Année 2020-2021

Sujet de projet Big Data Twitter
Twitter Insight

Twitter-Insight : Nous voulons mettre en place un système de collecte et d'analyse de tweets dans notre entreprise. Pour garantir un accès à l'information pérenne et libre, nous voulons créer notre propre système de stockage et d'analyse. Ce système sera interrogeable via une API qui permettra d'obtenir les réponses quasiment instantanément.

Les données que nous utilisons proviennent du site archive.org et sont téléchargeable gratuitement. Pour faire nos expérimentations, nous disposons déjà de toutes des tweets du mois de mars 2020.

Un tweet est un court message composé de 140 caractères plus 20 caractères pour stocker le nom de l'auteur du tweet. Chaque tweet est daté et identifié avec un numéro unique.

Dans ce projet, nous voulons étudier la mise en place d'un système d'analyse/stockage de ces tweets. Pour cela nous allons collecter au fil du temps tous les tweets pour ensuite utiliser les outils big data pour en extraire de l'information.

Les données que nous recevons sont en JSONNL (new line json) et contiennent au moins les informations suivantes :

Tweet ID, Date, heure, Auteur, Message, plus beaucoup d'autres informations

Chaque ligne stockée nécessite en moyenne 5Ko. Le nombre d'utilisateurs de Twitter actif chaque mois est actuellement de 330 millions. Ils génèrent en moyenne 504 millions de tweets par jours.

Nous voulons implémenter notre solution sur un cluster Hadoop dont le système de fichier HDFS est configuré avec une réplication de 3 et des tailles de blocs de 128Mo.

Question Infrastructure :

a) Les données que nous utilisons pour l'instant ne sont pas complètes car il manque des tweets. Combien de jours complets de collecte de données pouvons-nous stocker sur notre infrastructure de test (votre salle de TP) ?

b) Pour ce nombre de jour de collecte complet, combien de blocs de données seront disponibles sur chaque machine en moyenne ?

c) Afin de planifier nos achats de matériels futurs, calculez le nombre de machines total que nous devrons avoir dans notre cluster pour stocker 5 ans de tweets ?

Dans la suite nous listons une ensemble de réponses auxquelles pourrait répondre votre API. L'objectif du projet est de mettre en place tout le workflow du Master DataSet au serving layer. Tout votre système doit être pensé pour passer à l'échelle et pour pouvoir : (I) se mettre à jours rapidement à chaque nouvelle arrivage de donnée et (II)répondre quasiment instantanément aux requêtes de l'utilisateur. Vous développerez un petit front end d'interrogation de votre API à la google pour permettre la démonstration de l'efficacité de celle-ci.

Vous êtes libre de faire d'autres type de requêtes ou de ne pas toutes les faires. La notation portera sur la qualité des solutions mises en place d'un point de vue de leur scalabilité horizontale. Vous indiquerez pour toutes vos requêtes les temps de calculs nécessaire en fonction du nombre de tweets traité.(théorique et pratique)

Les technologies à utiliser sont :
Master Dataset : HDFS
Batch layer : Spark|MapReduce
Serving layer HBASE
Rest server : Nodejs|Python
Front end : HTML5/JS

Exemple d'analyse des hastags:

a) Permettre de récupérer pour un jours donné la liste des k hashtags les plus utilisés ainsi que leur nombre d'apparition (k entre 1 et 10000).
b) Permettre de récupérer les k hashtags les plus utilisés (k entre 1 et 10000) sur toutes les données.
c) Permettre de récupérer le nombre d'apparition d'un hashtag donné.
d) Récupérer tous les utilisateurs qui ont utilisé un hashtag

Exemple d'analyse des users:

a) Permettre de récupérer pour un utilisateur la liste de ses hashtag sans doublon.
b) Permettre de d'avoir le nombre de tweet d'un utilisateur.
c) Nombre de tweet par pays ou par langue

Exemple d'analyse des influencers :

a) Récupérer tous les triplets de hashtags ainsi que les utilisateurs qui les ont utilisés.
b) Donner k triplets de hashtags les plus utilisés (k entre 1 et 1000)
c) Trouver les influencers c.a.d les personnes avec le plus grand nombre de tweets dans les triplets que l'on a trouvé.
d) Trouver les faux influencer, personnes avec beaucoup de followers dont les tweets ne sont jamais retweeté.
e) Trouvez les sujets (hashtag) qui permettent d'avoir le plus de followers.
f) Retagger les tweet en cherchant les hashtags connus dans le texte des tweet.