanfile.com/blog/5-methods-of-protecting-data (Accessed: April 1, 2023).

Singh, R. (2022) *8 software engineering principles to live by*, *CalliCoder*. CalliCoder. Available at: https://www.callicoder.com/software-development-principles/ (Accessed: April 1, 2023).

Star schema vs snowflake schema and the 7 critical differences (2022) *Star Schema vs Snowflake Schema and the 7 Critical Differences*. Available at: https://www.keboola.com/blog/star-schema-vs-snowflake-schema (Accessed: April 1, 2023).

Star schema vs snowflake schema: 6 key differences (2022) *ThoughtSpot*. ThoughtSpot. Available at:

https://www.thoughtspot.com/data-trends/data-modeling/star-schema-vs-snowflake-schema (Accessed: March 24, 2023).

Zola, A. (2021) *What is a schema?*, *Data Management*. TechTarget. Available at: https://www.techtarget.com/searchdatamanagement/definition/schema (Accessed: March 24, 2023).

'Regulation (EU) 2016/679 of the European Parliament and the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and the on the free movement of such data, and repealing Directive 95/46/EC' (2016), Official Journal of the European Union L 119, pp. 1-88.