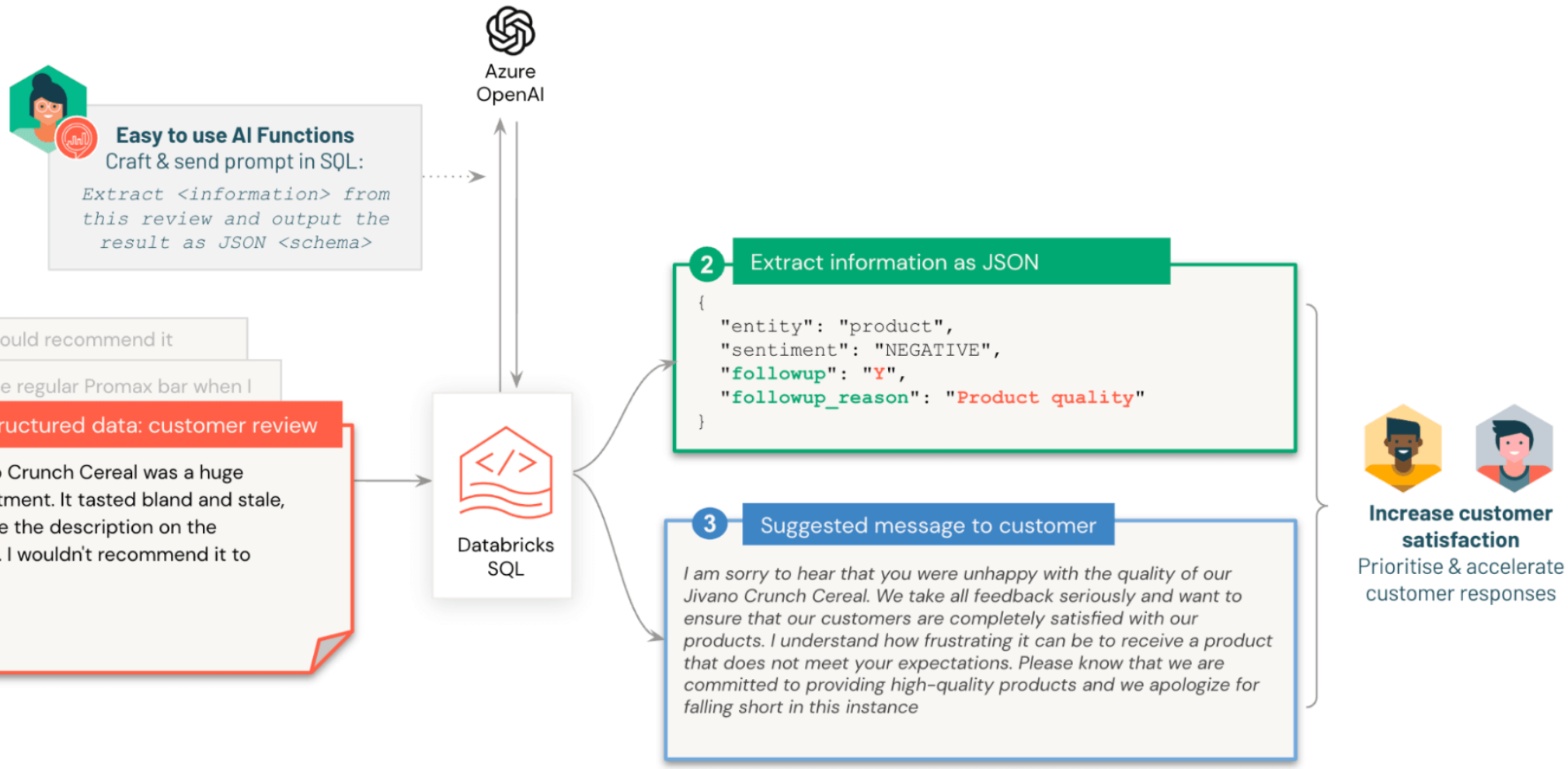
Apache Spark & Databricks

databricks



Apache Spark & Databricks

databricks

1. Apache Spark

Apache Spark

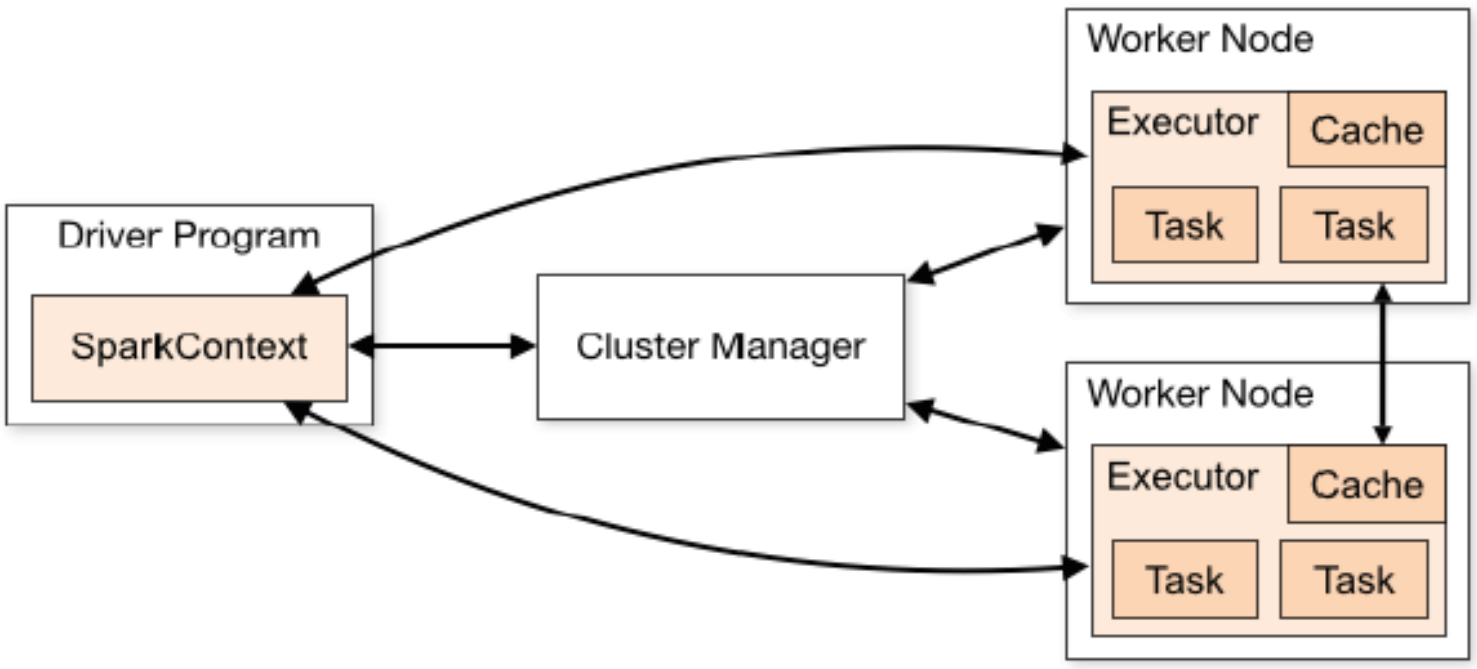
- **Sistema de procesamiento de datos distribuído de código libre**
  - Almacena los datos distribuidos en la **memoria RAM** de un clúster de nodos
  - Combina programación (Python o Scala) y estructuras tabulares (Dataframes, SQL)
- Escalabilidad **Big Data**:
  - Los programas Spark se ejecutan igual (sin ajustar el código) en 1 o N máquinas
  - Soporte a datos con alto Volumen, Variedad y Velocidad
- **Integración** con las mayoría de fuentes y destinos de datos
  - Ej. MySQL, Oracle, Data Lake, Cassandra, Azure Synapse, Kafka , HDFS, S3,….

Apache Spark & Databricks

databricks

1. Apache Spark

Arquitectura distribuida y totalmente escalable



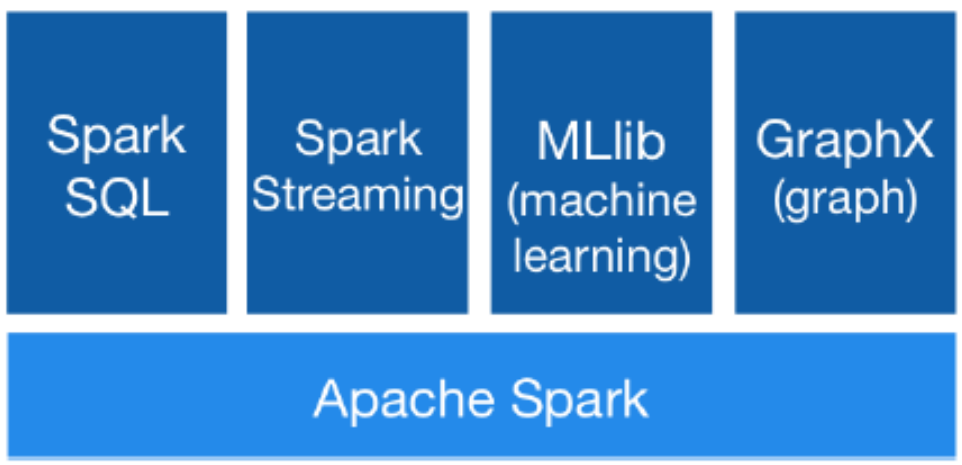
Apache Spark & Databricks

databricks

1. Apache Spark

Múltiples usos - Alta versatilidad

- Transformación de datos **batch o real-time**
- **Exploración** de datos RAW o procesados
  - Estructurados o no estructurados
- **Machine Learning**
  - Mlib o lenguaje R con ejecución distribuida
- **Gráfos**
- Base de datos **SQL** en memoria



Apache Spark & Databricks

databricks

2. Databricks

Databricks

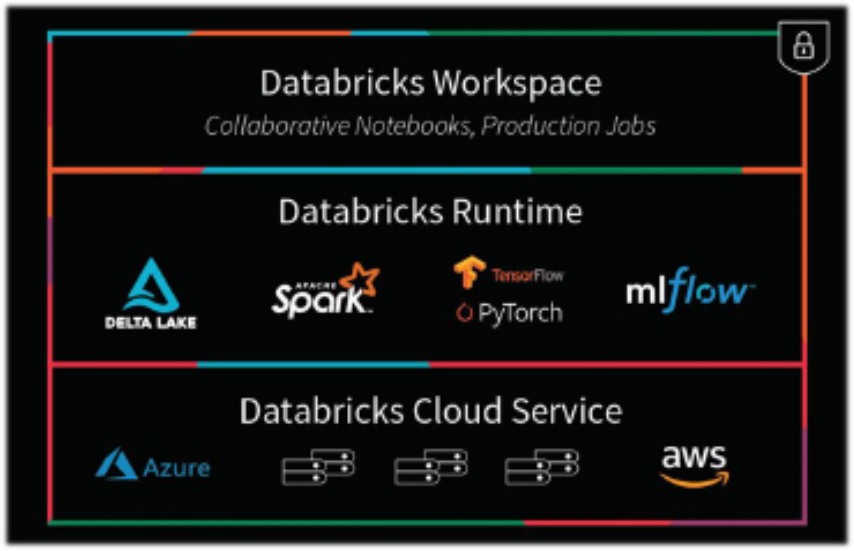
- **Versión enterprise de Spark como servicio cloud gestionado (PaaS)**
  - Principalmente usado en **Azure**, también disponible en **AWS** y **Google Cloud**
- **Evita tener que gestionar los recursos del clúster**
  - Las máquinas driver y worker(s) son las instancias de VM de Azure Compute
  - Creación e inicio de un clúster con N máquinas (workers) en minutos
  - Esas instancias no requieren mantenimiento, nos centramos en el análisis y transformación de datos
- **Añade importantes funcionalidades sobre Apache Spark**
  - Interfaz gráfica, librerías adicionales, optimización, integración con otros servicios cloud, …

Apache Spark & Databricks

databricks

2. Databricks

Principales diferencias frente a Apache Spark



- **Databricks Runtime**
  - Mismo núcleo que Apache **Spark, pero optimizado**
  - Librerias **Delta Lake**, Tensorflow, PyTorch y mlFlow
- **Databricks Notebooks**
  - Exploración de datos con cualquier nivel de estructura,

    PoC's, **compartición de resultados**, cuadros de mando…
  - No es necesario desplegar e integrar un herramienta de notebooks externa (ej. Jupyter o Zeppelin)
- **Integración nativa con servicios de Azure**
  - Blob Storage, **Data Lake, Data Factory, Azure SQL**, Synapse SQL, Event Hub, …

Apache Spark & Databricks

databricks

2. Databricks

Principales diferencias frente a Apache Spark

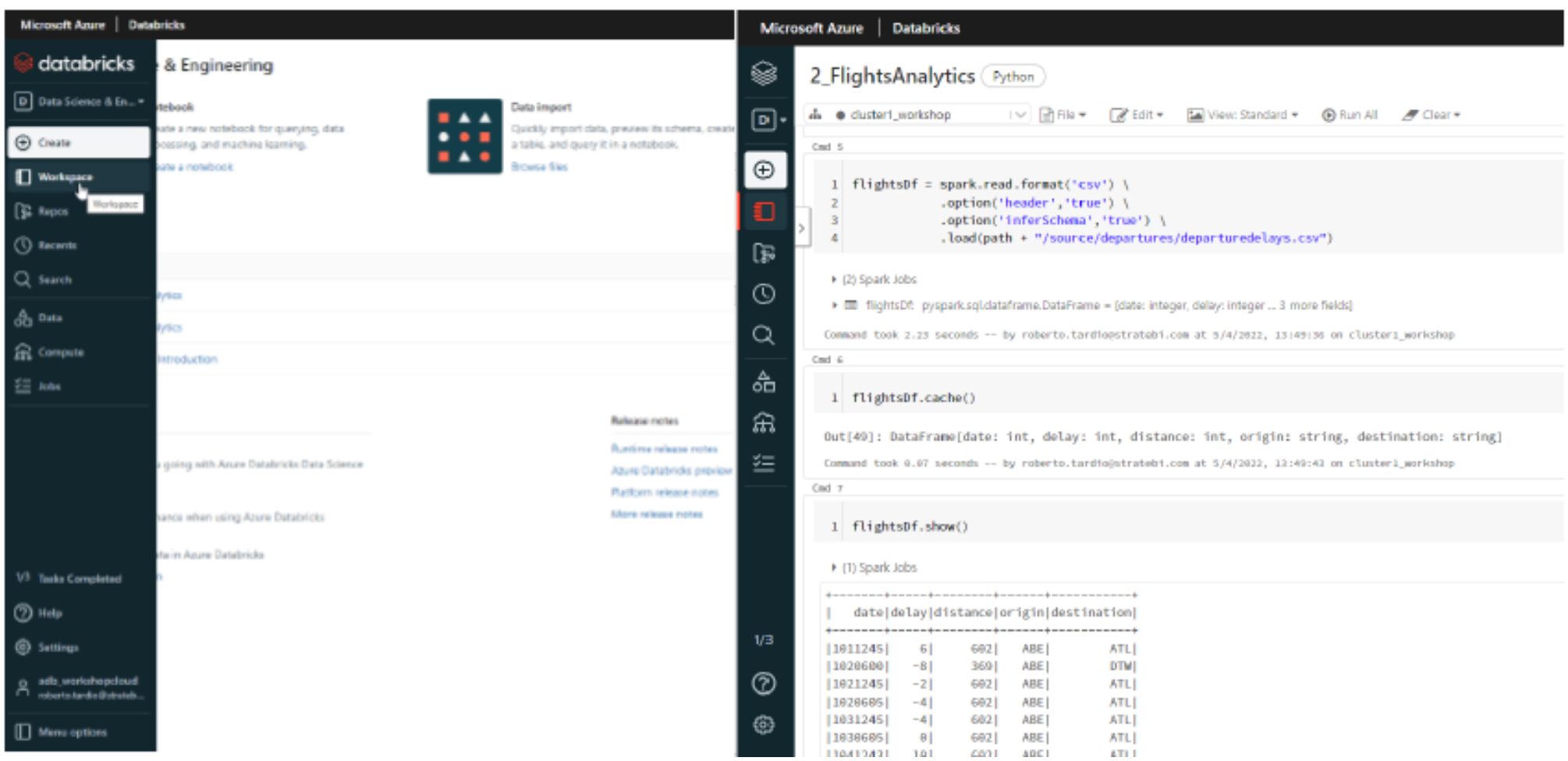
- **Delta Lake Gestionado**
  - Transacciones ACID, gestión de esquemas, versionado, optimizaciones,...
- **Automatización de Jobs**
  - Automatizar la ejecución de notebooks o procesos sin recurrir a otras herramientas (ej. Data Factory)
- **Seguridad avanzada**
  - Control de acceso (notebooks, clusters, Jobs o datos), auditoria, cifrado en reposo y transito,...
- **Integración**
  - J/ODBC (Power BI, Tableu, Looker, ...), Rest API, ...
- **Soporte experto**

Apache Spark & Databricks

databricks

2. Databricks

Workspace y Notebooks



Apache Spark & Databricks

databricks

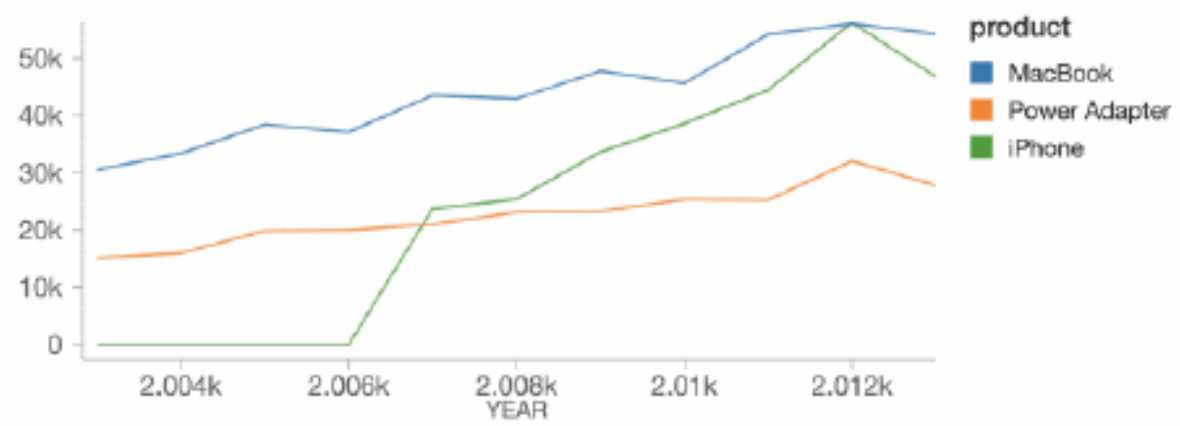
2. Databricks

Workspace y Notebooks

Cmd 1

```
1  %sql select * from sales where product!=""
2
```

▸ (2) Spark Jobs



Plot Options...

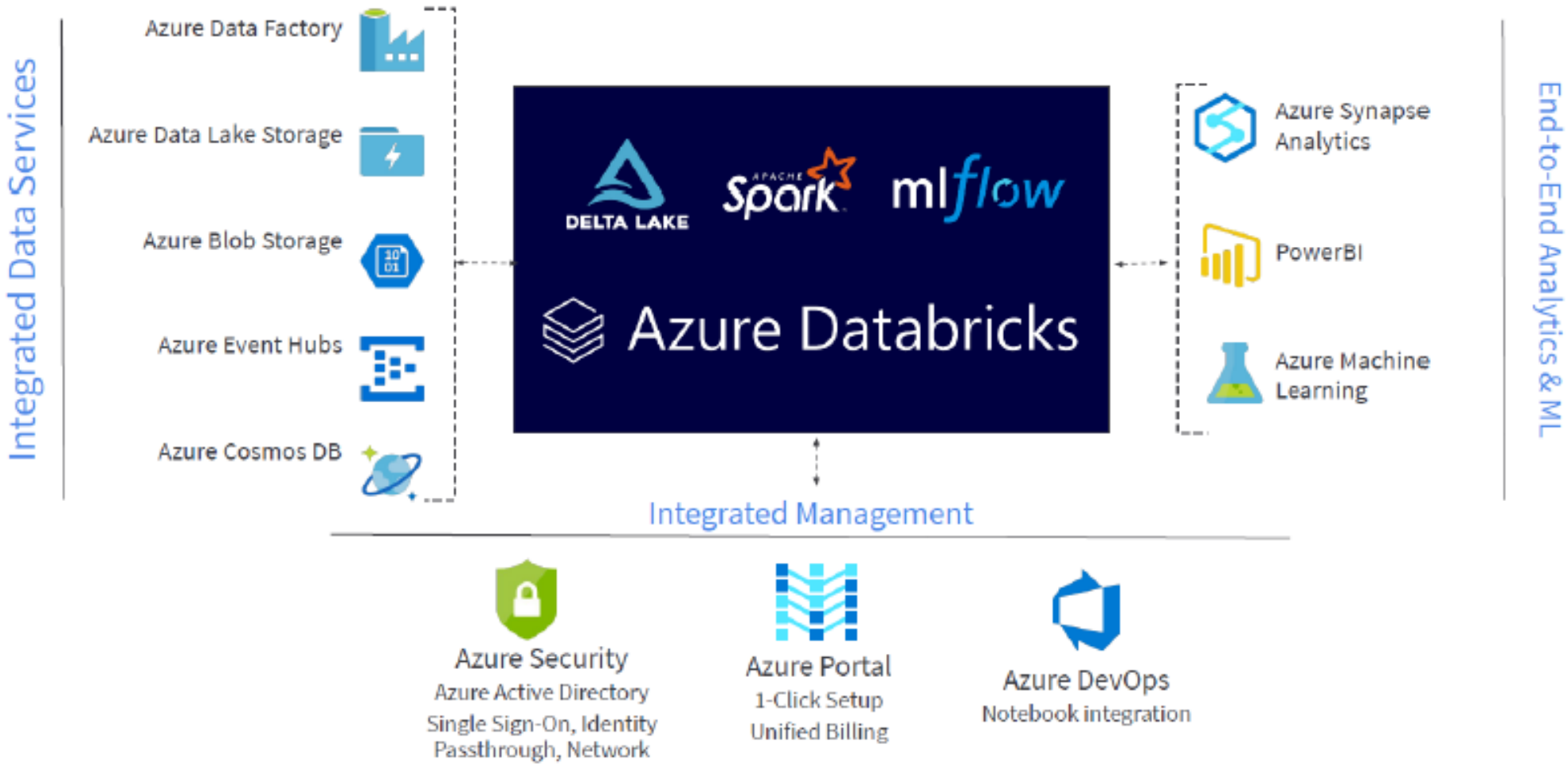
Command took 7.82 seconds -- by admin at 7/1/2019, 18:34:46 AM on test

Apache Spark & Databricks

databricks

2. Databricks

Integración nativa con múltiples servicios de Azure



Apache Spark & Databricks

databricks

2. Databricks

DELTA LAKE

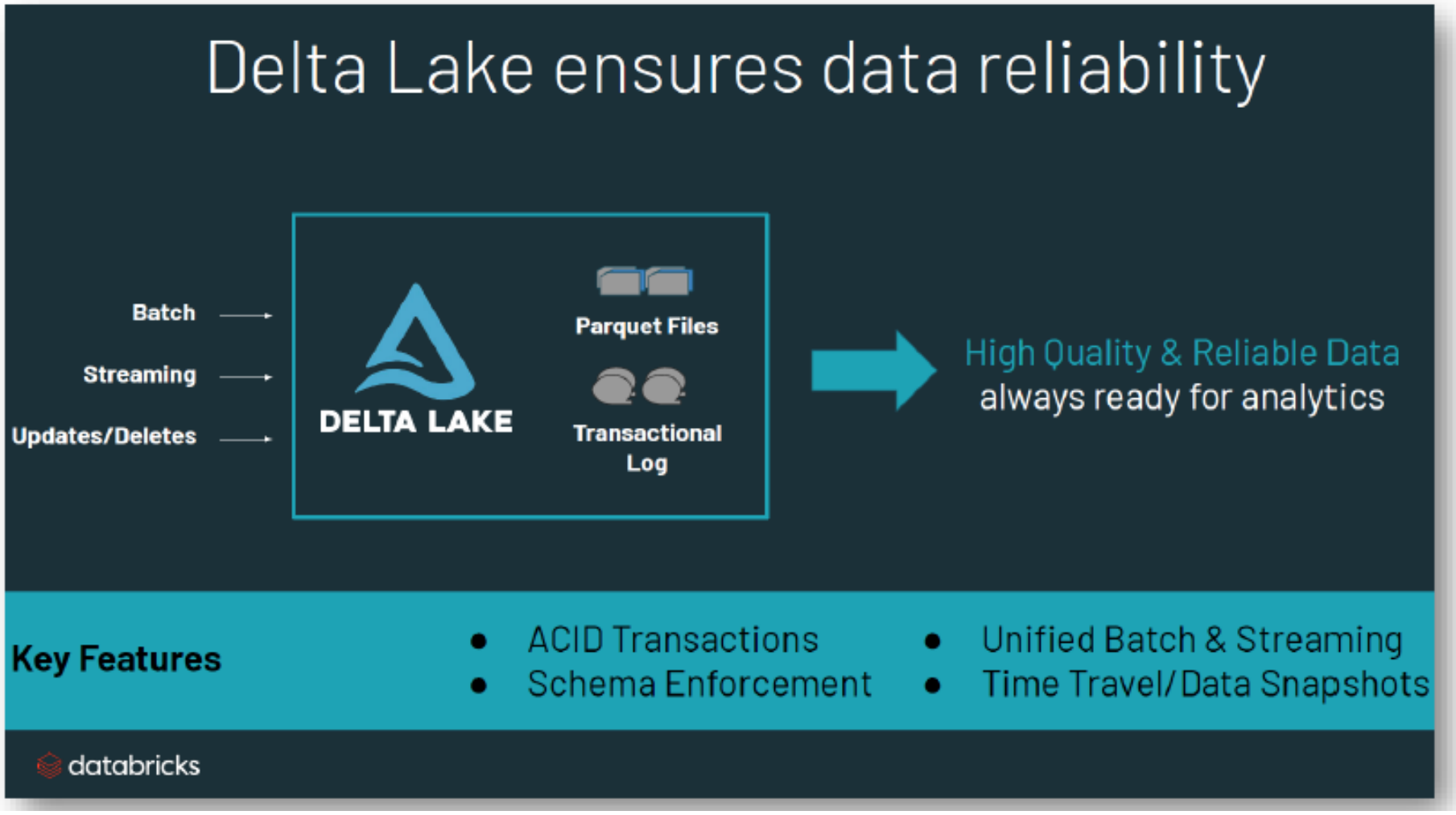
- **El formato Delta Lake** es un formato de código abierto para soportar transacciones ACID en los Datasets/Tablas de nuestro Data Lake
  - Basado en formato Parquet, añade soporte ACID
  - **Optimización automática** del almacenamiento
  - **Facilita la evolución de los esquemas** (adicción o eliminación de columnas)
  - **Indexación** para mejorar el rendimiento en filtrados y uniones
  - Soporte para sentencias **merge**
  - Open Source, uso en Spark como libreria adicional
    - En **Databricks** está integrado por defecto con el **Unity Catalog** para soportar el **concepto de Lakehouse**

Apache Spark & Databricks

databricks

2. Databricks

DELTA LAKE

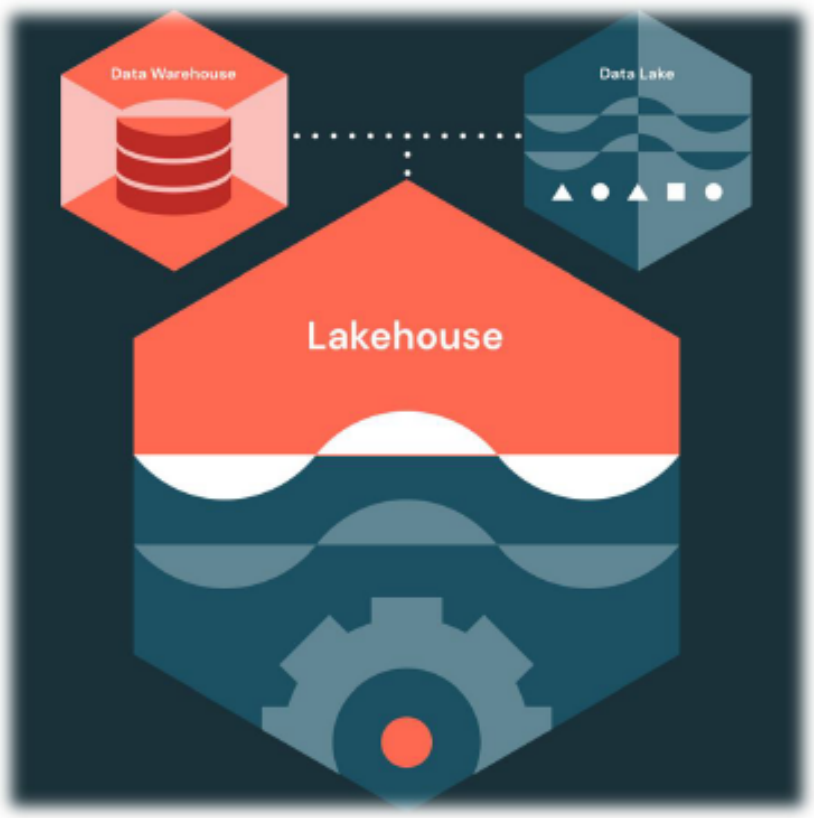


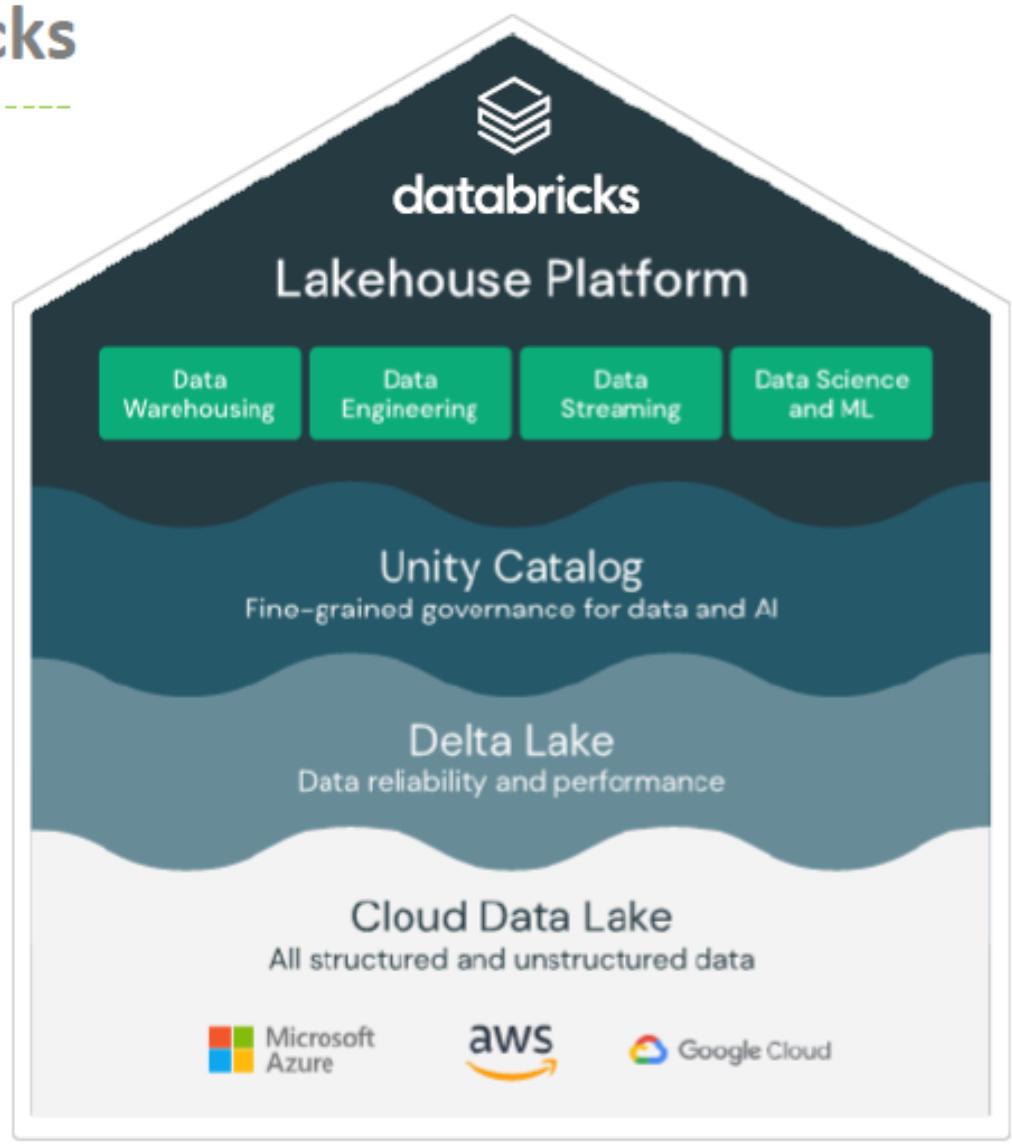
Apache Spark & Databricks

databricks

2. Databricks

- **Lakehouse**





Apache Spark & Databricks

databricks

2. Databricks

- **¿Data Warehouse, Lake o Lakehouse?**

Compute layer attributes — data lake vs. data warehouse vs. data lakehouse

Data Lake	Data Warehouse	Data Lakehouse
High performance for large jobs (TBs to PBs)	High concurrency	High performance for large jobs (TBs to PBs)
Economical	Scaling is exponentially more expensive	Economical
High operational complexity	Ease of use	Ease of use

Consumption layer attributes — data lake vs. data warehouse vs. data lakehouse

Data Lake	Data Warehouse	Data Lakehouse
Notebooks (great for data scientists)	Lack of support for data science/ML	Notebooks (great for data scientists)
Openness with rich ecosystem (Python, R, Scala)	Limited to SQL only	Openness with rich ecosystem (Python, R, Scala)
BI/SQL not 1st-class citizen	BI/SQL 1st-class citizen	BI/SQL 1st-class citizen

Apache Spark & Databricks

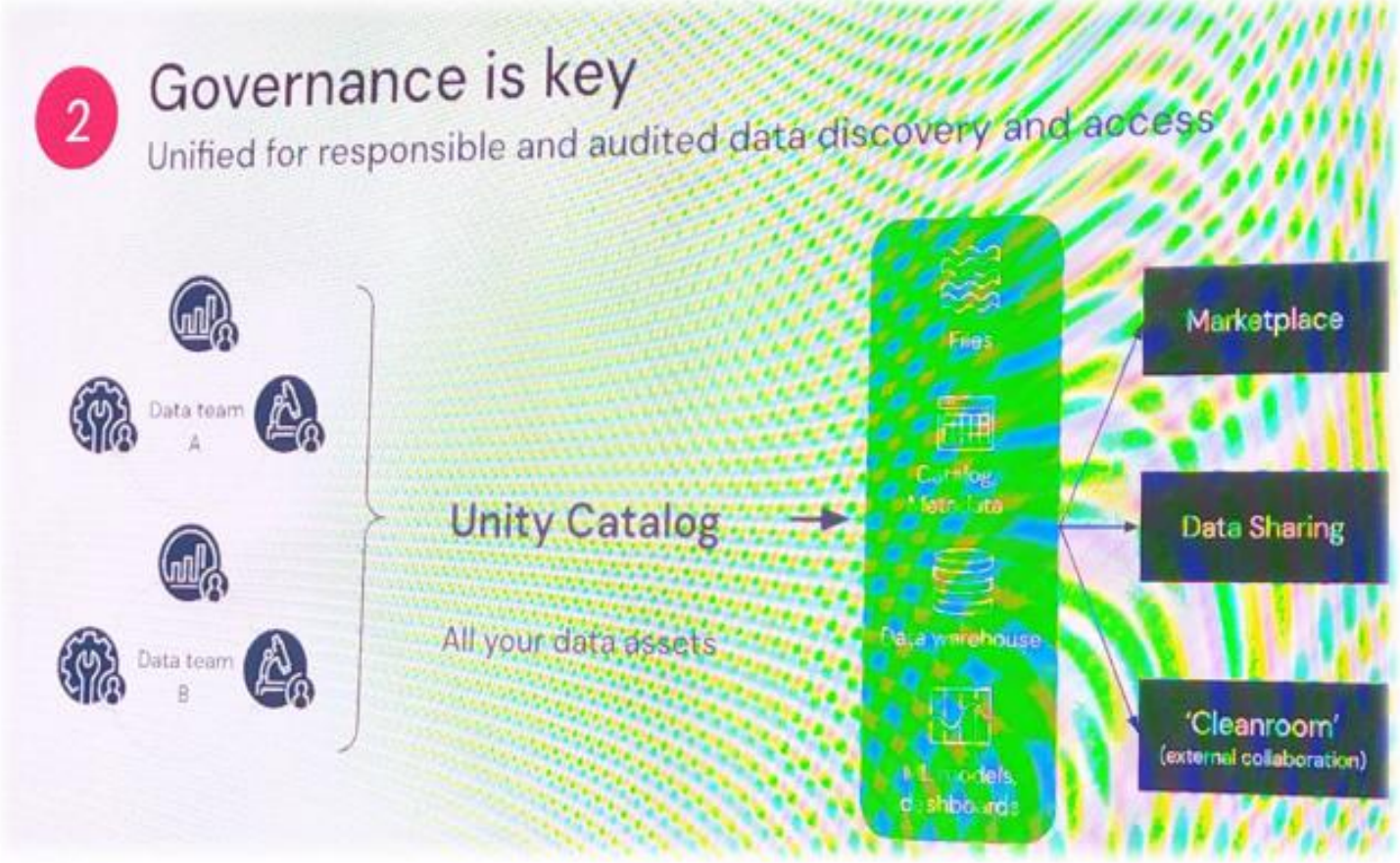
Lakehouse Day / Roadmap 2023

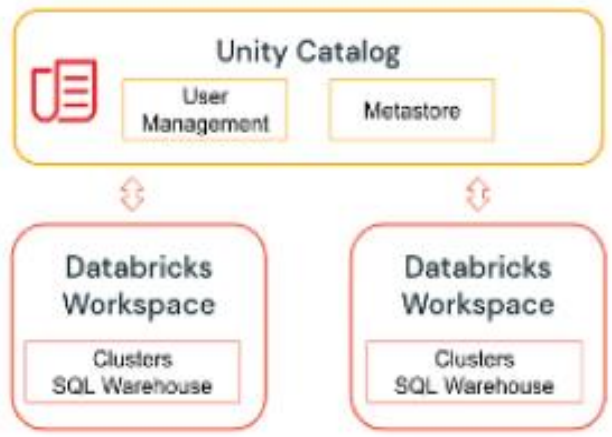
- **Lakehouse Day – Charlas destacadas**
- Ferroviario: **Cómo CAF utiliza la IA y los datos** para aumentar la productividad y reducir los costes
- **La Inteligencia del Fútbol: Cómo LaLiga Tech** está transformando la industria a través de los datos
- **Santalucía -** Arquitectura Lakehouse en el sector seguros: Casos de uso de Analítica Avanzada y BI
- **Gobernanza de datos** para el Lakehouse (Unity Catalog)
- **Roadmap**

Apache Spark & Databricks

databricks

Lakehouse Day / Roadmap 2023





Apache Spark & Databricks

databricks

4. Lakehouse Day / Roadmap 2023

Hello Dolly: Democratizing the magic of ChatGPT with open models

by Mike Conover, Matt Hayes, Ankit Mathur, Xiangrui Meng, Jianwei Xie, Jun Wan, Ali Ghodsi, Patrick Wendell and Matei Zaharia

March 24, 2023 in Company Blog

Share this post

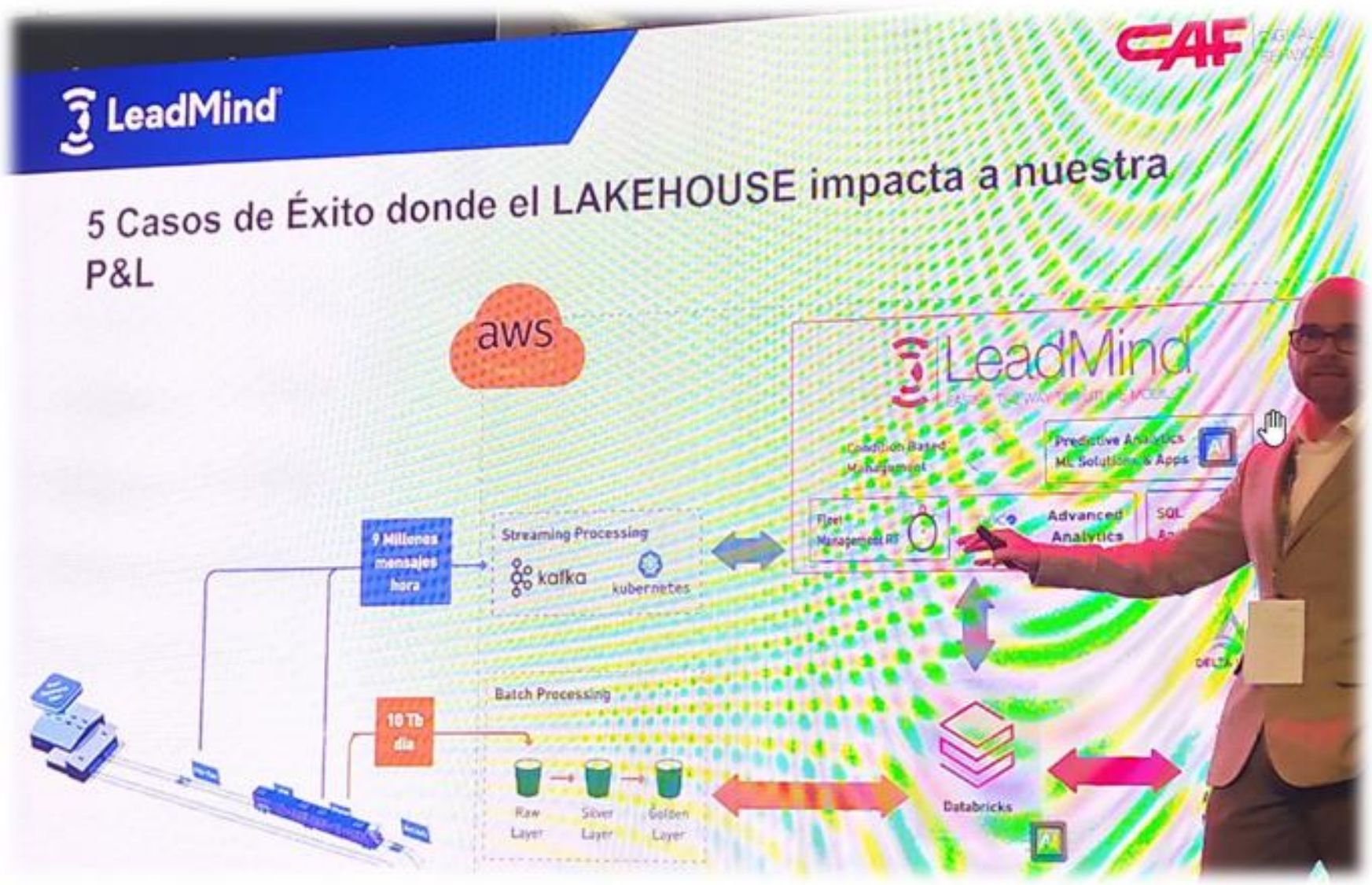
Summary

We show that anyone can take a dated off-the-shelf open source large language model (LLM) and give it magical ChatGPT-like instruction following ability by training it in 30 minutes on one machine, using high-quality training data. Surprisingly, instruction-following does not seem to require the latest or largest models: our model is only 6 billion parameters, compared to 175 billion for GPT-3. We open source the code for our model (Dolly) and show how it can be re-created on Databricks. We believe models like Dolly will help democratize LLMs, transforming them from something very

Apache Spark & Databricks

databricks

Lakehouse Day / Roadmap 2023



Apache Spark & Databricks

databricks

Lakehouse Day / Roadmap 2023

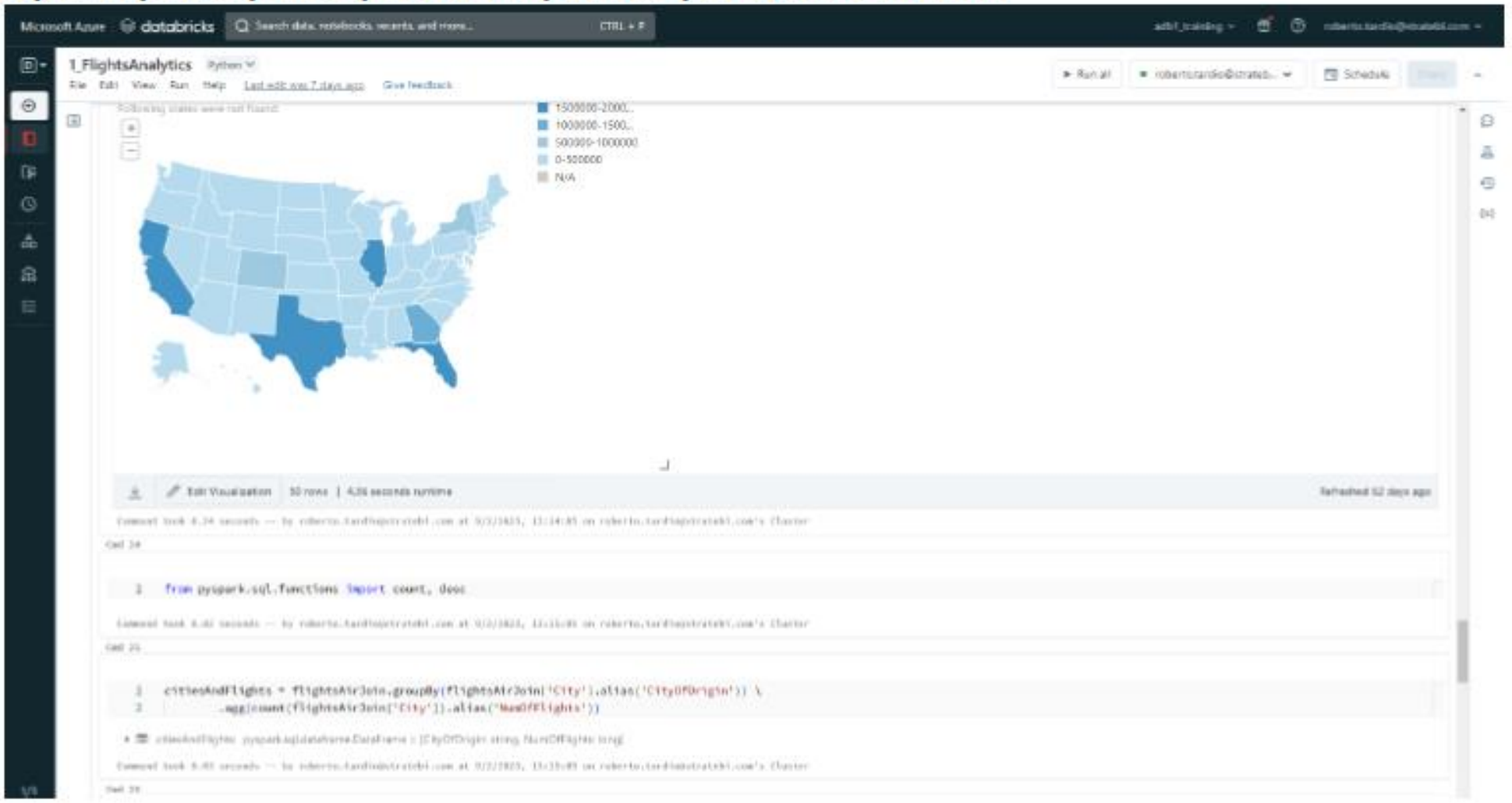
- **Funcionalidades destacables en el Roadmap**
- Modo Serverless: DB SQL, Model Serving, Workflows & Notebooks
- Tablas Streaming en el Lakehouse / Delta Live Tables
- Change data capture con Delta Live Tables
- Intelligent Workload Management: Mejoras en concurrencia de consulta, priorización, autoescalado inteligente
- Mejoras de los flujos de datos (orquestación): Condicionales, lanzamiento de otros jobs, …
- Soporte nativo para entornos de desarrollo (Visual Studio Code)
- Federación de consultas: Synapse, SQL Server, Redshift, PostgreSQL, Snowflake y MySQL
- Integración con nuevas herramientas: Go, Node.js, SQL API Rest, Amazon Quicksight

Apache Spark & Databricks

databricks

4. Lakehouse Day / Roadmap 2023

- **¿A qué esperas para aprender Apache Spark / Databricks?**

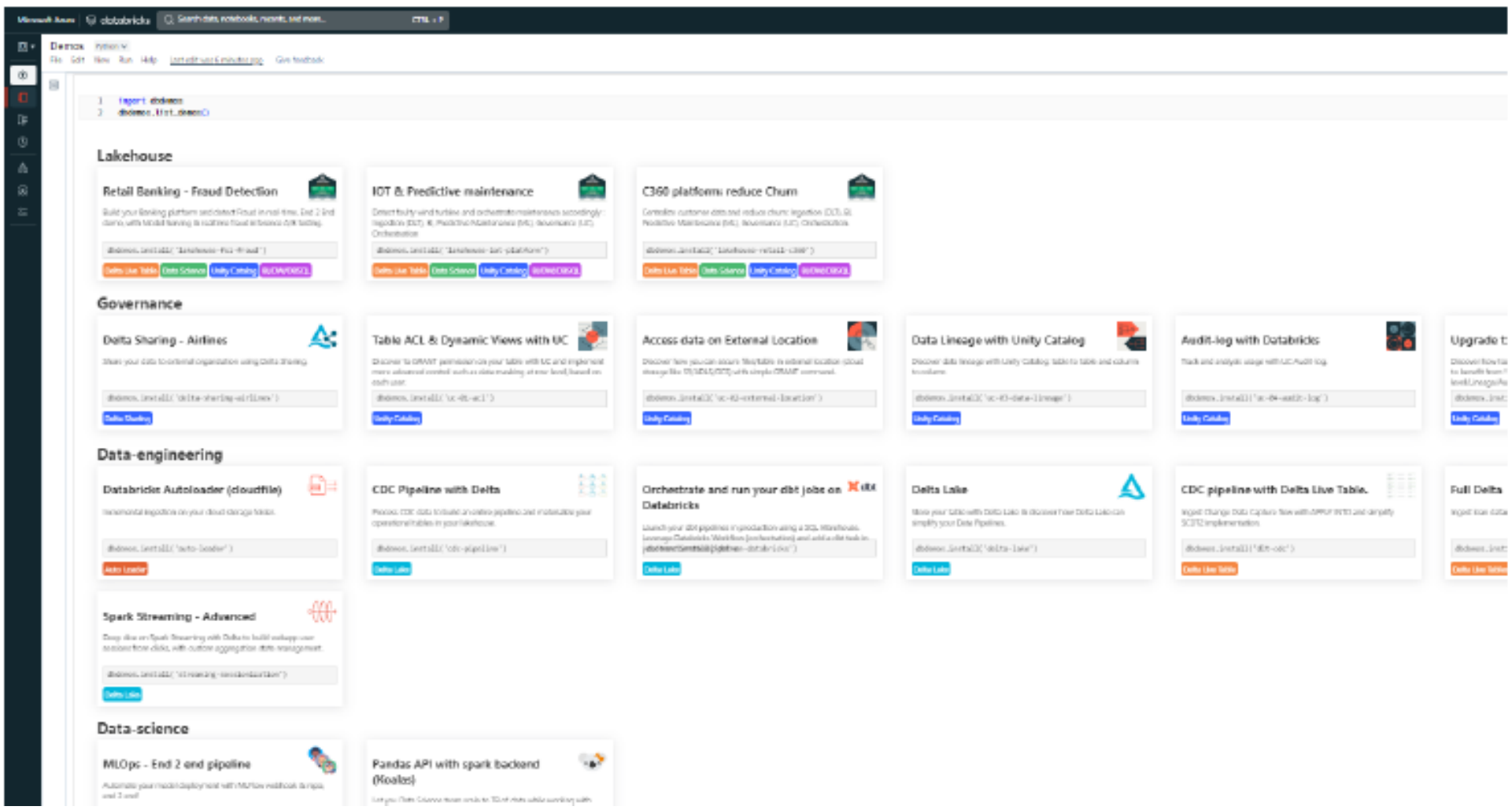


Apache Spark & Databricks

databricks

Lakehouse Day / Roadmap 2023

- **¿A qué esperas para aprender Apache Spark / Databricks?**



stratebi
open business intelligence