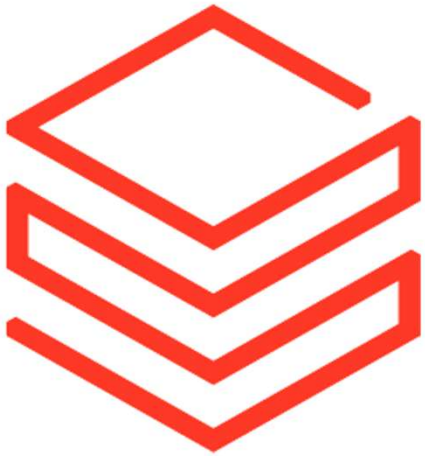
CESTE
Escuela Internacional de Negocios

Zaragoza (España)

Introducción a Databricks y su Arquitectura de Clusters



databricks

CESTE

¿Quién Soy?

David Bernabeu Ferrer – Data Platform Engineer en Plexus Tech

- AWS Solutions Architect Associate
- Ingeniero Informático y Máster en Ciencia de Datos por la Universidad de Alicante.
- Múltiples años de experiencia en AWS y plataformas de datos.

https://linkedin.com/in/david-bernabeu-data-engineer

Objetivos de la sesión

- Breve historia de Databricks
- Entender Databricks y su arquitectura.
- Identificar los tipos de clusters y su uso.
- Familiarizarse con el entorno de trabajo.
- Creación de un cluster y navegar por el entorno.

Historia de Databricks

- Fundada por creadores de Apache Spark (2013)
- Surgió del Proyecto AMPLab (https://amplab.cs.berkeley.edu/)
- Fundadores: Ali Ghodsi, Andy Konwinski.
- En 2017 fue anunciada como servicio de Azure (Azure Databricks)

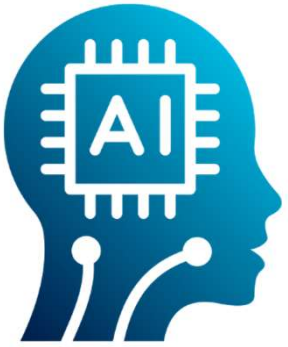
Actualidad Plataformas de Datos

ETLs

Data Warehouse

Business Intelligence

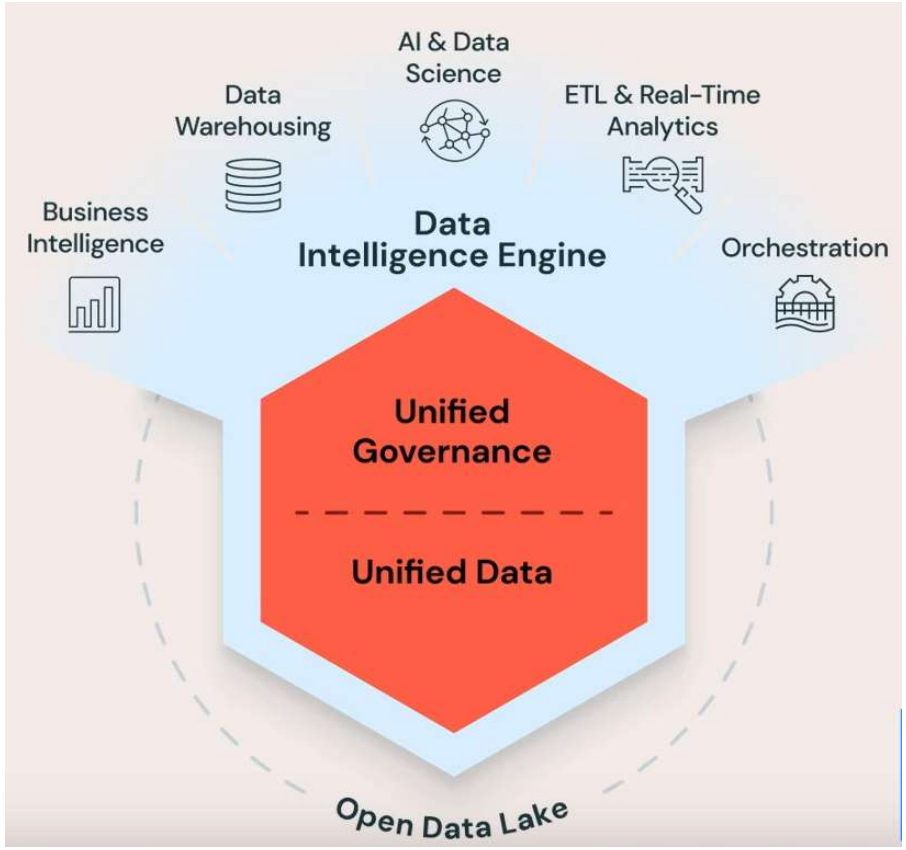
Plataforma



Inteligencia Artifical

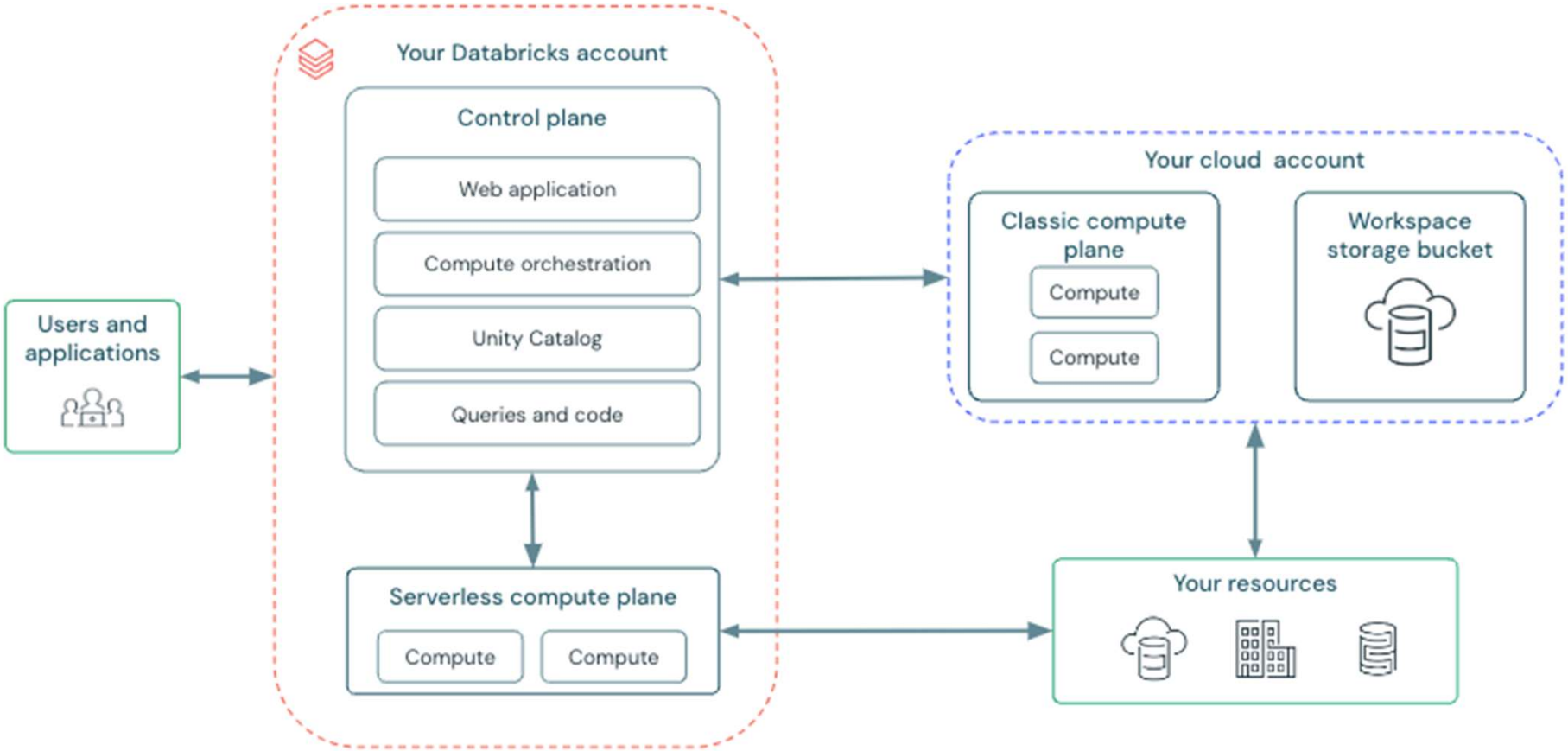
Gobernanza

¿Qué es Databricks & Lakehouse?

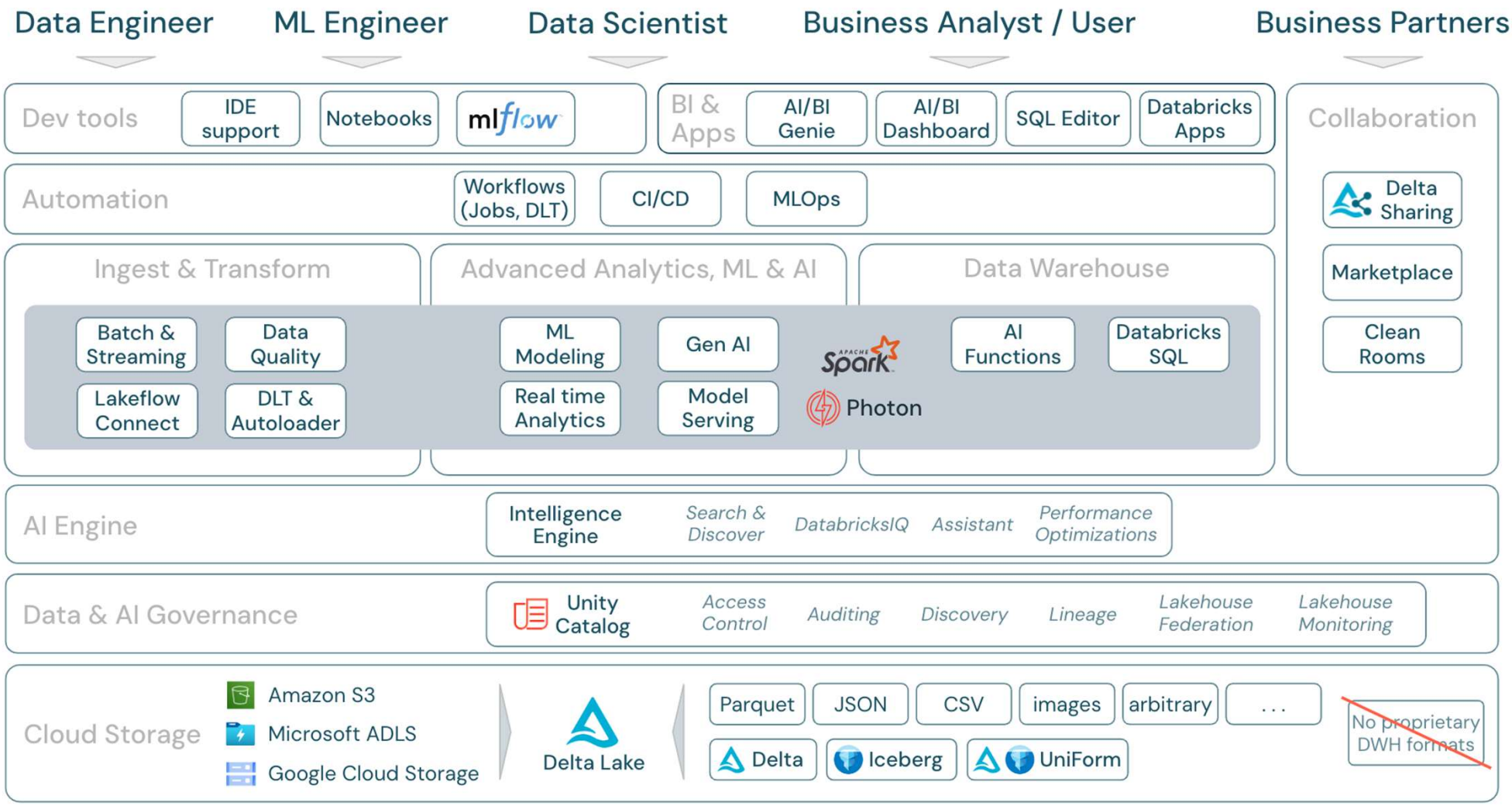


- Unificación de Plataforma de Data
- No vendor locking. Funcionamiento sobre AWS, Azure o Google Cloud.
- Uso de formatos OpenSource. Parquet.
- Uso de Delta Lake (próximas sesiones)

Databricks Architecture (I)



Databricks Architecture (II)



Clusters - Modos de acceso

Single User Access Mode

Shared Access Mode

No Isolation Shared Access Mode

Clusters - Tipos

All-Purpose

Job Clusters

Vector Search Clusters

SQL Warehouses

Clusters - Buenas prácticas

DEFINICIÓN PROPÓSITO DEL CLUSTER

DEFINICIÓN ROLES Y PERMISOS DE ACCESO Y USO.

CONFIGURAR AUTO-TERMINACIÓN

Clusters pools

RECURSO ADMINISTRADO.

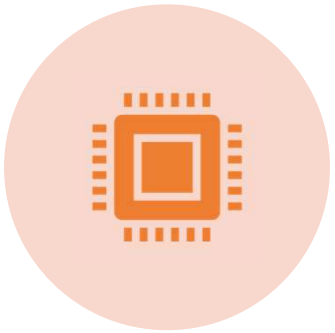
INSTANCIAS PRECONFIGURADAS Y READY-TO-GO.

REDUCEN TIEMPOS DE ESPERA.

PROPORCIONAN AGILIDAD Y ESCALABILIDAD

¿Cuándo utilizar Cluster Pools?

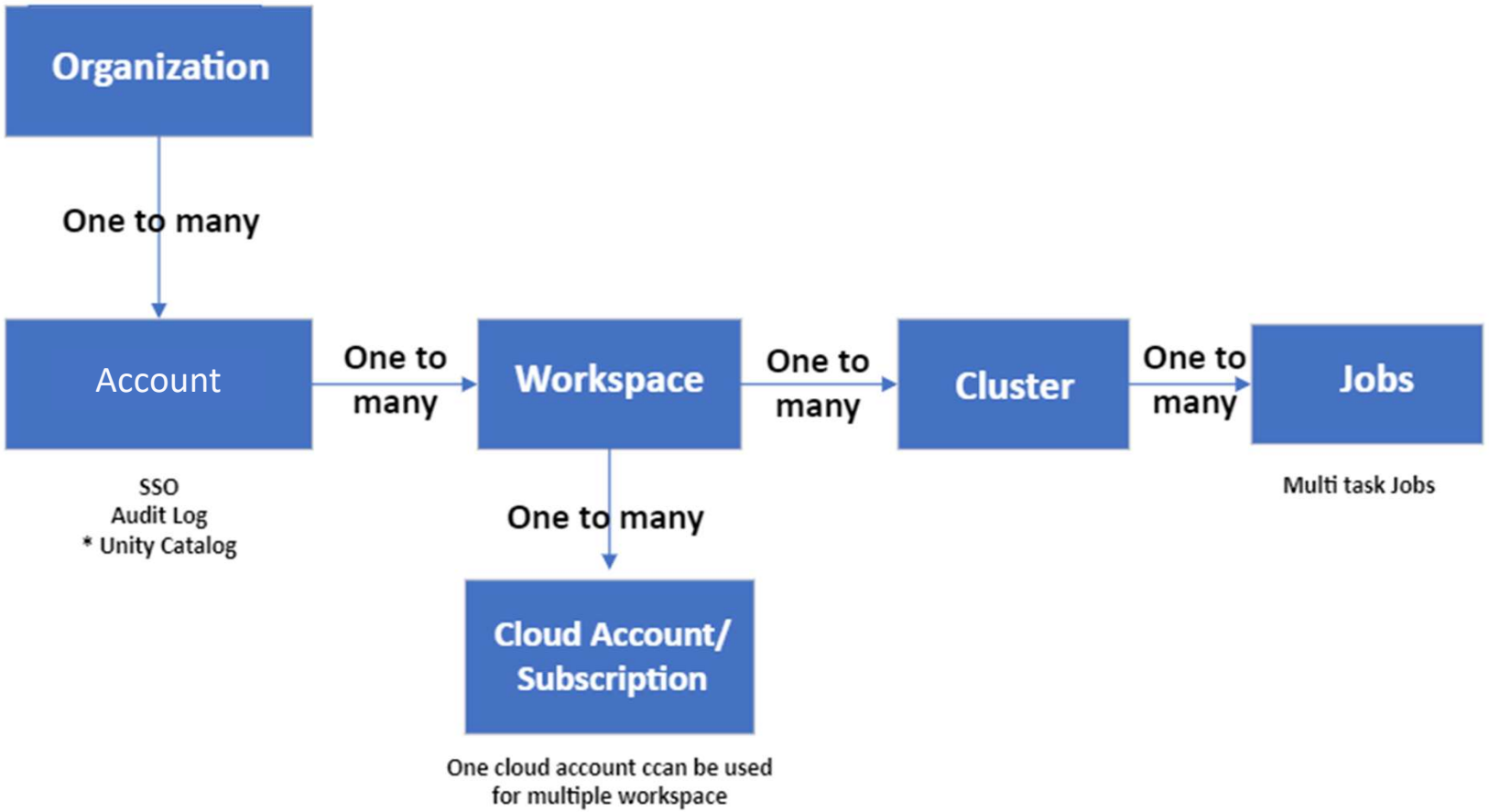
ENTORNO COLABORATIVO



PROCESAMIENTO ON-DEMAND

OPTIMIZACIÓN DE COSTES

Workspace


Organization
One to many
Account
One to many
Workspace
One to many
Cluster
One to many
Jobs
SSO
Audit Log
* Unity Catalog
One to many
Multi task Jobs
Cloud Account/ Subscription
One cloud account ccan be used for multiple workspace

Demo Time

Creación de cuenta en Databricks, configuración de clúster y navegación por el entorno.

https://www.databricks.com/try-databricks

Próxima sesión

¿Qué es Spark? Fundamentos.

Spark en Databricks

Ejercicios básicos de Spark, Dataframes y SparkSQL.

Creación de catálogos, esquemas y tablas.

Preguntas / Sugerencias

www.ceste.es