DWH migration from Oracle to BigQuery

How to migrate a Legacy DWH with a Cloud Native and Serverless approach

SECTOR

Retail, Luxury Goods

COMPETENCIES

Data Architecture

App Modernization

Cloud Migration

Data Engineering

SERVICE PROVIDER

Google Cloud

THE CHALLENGE

Companies are becoming data driven organisations, and for this reason the Data Warehouse (DWH) is one, if not the most, important part of the IT infrastructure.

Our client, which operates in the luxury goods sector, **started building their own DWH more than 20 years ago**. This incremental evolution process led to an accumulation of hotfixes and exceptions, without a proper infrastructure renewal to adapt to technological innovation and to the new business requirements of the market.

The challenge we faced was to **resolve stability and performance issues** and at the same time to refactor the code and to document parts of the infrastructure that, due to their legacy nature, were unknown to almost anyone. Specifically, there were critical pipelines that impacted the generation of key reports in the event of failure. Running these pipelines in the old infrastructure could take hours, which was unacceptable given the time-sensitive nature of the reports.

OUR SOLUTION

BIP xTech carried out an initial assessment phase, necessary to perimeter the scope of the modernization and to choose the proper technologies. We decided to create the new DWH on **Google Cloud Platform (GCP) using BigQuery and Dataform**.

BigQuery gives us **better performance** (compared to old on-premises hardware), integrates easily with **Business Intelligence tools**, complies with industry **security standards** and allows granular access control, and offers a **pay-as-you-go** pricing model (especially useful in the migration phase, when the amount of data to be processed is not known precisely).

Dataform is the orchestrator executing queries from BigQuery to BigQuery. It's been chosen because, in the legacy system, most of the transformations were already implemented in SQL language. This allowed BIP xTech to perform a comprehensive cleaning of the logic and refactoring without having to rewrite all the pipelines and **leveraging the current know-how** of the company's internal DWH team. The adoption of this tool only required training for the management of GCP console and the usage of version control tools (Git). In addition to maintaining continuity with the past, Dataform allowed us to exploit all the innovations brought over the years by software engineering, being based on the concept of **data transformations as code** (with full versioning support of transformations). It also natively integrates **assertions** (data quality checks, used to ensure data meets expectations) and **unit tests**, in such a way to have a real **single source of truth**. Equally critical is the ability to document the data schema and the transformations by **adding documentation to your Dataform code**. The documentation is automatically added to the Data Catalog within Dataform and can later be exported to other tools. Finally, the **dependency**

tree visualization of all the dependencies that are automatically detected by Dataform, allows the viewer to easily understand the relationship between each entity of the project and their order of execution.

Bringing the infrastructure to the cloud also lets us take advantage of all the native features offered by GCP, which would have required many ad-hoc developments to be implemented in the old on-premise environment. BIP xTech made use of "**table snapshot**" to save an ultra-light state of the table at a certain point in time, allowing instant data rollbacks or easily calculation of the delta between two different moments. We used **Cloud Workflow, allowing the management of the integration** between Dataform and other GCP services, and we monitored the whole infrastructure via **Cloud Alerting**, which **notifies the responsible personnel in real time** if any errors occur.

RESULTS

The adoption of BigQuery made it possible to **drastically reduce pipeline execution times**. The first pipeline analyzed saw a reduction **from two hours to a few minutes**. We also **reduced costs**, leveraging the potential of the "pay-as-you-go" model **by writing optimized queries that load only the strictly necessary portion of data** (partition pruning), and **eliminating infrastructure management costs**. This fully managed architecture is also automatically scalable and adapts to the growing needs of the company.

Thanks to Dataform we produced documentation integrated into the project allowing a substantial **speed up in the on-boarding process** of new team members. Furthermore, the data cleaning processes before the transformations and the checks after the transformations made it possible to **lower the number of unexpected errors**, and to have **descriptive logs** when they occur.

The enormous storage capacity of the cloud allows a pipeline to be re-run as if it were at a precise moment in the past, making it possible to **quickly fix any errors in the data** that are caused by bugs which are introduced in new versions of the code.

Finally, the use of version control tools **improves collaboration between team members** and regulates the **software lifecycle following industry best practices**, making it easier for future and continuous refactoring and modernization interventions, so we do not go back to the previous situation.

BENEFIT

90% reduction in pipeline execution time

Elimination of infrastructure management costs

Full history of changes of all pipelines available

Data Quality Assurance

Documentation integrated into the tool

Contact us

Davide Tresoldi

Senior Cloud Architect @ BIP xTech

davide.tresoldi@bip-group.com

———

Simeone Vilardo

Specialist @BIP xTech

simeone.vilardo@bip-group.com

bip. xTech