Blog

3 FAILURES OF THE MODERN DATA SCIENCE PLATFORMS

Submitted by:

John Santaferraro, Ferraro Consulting

purplecube.ai

April 15, 2023

Failures of the Modern **Data Science Platforms**

1 The Need for Modern Data Science Platforms

It has been 40 years since the inception of data science platforms with only a few surviving the explosion of data in both volume and variety. The popularization of digital engagement, SaaS, and cloud has rendered legacy platforms insufficient and opened the door for mass modernization.

2 The Success of Modern Data Science Platforms

In the last 10 years, we have seen a wave of modern data science platforms built specifically for new digital data types, unidirectional flow of data, and simplicity of data science of operations. Technology leaders are already using these platforms to accelerate and expand the use of machine learning in business processes. In turn, many organizations are already experiencing intelligent automation and continuous optimization.

3 The Failures of Modern Data Science Platforms

Along with growing success, there has been a frustration among data science professionals regarding insufficient data acquisition and preparation, along with the lack of end-to-end data orchestration. Most data scientists are still required to use multiple tools or rely on other parts of the data organization to operationalize data science insight. There are three failures of most modern data science platforms.

FAILURE NUMBER ONE: **Insufficient data acquisition**

First, modern data science platforms have focused on the simplification of MLOps by providing basic data acquisition capabilities in their platform. However, because their focus is more on MLOps, modern offerings struggle with acquiring all types of data at all latencies. In addition, most of the modern platforms have ignored the importance of rich, unified metadata to support data governance and to increase code reuse in current expansion and future migration. Unified Data Orchestration is designed to acquire many different types of data across the full spectrum FAILURE NUMBER ONE: of streaming data, data at rest, and APIs. Modern orchestration also includes a richer set of acquisition capabilities including change data capture for streaming and settled data, as well as, high performance ingestion to avoid bottlenecks.

FAILURE NUMBER TWO: **Insufficient data preparation**

Second, modern data science platforms have focused on the simplification of MLOps by providing minimal data preparation capabilities in their platform. However, because modern data science focuses on modern data, they tend to lack adequate data preparation capabilities that span all enterprise needs for data cleansing, transformation, and integration. Similar to acquisition failures, they also lack metadata capture and automation sufficient for the active use of metadata. Unified Data Orchestration is metadata centric, automating the capture of metadata and storing it for active use in automations, recommendations, governance, and data services. In addition, modern orchestration includes the ability to collaborate on data pipelines and reuse high quality work in similar use cases.

3.

FAILURE NUMBER THREE: **Insufficient data orchestration**

Third, modern data science platforms have focused on the simplification of MLOps by providing light orchestration capabilities in their platform, but they have completely missed the importance of unified data orchestration. Most have strength in only one or two of the following segments: structured data, semi-structured data, streaming data, historical data, data integration, data preparation, or data delivery. Unified data orchestration provides end to end orchestration for all data, at all latencies, for all analytical use cases, for all users, in all locations globally. In addition, modern orchestration includes the ability to optimize and distribute workloads to the platforms that best process specific workload types. This is entirely missing from most data science platforms.

Unified Data Orchestration for Data Science

Unified Data Orchestration gives data scientists a consistent means of data preparation, model development, and insight operationalization. With end to end data pipelines in a single platform, data scientists can focus more of the time and effort on developing, testing, and deploying models. This gives their organization a competitive advantage by speeding innovation cycles and enabling new business models at rates faster than their competitors.

Check out how PurpleCube Unified Data Orchestration Cloud is empowering data scientists to single-handedly operationalize data science insight.

1390 Market Street, Suite 200, San Francisco, California 94102, US

www.purplecube.ai