

SOBRE NOSOTRAS



MARIA VALENTINA ARIZA

Data Engineer, DataKnow



LAURA LÓPEZ BEDOYA

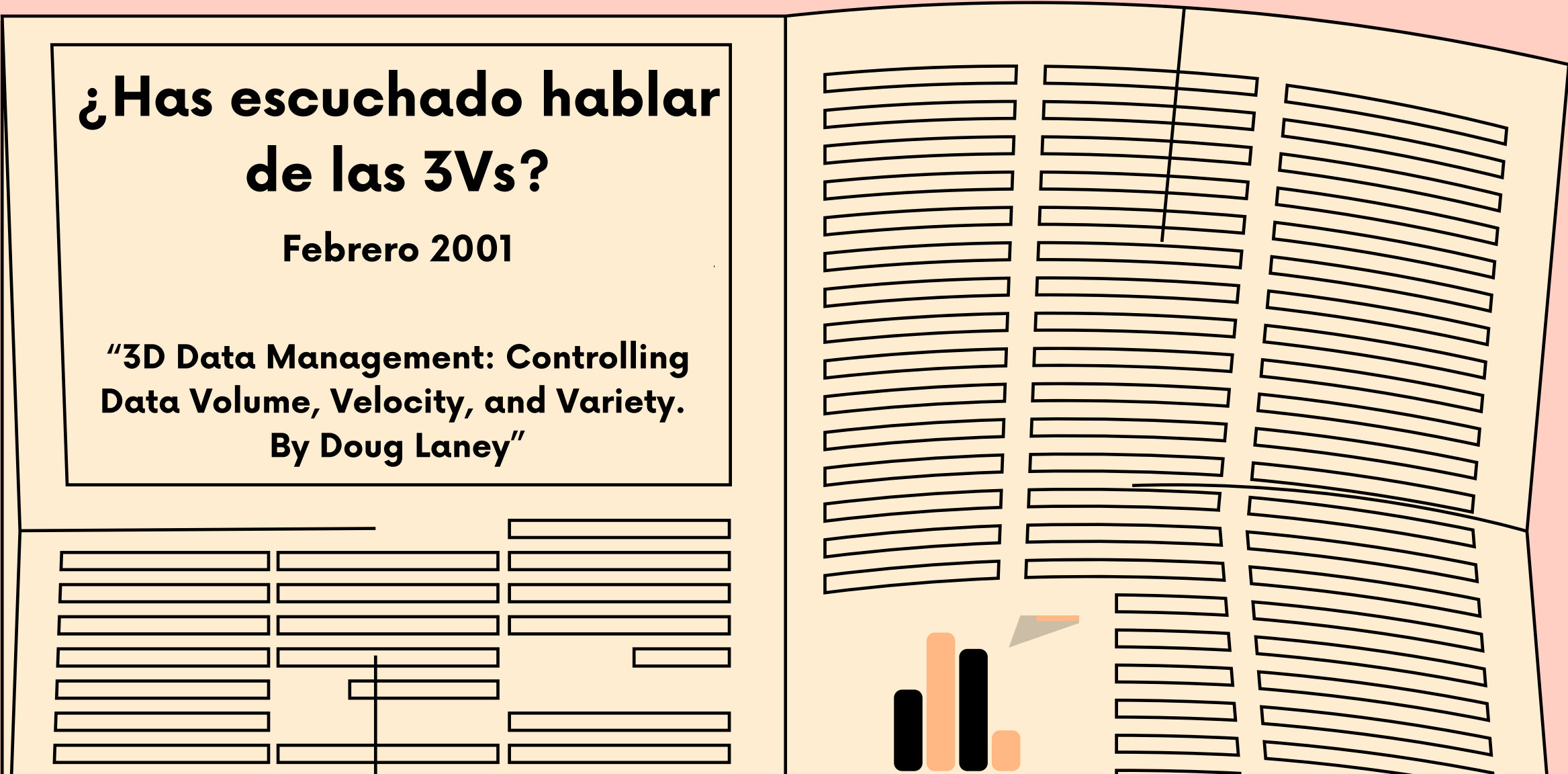
Data Science Analyst, Accenture

Agenda

- Big data
- Apache Spark
- Arquitectura de Spark
- PySpark
- Manos a la obra (Databricks)
- Recursos
-

BIG DATA

Hablamos de big data cuando tenemos grandes volúmenes de datos y no solo eso, son datos de diferentes fuentes, variados, y en ocasiones generados en tiempo real.



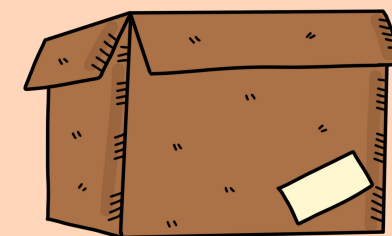
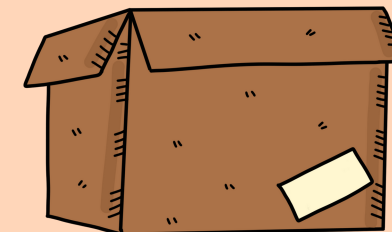
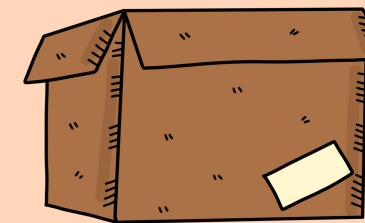
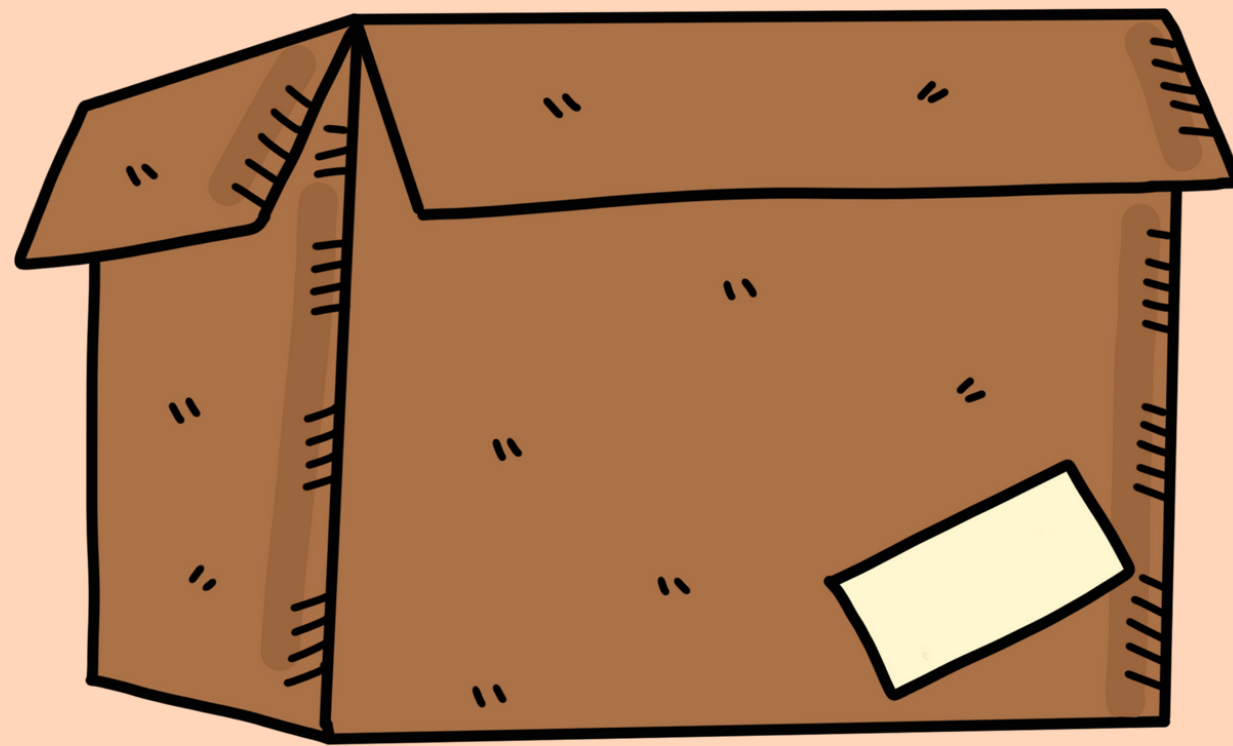
LAS 3Vs EN BIG DATA

1	Volumen
2	Velocidad
3	Variedad

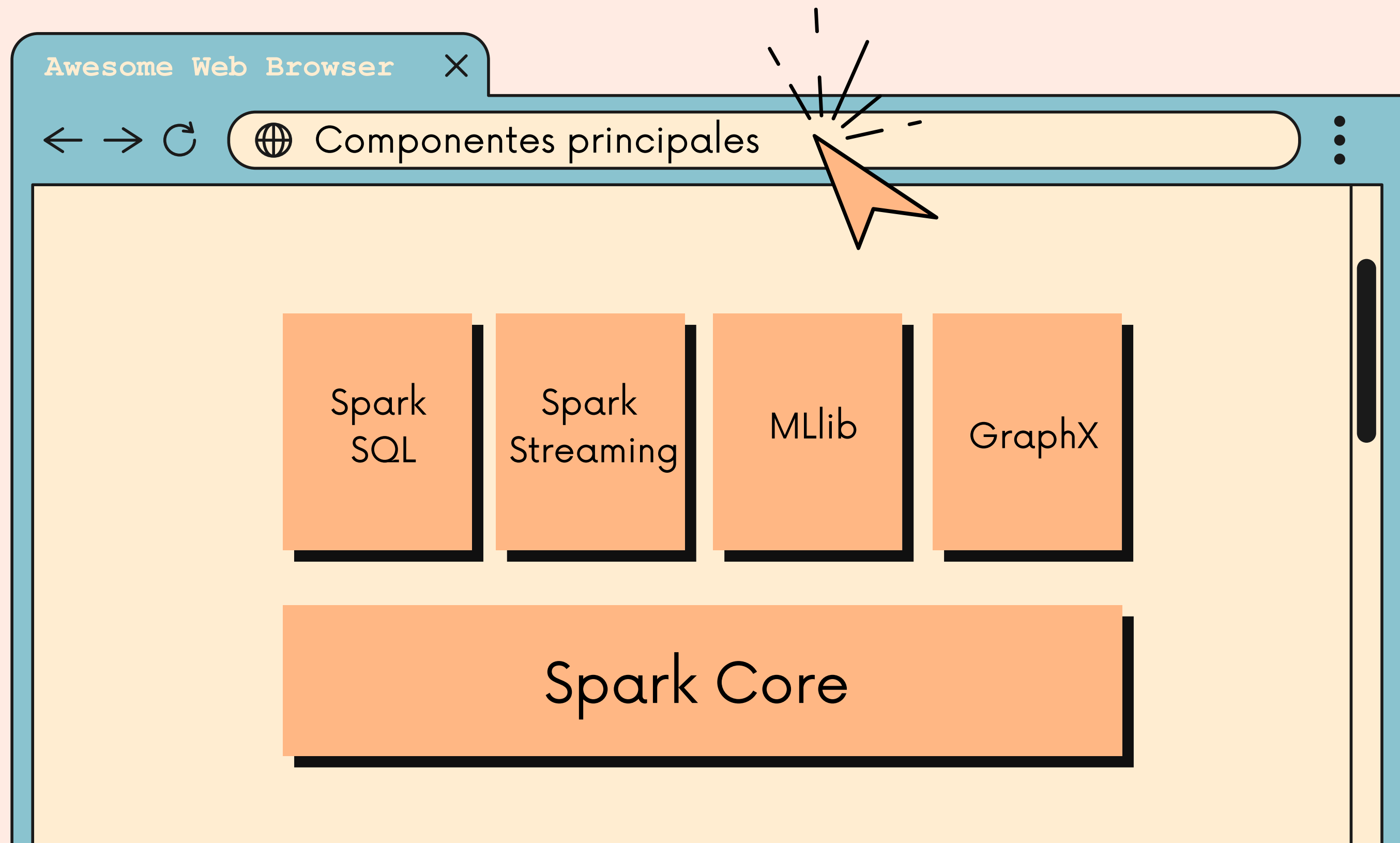


APACHE SPARK

Es un framework distribuido para ejecutar código en paralelo en muchas máquinas diferentes, conocido como procesamiento distribuido.

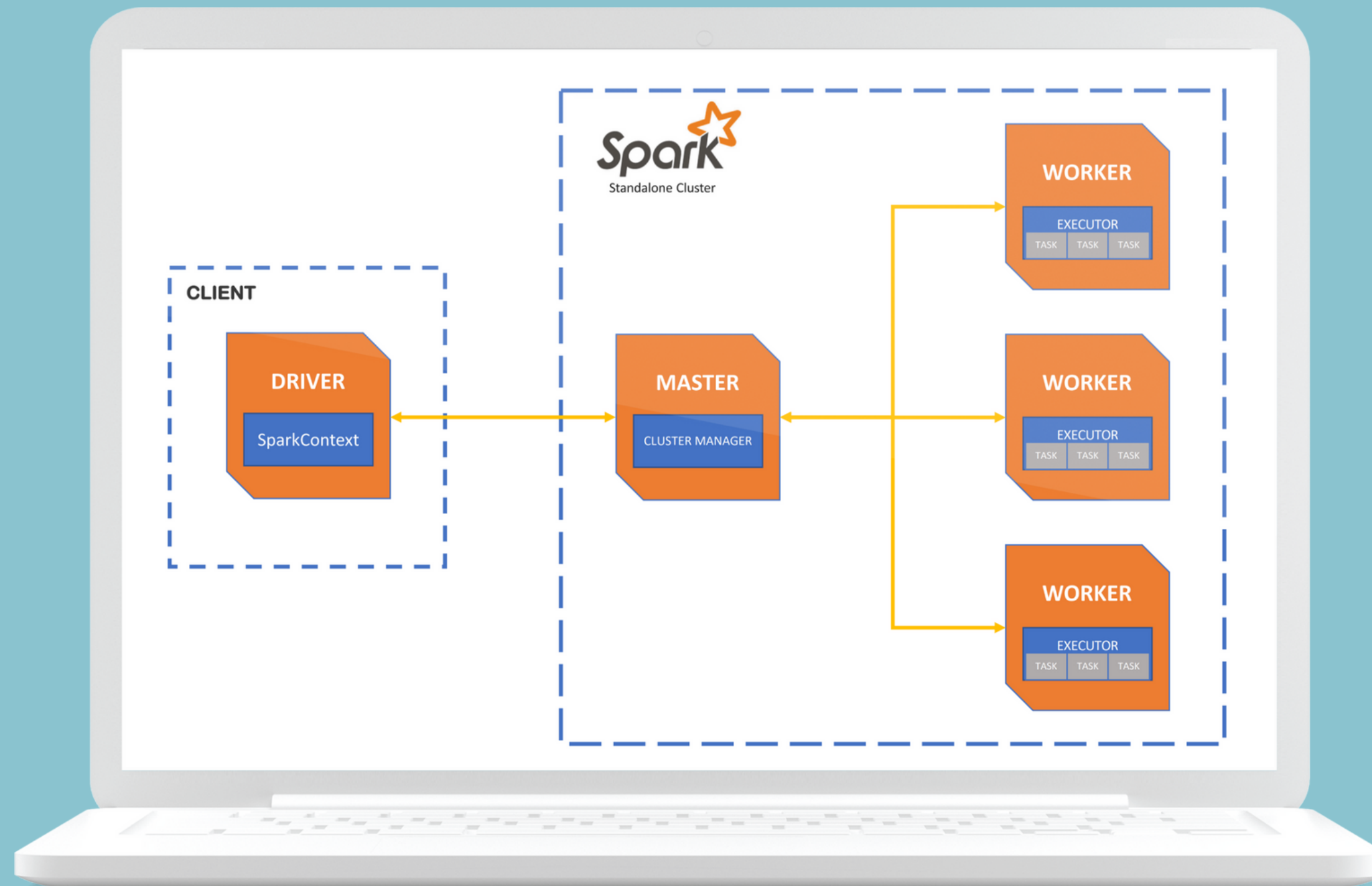


APACHE SPARK



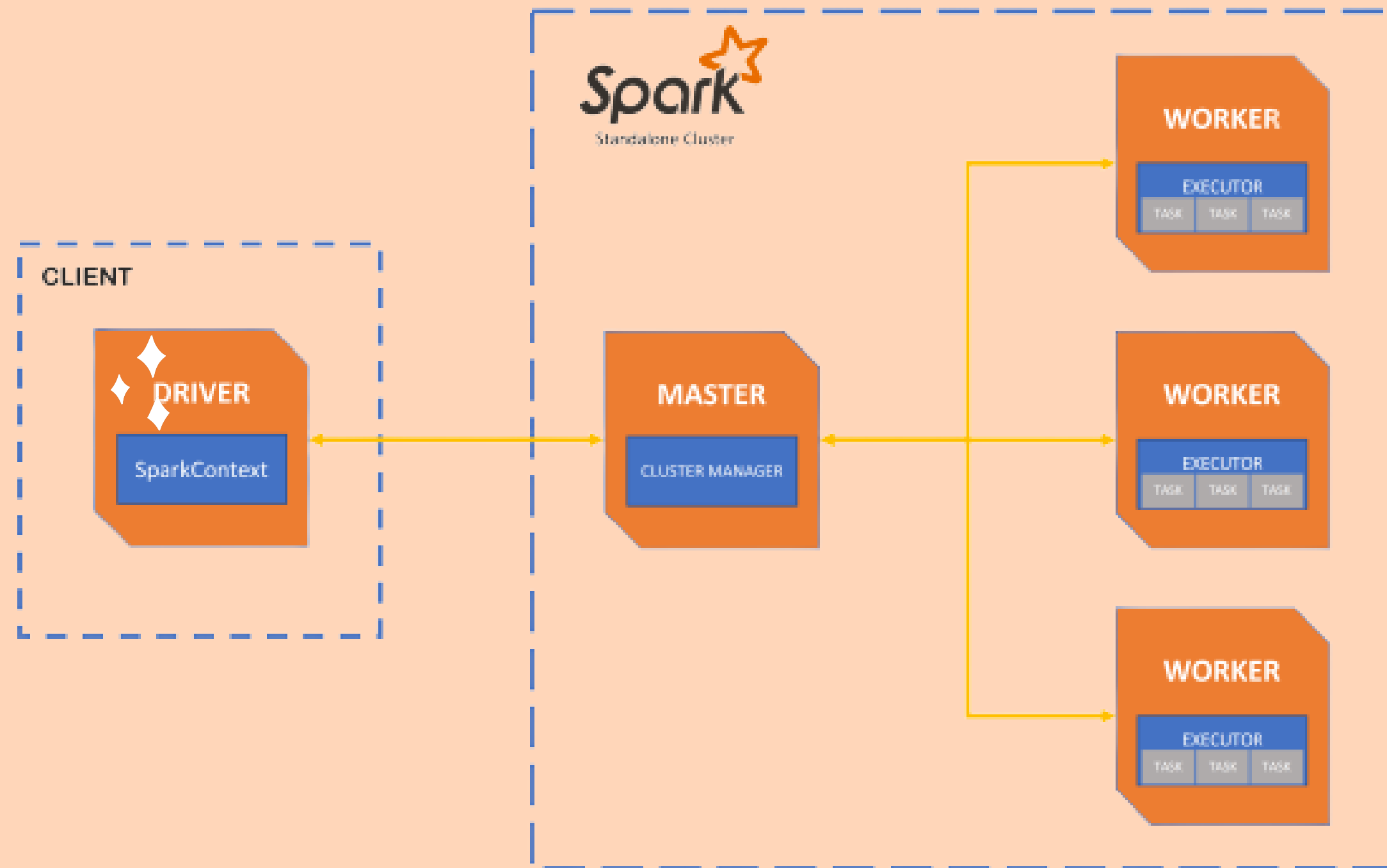
ARQUITECTURA SPARK

Apache Spark sigue una arquitectura maestro/esclavo con un administrador de clúster (Cluster Manager).



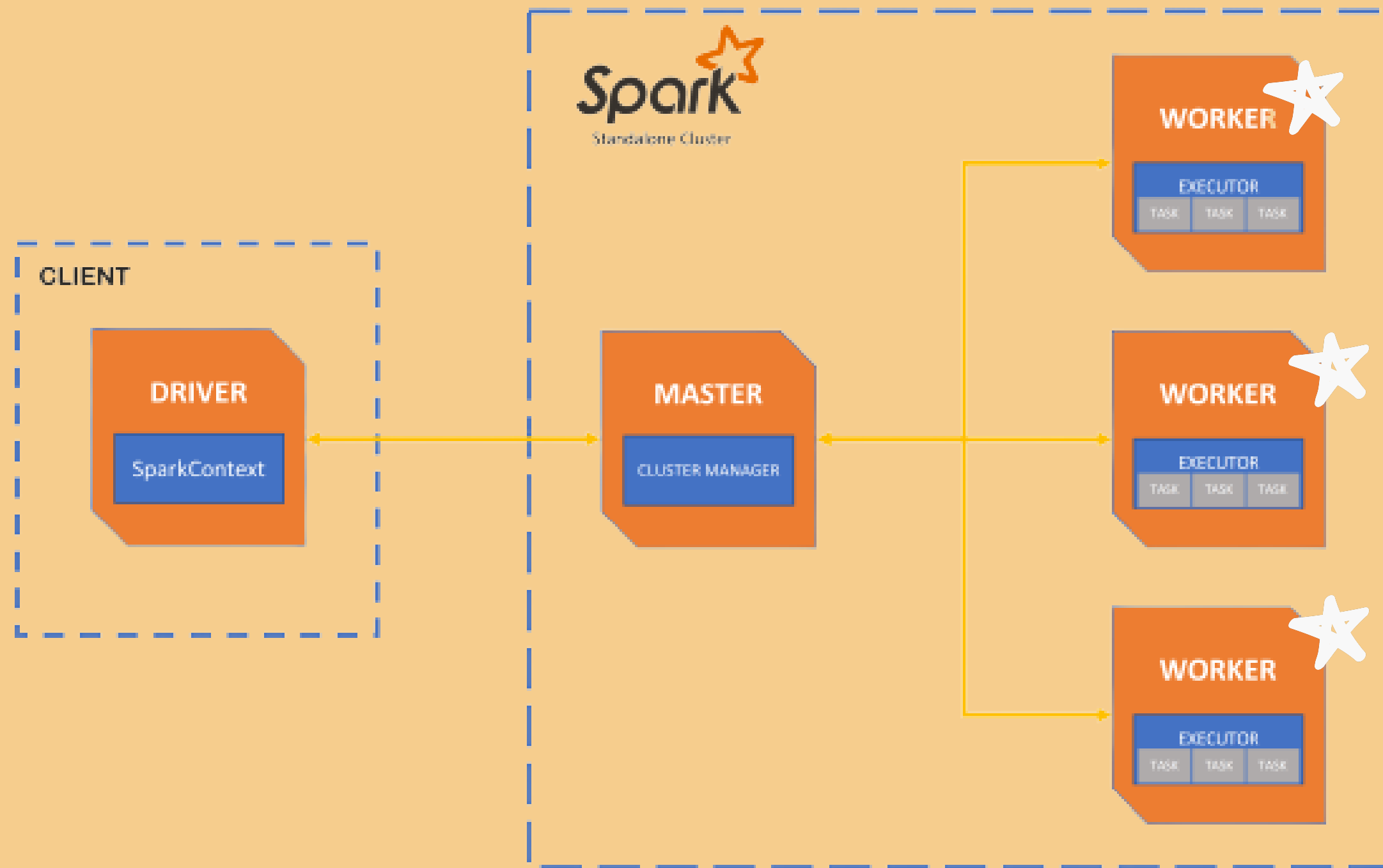
CONTROLADOR

El controlador (driver) es un proceso que se ejecuta en una de las máquinas y es responsable del ciclo de vida completo de la aplicación.



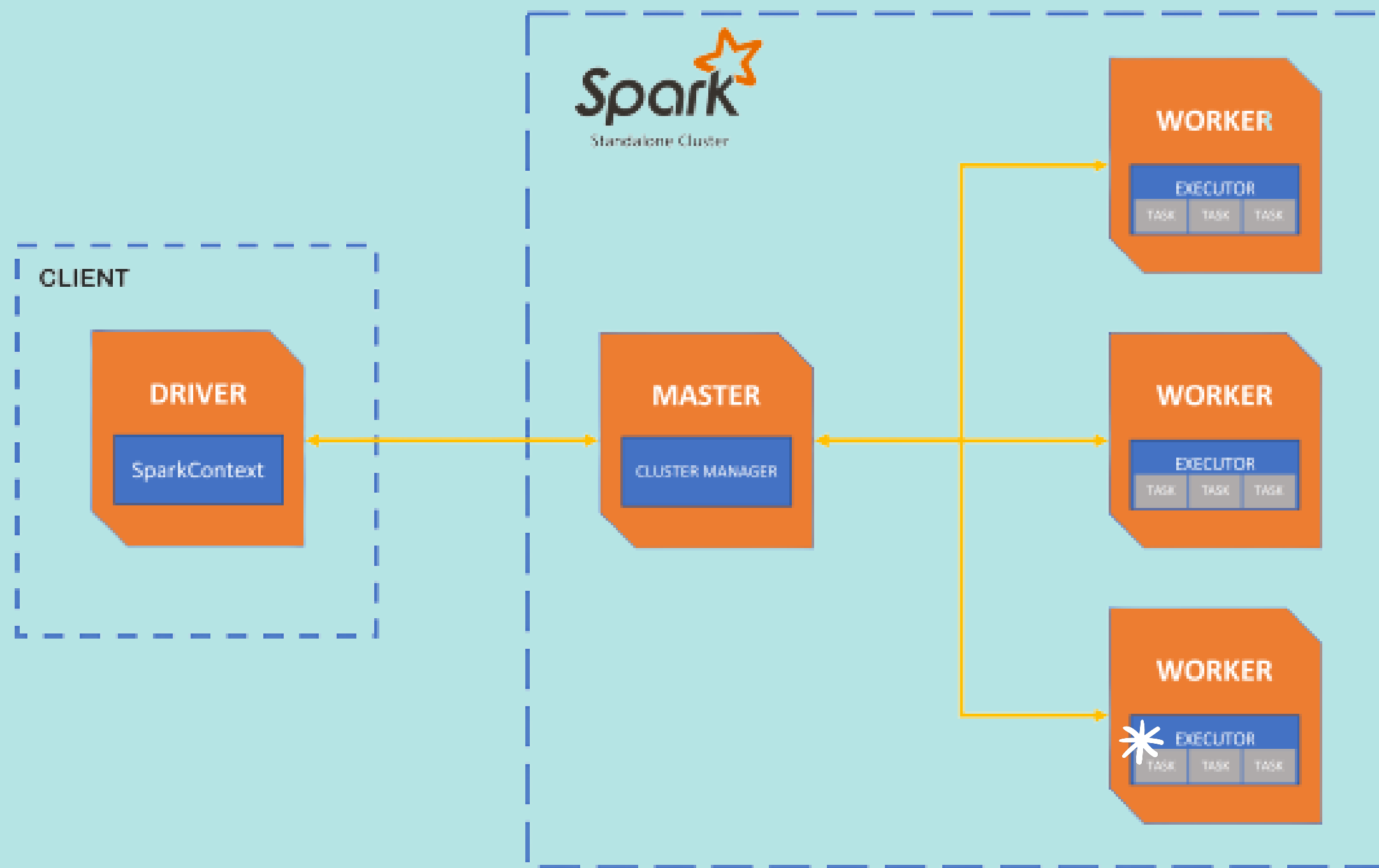
EJECUTORES

Los ejecutores son los procesos de trabajo que se ejecutan en otras máquinas del clúster. Puede haber varios ejecutores ejecutándose en la misma máquina.

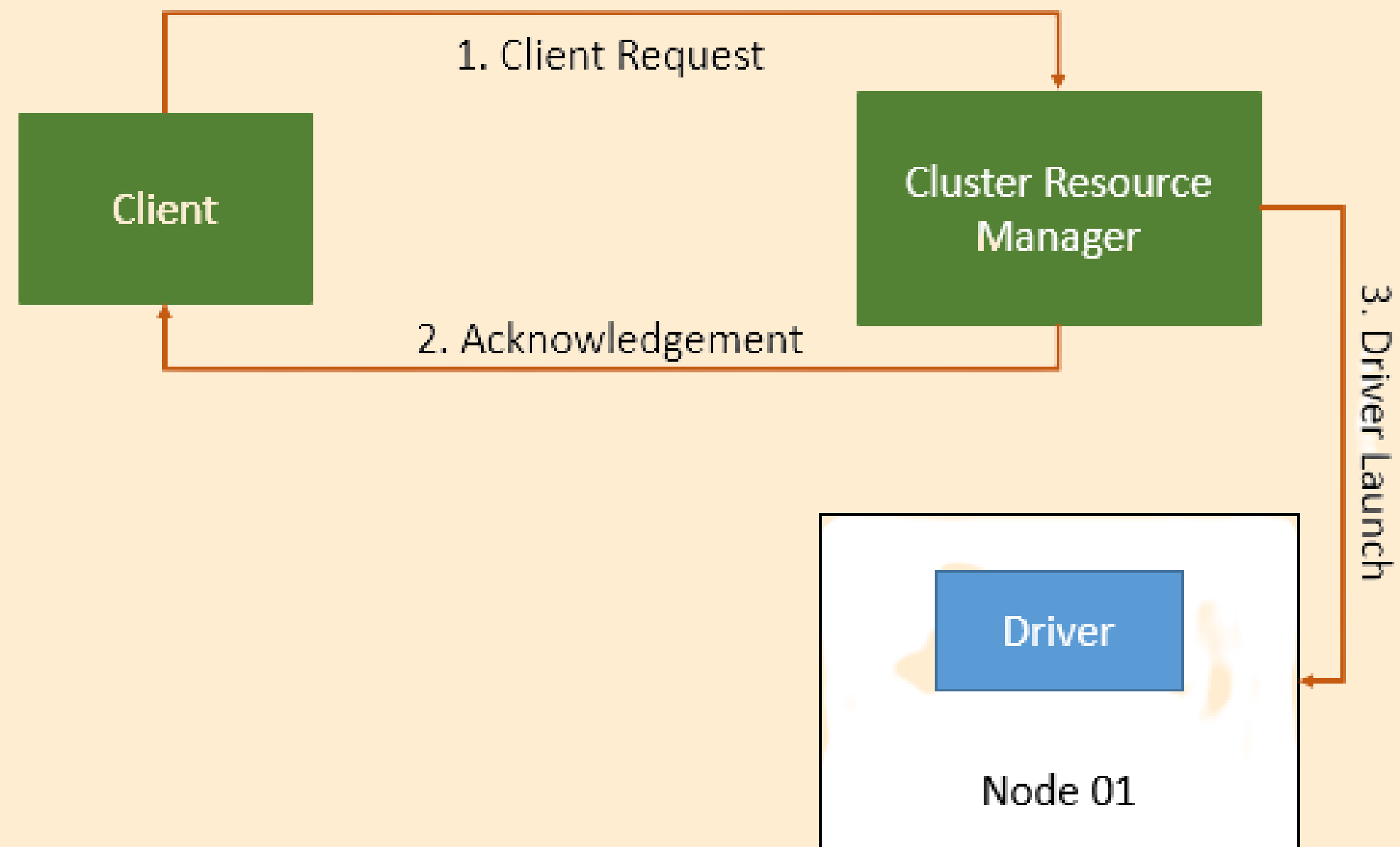


TAREAS

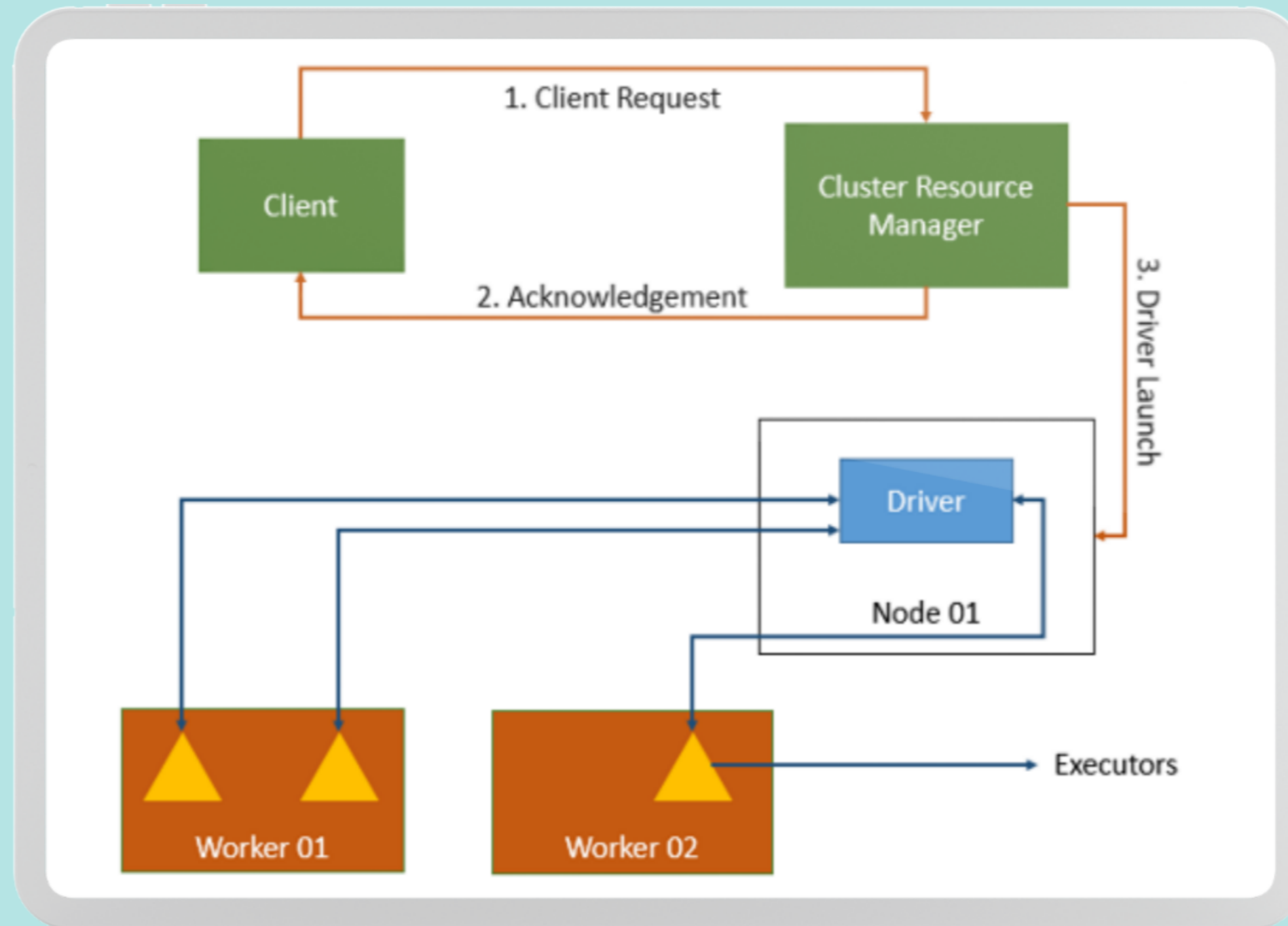
Una tarea es el trabajo a realizar en una partición (conjunto de filas). Estos son asignadas por el Driver a los ejecutores.



CICLO DE VIDA DE LA EJECUCIÓN



CICLO DE VIDA DE LA EJECUCIÓN



PYSPARK

PySpark es la biblioteca de Python para usar Spark que nos permite realizar tareas similares a Pandas.

NEW
NEW
NEW
NEW

~~~~~

**La comunidad Apache Spark ha lanzado una herramienta PySpark que se puede utilizar para admitir Python con Spark.**

~~~~~

~~~~~

**Además, soporta Spark SQL, DataFrame, Streaming y MLlib (para uso de Machine Learning).**

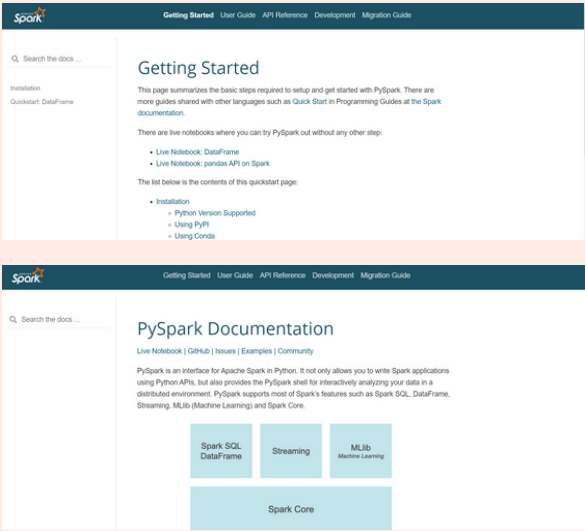
~~~~~



RECURSOS

Documentación oficial de Spark

Documentación Oficial de PySpark



DataCamp



Introduction to PySpark



Cleaning Data with PySpark



Machine Learning with PySpark

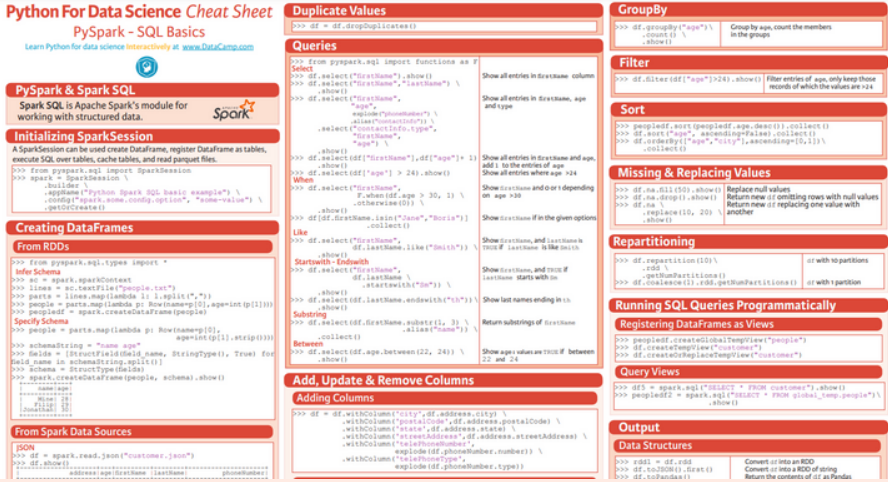


Feature Engineering with PySpark



Big Data Fundamentals with PySpark

Cheat Sheet



CONTACTAR CON NOSOTRAS



@valearizag



@lauralpezb

