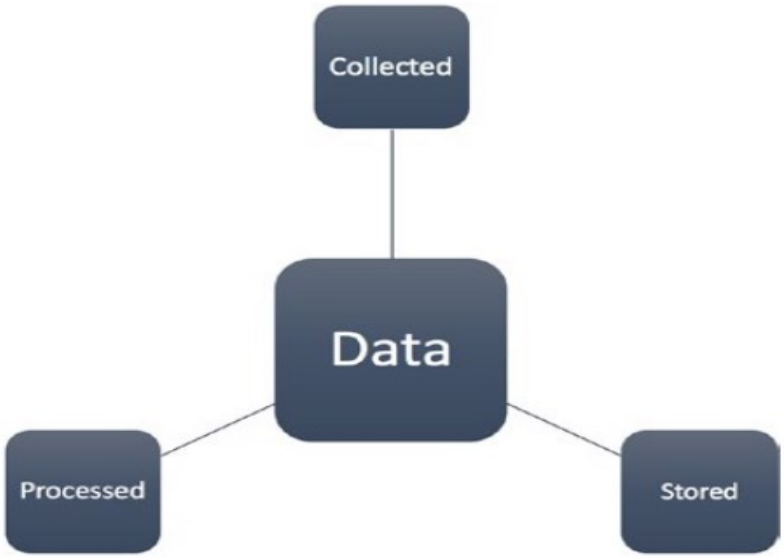
POINTS IMPORTANTS CHAPITRE 1 : INTRODUCTION TO SPARK

- Le projet Spark a été lancé pour résoudre les problèmes potentiels dans le cadre Hadoop MapReduce.
- Hadoop MapReduce : framework révolutionnaire pour gérer le traitement de données volumineuses mais une vitesse limite.
- Spark est donc capable d'effectuer des calculs en mémoire plus rapide que Hadoop MapReduce
- Les trois angles sous lequel les donnees peuvent etre visualisee sont : la manière dont elles sont collectées, stockées et traitées



- Les principaux composants de base de spark sont :

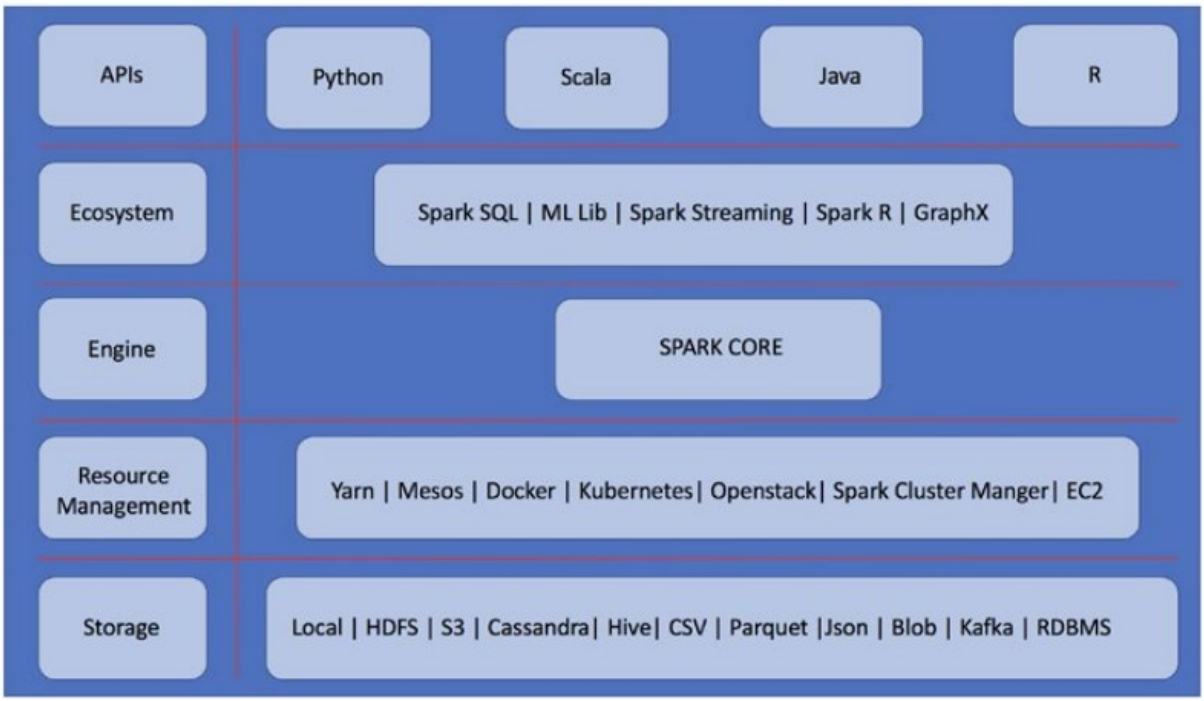
  • Stockage : Spark vous permet d'utiliser des bases de données relationnelles traditionnelles ainsi que NoSQL, telles que Cassandra et MongoDB.

  • La gestion des ressources : Les deux gestionnaires de ressources les plus utilisés sont YARN et Mesos. Le gestionnaire de ressources comporte deux composants principaux en interne :

    1. Cluster manager :de gérer les nœuds de travail et de leur attribuer des tâches, en fonction de la disponibilité et de la capacité du nœud de travail
    2. Worker

• Moteur et Écosystème : La base de l'architecture Spark est son noyau, qui est construit au-dessus des RDD (Resilient Distributed Datasets) et offre plusieurs API pour la construction d'autres bibliothèques et écosystèmes par les contributeurs Spark. Les bibliotheques defaut de spark sont : Spark SQL, Mllib, Structured Streaming, Graph X.

• API : Spark est disponible en quatre langues tels que scala,python,java et R



- Il existe plusieurs façons d'utiliser Spark :
  • Configuration locale
  • Dockers
  • Environnement infonuagique (GCP, AWS, Azure)
  • Briques de données