EUCLOID

databricks

Databricks:

Key Capabilities of a Modern & Open Data Platform

TARGET AUDIENCE

You should read this white-paper if you are:

- Migrating your data workloads to Cloud or from legacy platforms like Hadoop
- Using Spark and want a best-of-breed solution to automate its management
- Evaluating Databricks as a platform for your Data, Analytics & ML needs
- Exploring ways to reduce costs of your data infrastructure
- Looking to understand key capabilities of Databricks platform
- Looking for insights on how Databricks can address your specific use-cases like Governance

TABLE OF CONTENTS

Executive Summary

Increasingly large number of organizations today are composed of multiple and diverse data teams that need to operate in self-service mode so that they can effectively drive business growth. Concurrently data architectures also need to evolve in a manner that can enable these teams without causing platform sprawl or data swamps, while enabling flexibility, speed, and quality with proper governance capabilities in place. In this white paper we will look at Databricks platform and analyze its capabilities in detail with a lens to how it enables Modern Data Teams to get the best out of their data.

Challenges in Managing Big Data Today

Businesses today are generating increasingly vast amounts of data (structured and unstructured) and using it for downstream analytics & ML-driven projects. Due to vast amounts of data being captured and analyzed at increasing velocity, there is a growing need to store & manage this data effectively to address multiple challenges:

01 **Complexity & Costs**
Over time, data platforms become complex as multiple technologies are added incrementally over a period of time to address different requirements and processes. Furthermore, vast amount of data being stored in Data warehouses leads to spiraling costs in managing the data infrastructure, whether on-prem or on cloud.

02 **Data Quality**
As most data systems are built organically over time, this leads to inconsistencies in data due to multiple data formats, data standards, definitions and terminologies. These data quality issues result in sub-optimal downstream analytics and ML efforts.

03 **Data Access**
As use of data gets more widespread among all layers and functions in an organization, enabling access to data to multiple stakeholders becomes an effort-intensive and costly thus preventing businesses in getting the most value out of this valuable resource.

04 **Data Governance**
As data becomes more pervasive, there is a need to ensure availability, usability, integrity and security of data throughout its lifecycle. However, lack of clarity in ownership, inadequate policies in ensuring security of data can cause legal and financial liabilities.

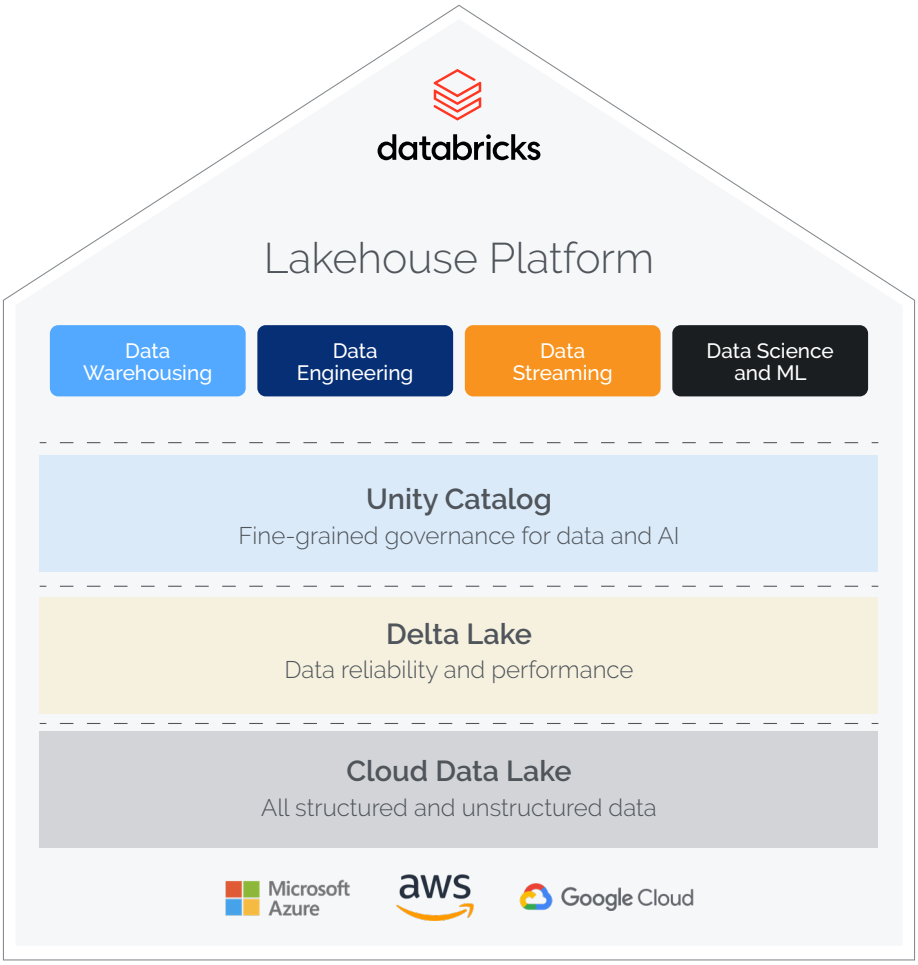
05 **Compliance**
Regulations like HIPAA and CCPA are forcing businesses to store customer data for longer periods of time. HIPAA for example requires healthcare providers to store Patient Health Information (PHI) for a minimum of six years. Furthermore, businesses need to make this data available on request either from regulatory authorities, stakeholders or their customers.

There is no silver bullet to address these challenges. However, we have seen with multiple customer implementations that a few data platforms are designed and built grounds-up with foundational elements that help you democratize data at scale while addressing cost, quality, governance and compliance needs in mind.

Databricks – a Modern Data Lakehouse Platform

Traditionally, data has been stored and managed using Data Lakes and/or Data Warehouses. Both these paradigms have their strengths and shortcomings. Data Lakes are a highly cost-effective and flexible means to ingest and store raw data but managing data quality and using this data for data analytics is time-consuming and costly. One the other hand, Data warehouses are ideal for Data Analytics, Machine Learning and Visualization capabilities but result in high complexity & costs due to management of data pipelines and data modeling.

Data Lakehouse is a new approach to data management that combines the best of scalability & flexibility of Data Lakes with quality & reliability of Data Warehouses. Databricks is a cloud-based data processing and analytics platform that was founded in 2013 by creators of Apache Spark – an open source distributed computing engine for processing large amounts of data. Databricks is built on top of Apache Spark and extends it into a Data Lakehouse platform through Delta Lake, Delta Live Tables, Unity Catalog & MLFlow.



With this architecture, Databricks addresses multiple challenges of a modern Data environment:

01 **Simplified Data management and processes:** Databricks is a unified analytics platform that provides data processing, data engineering, and data science capabilities in a single workspace through its notebook interface.

02 **Scalability & Efficiency:** Databricks is built on Apache Spark and hence inherits the ability to scale from a single machine to thousands of nodes seamlessly. This enables enterprises to process large amounts of data quickly and efficiently.

03 **Reduced Complexity & Costs:** Delta Lake works a data storage layer that sits on top of your cloud storage such as S3 or ADLS, hence reducing costs than a typical data warehouse. In addition, through Delta Live Tables ETL framework, users can create data pipelines with ease and ingest data from multiple sources like Kafka, Event Hubs, transactional data sources and others. Furthermore, Delta Live Tables are a serverless platform thus obviating the need for infrastructure management, resulting in reduced costs.

04 **Data Democratization & Collaboration:** Databricks provides a collaborative workspace for teams to collaborate through its notebook interface. Through this interface, teams to write and run code, visualize data, and share insights with others. Databricks notebooks support multiple languages including Python, R, SQL, and Scala.

05 **Data Sharing & Governance:** Unity catalog in Databricks is a centralized solution that makes it easy to manage access to data as well as understand Data Lineage and Data discoverability. In addition, Delta sharing protocol makes it very easy to share data with both internal and external stakeholders and enabling a layer of governance through Unity catalog.

06 **Security & Compliance:** Databricks provides enterprise-grade security, including encryption at rest and in transit, identity and access management, and compliance certifications such as SOC 2 Type II and HIPAA.

A large telecommunications conglomerate processes billions of transactions and terabytes of data everyday on Databricks.

Source: Databricks

In the following sections of this white paper, we will cover different capabilities of Databricks platform and how it addresses the challenges above in detail.

Delta Lake

Delta Lake is an open-source storage layer that is the foundation of Databricks platform. Delta lake runs on top of cloud storage such as Amazon S3, Hadoop Distributed File System (HDFS) or Azure Blob Storage. Delta Lake is fully compatible with Apache Spark APIs and stores data in Parquet format, a columnar storage format optimized for performance. Delta Lake offers the scalability of Data Lakes with the reliability of a data warehouse through following capabilities:

01 **ACID guarantees:** Delta Lake ensures that all changes are committed for durability and provides atomic transactionality. Hence Delta Lake ensures that there are no partial or corrupted data. Delta Lake also creates a transaction log that tracks all changes made to the data. This transaction log is stored in a separate directory and always ensures a correct view of data by serving as single source of truth.

02 **Schema Enforcement & Schema Evolution:** Any data inserted into Delta Lake goes through a check to ensure that it matches the Table Schema, thus preventing insertion of incorrect Data. In addition, Delta Lake allows the schema to be explicitly and safely evolved to meet new business or technical requirements.

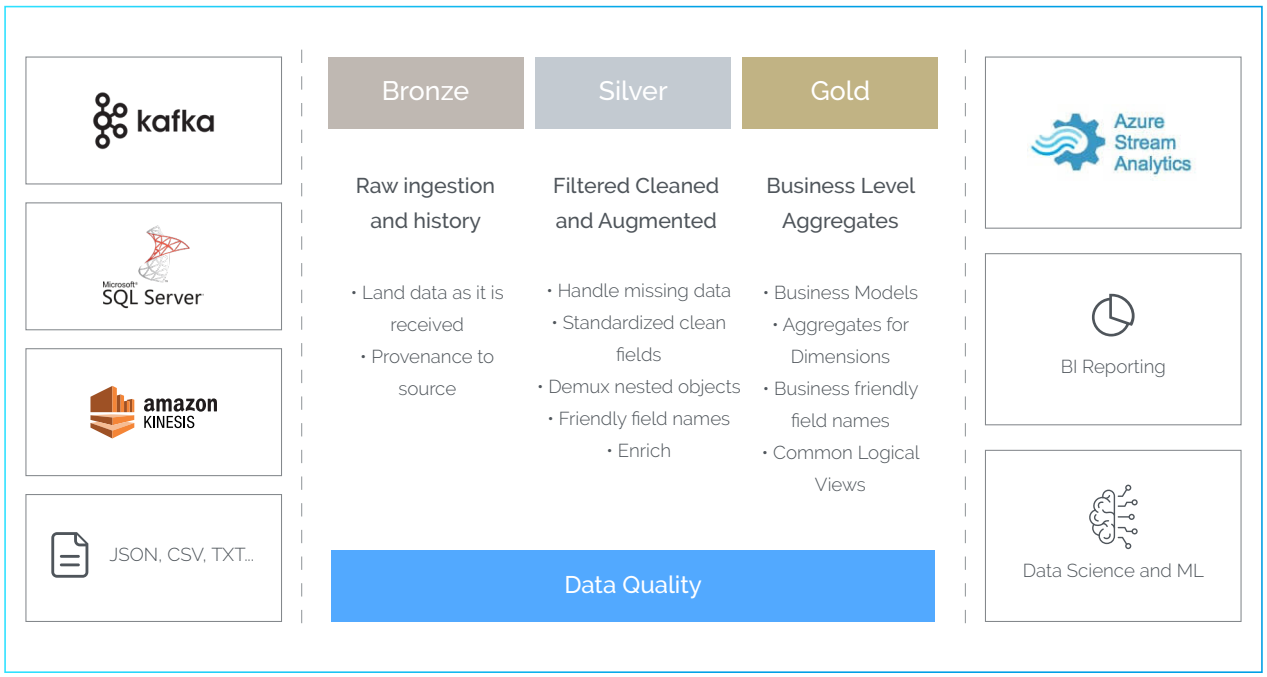
03 **Support for Delete,** Updates & Merge operations: Delta Lake supports these operations to enable Change-Data-Capture (CDC) use cases, Slowly-Changing-Dimension (SCD) operations and streaming upserts.

04 **Support for Batch and Streaming operations:** Delta Lake has an ability to work in both Batch and streaming source/sink modes across a wide variety of latencies. These capabilities enable use-cases like Real-time fraud detection.

There are a set of other useful capabilities enabled by Delta Lake's transaction log such as Time Travel and Data Lineage that we will cover in later sections of this white paper.

Medallion Architecture using **Delta Lake**

To address ever-evolving and complex business needs, Databricks recommends a Medallion architecture for your data platform. The goal of this approach is to progressively improve the quality and structure of data from one layer to the next (Bronze->Silver->Gold). Through Delta Lake, this architecture is easy to create and manage.



In this architecture, all the raw data from source systems resides in Bronze layer. The tables in Bronze layer have an "as-is" correspondence with source systems, along with other meta-data that is captured as part of the ingestion process. Key benefits of this layer are quick Change Data Capture, historical archive of source data, data lineage and auditability.

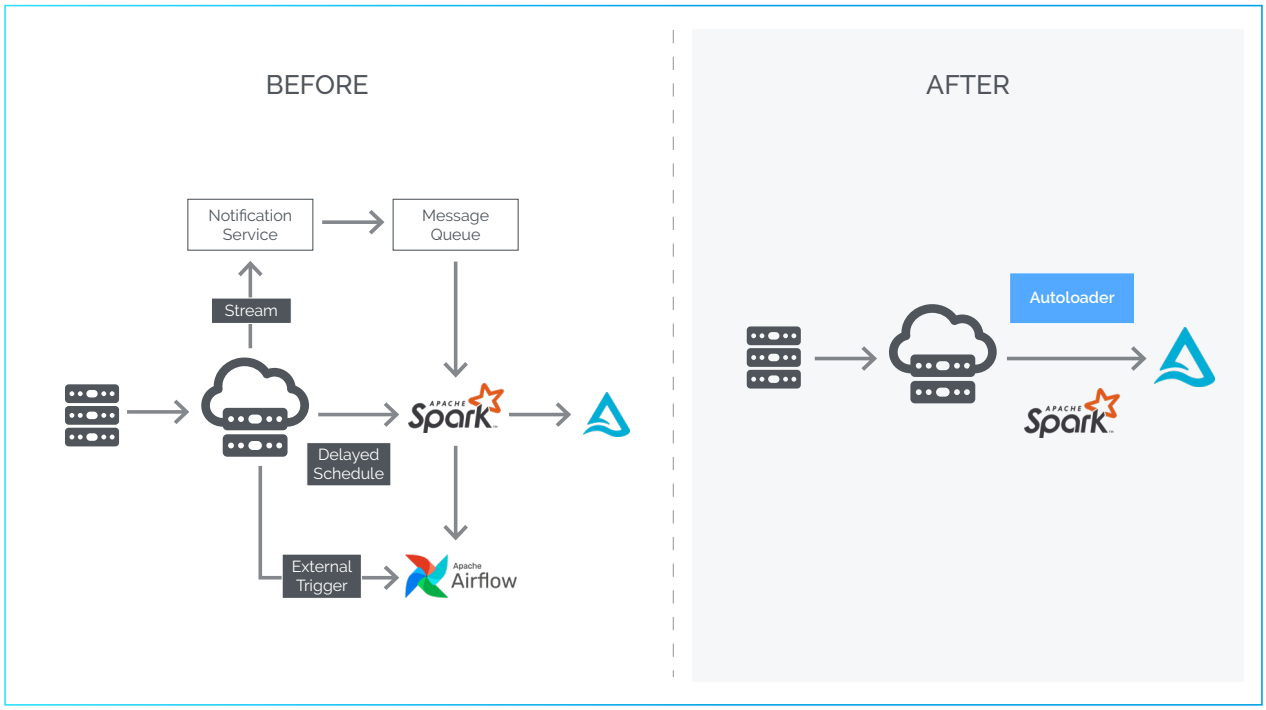
The Silver layer contains a cleansed version of the data from Bronze layer to provide a self-service layer or for ad-hoc reporting use-cases. This layer is primarily used by Analysts, Data engineers and scientists for their projects and analysis.

The Gold layer is a consumption ready layer for business reporting and final presentation layer for initiatives such as Customer Analytics, Segmentation, Product recommendations etc.

Autoloader

Introduced in February 2020, Autoloader enables incremental ingestion of Data into Delta Lake from a variety of cloud storage solutions like S3, ADLS or GCS. In the previous section, we saw how ACID transactions in Delta Lake enable data to be reliably read and queried. However, as more files get added to cloud storage manually, data ingestion into Delta Lake can have high end-to-end data latencies and a manual setup process to enable the ingestion that can cause failures.

Both of these issues are addressed via Autoloader, which is an optimized File Source called "cloudFiles" that will automatically process new files as they arrive. Autoloader thus enables scalability and ease of use in ingestion for both Batch loads and Streaming loads. Autoloader also enables downstream data transformation and pipelines using Delta Live Tables covered in the next section.



✖ Gets too complicated for multiple jobs

✔ Pipe data from cloud storage into Delta Lake as it arrives

✔ "Set and Forget" model eliminates complex setup

Delta Live Tables

Once data is ingested in Delta Lake via Autoloader in Bronze tables, it needs to be further transformed, enriched and processed for Silver and Gold layers via Data pipelines. However, creating fault tolerant and scalable data pipelines in any platform has historically suffered from three issues: Complexity in development, Poor Data quality and issues in error handling/recovery.

COMPLEX PIPELINE DEVELOPMENT	POOR DATA QUALITY	DIFFICULT PIPELINE OPERATIONS
• Hard to build and maintain table dependencies	• Difficult to monitor and enforce data quality	• Poor observability at granular, data level
• Difficult to switch between batch and stream processing	• Impossible to trace Data Lineage	• Error Handling and recovery is laborious

Delta Live tables is an ETL framework that makes it easy to build data pipelines for high quality data through declarative tools to develop and manage data pipelines, automated testing through "expectations" and automated error handling. A few key capabilities of Delta Live Tables are:

01 **Visualization DAGs:** Delta Live Tables creates visualization of Direct Acyclic graphs of pipelines to show dependencies and status of each pipeline. This enables to identify syntax errors in a pipeline and provide visual representation of lineage.

02 **Data Quality:** Delta Live Tables have inbuild data quality checks to fix, allow or fail the transformations based on severity of data rules or "expectations"

03 **Pipeline Observability:** Delta Live tables provide detailed view of historical runs of pipelines, run-time statistics such as records processed/rejected etc, logs and operational metrics for performance tuning.

04 **Automated Infrastructure Management:** Through Delta live Tables, complex and time-intensive tasks such as Orchestration, error handling, auto-scaling are automated.

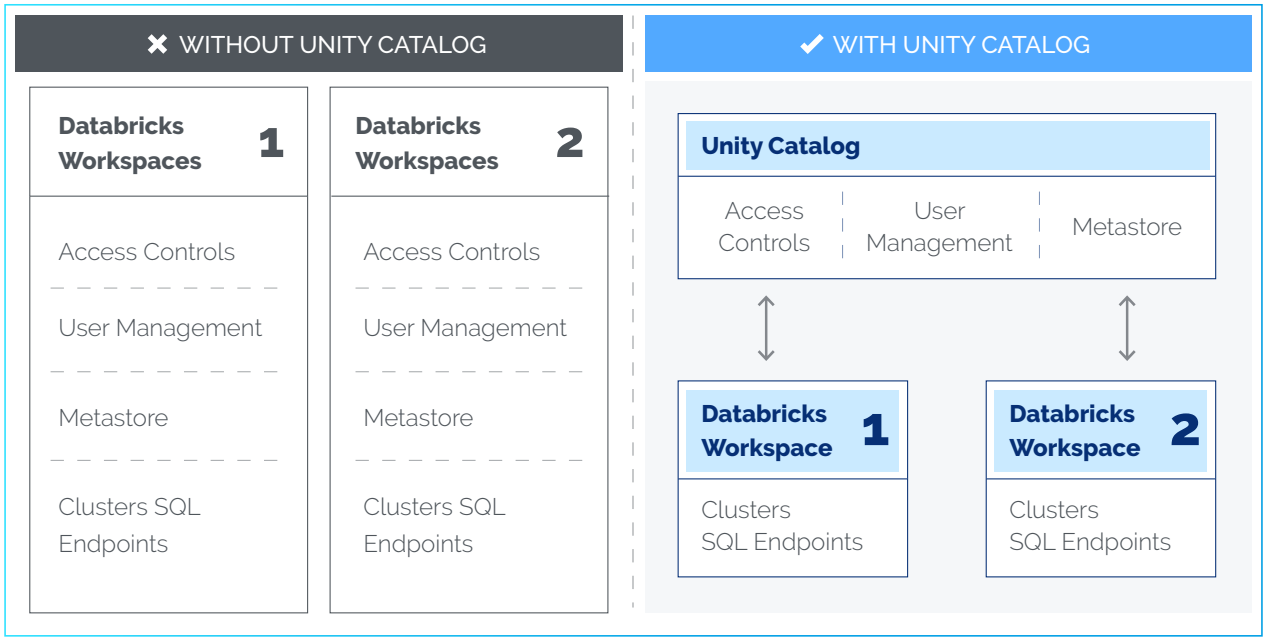
A major oil and gas company aggregates all sensor data into an integrated Data store. Through Delta Live tables they have improved efficiency in managing data at multi-trillion record scale, while continuously improving their AI capabilities.

Source: Databricks

Unity Catalog

As Data becomes more democratized in organizations, there is a need for adequate governance to ensure Data discoverability, lineage, collaboration and security. With Unity Catalog, these capabilities come out of the box with Databricks.

Unity Catalog is a metadata management service provided by Databricks that enables users to easily discover, understand the relationships between data assets and how they are used in different parts of their organization.



01 **Data discovery:** Unity Catalog enables users to search for data assets across different systems and view their metadata in a single interface. This feature helps users identify relevant data assets and understand their properties, such as schema, format, and location.

02 **Data lineage:** Unity Catalog tracks the lineage of data assets across different systems, enabling users to understand how data is transformed and processed as it moves through different stages of the data pipeline. This feature helps users identify issues and troubleshoot problems in their data pipelines.

03 **Data governance:** Unity Catalog provides a set of tools for managing data governance policies, such as access controls, data quality rules, and retention policies. This feature helps users ensure compliance with data regulations and maintain data integrity.

04 **Collaboration:** Unity Catalog enables users to collaborate with each other by sharing metadata and annotations on data assets. This feature helps users share knowledge and insights about data assets, leading to improved data quality and faster data discovery.

05 **Integration:** Unity Catalog integrates with a wide range of data systems, including databases, data lakes, streaming systems, and cloud storage. This feature enables users to manage metadata from different systems in a single interface, reducing the complexity of data management.

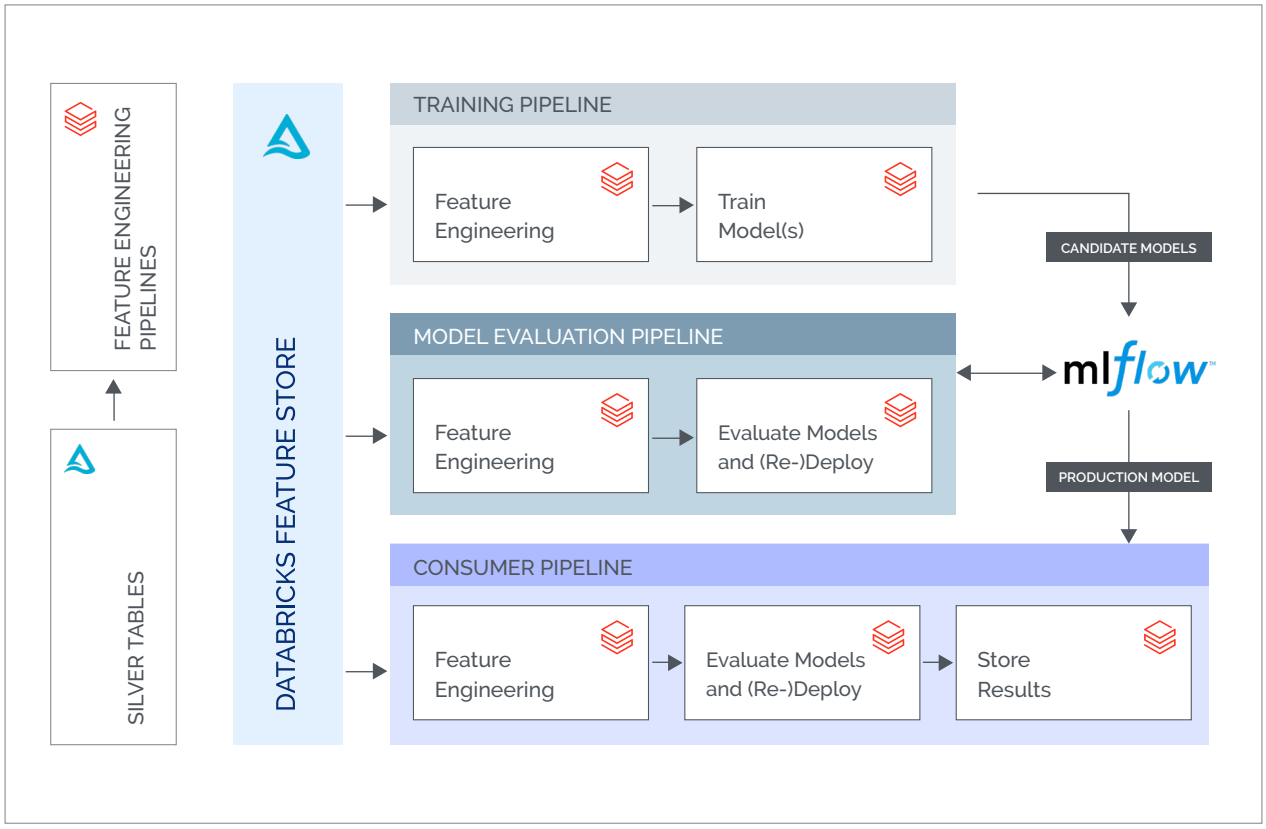
At a leading Tech unicorn, Unity Catalog simplified implementation of role-based access control, thus securing data at catalog, database, table and column level. Through Unity Catalog, they can easily provision appropriate level of access to users and can meet compliance and privacy policies.

Source: Databricks

ML Ops

Ever since advent of machine learning and data science a decade ago Industry has been looking for platforms which can facilitate model development and model implementation processes at scale in industrialized fashion. This requirement of scalability makes model life cycle management a complex problem. For example, organizations want their models to be dynamic which can be tuned quickly with changing requirements of market and customers. This leads to the problem of version control. Similarly, platforms are required to support collaborative development of models with many data scientists working together on a same set of models. While some of the tools like AWS Sage maker, Data robot etc. have solved some of the above problems, they come with their own set of integration and access management problems. These tools typically work as stand-alone platform which needs to be integrated with Data lakes and data warehouses.

Databricks MLOps and data science platform provides a very strong solution to above problems. Despite being a unified platform, MLOps capabilities of Databricks compare well with the stand-alone machine learning platforms. Also, since it is a unified platform, problems of integration and access management get eliminated by itself.



Databricks ML Ops platform is built on open-source library of MLflow. Data scientists can make use of MLflow functionalities either through Databricks notebook or through an intuitive User Interface. Leveraging MLflow Databricks provides a comprehensive suite of capabilities to support the entire machine learning life cycle, from experiment tracking and model management to model deployment. As part of model management capabilities, platform allows to create repository for thousands of different versions and variations of the models. In this repository each model is listed along with different performance parameters. Users can select any one of these models to be pushed in the production. Also, if they need to switch from one model to other in production it is extremely easy.

Model management functionality helps streamline the process around model development and model deployment. This allows for tighter collaboration between data teams, reduced conflict with DevOps and IT, and accelerated release velocity. It facilitates evolution of model development processes in a factory or assembly line model where different data scientists can work on different components in synchronized, collaborative and scalable manner to achieve the common goal. Beyond development process, Data scientists can deploy models in production very easily using batch inference on Apache Spark or real-time serving via REST APIs, making sure that the models are compliant with industry policies. To ensure necessary security deployment is allowed only through an authenticated Databricks token.

Ease of hyperparameter tuning of parameters is another key feature of Databricks machine learning platform. Data scientists can select the parameters on which they want to tune the models and then platform will automatically reiterate model development with different combinations of parameters. At the end of process, it will list all variations with their performance parameters along with the recommendation of the model which comes out on top.

ML Ops platform also facilitates experimentation in model development process. As a result, Model developers mostly need to work on feature engineering tasks only. Once they have cleaned up data and identified features, the platform will run experiments automatically. For these experiments, data scientists can select not just the features they want to run the experiments on, they can also choose different algorithms and methodology they want to experiment with.

Databricks also helps increase efficiency through faster model development and deployment cycles, scalability to manage thousands of models at once, and increased transparency in ML pipelines.

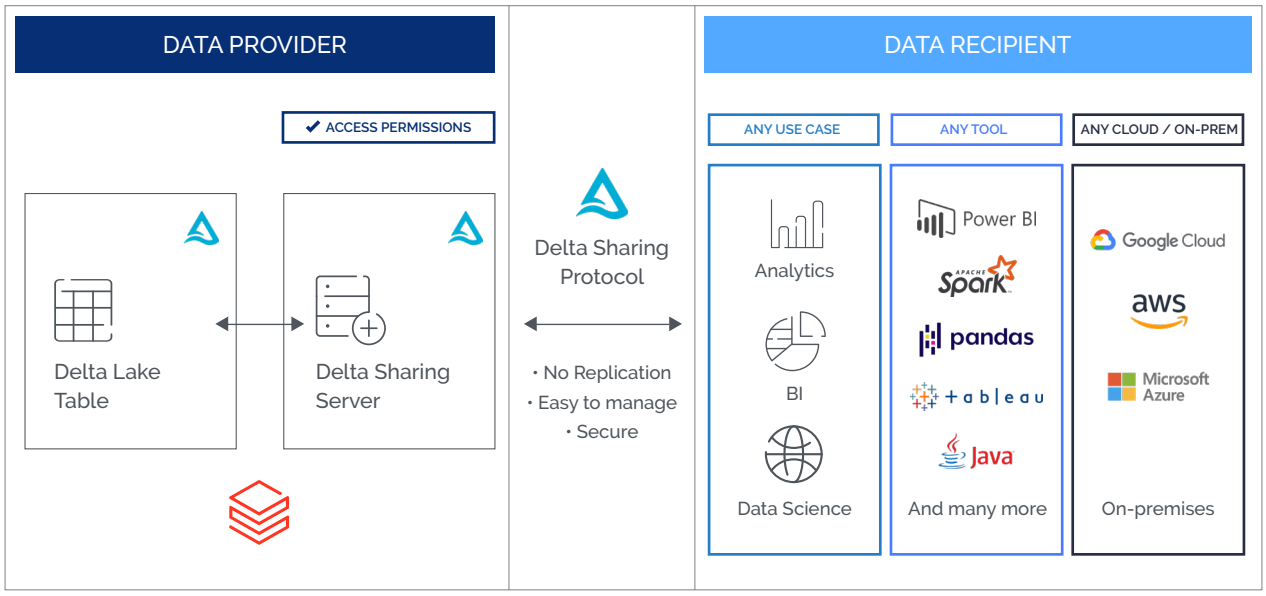
At a market leading Apparels firm, Databricks has enabled many business units to use the platform in an easy, self-service manner that was not possible earlier.

Source: Databricks

Delta Sharing

There is a growing need in businesses today to share data with multiple stakeholders, both internal and external. As an example, Marketing and Sales teams need to have a common view of data on marketing campaigns and downstream revenue to assess impact and ROI of marketing efforts. Or companies need to share real-time sales data with their suppliers to ensure a smooth supply chain and not cause revenue losses due to stock-outs. Most organizations have either developed a home-grown solution (like a SFTP or custom APIs) to share data or rely on proprietary solutions or use a Cloud storage solution like AWS S3 or Azure Blob. However, all of the above approaches do not simultaneously address the requirements of being cloud-agnostic, preventing lock-ins and ease of management & governance.

Delta Sharing is an open solution to share data from Lakehouse to any computing platform. It is designed in a manner that the recipient need not be on any cloud and data providers can share live data without a need to replicate it. Delta sharing is also natively integrated with Unity catalog thus making it easy to manage, govern and track all data sharing activities. Delta sharing also enables sharing of large datasets in parquet format on Delta lake without a need to move data.



At a leading online marketplace for securities transactions, Delta sharing has helped to streamline data delivery process for large data sets. This enables their clients to bring their own compute environment to ready freshly curated data with little to no integration work. Through Delta sharing, they also continuously expand its catalog of high quality data products.

Source: Databricks

Archival and Time Travel

Enterprises are capturing increasingly vast amounts of data (structured and unstructured) and using it for downstream analytics & ML-driven projects. Due to the vast amounts of data being captured and analyzed, there is a growing need to store & manage this data effectively to address multiple requirements:

01 **Compliance:** Regulations like HIPAA and CCPA are forcing businesses to store customer data for longer periods of time. HIPAA for example requires healthcare providers to store Patient Health Information (PHI) for a minimum of six years. Furthermore, businesses need to make this data available on request either from regulatory authorities, stakeholders or their customers.

02 **Disaster Recovery:** According to Statista, downtime can cost companies more than $5M/hour. Since data is mission-critical and hence businesses need to ensure recovery and continuation of data infrastructure against natural or human-induced disaster. Businesses need to ensure an adequate data retention strategy to counter these threats.

03 **ML Reproducibility:** In a Big-data environment, data is being constantly updated and changed. However, coming up with an effective ML model is an iterative process that requires data scientists to improve model accuracy against fixed datasets. Hence there is a need to have snapshots of the underlying data to improve the accuracy as well as reproduce experiments that were carried out in the past.

04 **Rollbacks:** It is common in complex systems that erroneous operations are performed, or deployment issues cause production servers to crash. These cause database integrity issues and hence need specific approaches to restore the data.

05 **Audit:** As use of data becomes more democratized in businesses and pipelines become more complex with multiple, evolving input sources, there is a need to audit changes in the destination data to ensure high data quality.

In addition to these needs, data teams also frequently need to share datasets with multiple internal teams as well as prepare for data migration scenarios as and when they come up. These use-cases add up over time and become very complex, effort & time intensive to manage.

Databricks has designed multiple capabilities in its platform such as Clones & Time Travel that address all these requirements above in a highly streamlined and efficient manner.

Clones

Introduced in September 2020, Clones is an easy way to create copies of tables in Data Lakehouse. Clones are created as a replica of a source table and inherit all characteristics of the source table including meta-data, schema, constraints, partionining, column description and statistics. However, clones are separate tables with their own history and lineage. Creating a clone is as simple as writing two lines of Python code below:

```
CREATE OR REPLACE TABLE  customer_order_clone
DEEP CLONE  customer_order
```

Databricks Deltalake has two types of clones: **Shallow Clones** and **Deep Clones.**

A **Shallow clone**, also known as a zero-copy clone, does not copy the underlying data of the source table, it only duplicates the meta-data. Hence a Shallow clone is not self-contained and is typically used for short-lived use cases such as testing and experimentation.

A **Deep clone,** on the other hand, replicates both the meta-data and the data files of the source table.

Clones are especially useful for Data archival to address compliance other related requirements. Typically, production tables retain data for specific time-periods such as last few months only. Hence if the production tables are getting updated in real time, retrieving older data becomes prohibitively time consuming and expensive. Clones are very effective in addressing these scenarios. Clones also have incremental cloning capability that enable faster, robust cloning and thus address failure scenarios effectively.

Clones also enable easy data sharing across multiple teams and users within the organization. Clones preclude the need to setup new pipelines to move the data to another store, as it is now easier to create a copy of relevant datasets.

Time Travel

Introduced in Databricks Delta Lake in February 2019, Delta automatically versions big data that is stored in Data Lake. This capability makes it very easy to roll-back data in case of accidental erroneous updates/deletes and reproduce ML experiments and reports. This is enabled by Multi-version Currency control protocol.

Time-travel feature has the capability to access different versions of data in two ways. First is using a Timestamp or Date String, in case you want to access the version at a specific instance in time. Second is using a version number. In Delta, every write has a version number that you can use to travel back in time.

Through Time-travel, you can look at the history of a table using "DESCRIBE HISTORY" command. Thus, you can audit data changes easily. Time-travel also enables to do a rollback in case some information is accidentally deleted or a table in erroneously updated.

Time Travel also enables easy reproducibility of ML Experiments, by integrations with MLFlow. Data scientists can log a time-stamped URL to the path as a parameter to track which version of data was used for a training model.

Conclusion

Databricks is a very powerful, integrated end-to-end Data, Analytics and AI platform that addresses a wide range of challenges that are posed by modern data teams. Built on open-source foundations like Apache Spark and Delta Lake, Databricks offers great flexibility and openness to enable both data and business teams to operate at scale in a self-service manner. Powerful capabilities built over time such as Delta Live Tables, Unity Catalog and MLOps places Databricks uniquely in Modern Data Stack as a solution for all data needs in a large or small business alike.

EUCLOID

Eucloid Data solutions is an award-winning Data & Growth consulting organization. With presence in multiple countries, Eucloid is one of the top Databricks partners in US. With experience in solving data challenges of industries like Retail, BFSI and Pharmaceuticals, along with deep tech expertise in Modern Data Stack, we help drive business growth through Data, Analytics and AI. Reach out to us at **sales@eucloid.com.**