Report Part 1 — Group 020

188.429 Business Intelligence (VU 4.0) 2025W

Assignment 1: Dimensional Modelling and ETL Implementation

Student A: Muhammad Sajid Bashir (52400204)
Student B: Eman Shahin (12432813)

5.1 Synthetic Tables (Table_X and Table_Y)

Purpose:
Two additional OLTP tables extend the snapshot with campaign-level business context. They link environmental programs to sensor devices, enabling campaign-based analysis of readings and maintenance.

Table_X – `tb_campaign`
Contains 6 rows describing campaigns (`id` , `campaign_name` , `objective` , `sponsor` , `start_date` , `end_date` , `etl_load_timestamp`).
Stores high-level information such as program purpose, sponsor, and duration.

Table_Y – `tb_campaign_device`
Contains 12 rows linking campaigns to devices (`campaign_id` , `device_id` , `assigned_from` , `assigned_to` , `priority` , `etl_load_timestamp`).
Creates a many-to-many bridge between `tb_campaign` and `tb_sensordevice` .

Integration into OLAP Schema:
`tb_campaign` was transformed into `dim_campaign` .
`tb_campaign_device` connects device and campaign data during ETL. Both fact tables include `sk_campaign` , allowing analysis by campaign, sponsor, and timeframe.

5.2 Business / Analytic Questions

Questions below rely on the star schema (`ft_reading` , `ft_service` , `dim_timeday` , `dim_device_geo` , `dim_parameter` , `dim_campaign`) and cannot be answered from the normalized OLTP snapshot.

Student A (Readings & Campaigns)

1. How did average parameter values (PM10, NO_2) change by month or week per campaign in 2023?

2. Which cities or countries recorded the highest exceedances during campaign periods?
3. Which devices generated the largest data volume per campaign?
4. Do publicly sponsored campaigns achieve higher data quality scores?

Student B (Services & Technicians)

1. Which devices required the most maintenance during campaigns?
2. Do technicians with the latest SCD2 role achieve higher service quality?
3. Which campaigns caused the longest maintenance times per city?
4. Is low reading quality associated with more frequent maintenance?

5.3 Star Schema Diagram

The AirQ Data Mart (`dwh_020`) follows a **Star Schema** with two fact tables:

- `ft_reading` — environmental sensor readings
- `ft_service` — maintenance and technician events

Shared conformed dimension: `dim_device_geo` (Country → City → Device).
Other key features:

- One SCD Type 2 dimension: `dim_technician_role_scd2`
- Hierarchical dimensions: `dim_timeday` , `dim_device_geo`
- Synthetic dimension: `dim_campaign`
- All tables include `etl_load_timestamp` .

Figure: *AirQ_ERD_dwh_020.png — Star schema with facts and connected dimensions.*

5.4 Fact Tables

Fact Table 1 — `ft_reading` (Student A)

Captures all air-quality readings per device per day for campaign-level analysis.

Grain: One row per reading event per device per day.
Measures:

- `recorded_value` (SUM/AVG) — pollutant concentration
- `data_volume_mb` (SUM) — data collected in MB
- `data_quality_score` (AVG) — average data quality (1–5)
- `exceedance_flag` (COUNT) — threshold exceedances
  Dimensions: `dim_timeday` , `dim_device_geo` , `dim_parameter` , `dim_readingmode` , `dim_campaign` .

Fact Table 2 — `ft_service` (Student B)

Represents maintenance and technician activity per device per day.

Grain: One row per service event per device per day.
Measures:

- `service_cost_eur` (SUM) — total service cost
- `duration_minutes` (SUM) — total maintenance duration
- `service_quality_score` (AVG) — service quality (1–5)
- `underqualified_flag` (COUNT) — below-skill assignments
  Dimensions: `dim_timeday`, `dim_device_geo`, `dim_servicetype`, `dim_technician_role_scd2`, `dim_campaign`.

5.5 Dimension Tables

Eight dimensions describe temporal, spatial, operational, and campaign context.
They include one conformed dimension, two hierarchical dimensions, one SCD2, and one synthetic dimension, each with `etl_load_timestamp`.

Dimension	Hierarchy	Key Features
dim_timeday	Year → Month → Day	Daily time hierarchy for all facts
dim_device_geo	Country → City → Device	Conformed, shared by both facts
dim_parameter	Group → Family → Parameter	Describes environmental measures
dim_servicetype	Category → Type → Subtype	Defines service classifications
dim_technician_role_scd2	(historical tracking)	SCD2 for technician roles
dim_readingmode	Flat	Reading configuration lookup
dim_campaign	Campaign → Sponsor → Objective	Synthetic dimension from Table X/Y
dim_alertstatus	Flat	Optional minor lookup dimension

Hierarchical dimensions support drill-down and roll-up analytics, SCD2 preserves technician history, and the synthetic campaign dimension links business initiatives to operational data.

5.6 Snowflake vs. Star

Student A:
The Star schema is more efficient for analytical workloads. It provides faster aggregations and simpler joins across large datasets. Since AirQ data mainly supports time and campaign-based analysis, denormalized dimensions improve usability without meaningful redundancy.

Student B:
The Star schema ensures clarity, performance, and traceability of ETL processes. While a Snowflake schema could slightly reduce duplication, it would complicate queries and reduce speed. For analytical reporting, the Star model is the optimal balance between simplicity and efficiency.

5.7 ETL and Validation Summary

ETL Overview:
ETL scripts in `/etl/` sequentially loaded data from `stg_020` to `dwh_020`.
Steps included creating time and lookup dimensions, applying SCD2 logic, integrating campaigns, and building two fact tables.
All tables include surrogate keys and audit timestamps.

Validation Summary:
Seven SQL-based post-ETL checks verified data integrity and consistency.

#	Check	Result	Interpretation
1	Row count match (staging vs DWH)	OK	All dimensions fully loaded
2	Attribute consistency	OK	No mismatched names
3	Referential integrity	OK	No orphaned facts
4	SCD2 validity	OK	Correct versioning of technician roles
5	Measure range	OK	All measures within valid limits
6	Null/domain checks	OK	No invalid or missing values
7	Campaign linkage	OK	Synthetic data correctly integrated

Conclusion:
The ETL pipeline executed successfully. The `dwh_020` schema is referentially complete, historically accurate, and analytically ready.

5.8 Reflection and Lessons Learned

Student A – Muhammad Sajid Bashir (52400204):
This project strengthened my understanding of dimensional modelling and consistent grain

design. Implementing SQL-based ETL highlighted the value of automation and reproducibility. SCD2 implementation clarified how history tracking enhances analytical depth. Team collaboration emphasized shared ownership and quality control.

Student B – Eman Shahin (12432813):
I learned how SQL-driven ETL pipelines enforce structure and transparency. Implementing `ft_service` improved my grasp of additive measures and performance validation. Running integrity checks showed how data quality directly supports analytics. Effective teamwork ensured a coherent, maintainable solution.

End of Report