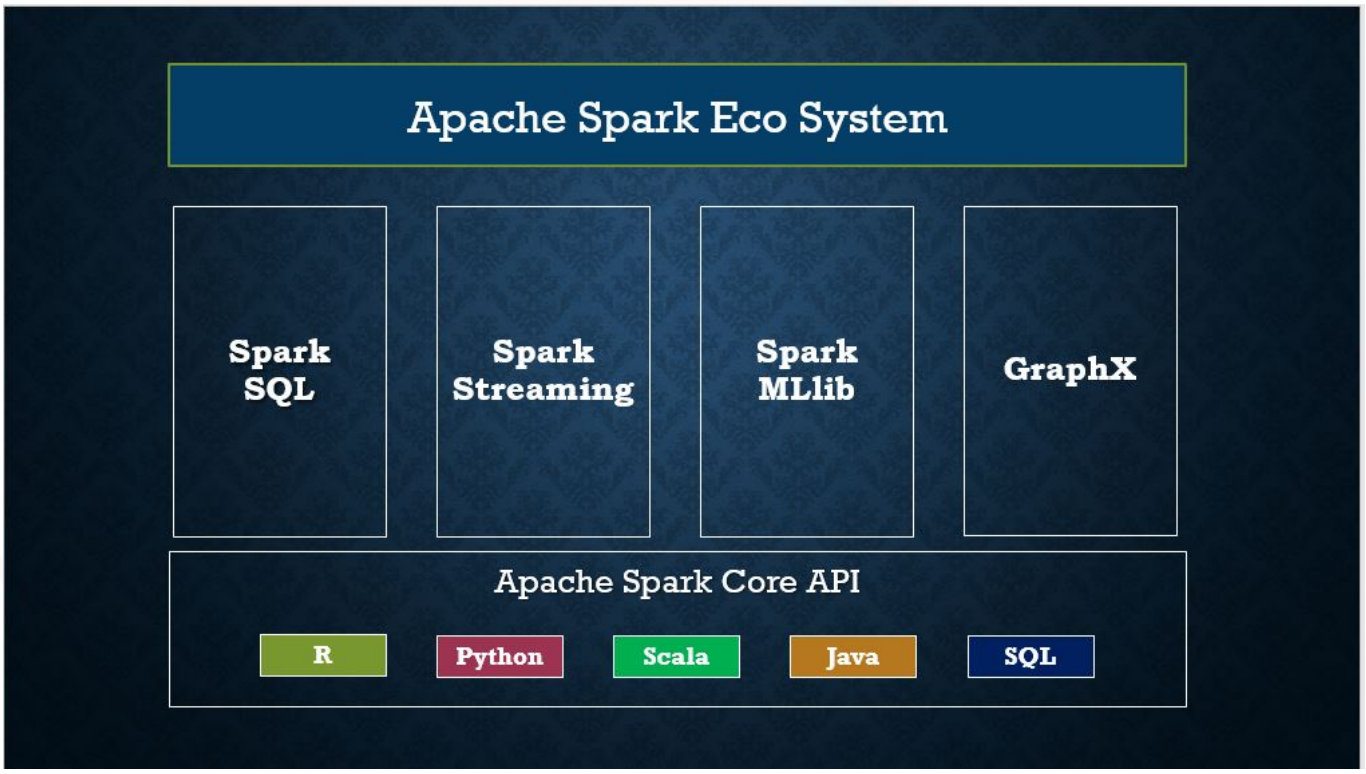
Spark : Les modules Spark

semifir

Note: ref = `https://sparkbyexamples.com`

Les Modules Spark

Spark est subdivisé en differents modules répondant chacun à un besoin spécifique :



Spark Core

> Module qui fournis les fonctionnalités de base de spark.

Fournis l'ensemble des fonctionnalités permettant :

- Definir les inputs / outputs du job spark
- Manipulater et transformer les RDD
- Lancer des jobs spark
- Etablir le contexte

Spark Core

Il contient toutes les API permettant de piloter ces fonctionalités dans les differents langages de programmation supportés par spark :

- R
- Python
- Scala
- Java
- SQL

Spark SQL

> Module qui permet de manipuler des données structurées.

Il permet de lire et de manipuler des données avec des fonctionalités de requetage identique à celle retrouver dans les requetes SQL.

Spark SQL

- Les données seront structurées sous une forme de structure multidiensionnelle rappelant les table SQL appelée **DataFrame**.
- Les manipulations, appelées **Transformation**, seront effectuées de manieres distribuées sur les noeuds du cluster.
- Les fonctions de transformation sont utilisables dans les differents langages de programmation supportés par spark.

Spark Streaming

> Module qui permet le traitement des flux de données.

Il permet de prendre en charge les données issues de source en temps réél (kafka, Flume, Kinesis,sockets reseaux, etc..) et de les traiter en temps réél.

Spark Streaming

Couplé avec les fonctionnalités de Spark SQL, il permet de traiter les données en temps réél et de les manipulées avec les memes fonctionalités que celles de Spark SQL.

Spark Streaming

- Les flux de données définis sous une forme de séquences de RDD appelé **DStreams** (Discretized Streams).
- Les données sont ensuite injectées et traitées par petits lots appelés **micro-batches**.
- Leurs fréquences sont définies par l'utilisateur par le biais d'un paramètre appelé **batch interval**.

Spark MLlib

> Module qui permet de fournir des fonctionalités de machine learning distribuées.

Libairie adapté pour le traitement distribuées de données volumineuses necessaire au pipeline d'apprentaissage du machine learning.

Spark MLlib

Il inclut des fonctionalités de :

- Traitement de données
- Classification
- Clustering
- Régression
- Normlaisation

Spark GraphX

> Module qui permet de manipuler des données sous forme de graphes.

Il contient des fonctionalités permettant de gerer manipuler des données graphes en les distribuants.

Spark GraphX

Il contient également des fonctions de visualisations et de stockages distribuées de données graphes.

Son role est de fournir une infrastructure de gestion des données graphes de grandes tailles, souvent couplées à des algorithmes de machine learning pour les analysées.

Installation de Spark