Automated deployment of BigData Cluster in the Cloud

Hands-on part 2 - School on Open Science Cloud 2018

Tracolli Mirco - INFN sec. Perugia

Outline

- Clusters, Cloud and BigData
- Apache Spark for distributed tasks
- Custom Cluster with DODAS

Cluster

A group of coupled computers that work together.

We need that to increase our computational power.

It’s not trivial to configure and manage.

Cloud

An high level view of services provided by cluster of computers.

It gives a more user friendly approach to interact with calculus computing resources.

BigData

The nightmare of data analysts and treasure for Machine Learning people.

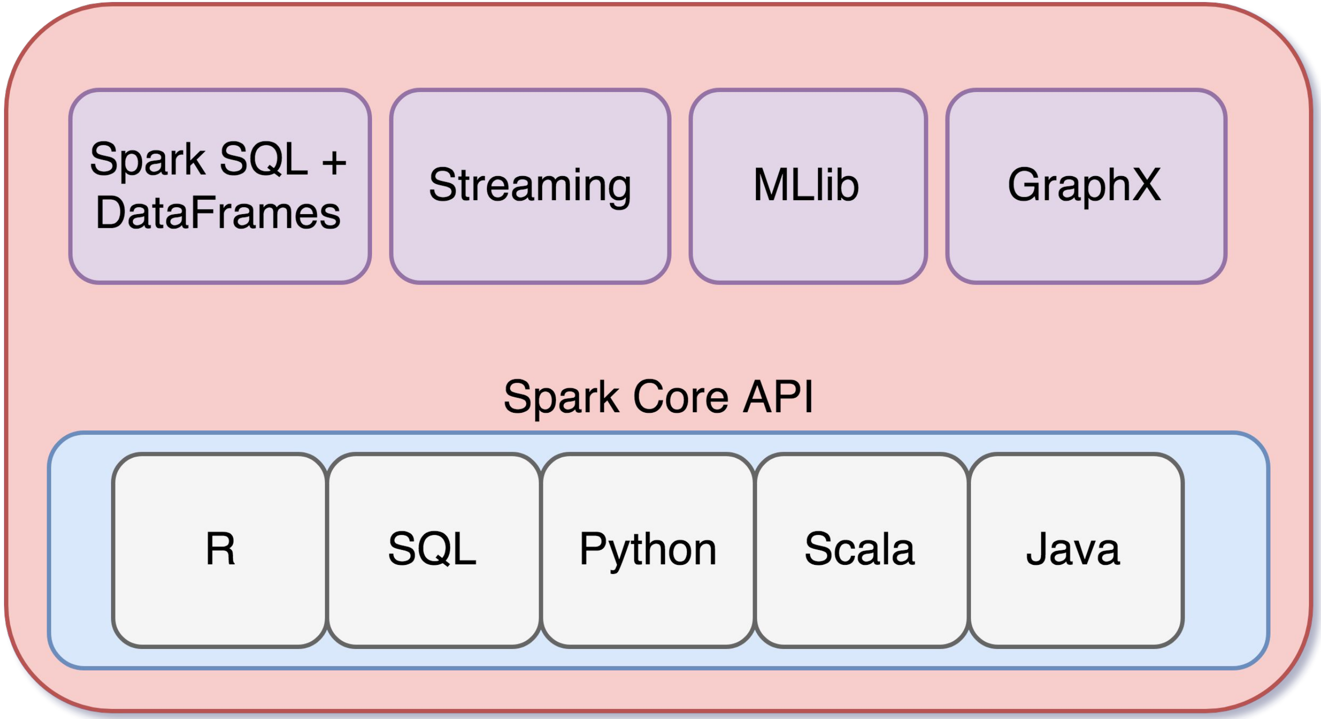
They need an appropriate environment to be managed and it’s also a problem the storing of the data themselves.

Apache Spark

Apache Spark is an open-source powerful distributed querying and processing engine.

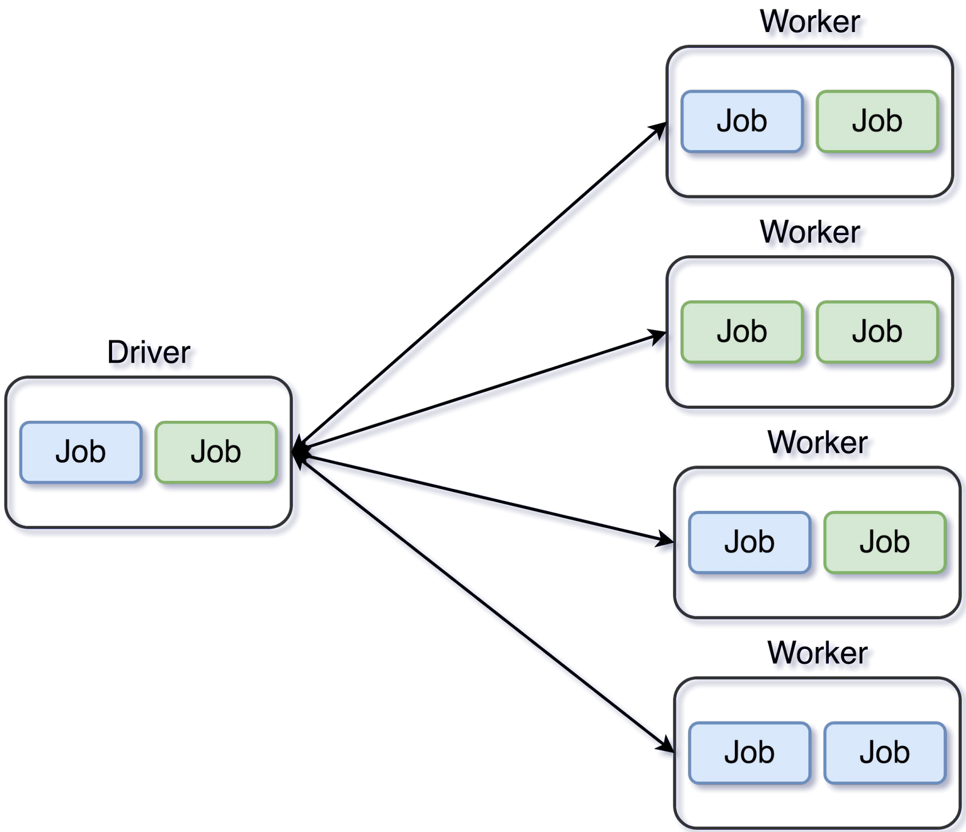
It’s quite easy to write a task for this engine (using its API) and it allows you to process a large amount of data.

Spark Ecosystem



Spark Ecosystem

Spark Workflow

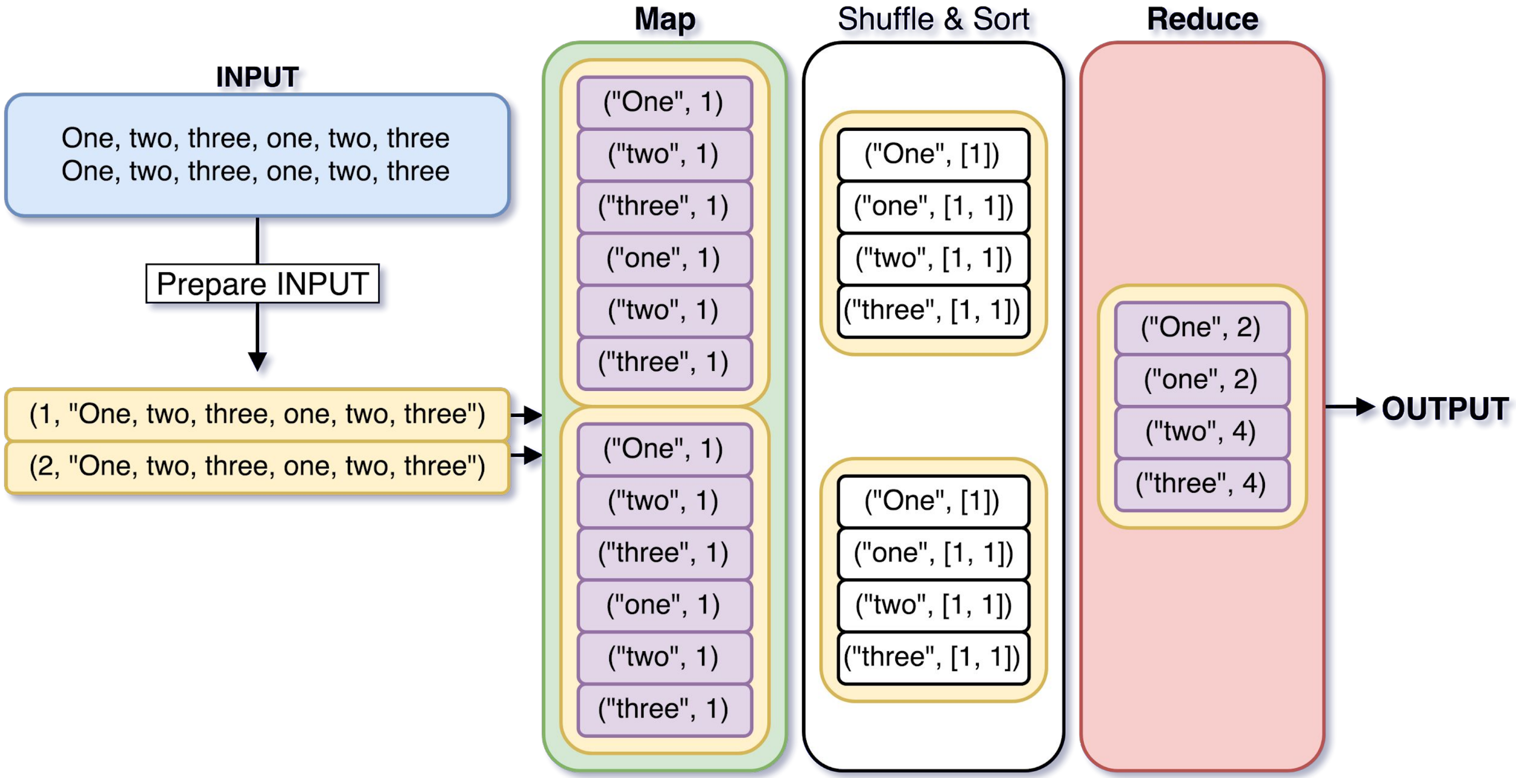


MapReduce

Spark workflow matches the MapReduce programming model. This model is a paradigm with which you split and manipulate your data in a certain way, using principally the two function that give the name to the model:

- Map: apply something to all the elements
- Reduce: select and extract some elements

MapReduce - Example



Custom Cluster with DODAS

Dynamic On Demand Analysis Service: it's a Platform as a Service tool built combining several solutions and products developed by INDIGO-DataCloud. Currently, it’s a Thematic Service in the context of EOSC-hub H2020 project.

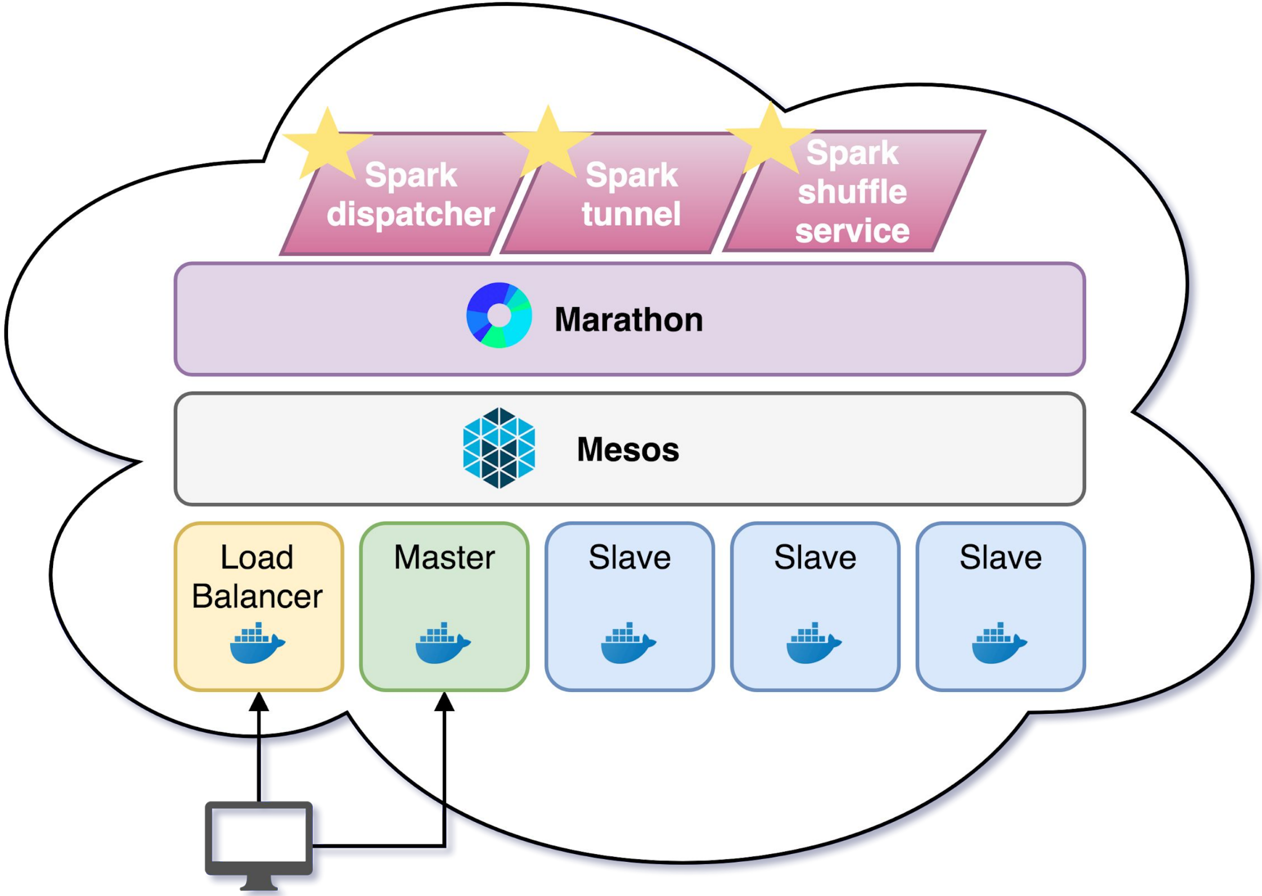
In detail, DODAS is a service for generating over cloud resources an on-demand container based solution.

Custom Cluster with DODAS

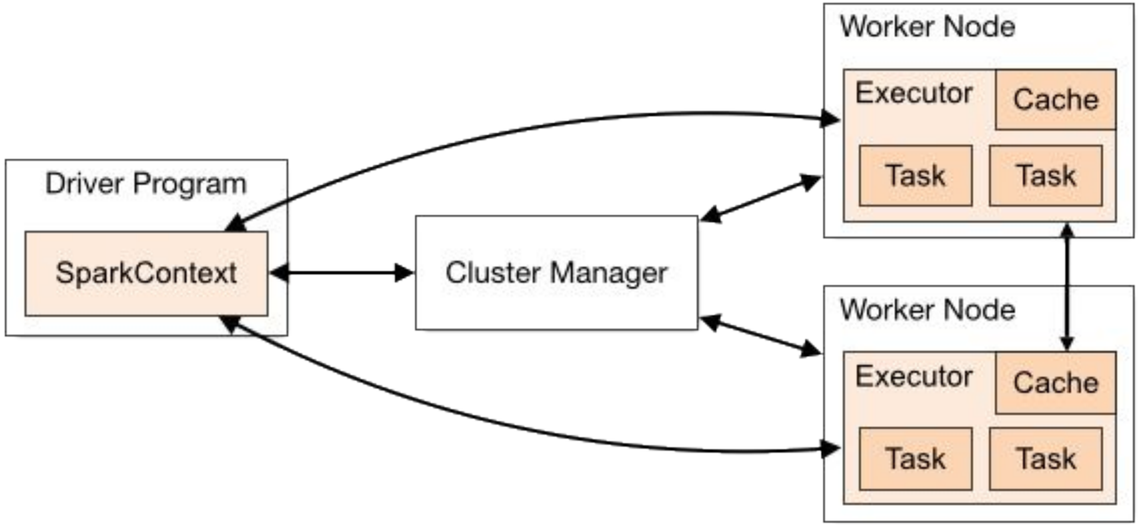
The cluster will use as resource manager Apache Mesos.

Plus we will use some custom Docker containers managed by Apache Marathon, a framework for Mesos.

Cluster Schema



Spark Workflow in our cluster



Documentation

https://dodas-ts.github.io/SOSC-2018/