République Algérienne Démocratique & Populaire
Ministère de l'Enseignement Supérieur & de la Recherche Scientifique
Ecole Supérieure en Informatique 08 mai 1945 de Sidi Bel Abbes
Année universitaire : 2023-2024 **Semestre 1**
3 SC : SIW & ISI **Module : Big Data**

TP 5 : Spark GraphX et GraphFrames

Nous allons utiliser un Dataset composé de deux fichiers :

- Le premier fichier **"station_data.csv"** représente des stations.
- Le deuxième fichier **"trip_data.csv"** représente des voyages (trips) à travers des vélos.

1) Créer un DataFrame pour chaque fichier.

Construction du graphe

La première étape consiste à construire le graphe. Pour ce faire, nous devons définir les sommets et les arcs, qui sont des DataFrames avec des colonnes spécifiquement nommées.

Dans notre cas, nous créons un graphe orienté. Ce graphe pointera de la source à la destination.

Dans le contexte de ces données de trajet à vélo, cela indiquera le lieu de départ d'un trajet vers le lieu de fin d'un trajet. Pour définir le graphe, nous utilisons les conventions de nommage des colonnes présentées dans la bibliothèque GraphFrames. Dans la table des sommets, nous définissons notre identifiant comme **id** (dans notre cas, il s'agit de la colonne **name**), et dans la table des arcs, nous renommons l'ID de sommet source (la colonne **Start Station**) de chaque arc comme **src** et l'ID de destination (la colonne **End Station**) comme **dst**.

2) Dans la table des sommets (le DataFrame qui représente le fichier station_data.csv), renommer la colonne **name** en **id**.
3) Dans la table des arcs (le DataFrame qui représente le fichier trip_data.csv), renommer la colonne **Start Station** en **src** et la colonne **End Station** en **dst**.
4) Créer un GraphFrame, qui représente notre graphe.

Interroger le graphe

Répondre aux requêtes suivantes :

5) Retourner le nombre de voyages effectués entre chaque source et destination triés par ordre décroissant (selon le nombre de voyages).
6) Retourner le nombre de voyages qui se commencent ou se terminent à partir de la station 'Townsend at 7th' triés par ordre décroissant (selon le nombre de voyages).
7) Retourner les sommets qui n'ont jamais été une destination d'un voyage qui commence à partir de ' Spear at Folsom'.
8) Retourner la station qui a le nombre maximum de voyages entrants.
9) Retourner le voyage qui a la plus grande durée.

Les sous-graphes

10) Créer un sous-graphe qui ne contient que les voyages qui se commencent ou se terminent à ‘Townsend at 7th’.

Recherche de motifs

11) Retourner tous les chemins qui forment un motif en "triangle" entre trois stations.
12) Retourner tous les chemins qui passent par trois sommets et qui commencent à partir de 'Townsend at 7th'.