Interroger et analyser des très grands graphes dans un environnement distribué.

Encadrant : Hubert Naacke

Contexte : Projet EPIQUE, workflow de textmining pour extraire des domaines scientifiques (topics) dans des grands corpus de documents. http://www-bd.lip6.fr/wiki/site/recherche/projets/epique/start

Ce PLDAC s'intéresse au problème d'analyser des très grands graphes pondérés pour extraire des motifs permettant de mieux comprendre l'information décrite par ces graphes. Il s'agit d'extraire des sous graphes qui satisfont certains critères sur le poids et arcs, le degré sortant et la longueur des chemins, etc. Ce type de traitement pose un problème de performance et devient trop long pour des très grands graphes. Pour gagner en efficacité, on propose d'utiliser une plateforme de calcul distribué telle que Spark. La capacité de calcul de cette plateforme est proportionnelle au nombre de machines utilisées ce qui offre un fort potentiel. Toutefois le gain en performance avec une telle plateforme s'avère difficile à obtenir et cela nécessite de repenser les algorithmes de manipulation de graphe pour les adapter aux spécificités d'un environnement distribué. Dans ce projet PLDAC, vos tâches seront les suivantes :

- Etudier la solution de référence proposée dans le projet EPIQUE, comprendre pour quelle raison cette solution n'exploite pas de manière optimale la puissance de calcul d'une plateforme distribuée.
- Proposer une solution pour paralléliser davantage les extractions de sous graphes.
- Implémenter votre solution sur le cluster Spark du laboratoire LIP6 et mesurer les performances obtenues

Prérequis : UE de M1 MLBDA et SAM, langages python et scala.

Biblio :

[6] Li KE, Bernd Amann, Hubert Naacke, Exploring the Evolution of Science with Pivot Topic Graphs: 3rd International Workshop on Big Data Visual Exploration and Analytics EDBT/ICDT 2020 **Lire le pdf** https://bigvis.imsi.athenarc.gr/bigvis2020/papers/BigVis2020_paper_3.pdf