DMCI Google Partner

ESPECIALIZACIÓN

Ingeniería de datos con Azure

Curso: Microsoft Azure & ETL Fundamentals

Docente: Richard Tadeo Zenteno

REGLAS

Se requiere **puntualidad** para un mejor desarrollo del curso.

Para una mayor concentración **mantener silenciado el micrófono** durante la sesión.

Las preguntas se realizarán **a través del cha**t y en caso de que lo requieran **podrán activar el micrófono**.

Realizar las actividades y/o tareas encomendadas en **los plazos determinados**.

Identificarse en la sala Zoom con el primer nombre y primer apellido.

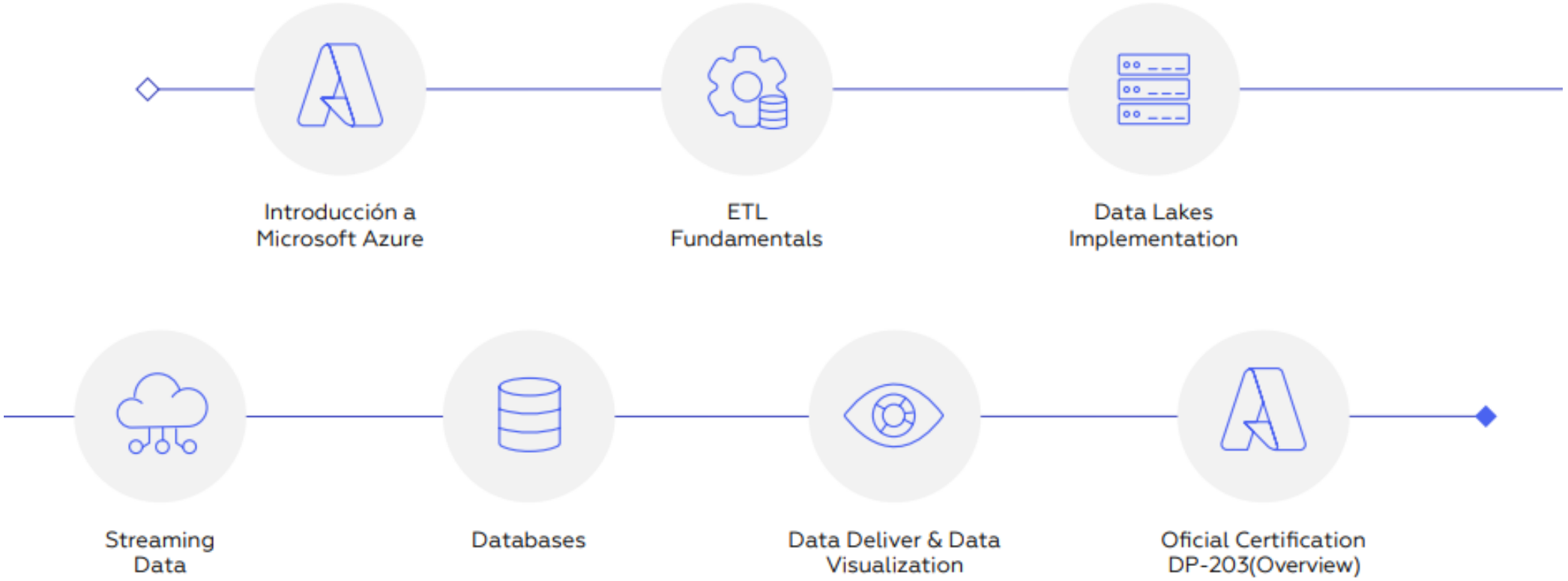
ITINERARIO

07:00 PM – 07:30 PM	***Soporte técnico DMC***
07:30 PM – 08:50 PM	***Agenda***
08:50 PM – 09:00 PM	***Pausa Activa***
09:00 PM – 10:30 PM	***Agenda***

Horario de Atención Área Académica y Soporte

Lunes a Viernes 09:00 am a 10:30 pm / Sábado 09:00 am a 02:00pm

MALLA CURRICULAR



CERIFICACIÓN FINAL
por **Aprobación** de la Especialización en **Ingeniería de Datos con Microsoft Azure** (48 horas académicas)

CONTENIDO

Introducción a Microsoft Azure

- Introducción a Cloud Computing. Proveedores de servicios Cloud, On-Premise vs. On-Cloud, principales servicios, descripción de los modelos de costos.
- Identify and Access Management (IAM). Overview de los roles principales, ejemplos de gestión de permisos.

ETL Fundamentals

- Introducción a las soluciones ETL. Definición, descripción de sus etapas.
- Introducción a los servicios Azure Data Factory y Data Flow. Características generales, casos de uso.
- Taller: Implementación de un ETL Básico con Azure.

Data Lakes Implementation

- Introducción a Data Lakes. Definición, arquitectura, capas (Raw, Stage, Analytics).
- Introducción a los servicios Azure Blob Storage y Storage Account.
- Taller: Implementación de un Datalake en Azure.

CONTENIDO

Streaming Data

- Introducción a procesamiento de datos Batch y Streaming. Diferencias Near-Real-Time y Real-Time.
- Introducción a IoT. Definición, uso de sensores, aplicaciones.
- Revisión de servicios: Azure EventHubs y IoTHub. Características generales, ejemplos de implementación y uso.
- Taller: Manejo de Streaming al Data.

Databases

- Introducción a las bases de datos Relacionales y No-Relacionales. Definición, características, casos de uso.
- Azure SQL Database for MariaDB. Descripción y características generales.
- Azure SQL Database for PostgreSQL. Descripción y características generales.
- Azure SQL Database for CosmosDB. Descripción y características generales.
- Taller: Diseño de una base de datos relacional y técnicas para poblarla.

Data Deliver & Data Visualization

- Azure Synapse Analytics. Propósito del servicio, características generales.
- Fabric. Propósito del servicio, características generales.
- Taller: Conexión de Power BI a servicios de datos de Azure.

AGENDA

01

Databricks ETL en Batch

02

Laboratorio 04: Implementación de un ETL Básico con Databricks

¿Qué es Spark?

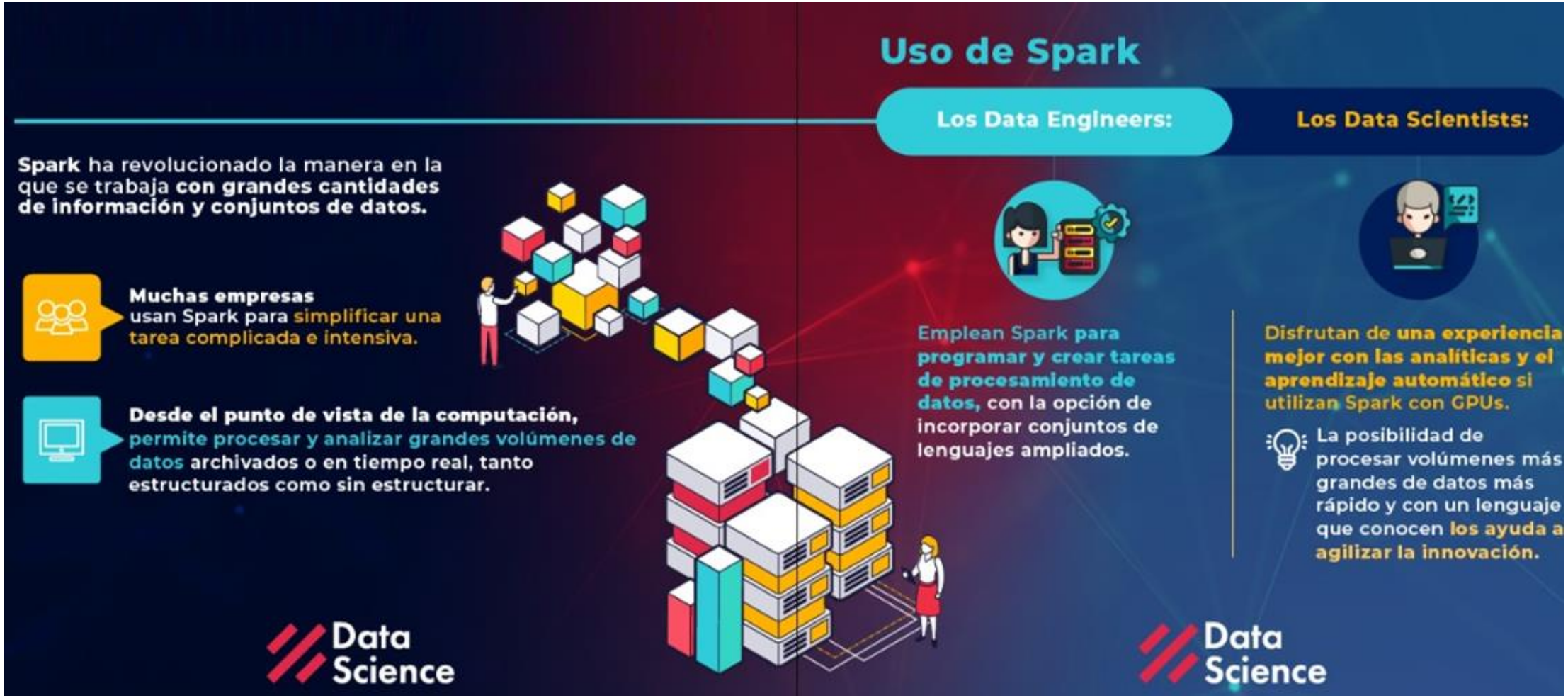
Spark es un motor unificado de procesamiento de datos, desarrollado bajo un framework opensource.

Presenta mejoras sobre la implementación de MapReduce de Hadoop.

Diferencias:

- Datos
  - **Hadoop** mantiene los datos en disco.
  - **Spark** mantiene datos en memoria para procesarlos.
- Ejecución de Proceso
  - **Hadoop** ejecuta tareas en 2 etapas.
  - **Spark** planea y optimiza DAGs.

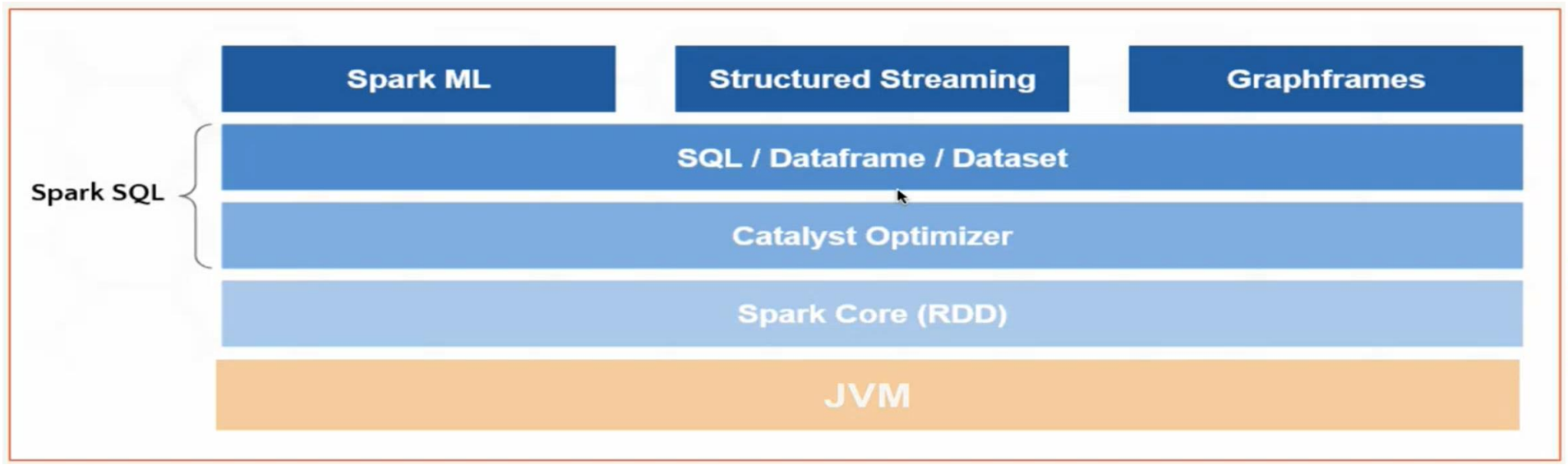
¿Las empresas como utilizan Apache Spark?



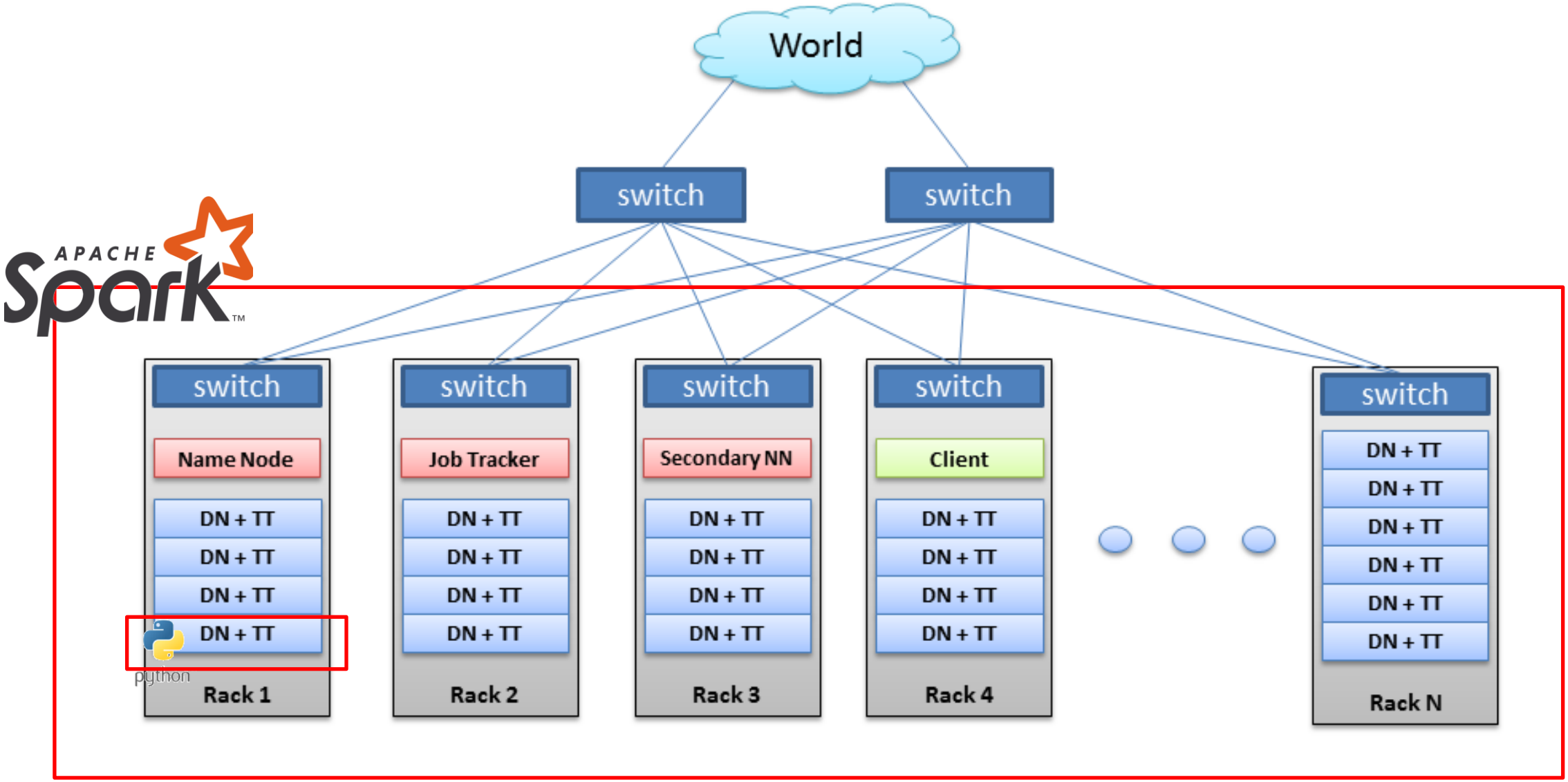
Arquitectura Spark

Spark esta escrito en Scala y corre en la maquina virtual de Java (JVM)

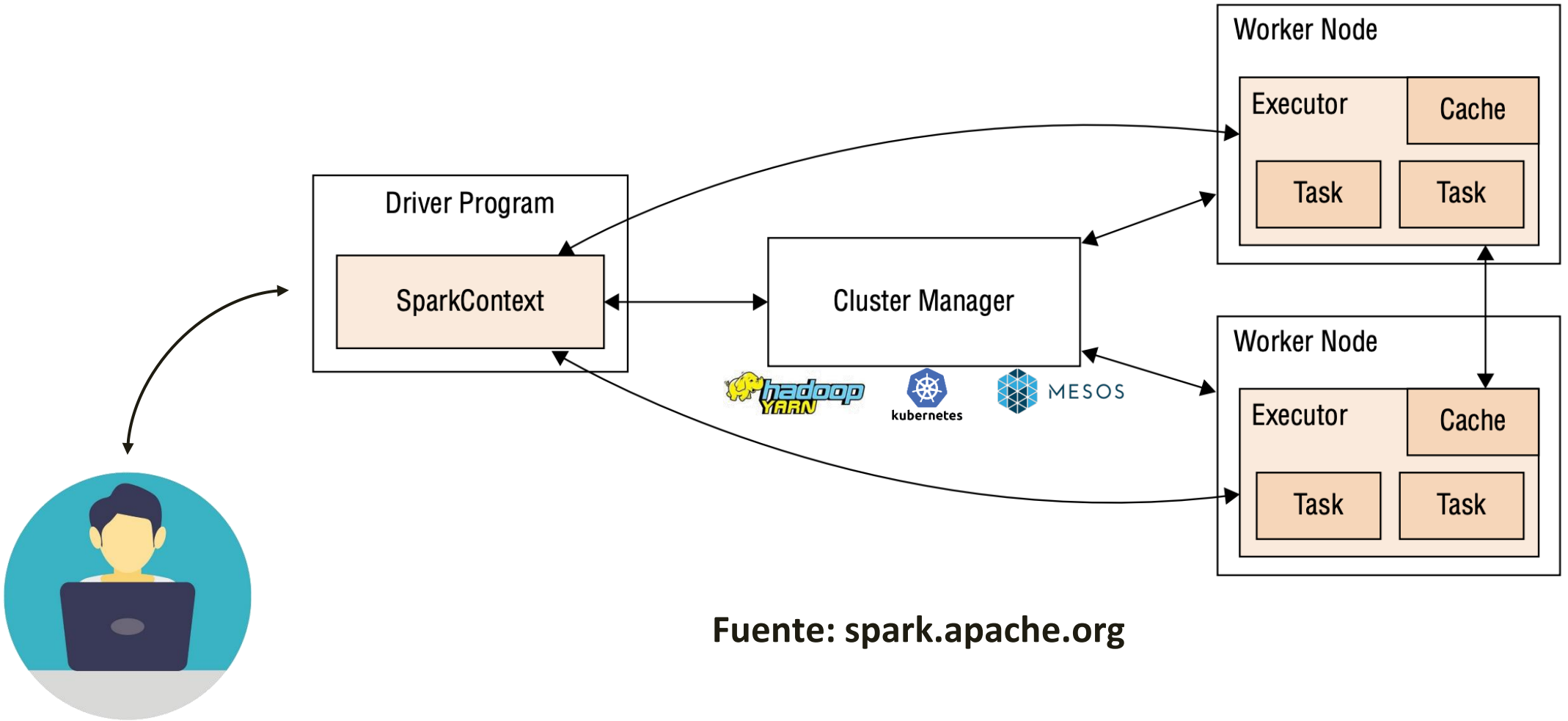
Componentes de **Spark**



Arquitectura Física - Clúster

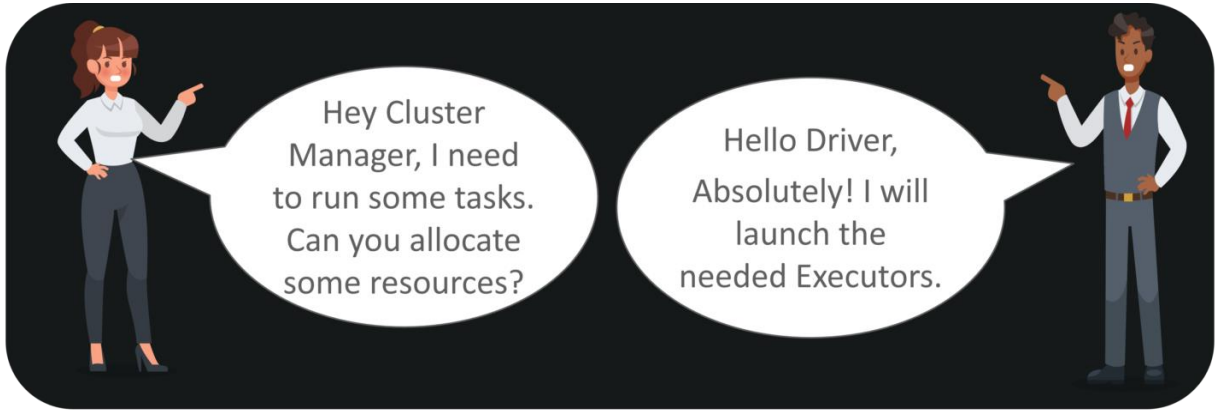


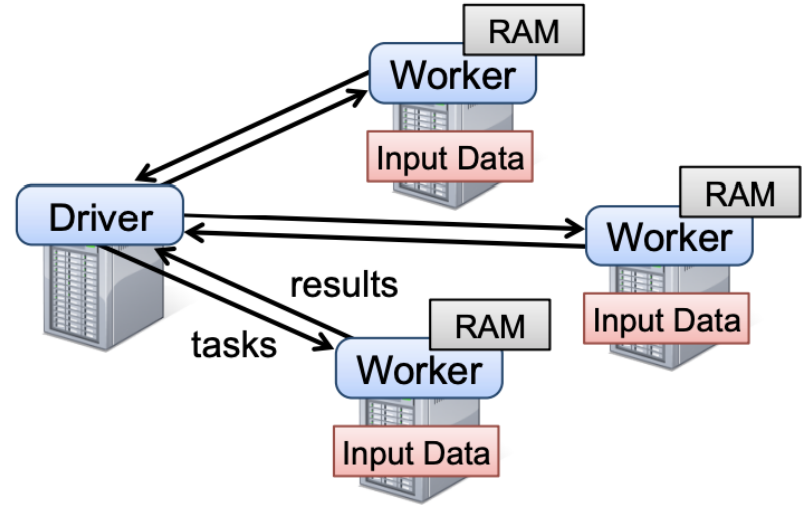
Arquitectura Spark



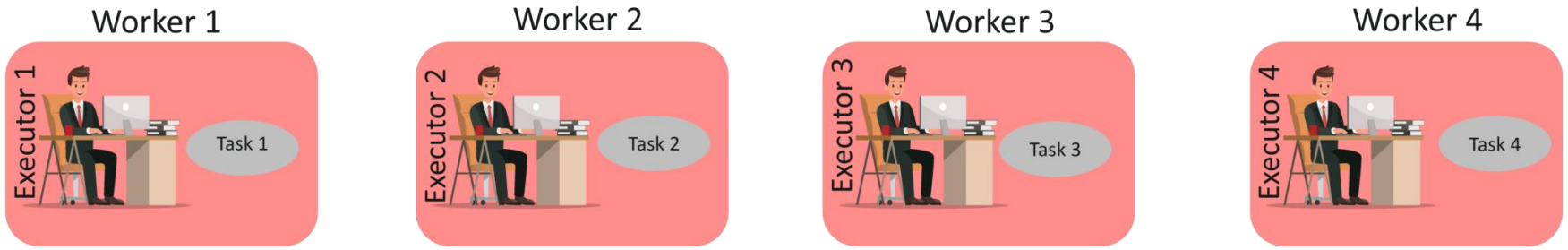
Fuente: spark.apache.org

Arquitectura Spark





Fuente: spark.apache.org



Fuente: edureka.co

Diferencia entre lenguajes de programación

Pyspark, Scala	SQL
Performance igual que SQL	Performance igual que PYSPARK
El código es modular	Código no Modular
Código mantenible	Código secuencial
Se puede aplicar pruebas unitarias	No aplica pruebas unitarias
Generar funciones reutilizables	
. Cuenta con API para incrementar la adaptabilidad con otras aplicaciones	
Código funcional	

Spark en el Cloud

Google Partner

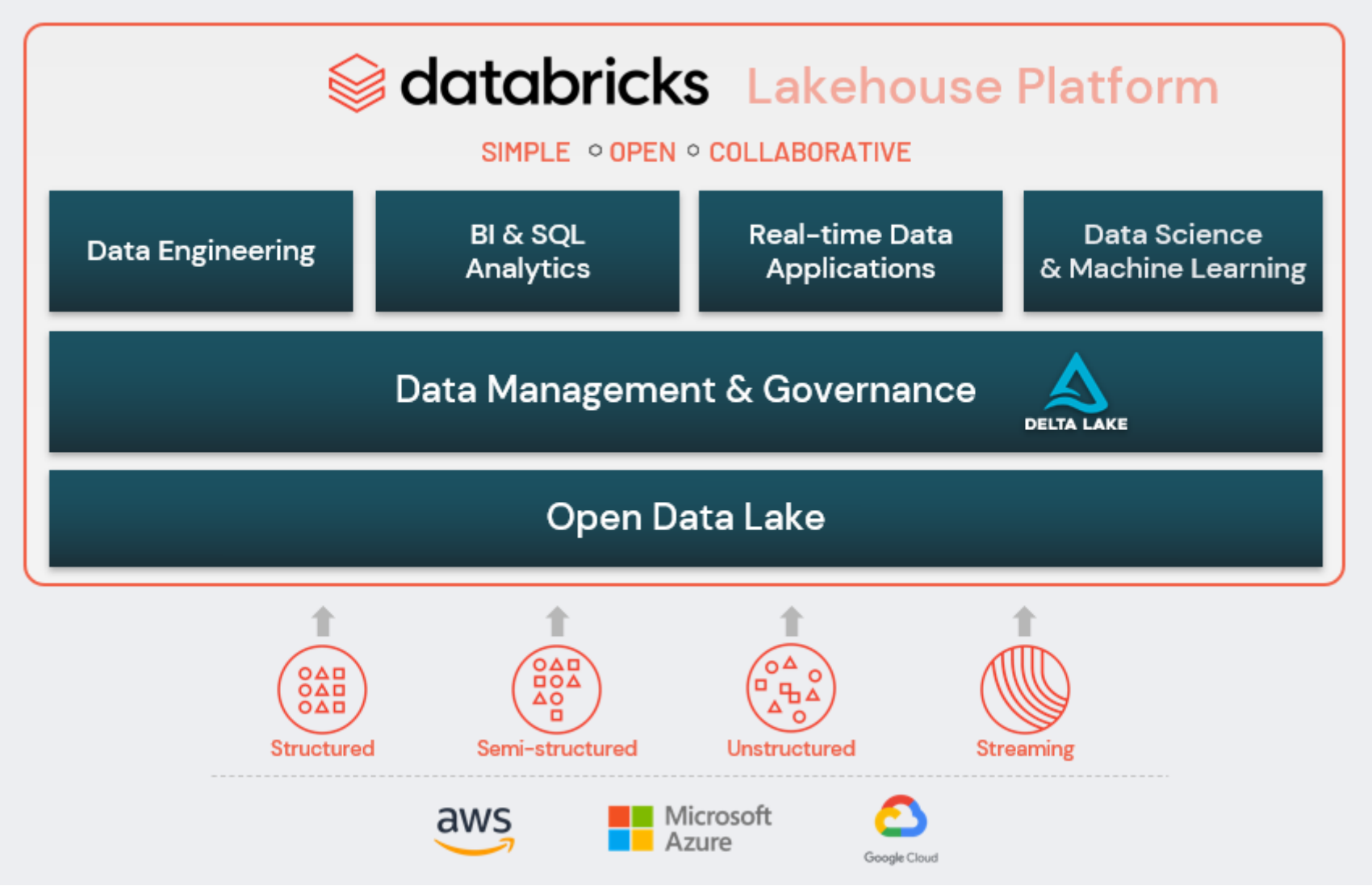
databricks

One Platform | Multi-Cloud | Open Source

APACHE Spark™ | DELTA LAKE | mlflow™ | Koalas

A Unified Data Analytics Platform for accelerating innovation across data engineering, data science, and data analysts

Lakehouse en Databricks



Data Lakehouse

Data Lake

DELTA LAKE

An open approach to bringing **data management and governance** to data lakes

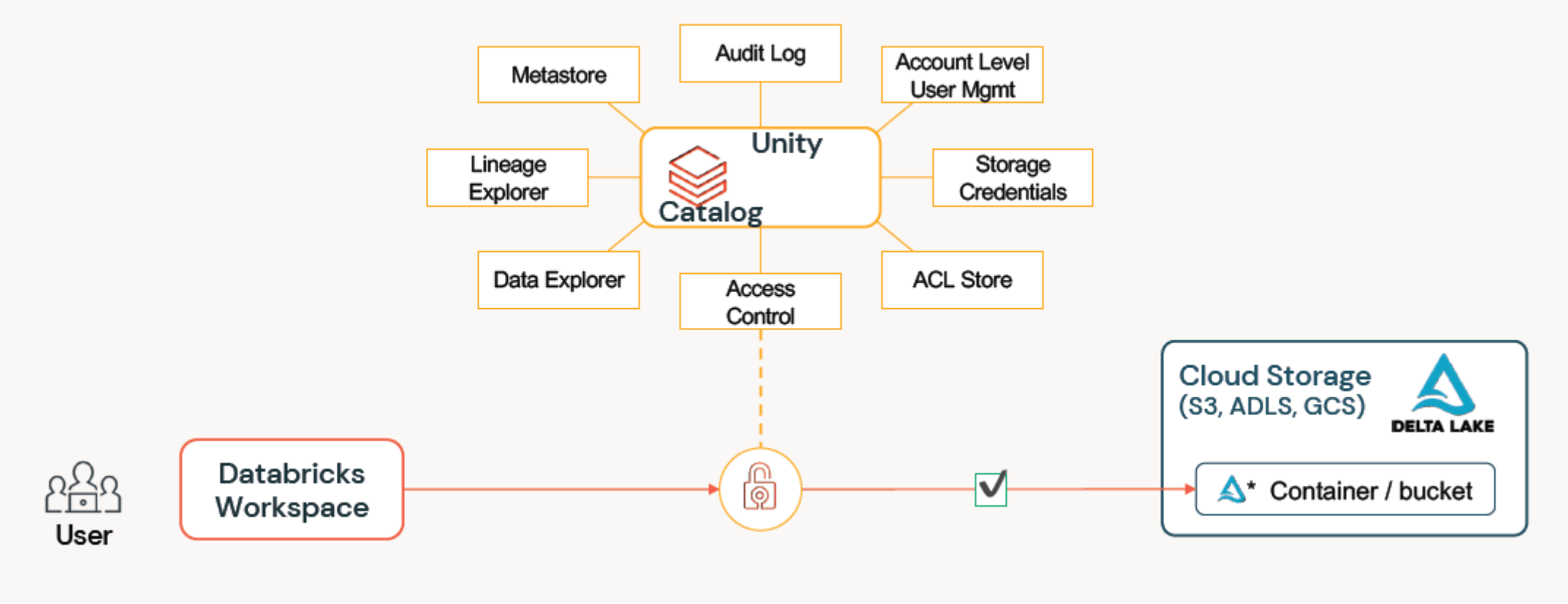
Better reliability with transactions

48x faster data processing with indexing

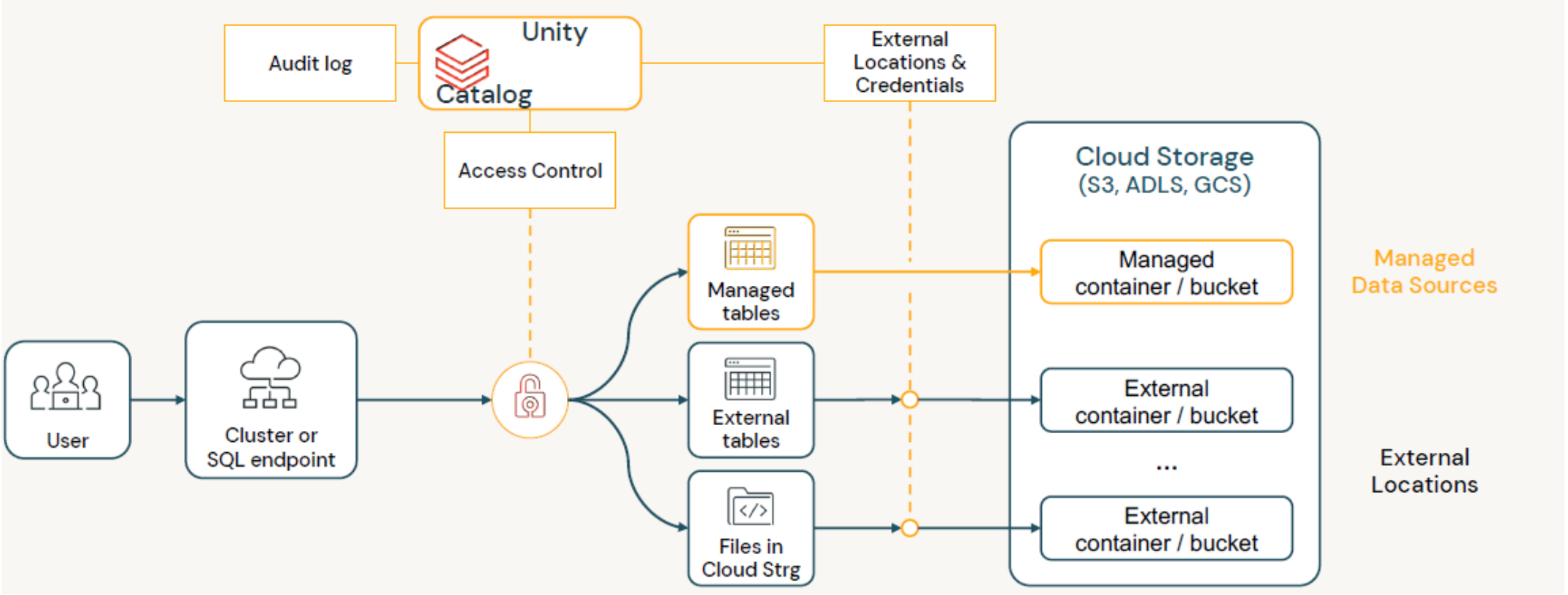
Data governance at scale with fine-grained access control lists

Data Warehouse

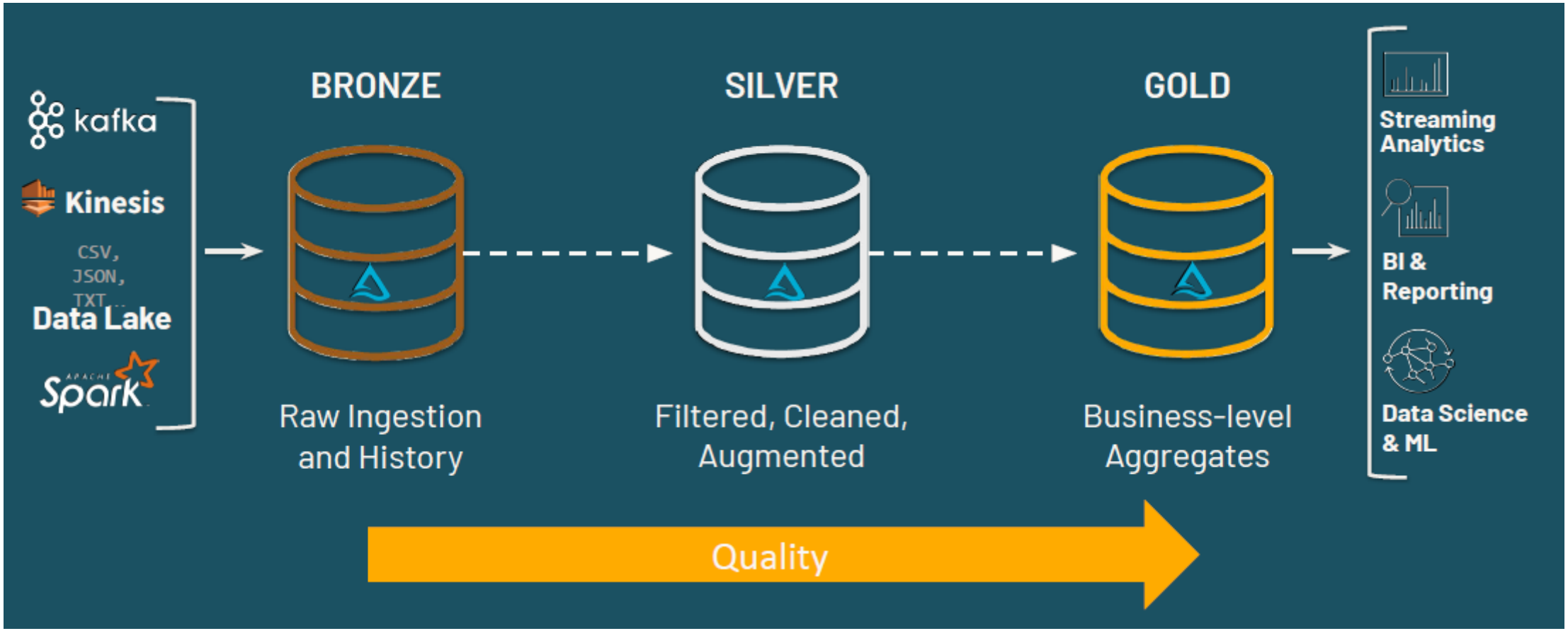
Unity Catalog - Architecture



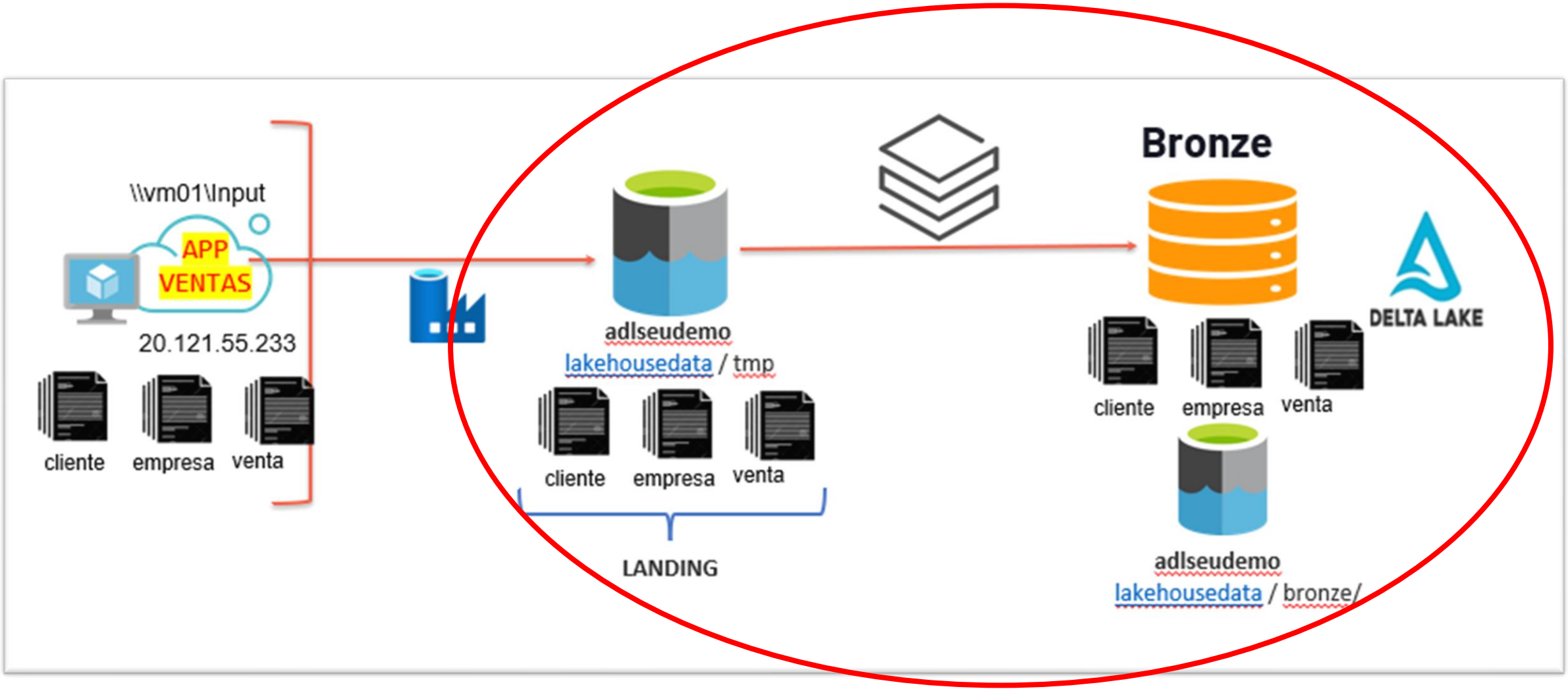
Unity Catalog - Architecture



Patrón de diseño de Arquitectura Delta



LABORATORIO: ETL con Databricks



RONDAS DE PREGUNTAS

¡GRACIAS!