PROPOSITION ACTION POUR LA FEDERATION LI-LIFO

1- Titre :

Extraction de motifs séquentiels à grande échelle dans un environnement Hadoop / MapReduce

2- Description succincte (thématiques et objectifs)

Depuis 25 ans, l'extraction de motifs intéressants joue un rôle très important dans le domaine de la fouille de données. Cependant celle-ci reste confrontée à des temps de traitement qui peuvent être importants sur de grandes masses de données. Dans ce contexte, différents algorithmes d'extraction de motifs fréquents on été implantés dans un environnement Hadoop / MapReduce ou Spark. Néanmoins, la plupart des travaux existants portent sur l'extraction de motifs ensemblistes (itemsets) et très peu de travaux existent sur l'extraction de motifs plus complexes (séquences, graphes, etc.).

Dans ce cadre, l'objectif de cette action est de proposer de nouveaux algorithmes et outils pour l'extraction de motifs séquentiels sur de grandes masses de données en s'appuyant sur une complémentarité forte entre différents équipes de la fédération ICVL, à savoir :

- une expertise de l'équipe BdTln du LI dans le domaine de l'extraction de motifs locaux et de leur optimisation (mais dans un environnement non parallèle) ;

- une expertise des équipes PaMDA du LIFO et OC du LI dans les domaines du parallélisme et du calcul distribué, en particulier des plateformes de développement de type Hadoop / MapReduce.

3- Participants (personnes impliquées avec précision sur appartenance)

Pour l'équipe BDTLN (LI):

- Arnaud GIACOMETTI (PR),
- Dominique LI (MCF),
- Arnaud SOULET(MCF),

Pour l'équipe OC (LI):

- Patrick MARTINEAU (PR)

Pour l'équipe PaMDA (LIFO):

- Mostafa BAMHA (MCF),

Pour l'équipe LMV (LIFO):

- Ali ED-DBALI (MCF)

4- Historique des collaborations (s'il existe des collaborations passées et des résultats déjà obtenus)

- Collaboration existante au LI entre les équipes BDTLN et OC sur cette thématique
- Dépôt de projets: H2020 LASCAR (LI) et APR-IA GIRAFON (LI+LIFO)

5- Proposition de travail (description plus détaillée des collaborations envisagées et des résultats attendus)

Des collaborations ont déjà commencé à travers les projets récemment soumis (H2020 LASCAR (LI) et APR-IA GIRAFON (LI+LIFO)).
Dans un premier temps, les travaux vont se concentrer sur la conception et la preuve formelle de nouveaux algorithmes pour l'extraction de motifs séquentiels sur de grandes masses de données. Ensuite, des benchmarks devront être préparés pour évaluer et comparer les performances des différents algorithmes proposés, mais aussi comparer les forces et faiblesses des différentes variantes existantes du modèle Hadoop / MapReduce. Enfin, les compétences de l'équipe OC devront permettre d'identifier comment pourraient être améliorés les mécanismes de répartition de charge du modèle Hadoop / MapReduce et de ses variantes, mais aussi si de nouvelles variantes pourraient être développées face aux contraintes particulières de l'extraction de motifs (la taille potentiellement exponentielle des espaces de recherche à parcourir).

6- Prospectives (Interactions possibles avec autre action? Auriez vous des idées sur les thématiques d'un axe qui pourraient englober cette action?

L'action se concentrera en priorité sur l'extraction de motifs séquentiels. Mais elle pourrait ensuite être étendue au traitement et à l'exploration d'autres types de données ou à d'autres domaines d'application, telle que l'analyse et la fouille de grands graphes. Enfin, à moyen terme, les exigences de généricité et d'efficacité imposent le recours à des langages de requêtes pour la préparation des données à analyser. Dans ce cadre, des interactions avec l'action intitulée "Interrogation Cohérence Masse Données" pourraient être développées afin d'identifier les primitives nécessaires dans le développement d'un langage de requêtes pour l'interrogation et l'analyse des données distribuées.

7- Bibliographie:

- **Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth.** Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., & Hsu, M. C. In the 29th International Conference on Data Engineering (ICDE), pp.

0215-0215. April 2001, IEEE Computer Society.

- **Polynomial-Delay and Polynomial-Space Algorithms for Mining Closed Sequences, Graphs, and Pictures in Accessible Set Systems.** Hiroki Arimura, Takeaki Uno, SDM 2009: 1088-1099.
- **Large-scale frequent subgraph mining in MapReduce.** W. Lin, X. Xiao, and G. Ghinita. In ICDE, pages 844–855, 2014.
- **Mind the gap: Large-scale frequent sequence mining.** Miliaraki, K. Berberich, R. Gemulla, and S. Zoupanos. In SIGMOD, pages 797–808, 2013.
- **Frequent itemset mining for big data.** S. Moens, E. Aksehirli, and B. Goethals. In BigData Conference, pages 111–118, 2013.