Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

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 [1]

Aim

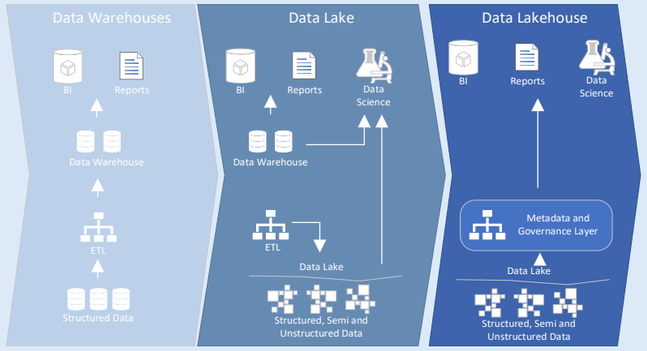
To demonstrate the feasibility and benefits of a Lakehouse Architecture in enhancing data management processes for water network development.

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 [2]

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 [3]

Performance and Scalability:

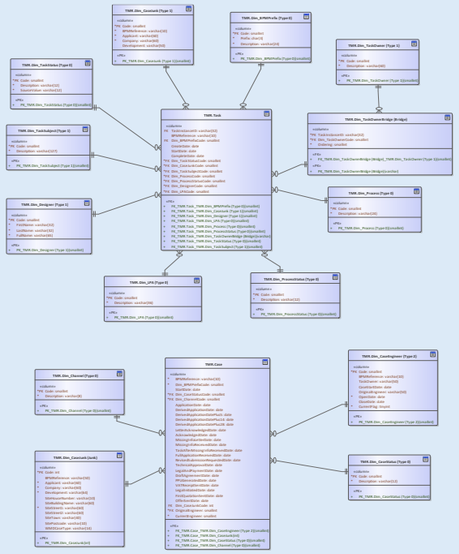
Cloud-based solutions like Databricks offer significant improvements in performance and scalability compared to traditional on-premises systems. Databricks enables faster data processing and can handle larger data volumes efficiently.

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.

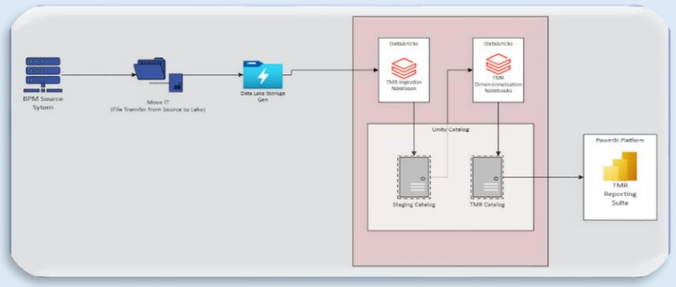


Dimensional Modelling [4]

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.

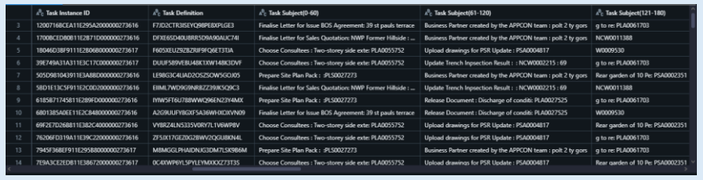
Solution Architecture Design [5]

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.

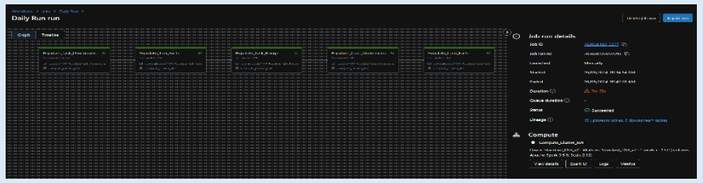


Raw to Dimensionalised [6]

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.



	Task Instance ID	Task Definition	Task Subject(0-60)	Task Subject(61-120)	Task Subject(121-180)
3	1200716BCEA11E295A200000027361 6	F7D2C7E3SEYQ9BPE8DFGLE3	Finalise Letter for Issue BOS Agreement: 39 st pauls terrace	Business Partner created by the APPCON team : polt 2 ty goes	g to re: PLA0061703
4	170B8CD8B9117E2B710000000273616	DFX6EQDABJBBFQ9AN9GALC74I	Finalise Letter for Sales Quotation: NWP Former Hillside : ...	Business Partner created by the APPCON team : polt 2 ty goes	NCW0011388
5	1B0A6C1B9B9117E2B9B000000273617	F6D5XEUZN2BZ8FBFQNETTJTA	Choose Consultees : Two-storey side extn: PLA0055752	Upload drawings for PSR Update : PSA0004817	W0009530
6	39E749A31A3119E3C17C000000273617	DUUF5B9VEBJ4B1XWY14BK0DVF	Choose Consultees : Two-storey side extn: PLA0055752	Update Trench Inspection Result : : NCW0030215 : 69	g to re: PLA0061703
7	505D9B1043911E3A8B000000273616	LE9B3C4LIAD2O5ZSOW5OCJI05	Prepare Site Plan Pack : : PLS0027273	Business Partner created by the APPCON team : polt 2 ty goes	Rear garden of 10 Pe: PSA0002351
8	58D1E13C5F911E2C0D200000027361 6	EBML7WD9GJBNRBZZ39UR5Q9C3	Finalise Letter for Sales Quotation: NWP Former Hillside : ...	Update Trench Inspection Result : : NCW0030215 : 69	NCW0011388
9	61858717458119289F0000000273616	RMW5FT6U7BBWWC9K6NZGYH4MX	Prepare Site Plan Pack : : PLS0027273	Release Document : Discharge of condit: PLA0027525	g to re: PLA0061703
10	6B0138SA0EE1E2C84800000027361 6	A2G9UUFY8GXFSA36WH0XDVN09	Finalise Letter for Issue BOS Agreement: 39 st pauls terrace	Release Document : Discharge of condit: PLA0027525	W0009530
11	69F267D268B11E38ZC4000000273616	VYBIZ4LNS3I3SV0RY7L1V6WP8V	Choose Consultees : Two-storey side extn: PLA0055752	Upload drawings for PSR Update : PSA0004817	Rear garden of 10 Pe: PSA0002351
12	7620A1F0319A11E39K2200000027361 6	ZFSUXTYGEZGBWV2QGUIBXN4L	Choose Consultees : Two-storey side extn: PLA0055752	Upload drawings for PSR Update : PSA0004817	g to re: PLA0061703
13	794SF368D9119E29SB000000273617	MBMGCLPHADNUC3DM7L5R8B6M	Prepare Site Plan Pack : : PLS0027273	Business Partner created by the APPCON team : polt 2 ty goes	g to re: PLA0061703
14	7E9A3C2ED8113E38F200000027361 7	0C4XWP6YL5PVEYMXXXZ7T3J5	Choose Consultees : Two-storey side extn: PLA0055752	Upload drawings for PSR Update : PSA0004817	Rear garden of 10 Pe: PSA0002351



	DerivedApplicationDate	Prefix	BPMReference	TaskInstanceID	CaseStartDate	CaseStatus	Channel	Company	Development
1	2016-06-15	BPM	BPM99999	6D5D360A8496118ZAC9800000027361	2016-06-08	In Progress	ONLINE	Lagu	Whispering Pines
2	2016-06-15	BPM	BPM67890	8C626D8E2B9119JEC3480000027361	2016-05-28	Confirmed	ONLINE	Ferndburg	Sunset Ridge
3	2016-06-15	BPM	BPM98765	CF43094C52841IE3BC51000000273617	2016-06-04	Confirmed	ONLINE	Twitterbeat	Meadowbrook Mew
4	2016-06-15	BPM	BPM98765	9D6D8B8C39119B49JD2000000273616	2016-06-14	Rejected	ONLINE	Wordify	Cedar Ridge Estates
5	2016-06-15	BPM	BPM24680	5D0BF18CE3751IE8E6A000000273616	2016-05-18	Legal	OFFLINE	Avaeno	Maple Grove Estates
6	2016-06-15	BPM	BPM99999	9C205A8AD56A31E8B3D000000273616	2015-12-18	Confirmed	ONLINE	Quantiba	Pinecone Heights
7	2016-06-15	BPM	BPM24680	A8AE166FFC411E289C5000000273616	2015-12-22	Rejected	OFFLINE	Yakiki	Cedar Ridge Estates
8	2016-06-15	BPM	BPM24680	H99198681BA13E39J3J000000273616	2016-04-29	Rejected	ONLINE	Skywoodle	Willow Creek Village
9	2016-06-15	BPM	BPM11111	DC721FE25E111E38D4E000000273617	2016-02-09	In Progress	ONLINE	Tavo	Golden Hills Estates
10	2016-06-15	BPM	BPM98765	CA29F9B9E39119JC36A000000273616	2015-10-24	Confirmed	ONLINE	Lazzy	Oakwood Park
11	2016-06-15	BPM	BPM11111	647EBAD9B1331IE4865F000000273616	2016-04-28	Completed	OFFLINE	Gigazoom	Golden Hills Estates
12	2016-06-15	BPM	BPM11111	3C91C741301B13E3BAB000000273616	2016-05-06	In Progress	OFFLINE	Wikizz	Cedar Ridge Estates

Project Outcomes [7]

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.