datapath

Databricks Workspace: Ciencia e Ingeniería de Datos

www.datapath.ai

Encuesta de clase

[Databricks Data Engineer Associate (google.com)]

- Ingeniero Informático y de Sistemas (USMP)
- 6 años de experiencia
- Experiencia en empresas de sector Banca, Seguros, Consultoría de TI.
- Actualmente me desempeño como **Senior Data Engineer** en el equipo de YAPE

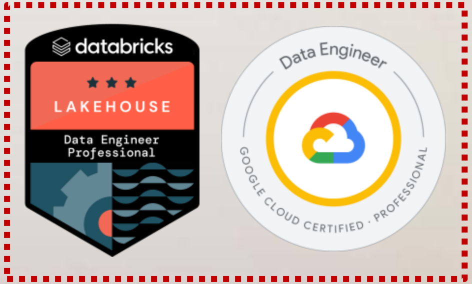
Jhon Rodriguez

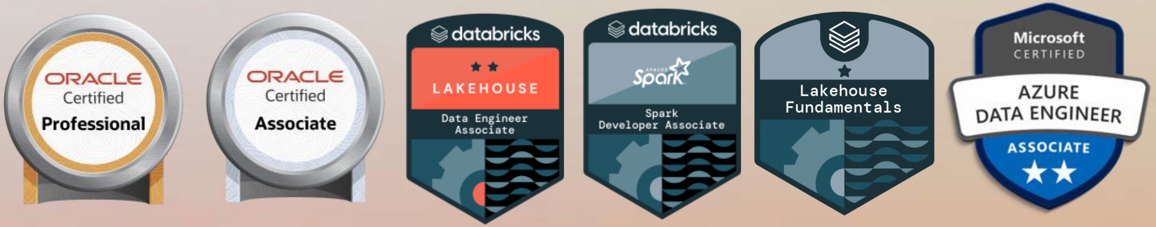
/jhonrodriguezb/

jerodriguez1608@gmail.com

www.datapath.ai

datapath





Objetivos de la clase

1. Comprender que es Databricks
2. Estar en la capacidad de describir la Arquitectura
3. Familiarizarnos con el Workspace de Ciencia e Ingeniería de Datos.

Agenda

1. Introducción
2. Qué es un Data Lakehouse
3. Qué es Databricks
4. Workloads

datapath

Reglas del Juego

- Mantener el micrófono apagado en caso no vayan a hablar.
- Preguntar en caso que tengan dudas
- Mantenerse atento a la clase.

Modo de evaluación

datapath

%

Evaluación continua

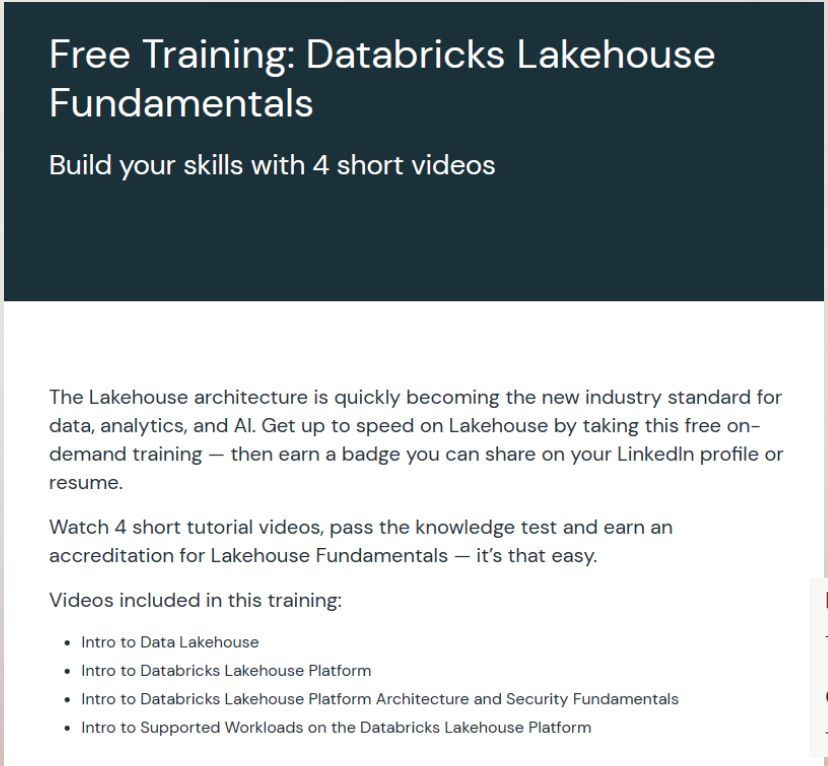
Ejercicios, challenges y/o test.

%

Examen Final

Caso con el cual se busca consolidar lo aprendido haciendo uso de las herramientas y aprendizajes obtenidos a lo largo del curso y/o programa.

Certified Lakehouse F.



Free Training: Databricks Lakehouse Fundamentals

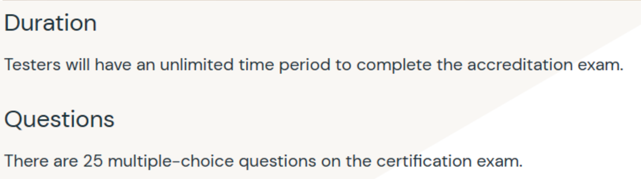
Build your skills with 4 short videos

The Lakehouse architecture is quickly becoming the new industry standard for data, analytics, and AI. Get up to speed on Lakehouse by taking this free on-demand training — then earn a badge you can share on your LinkedIn profile or resume.

Watch 4 short tutorial videos, pass the knowledge test and earn an accreditation for Lakehouse Fundamentals — it's that easy.

Videos included in this training:

- Intro to Data Lakehouse
- Intro to Databricks Lakehouse Platform
- Intro to Databricks Lakehouse Platform Architecture and Security Fundamentals
- Intro to Supported Workloads on the Databricks Lakehouse Platform



Duration

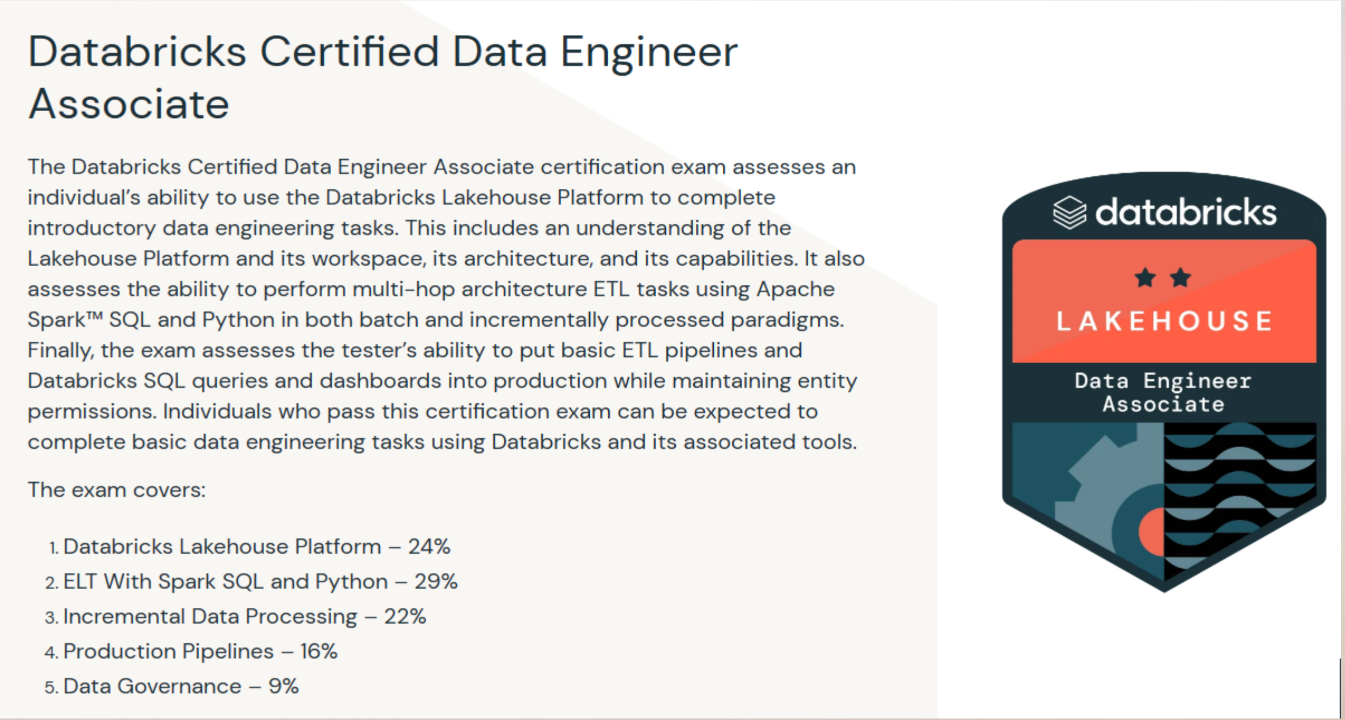
Testers will have an unlimited time period to complete the accreditation exam.

Questions

There are 25 multiple-choice questions on the certification exam.

www.datapath.ai

Certified Data Engineer



Databricks Certified Data Engineer Associate

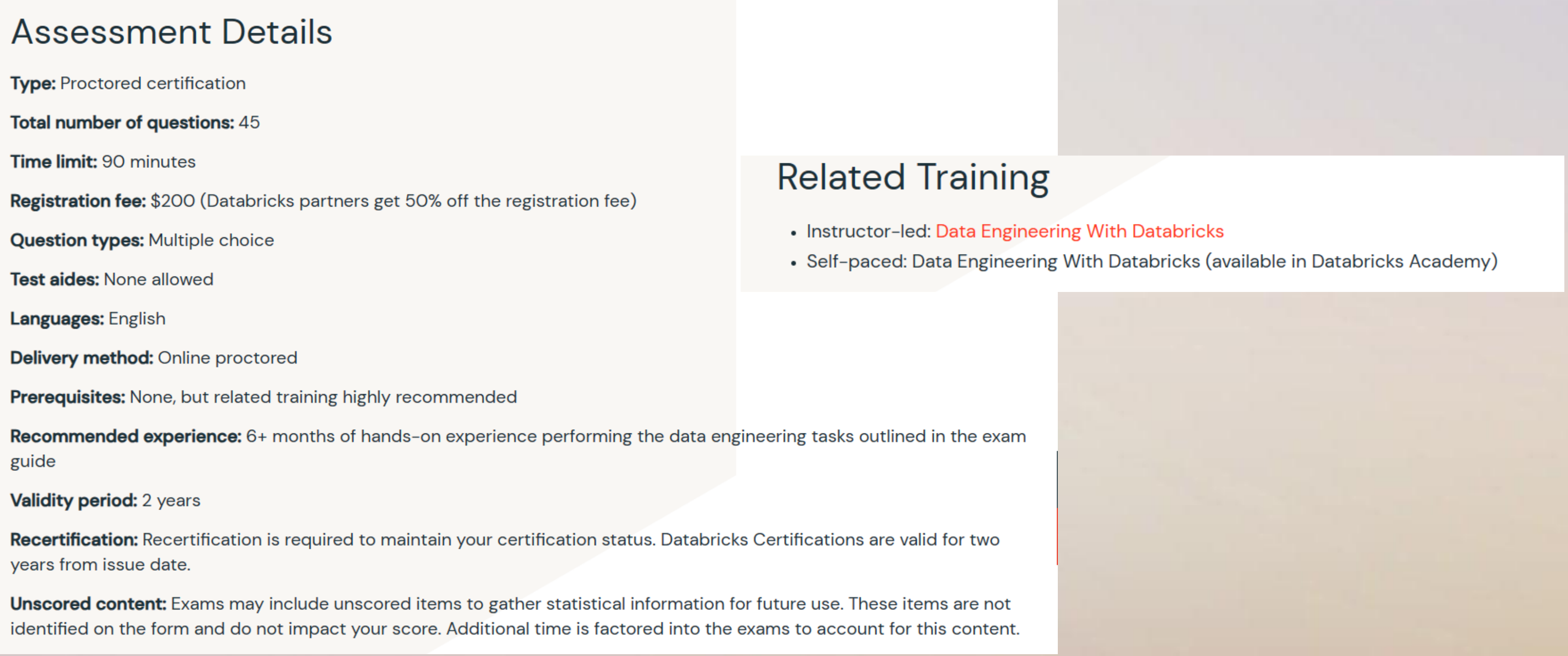
The Databricks Certified Data Engineer Associate certification exam assesses an individual's ability to use the Databricks Lakehouse Platform to complete introductory data engineering tasks. This includes an understanding of the Lakehouse Platform and its workspace, its architecture, and its capabilities. It also assesses the ability to perform multi-hop architecture ETL tasks using Apache Spark™ SQL and Python in both batch and incrementally processed paradigms. Finally, the exam assesses the tester's ability to put basic ETL pipelines and Databricks SQL queries and dashboards into production while maintaining entity permissions. Individuals who pass this certification exam can be expected to complete basic data engineering tasks using Databricks and its associated tools.

The exam covers:

1. Databricks Lakehouse Platform – 24%
2. ELT With Spark SQL and Python – 29%
3. Incremental Data Processing – 22%
4. Production Pipelines – 16%
5. Data Governance – 9%

www.datapath.ai

Getting Certified



Assessment Details

Type: Proctored certification

Total number of questions: 45

Time limit: 90 minutes

Registration fee: $200 (Databricks partners get 50% off the registration fee)

Question types: Multiple choice

Test aides: None allowed

Languages: English

Delivery method: Online proctored

Prerequisites: None, but related training highly recommended

Recommended experience: 6+ months of hands-on experience performing the data engineering tasks outlined in the exam guide

Validity period: 2 years

Recertification: Recertification is required to maintain your certification status. Databricks Certifications are valid for two years from issue date.

Unscored content: Exams may include unscored items to gather statistical information for future use. These items are not identified on the form and do not impact your score. Additional time is factored into the exams to account for this content.

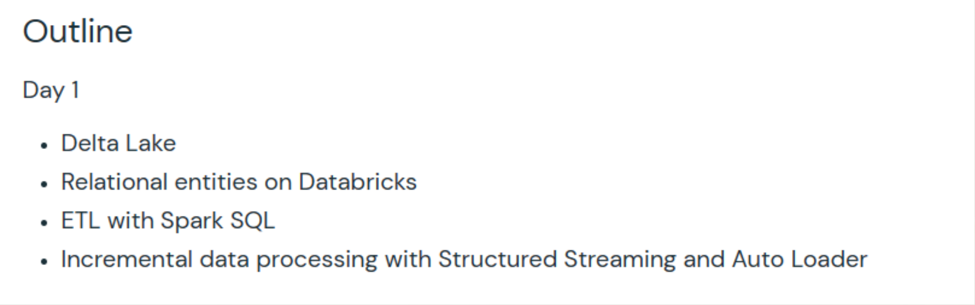
Related Training

- Instructor-led: Data Engineering With Databricks
- Self-paced: Data Engineering With Databricks (available in Databricks Academy)

www.datapath.ai

Databricks Training

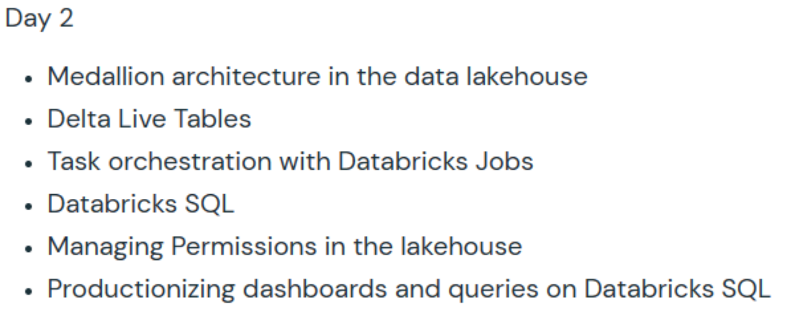
datapath



Outline

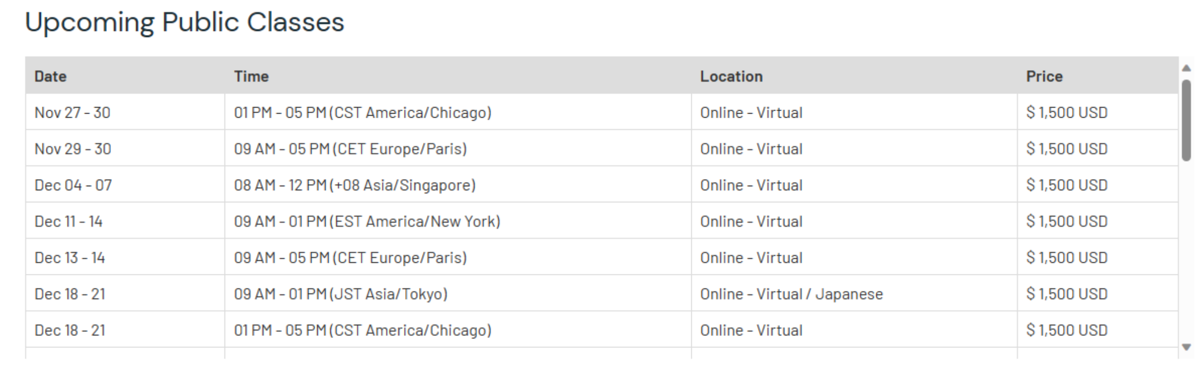
Day 1

- Delta Lake
- Relational entities on Databricks
- ETL with Spark SQL
- Incremental data processing with Structured Streaming and Auto Loader



Day 2

- Medallion architecture in the data lakehouse
- Delta Live Tables
- Task orchestration with Databricks Jobs
- Databricks SQL
- Managing Permissions in the lakehouse
- Productionizing dashboards and queries on Databricks SQL



Upcoming Public Classes

Date	Time	Location	Price
Nov 27 - 30	01 PM - 05 PM (CST America/Chicago)	Online - Virtual	$ 1,500 USD
Nov 29 - 30	09 AM - 05 PM (CET Europe/Paris)	Online - Virtual	$ 1,500 USD
Dec 04 - 07	08 AM - 12 PM (+08 Asia/Singapore)	Online - Virtual	$ 1,500 USD
Dec 11 - 14	09 AM - 01 PM (EST America/New York)	Online - Virtual	$ 1,500 USD
Dec 13 - 14	09 AM - 05 PM (CET Europe/Paris)	Online - Virtual	$ 1,500 USD
Dec 18 - 21	09 AM - 01 PM (JST Asia/Tokyo)	Online - Virtual / Japanese	$ 1,500 USD
Dec 18 - 21	01 PM - 05 PM (CST America/Chicago)	Online - Virtual	$ 1,500 USD

www.datapath.ai

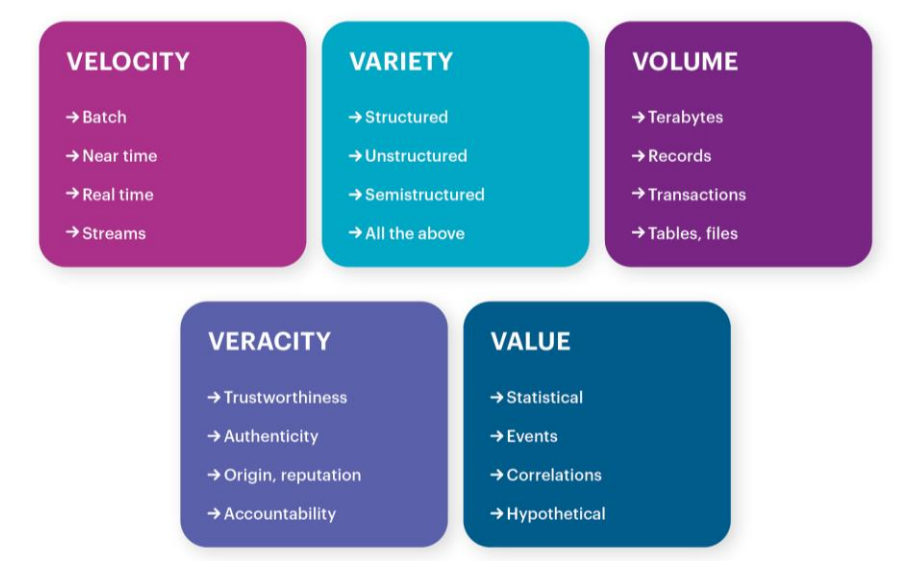
Introducción

Big Data

El Big Data consiste en un proceso que analiza e interpreta **grandes volúmenes de datos**, tanto **estructurados** como no **estructurados**. El Big Data sirve para que los datos almacenados de forma remota puedan ser utilizados por las empresas como base para su toma de decisiones.

datapath

Las 5 V's del Big Data



VELOCITY

- → Batch
- → Near time
- → Real time
- → Streams

VARIETY

- → Structured
- → Unstructured
- → Semistructured
- → All the above

VOLUME

- → Terabytes
- → Records
- → Transactions
- → Tables, files

VERACITY

- → Trustworthiness
- → Authenticity
- → Origin, reputation
- → Accountability

VALUE

- → Statistical
- → Events
- → Correlations
- → Hypothetical

www.datapath.ai

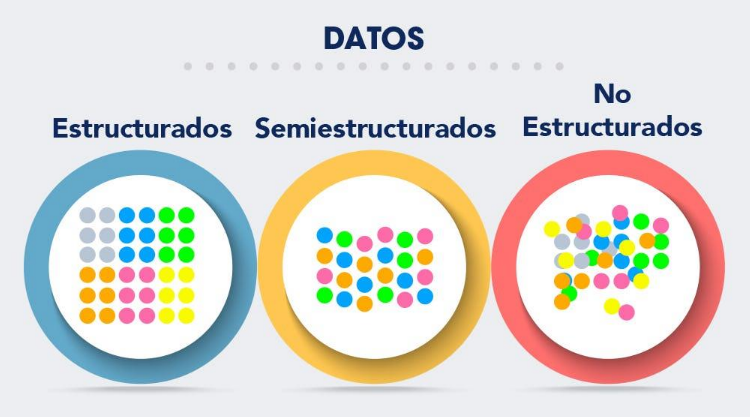
datapath

Tipos de datos

1). Estructurado: Libros Excel, CSV, Tablas BD relacional. (Filas y Columnas)

2). Semiestructurado: Archivos JSON, Paginas Web, XML (Jerarquías)

3). No Estructurados: Videos, imágenes, audio, PDF.



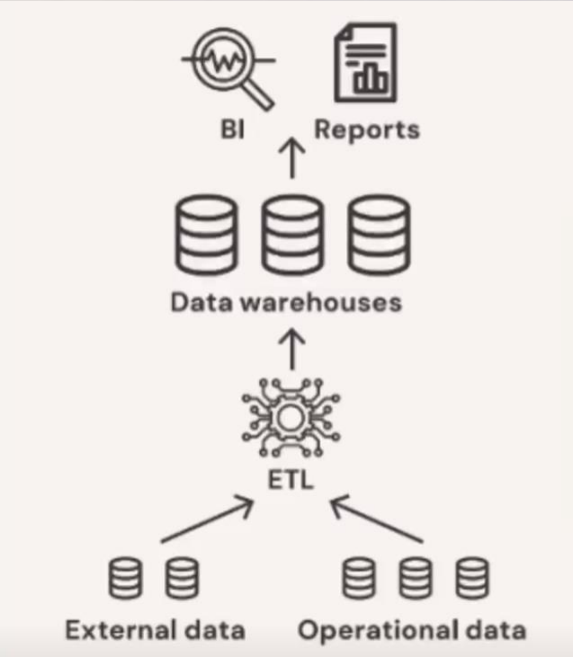
Data Warehouse (1980s)

Ventajas:

- Business Intelligence
- Tareas analíticas
- Datos limpios y estructurados
- Esquemas predefinidos

Desventajas:

- No soporta dato semiestructurados y no estructurados.
- Esquemas flexibles
- Alto tiempo de procesamiento



datapath

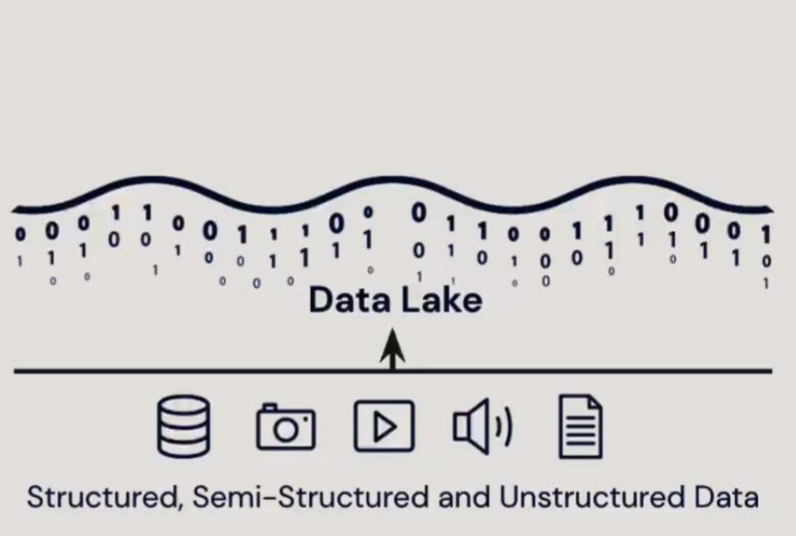
Data Lake (2000)

Ventajas:

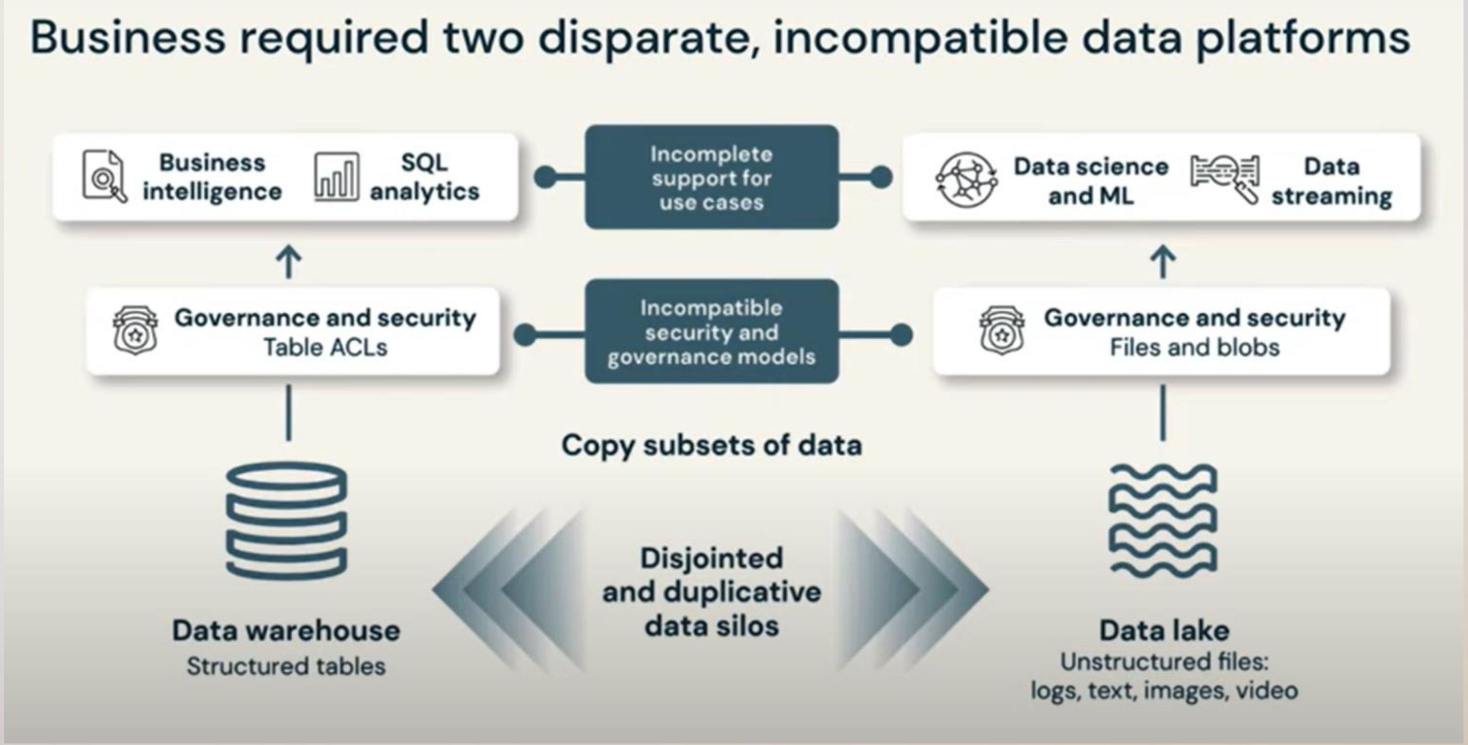
- Almacenamiento de datos flexible
- Soporte para cargas Streaming
- Eficiente costo en la nube
- Soporte para proyectos IA y ML.

Desventajas:

- No soporta transacciones
- Baja credibilidad en los datos
- Bajo rendimiento para tareas de análisis
- Problemas en el gobierno de datos



www.datapath.ai



Business required two disparate, incompatible data platforms

Business intelligence

SQL analytics

Incomplete support for use cases

Data science and ML

Data streaming

Governance and security
Table ACLs

Incompatible security and governance models

Governance and security
Files and blobs

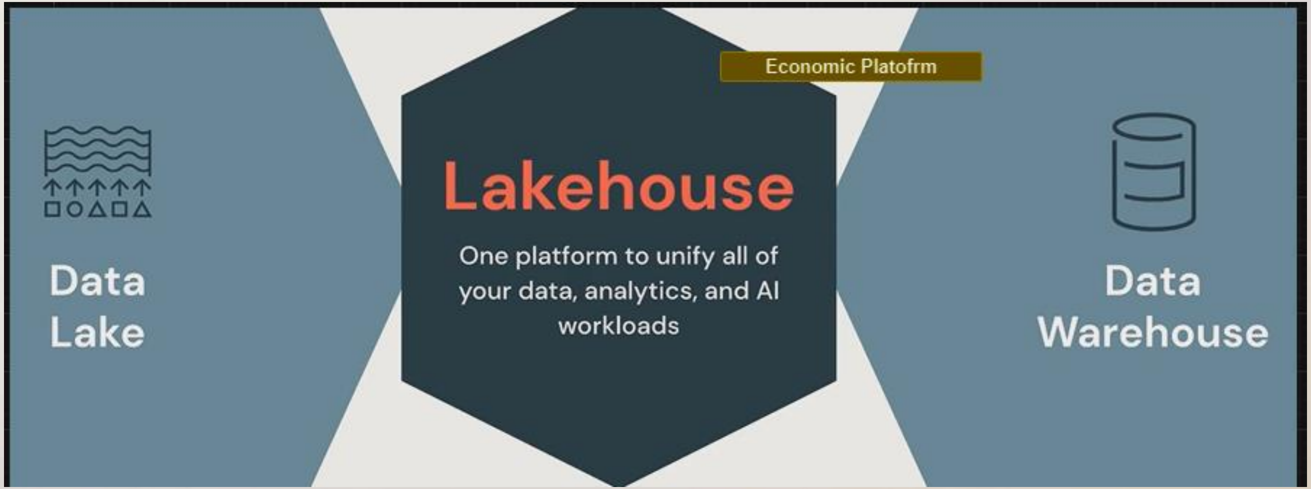
Copy subsets of data

Data warehouse
Structured tables

Disjointed and duplicative data silos

Data lake
Unstructured files:
logs, text, images, video

Data Lakehouse



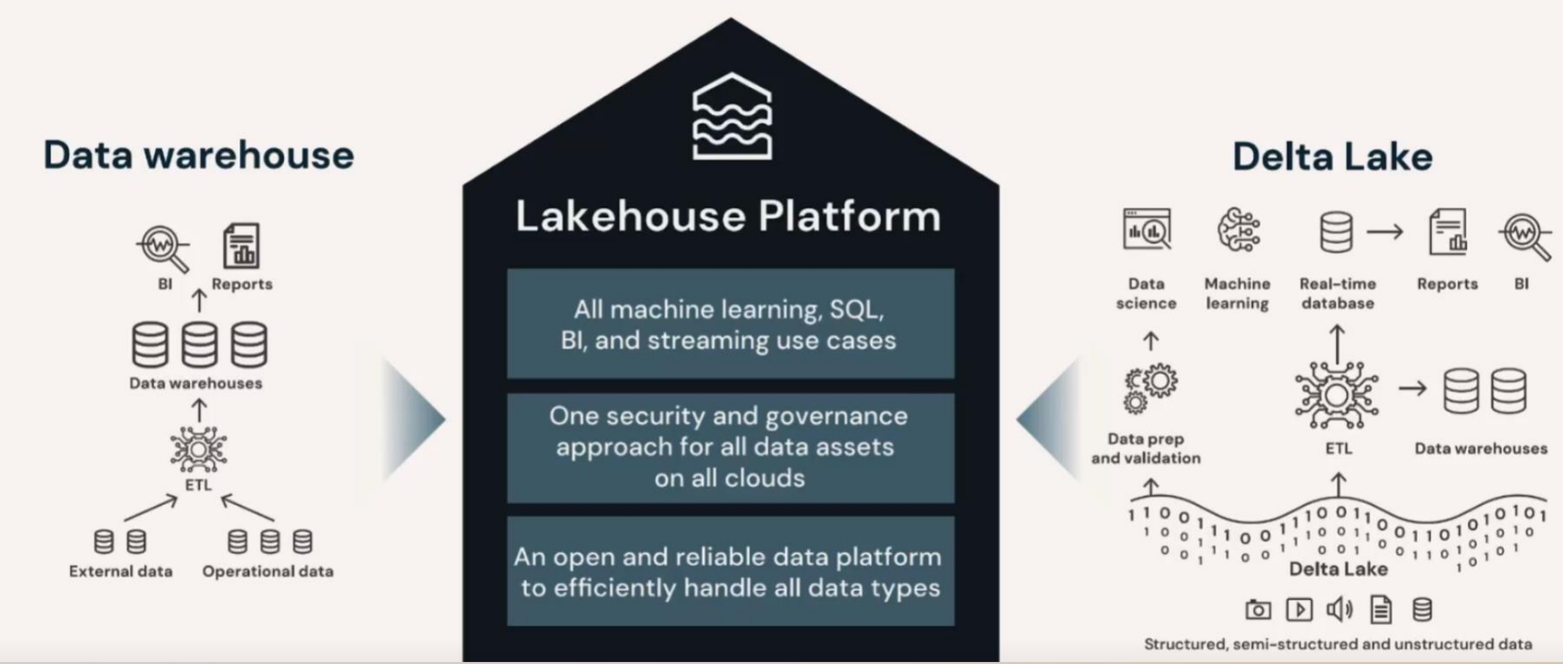
Economic Platofrm

Lakehouse

One platform to unify all of your data, analytics, and AI workloads

www.datapath.ai

Data Lakehouse



Data warehouse

BI

Reports

Data warehouses

ETL

External data

Operational data

Lakehouse Platform

All machine learning, SQL, BI, and streaming use cases

One security and governance approach for all data assets on all clouds

An open and reliable data platform to efficiently handle all data types

Delta Lake

Data science

Machine learning

Real-time database

Reports

BI

Data prep and validation

ETL

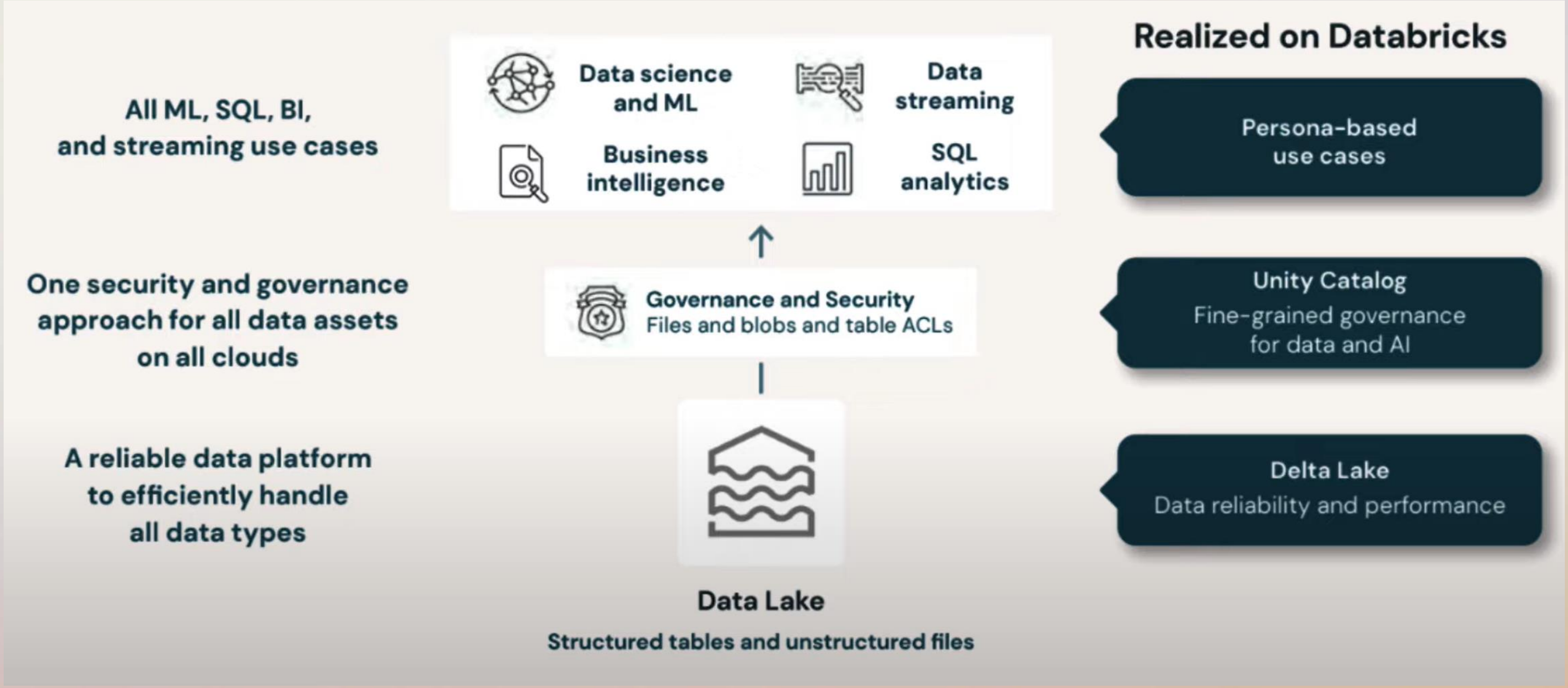
Data warehouses

Delta Lake

Structured, semi-structured and unstructured data

datapath

Data Lakehouse



All ML, SQL, BI, and streaming use cases

One security and governance approach for all data assets on all clouds

A reliable data platform to efficiently handle all data types

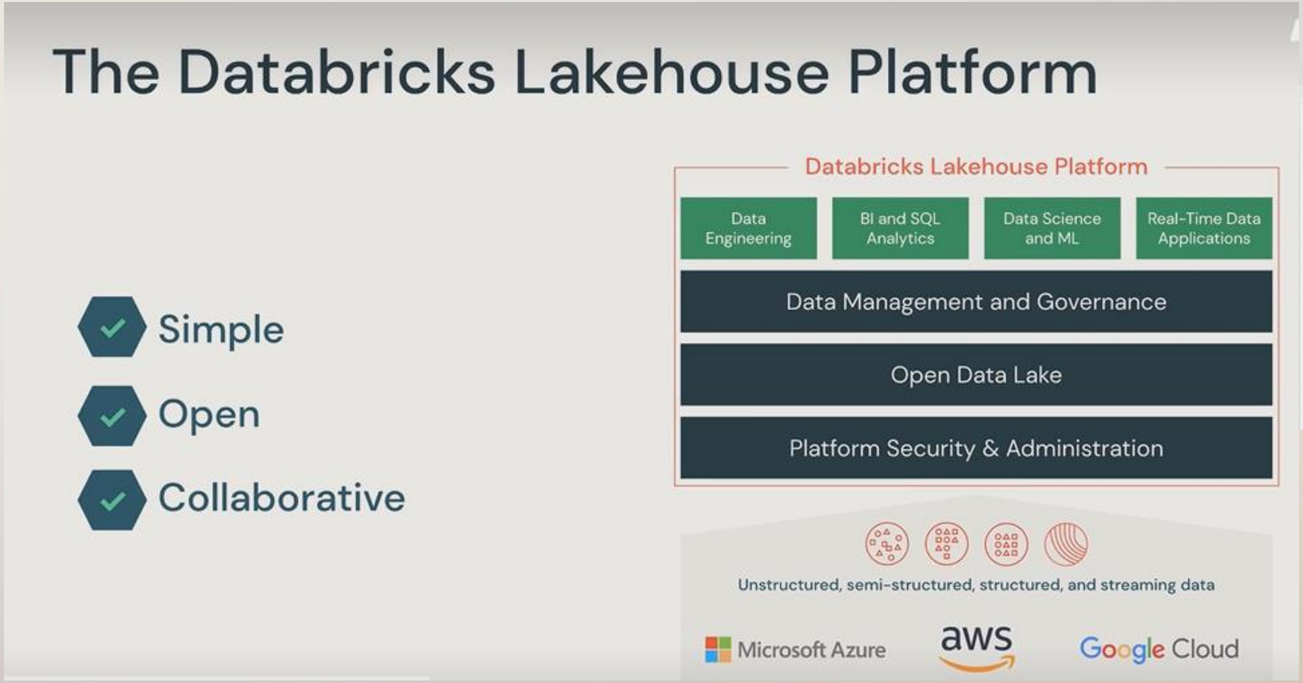
Realized on Databricks

Persona-based use cases

Unity Catalog
Fine-grained governance for data and AI

Delta Lake
Data reliability and performance

Data Lakehouse



The Databricks Lakehouse Platform

- ✓ Simple
- ✓ Open
- ✓ Collaborative

Databricks Lakehouse Platform

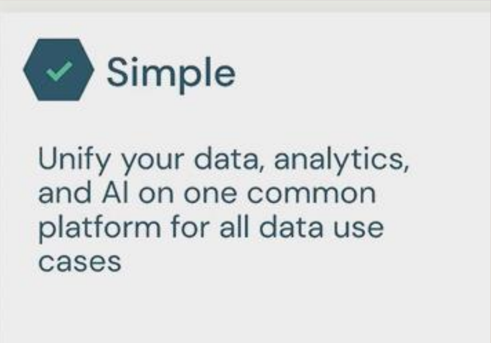
Data Engineering	BI and SQL Analytics	Data Science and ML	Real-Time Data Applications

Data Management and Governance

Open Data Lake

Platform Security & Administration

Unstructured, semi-structured, structured, and streaming data

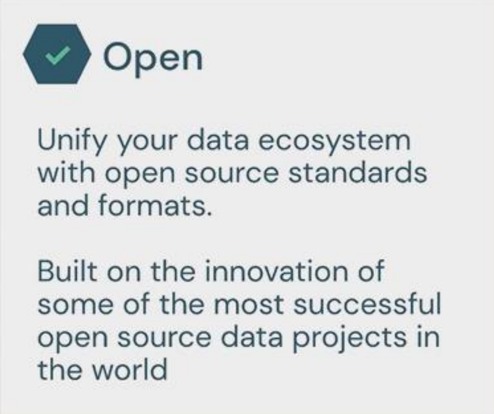


✓ Simple

Unify your data, analytics, and AI on one common platform for all data use cases

datapath

Data Lakehouse



Open

Unify your data ecosystem with open source standards and formats.

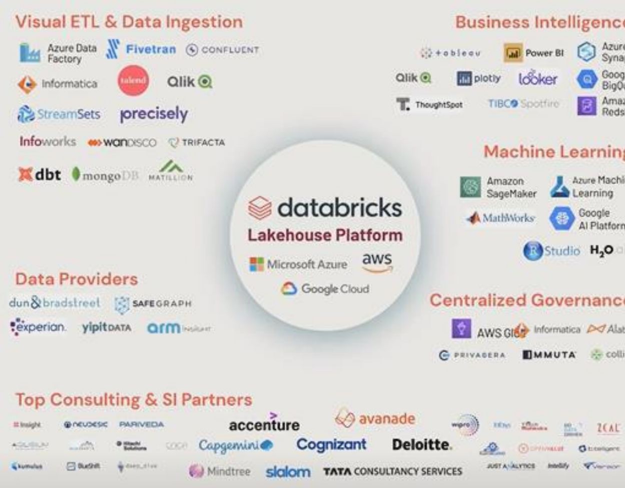
Built on the innovation of some of the most successful open source data projects in the world

30 Million+

Monthly downloads

450+

Partners across the data landscape



Visual ETL & Data Ingestion

Azure Data Factory, Fivetran, CONFLUENT, Informatica, talend, Qlik, StreamSets, precisely, Infoworks, wanDISCO, TRIFACTA, dbt, mongoDB, MATILLION

Business Intelligenc

tableau, Power BI, Azure Syna, Qlik, plotly, looker, Goog BigQ, ThoughtSpot, TIBCO Spotfire, Amaz Reds

Machine Learnin

Amazon SageMaker, Azure Machi Learning, MathWorks, Google AI Platform, RStudio, H2O

databricks

Lakehouse Platform

Microsoft Azure, aws, Google Cloud

Data Providers

dun&bradstreet, SAFEGRAPH, experian, yipitDATA, arm insight

Centralized Governanc

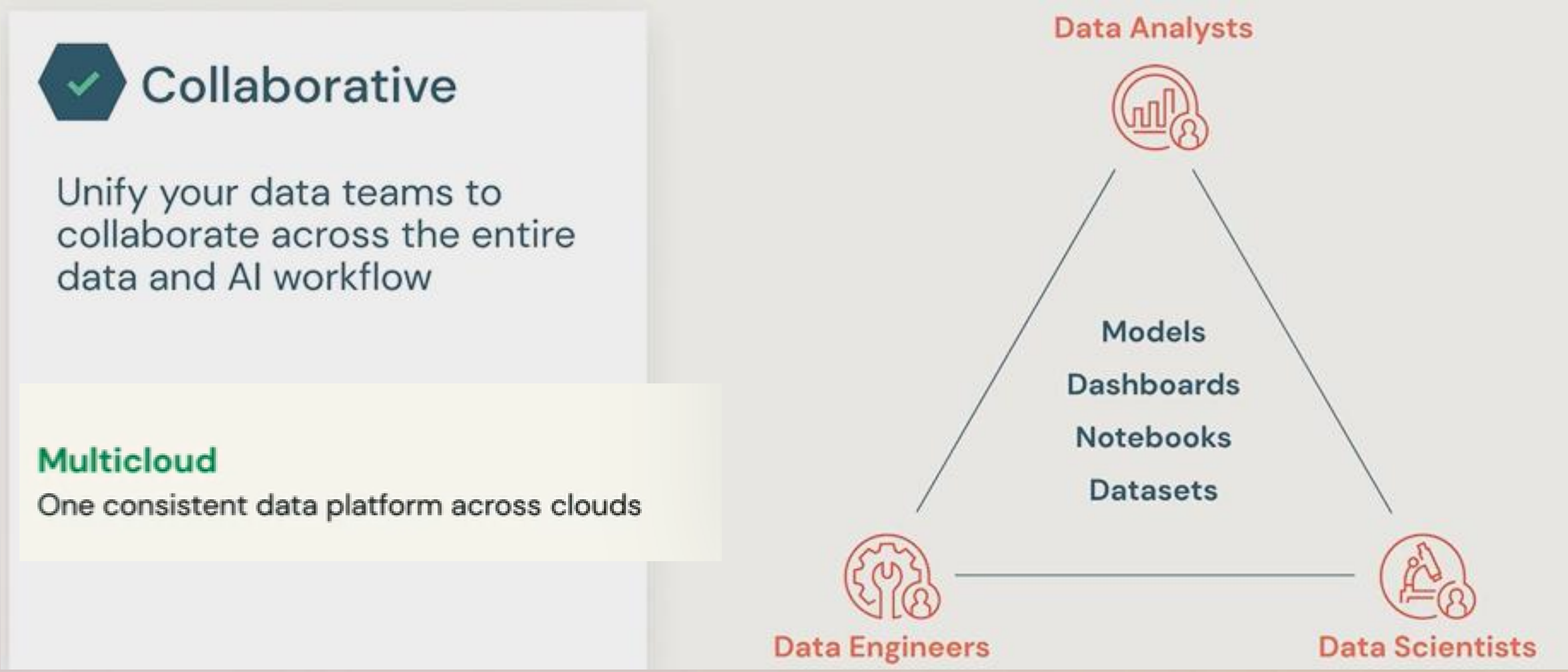
AWS Glue, Informatica, Alat, PRIVACERA, IMMUTA, colli

Top Consulting & SI Partners

Insight, neudesic, PARIVEDA, accenture, avanade, wipro, Capgemini, Cognizant, Deloitte., kumulus, BlueShift, Mindtree, slalom, TATA CONSULTANCY SERVICES, JUST ANALYTICS, Intellify

www.datapath.ai

Data Lakehouse



Collaborative

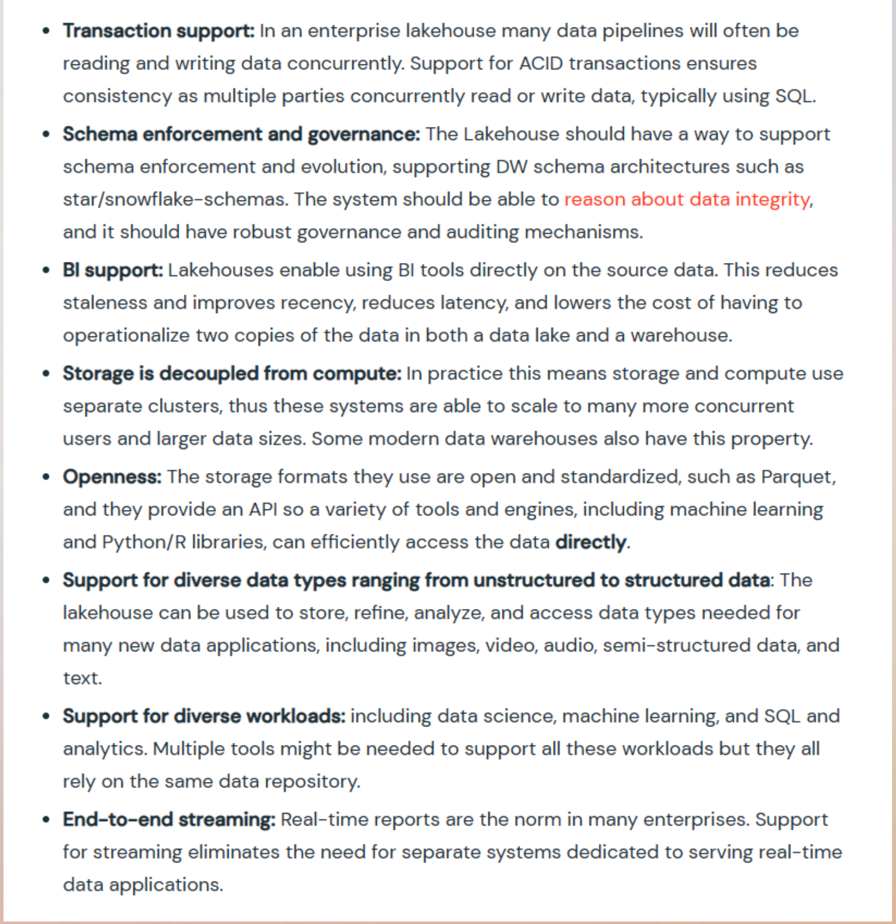
Unify your data teams to collaborate across the entire data and AI workflow

Multicloud

One consistent data platform across clouds

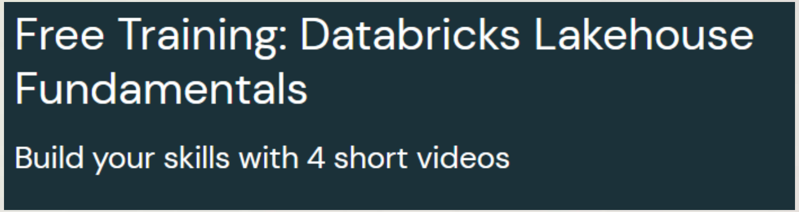
Data Lakehouse

- Soporta Transacciones
- Gobierno de datos
- Soporta BI
- Desacopla el almacenamiento y procesamiento.
- Soporta diversos formatos de almacenamiento.
- Posee diversos Workloads
- End-to-end streaming.



- **Transaction support:** In an enterprise lakehouse many data pipelines will often be reading and writing data concurrently. Support for ACID transactions ensures consistency as multiple parties concurrently read or write data, typically using SQL.
- **Schema enforcement and governance:** The Lakehouse should have a way to support schema enforcement and evolution, supporting DW schema architectures such as star/snowflake-schemas. The system should be able to reason about data integrity, and it should have robust governance and auditing mechanisms.
- **BI support:** Lakehouses enable using BI tools directly on the source data. This reduces staleness and improves recency, reduces latency, and lowers the cost of having to operationalize two copies of the data in both a data lake and a warehouse.
- **Storage is decoupled from compute:** In practice this means storage and compute use separate clusters, thus these systems are able to scale to many more concurrent users and larger data sizes. Some modern data warehouses also have this property.
- **Openness:** The storage formats they use are open and standardized, such as Parquet, and they provide an API so a variety of tools and engines, including machine learning and Python/R libraries, can efficiently access the data **directly**.
- **Support for diverse data types ranging from unstructured to structured data**: The lakehouse can be used to store, refine, analyze, and access data types needed for many new data applications, including images, video, audio, semi-structured data, and text.
- **Support for diverse workloads:** including data science, machine learning, and SQL and analytics. Multiple tools might be needed to support all these workloads but they all rely on the same data repository.
- **End-to-end streaming:** Real-time reports are the norm in many enterprises. Support for streaming eliminates the need for separate systems dedicated to serving real-time data applications.

Lakehouse certified

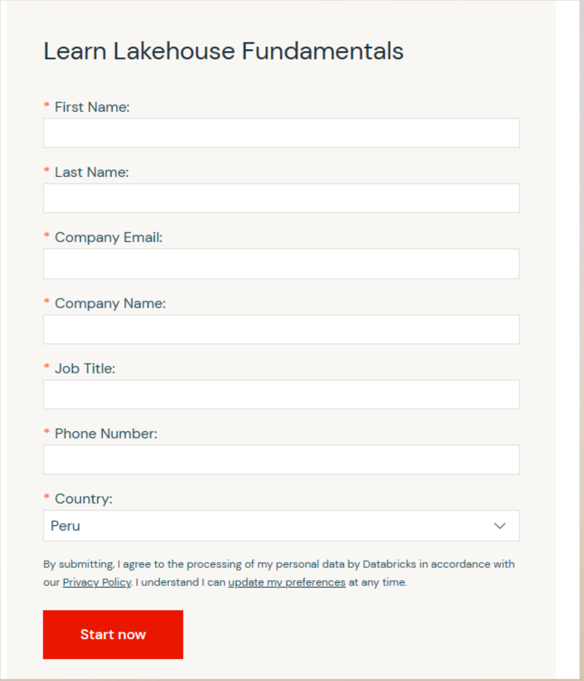


Free Training: Databricks Lakehouse Fundamentals

Build your skills with 4 short videos

[Lakehouse Fundamentals | Databricks]

Enviar un resumen por cada video en un Word , cualquier pregunta sobre algún tema que encuentren en el video, lo revisamos la siguiente clase



Learn Lakehouse Fundamentals

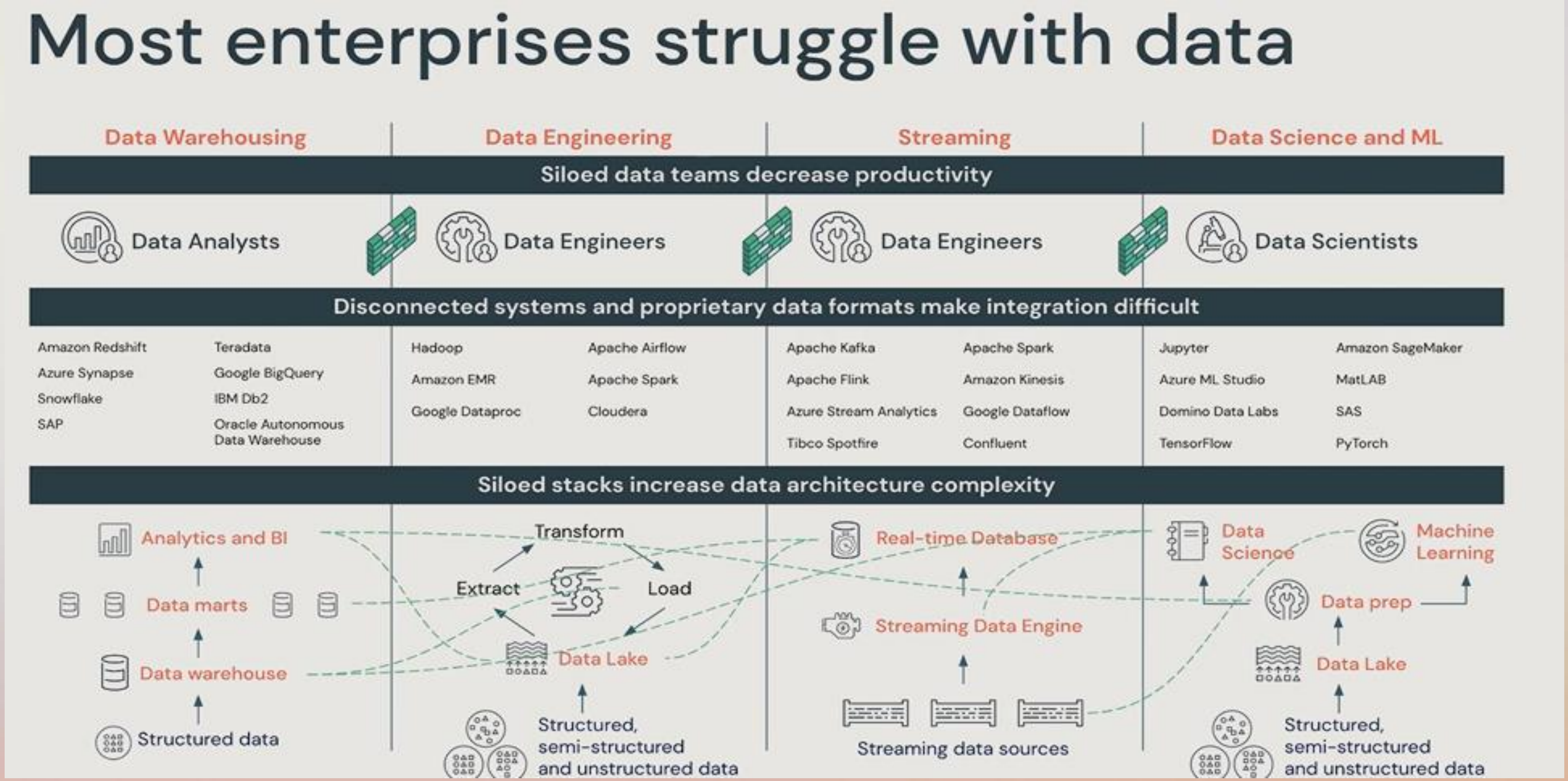
* First Name:	
* Last Name:	
* Company Email:	
* Company Name:	
* Job Title:	
* Phone Number:	
* Country:	Peru

By submitting, I agree to the processing of my personal data by Databricks in accordance with our Privacy Policy. I understand I can update my preferences at any time.

www.datapath.ai

datapath

Data Lakehouse

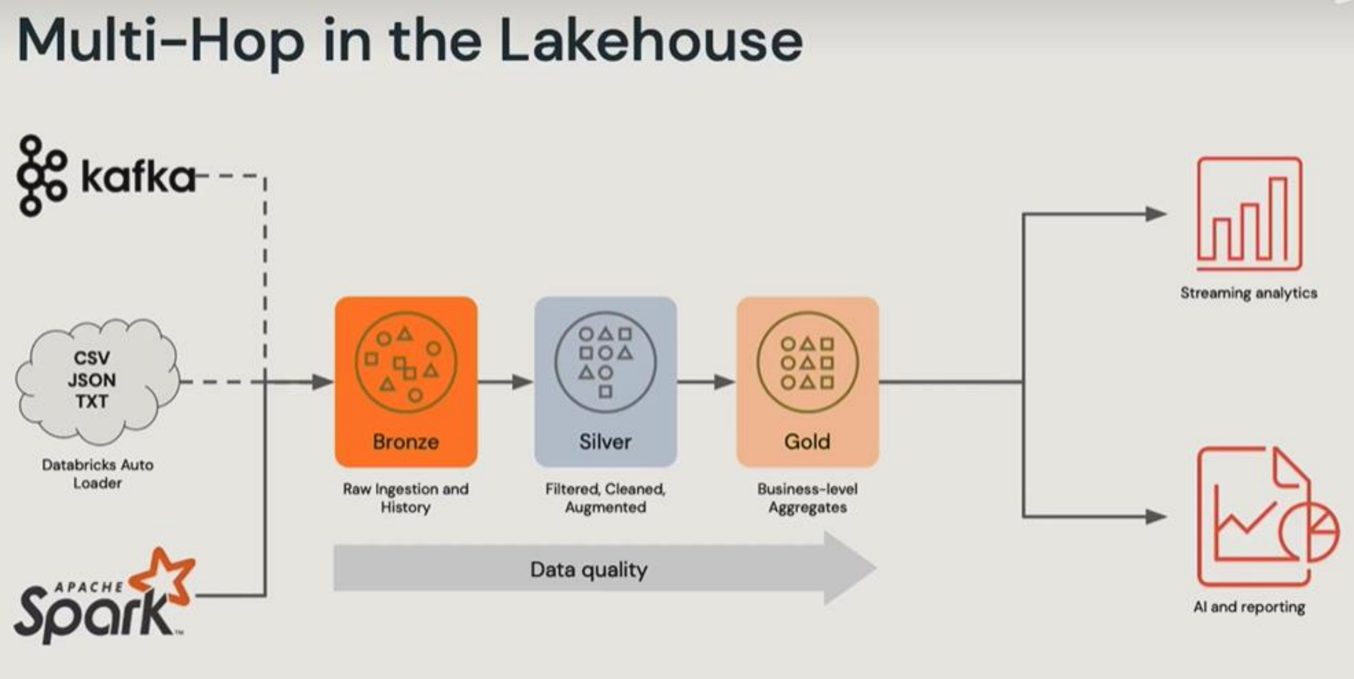


Most enterprises struggle with data

Data Warehousing		Data Engineering		Streaming		Data Science and ML	
Siloed data teams decrease productivity							
Data Analysts		Data Engineers		Data Engineers		Data Scientists	
Disconnected systems and proprietary data formats make integration difficult							
Amazon Redshift	Teradata	Hadoop	Apache Airflow	Apache Kafka	Apache Spark	Jupyter	Amazon SageMaker
Azure Synapse	Google BigQuery	Amazon EMR	Apache Spark	Apache Flink	Amazon Kinesis	Azure ML Studio	MatLAB
Snowflake	IBM Db2	Google Dataproc	Cloudera	Azure Stream Analytics	Google Dataflow	Domino Data Labs	SAS
SAP	Oracle Autonomous Data Warehouse			Tibco Spotfire	Confluent	TensorFlow	PyTorch

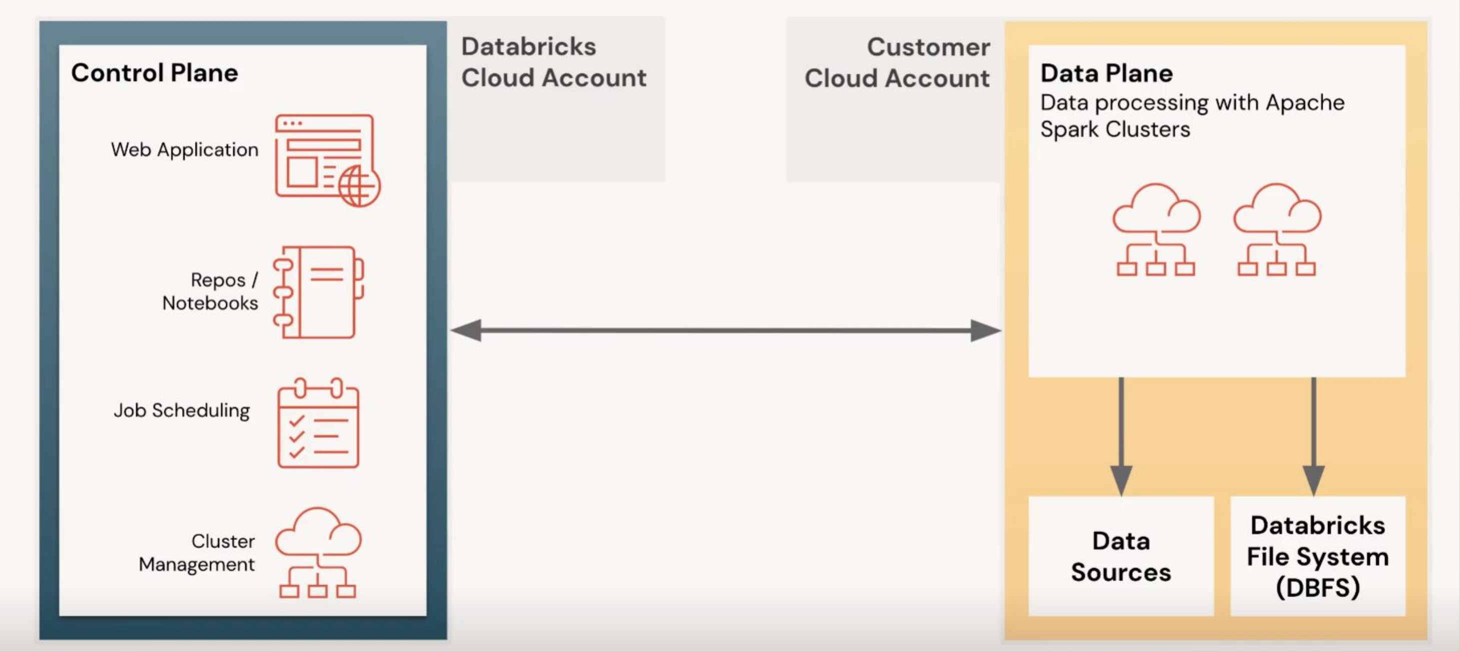
Siloed stacks increase data architecture complexity

Arquitectura de Databricks



Multi-Hop in the Lakehouse

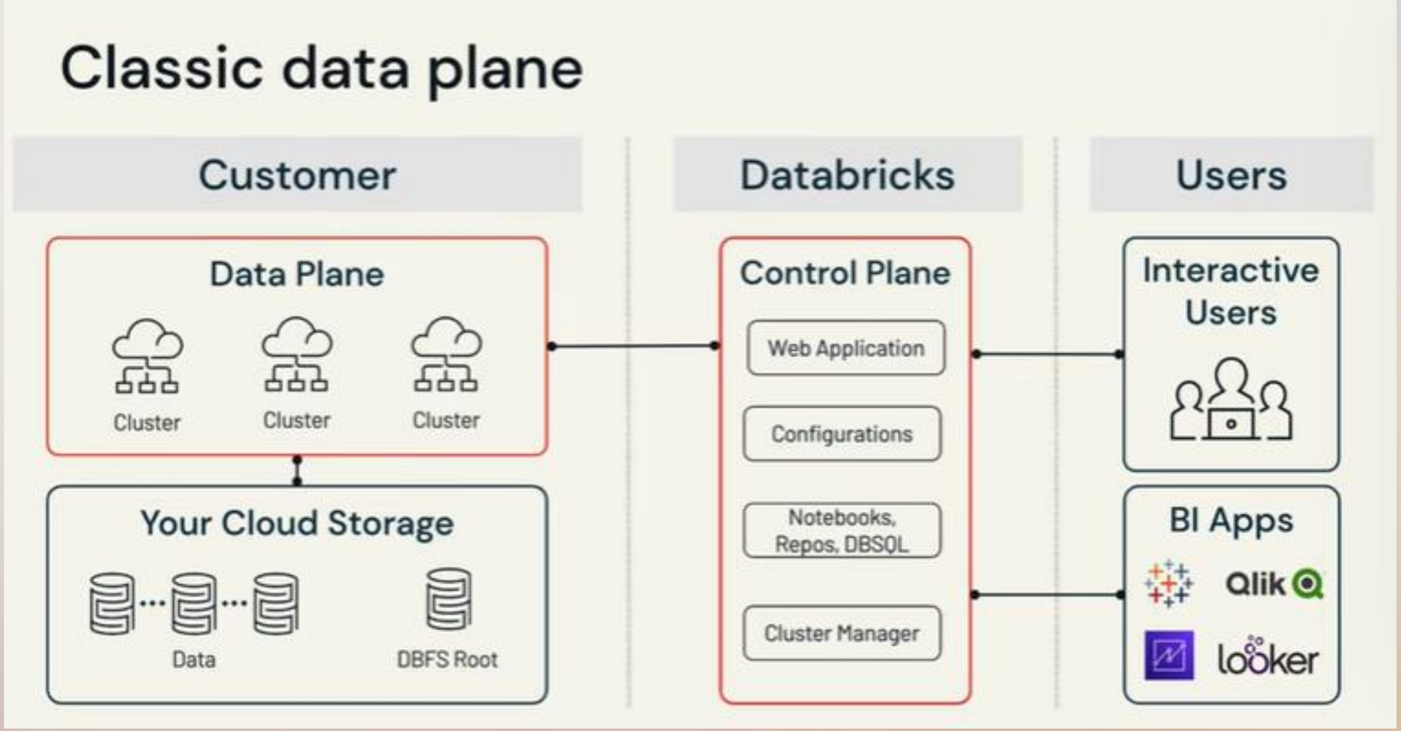
Arquitectura de Databricks



www.datapath.ai

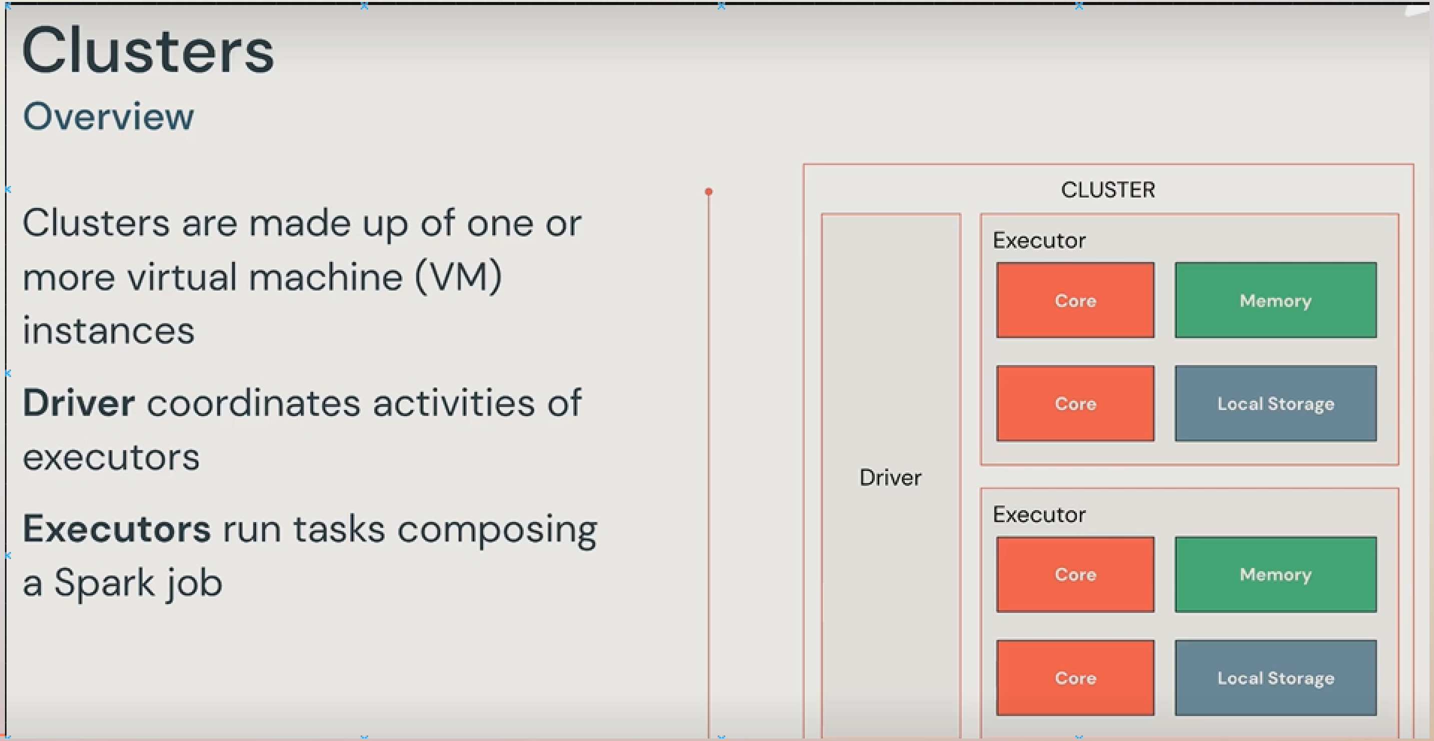
datapath

Arquitectura de Databricks



www.datapath.ai

Clúster



Clusters

Overview

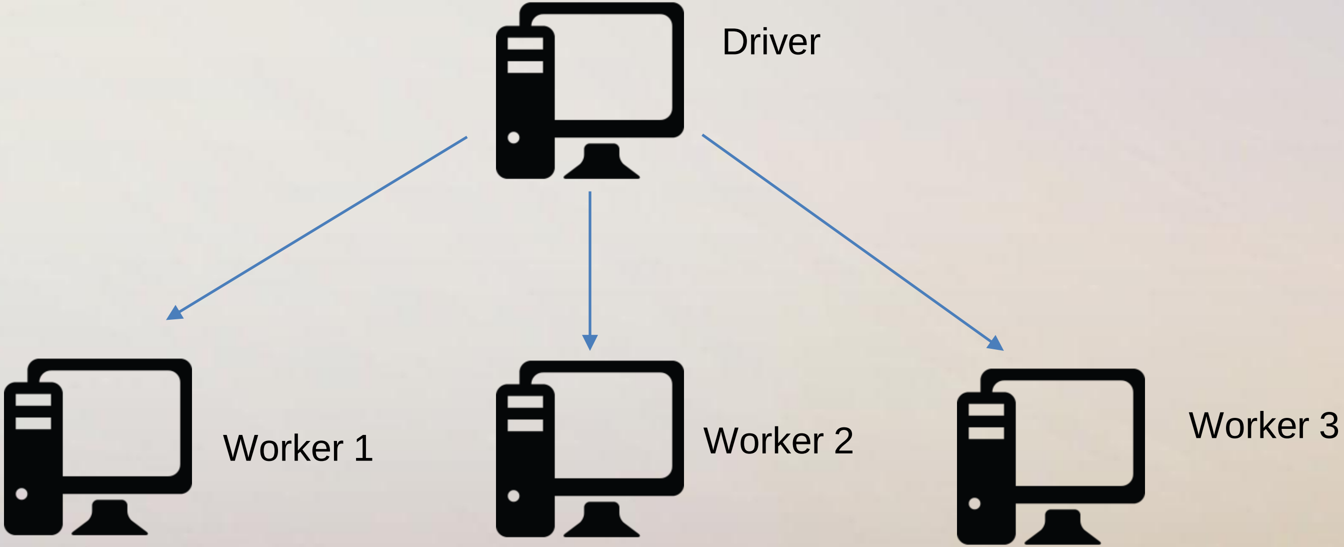
Clusters are made up of one or more virtual machine (VM) instances

Driver coordinates activities of executors

Executors run tasks composing a Spark job

www.

Clúster



Tipos de Cluster

1. **All-purpose Cluster**

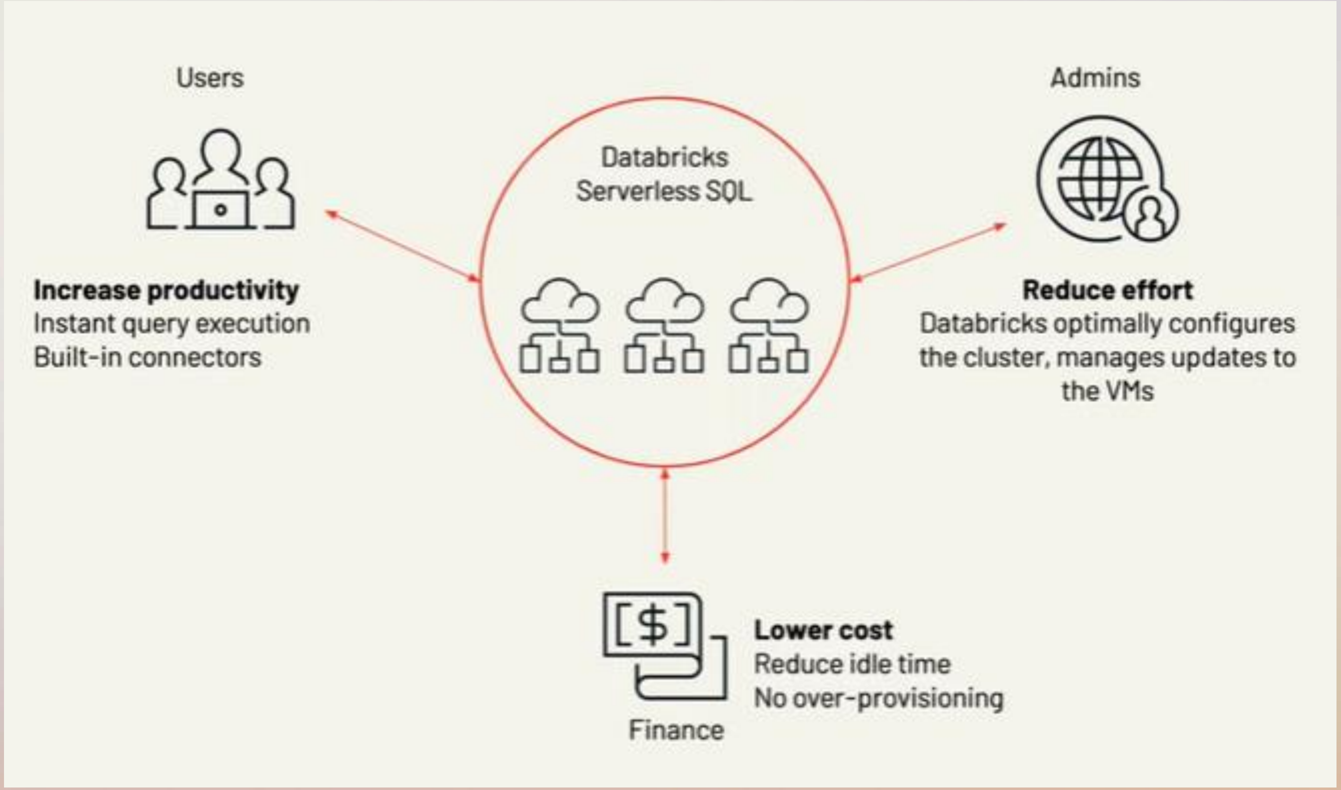
- Analiza los datos de manera colaborativa usando notebooks interactivos
- Crea Clusters desde el Workspace o API.
- Retiene hasta 70 clusters por hasta 30 días.

2. **Job Clusters**

- Ejecuta Jobs automatizados
- El programador de Jobs crea job clusters para la ejecución.
- Retiene hasta 30 clusters.

datapath

Arquitectura de Databricks



Users

Increase productivity
Instant query execution
Built-in connectors

Databricks
Serverless SQL

Admins

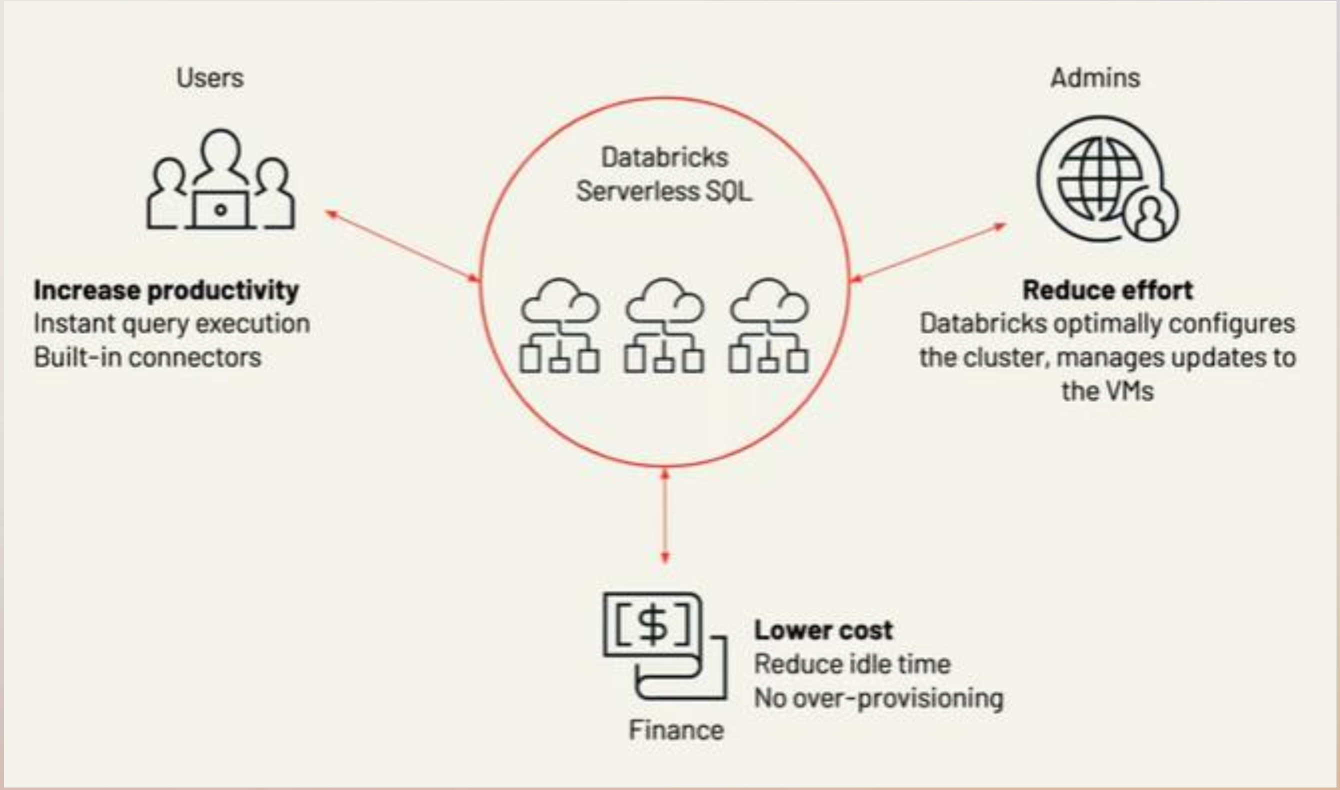
Reduce effort
Databricks optimally configures the cluster, manages updates to the VMs

Lower cost
Reduce idle time
No over-provisioning

Finance

www.datapath.ai

Arquitectura de Databricks



Users

Increase productivity
Instant query execution
Built-in connectors

Databricks
Serverless SQL

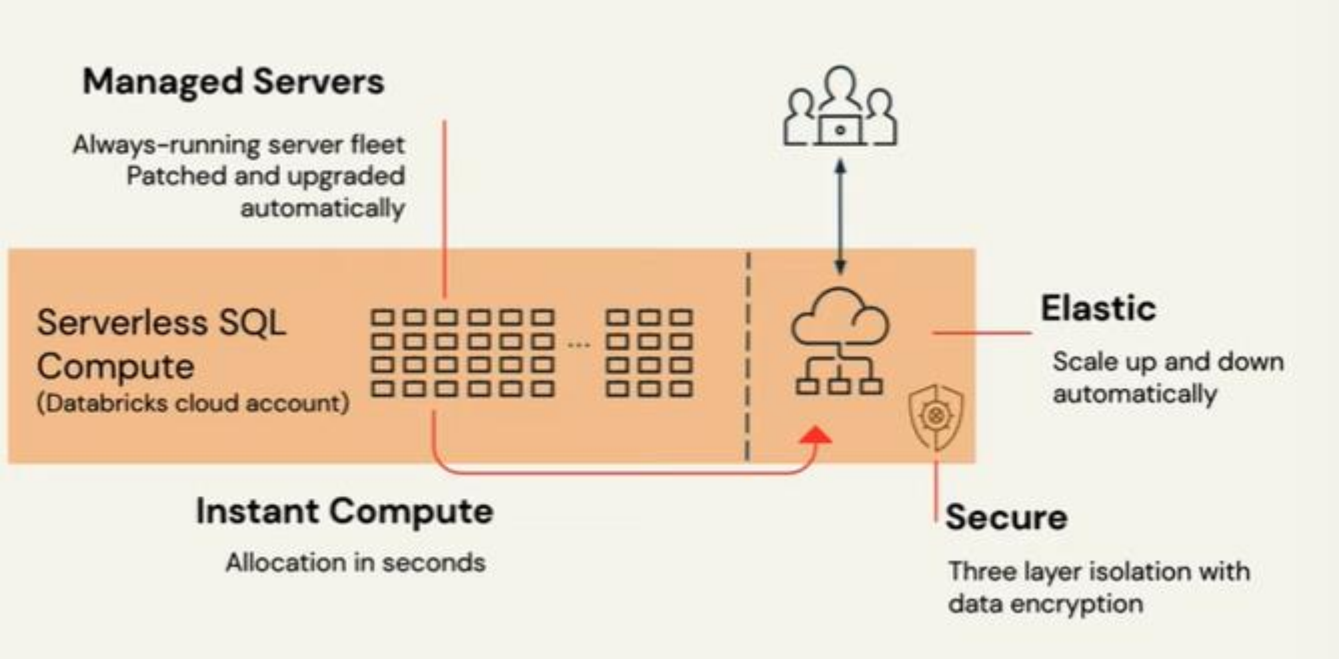
Admins

Reduce effort
Databricks optimally configures the cluster, manages updates to the VMs

Finance

Lower cost
Reduce idle time
No over-provisioning

Arquitectura de Databricks



Managed Servers

Always-running server fleet
Patched and upgraded automatically

Serverless SQL Compute
(Databricks cloud account)

Elastic

Scale up and down automatically

Instant Compute

Allocation in seconds

Secure

Three layer isolation with data encryption

datapath

Laboratorio 1

1. Creación de cuentas Azure.
2. Desplegar un Databricks Workspace.
3. Recursos de cómputo (Clusters)
4. Desarrollo de código de con los Notebooks de Databricks
5. Repositorios.

datapath

¿PREGUNTAS?

Aprende, aplica y crece

www.datapath.ai