TREXIN INSIGHT PAPER

February 13, 2024

DATABRICKS AT TREXIN

Leveraging Databricks to overcome data challenges.

TREXIN'S PERSPECTIVE

Together, our team at Trexin has found that there is one inevitable truth about data, it is messy. Data integrity poses a significant challenge as it encompasses issues such as misspellings, inaccuracies in numerical values, missing data, poor-quality data, discrepancies in data identified as the "source of truth," and retention of data deemed necessary but underutilized. Data is very tedious as it must be accurate, cleansed, reliable, trustworthy, dependable, valid, and definitive. If the data is incorrect, it affects not just potentially one system, it has repercussions across multiple systems, influencing downstream usage, user interaction, web interfaces, daily loads, and ad-hoc reporting.

Founded in 2013, Databricks supplies a solution, driven with a Spark engine, designed to assist organizations in leveraging their data by bringing together different departments and all types of users to work in the same space. To celebrate our recent partnership with Databricks, we have decided to craft a three-part series of Trexin Insight Papers (TIPs) regarding Databricks at Trexin. In this first paper, we will discuss how Databricks provides an extensive platform to drive innovation, improve decision-making, and enhance their competitiveness in their market space. To learn more about Trexin's experience with Databricks and understand how Databricks employs a Lakehouse strategy, blending the advantages of both a data lake and a data warehouse, keep an eye out for the upcoming TIPs in this series.

CHALLENGES WITH DATA

In the past, accomplishing your goals while maintaining a clean data set and a source of truth for your data involved relying upon multiple techniques employing many applications. Storage of your data was spread across multiple servers, which were redundant backups of storage in the event of a breach, natural disaster, or bad programming. Programming languages have evolved to the point where a user can accomplish the same task using SQL, R, Python, Java, Bash, or PySpark. Scheduling tasks and pipelines has become a guessing game and an organization's preference for a given tool.

One of the biggest challenges of today is navigating how goals can be achieved. An even greater challenge is bringing all the resources an organization has together on the same task. Some users work in SQL and work directly from the database, others work from a server, or secondary IDE, accessing the database through initialization files and providing user credentials. Other resources are used for testing, performing pipeline scheduling, or setting up access/pre and post validation reports. This can create conflicts when communicating about a project's data. One group just needs the database and table location while the other methods require access and permissions through an IDE. Both groups will access the data, but communicate their interactions differently.

USING DATABRICKS TO OVERCOME CHALLENGES WITH DATA

How do we bring all these resources together? Using Databricks and the cloud! Databricks allows users of multiple programming languages to accomplish their goals. SQL, BASH, Scala, and Python scripts can all be run directly from the simple user interface Databricks provides. Not sure about computer resources needs? Databricks allows for users to spin up clusters using Spark and set their size for use. Whether it's a very small job, reading HL7 data, or a huge batch job reading 1 billion rows, Databricks offers the user the resources needed to accomplish their goals, while maintaining cost-effectiveness.

Is your organization leaning heavily into AI and machine learning? Databricks has specific modules available to help users create, maintain, and deploy models easily and quickly. Enjoying your time confirming your scheduler can access the code and databases while running the scripts and not fail? Databricks' built-in scheduling feature allows users to monitor scheduled jobs in real time. Maybe your goals are reporting? Databricks has a reporting module that can be used to create many reports organizations employ today.

What about storage? Databricks connects directly to your cloud network, instantly providing a single source of truth location for the data and users. Code repos? Databricks can connect directly to your preferred code repository and users can maintain code directly from Databricks. Provided here are just a few examples among the wide array of use cases that Databricks offers.

- Data Warehousing
- Data Processing / ETL Pipelines
- Data Analytics:
    - Real-Time Analytics
    - Predictive Analytics
    - Customer Analytics
- Optimization of Business Process Flows
- Fraud Detection and Prevention
- Organizational Market Areas:
    - Healthcare Data Analysis
    - Financial
    - E-commerce
    - Sports
    - Supply Chain Optimization
- Sentiment Analysis
- Risk Assessment and Mitigation

Databricks is a transformative platform that brings numerous benefits to organizations. By providing a unified, secure, and powerful data analytics environment, Databricks empowers organizations to leverage their data assets fully, make informed decisions, and gain a competitive edge in today's data-driven landscape.

Trexin boasts extensive expertise in Databricks workspaces, offering a comprehensive solution for all data-related needs. For anyone looking to overcome data challenges with Databricks, our seasoned Advisors are ready to provide detailed insights and support.


This TIP was written by Michael Litwin and Mia Sabin. Michael and Mia welcome comments and discussion on this topic and can be reached at michael.litwin@trexin.com and mia.sabin@trexin.com.