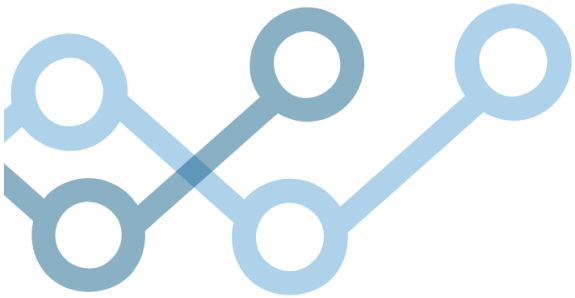
sophinea

"BRINGING CLARITY TO YOUR DATA"

STORM KING ANALYTICS

New York State Insurance Fund (NYSIF)

Analytics, Business Intelligence and Data Science Solutions
RFI 2023-58-IT



Prepared by:

Storm King Analytics
Dan Evans
Founder/CEO
dan.evans@stormkinganalytics.com
&
Sophinea Corporation
Brian Thamm
Founder/CEO
bthamm@sophinea.io

Introduction

Storm King Analytics, a Service Disabled, Veteran Owned Small Business (SDVOSB) and Sophinea Corporation (Small Business) are pleased for the opportunity to respond to the subject RFI in collaboration with Databricks.

Statement of Understanding

The NYSIF Data Warehouse & Operational Data Store currently has 22 TB of data with a roughly 10% increase each year in size. NYSIF's objective is to develop a solution that is:

1. Easy to manage
2. Secure
3. Provides a robust tool set to be able to perform simple and advanced data preparation and analysis
4. Provides the capability to serve as a data store for various Business Intelligence (BI) tools
5. Is cost effective

Key Considerations

Storm King Analytics and Sophinea have supported numerous, mission critical analytics projects using highly sensitive and regulated data. These projects include supporting graph analysis of networks in support of the Department of Defense (DoD), leading the deployment of the first global BI initiative at the Department of State (DoS) with the deployment of Tableau Server on the United States Refugee Assistance Program (USRAP), and supporting the automated reconciliation and analysis of hundreds of thousands of financial transactions on the DoS Visa program. All of these programs require scalable, reliable, and cost effective solutions to analytics.

Through our experience, we understand the challenges NYSIF is facing. There are countless initiatives ongoing at any point in time, many departments have various tool sets and experience that has been acquired over the years, and there is a need to enable users to answer questions expediently while also ensuring the optimal use of scarce data science expertise.

From this experience, there are various key attributes of a data science platform.

1. It needs to operate on open standards
2. It needs a strong set of data governance tools

3. It needs to operate as a platform
4. It needs the ability to operate as multi-cloud and on-prem

Open Standards

One of the biggest challenges early adopters of data analytics platforms have experienced is vendor lock in. From SAP, to Oracle, to Palatir, there are significant migration costs once the data is ingested and workflows are developed using their proprietary platform. Investing in proprietary platforms results in a long term relationship with the vendor and a reliance on their ability to keep up with rapid advancement on data analytics and confidence that the vendor's licensing costs do not become cost prohibitive. This can work, but we have seen too many examples of vendor obsolescence and licensing cost explosion hindering clients with their mission. Unlike proprietary platforms, the open source community continues to lead in the domain of data analytics. Indeed, Python, R, Scala, and SQL continue to be some of the most robust programming languages for data ingestion, preparation, and analytics. In addition, JDBC continues to provide the basis for flexibility when connecting to BI tools, such as TIBCO Spotfire, MS Excel, MS Power BI, Tableau, Qlik, Thoughtspot, and others.

Data Governance

Data Governance can be one of the most challenging aspects of building a data analytics platform. Data Governance can often be split into two domains: (1) Security and (2) Data Integrity. When choosing a platform, it is important to consider both domains. From the perspective of security, the data analytics platform should be able to enable administrators to provision fine grained access to the underlying data while also limiting the size and number of data abstracts required for analysts to perform analysis. One of the frequent conversations we have with clients is related to the reality that data becomes increasingly difficult to secure as it is downloaded from source systems. In addition, data integrity is incredibly important and provides the foundation for data trustworthiness. The ability to develop and maintain standards ensures the analysis performed has integrity. A common approach to this problem is the development of a medallion architecture. A medallion architecture is a data design pattern used to logically organize data in a data store, with the goal of incrementally and progressively improving the structure and quality of data as it flows through each layer of the architecture (from Bronze ⇒ Silver ⇒ Gold layer tables). This type of architecture is efficient and also promotes the use of a strong data foundation across all models.

DataOps

Operationalizing data analytics is a common reason for failed data projects. There are numerous breakdowns that can occur throughout the data process, to include, an inability to access data, compute resources to process batch and streaming data, and promote data products into production. Data Operations (DataOps), takes many of the proven principles of

DevOps and applies it to data analytics. DataOps is a collaborative and agile approach that integrates data engineering, data science, and operations to improve the efficiency, quality, and speed of data-related processes. It aims to streamline the end-to-end data lifecycle, from data acquisition and integration to analysis and deployment. Using a DataOps model removes many of the bottlenecks that block and hinder the delivery of data analytics products. Thus, selecting a platform that enables DataOps can support higher success with data analytics initiatives.

Deployment Flexibility

Deployment flexibility is important for a few reasons. First, clients may currently be operating on legacy, on-prem infrastructure but want a future-proof solution that enables a seamless migration to the cloud. The other reason is that organizations are more commonly operating between cloud and on-prem workloads, and also perhaps in a multi-cloud environment. As such, an analytics platform should be able to operate in all of these environments as well.

Proposed Solution

Storm King Analytics and Sophinea would propose the use of Databricks as a solution to addressing NYSIF's data analytics requirements. This solution addresses NYSIF's requirements for developing a scalable, flexible, and secure data platform that can operate across multiple infrastructures. As a Databricks partner with Databricks certified staff, our team is well positioned to support NYSIF with the development of world class data analytics capabilities.

Cost Effective Data Storage

Databricks Lakehouse provides a cost effective means to store data that uses an open source, non-proprietary Delta Lake to ensure atomicity, consistency, isolation and durability (ACID) quality guarantees on the data stored while also using object storage such as AWS S3 to drastically reduce cost.

Data Analytics SDLC

Databricks enables the use of DataOps to enable the effective promotion of data analytics products through sandbox/development, test/certification, and into production. All work products can be managed through the use of code and scripts, to include the creation of compute resources using YAML files to medallion data processing frameworks. This enables the effective use of software development and DevOps best practices in combination with git source control to reduce common manual errors of promotion.

Performance and scalability

Databricks utilizes the infinite scalability and performance capabilities of the cloud, enabling NYSIF as much (or as little) resources necessary to perform a wide range of data analytics tasks.

Accessibility and security requirements

Databricks meets the highest security standards. It can be accessed using a SaaS model while meeting Department of Defense (DoD) Impact Level 5 (IL5), SOC2, NIST, and PCI requirements. Databricks can also be installed within the existing security boundaries of NYSIF's cloud infrastructure.

Real Time or near-Real Time data collection

Databricks supports realtime and near-real time data collection and processing tools such as Apache Kafka, RabbitMQ, and Apache Spark Structured Streaming.

Data Preparation/ingestion (ETL, ELT)

Databricks supports popular data ingestion tools such as Fivetran, Airbyte, and Matillion and popular Python and R libraries to ingest data. In addition, Databricks supports tools such as Alteryx, python pandas, and R tidyverse to manipulate and prepare data.

Support For Structured, unstructured, semi structured data

Databricks Lakehouse enables the best of all worlds as it relates to data storage. As the Lakehouse utilizes object storage, NYSIF would be able to analyze structured, unstructured, semi structured data without needing multiple data storage tools.

Data Exploration and Visualization

Using Databricks JDBC connectors, NYSIF would be able to easily connect to a wide range of data analysis tools, to include TIBCO Spotfire, MS Excel, MS Power BI, Tableau, among others.

User Friendly Interface

Databricks' AutoML provides an easy-to-use interface to manage machine learning models, to include the ability to perform:

- Feature engineering
- Model creation and training
- Model testing

- Deployment
- Monitoring
- Maintenance
- Data Model governance
- Business value tracking

In addition, Databricks AutoML seamlessly integrates with the opensource MLFlow project.

Explainable AI (XAI)

Databricks proves a robust set of tools to track data lineage from source data through all steps of data processing in support of Artificial Intelligence (AI) use cases. When Databricks is used in conjunction with explainable AI techniques, it allows data scientists and machine learning practitioners to develop, analyze, and interpret machine learning models in a more transparent and interpretable manner.

Advanced Analytics

Databricks supports all popular open-source python and R advanced analytics libraries, to include Pandas, scikit-learn, TensorFlow, PyTorch, caret, and randomForest. These libraries enable a wide range of advanced analytics capabilities, to include:

- Descriptive Analytics (“What”):
    - Data mining
    - Cluster analysis
    - Time series analysis
- Diagnostic Analytics (“Why”)
    - Cohort analysis
    - Retention analysis
    - Regression analysis
    - Sentiment analysis
- Predictive Analytics (“If”)
    - Machine learning
    - Predictive modeling
- Prescriptive Analytics (“How”)
    - Complex event analysis
    - Machine learning
    - Artificial Intelligence

Delivery of Analytics

Databricks and their integrated notebooks provide a user-friendly interface to explore data. Databricks further allows the seamless operationalization of these notebooks into testing and production environments.

Collaboration

Databricks provides the ability to collaborate with other analysts and end users in near real time. This capability enables faster iteration on analysis and faster time to insight.

Data Coherence

Using a medallion data architecture in combination with Databricks leads to greater coherence as data models can be shared and built upon to address various business questions.

Multiple Personas

Databricks enables the development of fine grained controls based on the user's role. This includes, but is not limited to controlling and provisioning read, editing, and run permissions on notebooks and also administering the provisioning of compute resources and granting access to the platform.

Multiple Programming Languages

Databricks natively supports Python, R, SQL, and Scala.

Semantic Layer Support

Databricks enables the development of a semantic layer on the Lakehouse. In addition, popular 3rd party tools such as dbt and Atscale seamlessly integrate with Databricks.

Metadata Management/Data Lineage

Databricks Unity Catalog and Delta Live tables enable metadata management and data lineage. 3rd party tools, such as Alation, Informatica, and Collibra also integrate seamlessly with Databricks.

Business Data Dictionary

Databricks Unity Catalog enables the development and maintenance of a data dictionary. 3rd party tools, such as Alation, Informatica, and Collibra also integrate seamlessly with Databricks.

Data Governance

The use of Databricks as a single platform and source of truth consolidates data governance activities to a single platform.

Data Access

Databricks is capable of easily connecting to a wide array of data sources. In addition, integration of tools such as Immuta enable fine grained internal access policies on internal datasets.

Security

Databricks enables the use of multi-factor authentication for users. All access is logged and easily accessible through query using the Databricks platform. Source control using git repos is an out of the box feature of Databricks as is support for Active Directory. Databricks supports data encrypted at rest and in transit.

Cost control, monitoring, auditing

All costs can be accessed on demand by an administrator. Groups and tagging can also be used to track costs by attributes such as projects, roles, departments, etc.

On Prem or Cloud implementation

Databricks supports SaaS and Cloud hosting within NYSIF’s cloud environment.