Approach Note: Migrating from Netezza to Big Data Platform

Ganesh Sankaran

LatentView Analytics

Table of Contents

[Business Context 3](#_Toc490816777)

[Current Situation 3](#_Toc490816778)

[Architecture Overview 3](#_Toc490816779)

[Issues with the Current Situation 5](#_Toc490816780)

[Requirements of the new platform 6](#_Toc490816781)

[Proposed Solution 6](#_Toc490816782)

[Guiding Questions 6](#_Toc490816783)

[Solution Approach 7](#_Toc490816784)

[Deliverables 8](#_Toc490816785)

[Timelines 9](#_Toc490816786)

[Pricing 9](#_Toc490816787)

# Business Context

The data management solutions and support team in the GADS organization in MetLife is responsible for gathering data from different source systems, processing them, making them available for analysis, reporting, modeling and applications.

The team is undertaking a data refactoring exercise that aims to achieve the following goals: reduce redundancy in ETLs (avoid having ETL logic in different layers), and align with the new MetLife processing and technology architecture as well as the next gen data science tools. In the new architecture, there is a designated RDZ (raw data zone) to be the staging area for all source data across the enterprise, from where data is transformed and loaded into IDZ (integrated data zone). Data can be delivered from IDZ to various databases and downstream applications.

One of the critical infrastructure components in the Dental BU is the Research & Experimentation Data Mart that runs on Netezza. There are several issues with the Netezza platform in its current form ([outlined in the next section](#_Issues_with_the)). As part of a phased approach in data refactoring, the team wants to migrate away from Netezza, and move to a new platform. The New Data Platform must have the capabilities to run multiple workloads, including data science processes (modeling, scoring, experimentation), visualization, reporting and ad-hoc querying, while also allowing write-backs and seamless closed-loop communication with the rest of the ecosystem. The New Data Platform (“Research, Experimentation, Visualization, Analytics”) needs to leverage Big Data technologies for overall alignment with Enterprise Architecture, scalability, performance, maintainability, and efficiency.

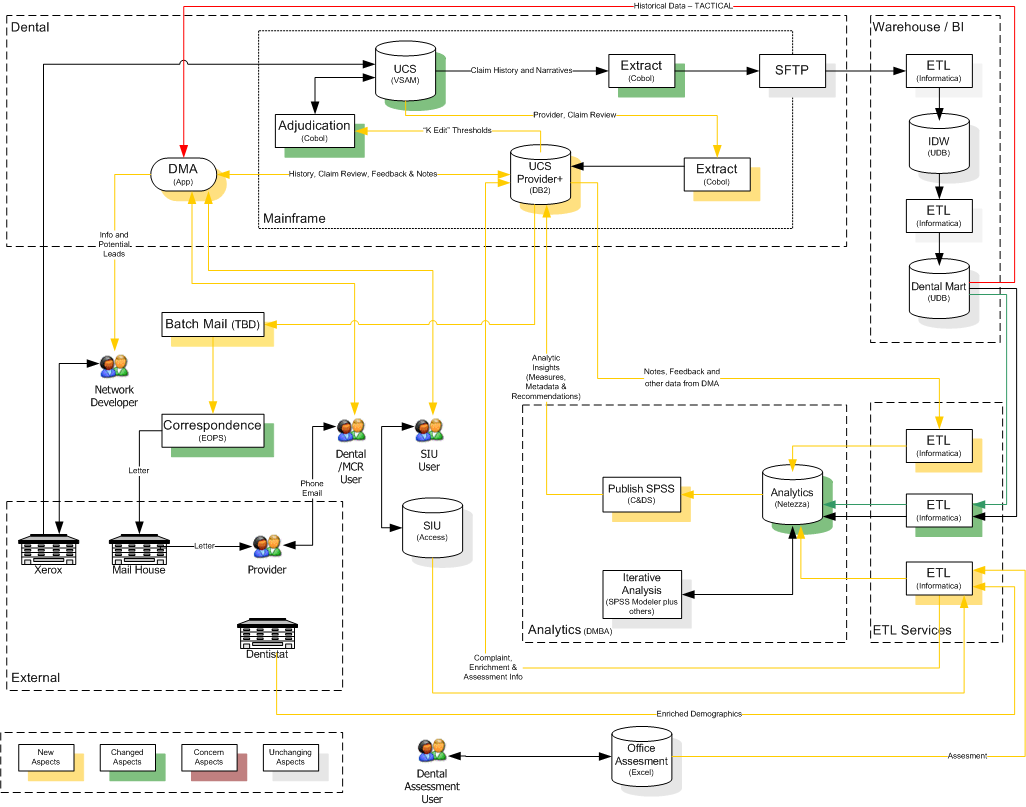
The objective of this document is to describe a solution approach to help MetLife migrate from Netezza to the new “Research, Experimentation, Visualization & Analytics” environment (called the “New Data Platform” henceforth).

We propose a consulting exercise to define the roadmap (or leverage existing MetLife roadmaps), collect detailed requirements, identify the technology and data flow architecture and create a detailed project plan for migrating to the New Data Platform.

# Current Situation

This section describes at a high level, the major source systems and the data flows in the Dental BU, with an emphasis on the data flows into the Research & Experimentation Data Mart.

## Architecture Overview

1. The operational applications / systems that provide data for the Research & Experimentation Data Mart are: UCS, Adjudication, UCS Provider+, and DMA (App)
2. Data from UCS is extracted and published to SFTP. From here, an ETL process (Informatica) loads the data into a staging area, performs transformations and the transformed data is loaded into IDW (UDB). In a subsequent processing step (Informatica), data is transformed and loaded into Dental Mart (UDB). Data from Dental Mart is loaded into DMA (App)
3. Another ETL process (Informatica) extracts data from UCS Provider+, and transforms and loads it into the Research & Experimentation Data Mart (Netezza)
4. There are several other ETL process (Informatica) that load, transform, aggregate and integrate data from various systems into the Research & Experimentation Data Mart:
   1. SIU (Special Investigation Unit)
   2. Office Assessment
   3. External Datasets
5. From the analytics data mart, data science teams use SPSS to develop and run their own ETL processes. These processes source data from the Research & Experimentation Data Mart, do transformations, summarizations and aggregations, and finally write back to the Research & Experimentation Data Mart
6. The Research & Experimentation Data Mart is also used by data science teams for building, validating and scoring predictive models, and running experiments
7. The Research & Experimentation Data Mart is also used to develop reporting applications (Qlik). Some of the Qlik reports have complex queries with embedded business logic
8. Data from the Research & Experimentation Data Mart is published into UCS Provider + through SPSS C&DS; data is also pushed into DMA (App), thus making it a closed loop process

## Issues with the Current Situation

There are several known issues with the current scenario:

1. ETL Complexities –
   1. ETL processing happens in several different places: UCS 🡪 IDW, IDW 🡪 Dental Mart, Dental Mart 🡪 Research & Experimentation Data Mart, UCS Provider+ 🡪 Research & Experimentation Data Mart, etc. leads to maintenance complexities
   2. ETL processing logic is distributed across many different tools: Informatica, SPSS and sometimes even embedded in Qlik leads to inconsistencies of business definitions and no single source of truth
   3. Many of the SPSS Streams do not have nuggets, and they are used for data management activities, resulting in potential inefficiencies, redundancies, lack of visibility into the definitions
2. Technology Complexities –
   1. **Netezza:** Netezza was not intended to serve as a production instance for certain downstream applications, but only as an experimentation database
      1. A mix of development, test and production SPSS streams cause maintenance and redundancy issues
      2. The lack of adequate back-up or disaster recovery options results in potential risks for the business
   2. **Next Gen Data Science Platform:** The data science team is likely to move away from SPSS and move towards a different NextGen Analytics platform (open source Python, R)
      1. The SPSS Streams need to be migrated to a different platform to ensure maintainability of the solution

## Requirements of the new platform

In order to address some of the issues above, the New Data Platform needs to be built with the following requirements:

1. **Align with MetLife Enterprise Architecture:** MetLife is looking to move to a more streamlined enterprise data architecture (with the designated RDZ and the IDZ). The new Data Platform needs to be compatible with this architecture
2. **Write-back Support:** The New Data Platform should support write-back from other applications to ensure closed loop business intelligence
3. **Multiple Workload Support:**
   1. Batch processing, work well with Informatica
   2. Ad-hoc SQL querying interface for multi-user analytical queries on large volumes of data, both known and unknown query patterns
   3. Reporting & Visualization tools such as Qlik
   4. Continue the support for data science process – model development, scoring, test & learn experimentation
   5. In the future, support for real time analysis
   6. Achieve all this with guarantees to ensure SLA adherence
4. **Robust DR and back-ups support**

# Proposed Solution

LatentView proposes a consulting exercise to define the roadmap and implementation plan for the new Data Platform. The objective of the consulting exercise is to identify objectives, create a roadmap, discover all the unknowns, collect requirements, evaluate and choose the tools, define the architecture, design the data flows and data models, and create a detailed project plan.

## Guiding Questions

The goals for the consulting exercise would be to develop answers to the following key questions:

1. What is the technology Architecture for the New Data Platform? What are the appropriate technology choices? What are the trade-offs between various choices?
2. What is the data processing architecture for the New Data Platform? How does data flow into the New Data Platform? How does the New Data Platform serve its end users (data scientists, reporting applications and operational systems)? How do users access data in the New Data Platform?
3. How do we implement the migration from the current state to the future state? What does the data model look like? How do we load the existing data? How do we migrate the logic embedded in SPSS Streams into Informatica? How do we design the ETLs? How does the data flow from the New Data Platform into other systems?
4. How does the data processing architecture for the New Data Platform adapt to the expected changes to the MetLife Enterprise Data Architecture in the medium term? What happens to the SPSS Streams once SPSS is decommissioned? How do we replace C&DS with a different solution, if required?

## Solution Approach

LatentView will follow its proven DARE (**D**iscover, **A**nalyze, **R**ecommend, **E**ngineer) framework for this exercise. DARE is a structured methodology that has evolved from experience with many customers in helping them move towards an optimal state of BI infrastructure.



**Discover**

* Conduct discussions with enterprise architects, source system owners, program owners, review documents to understand drivers for refactoring, issues with current infrastructure and expected results to achieve
* Understand data sources, ETL processes, data marts, enterprise data model, data quality, security, standards and processes, tools and technologies. Specifically, understand the following in MetLife environment:
  + Source systems – UCS, UCS Provider +, DMA, SIU, external sources
  + Data Marts – DMA, IDW, DentalMart, SQL Server infrastructure
  + ETL Processes – Informatica, Shell Scripts, SPSS Streams, Qlik queries
  + Data Science Processes – Modeling, Scoring, Experimentation
  + Reporting & Information Delivery – Qlik, downstream systems
  + Loop back to Source System Applications – DMA, UCS Provider+
* Understand the evolution of MetLife enterprise data architecture roadmap

**Analyze**

* Analyze source system dependencies, gaps in data model, information delivery, ETL
* Review documents or Reverse Engineer the ETL processes to map data flows from Source to Staging Areas to Target
* Collect sample data from sources, estimate data volumes, analyze process metadata for each of the sources as well as for each of the ETL processes outlined above in the Discover phase
* Understand in detail the subsystems of ETL: the source to target mappings, data extract logic, dimension management strategies, ongoing load process, management processes

**Recommend**

* Roadmap for migrating to the new Architecture
* Conceptual design of the New Data Platform
* Technology architecture, tool choices and evaluation
* Backroom and front-room architecture in alignment with MetLife enterprise data architecture
* Design of the dimensional model, logical and physical design, design of the ETL
* Strategies for migrating to the New Data Platform
* Prepare detailed project plan, estimate of pricing and timelines

**Engineer**

* This is not in scope for the consulting exercise

## Deliverables

The deliverables from the proposed exercise will be as follows:

1. Conceptual Solution design for the New Data Platform. Technology Architecture document. Making the right tool choices
2. Logical & physical design of the dimensional model for the New Data Platform
3. Design of the ETL for populating the New Data Platform, including data extraction, cleansing, conforming and integration, data delivery, and management
   1. Source to Target mappings
   2. Strategies for Dimension Management
   3. Drill down by target table
   4. One time data migration process
   5. Incremental load process
4. Future Adaptation strategies for the New Data Platform

## Timelines

We expect the consulting exercise to be complete in 8 weeks

## Pricing

LatentView proposes a total cost of USD 45,000 for this consulting exercise