**Rapport : Projet PySpark – Recherche des Amis Communs**

**1. Introduction**

Ce projet a pour but d’appliquer les concepts de **programmation distribuée avec Apache Spark**, en résolvant un problème courant des réseaux sociaux : **trouver les amis communs entre deux utilisateurs**. L’objectif principal est d’illustrer l’utilisation de **PySpark** (l’API Python de Spark) pour effectuer un traitement efficace sur un graphe social.

**2. Objectifs du projet**

* Apprendre à manipuler **PySpark (API Python de Spark)**.
* Modéliser un **graphe social** à partir d’un fichier texte.
* Identifier **les amis communs** entre deux utilisateurs donnés.
* Utiliser les concepts de **MapReduce** et de **traitement distribué**.
* Comprendre le fonctionnement des **RDDs (Resilient Distributed Datasets)**.

**3. Présentation d'Apache Spark**

### ****Qu'est-ce que Spark ?****

Apache Spark est un moteur de traitement distribué open-source conçu pour des analyses de données à grande échelle. Il offre des performances élevées grâce à son traitement en mémoire (in-memory processing).

### ****Concepts Clés****

* **RDD (Resilient Distributed Dataset)** : Structure de données distribuée et tolérante aux pannes.
* **DataFrame** : Abstraction de données tabulaires optimisée pour les opérations SQL.
* **MapReduce** : Modèle de programmation pour le traitement parallèle de données.
  + **Map** : Transforme les données en paires clé-valeur.
  + **Reduce** : Agrège les résultats par clé.

### ****Avantages de Spark****

* **Rapidité** : Jusqu'à 100x plus rapide que Hadoop MapReduce.
* **Polyvalence** : Supporte Python, Java, Scala, et R.
* **Intégration** : Compatible avec Hadoop, Kafka, et d'autres outils Big Data.

**4. Présentation des données**

**Format du fichier d'entrée :**

Chaque ligne du fichier texte suit la structure suivante :

<user\_id> <Nom> <friend\_id1>,<friend\_id2>,...

Exemple :

1 Sidi 2,3,4

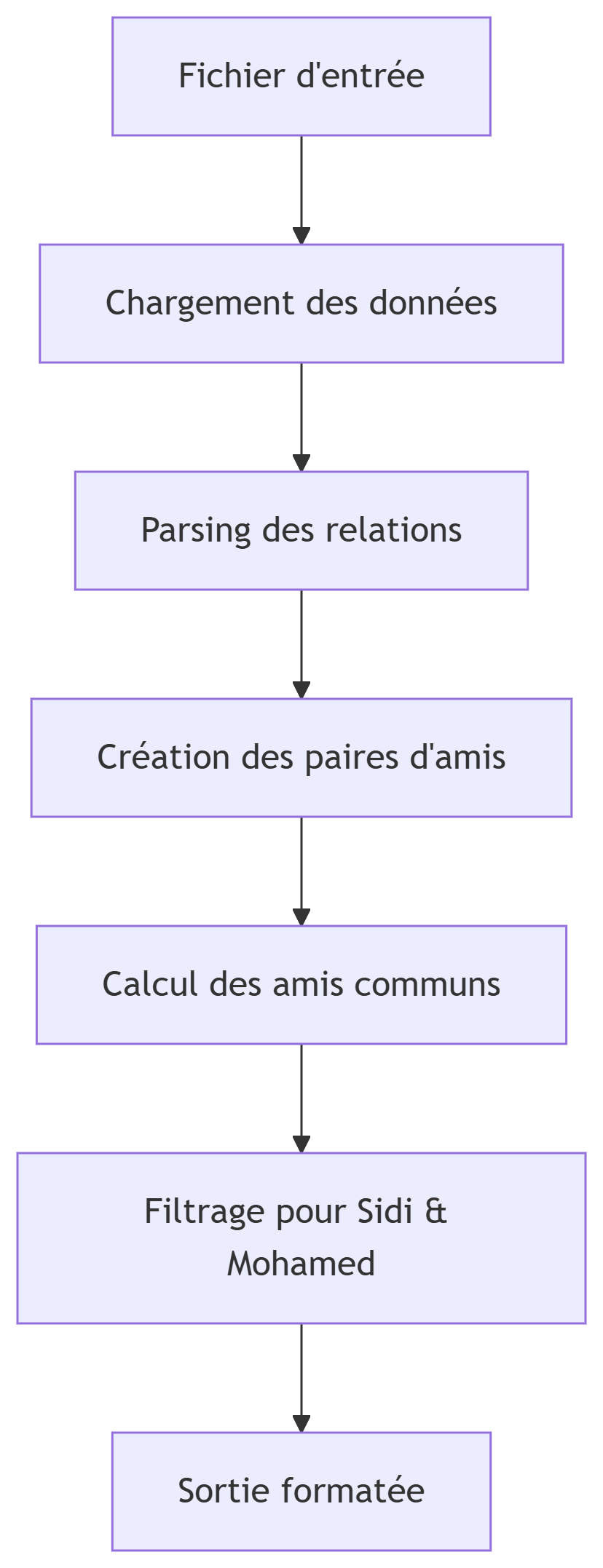
2 Mohamed 1,3,5

Ce qui signifie :

* L'utilisateur **1 (Sidi)** est ami avec les utilisateurs 2, 3 et 4.
* Les relations sont **symétriques** : si A est ami avec B, alors B est ami avec A.

## ****5. Méthodologie et Étapes du Projet****

### ****Schéma Global****



**6. Architecture de traitement avec Spark**

Spark utilise une architecture distribuée où les **RDDs** sont répartis sur plusieurs nœuds. Le processus suit le paradigme **MapReduce** :

* **Map** : transforme les données ligne par ligne.
* **Reduce** : agrège les résultats intermédiaires.

Dans ce projet :

* **Map** sert à extraire les paires d’utilisateurs.
* **Reduce** combine les listes d’amis pour trouver l’intersection.

## 7****. Implémentation Technique****

### ****Extraits de Code****

#### ****Chargement des données****

df = spark.read.text("social\_graph.txt") \

.select(

split(col("value"), " ")[0].cast("integer").alias("user\_id"),

split(col("value"), " ")[1].alias("name"),

split(split(col("value"), " ")[2], ",").cast("array<integer>").alias("friends")

)

#### ****Création des paires normalisées****

from pyspark.sql.functions import explode

pairs\_df = df.select(

"user\_id",

"name",

explode("friends").alias("friend\_id")

).withColumn(

"pair",

array(sort\_array(array(col("user\_id"), col("friend\_id"))))

#### ****Calcul des amis communs****

common\_friends\_df = pairs\_df.groupBy("pair") \

.agg(collect\_set("friend\_id").alias("common\_friends")) \

.filter(size(col("common\_friends")) > 0)

**8. Résultats et Analyse**

**Données de Test**

| **user\_id** | **nom** | |  | **amis** |
| --- | --- | --- | --- | --- |
| 1 | Sidi | |  | 2, 3, 4 |
| 2 | Mohamed | |  | 1, 3, 5 |
| 3 | Ali |  | | 1, 2, 4 |

**Résultat Obtenu**

[1<Sidi>2<Mohamed>3]

Cela signifie que Sidi (1) et Mohamed (2) ont **Ali (3)** comme ami commun

**9. Conclusion**

Ce projet a permis d’explorer :

* Le **traitement parallèle avec Spark**
* L’implémentation de **graphes sociaux**
* L’utilisation efficace de **PySpark** pour modéliser un problème réel

La méthode utilisée est généralisable à d'autres applications de graphes, comme la détection de communautés ou la recommandation de nouveaux amis.