

Une Plateforme ETL parallèle et distribuée pour l'intégration de données massives

Mahfoud Bala*, Oussama Mokeddem*,
Omar Boussaid**, Zaia Alimazighi***

*LRDSI, Université Saad Dahleb, Blida 1, Algérie
{mahfoud.bala, mokeddem.dev}@gmail.com,

**ERIC, Université Lumière, Lyon 2, France
omar.boussaid@univ-lyon2.fr,

***LSI, USTHB, Alger, Algérie
zalimazighi@usthb.dz

Résumé. Nous nous intéressons, dans ce papier, à l'impact des données massives dans un environnement décisionnel et plus particulièrement sur la phase d'intégration des données. Dans ce contexte, nous avons développé une plateforme, baptisée *P-ETL (Parallel-ETL)*, destinée à l'entreposage de données massives selon le paradigme *MapReduce*. *P-ETL* permet le paramétrage de processus *ETL* (workflow) et un paramétrage avancé relatif à l'environnement parallèle et distribué. Ce papier décrit la plateforme *P-ETL* en vue d'une démonstration. Face à des jeux de données allant de $244 * 10^6$ à $7,317 * 10^9$ tuples, les expérimentations menées ont montré l'amélioration significative des performances de *P-ETL* lorsque la taille du cluster et le nombre des tâches parallèles augmentent.

1 Introduction

Les données massives, appelées communément "*big data*", impactent directement le processus *ETL (Extracting-Transforming-Loading)* vu que celui-ci est le premier composant du système décisionnel confronté à ces données. Peu de travaux ont traité sur la problématique des données massives dans le processus *ETL*. Liu et al. (2011) ont proposé une approche parallèle/distribuée appelée *ETLMR* consistant à améliorer les performances de la phase de transformation (T) et de chargement (L) de l'*ETL* et ce en adoptant, pour chacune des deux phases, des stratégies de distribution appropriées. Les expérimentations de Misra et al. (2013) ont montré que le paradigme *MapReduce* est prometteur et que les solutions *ETL* basées sur des frameworks open source tel que *Apache Hadoop* sont plus performantes et moins coûteuses par rapport aux solutions *ETL* commercialisées. Contrairement aux travaux de Liu et al. (2011), ceux de Misra et al. (2013) considèrent la phase d'extraction (E) de l'*ETL* très coûteuse ; celle-ci a été traitée dans un environnement parallèle/distribué selon le paradigme *MapReduce*. (Liu et al., 2012) est une démonstration du prototype *ETLMR*. Dans (Liu et al., 2014), les auteurs proposent une plateforme *CloudETL* basée sur *Apache Hadoop* et *Apache Hive* où les performances ont été nettement améliorées par rapport à celles d'*ETLMR* (Liu et al.,