Datamart ETL Solution – Executive One-Pager

# 1. Project Overview

Project Name: Datamart ETL Solution

Business Objective: Implement a scalable, medallion architecture-based ETL solution to consolidate and process retail point-of-sale (POS), inventory, and financial data from multiple sources. This enables comprehensive analytics and reporting across various business dimensions.

## Business Logic:

* Ingest raw data from multiple retail systems
* Maintain data lineage and point-in-time recovery
* Optimize data for analytical query performance
* Support both batch and selective data processing
* Transform transactional data with business rules, including filtering, calculations, classification, enrichment, bundle explosion, and net metric calculations
* Multi-dimensional aggregation supporting eight analytical perspectives

# 2. Architecture Design

## High-Level Architecture:

Source Systems → Bronze Layer → Silver Layer → Gold Layer → Analytics

## Source Data:

* ERP outbound, returns, and POS systems (cloud storage)
* Multiple retail data systems
* Customer and product master data
* Order and shipment staging data
* Stratification and product management systems

## Data Retention Plan:

* Bronze Layer: Point-in-time backups with hierarchical date structure
* Silver Layer: Mix of persistent tables and temporary views
* Gold Layer: Aggregated tables with historical data
* Backups stored in partitioned Parquet files

## Cloud Services Used:

* Cloud-based compute platform
* Cloud storage for data lake
* Analytics and configuration management services
* Secure credential management

## Resource Group Structure:

* Development, test, and production environments
* Catalogs for raw, enriched, and aggregated data
* Planned integration with analytics services

# 3. Deployment Strategy

## CI/CD Pipelines:

* Configuration-driven deployment
* Modular functions for load, backup, and optimization
* SQL-based enrichment and aggregation
* Standardized processing patterns for Gold Layer

## Environment Strategy:

* Development environment operational
* Test and production environments planned

# 4. Resource Configuration

## Compute:

* Job clusters for scheduled operations
* Interactive clusters for development
* Autoscaling for optimization tasks

## Storage:

* Primary storage in cloud data lake
* Delta tables for all layers
* Backups in partitioned Parquet format

## Databases:

* Catalog-enabled data management
* Multiple tables across Bronze, Silver, and Gold layers
* Aggregated tables for various business perspectives
* Configuration management planned via analytics service

# 5. Security & Compliance

* User-based workspace access control
* Role-based access to catalogs and schemas
* Secure credential storage (implementation details pending)

# 6. Cost Management

* Query performance optimization
* Storage efficiency via partitioning
* Scheduled optimization to reduce compute costs
* Configuration-driven and incremental processing
* Standardized patterns for consistent performance

# 7. Operations & Maintenance

* Timestamp tracking for all operations
* Error logging and execution status tracking
* Conditional execution to prevent empty runs
* Automated backups and selective recovery
* Original data format preservation

# Key Performance Indicators

* 15+ Bronze, 12+ Silver, 8 Gold tables managed
* 400M+ records processed incrementally
* Automated duplicate detection and data quality checks
* Storage optimization and backup granularity
* Comprehensive business intelligence coverage
* Multi-dimensional analysis and historical data from July 2021 onward

# Advanced Features

* Three-stream processing for POS, outbound, and return data
* Multi-source aggregation and conditional filtering
* Dimensional flexibility and BOM explosion tracking
* Consistent architecture and reusable components
* Parameterized processing for easy modification

# Next Steps

* 1. Complete analytics integration for configuration management
* 2. Standardize unique identifiers across Silver tables
* 3. Document source and logic for combined item attributes
* 4. Implement incremental processing for additional layers
* 5. Enhance configuration tracking
* 6. Document prepack logic
* 7. Establish test and production environments
* 8. Implement automated data quality monitoring
* 9. Optimize clustering strategies for multi-dimensional queries
* 10. Implement automated performance monitoring
* 11. Create business user documentation for analytical use cases

# Business Impact

* Delivery of daily and weekly KPIs at account and product levels
* Multi-level customer and inventory analysis
* Bundle analytics and historical trending
* Operational benefits through automation and optimization