

13 octobre 2018
Créé par : Philippe Lawson

Recherche d'Experts

Data Lake

Préliminaires

Les datasets contiennent l'intégralité des données utiles contenues dans les fichiers bruts. En effet, bien que la recherche d'experts n'utilise qu'un sous ensemble des attributs, le choix a été de stocker l'ensemble des attributs. Ainsi ce jeu de données pourra être utilisé à des fins différentes par d'autres applications.

Namespaces Wikipédia

Il y a environ 32 [namespaces](#) dans Wikipédia. Seul le namespace 0 correspond aux articles. Le programme de sélection d'experts ne traite que les données du namespace 0.

Pseudo utilisateurs

Un certain nombre de traitements automatisés opèrent des modifications sur les contenus Wikipédia. Il s'agit de « bot », de scripts de conversion, etc. Le programme de sélection d'experts ne décompte pas ces utilisateurs fictifs.

Adresse IP

Certaines modifications n'ont pas de username, elles ont une adresse IP. Le programme de sélection d'experts ne les inclut pas dans le décompte des révisions.

Requêtes Spark SQL et résultats

Vues temporaires

Trois vues temporaires sont créées en appliquant les filtres décrits dans le chapitre Préliminaires.

revision

Vue contenant l'ensemble des révisions.

Colonnes	Commentaires
page_id	Identifiant de la page
page_title	Titre de la page
contributor	Username du contributeur

pagelink_to_me

Vue des identifiant des pages pointant vers le document de référence.

Colonnes	Commentaires
pl_from	Identifiant de la page

pagelink_from_me

Vue des titres de pages pointés par le document de référence.

Colonnes	Commentaires
pl_title	Titre de la page

Requête

```
SELECT rv.contributor contributeur,  
COUNT(rv.contributor) quantite FROM revision rv  
WHERE rv.page_title = '<titre>' or rv.page_id in (SELECT pl_from FROM pagelink_to_me)  
or rv.page_title in (SELECT pl_title FROM pagelink_from_me)  
group by contributeur order by quantite desc
```

Avec <titre> étant le titre du document de référence.

Les résultats

Cinéma surréaliste

```
Page Id: 2785024, Title: Cinéma surréaliste  
  
Contributeur      |Nb_révisions  
-----  
Rflock            |3255  
Arcane17          |1623  
Vatekor           |1464  
-----
```

Anthropologie marxiste

```
Page Id: 1590508, Title: Anthropologie marxiste  
  
Contributeur      |Nb_révisions  
-----  
Noelbabar         |277  
Horowitz          |123  
Elnon             |113  
-----
```

Dennō Senshi Porigon

Page Id: 6775311, Title: Dennō Senshi Porigon

Contributeur	Nb_révisions
TiboF	1911
Céréales Killer	1149
Nicko	1097