The Snowflake vs. Databricks breakdown

Which data platform fits best with the needs of your organization?

vs.

Two of the most dynamic and fastest growing companies in the big data world — Snowflake and Databricks, were built around innovative concepts.

Both companies offer expansive sets of consistently updated features within a unique design and architecture. Simply put, each platform stores data, ingests and transforms data, and produces analytics.

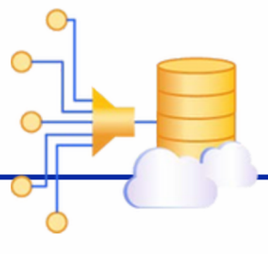
Within those main functions, Snowflake and Databricks have a range of capabilities that better fit the strategies of individual organizations.

Wavicle's expert cloud data consultants created a comprehensive guide for you to use as a comparison-based starting point for evaluating which platform better suits your needs.

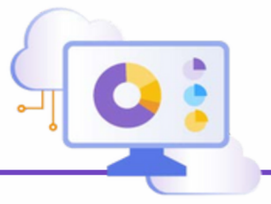
Snowflake vs. Databricks

1 Value & Architecture

2 Storage



3 Ingestion & transformation



4 Data analytics

5 Additional features

1 Platformvalues

Snowflake: The Data Lake

- One platform
- One copy of data
- Many workloads Near-zero
- maintenance Near-unlimited
- performance and scale
- Secured and governed access to data

Databricks: The Lakehouse Platform

- Unified Analytics platform
- Open format storage layer with Delta Lake
- Structured transactional layer
- High performance query engine
- One platform for every use case

Wavicle insights

Snowflake’s values ofspeed, scalability, and sharing are built throughout. For a rapidly expanding organization that needs to handleasignificant number of concurrent workloads and share data across multiple partners efficiently and securely,Snowflake is a strong choice. For Databricks, the foundation of data science is evident in the platform's value pillars. Databricks is suited for a wide-variety of machinelearning cases. Organizations focused on scalable data engineering, collaborative data science, and transforminglarge volumes of unstructured data should be intrigued by Databricks.

Architecture

Snowflake: A three-tier design

1.Centralized storage

2.Multi-cluster compute

3.Cloud services

- Optimization
- Management
- Transactions
- Security and governance
- Metadata
- Sharing and collaboration
-

Databricks:

Infrastructure and governance supported with Data Mesh, an organizational and architectural paradigm. It has an emphasis on decentralized data ownership.

- Decentralized data teams and ownership
- Data products driven by domain driven design
- Self serve data infrastructure
- Global federated governance

Data Mesh blueprint:

- Streaming phases
  - **Bronze:** Raw operational data
  - **Silver:** Curated atomic and cleansed data
  - **Gold:** Aggregated data (data marts) for specific datasets
- Data catalog
- Integration

Architecture features

Snowflake:

- Snowflake is available on AWS, Azure, or GCS
- Data stored in Snowflake storage
- Data can be accessed from S3, Azure Blob Storage or GCS
- Data loaded to Snowflake is indexed and partitioned during ingestion
- De-coupled compute and storage
- Virtual warehouses (VWH) can be instantly scaled from SQL command line or web-based GUI. VWH can also be configured to auto scale.

Databricks:

- Databricks is a native component of Azure and is also available on AWS
- Delta Lake sits on top of your existing data lake, delivering reliability, security, and performance
- Accessed by SQL / ML layers
- Data can be accessed from Amazon S3, Azure Storage, or GCS

Wavicle insights

Both of the platformscan be spun up on AWS, Azure, and GCP platforms. Snowflake does not require any pre- planning or maintenance to start, eliminating the need for a database administrator in many cases. It automatically runs across three availability zones, allowing for replication to an alternate cloud. Fully elastic autoscaling, a hallmark feature of Snowflake, means increasing or decreasing the size of an instance can be completed easily. For creating a Databricks cluster, there's three different cluster modes: Standard, High Concurrency, and Single Node. For the user, deciding which cluster mode to use can be a challenge but is the key to managing cost and performance. Databricks also features autoscaling by leveraging reporting statistics to scale up, or, remove workers in the cluster. To use and maintain Databricks, users need to have some level of knowledge surrounding cloud infrastructure components and how they work together. Snowflake's architecture means a rapid rollout to start, with levels of automation. This makes it a great choice for an organization that may not have the initial bandwidth or expertise in the platform. The customizable options of clustering for Databricks are a very attractive feature but requires strong competency in the platform and users must choose between cost and performance during configuration.

Storage

2Data warehouses and data lakes

Snowflake Cloud Data Warehouse:

Snowflake's Cloud Data Warehouse is SaaS-based and built on top of Amazon Redshift or Microsoft Azure cloud infrastructure. Users do not need to install, configure, or manage hardware or software. Storage, compute, and services are independently elastic and give users flexibility for what they need most.

With the introduction of Snowpark, Snowflake customers can write queries using proceduralprogramming languages.

Snowflake for Data Lakes:

The company totes its answer to data lakes as a flexible solution to enable or enhance data lake strategy. What does that mean for potential customers? A centralized repository for structured and ~~unstructured data~~ alike, with the latter functionality currently in Public Preview.

Databricks Delta Lake:

Databricks's answer to a data warehouse is well beyond the traditional model. The Delta Lake is an open format storage layer and a single home for structured, semi-structured, and unstructured data. It provides traditional warehouse features like he mas for each table. All data in Delta Lake is stored in open Apache Parquet format, allowing data to be read by any compatible reader.

Wavicle insights

Bothplatforms are leading the collision of traditional ~~data warehouses~~ and data lakes. The overlapping capabilities and names can become blurred. Organizations that don't have the time or resources for setup, maintenance, and support of servers should consider Snowflake.

If management of a data lake and data warehouse is an issue for an organization, Databricks can help solve the problem, along with its advanced analytics and AI/ML capabilities.

Access

Snowflake:

Democratized data access and simplified and controllable data governance are a hallmark feature of Snowflake. The flexibility and security policies are designed to boost innovation.

Snowflake's platform features designed into the accounts and users, column-level security, row access policies, audit logging for access history and object tagging for sensitive data for compliance, discovery, protection, and resource usage.

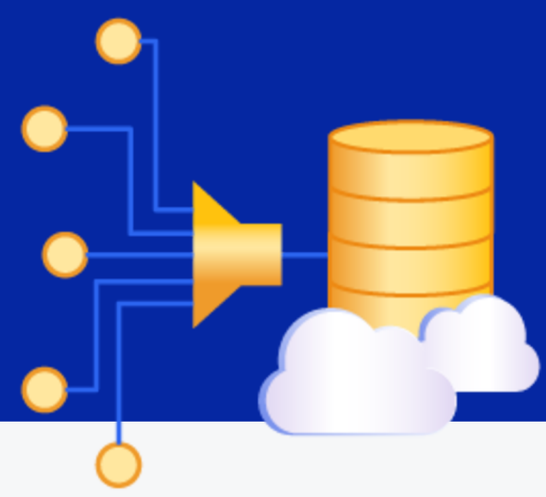
Databricks:

Databricks provides access control down to the storage layer by leveraging AWS security controls within the platform. At the same time, Databricks provides access control to compute resources, API provisioning and permission management, audit logging with Amazon Cloud Trail, and Amazon CloudWatch.

Wavicle insights

Snowflake's emphasis on democratized access and security are a big plus for the platform. However, that strength comes with a variable—difficult to manage operational governance and CPU cost. With easier control of compute resources, Databricks provides more transparent cost and relies on AWS for its security functions. If an organization needs day one access to sensitive data across various units at scale, Snowflake is a great choice. If more efficiently managed spend and familiar AWS features are appealing, Databricks can be quickly operationalized.

Ingestion & transformation



3Pipeline

Snowflake Data Engineering and Snowpipe:

Aswith most features of Snowflake, building the pipeline into the platform is about speed, efficiency, andease of use. With Snowflake, data engineers canspend little to no time managing workloads, making it scalable to handle concurrency and computation requirements. Snowflake's Snowpipe enables loading data from files as soon as they're available in a stage for event-based real time ingestion into the table.

- **Amazon:** Snowflake can ingest data from all AWS.
- **Microsoft:** Snowflake can ingest data directly from Azure storage.
- **Google:** Snowflake can ingest data directly from GCS.

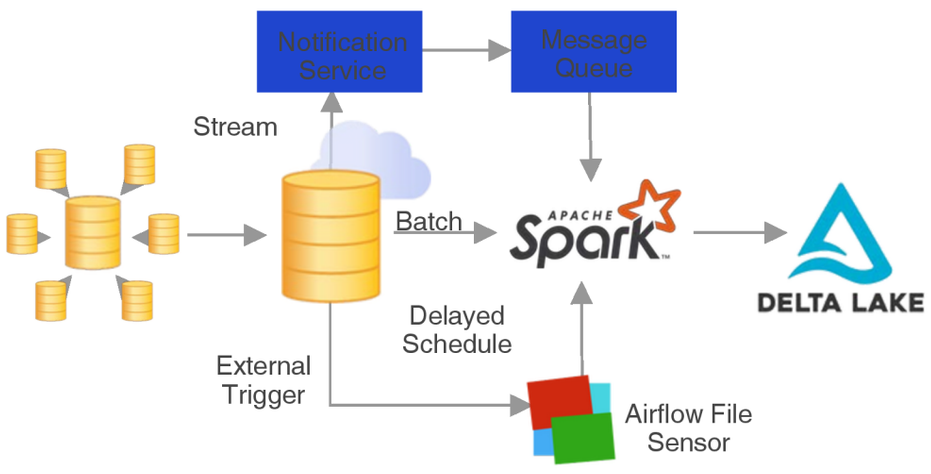
Databricks Autoloader:

An efficientsolutionthatincrementally processes new data files as they arrive into cloud storage.

- **Amazon:** Databricks can ingest data from all AWS.
- **Microsoft:** Databricks can ingest data directly from Azure storage.
- **Google:** Databricks can ingest data directly from GCS.

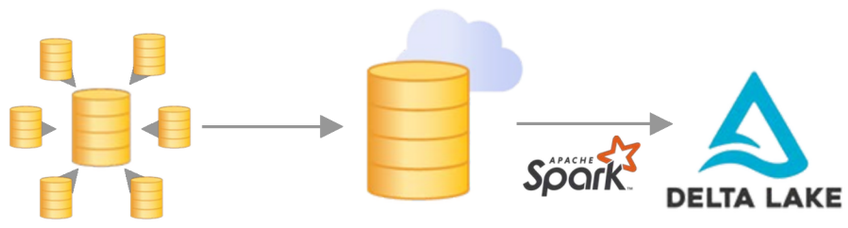
Databricks Autoloader

Before



- Gets too complicated for multiple jobs

After



- Pipe data from cloud storage into Delta Lake as it arrives
- "Set and forget" model eliminates complex setup

Pipeline integrations

Snowflake:

Snowflake features Snowpipe integrations from the following cloud storage services. Snowpipe loads data into Snowflake as soon as that data is available in the staging layer.

- Amazon Web Services: Amazon S3
- Google Cloud Storage
- Microsoft Azure Blob Storage
- Microsoft Data Lake Storage Gen2
- Microsoft General-purpose v2

Snowflake Account Host	Amazon Web Services	Google Cloud Platform	Microsoft Azure
Amazon S3	✓	—	—
Google Cloud Storage	✓	✓	—
Microsoft Azure Blob Storage	✓	—	✓
Microsoft Data Lake Storage Gen2	✓	—	✓
Microsoft Azure General-purpose v2	✓	—	✓

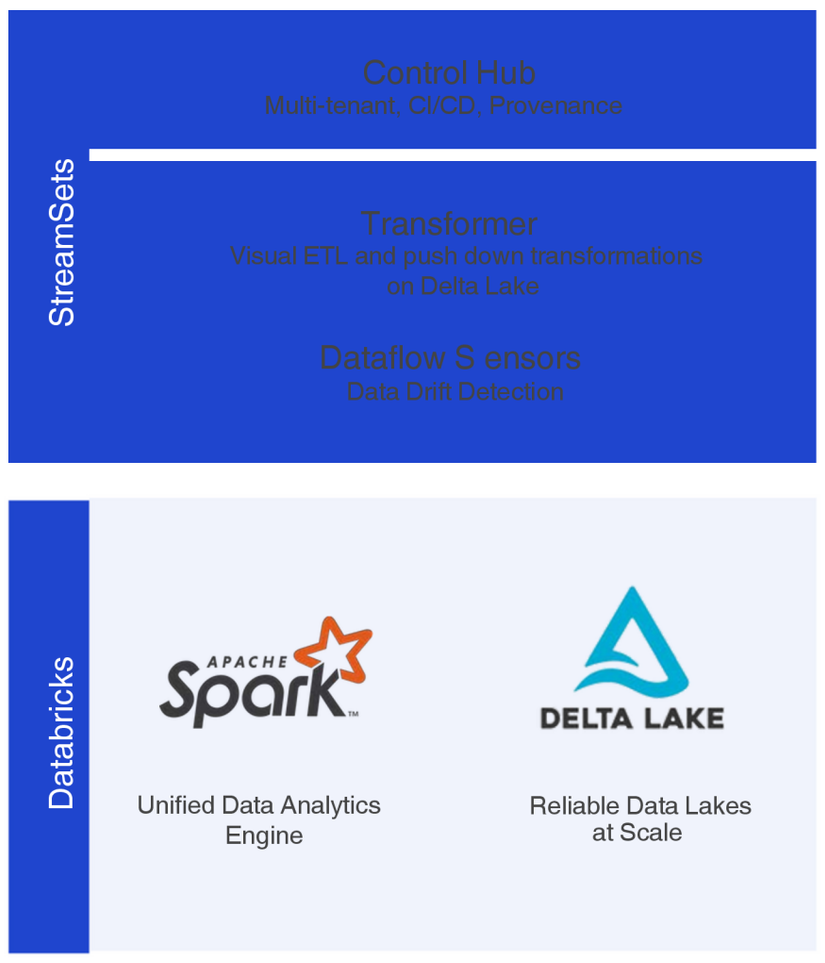
Supported cloud storage services The following table indicates the cloud storage service support for Snowpipe REST API calls from Snowflake accounts hosted on each cloud platform:

Snowflake Account Host	Amazon Web Services	Google Cloud Platform	Microsoft Azure
Amazon S3	✓	✓	✓
Google Cloud Storage	✓	✓	✓
Microsoft Azure Blob Storage	✓	✓	✓
Microsoft Azure General-purpose v2	✓	✓	✓
Microsoft Data Lake Storage Gen2	✓	✓	✓

Databricks:

Databricks automates streaming data ingestion and transformation with StreamSets. The partnership provides a fast and easy to use drag and drop interface. It allows users to design, test, and monitor batch and streaming ETL pipelines without the need for coding or specialized skills.

Databricks + StreamSets:



Wavicle insights

Bothplatforms are designed for fast, easy, and multiple sourced ingestion. ELT/ETL tools like Matillion, Talend, andSnap Logic can be used on both platforms to easily ingest and migrate data.

Themultitude of ingestion capabilities for both platforms means excellent flexibility for the major cloud providers. Customers of Amazon, Microsoft, or Google should be comfortable with either platform.

Performance

Snowflake:
Inhead-to-head comparisons conducted by independent companies and with minimal configurations or tuning, Snowflake out performed other cloud data warehouses on query time and related costs. This means Snowflake is almost a serverless solution.

Databricks:

Accordingto the ~~Transaction Processing Performance Council~~, Databricks SQL is now the record holder for data warehouse performance. For data scientists, the performance clusters of Databricks allow large-scale data batch processing and real-time stream data processing. The ability of its ML, deep learning, and graph analysis are exactly what you would expect from the founders of Apache Spark and MLflow.

Wavicle insights

Asboth platforms continue to improve at a rapid pace, performance will be a continued debate. While the tests mayshow contradicting results, they are impacted by use case, configuration of systems, code, and structure of underlying data.

Both platforms are top-of-class performers. For pure speed involving query time, Snowflake's near serverless solution continues as a standard of pure performance. The performance of large batch processing and ML for Databricks makes it the pinnacle of data science related performance.

Data sharing

Snowflake:

Anothero ne o f the pillars for the creation of Sownflake was data sharing. Data can be warehouses . This allows for sharing data

Secu red data sharing a cr oss Snowflake

Tables

- External tables
- Secure views
- Secure materialized views
- Secure user-defined functions

Databricks:
Databricks in its current form does not allow for cloning of data, only copying. With the introduction of Delta Sharing, Databricks users can share secured and real-time large datasets for sharing data cross products. This allows for sharing any data set in Delta Lake or Apache Parquet formats.

Wavicle insights

The exciting addition of the first version of Delta Sharing is a major upgrade in this category for Databricks, but it's still limited in its scope. The future plans call for sharing objects, such as streams, SQL views, or arbitrary files. Snowflake's cloning and wide-spread data sharing capabilities make it a great choice for organizations that need to share data with a wide variety of partners, vendors, or customers.

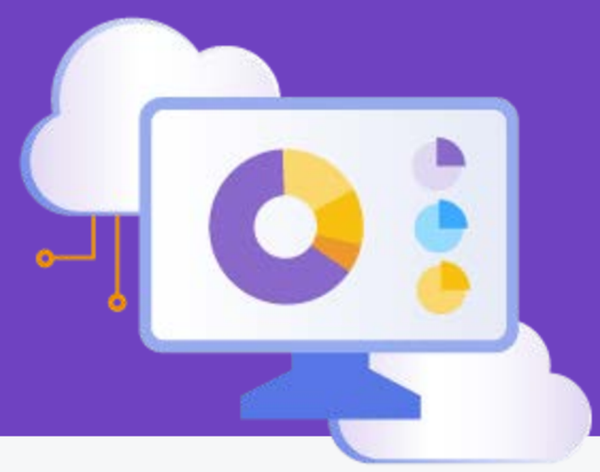
Data format

Snowflake:
Snowflake handles structured and unstructured data natively. Semi-structured JSON, Avro, ORC, Parquet, or XML can be loaded into a single field. The Query API allows parsing unstructured data at speed and scale.

Databricks:
Databricks default data format is Parquet and all data stored in the Delta Lake is stored in Parquet format. Databricks can read semi-structured data like JSON. By using the combination of Databricks & Labelbox, you can effectively handle unstructured data. With Sparser, Databricks users can rapidly parse unstructured data formats in Apache Spark.

Wavicle insights

Bothp latformsareableto handle a wide range of data formats. This really comes down to preference and experience.



4 BI and visualization

Snowflake:

Snowflake is compatible with several BI and ~~visualization~~ tools such as Tableau, PowerBI, and ThoughtSpot.

Databricks:

Databricks comes with built-in BI functionality but it is not the strongest feature. It's compatible with tools like Tableau and ThoughtSpot for analyzing data lakes at scale.

Wavicle insights

Bothplatformsintegratewell with leading BI and visualization tools. There isn't a distinct advantage for either, unless you need to handle significant numbers of concurrent users, then Snowflake would be a better choice.

AI/ML

Snowflake:

Snowflake is designed to support machine learning and in conjunction with tight integrations to Spark, R, Qubole, and Python. Snowflake performance means scaling up or down but it also takes on data curation responsibilities and reduced data-related burdens from ML tools.

In 2021, Snowflake introduced Snowpark, so developers can build queries using DataFrames right in their code, without having to create and pass along SQL strings.

Built-in:

- Spark
- Python
- Java
- Seashop
- Node.js

Integration is available with:

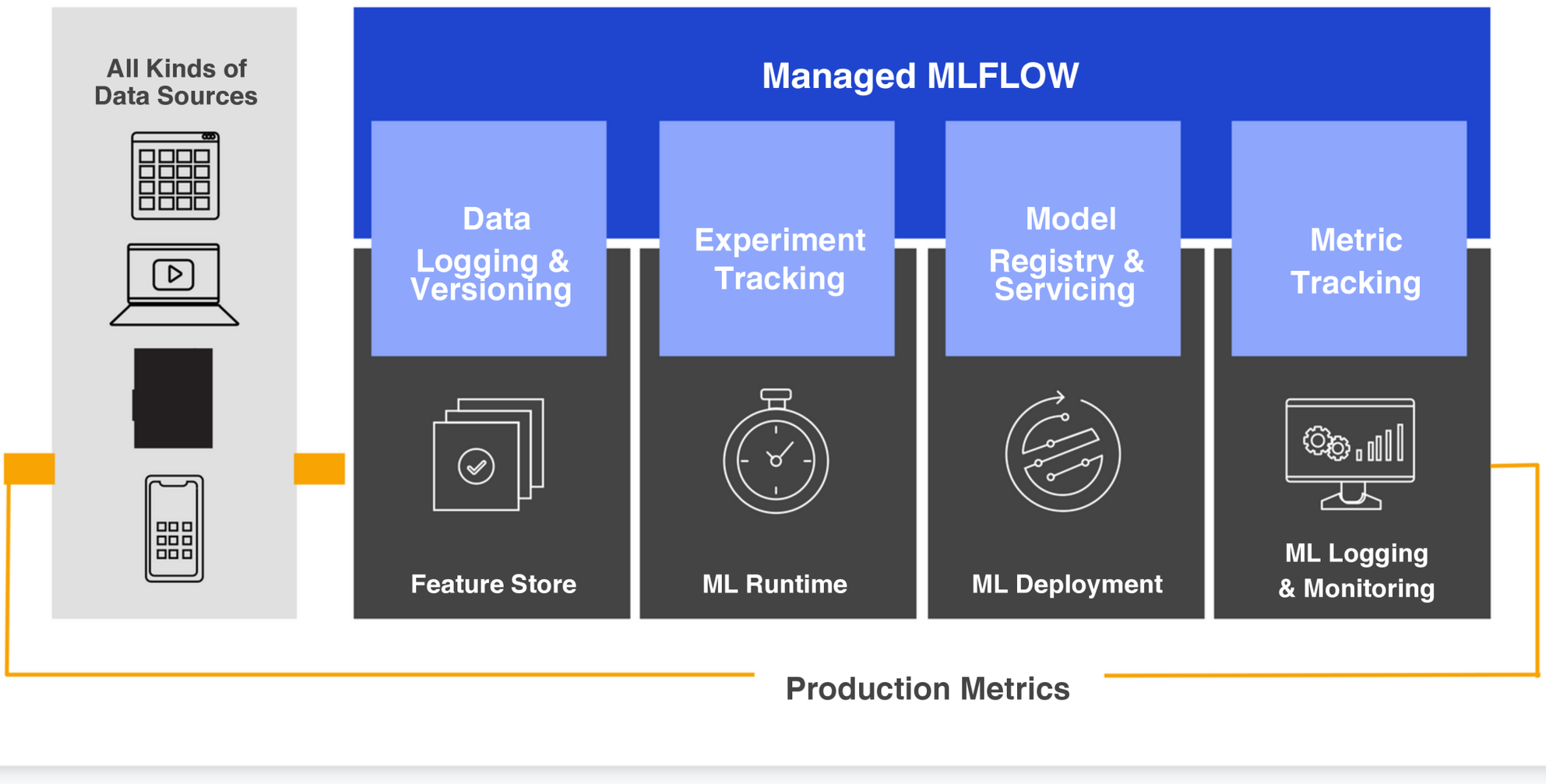
- Dataiku
- Data Robot
- Amazon Sagemaker

Databricks:

Built on top of MLflow, Managed MLflow, Databricks' open-source platform, manages the complete ML lifecycle, including experimentation, reproducibility, deployment, and a central model registry with enterprise reliability, security, and scale.

- Built-in Spark
- Managed MLFlow
- ML Runtime
- Collaborative Notebooks
- Feature Store
- AutoML

Data foundation for the full ML Lifecycle



Wavicle insights

Databricks was designed from its creation to be the most powerful, efficient, and collaborative environment for machine learning and that remains the truth. Even with the introduction of a model like Snowpark for additional developer languages, Databricks is still the premier platform for AI/ML. Organizations with a strong need for ML within their caseloads should look to Databricks or a combination of the two.

ML integration

Snowflake:

Snowflake can access code directly from Jupyter, Notebooks, or JAR files from within the platform.

Databricks:

For the more hands-on-the-keys crowd, Databricks has built-in ML functionality for Jupyter and Notebooks.

Wavicle insights

The built-in ML functionalities of Databricks makes it the most efficient and collaborative environment for developers with heavy use of ML.

UI

WithSnowflake's platform meant for a variety of end users, the UI is easier to navigate. As for Databricks, it is designed for ultimate function over form.

Scalability

Snowflake:

Storage, compute, and services are independently elastic. Users can spin up separate virtual warehouses instantly to support ETL, ELT, and BI workloads with no resource contention.

Databricks:

Users can enable clusters for auto-scaling based on workload with serverless pools to deal with concurrency.

Wavicle insights

For scalability, each platform has very distinct characteristics. As mentioned, the independent elasticity of Snowflake creates a top-of-class model for scalability and for organizations where it's a top priority, Snowflake is a strong choice.

Additional features

5 Snowflake: ▪

Time travel to query data from different points in time

- Clone and restore data from tables, schemas, or entire databases for a point in time
- Restore tables from a point in time or before updates were made
- Geo-spatial data for calculating distance is built into Snowflake

Databricks:

- Supports Python, Scala, R and, SQL OOB
- Optimized for machine and deep learning
- Manage a machine learning pipeline

Pricing and cost optimization

Snowflake:

- Usage based on a combination of time and compute
- Auto-scaling and increasing VM sizing during SQL processing can streamline costs

Databricks:

- Minimal users model – lower cost
- Enterprise level users – higher cost
- Auto-scaling configurations

Interoperability

Despite the differences, Snowflake and Databricks have a high-level of interoperability. Snowflake can read data from Databricks for analysis and visualization. Databricks fills the role of a connector that can read and process data within the platform and push results to Snowflake. In an ideal world, organizations across the board could utilize both platforms for their advantages.

Wavicle insights

Organizations across various industries utilize both platforms for their distinct advantages. This “best of both worlds” stack sets up data engineers and data scientists alike in fast, scalable, and collaborative environments. Wavicle has experience enacting this powerful stack simultaneously for clients.

The choice

Well, the truth is that it will take much more than a guide to determine which platform, Snowflake or Databricks, is the best fit for your organization. Many organizations leverage both platforms for their unique capabilities in a powerful stack. Each platform is a pathway for storing, ingesting, transforming, and analyzing data. Regardless of which way you are leaning, Wavicle can help you make the best possible choice based on your business strategy and goals. With our deep technical expertise and our proprietary accelerators, we migrate data quickly and integrate Snowflake, Databricks, or both into your technology stack. Are you looking to add Snowflake, Databricks, or both into your organization? Our expert cloud consultants bring proven experience with each to ensure you get the most out of the platforms.

Learn how