Avancement de travail

Présenté par:
Takwa Ben Smida

Proposé par:
Pr. Engelbert Mephu Nguifo
Dr. Sabeur Aridhi

Plan

1. Motivations
2. La plateforme Spark
4. Graphx
5. objectifs

Motivations

- MapReduce:
  - Plateforme pour le traitement massive des données
  - N'en fournit pas pour accéder à la mémoire partagée.
  - S'appuie sur le partage de données et il ne fournit pas l'abstraction pour une réutilisation des données.

→ Inefficace pour une importante classe d'applications:

- Apprentissage automatique (Clustering);
- Algorithmes de graphe (PageRank);
- Algorithmes d'extraction de données interactives …

Plan

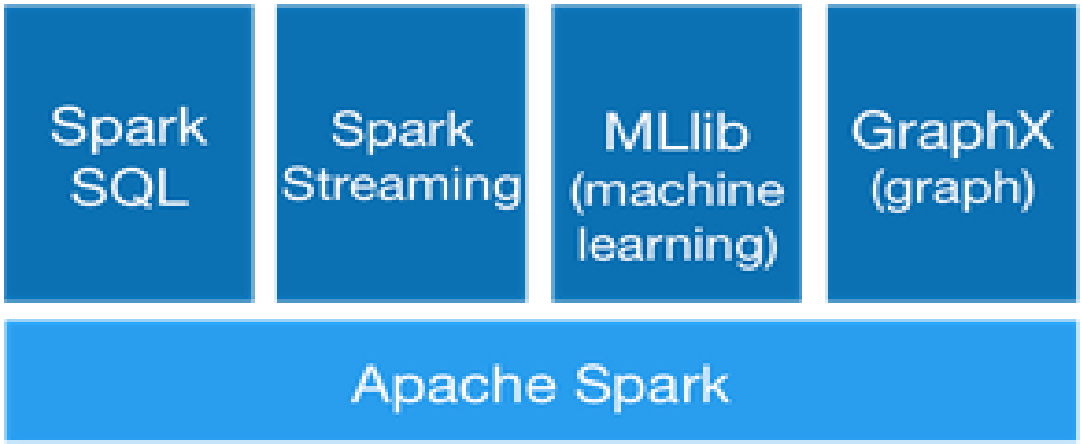
1. Motivations
2. La plateforme Spark
4. Graphx
5. objectifs

Spark

- Un modèle de programmation adapté au traitement de données à grande échelle;
- Open source , développé par la Fondation Apache;
- Une alternative *in-memory* plus rapide que le traditionnel MapReduce de Hadoop;
  - *fonctionne jusqu'à 10x sur disque et jusqu'à 100x en mémoire.*
- Divers langages de programmation : ***Java, Scala ou Python***.

Spark

- Utilisé pour une grande variété d'usages:
  - capable de se connecter à des bases SQL: ***Spark SQL***
  - une analyse permanente des données en temps réel: ***Spark Streaming***
  - les calculs en profondeur impliquant l'apprentissage automatique: ***Mllib***
  - Traitement des graphes: ***Graphx***



RDD: Resilient Distributed Datasets

- une collection partitionnée d'enregistrements en lecture seule;
- Réutilise efficacement les données dans une large gamme d'applications;

RDD: Resilient Distributed Datasets

- Tolérant à la panne;
  - Si une partition de RDD est perdue:
    - le RDD dispose d'informations sur la manière dont il a été produit pour recalculer la partition manquante.
    - les données perdues peuvent être récupérées rapidement sans avoir à recourir aux mécanismes de réplication.

Plan

1. Motivations
2. La plateforme Spark
4. Graphx
5. objectifs

Graphx

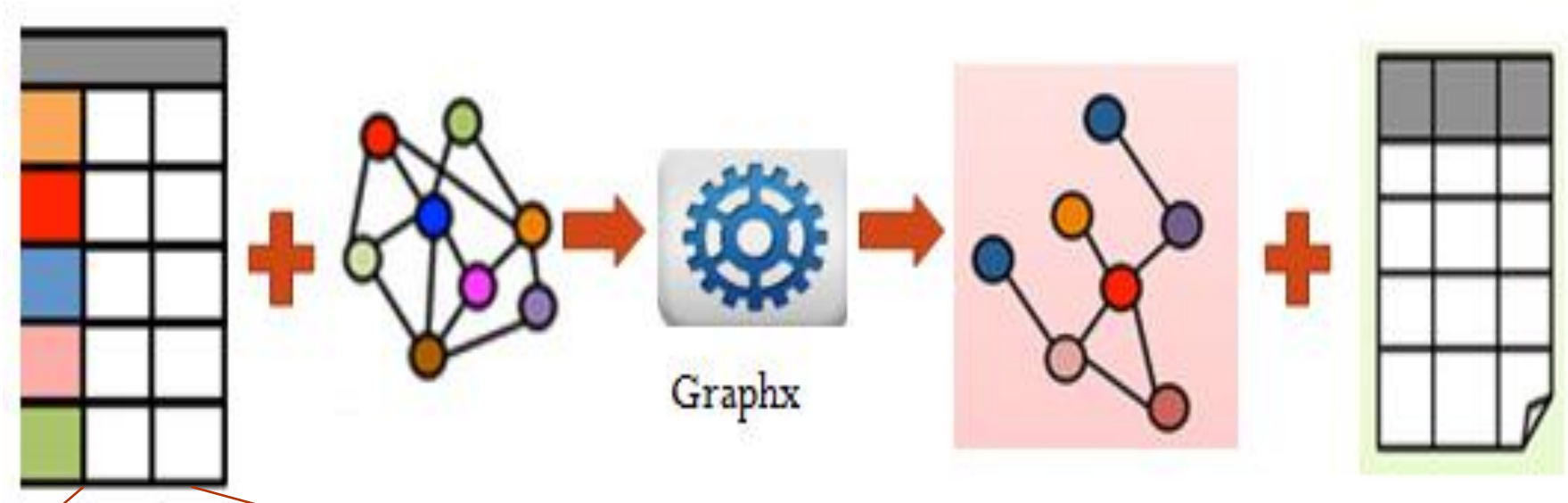
Définition

- une bibliothèque de graphes qui tourne au-dessus de Spark Apache;
- Se base sur RDD de Spark.
- Les utilisateurs peuvent construire, transformer, restreindre des graphes.
  - Un graphe est défini par:

-G(V, E ,P) , avec P= (P_V , P_E).

Graphx

Définition



Id	Property (V)
3	(rxin, student)
7	(jgonzal, postdoc)
5	(franklin, professor)
2	(istoica, professor)

SrcId	DstId	Property (E)
3	7	Collaborator
5	3	Advisor
2	5	Colleague
5	7	PI

Graphx

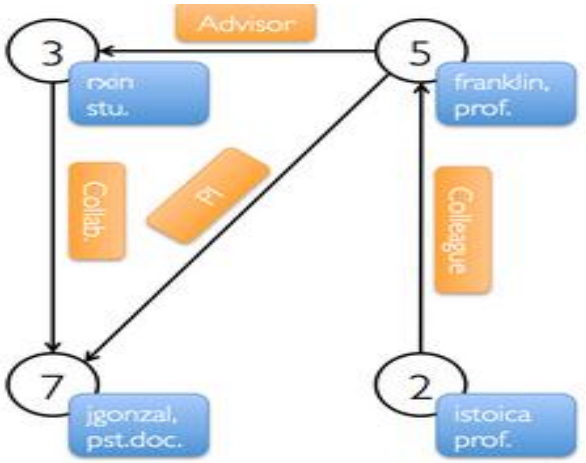
Opérateurs de Graphx

- Map
- Filter
- ReduceByKey

- Subgraph

Id	Property (V)
3	(rxin, student)
7	(jgonzal, postdoc)
5	(franklin, professor)
2	(istoica, professor)

SrcId	DstId	Property (E)
3	7	Collaborator
5	3	Advisor
2	5	Colleague
5	7	PI



Plan

1. Motivations
2. La plateforme Spark
4. Graphx
5. objectifs

Objectifs

- Une évaluation de performance : SPARK et MapReduce:
  - Variation de la taille de données;
  - Variation du nombre de nœuds;

- Proposer un algorithme d'extraction des sous graphes fréquents à partir d'un seul graphe en utilisant Apache SPARK;

Merci pour **votre attention**