

Seamless Transition: Migrating On-Premises Reporting Solutions to The Cloud for Agility and Scalability

Project by Iwan Dyke
UWTSD (University of Wales Trinity St David)
Course: DA6CDI Computing (Data & Information Systems)

Project Objectives

Aim

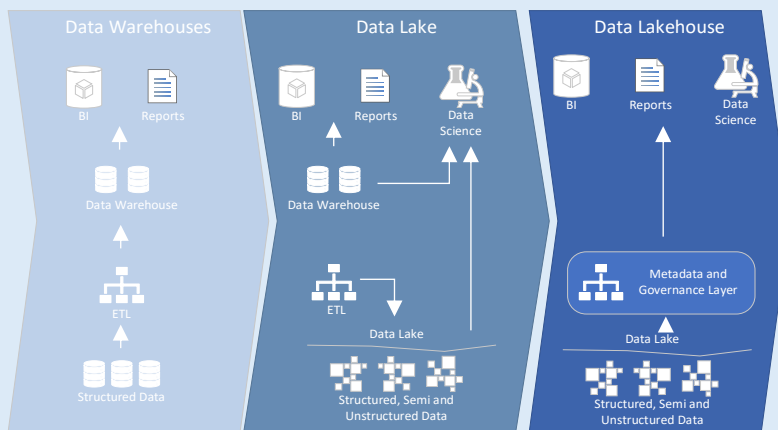
To demonstrate the feasibility and benefits of a Lakehouse Architecture in enhancing data management processes for water network development.

Objectives

- Evaluate data management practices in water network development.
- Design and implement a prototype Lakehouse Architecture using Databricks and the Azure Datalake
- Conduct an analysis of processing times, scalability and cost efficiency between On-premises Data Warehouse technologies and cloud based.

Evolution of Data Management

This infographic illustrates the evolution from traditional Data Warehouses to Data Lakes and finally to Data Lakehouses. Data Warehouses are designed for BI and reporting using structured data, whereas Data Lakes accommodate both structured and unstructured data for diverse analytics. The modern Data Lakehouse integrates the benefits of both, providing a unified platform with robust governance and performance optimization for all data needs.



Research Key Findings

Performance and Scalability:

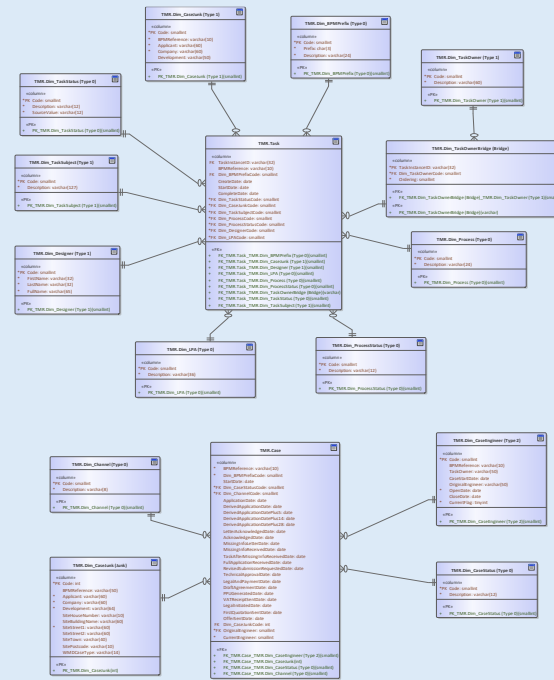
Cloud-based solutions like Databricks offer significant improvements in performance and scalability compared to traditional on-premises systems. Databricks enables faster data processing and can handle larger data volumes efficiently.

Cost Efficiency:

Transition to cloud platforms results in considerable cost savings. The pay-as-you-go model of cloud services eliminates the need for substantial upfront investments and reduces ongoing maintenance costs.

Data Management and Governance:

Features like Unity Catalog ensure centralized access control, auditing, and data lineage. Enhanced data governance and management capabilities are critical benefits of adopting cloud solutions.

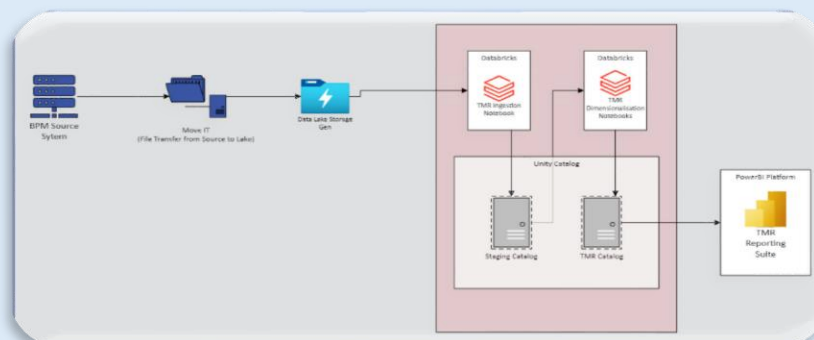


Dimensional Modelling

Dimensional Modelling created by Ralph Kimball aims to improve the efficiency of data retrieval. Where normalisation aims to reduce data redundancy for data ingestion, dimensional models can be considered the opposite. Data redundancy is intentionally introduced to the data model improving query performance for data retrieval. Dimension tables contain contextual information such as strings and other attributes. For optimisation purposes FACT tables should only hold integers, keys, flags and dates. This is because a JOIN on an integer or primary key is infinitely faster than on a string due to how strings are comprised. Some would argue that dates should be contained within a separate date dimension however this is left to personal preference in industry.

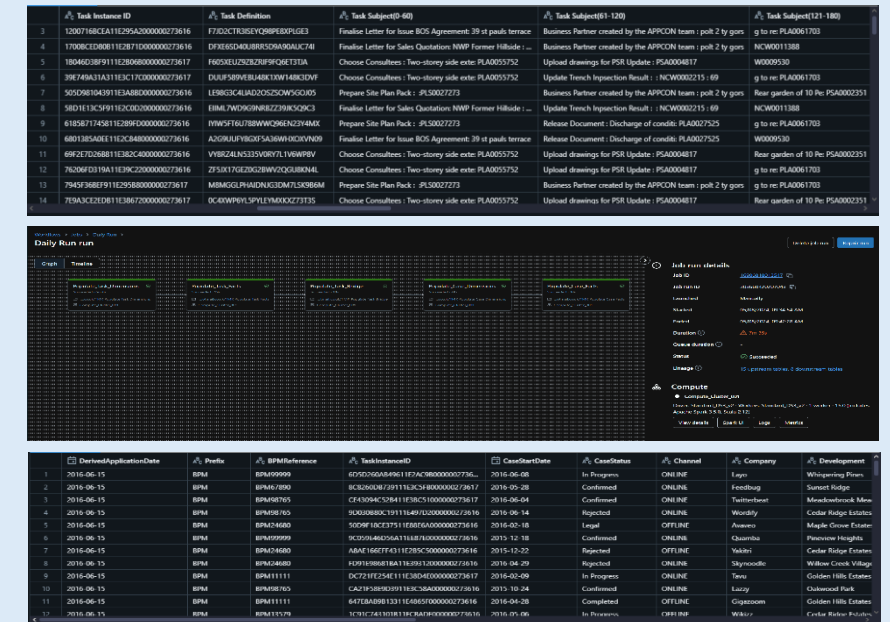
Solution Architecture Design

This infographic illustrates the workflow for migrating on-premises data to a cloud-based Lakehouse Architecture. Data from the BPM source system is transferred to the Data Lake Storage using Move IT. Databricks handles data ingestion and transformation through notebooks, while Unity Catalog ensures robust data governance. The transformed data is then utilized by the TMR reporting suite on the Power BI platform for enhanced reporting and analytics. This streamlined process enhances data management, scalability, and reporting capabilities.



Raw to Dimensionalised

This diagram illustrates the transformation of raw data into a dimensionalised format using Databricks. The top section shows the initial raw data, including task instances and their definitions. The middle section captures a successful run of the Databricks job, which processes and transforms the data. The bottom section displays the final, dimensionalised data, structured for efficient querying and analysis. This process enhances data usability and ensures readiness for advanced analytics and reporting, demonstrating the effectiveness of the implemented Lakehouse Architecture.



Project Outcomes

Summary of Work: Successfully demonstrated the feasibility and benefits of transitioning to a cloud-based data platform using Databricks and Azure Data Lake.

Conclusions:

Enhanced data governance and scalable data management. Significant cost and performance benefits of cloud migration.

Detailed Results:

- Processing Time Reduction: Reduced from 52 minutes to 8 minutes (approximately 70% reduction).
- Cost Savings: Decreased from £3 to £0.66 (approximately 70% reduction).

Future Work:

Implement data quality checks using Unity Catalog. Enhance real-time data processing capabilities.