PROJECT REPORT

Project Team: Piotr Sieminski, Rhiannon Johns, Daniel Bösch

INTRODUCTION

Our client, a successful barbershop located in the UK, requires measurement of the shop's turnover to plan the budget, manage taxes and human resources. Furthermore, employee performance needs to be evaluated regularly since it is crucial for the clients success.

Used Abbreviations:

1. GCP - Google Cloud Platform
2. DWH - Data Warehouse
3. UI - User interface
4. DBT - Data transformation tool (open source)
5. ETL - Extract, transform and load

CURRENT SITUATION

Currently, our client performs the time-consuming task of manually arranging, processing, and storing data in spreadsheets to create reports daily. This data comes from four separate sources: cash registers, customers, employees, and service. The system's weakness lies in the disconnected data sources and the manual process.

DATA PROCESS ARCHITECTURE

To tackle the lack of integration and manual process, the most beneficial solution is a fully automated and optimised cloud platform. This will enable the client to focus on other core areas whilst still generating reports and insights vital to the running of the business.

Our solution starts with setting up an e-commerce web application service for handling transactions, customer data, and services; it will be the single point of truth for raw data.

The next step is to set up a GCP account with BigQuery (free tier) because this will suit the small volume of data the client will be storing and processing. It is also equipped with top-tier security measures, managed directly by Google.

After the GCP account is created, data loads through an API from the e-commerce web application to the GCP.

Whilst the technical preconditions are now established, our data would still be considered raw data. Thus, a data pipeline must be implemented for transforming and cleaning the data.

The software service, DBT, will be used for the data engineering workflow - creating the ETL for these tasks. Furthermore, this tool has version control, documentation, and a testing framework which we will utilise to develop a reliable data pipeline resulting in a DWH.

Since data quality is crucial, we will implement a data testing strategy to detect inconsistencies such as outliers and formatting errors. Additionally, we will create user profiles for all used services to guarantee safe access management.

After this, the data is considered clean; automated queries will be created to gather the data. The data will be available via Google Sheets and guarantee our client a flexible way to analyse the data.

Figure 1 represents data processing similar to the architecture suggested above:

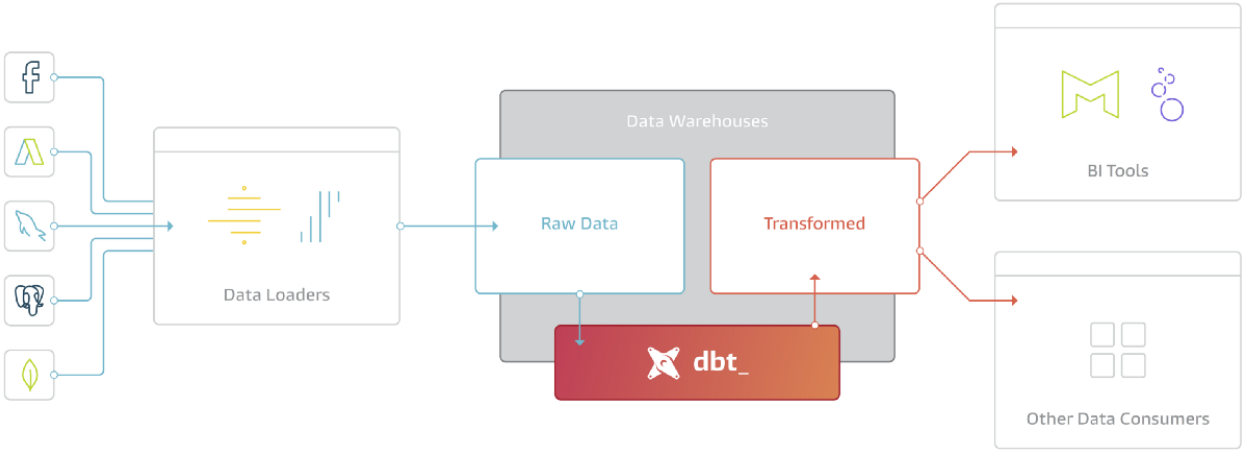


Figure 1: *Data process architecture. (Handy & Lantz, 2017)*

PIPELINE OVERVIEW

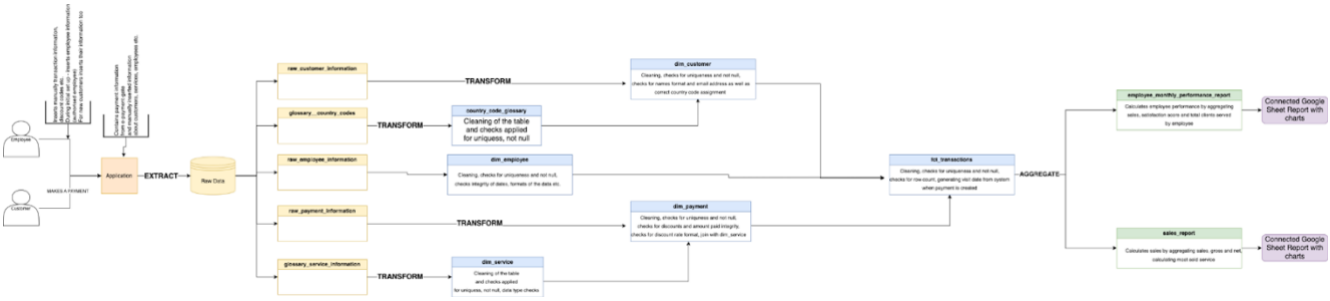


Figure 2: *Pipeline Overview*

Full pipeline overview

PIPELINE DESCRIPTION

1. Human input into the application
    a. Authorised employee inputs customer information into the UI of the App
        i. Data automatically exported into raw data storage facilities
    b. Customer completes payment
        i. Payment metadata from the payment gate gets automatically exported into raw data storage facilities
2. GCP loads raw data tables based on the extract from the web application
3. Transformation step of the raw tables:
    a. Naming convention is applied to all table columns
    b. Columns checked for uniqueness and null values
    c. Data type checks are applied to ensure consistency across DWH
    d. Unnecessary information is truncated
    e. Source freshness checks are applied to ensure new data gets loaded according to set schedule
    f. Allowed values test is applied to ensure the correctness of the information
4. Transformation of aggregated tables:
    a. Checking for row count (if no rows were omitted or duplicated)
    b. Data quality check from 3 are applied
5. Google Sheets Reports:
    a. Connected Google Sheets solution in GCP automatically populates Google Sheets with new data
    b. Manually created visualisations in other tabs of said sheets automatically update as the new data flows in

Additional notes:

1. All tests are automatically handled and orchestrated by DBT

LOGICAL DATABASE

The entities are customers, employees, services, and payments. The database design based on these entities have the dimensions:

- dim_services
- dim_payment
- dim_customer
- dim_employee

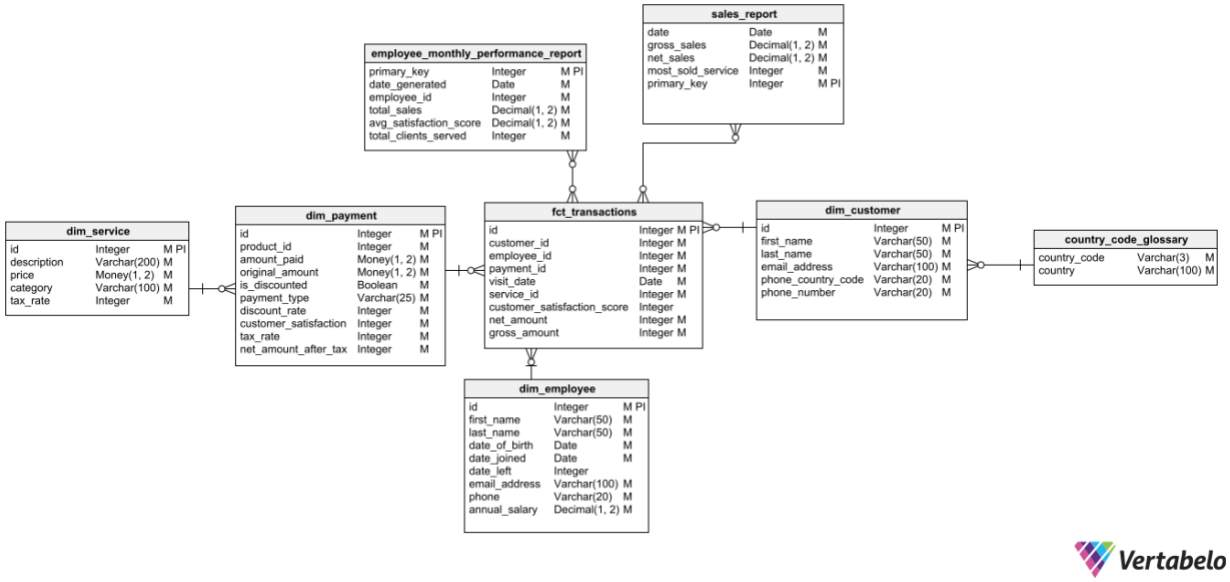
Every dimension table contains several dimension attributes (Kimball & Ross, 2013: 13). The chosen dimension attributes explain the type of service, payment, customer information, and the employee who provided the service.

The sales are stored in the fact table, 'fct_transaction', which contains the data that needs to be analysed (TechTarget, 2012). This table has foreign keys that are linked to the keys in the dimension tables (TechTarget, 2012). This database architecture is known as a star schema (Kimball & Ross, 2013: 10-13).

The data types and formats can be seen in the visual representation of the database design and can be categorised into:

- Integer: this is used for keys, foreign keys, counts, and categories; it is used when a whole number can represent the data information.
- Varchar(length): this is used for texts such as descriptions and country names.

- Money(i, j): this is used for storing currency amounts with precision i and scale j (HCL, 2022).
- Boolean: this is used as a true or false flag for the column is_discounted.
- Decimal (i, j): this is used for annual_salary, total sales, avg_satisfaction, gross_sales and net_sales; with precision i and scale j.
- Date: is used for date items that contain dates.



Full architecture overview

The following reports are built on this star schema:

- employee_monthly_performance_report
- sales_report

PROPOSAL

Pos	Task	Effort (days)
1	Set up e-commerce web application	5
2	Set up a GCP with BigQuery	1
3	Set up automated data loads from e-commerce to GCP through API	1
4	Set up DBT	1
5	Creating a logical database design	1
6	Creating a data pipeline for transforming and cleaning data (ETL) with DBT	3
7	Creating a test strategy to ensure high data quality and detecting data inconsistencies and outliers	1
8	Setting up access management policies	1
9	Creating automated reports, displayed via Google sheets	2
	TOTAL	**16**

We are happy to offer the above solution excluding VAT for **£9,600** (equivalent to 16 days). This offer includes pos. 1 to 9. This offer does not include the fees for the used services; the used services require a payment plan which is the responsibility of the client.

Additionally, we offer an optional package with three days support and maintenance excluding VAT for **£1,800** for access to support and maintenance as needed.

REFERENCES

Handy, T. and Lantz, J. (2017) *What, exactly, is DBT?*, *Transform data in your warehouse*. Available at: https://www.getdbt.com/blog/what-exactly-is-dbt/ (Accessed: February 18, 2023).

Kimball, R. & Ross, M. (2013) *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. 3rd ed. Indiana: John Wiley & Sons Inc.

HCL, T. L. (2022) MONEY(p,s) data type. Available from: https://www.ibm.com/docs/en/informix-servers/14.10?topic=types-moneyps-data-type [Accessed: February 19, 2023].

TechTarget, C. (2012) DEFINITION - fact table. Available from: https://www.techtarget.com/searchdatamanagement/definition/fact-table [Accessed: February 19, 2023].