

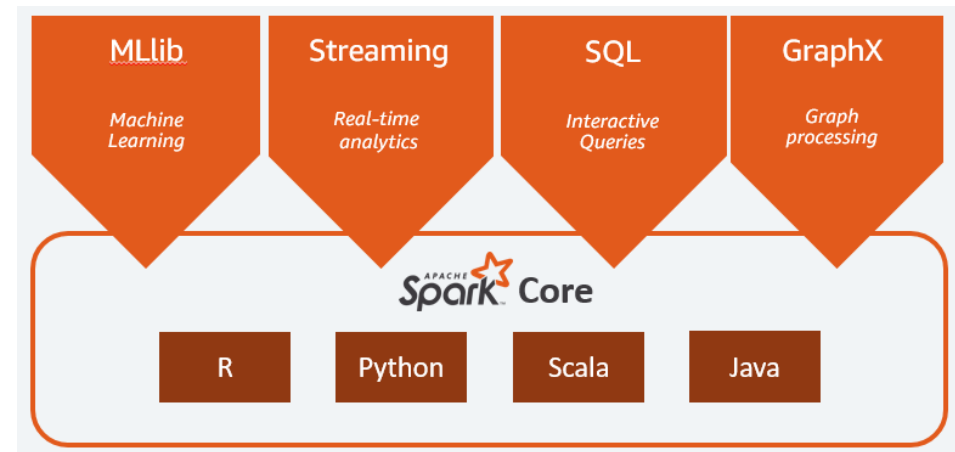
01

DATABRICKS



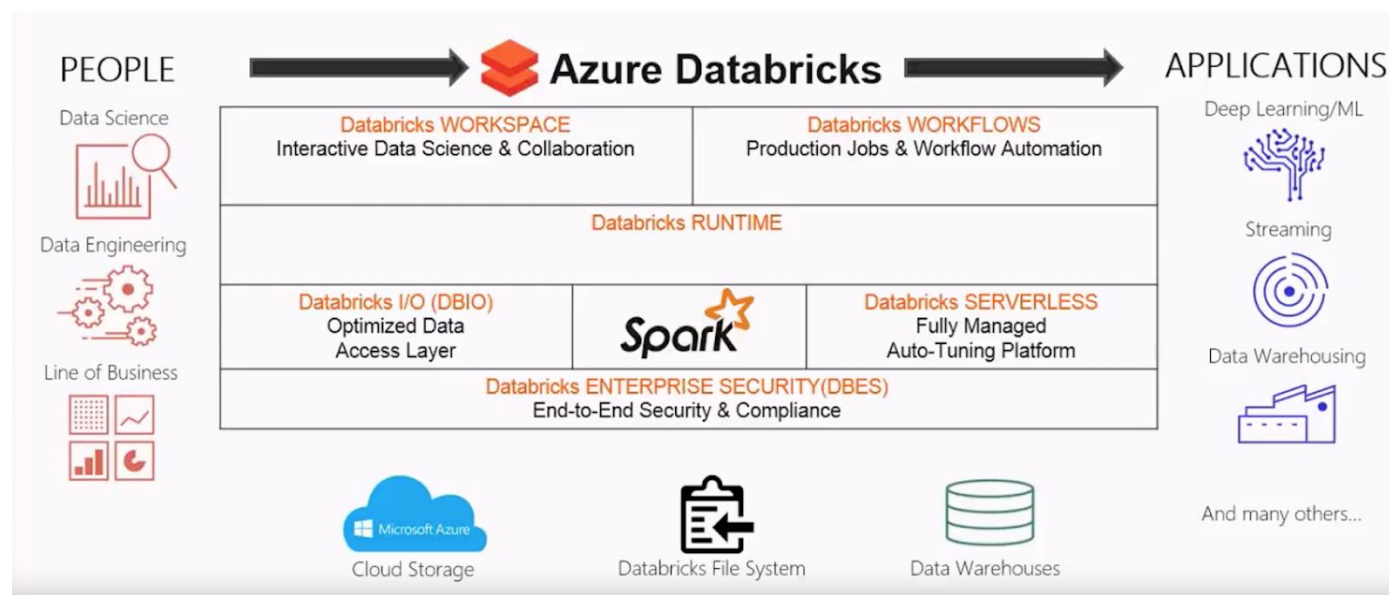
Que es Apache Spark?

- Computación en clúster de código abierto
- Procesa rápidamente grandes cantidades de datos El estándar para "Big Data"
- Desarrollado en UC Berkeley en 2009



Que es Databricks?

- Databricks es una empresa fundada por el equipo que creó Spark
- Asociada con Microsoft para crear **Azure Databricks**



Fuente: <http://drsalbertspijkers.blogspot.com/2018/04/microsoft-azure-databricks.html>

Terminologia

1. Workspaces
2. Notebooks
3. Librerias
4. Tablas
5. Clusters
6. Jobs

databricks

Home

Workspace

Recent

Tables

Clusters

Jobs

Search

Log ETL: Prod - Delta Update

[< All Jobs](#)

Delete

Log ETL: Prod - Delta Update

Job ID: 133

Task: Notebook at [/Users/etl@databricks.com/log-etl/prod/common/spark-log-delta](#) - [Edit](#) / [Remove](#)

- Dependent Libraries:** [Add](#)
 - process-2.1-assembly_deploy-c8ff10d3-20170331151625.jar - (Jar) [Remove](#)

Cluster: Driver: c3.4xlarge, Workers: c3.4xlarge, 330 GB, Spot, fall back to On-Demand, Spark Master (Experimental, Scala 2.11), us-west-2a [Edit](#)

Schedule: Every 2 hours (US/Pacific) [Edit](#) / [Remove](#)

Advanced ▾

Alerts: [Edit](#)

- On error: etl-alerts@databricks.com

Maximum Concurrent Runs: 1 [Edit](#)

Timeout: 60 minutes [Edit](#) / [Remove](#)

Retries: None [Edit](#)

Active runs

Run	Start Time	Launched	Duration	Status
Run Now				

Completed in past 60 days

[Latest successful run \(refreshes automatically\)](#)

< Previous 20

Next 20 >

Run	Start Time	Launched	Duration	Status
Run 52498	2017-04-04 10:00:00 PDT	By scheduler	7m 41s	Succeeded
Run 52497	2017-04-04 08:00:00 PDT	By scheduler	12m 37s	Succeeded

Exercícios Aplicados

Se realizarán proyectos prácticos aplicando lo aprendido en la teoría. Para ello deberéis utilizar el Notebook Ejercicios.ipynb