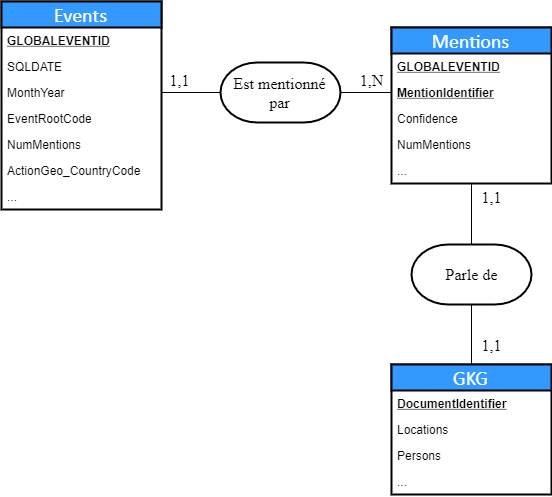
Réalisation : Mansour Hmidan, Sofien Bentaher, Ahmed Ben Amor, Abdelkrim Mahlal , Haythem Chettaoui, Djafar Hamiche, Kamel Boucherba, Aymen Cheikh , Lamine Ouachek

GDELT : l’architecture

GDELT est la base de données ouverte la plus grande, la plus complète et la plus haute résolution de la société humaine jamais créée. Création d'une plate-forme qui surveille les médias d'information du monde entier dans presque tous les coins de chaque pays dans des formats imprimés, diffusés et Web, dans plus de 100 langues, à chaque instant de chaque jour et qui s'étend du 1er janvier 1979 à nos jours, avec mises à jour, ont nécessité un éventail sans précédent d'innovations techniques et méthodologiques, de partenariats et de tout nouvelles mentalités pour rassembler tout cela et en faire une réalité. La création d'une base de données d'un quart de milliard d'enregistrements géoréférencés couvrant le monde entier sur 30 ans, couplée aux réseaux massifs qui connectent toutes les personnes, organisations, lieux, thèmes et émotions sous-jacents à ces événements, a nécessité non seulement de résoudre des défis sans précédent pour créer la base de données, mais aussi une "réimaginassions" de la façon dont nous interagissons et pensons aux données à l'échelle de la société.

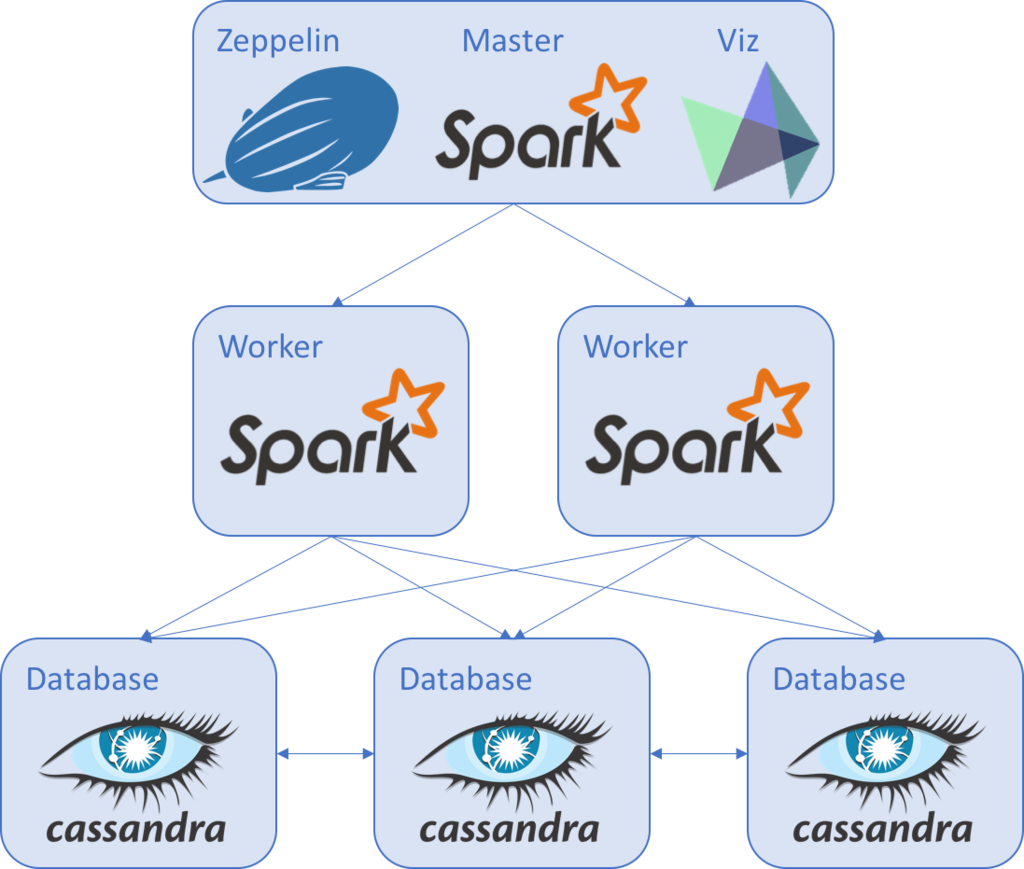
1. **Modèle de données**



1. **Architecture**

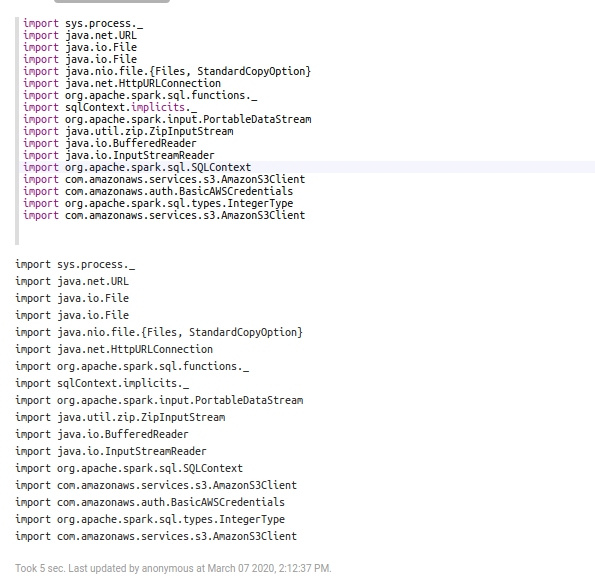
Pour ce projet, on a opté pour l'architecture suivante:

* Utiliser Spark comme ETL avec son langage natif et le traitement et les jointures intermédiaires via spark SQL
* Déposer les fichier GDELT dans S3 sous format parquet pour accélérer le processus de lecture
* Utiliser zepplin notebook



1. **Préparation de donnée**

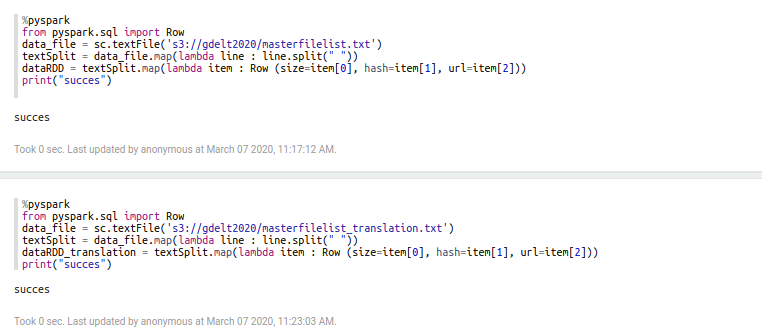
- Importation des packages nécessaires

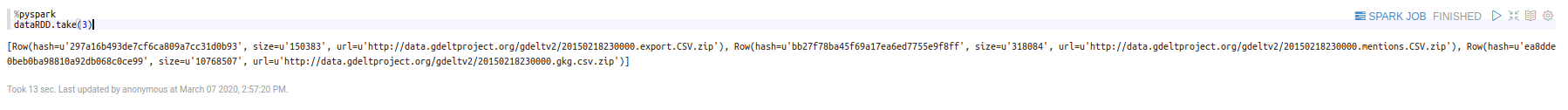


- Telechargement des ficher masterfilelist.txt et masterfile\_translation.txt



- Recuperer le fichier masterfile.txt et le spliter dans un RDD avec 3 colonne qui sont size, hash et url





- Convertion les donnée du RDD à Data Frame

