POINTS IMPORTANTS DU CHAPITRE 14 : HELLO APACHE SPARK

- Apache Spark est un moteur de calcul distribué. Cela signifie que Spark peut exécuter des calculs sur plusieurs machines.
- Spark est fortement associé à Hadoop. Hadoop, comme mentionné dans le premier chapitre, est un consortium de services qui se répartissent en différentes catégories :

  • Calcul (Spark, MapReduce, Hive, Impala, Storm,
  Flink, Samza, Drill et Presto)
  • Stockage (HDFS, Kudu, HBase, MongoDB, Cassandra et
  Gremlin)
  • Sécurité (Sentry, Knox et Ranger)
  • Gestion des métadonnées (Hive Metastore, HCatalog,
  Atlas et Cloudera Navigator)
  • Files d'attente de messages (Kafka, EventHub et Kinesis)
  • Intégration (NiFi, StreamSets et Flume)
  • Gestionnaire de cluster (YARN et Mesos)

- Lien avec spark et Yarn

  YARN, comme vous pouvez le voir dans la liste précédente, est un gestionnaire de cluster. Les gestionnaires de cluster jouent un rôle crucial dans l'anatomie globale du cluster.

  Comme beaucoup d'autres composants, YARN a deux composants :

  • Gestionnaire de ressources
  • Gestionnaire de nœuds

- Spark se compose de deux types de processus :

  • Processus pilotes

  • Processus d'exécution

- le plus rapide et le plus gratuit pour commencer à utiliser Apache

  Spark est via :

  • VM Cloudera QuickStart

  • Édition communautaire de Databricks

NB : ce chapitre parle plus de spark et plus encore de l'installation de son environnent. Nous l'avons déjà installé au niveau des premiers chapitres et mise en évidence au niveau de exercices de ce chapitre-là.