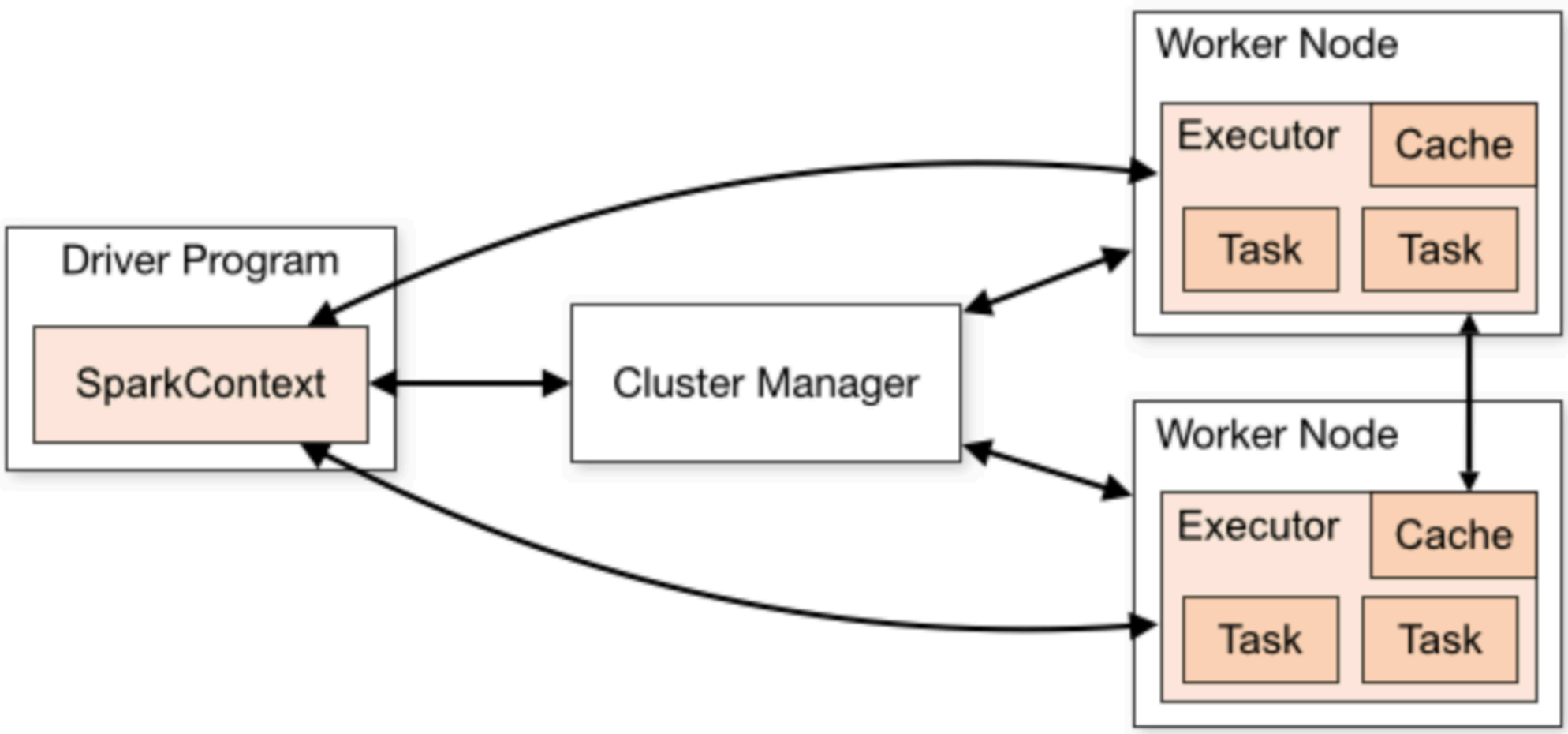
K-means in spark

Thanks to Paul-Henri PERRIN, PhD student at Data Science team @ Dauphine

Spark run time support



Spark-submit, example

```
>spark-submit \
--master yarn
--deploy-mode cluster \
--executor-cores 4 \
--num-executors 11 \
--executor-memory 2g \
--conf spark.yarn.executor.memoryOverhead=2g \
--conf spark.driver.memory=5g \
--conf spark.driver.cores=2 \
--conf spark.yarn.jars="file:///home/cluster/shared/vms/spark-current/jars/*.jar" \
kmeans-dario-x.py
```

Basic notions

- Borrowed from Francis Back's nice notes at ENS
  (https://www.di.ens.fr/~fbach/courses/fall2013/lecture3.pdf)

K- means clustering is a method of vector quantization. K-means clustering is an algorithm of alternate minimization that aims at partitioning n observations into K clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype to the cluster (see Figure 3.1).

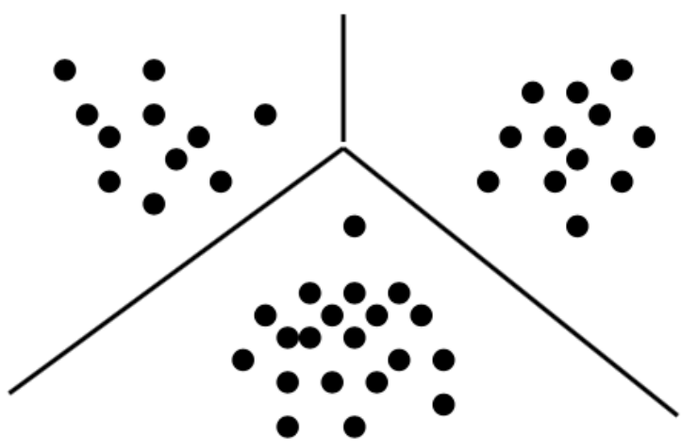


Figure 3.1. Clustering on a 2D point data set with 3 clusters.

Notations

We will use the following notations:

- $x_i \in \mathbb{R}^p$, $i \in \{1, ..., n\}$ are the observations we want to partition.
- $\mu_k \in \mathbb{R}^p$, $k \in \{1, ..., K\}$ are the means where μ_k is the center of the cluster k. We will denote μ the associated matrix.
- z_i^k are indicator variables associated to x_i such that $z_i^k = 1$ if x_i belongs to the cluster k, $z_i^k = 0$ otherwise. z is the matrix which components are equal to z_i^k.

Finally, we define the *distortion* $J(\mu, z)$ by:

$$J(\mu, z) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_i^k \|x_i - \mu_k\|^2.$$

The K-means algorithm

The aim of the algorithm is to minimize $J(\mu, z)$. To do so we proceed with an alternating minimization :

- Step 0 : We choose a vector μ
- Step 1 : we minimize J with respect to z : $z_i^k = 1$ if $\| x_i - \mu_k \|^2 = \min_s \| x_i - \mu_s \|^2$, in other words we associate to x_i the nearest center μ_k.
- Step 2 : we minimize J with respect to μ : $\mu_k = \frac{\sum_i z_i^k x_i}{\sum_i z_i^k}$.
- Step 3 : we come back to step 1 until convergence.

Data

- We will use a simple (classical) data set describing features of flowers (available at https://www.dropbox.com/s/9kits2euwawcsj0/iris.data.txt)

Le jeu de données [modifier | modifier le code]

Fisher's *Iris* Data

longueur des sépales (en cm) *(Sepal length)*	largeur des sépales (en cm) *(Sepal width)*	longueur des pétales (en cm) *(Petal length)*	largeur des pétales (en cm) *(Petal width)*	Espèce *(Species)*
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*

source : https://fr.wikipedia.org/wiki/Iris_(jeu_de_données)