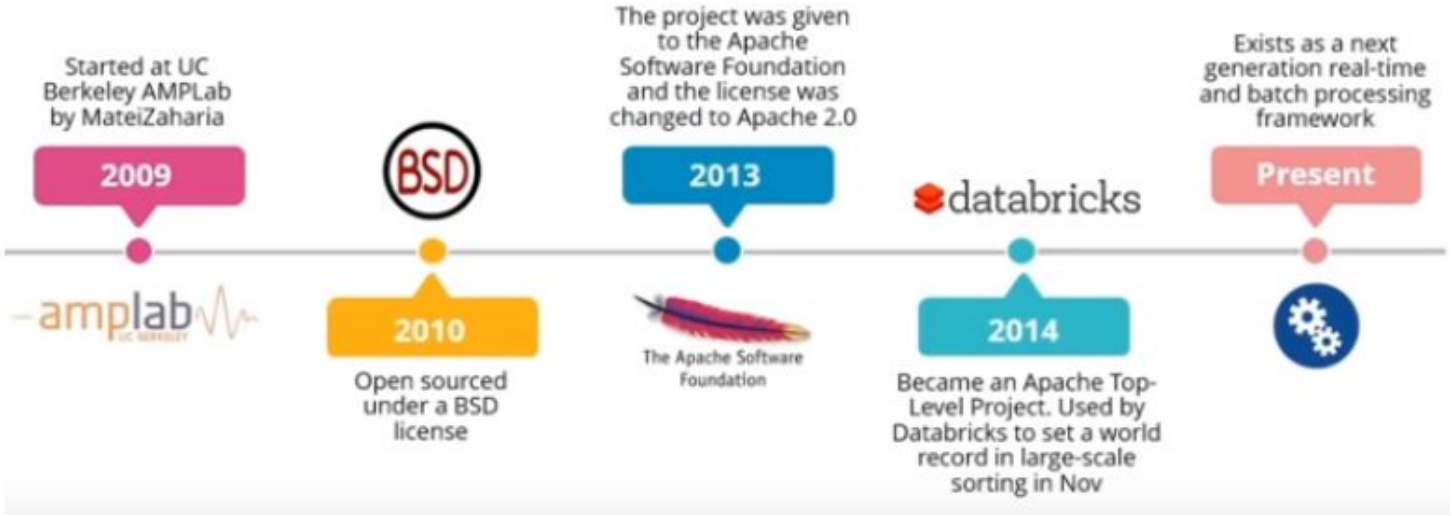
Automatización e Integración datos para IA
Docente. Wilmer López López
Estudiante. Fernando Andrés Roa Martín

UNIVERSIDAD CENTRAL

DEFINICIÓN E HISTORIA

Framework de programación de código abierto para el procesamiento de datos masivos de forma distribuida. “Divide y vencerás”.



http://www.igfasouza.com/blog/what-is-apache-spark/

UNIVERSIDAD CENTRAL

CARACTERÍSTICAS

- Trabaja en RAM. Velocidad 100x vs Hadoop.
- Flexible. Admite lenguajes como Java, Scala, Python, R y SQL.
- Procesamiento en tiempo real.
- Tolerante a fallos.

COMPONENTES

https://www.youtube.com/watch?v=znBa13Earms Min 9:32

UNIVERSIDAD CENTRAL

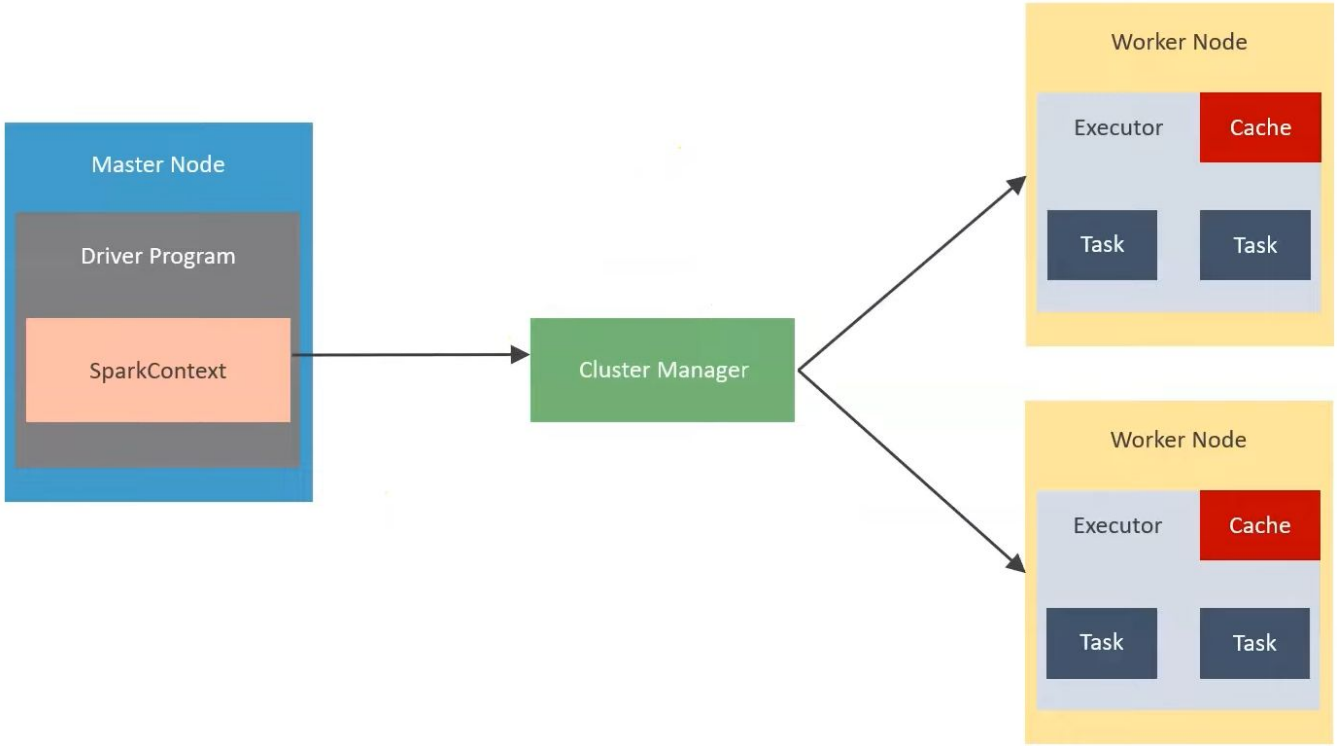
RESILIENT DISTRIBUTED DATASETS - RDD

Son colecciones de objetos inmutables y tolerantes a fallos, distribuidos en los nodos del cluster, que pueden ser operados en paralelo.

Se pueden realizar dos operaciones sobre los RDD:

- Transformation.
- Action.

ARQUITECTURA



https://www.youtube.com/watch?v=znBa13Earms Min 29:25

UNIVERSIDAD CENTRAL

SPARK CLUSTER MANAGERS

Los administradores de cluster pueden ser:

- Standalone.
- Mesos.
- Hadoop YARN.
- Kubernetes.

https://www.youtube.com/watch?v=znBa13Earms Min 33:02

UNIVERSIDAD CENTRAL