Microsoft

RIFUMO
EMPOWERMENT HOLDINGS (PTY) Ltd

Data Platform Modernization and Migration

Table of Contents

1 Executive Summary

1.1 Executive Summary and Key Introduction

This proposal outlines a comprehensive approach to modernizing legacy SSIS and SSAS workloads by migrating them to Microsoft Fabric. The solution leverages Fabric's unified analytics platform to deliver a governed, scalable, and real-time data platform.

1.2 Scope of Work

- Assessment & Roadmap: Inventory SSIS/SSAS assets, profile data, map dependencies, and produce a migration plan, estimates, and a prioritized backlog.
- Fabric Landing Zone: Provision workspaces, OneLake, Dev/Test/Prod; configure RBAC, Managed Identities, Key Vault; set up CI/CD and deployment pipelines.
- ETL Migration (SSIS → Fabric): Refactor control flows to Fabric Data Factory pipelines; data flows to Dataflows Gen2 or Notebooks; implement parameterization, error handling, retries, and observability.
- Semantic Models (SSAS → Fabric): Convert tabular models (TMDL/Tabular Editor) to Fabric semantic models; implement Direct Lake; RLS/OLS and object-level security; validate measure parity.
- Reporting & Power BI: Repoint to Direct Lake/composite models; implement deployment pipelines; tune performance and governance artifacts (shared datasets, endorsed content).
- Governance & Quality: Register assets in Purview; classification and sensitivity labels; lineage; data quality rules; data dictionary and stewardship model.
- Cutover & Hypercare: Parallel run and reconciliation, UAT sign-off, production cutover, decommission SSIS/SSAS, knowledge transfer and runbooks.

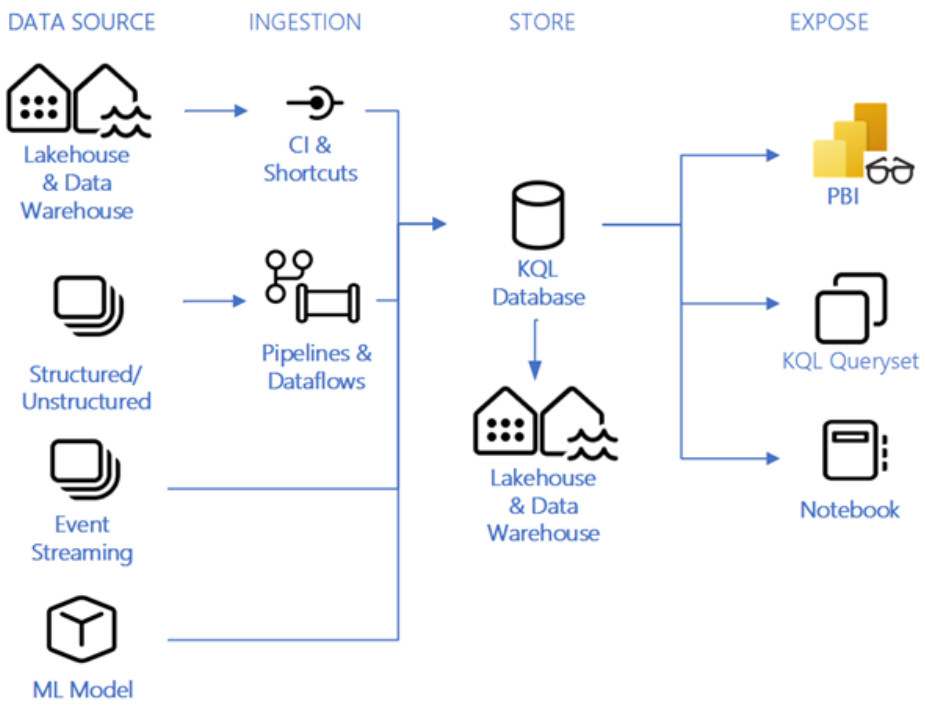
2 Solution Overview

2.1 Data Integration and Dashboards

The architecture leverages Microsoft Fabric as the foundational platform, integrated with Databricks for advanced data engineering, machine learning, and real-time analytics, along with dedicated workspaces to optimize collaboration, management, and data governance.

2.1.1 Layered Architecture Overview

The architecture will be divided into several layers to ensure scalability, security, and seamless integration:



2.1.1.1 Data Ingestion Layer

- **Sources:** Diverse data sources, including on-premise databases (e.g., Oracle, Informix, SQL Server), cloud-based services, streaming platforms, and APIs.
- **Tools:** Microsoft Fabric (Data Factory) for automated data pipelines and Databricks for real-time streaming ingestion (Kafka or Event Hubs).
- **Capabilities:** Supports both batch and streaming ingestion to accommodate diverse use cases.

2.1.1.2 Data Lakehouse Storage Layer

- **Microsoft Fabric OneLake:** Acts as the unified storage layer for structured, semi-structured, and unstructured data.
- **Databricks Delta Lake:** Integrated with OneLake for efficient storage management with features like ACID transactions and schema enforcement.
- **Partitioning and Indexing:** Optimized data storage using partitioning and indexing techniques to enhance query performance.

2.1.1.3 Data Processing and Transformation Layer

- **Databricks Workspaces:** Dedicated environments for ETL development, data engineering, and machine learning model training.
- **Microsoft Fabric Synapse Analytics:** For big data processing, SQL-based analytics, and advanced data transformations.
- **Orchestration Tools:** Integration of Databricks workflows and Fabric pipelines for automation and coordination of complex data processes.

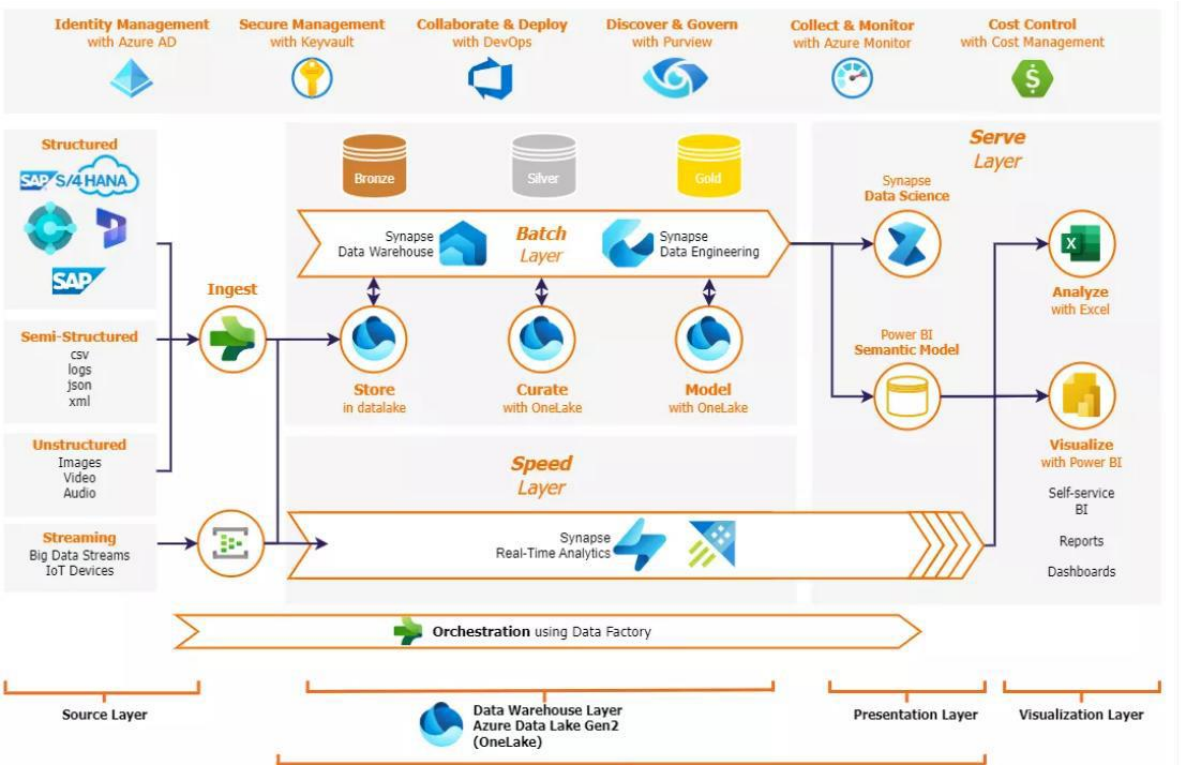
2.1.1.4 Data Governance Layer

- **Microsoft Fabric Compliance Center:** Provides centralized governance with role-based access control (RBAC), encryption, and regulatory compliance (e.g., POPIA).
- **Databricks Unity Catalog:** Ensures consistent metadata management and fine-grained security policies across Databricks environments.
- **Auditing and Monitoring:** Tools like Microsoft Purview and Databricks monitoring features for enhanced governance and observability.

2.1.1.5 Analytics and Visualization Layer

- **Microsoft Power BI:** Integrated with OneLake and Synapse Analytics for real-time data visualization, reporting, and dashboard creation.

- **Databricks SQL:** Enables direct querying of Delta Lake for analytics use cases requiring high performance and scalability.
- **Real-Time Analytics:** Databricks facilitates streaming analytics by processing data in near real-time and rendering insights via Power BI.



2.1.1.6 Key Components

The following components will play a pivotal role in the architecture:

Microsoft Fabric

- **Data Factory:** Automates data ingestion and ETL tasks across diverse sources.
- **OneLake:** Serves as the central storage hub, enabling seamless data management and universal access.
- **Synapse Analytics:** Supports complex data processing and analytical workloads.
- **Power BI:** Provides rich visualization and reporting tools tailored for end users.

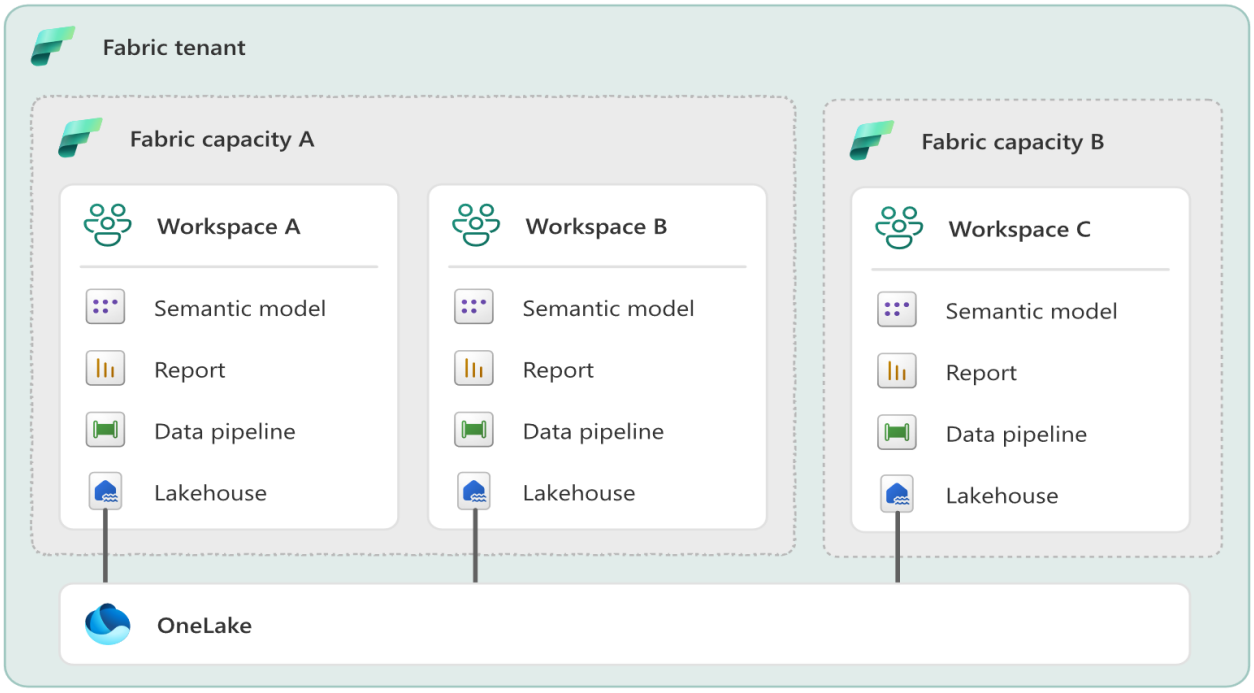
Databricks

- **Delta Lake:** Ensures efficient, reliable, and scalable storage of data.

- **Workspaces:** Offers isolated environments for data engineers and scientists to collaborate on data pipelines, models, and experiments.
- **Streaming Frameworks:** Real-time data ingestion using Kafka, Event Hubs, or Delta Live Tables.

Workspaces

- **Dedicated Databricks Workspaces:** Create specialized environments for machine learning, ETL processes, and data analytics.
- **Microsoft Fabric Workspaces:** Allow teams to manage datasets, analytics reports, and shared dashboards effectively.



2.1.1.7 Workflow Example

Step 1: Data Ingestion

- Raw data is ingested into the architecture via Microsoft Fabric's Data Factory pipelines (batch data) or Databricks streaming mechanisms (real-time data).

Step 2: Storage

- Ingested data is stored in Fabric OneLake and Databricks Delta Lake. Metadata is captured in Databricks Unity Catalog for governance.

Step 3: Processing

- Data engineers use Databricks Workspaces for ETL operations and transformation, with workflows orchestrated through Microsoft Fabric pipelines.

Step 4: Analytics

- Transformed data is made available to Power BI for visualization and reporting. Analysts access data directly via Databricks SQL for ad-hoc queries.

Step 5: Governance and Security

- Role-based access control (RBAC), auditing, and encryption are enforced by Microsoft Fabric Compliance Center and Databricks Unity Catalog.
- 9.Benefits of This Architecture
- **Unified Platform:** Integrates Microsoft Fabric and Databricks seamlessly for streamlined data management.
- **Scalability:** Accommodates large datasets and diverse analytics requirements.
- **Flexibility:** Offers dedicated workspaces for specialized tasks, improving collaboration and productivity.
- **Real-Time Capabilities:** Enables near real-time data processing and analytics.
- **Security and Governance:** Ensures robust protection and compliance with industry standards.

2.1.2 Methodology and Project Approach: Data Platform Modernization

2.1.2.1 Phase-Wise Implementation Plan

Each phase lays the foundation for subsequent stages. The roadmap ensures technical maturity and adoption readiness.

Phase 1: Discovery & Assessment

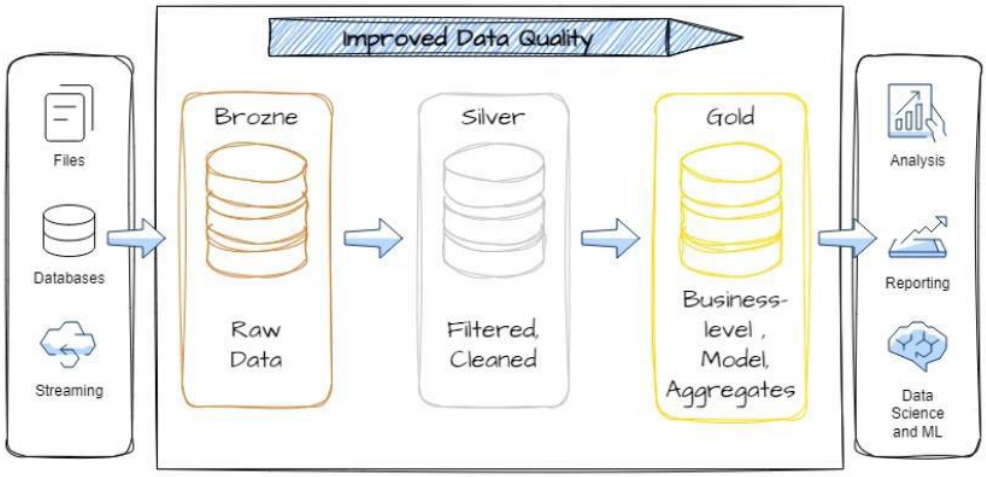
- Inventory of legacy and current systems.
- Data profiling, quality checks, and lineage mapping.
- Catalog metadata and identify integration touchpoints.

Phase 2: Platform Design and Architecture

- Deploy Microsoft Fabric with OneLake, Dataflows Gen2, and Spark notebooks.
- Implement Medallion architecture:
    - Bronze: Raw data ingestion.
    - Silver: Cleansed, conformed data.
    - Gold: Business-ready models for reporting.

Phase 3: Data Integration and Pipeline Engineering

- Use Fabric Pipelines to ingest legacy data and automate ETL.
- Define pipeline schedules, error handling, and dependency logic.
- Integrate batch and real-time ingestion patterns.



Medallion Architecture

2.1.2.2 Medallion Architecture Overview (Fabric Lakehouse)

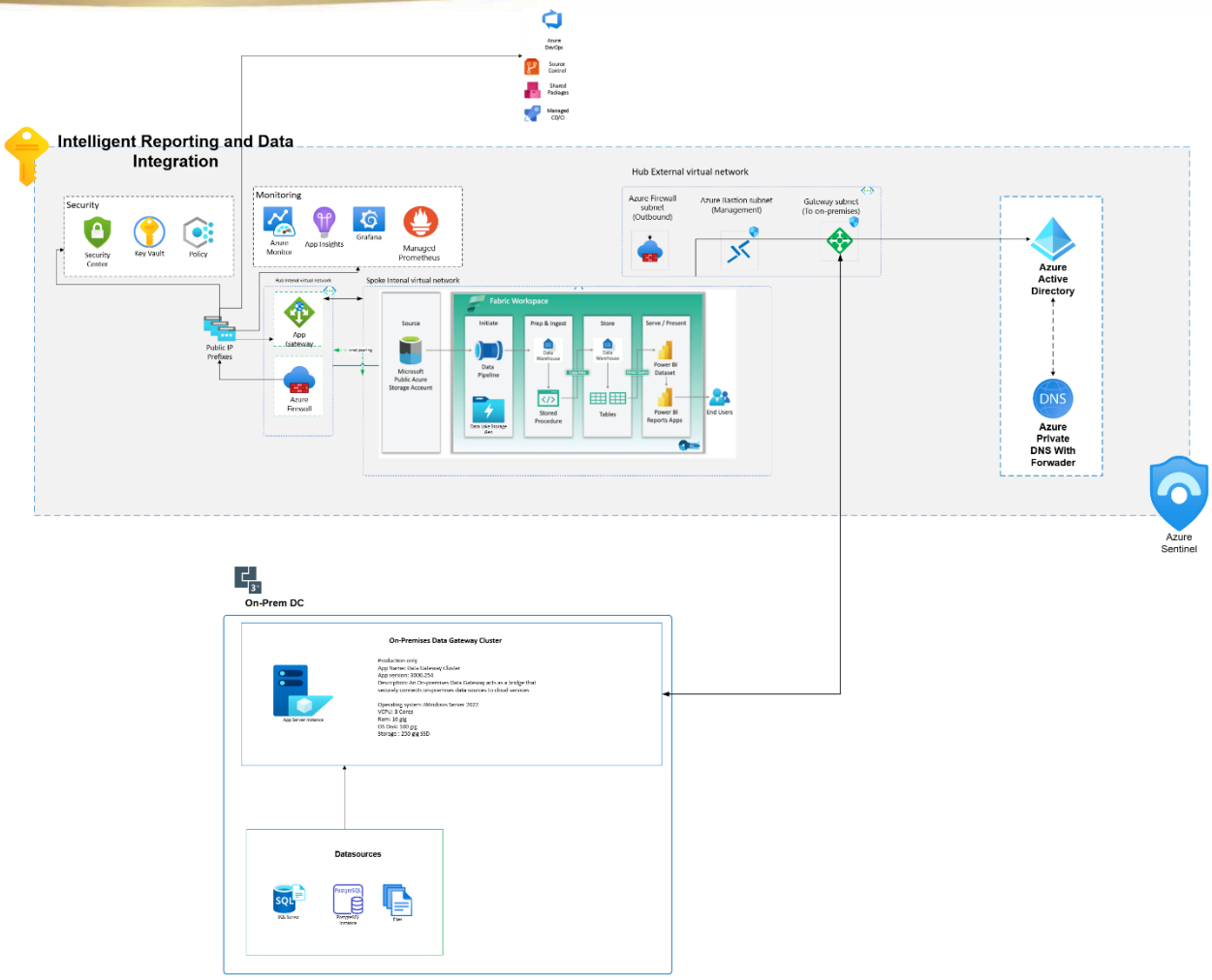
Layer	Purpose	Key Tools in Fabric
Bronze	Raw data ingestion and storage	Dataflows Gen2, Data Pipelines, OneLake
Silver	Cleaned and transformed business data	Lakehouse SQL, Notebooks, Spark
Gold	Business-ready, aggregated/curated datasets	Lakehouse SQL, Power BI, Warehouse

2.1.2.3 Data Gateway Design

Secure and scalable gateway infrastructure is critical to hybrid integration.

Design Principles:

- Implement Microsoft On-premises Data Gateway for SAP
- Isolate gateway configuration per environment (Dev/Test/Prod).
- Monitor availability and performance using gateway metrics.



2.1.2.4 Data Cleansing, Standardization, and Verification

Maintaining data quality is essential to usability and trust.

Techniques Used:

- Referential Integrity Checks to validate relationships between entities.
- Rule-based Validation against business logic
- Duplicate resolution and enrichment routines.
- Anomaly detection using AI modelling within Fabric notebooks.

2.1.2.5 Data Governance and Security Strategy

Compliance and control form the backbone of enterprise-grade modernization.

Governance Tactics:

- Implement Microsoft Purview for metadata management and lineage.
- Apply sensitivity labels and RBAC across workspaces.
- Utilize Azure Key Vault for secure credential and key storage.
- Enable audit logging for traceability and monitoring.

2.1.2.6 Workspace Design and Semantic Modelling

Modular workspace design allows structured development and controlled access.

Design Elements:

- Create separate workspaces by data domain.
- Enforce naming conventions and role-based access.
- Define semantic models with calculated measures, KPIs, and hierarchies.
- Use deployment pipelines for lifecycle promotion from Dev to Prod.

Real-Time Reporting Architecture Using Power BI

Responsive, self-service analytics empower users across the organization.

Architecture Overview:

- Direct Lake Mode ensures high-speed dashboard rendering.
- Build composite models for learner, financial, and operational domains.
- Embed dashboards into existing portals for seamless user experience.
- Configure filters, drill-through, and navigation for decision support.

- Training, Support, and Change Enablement

Sustainable change requires education, documentation, and ongoing engagement.

Approach to Enablement:

- Design targeted training tracks for analysts, managers, and IT staff.
- Provide embedded help features and glossary terms within dashboards.
- Collect feedback for iterative improvement.
- Schedule quarterly adoption review sessions.

- Performance Optimization and Monitoring

Ongoing measurement ensures platform health and data reliability.

Monitoring Strategies:

- Use Fabric's built-in metrics for pipeline performance.
- Track dashboard load times and user engagement.
- Implement alerting for data failures and ingestion anomalies.
- Define thresholds for data freshness and reporting lag.

2.1.3 Migration Approach & Methodology

We deliver in two-week sprints using a Strangler-Fig approach: wrap legacy endpoints, build new pipelines and models in Fabric, run in parallel for reconciliation, then switch by domain. Continuous testing, security, and observability are applied throughout.

2.1.3.1 SSIS → Fabric Patterns

Control flows → Fabric Data Factory pipelines; data flows → Dataflows Gen2 or Notebooks.

Externalize configs and secrets to Key Vault; parameterize per environment (Dev/Test/Prod).

Observability via Fabric monitoring, Application Insights custom events, and alerts; robust retries and compensation.

2.1.3.2 SSAS → Fabric Semantic Models

Inventory measures, partitions, and roles; convert model metadata (TMDL/Tabular Editor).

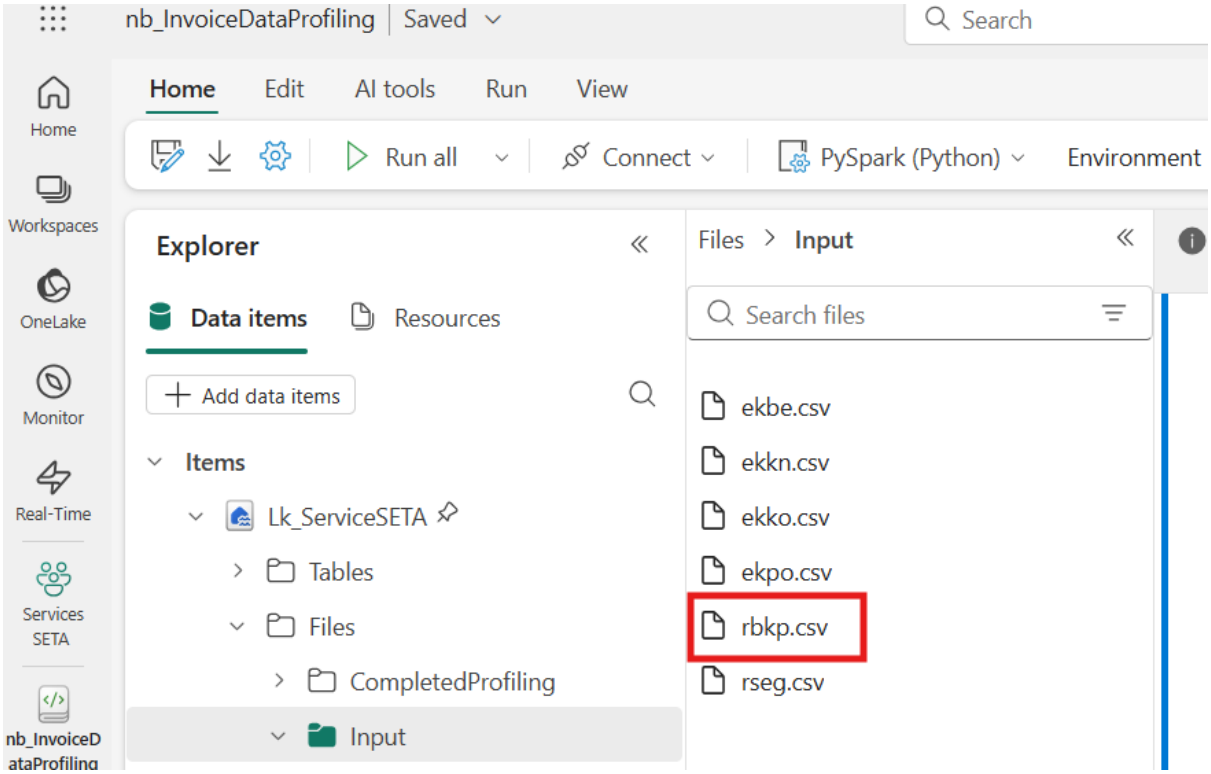
Optimize for Direct Lake (star schemas, aggregations, partition pruning).

Implement RLS/OLS; validate DAX parity and calculation accuracy with test packs.

2.1.4 Demo

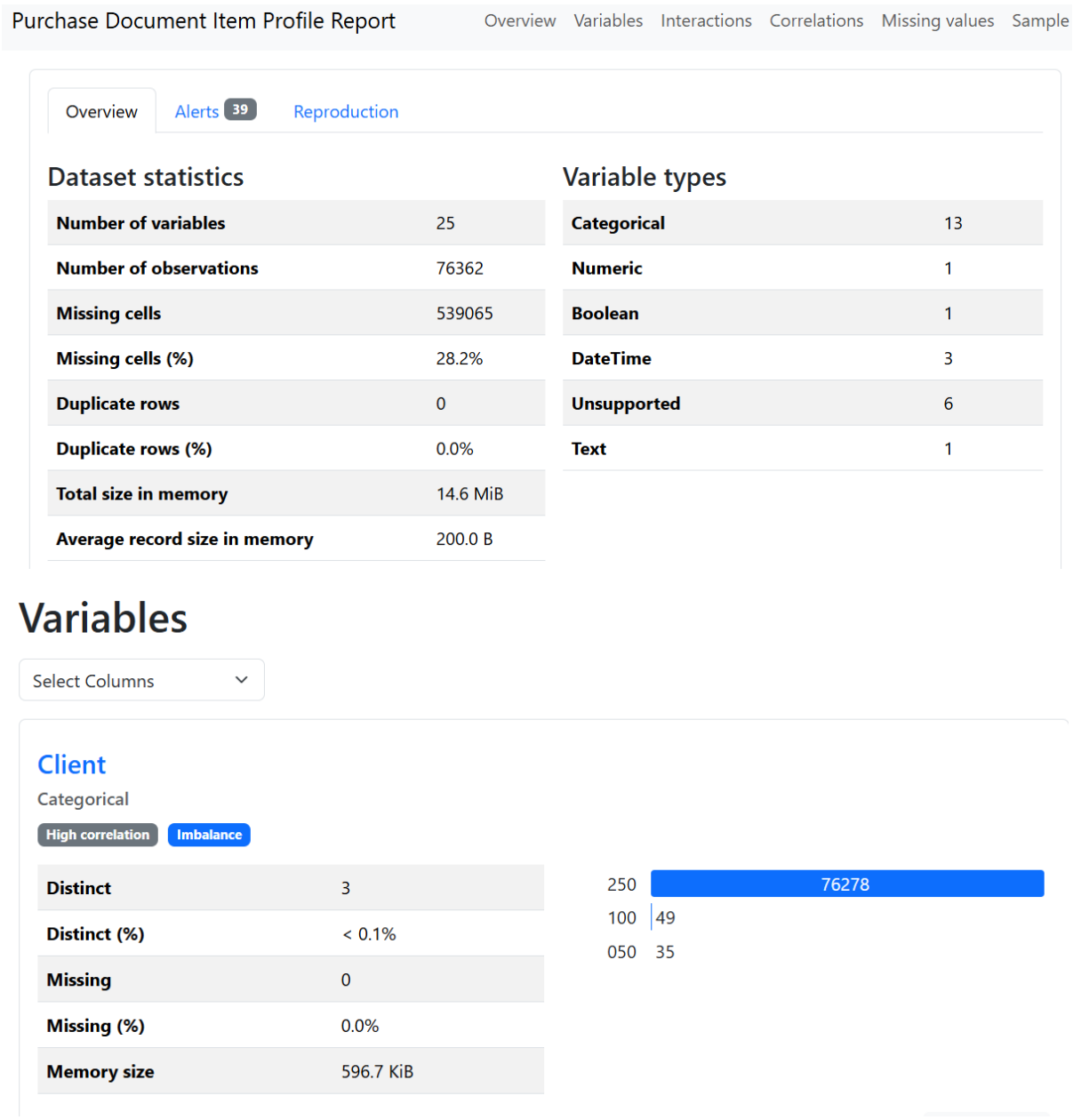
Bronze Layer

This layer is taking source data from legacy system (e.g Oracle) into the datalake (first layer of medallion). See below the RBKP file dropped into the datalake.



Data Profiling

Using Python Spark tool we create data profiling results for different source files. The results are stored in the OneLake folder as html files. See below examples:

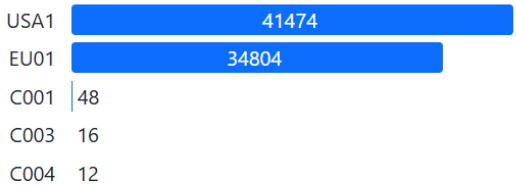


CompanyCode

Categorical

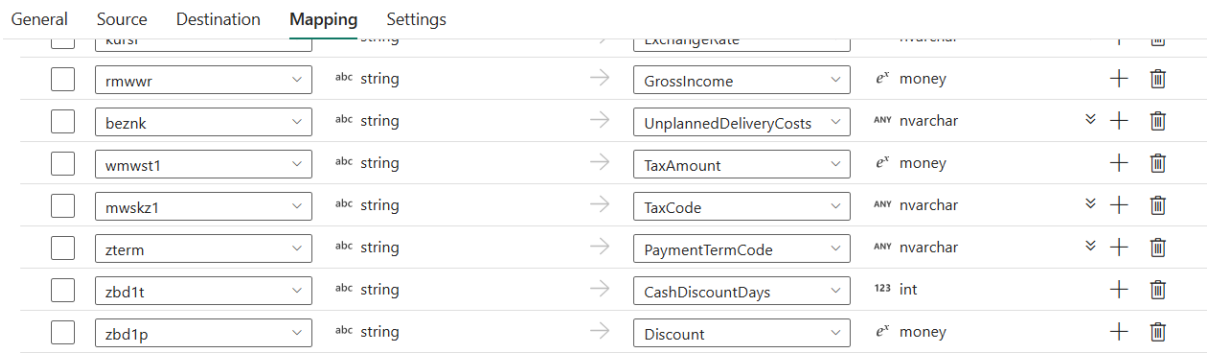
High correlation | Imbalance

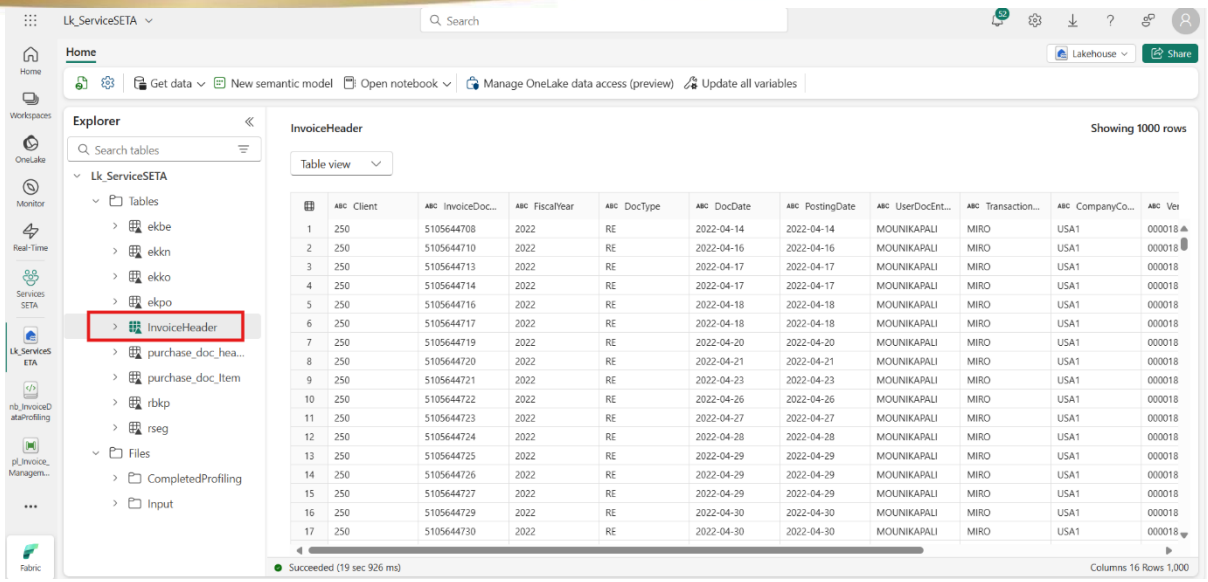
Distinct	6
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	596.7 KiB



Silver Medallion Layer

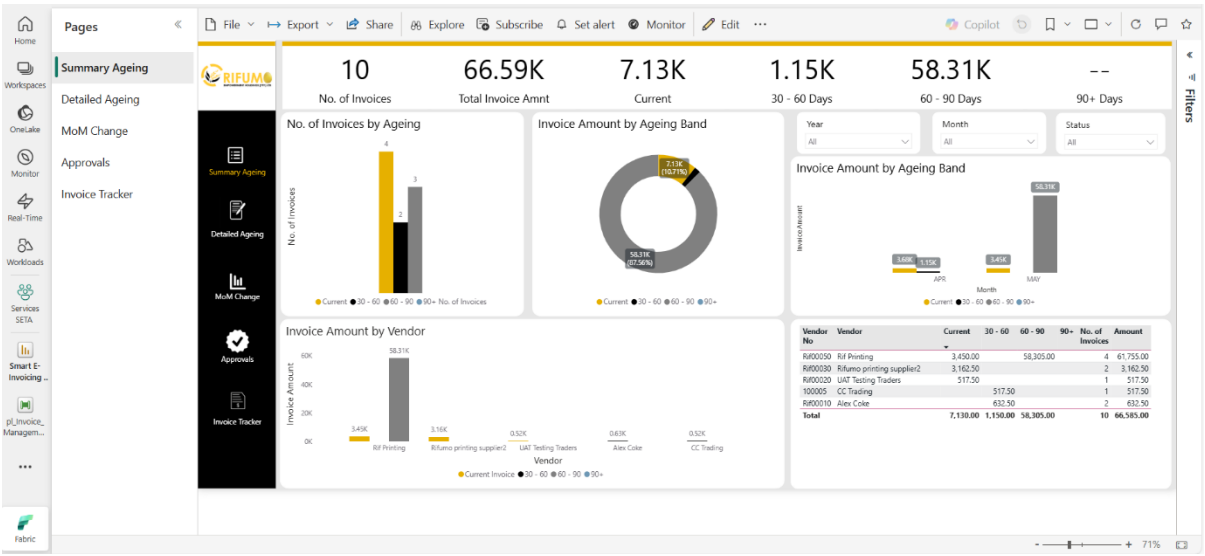
After applying any business rules, mapping, and data cleanup using stored procedures based on data profiling and data dictionary, the files are stored as tables in the database, see below:

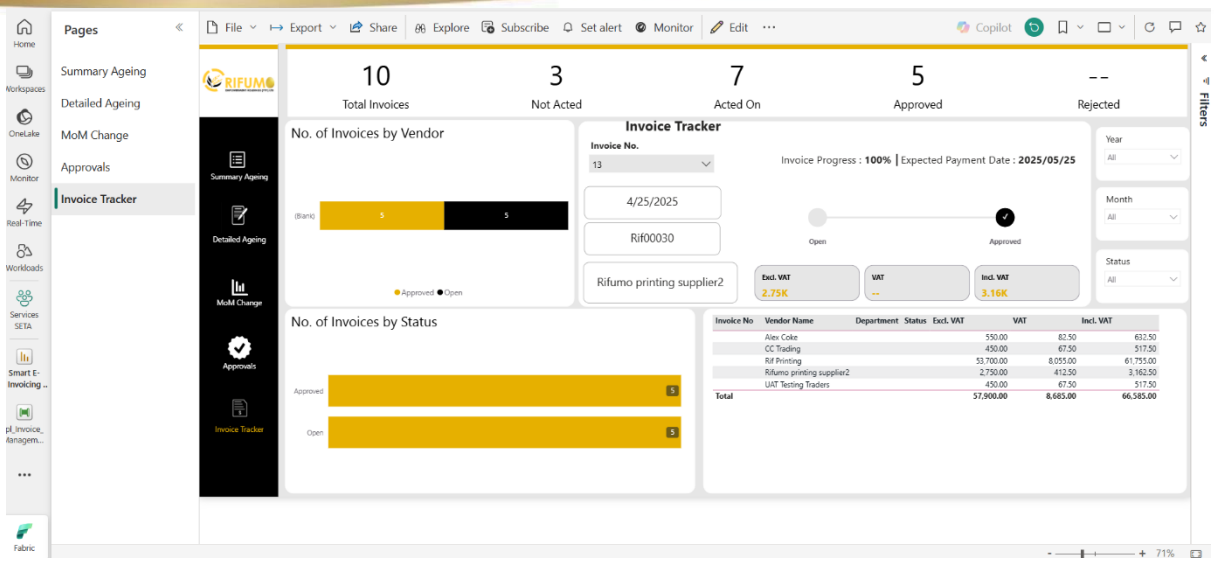




Gold Medallion Layer

In the 3rd layer of the medallion architecture, the data is then reported using power BI which is integrated in the fabric workspace.





2.1.4.1 Data Handling and Protection

Data Obfuscation

Data obfuscation is the process of replacing sensitive information with data that looks like real production information, making it useless to malicious actors. It is primarily used in test or development environments—developers and testers need realistic data to build and test software, but they do not need to see the real data.

During the Customer Review Sessions of the Source-to-Target and Report Mappings recommendations for Sensitive and Private Data will be reviewed for Data Obfuscation Decisions.

There are three primary data obfuscation techniques:

I. **Masking-Out**

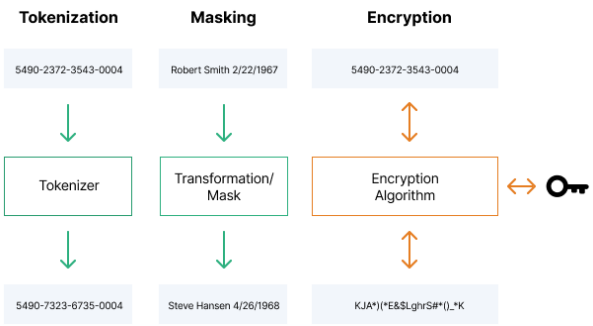
Is a way to create different versions of the data with a similar structure. The data type does not change, only the value change. Data can be modified in a number of ways, for example shifting numbers or letters, replacing words, and switching partial data between records.

II. **Data encryption**

Uses cryptographic methods, usually symmetric or private/pub key systems to codify the data, making it completely unusable until decrypted. Encryption is very secure, but when you encrypt your data, you cannot manipulate or analyze it.

III. **Data tokenization**

Replaces certain data with meaningless values. However, authorized users can connect the token to the original data. Token data can be used in production environments, for example, to execute financial transactions without the need to transmit a credit card number to an external processor.



Key reasons organizations rely on data obfuscation methods:

- **Third parties can't be trusted**—sending personal data, payment card information or health information to any third party is dangerous. There is a dual risk—it increases the number of people who have access to the data beyond the organization's control, and it exposes the organization to violations of regulations and standards.
- **Business operations may not need real data**—any use of customer, employee, or user data is risky because it exposes the data to employees, contractors, and others. Many business processes, such as development, testing, analytics, and reporting, do not necessarily need to process real personal data. By obfuscating the data, the organization can maintain the business process but eliminate the risk.
- **Compliance—**many compliance standards require data to be obfuscated under certain conditions.

Data Security

As the data being stored and presented is extremely sensitive role-based security has been requested for the Power BI data model.

- Role based security will be implemented on an organizational hierarchy perspective.
- Where if a user has access to the Head Office Layer, he/she would be able to see everything at head office and below,.
- Only users with Active Directory Accounts will be able to access the Power BI reports, and the organisation will need to provide a list of which user groups (Not specific individual users) need to have access at which levels.

3 Data Governance, Security & Compliance

Catalog & Lineage: Register all data products in Microsoft Purview; define domains/collections; enable end-to-end lineage.

Security: Entra ID RBAC/ABAC, Managed Identities; encryption at rest and in transit; secrets in Key Vault.

Compliance: Sensitivity labels; retention policies; immutable audit logs for critical events; POPIA-aligned controls.

Operations: SRE runbooks, alerting, and dashboards; measurable SLOs and error budgets.

4 Timelines and Pricing

Phase 1 – Assessment (2–4 weeks): asset catalog, profiling, migration backlog, and estimates.

Phase 2 – Foundations (2 weeks): Fabric landing zone, Dev/Test/Prod, CI/CD, standards.

Phase 3 – Migration Wave 1 (4–6 weeks): priority SSIS pipelines and first semantic model.

Phase 4 – Migration Wave 2 (4–6 weeks): remaining pipelines/models; performance tuning.

Phase 5 – Cutover & Hypercare (2 weeks): parallel run, reconciliation, decommission.

Milestone	% of Contract (ex VAT)	Amount (USD)
Mobilization & Architecture sign-off	**20%**	**$ 229,240.87**
Foundations ready (Fabric tenant, LZ, CI/CD)	**20%**	**$ 229,240.87**
Core migration wave complete (ETL + first semantic models)	**30%**	**$ 343,861.30**
UAT sign-off across priority domains	**20%**	**$ 229,240.87**
Go-Live & Hypercare complete	**10%**	**$ 114,620.43**

4.1 Pricing & Commercials

Item	Amount (ZAR)	Amount (USD)
Professional services (ex VAT)	R 20,230,506.54	$ 1,146,204.34
VAT 15% (if applicable)	R 3,034,575.98	$ 171,930.65
Total (incl. VAT)	R 23,265,082.52	$ 1,318,134.99

Azure Marketplace Plans (USD, ex VAT):

Plan	Description	Duration	Price (USD)
Assessment FastStart	Fixed-scope SSIS/SSAS assessment & migration roadmap	2–4 weeks	$ 114,620.43
Migration Foundation	Landing zone + first wave ETL & one semantic model	6–8 weeks	$ 343,861.30
Full Modernization	All waves migration, governance, cutover & hypercare	12–16 weeks	$ 687,722.60

4.2 Assumptions, Dependencies & Exclusions

Azure consumption and Fabric capacity are excluded unless stated otherwise.

Access to source systems, SMEs, and interface specifications provided on schedule.

Delivery primarily remote; T&E billed at cost if on-site is required (pre-approved).

Security reviews and penetration testing beyond standard hardening can be scoped separately.

4.3 Risks & Mitigations

Risk	Likelihood	Impact	Mitigation
Legacy SSIS packages rely on undocumented business logic	**Medium**	**High**	**Reverse-engineer with SMEs; create test harness; stage migration; parallel run for 2 sprints**
Model performance issues at scale	**Medium**	**High**	**Star schema; aggregations; incremental refresh; Direct Lake where feasible**
Capacity under-sized leading to throttling	**Low**	**High**	**Monitor Fabric metrics; tune pipelines; scale reservations as needed**
Stakeholder availability for UAT	**Medium**	**Medium**	**Block calendars early; UAT playbooks; BA-led daily triage**

Security audit findings late in cycle	**Low**	**High**	**Security checkpoints from Sprint 2; Defender & Purview scans continuously**

4.4 Team & RACI (Summary)

Programme Management (PM): governance, schedule, risk & stakeholder management.

Solution/Technical Leads (SA/TL): architecture, standards, performance & security patterns.

Lead Data Engineer (LDE) & Data Engineers (DE): ingestion, transformation, orchestration.

Lead BI Developer (LBD) & BI Developers (BD): semantic models, DAX, dashboards.

Data Analysts (DA) & Business Analysts (BA): requirements, mapping, testing, change & training.

Site Reliability/DevOps (SRE/SD): CI/CD, monitoring, IaC, environments.

.