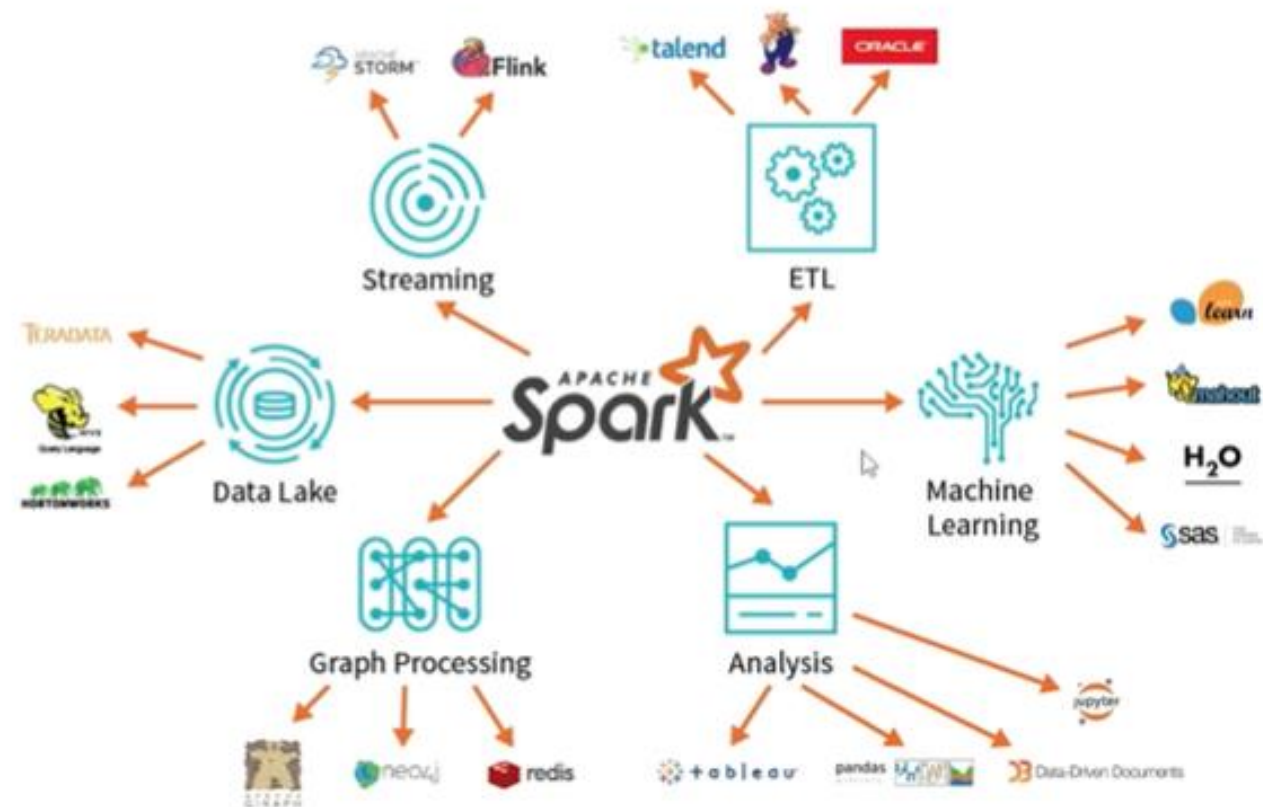


Spark es una solución **Big Data** de **código abierto**. Desarrollado por el laboratorio RAD de **UC Berkeley** (2009).

Se ha convertido en una **herramienta de referencia** en el campo del Big Data.

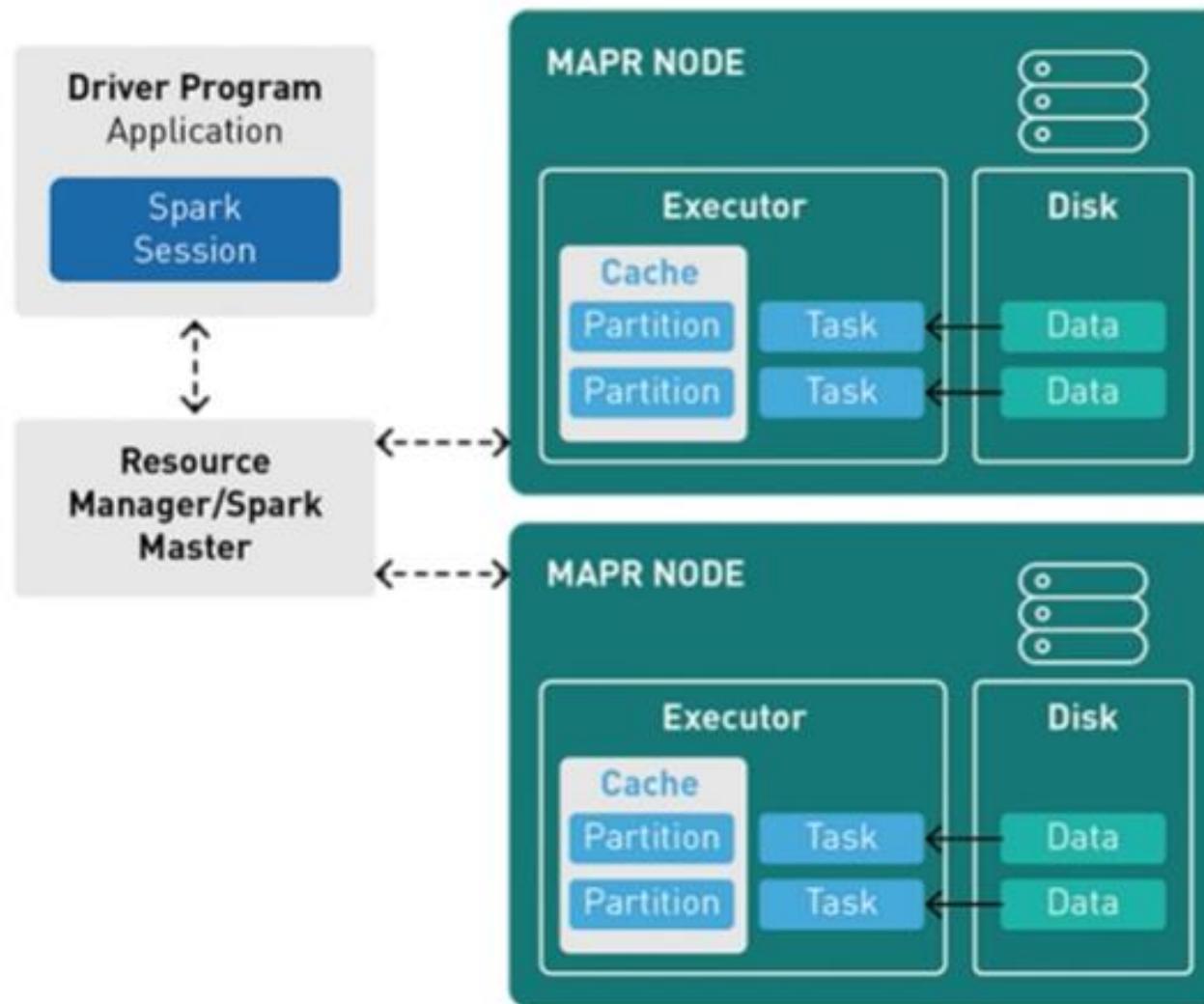


Más fácil y rápida que Hadoop MapReduce.

Diferencias:

- **Spark** mucho **más rápido** al almacenar en caché los datos en la **memoria** vs **MapReduce** en el **disco duro** (más lectura y escritura)
- Spark optimizado para un mejor **paralelismo**, utilización **CPU** e inicio más rápido
- Spark tiene modelo de **programación funcional** más rico
- Spark es especialmente útil para **algoritmos iterativos**





Streaming

MLlib
For Machine Learning

GraphX
For Graph Computing

**Spark SQL &
DataFrames**

Spark Core API

R

Python

Scala

SQL

Java



Data science and Machine learning



SQL analytics and BI



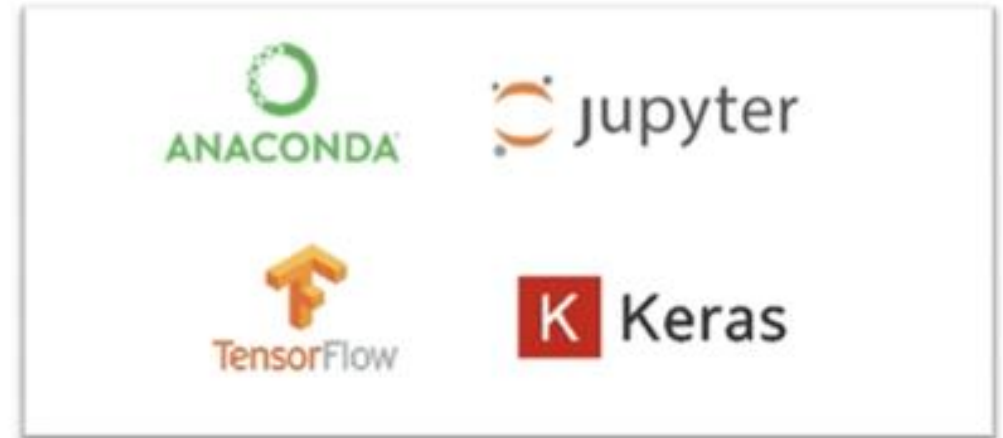
Storage and Infrastructure



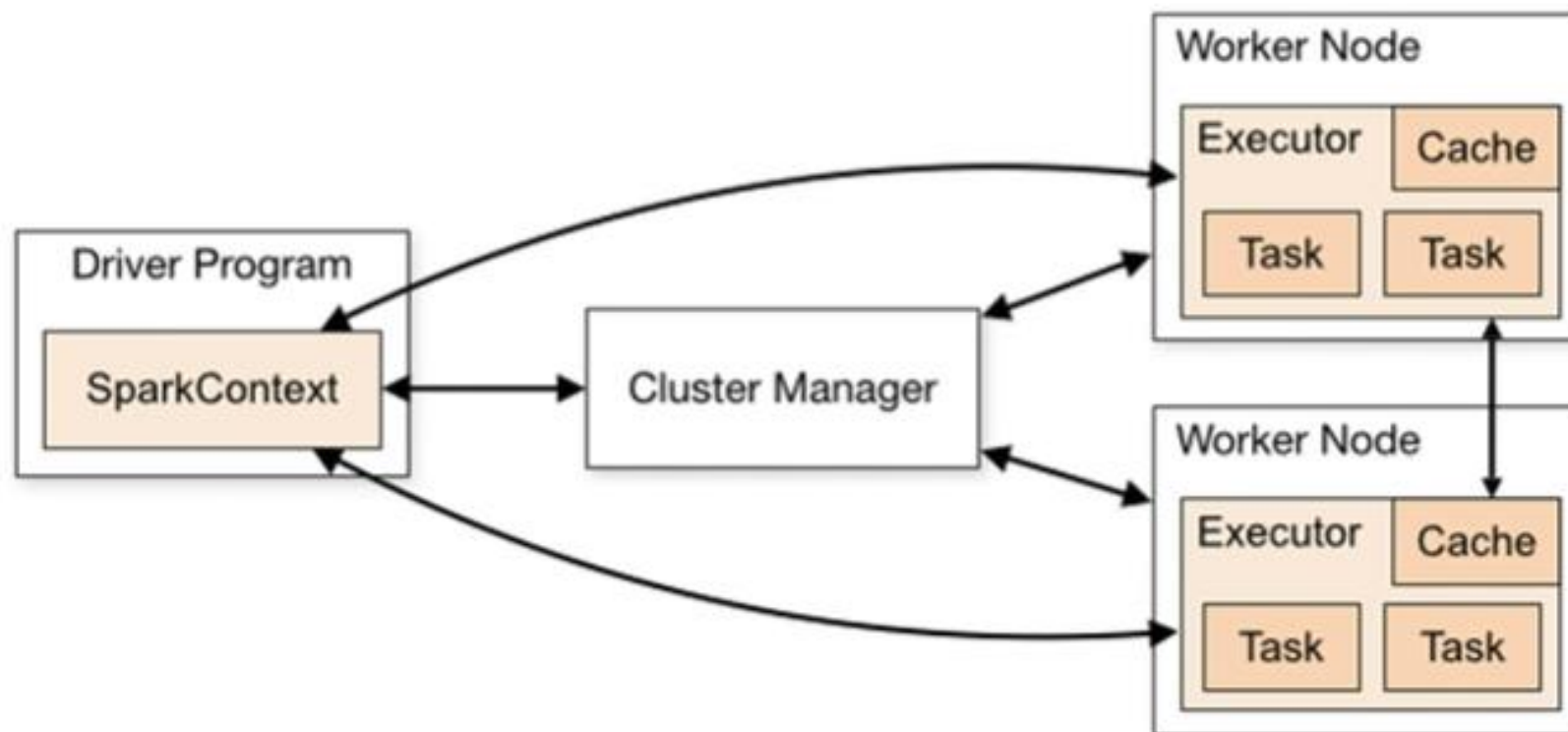
PySpark es una biblioteca Spark **escrita en Python** para ejecutar la aplicación Python usando las **capacidades de Apache Spark**.

Ventajas de PySpark:

- **Fácil** de aprender
- Amplio conjunto de librerías para **ML y DS**
- Gran apoyo de la **comunidad**



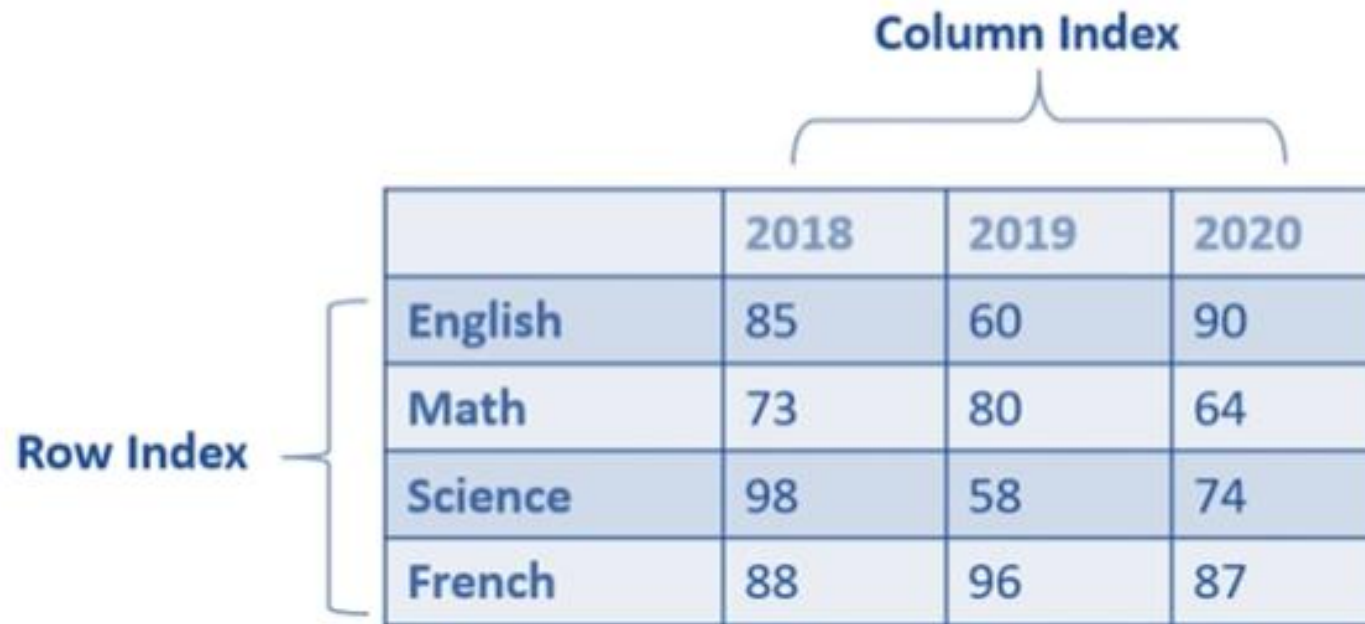
Apache Spark funciona en una **arquitectura maestro-esclavo**. Las **operaciones** se ejecutan en los **trabajadores**, y el **Cluster Manager** administra los recursos.



Spark admite los siguientes administradores de clústeres:

- **Standalone** : administrador de clúster simple
- **Apache Mesos** : es un administrador de clústeres que puede ejecutar también Hadoop MapReduce y PySpark.
- **Hadoop YARN** : el administrador de recursos en Hadoop 2
- **Kubernetes**: para automatizar la implementación y administración de aplicaciones en contenedores.

Los **DataFrames** son de naturaleza **tabular**. Permiten varios formatos dentro de una misma tabla (**heterogéneos**), mientras que cada variable suele tener valores con un único formato (**homogéneos**).
Similares a las tablas SQL o a las hojas de calculo.

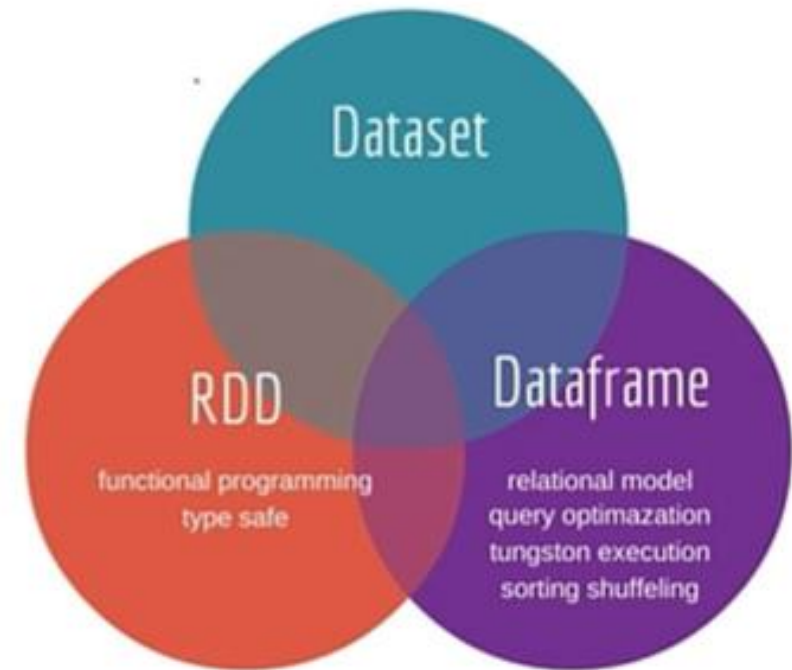


The diagram illustrates a DataFrame as a table. A bracket on the left labeled "Row Index" points to the subject names in the first column. A bracket on top labeled "Column Index" points to the years in the first row. The table contains the following data:

	2018	2019	2020
English	85	60	90
Math	73	80	64
Science	98	58	74
French	88	96	87

Algunas de las ventajas de trabajar con Dataframes en Spark son:

- Capacidad de procesar una **gran cantidad de datos** estructurados o semiestructurados
- Fácil **manejo de datos** e imputación de valores faltantes
- Múltiples formatos como **fuentes de datos**
- Compatibilidad con **múltiples lenguajes**



Los **DataFrames** de Spark **se caracterizan** por: ser distribuidos, evaluación perezosa, inmutabilidad y tolerancia a fallos.

