Azure Databricks

Who are we ?

Laurent Leturgez

I'm a data and cloud Architect and Spark lover. I worked many years as an Oracle consultant and expert, and now I work with Cloud solutions devoted to solve complex problems with high volumes of data.

Alexandre Bergere

I am a Data Analyst & Solution Architect indepedent - ☁ MCSE, Cosmos DB & Delta lover.
I developed my skills through various clients' projects, teaching at the University and personal proof of concepts.
I'm also the Co-Founder of DataRedkite, a product which can quickly give to its user a good management of data in Microsoft Azure DataLake.

PREMISEO
DATA, SYSTEMS, CLOUD

DataRedKite

Our Guest !

Matthieu Lamairesse

Solutions Architect at Databricks

I'm a Data Analytics geek.
I got started in the BI would and quickly moved the the Big Data world when it became a thing. I have over 8 years of experience in Data Science and Big Data acquired in successively at different startups such as Hortonworks/Cloudera and of course Databricks

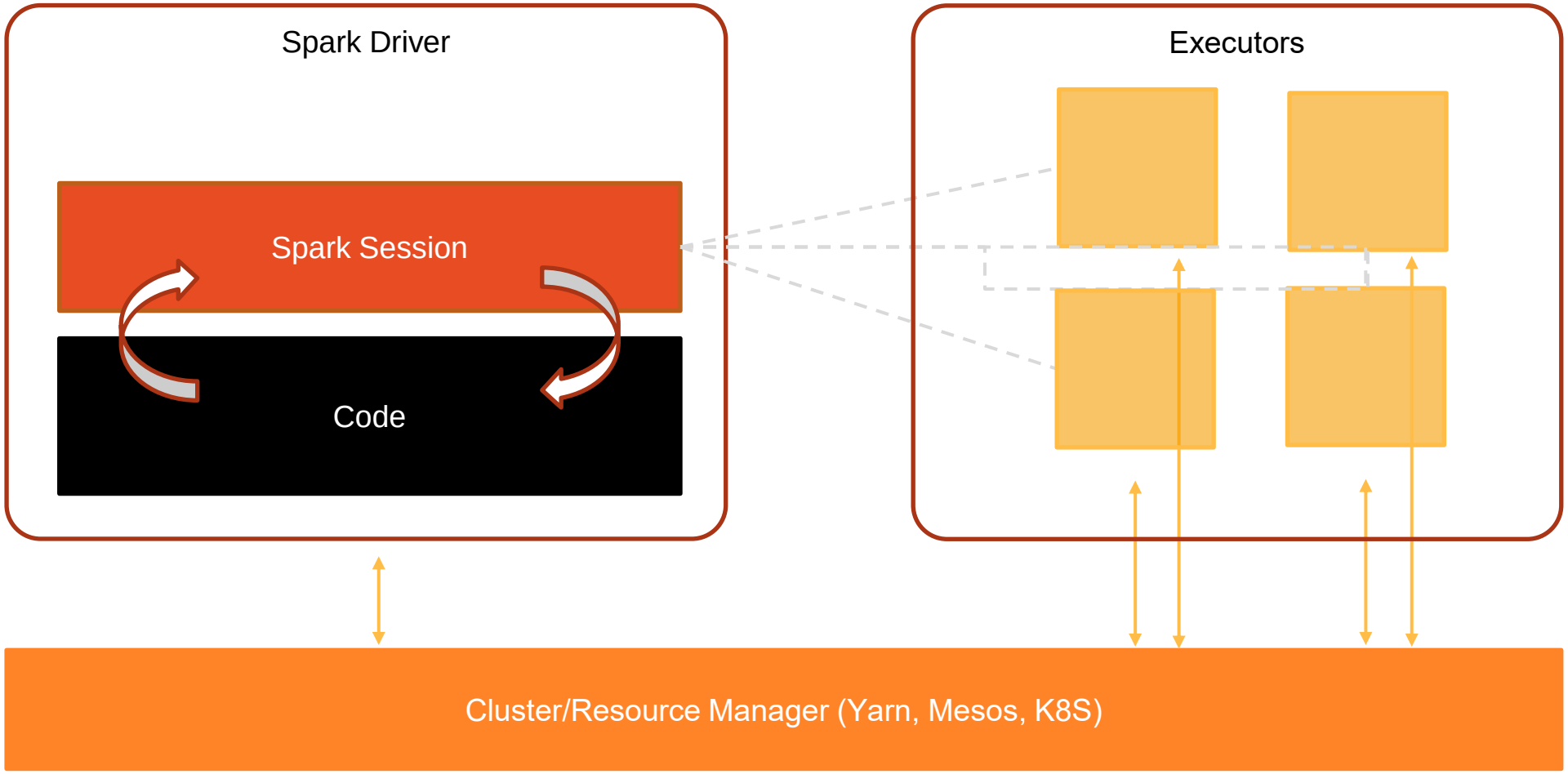
Azure Databricks

Azure Databricks

- A managed platform
- Databricks for Data Engineer
- Databricks for Data Scientists
- Databricks for Data Analysts

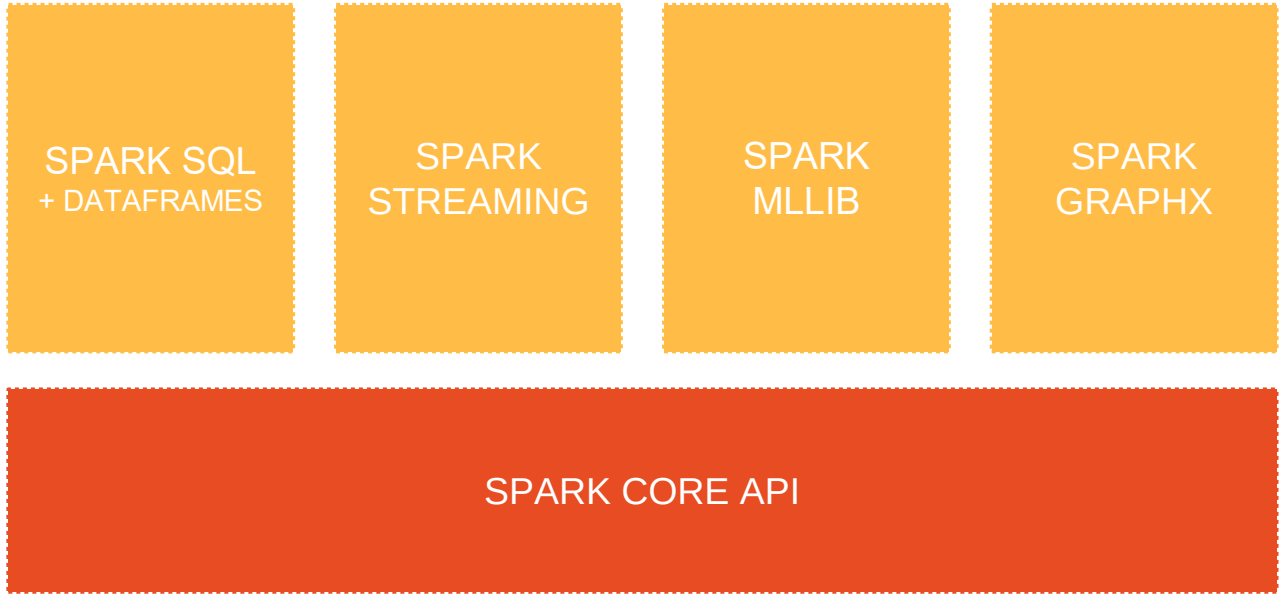
Spark

What's Spark ?


Spark Driver
Spark Session
Code
Executors
Cluster/Resource Manager (Yarn, Mesos, K8S)

What’s Spark ?

Combine **SQL**, **streaming**, and **complex analytics**.



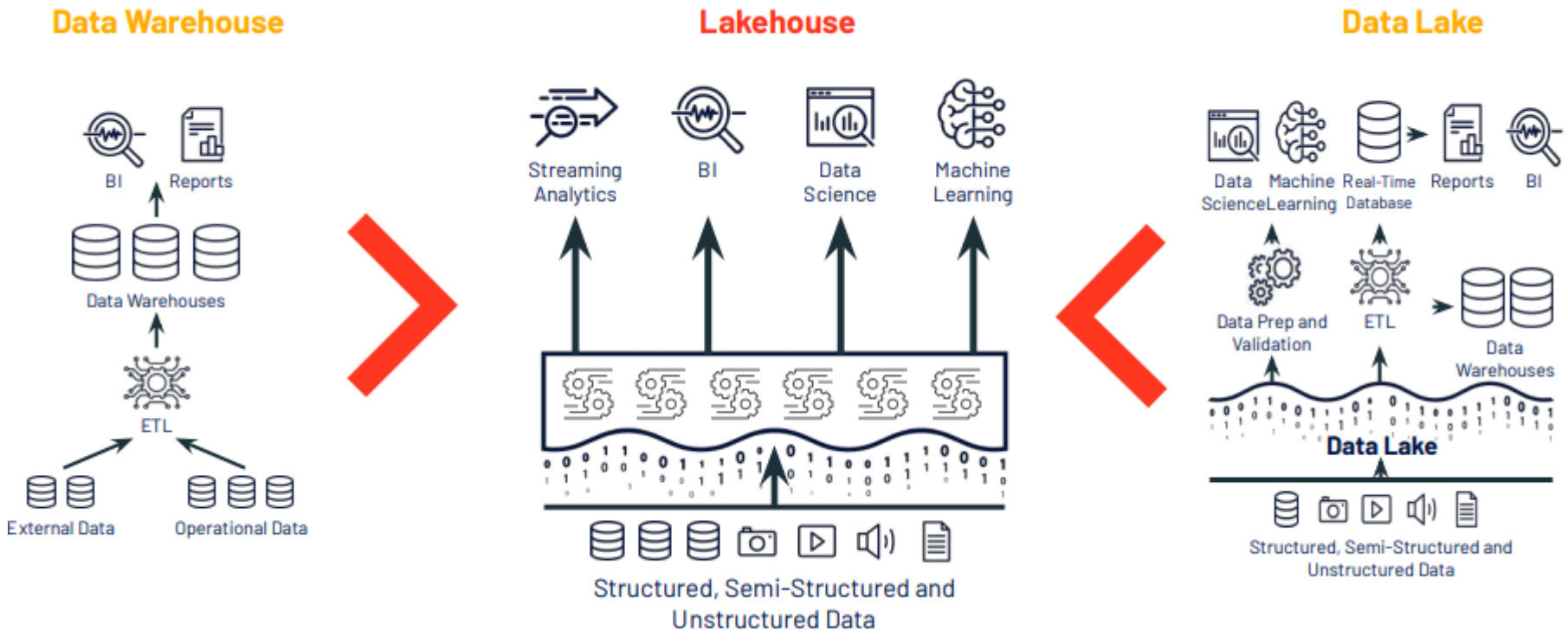
Databricks

Spark’s founders

Accidental Billionaires: How Seven Academics Who Didn’t Want To Make A Cent Are Now Worth Billions

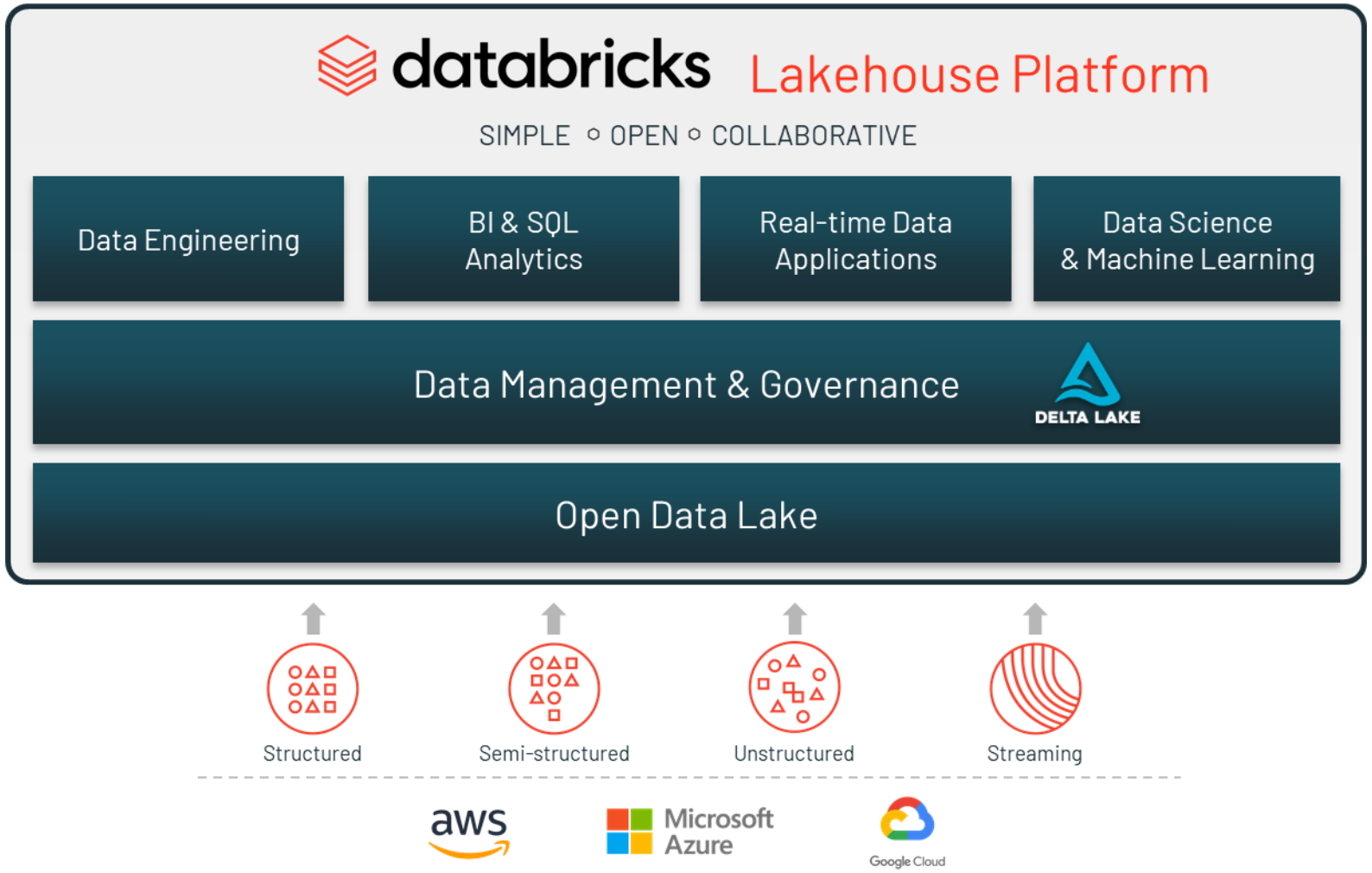
https://www.forbes.com/sites/kenrickcai/2021/05/26/accidental-billionaires-databricks-ceo-ali-ghodsi-seven-berkeley-academics/?sh=59e3a3247008

Lakehouse = Data Lake + Data Warehouse



http://cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf

Azure Databricks : A unified plateform



Azure Databricks

Azure Databricks is a data analytics platform optimized for the Microsoft Azure cloud services platform. Azure Databricks offers two environments for developing data intensive applications:

D Data Science & Engineering

- Azure Databricks Workspace: provides an interactive workspace that enables collaboration between data engineers, data scientists, and machine learning engineers.

M Machine Learning

- Azure Databricks Machine Learning : Provides a complete toolset for MLOps based on the popular MLFlow Project. With it's hosted services it's easy to capture and compare training experiments, register models and feature, deploy and follow models in production

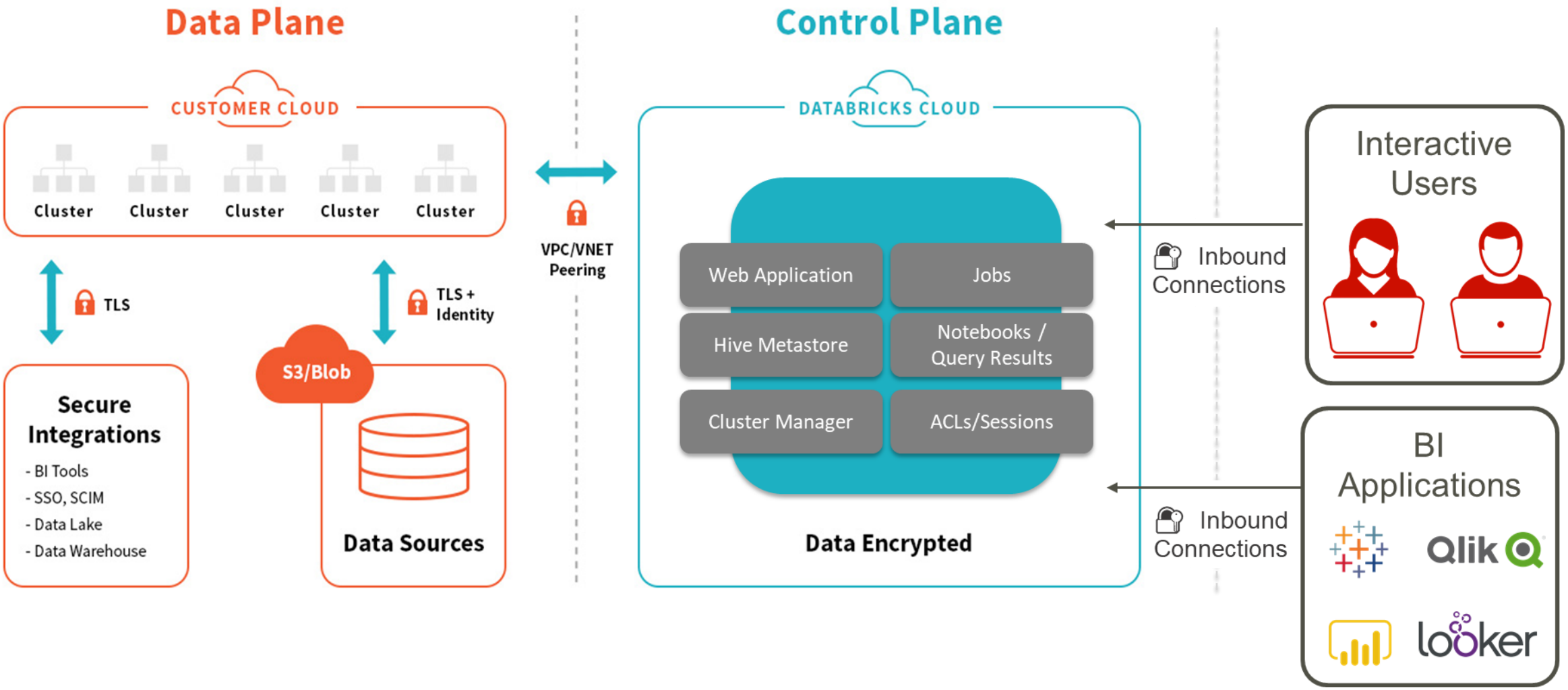
S SQL

- Azure Databricks SQL Analytics : provides an easy-to-use platform for analysts who want to run SQL queries on their data lake, create multiple visualization types to explore query results from different perspectives, and build and share dashboards.

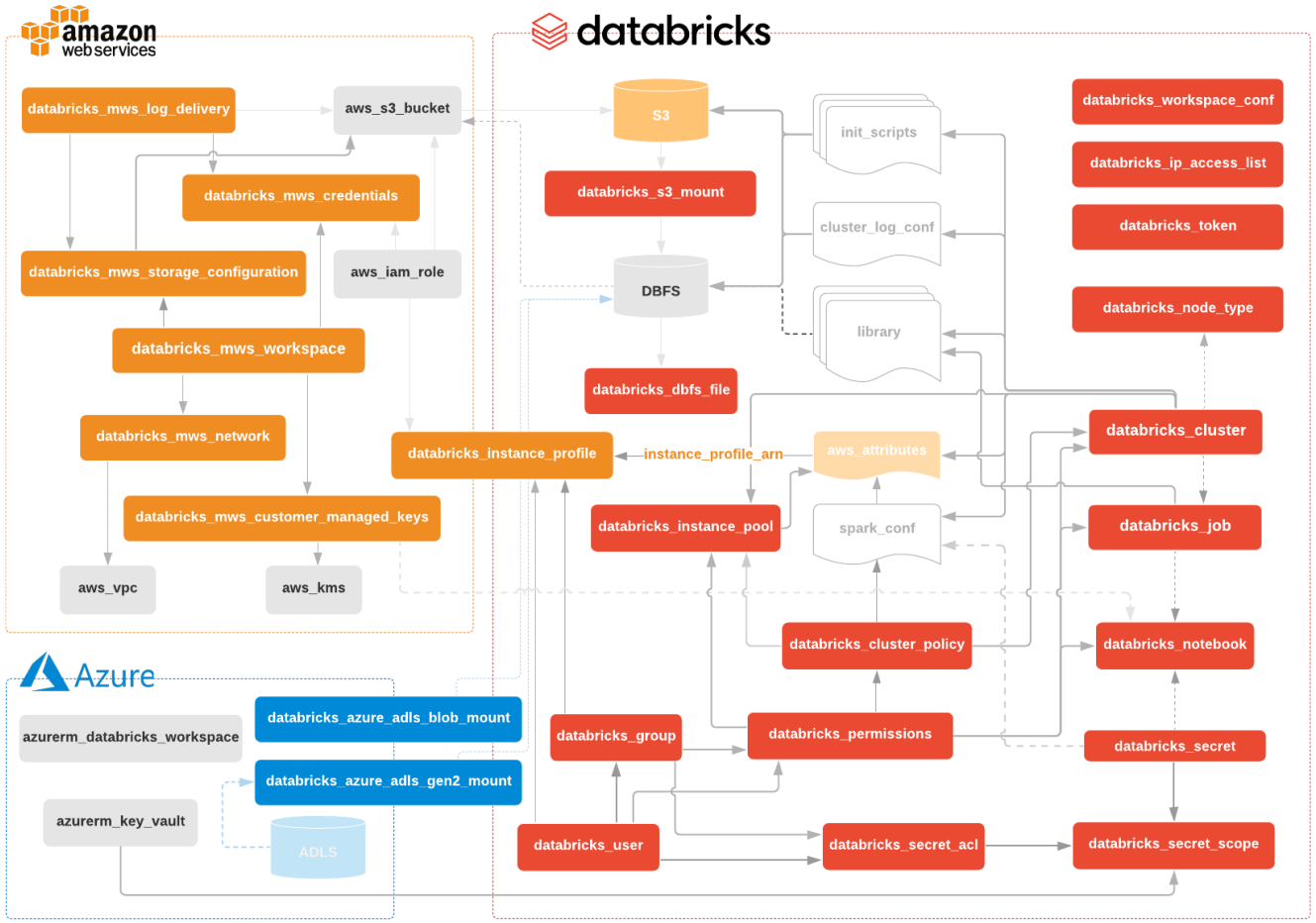
Azure Databricks

Demo time

Control Plan – Data Plan



Terraform provider databricks



https://github.com/databrickslabs/terraform-provider-databricks
https://registry.terraform.io/providers/databrickslabs/databricks/latest/docs

Databricks Platform features

- Databricks File System (DBFS) : A filesystem abstraction layer over a blob store. It contains directories, which can contain files (data files, libraries, and images), and other directories. DBFS is automatically populated with some datasets that you can use to learn Azure Databricks.

- Spot instances : Spot instances allow you to use spare computing capacity and choose the maximum price you are willing to pay.

- Pools : pools reduce cluster start and auto-scaling times by maintaining a set of idle, ready-to-use instances.

Databricks for Data Engineer

Demo time

Databricks Engineer features

- Metastore : Data Catalog providing a data access abstraction on top of the datalake structures in familiar concepts such as Database.

- Unity Catalog : Fine-grained Governance for Data and AI on the Lakehouse ***(preview)***.

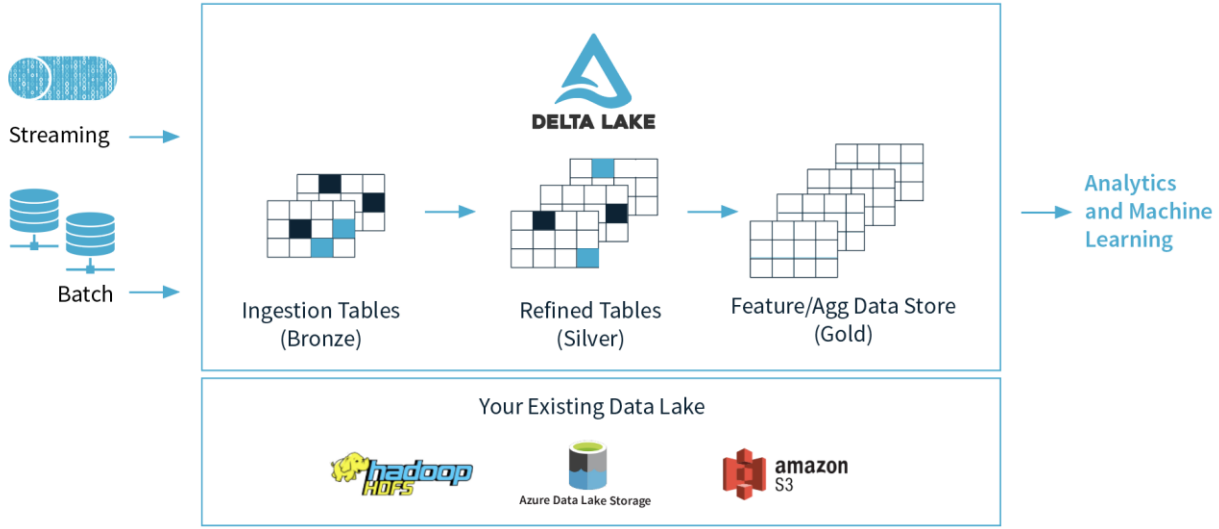
- GitHub version control : set up version control for notebooks using GitHub through the UI, Databricks CLI or Workspace API.

- Databricks connect : allows you to connect your favorite IDE (Eclipse, IntelliJ, PyCharm, RStudio, Visual Studio), notebook server (Jupyter Notebook, Zeppelin), and other custom applications to Databricks clusters.

Delta Lake

Delta Lake is an open-source storage layer that brings ACID transactions to Apache Spark™ and big data workloads.

1. Hard to append data
2. Modification of existing data difficult
3. Jobs failing mid way
4. Real-time operations hard
5. Costly to keep historical data versions
6. Difficult to handle large metadata
7. “Too many files” problems
8. Poor performance
9. Data quality issues



Delta Lake features

- ACID : Make every operation transactional and ensure that readers never see inconsistent data.
- Time travel (data versionning) : All transactions are recorded and you can go back in time to review previous versions of the data - Data versioning for reproducing experiments, rolling back, and auditing data.
- Indexing : Automatically optimize a layout that enables fast access (Partitioning, Data Skipping, Z-ordering).
- Schema enforcement and evolution : prevents users from accidentally polluting their tables with mistakes or garbage data and enables them to automatically add new columns.
- Delta live Tables : build and manage reliable data pipelines that deliver high quality data on Delta Lake.
- Delta Sharing : An Open Protocol for Secure Data Sharing.

Databricks for Data Scientists

Demo time

MLflow features

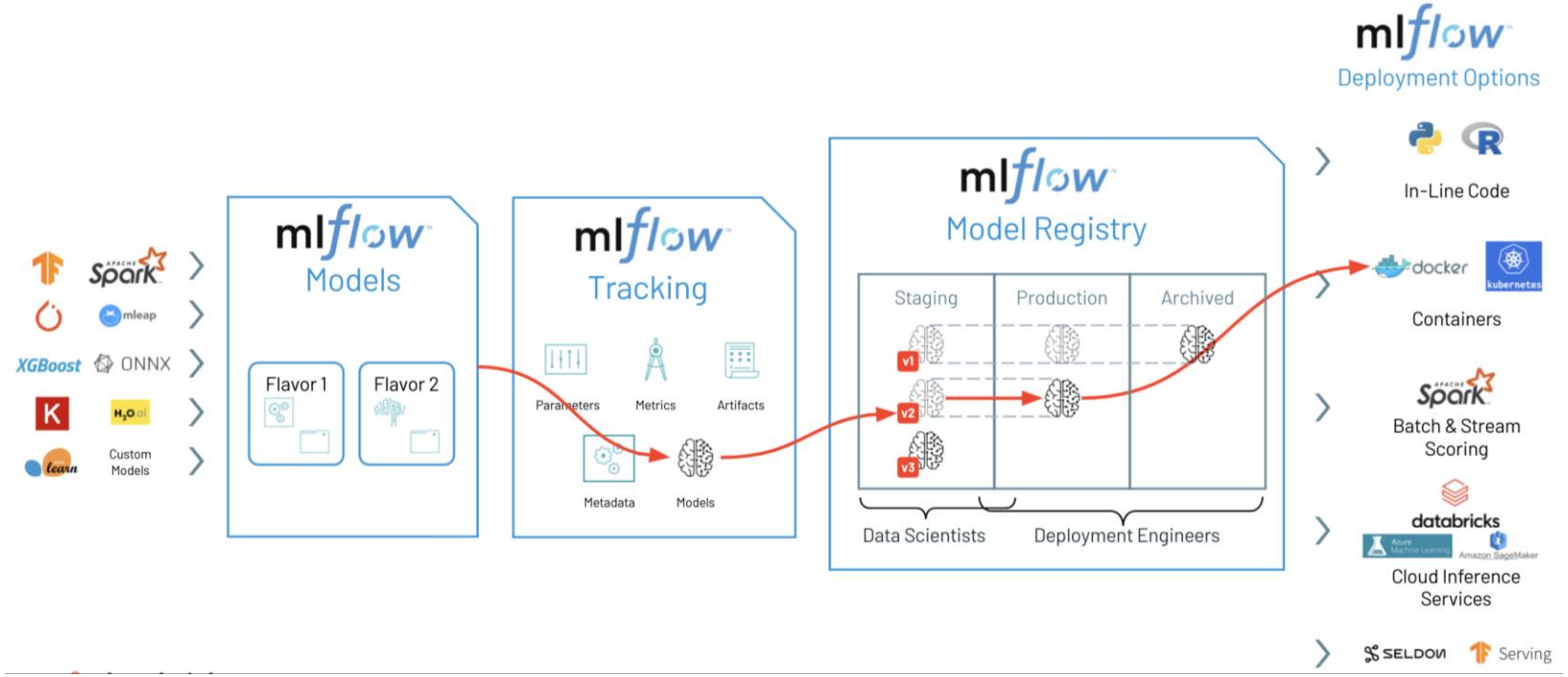
MLflow is an open source platform to manage the ML lifecycle, including experimentation, reproducibility, deployment, and a central model registry.

- MLflow Tracking: Automatically log parameters, code versions, metrics, and artifacts for each run using Python, REST, R API, and Java API.

- MLfow Projects: Package data science code in a format to reproduce runs on any platform.

- MLflow Models: A standard format for packaging machine learning models that can be used in a variety of downstream tools—for example, real-time serving through a REST API or batch inference on Apache Spark.

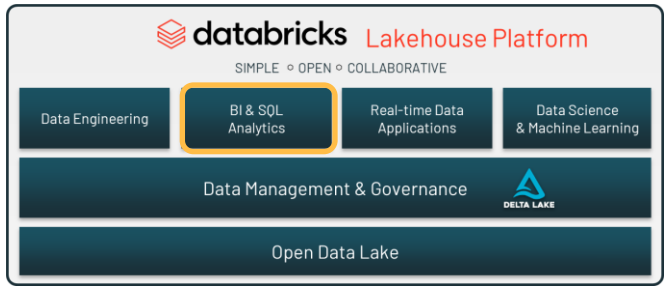
- MLflow Model registry: collaborative hub where teams can share ML models, work together from experimentation to online testing and production, integrate with approval and governance workflows, and monitor a ML deployments and their performance (versioning, ci/cd workflow …).



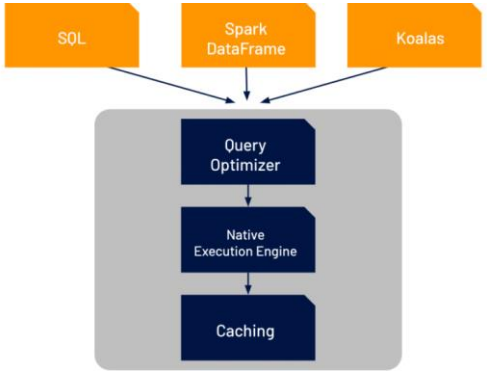
Databricks for Data Analysts

Demo time

Databricks SQL features



- Delta Engine : a new query engine designed for speed and flexibility. It's built from the ground up to deliver fast performance on modern cloud hardware for all data use cases across data engineering, data science, machine learning, and data analytics.



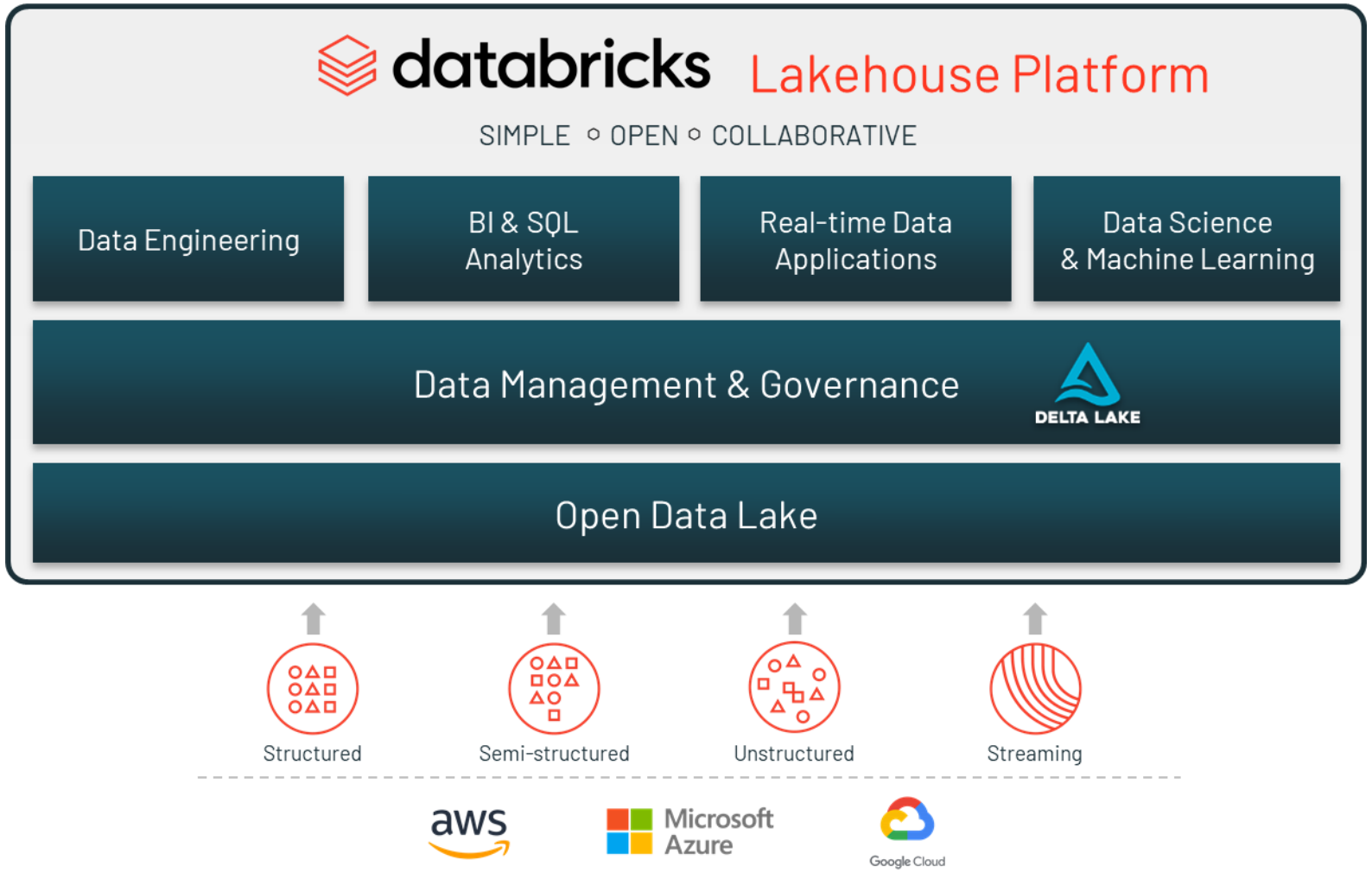
- Query Editor : where users can explore their databases, write SQL queries with intelligent auto-complete, and view query output in either a tabular display

- Dashboarding : Easily Create Visualizations and Share Dashboards through Databricks.

- SQL Endpoint : provides easy connectivity to other BI and SQL tools via ODBC/JDBC connections (Power BI, Tableau, Qlik …)

Delta Engine

Azure Databricks : A unified analytics plateform

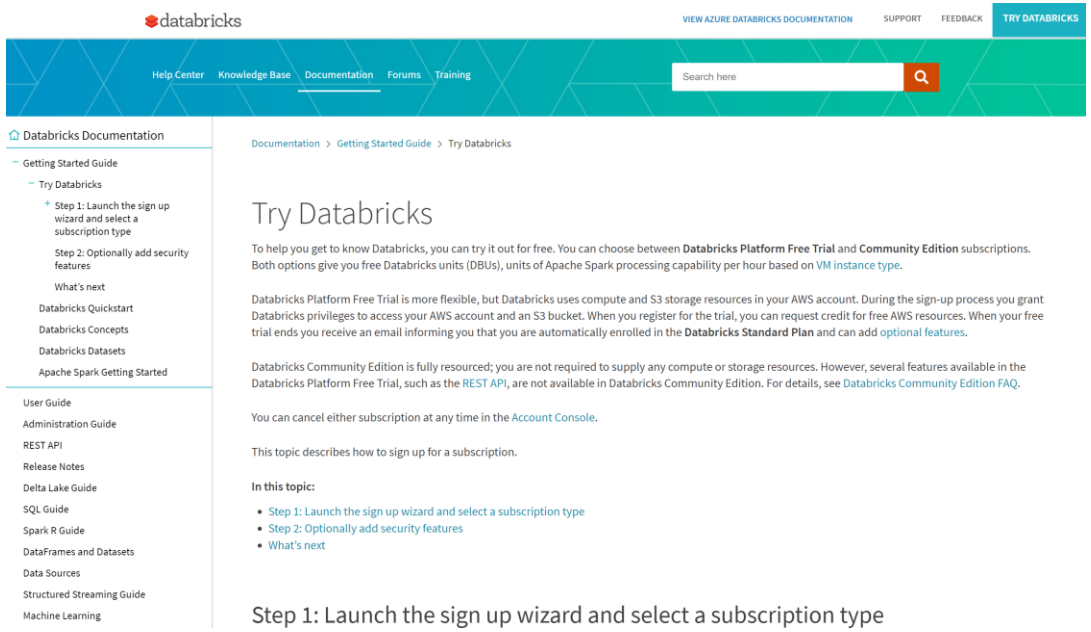


Resources

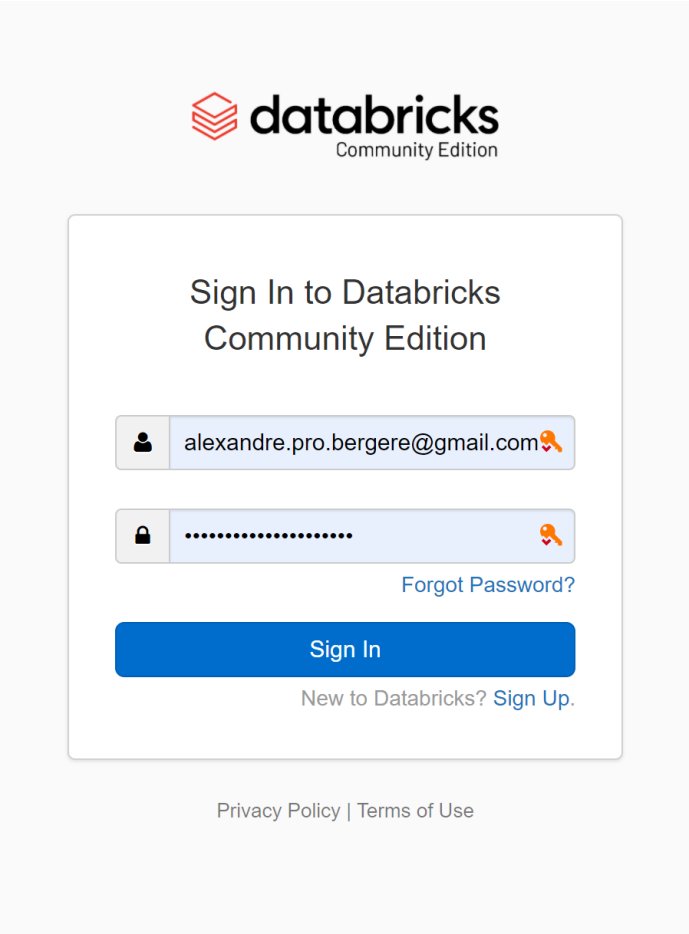
Databricks

You can find the main sources from here:

- docs.databricks.com
- Databricks academy.
- SparkAISummit video archive.
- delta.io
- delta slack channel
- Azure Databricks Best practices
- The path Perform data engineering with Azure Databricks in Microsoft Learn.



Databricks community

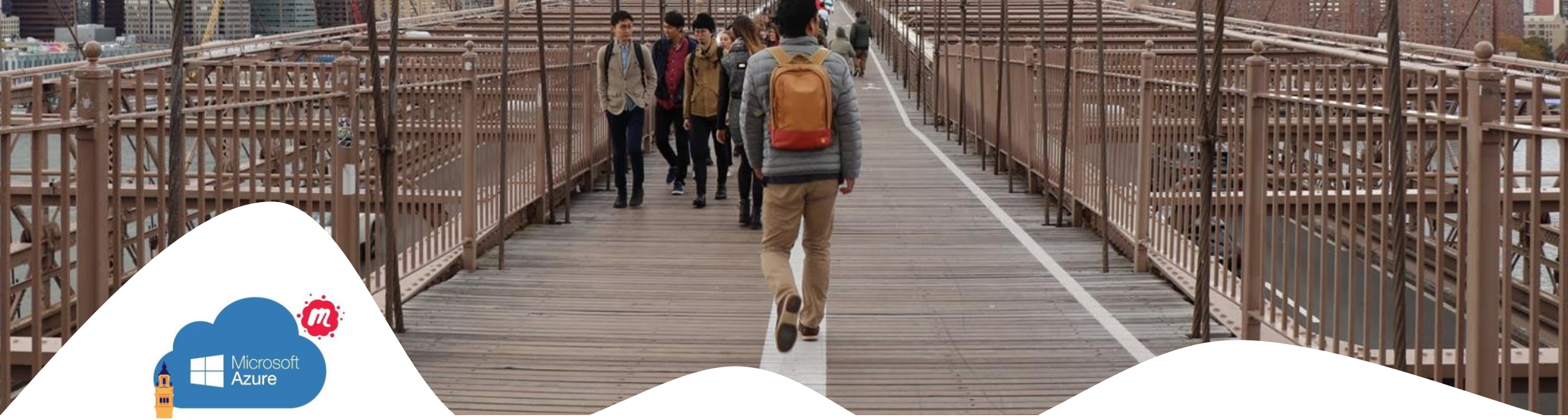


https://community.cloud.databricks.com

Fill the form

Your turn !

https://forms.office.com/r/iVdGaV70jz



Next Session: Data lake

Azure Data Lake Storage
Massively scalable, secure data lake functionality built on Azure Blob Storage

Thank you

PREMISEO
DATA, SYSTEMS, CLOUD

https://premiseo.com/

DataRedKite

dataredkite.com