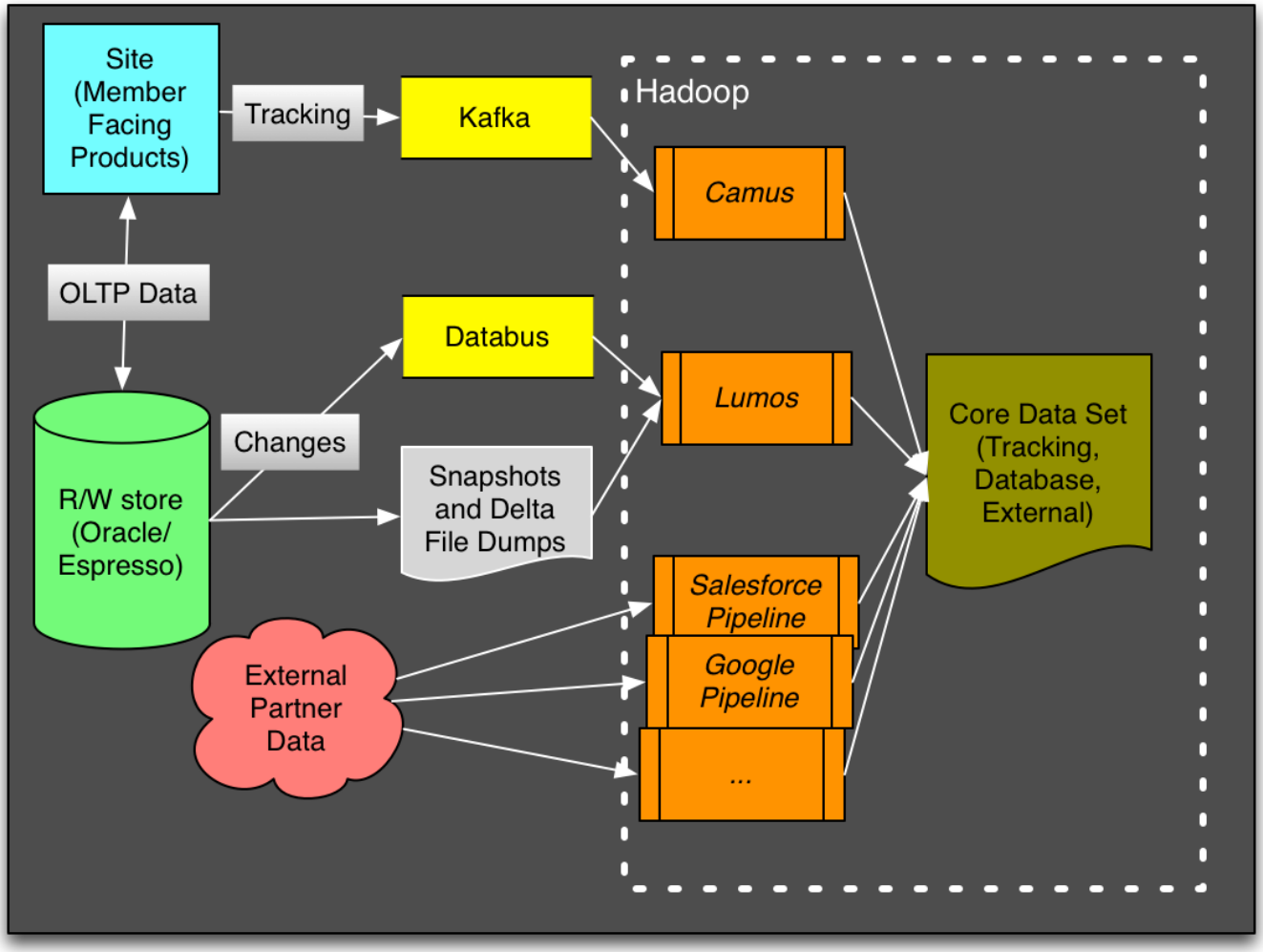
Avec Gobblin, LinkedIn digère mieux le Big Data

Les réseaux sociaux dévoilent petit à petit leurs recettes pour faire fonctionner leurs infrastructures hors norme (lire Facebook repense son architecture réseau), mais également pour s'adapter à la quantité de données produites et à analyser. C'est dans ce cadre que LinkedIn a donné des éléments complémentaires sur **Gobblin**, son **framework d'ingestion de Big Data**. Dans un article intitulé, « *Gobblin'Big Data with ease* », Lin Qiao, ingénieure, a expliqué comment LinkedIn a simplifié l'ingestion de grandes quantités de données à destination de datawarehouses basés sur Hadoop.

Tout démarre par la récolte des données pour créer un « *jeu de données interne* », souligne l'ingénieure. Ce dataset comprend une multitude d'informations : les profils des membres, les actions des utilisateurs comme un commentaire ou un article, etc. Ces renseignements sont issus des bases de données de LinkedIn (Oracle, MySQL et Espresso). Il faut ajouter à cela le informations issues du système des logs d'évènements (géré via un outil maison baptisé Kafka), celle du module Databus capable de gérer les modifications des données de manière incrémentale et encore celles provenant de données d'applications tierces comme Salesforce, Google ou Twitter. Au total, Linkedin a créé des tuyaux de données capables de **gérer des centaines de To par jour** qui sont ensuite reversées dans un cluster Hadoop (ch schéma ci-dessous).

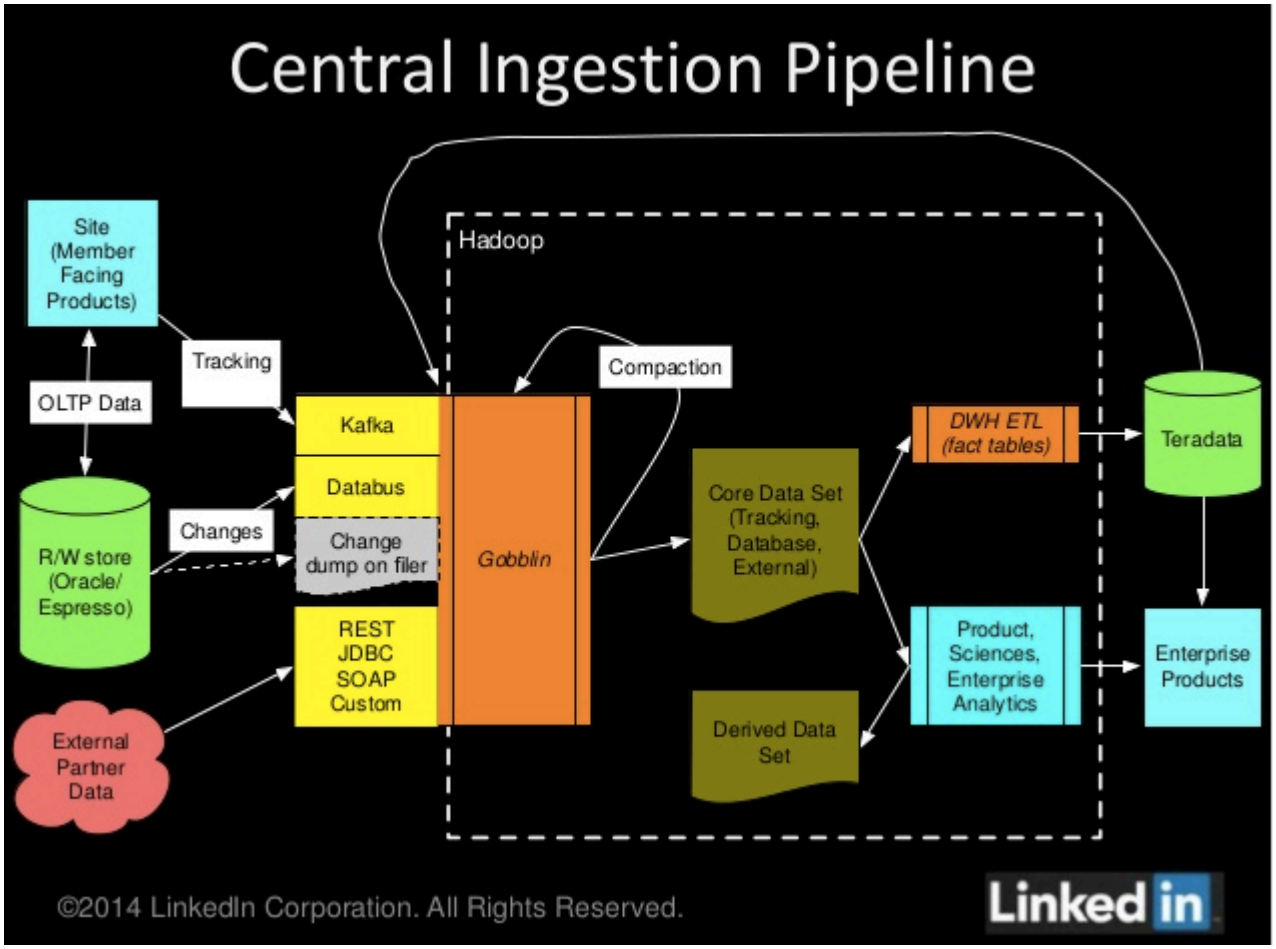


Unifier les pipelines et créer une passerelle avec Hadoop

Cependant la diversité des types de données présente une contrainte pour LinkedIn, car il y a plusieurs variables à prendre en compte. Il faut choisir les sources que l'on va capter (évènements, fichiers de logs, etc.), les formats de flux (par lots ou en continu) et aussi les protocoles utilisés (REST, Kafka, Camus, Lumos – trois projets propres à LinkedIn – ou des API spécifiques). Au final avec ce système, Lin Qiao constate l'existence « *de 15 types de pipelines d'ingestion de données avec la problématique d'assurer un même niveau de qualité, d'interopérabilité et de fonctionnalité sur les données* ».

C'est **pour simplifier ce processus que Gobblin a été créé**. Il a pour objectif d'unifier les tuyaux en un seul et ainsi alimenter directement le cluster Hadoop. Gobblin intègre « *des adaptateurs prêts à l'emploi pour l'ensemble des sources de données comme Salesforce, MySQL, Google, Kafka et Databus, etc* » (cf schéma ci-dessous). Le framework va s'appuyer sur Yarn Manager pour faciliter « *l'ingestion par lots ou de manière continue* ». Aujourd'hui, Gobblin est utilisé pour gérer seulement des dizaines de To, annonce l'ingénieure. « *Nous sommes actuellement en train de migrer nos datasets externes et internes dans Gobblin, pour tester les API internes, la plateforme et le support. Au début 2015, nous prévoyons de migrer certains de nos pipelines dans Gobblin* ». Par ailleurs, LinkedIn prévoit de **mettre le framework en Open Source** pour que les développeurs puissent proposer des jeux de données

supplémentaires et d'autres connecteurs.



A lire aussi :
LinkedIn, toujours dans le rouge, malgré des revenus en hausse
LinkedIn géolocalise les compétences IT

Crédit Photo : phipatbig-Shutterstock