Building the modern Data stack

Observations & Learnings from the trenches

Executive Summary

This blog captures our insights and learnings from using both Snowflake and Databricks in the field. We hope our analysis of current trends and our experiences in building highly scalable and complex data applications, will assist you in making the right selection for your specific Data & AI needs.

We have compared the key capabilities of **Databricks and Snowflake** across multiple areas, to enable one to look at both platforms holistically and not get enamored by “shiny features” that can seldom stand the test of time.

Introduction & background

Enterprises today are embracing the cloud to modernize their Data & Analytics stack. The velocity with which enterprises are migrating their legacy data-intensive applications has skyrocketed due to huge benefits offered by cloud providers, such as elastic computing, performance at scale, and highly secure infrastructure. Resources with a plethora of features, based on open standards and frameworks have helped in developing large developer communities that have the skills and competencies to build mature applications. With an insatiable appetite for data, business users continue to invest in self-service applications built on these open standards that are hugely strategic to their business. At the same time, IT-driven applications built by developers who are fiercely loyal to these open standards have further reduced the barriers to cloud adoption. Java, Python, and SQL/No SQL-based workflows are increasingly common in enterprises of all sizes. The evolution of several open-source frameworks such as Apache Spark has enabled data applications whose upper limits of scalability and data processing have not been reached yet. However, this has led to the proliferation of a fragmented set of tools and frameworks, creating maintenance nightmares for developers and platform engineering teams

Fast forward to current times and, it brings us to two heavyweights in the big data ecosystem, **Databricks, and Snowflake,** who are working tirelessly to simplify the development and deployment process while empowering applications to process gigabytes to petabytes of data on demand.

Data Warehouse, Data Lake & the Lakehouse for modern Data-intensive applications

With the rise of cloud computing, machine learning, real-time analytics, and self-service analytics over the last decade, the data and analytics landscape has seen a tectonic shift. Organizations big and small have embarked on migrating existing legacy on-prem data warehousing solutions to more open, scalable, and elastic data lake solutions. The acceleration is tilting toward open data formats due to the ubiquity of Spark ecosystems and the adoption of Spark by the developer community.

Databricks, the original creators of the Spark Engine, and the Delta format has been the major catalyst driving these transformations. These key innovations have enabled enterprises to overcome the shortcomings of the relational database's inability to handle unstructured and semi-structured data, which is critical for processing large amounts of data for Machine Learning, Visualization, and Predictive modeling workloads. Enterprises are prioritizing digital transformation initiatives to meet the needs of their customers for simplicity, quicker time to solutions, and improved performance to make informed decisions based on faster data insights. To handle high throughput and low latency applications, elastic, powerful data processing and compute capabilities are required to support real-time streaming events, perform predictive modeling, and allow scalable visualization solutions.

Data management is a key focus area in organizations to create a consistent and trusted single source of truth for all Analytics (Descriptive, Diagnostic, Predictive, and Prescriptive) requirements. Centralized data lake solutions solve such complex data management needs, especially for large enterprises that span the globe with complex and conflicting security and compliance requirements. For these global organizations, quick access to systems of records are key while ensuring customer data is stored securely and in compliance with local regulations.

Organizations are migrating to cloud-based solutions and moving towards managed, serverless services to reduce infrastructure and operational overheads, allowing engineering teams to focus on simplifying code development and maintenance, and shift their efforts to delivering business value. Databricks and Snowflake are currently the two major vendors addressing these needs and enabling enterprises to deploy to a multi-cloud, scalable, reliable, elastic, managed, secure, and governed platform to rapidly unlock value from data.

While both Databricks and Snowflake are trying to provide an easy migration path to their platforms, selecting the right vendor is dependent on the organization's priorities. Recent trends very strongly indicate that the future is moving toward multi-cloud, managed, lakehouse platforms. Lakehouses empower one to make data-driven decisions faster using AI/ML and real-time analytics. Databricks, a SaaS-based platform, built on Spark, is better suited for organizations that require end-to-end data and analytics capabilities such as data ingestion, curation, governance, feature engineering, data warehousing, machine learning, and near real-time streaming. Furthermore, the AI/ML landscape is rapidly expanding to make automated decisions based on consumer demands. The Databricks machine learning platform accelerates the development, experimentation, and deployment of machine learning models. Snowflake, on the other hand, is a cloud-agnostic, multi-cluster, shared data platform, ideal for data warehousing applications such as business intelligence for enterprise-wide visualization and reporting.

Warehouse or Lakehouse?

Databricks and Snowflake, both provide solutions to some of the most pressing challenges in the data, analytics, artificial intelligence (AI), and machine learning (ML) landscapes. We outline some of our observations comparing and contrasting the features of both tools that are important to enterprises looking to invest in their next data platform. We have selected the following key dimensions for our analysis:

- **Architectural capabilities for building complex data applications**

- **Key features for very specific use cases**
- **Scalability & Reliability**
- **Security & Compliance**
- **Data protection & Performance**
- **Data analytics use cases & pricing model**

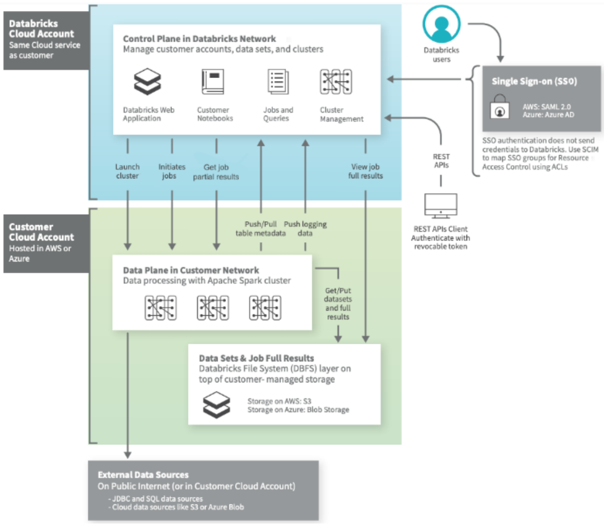
There are major differences in how each solution addresses specific requirements. The two most common strategies used by the organization to store and process data are either data lakes or data warehouses. Data Lakes are used for storing unstructured, semi-structured, and structured data as well as processing large volumes of data suitable for machine learning workloads (key strength of Databricks). A Data Warehouse is used to store structured data to support all BI reporting workloads (key strength of Snowflake). Whereas a Data lakehouse (pioneered by Databricks) is a new paradigm that combines the features of both a data lake and a data warehouse, making it suitable for machine learning, business intelligence (BI), and real-time analytics workloads.

Architectural capabilities for building complex data applications

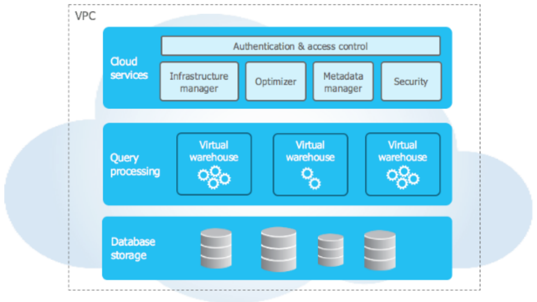
	Databricks	**Snowflake**
While both platforms enable the building of robust data applications, we assess that Databricks is the clear leader when it comes to supporting complex data sources, supporting a variety of analytics capabilities, including the ability to process *fast* data	★★★★☆	★★★☆☆

Databricks	Snowflake
The "Lakehouse" architecture unifies data lake and data warehouse capabilities	A hybrid of traditional shared disk and shared nothing database architecture
Separation of compute and storage. Distributed in-memory cluster computing	Multi cluster shared data architecture
Cloud agnostic supports major cloud providers - AWS, Google, Azure	Cloud agnostic supports major cloud providers- AWS, Google, Azure
Built on open-source technologies- Apache Spark, MLFlow, Delta Lake, Koalas, Delta sharing	Closed platform and proprietary technology
Supports multiple languages- Python, R, SQL, Scala, Java	Supports mainly SQL and Python language using Snowpark
Scalable platform for real-time streaming, Business intelligence, machine learning workloads	Scalable platform for BI reporting and ad-hoc analytics
Databricks platform consists of a Control and Data plane. The control plane is hosted on the Databricks cloud, supporting web applications, job scheduling, security, and Notebooks/Repos. The data plane is managed by the customer for data processing and data storage	Cloud services, Query processing, and database storage are the three layers in the snowflake architecture. Snowflake is a Software as a Service (SaaS) fully managed platform.
Storage layer - Delta Lake is optimized to compress and store columnar data. Optimizations	Storage layer – Data is stored as micro-partitions and supports clustering keys

using ZOrder, Adaptive query execution are available	
Medallion architecture is implemented using managed, serverless solution using Delta Live tables	One has to build data pipelines using Snow pipe and Snow SQL features.


Databricks Cloud Account
Same Cloud service as customer
Control Plane in Databricks Network
Manage customer accounts, data sets, and clusters
Databricks Web Application
Customer Notebooks
Jobs and Queries
Cluster Management
Databricks users
Single Sign-on (SSO)
AWS: SAML 2.0
Azure: Azure AD
SSO authentication does not send credentials to Databricks. Use SCIM to map SSO groups for Resource Access Control using ACLs
REST APIs
REST APIs Client Authenticate with revocable token
Launch cluster
Initiates jobs
Get job partial results
View job full results
Push/Pull table metadata
Push logging data
Customer Cloud Account
Hosted in AWS or Azure
Data Plane in Customer Network
Data processing with Apache Spark cluster
Get/Put datasets and full results
Data Sets & Job Full Results
Databricks File System (DBFS) layer on top of customer- managed storage
Storage on AWS: S3
Storage on Azure: Blob Storage
External Data Sources
On Public Internet (or in Customer Cloud Account)
- JDBC and SQL data sources
- Cloud data sources like S3 or Azure Blob

Source: Databricks architecture



Source: Snowflake architecture

Key features for very specific use cases

	Databricks	Snowflake
Databricks, with its abilities to provide extensive and native support for both data engineering and ML workloads is the clear leader here. Organizations leveraging Databricks have been able to consolidate their technology stack and tools to a single platform for all of their Data & AI workloads	★★★★☆	★☆☆☆☆

Databricks	Snowflake
Supports real-time streaming of data – using structured streaming, Autoloader and Delta live table features	Supports limited real-time streaming using Snow Pipe and only for cloud files. Has a unistore feature for transactional workloads
Provides CLI, Web UI, notebook experience and connectors for data science, data engineers and data analysts	Provides CLI, Web UI, and connectors for data analyst and business users
Provides full end-to-end ML deployment - Feature store - Experimentation - Model Development - Model Training and Tuning	Limited support for Machine learning use cases. Provided thru partner integrations and Snowpark for data science.

- Model Deployment - Model Serving - Model Management – registry, artifacts	
Provides Data management capability using “Unity Catalog” to manage the data assets, data access, and data lineage	Provides inbuilt features for data access control. Limited support on data lineage
Delta sharing is available to share the data securely with different consumers. First, open protocol for secure data sharing.	Data sharing feature to share the data securely with other snowflake accounts. Snowflake features

Scalability & Reliability

	Databricks	Snowflake
Both platforms in our experience are highly scalable & reliable. In addition to supporting multiple cluster types, they both have serverless capabilities, that make the process of infrastructure provisioning and configuration, completely transparent to the end user.	★★★★☆	★★★★☆

Databricks	Snowflake
Autoscaling feature is available to scale out based on the data volume and performance needs. Auto suspend and Auto resume features are also available	Autoscaling is available to scale up to max 10 clusters. Auto suspend and Auto resume features are available
Provides different types of compute options Job cluster – Scheduling data pipelines All-purpose cluster – Concurrent users for interactive workloads and reporting	Multiple clusters can be created for different scenarios, data loading, data unloading, reporting, data analysis, etc.,
Databricks SQL provides different cluster types. Flexibility to choose the compute cluster node types based on the requirements (Compute, storage, memory intensive, general purpose)	Clusters are available in T-shirt sizing (XS, S, M, L, XL,2XL,3XL,4XL). No abilities to choose the node types in the cluster
Users access using Web UI, CLI, JDBC, ODBC, Python, SQL connectors	Well defined UI, CLI, SQL, and JDBC/ODBC connectors for accessing snowflake platform

Security and compliance

	Databricks	Snowflake
Highly comparable and very similar capabilities, providing common features such as Data Encryption at rest, Role-based Access Control with support for multi-factor authentication, row, column and masking features for managing PII data	★★★★☆	★★★★☆

Data protection and Performance

Databricks	Snowflake
★★★★☆	★☆☆☆☆

This is where Databricks really shines and leaves its closest competition miles behind. Its Delta storage format is completely open-sourced, does not require data to be copied or moved, and supports any data in any format. With its relentless focus on delivering high-performance capabilities, Databricks is now very very close to offering similar, if not better performance than traditional data warehouse platforms.

Databricks	Snowflake
Support for unstructured, semi-structured, and structured data types	Supports semi-structured, structured data types. Limited support for unstructured data
An open ecosystem with frameworks like Apache Spark, Delta Lake, Delta sharing, ML flow, Koalas, etc.,	Closed Ecosystem
No vendor lock-in	Vendor lock-in
Combination of managed and custom features with the flexibility to tune the configuration for improving the performance. Benchmark results	Serverless and fully managed. Provides limited capability to tune performance
Provides features like time travel and Clone (Deep clone and Shallow clone) features	Supports time travel, zero clone, and fail-safe modes
Capable of handling Complex transformations and data enrichments	More suitable for simple data transformations

Use cases and pricing

Databricks	Snowflake
★★★★☆	★☆☆☆☆

The biggest advantage of the Databricks pricing model is that it does not increase exponentially over time (like Snowflake's) with increasing data volumes and usage.

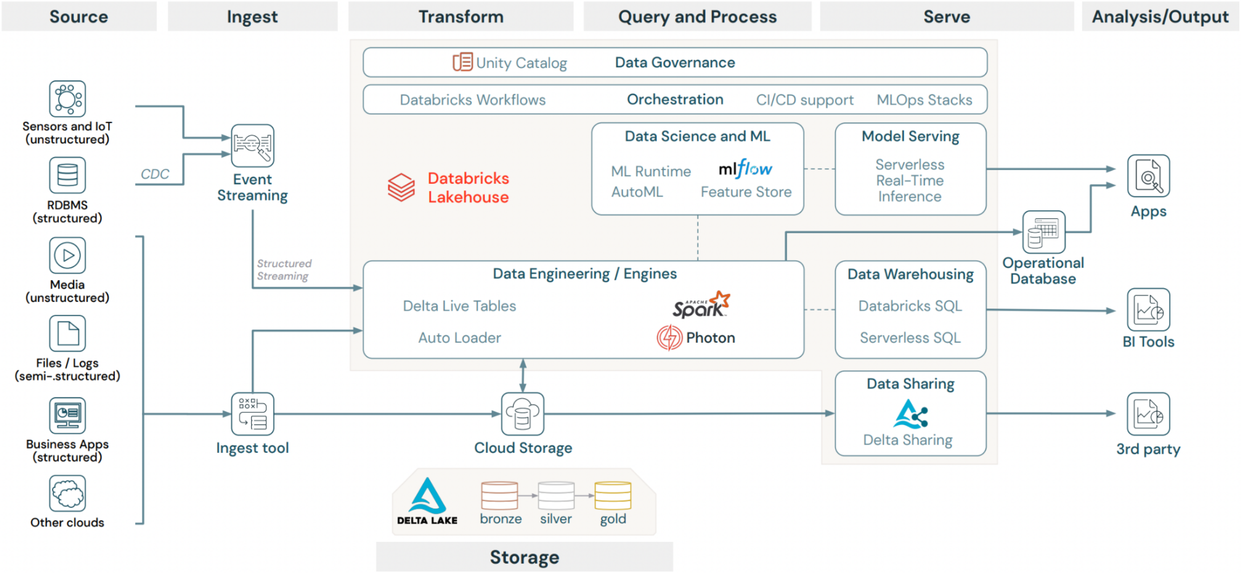
Databricks	Snowflake
Unified collaborative ecosystem for multiple use cases. - ELT/ETL - Machine learning - Real-time streaming - Data warehousing - Lakehouse - Partner integrations	Supports below use cases. - Data warehousing - ELT - Micro batch and batch-based workloads - BI analytics - Integration with partner tools

Compute cost is based on DBUs (Databricks processing units). Storage cost is separate. Databricks pricing	Per second billing for the compute based on the cluster types. Storage cost is separate. Snowflake pricing
Cost depends on the workloads and the data volume with the flexibility to optimize the cluster types using spot instances	Cost depends on the size of the virtual warehouse and the snowflake editions

Databricks is a feature-rich, developer-centric tool. It offers an incredibly flexible development and deployment platform that is not limited to just JDBC/ODBC connectivity but provides native support through Python, Java, and Scala language frameworks for solving today's most complex Data & AI challenges.

Based on our experience, Databricks Lakehouse is the more superior choice as the modern data platform. It is one platform, based on open standards, serving more users and use cases and solves today's most complex Data & AI challenges with a lower TCO.

Lakehouse Architecture:



Author: Sandeep Arabatti, Co-founder, Computomic
sandeep@computomic.com

About Computomic

Computomic helps companies realize value from data by modernizing their existing data stack to modern data platforms. We use solution and migration accelerators that save companies up to 80% of time and money when migrating from legacy data platforms to the modern data stack. We are a rapidly growing company with a global footprint focused on solving Big Data, Analytics, and Financial Crime Prevention challenges.