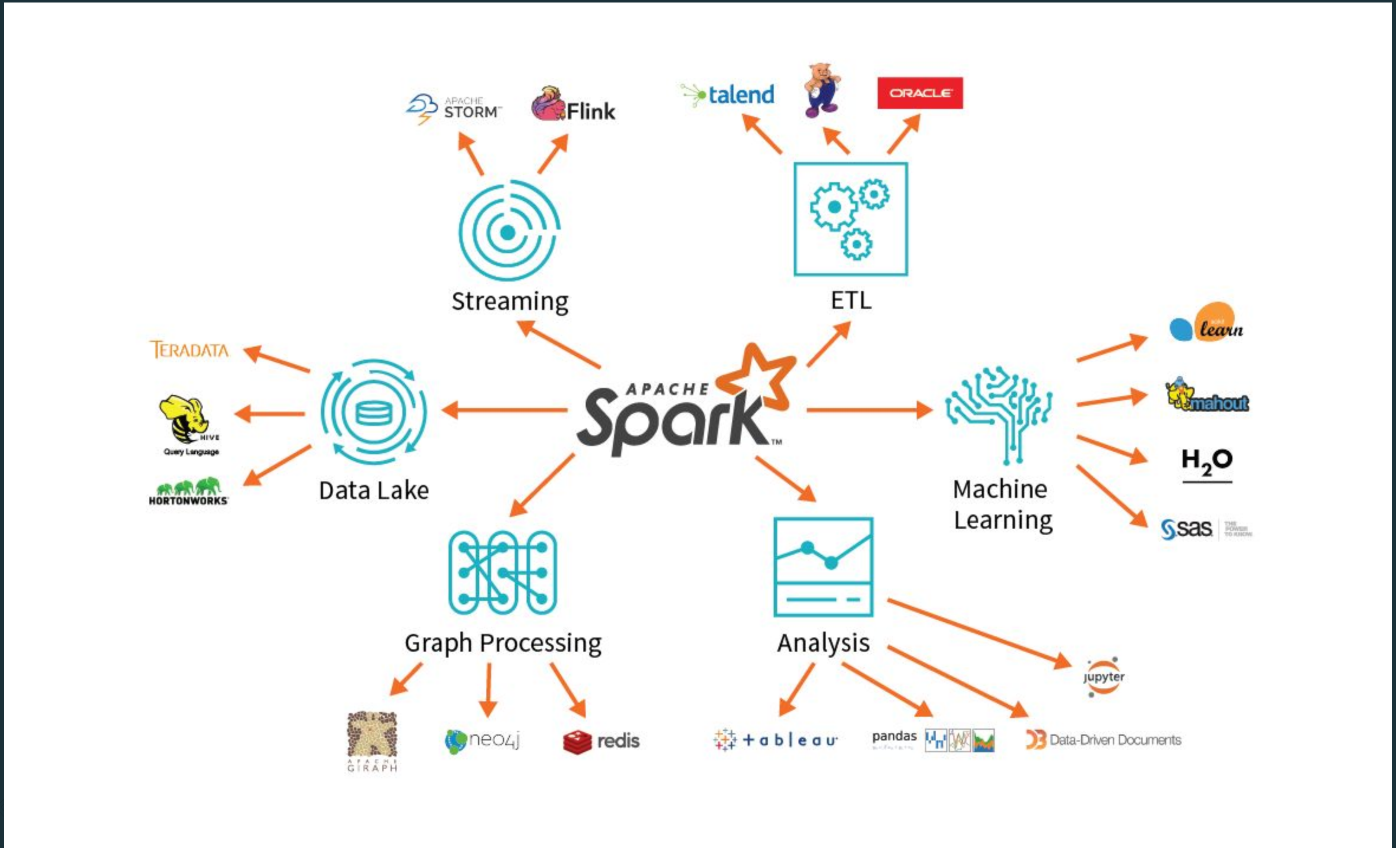
Intro to Apache Spark

databricks

Unified Analytics Engine



Apache Spark

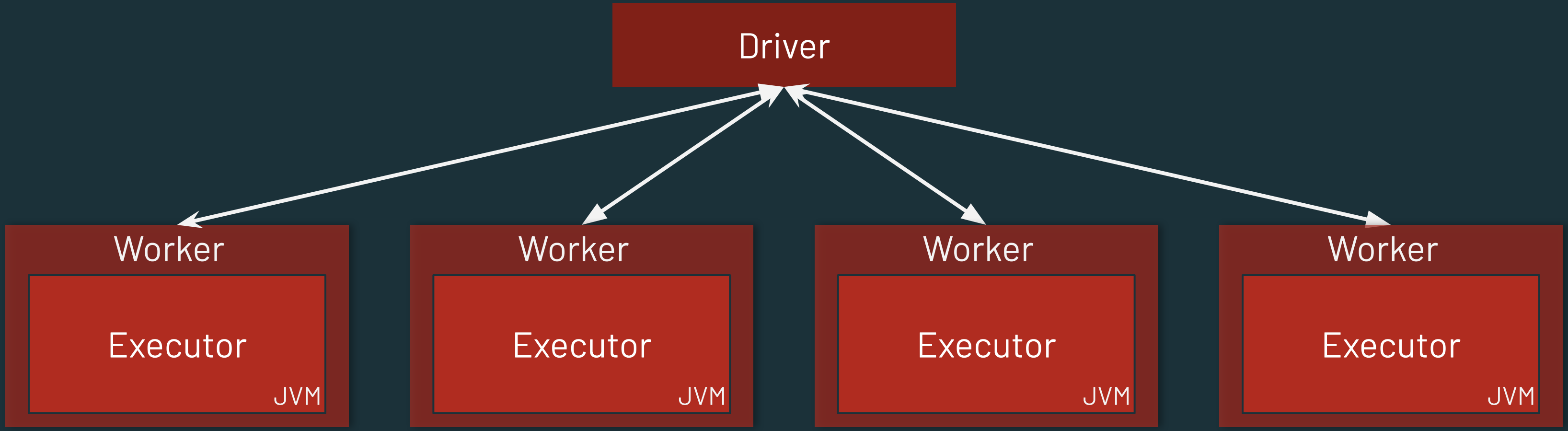
> *“Unified analytics engine for big data processing, with built-in modules for streaming, SQL, machine learning and graph processing”*

- Research project at UC Berkeley in 2009
- APIs: Scala, Java, Python, R, and SQL
- Built by more than 1,400 developers from more than 200 companies

How to process lots of M&Ms?

M&Ms

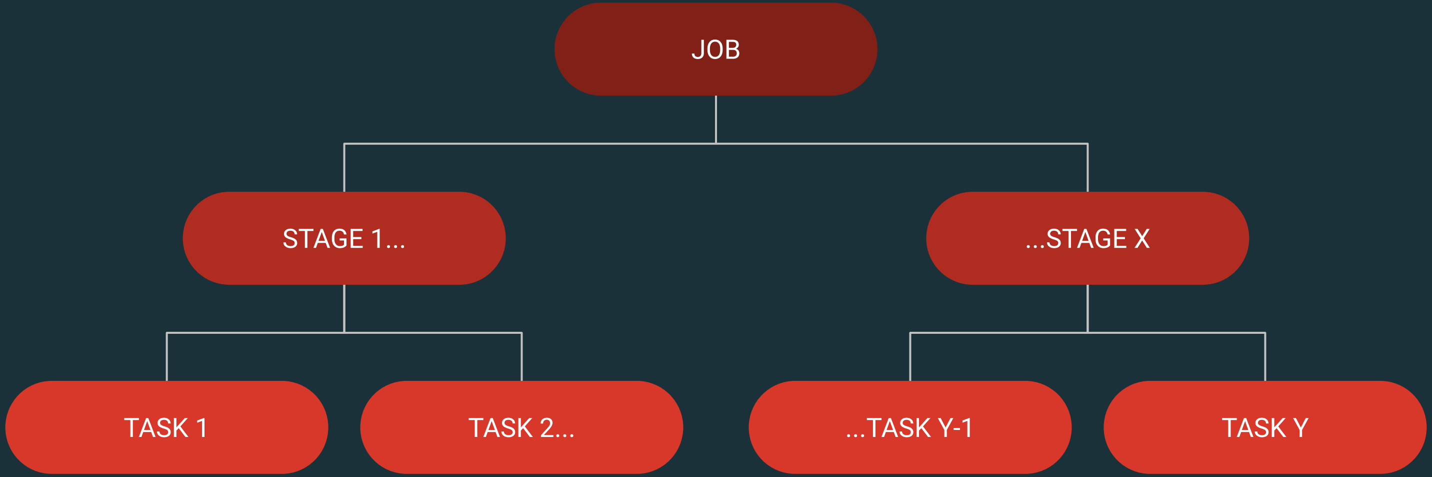
Spark Cluster



One Driver and many Executor JVMs

Back to M&Ms

Spark Jobs



Each Spark **Job** is broken down into a number of **Stages**. Each Stage is broken down into a number of **Tasks**.

A **Stage** boundary occurs when data must be exchanged between nodes. This is called a **Shuffle**.