|  |
| --- |
|  |
| Data Lake |
|  |
| Recherche d’experts |

**Dan & Briggs**

13 octobre 2018

Créé par : Philippe Lawson

Data Lake

Recherche d’experts

# Préliminaires

Les datasets contiennent l’intégralités 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](https://en.wikipedia.org/wiki/Wikipedia:Namespace) 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’expert ne prend pas en compte ces utilisateurs fictifs.

## Adresse IP

Certaines modifications n’ont pas de username, elles ont une adresse IP. Le programme de sélection d’experts ne prend pas en compte ces 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)

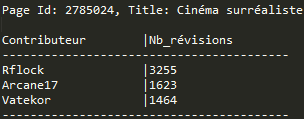
or rv.page\_title in (SELECT pl\_title FROM pagelink\_from\_me)

group by contributeur order by quantite desc

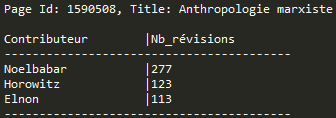
Avec <titre> étant le titre du document de référence.

## Les résultats

### Cinéma surréaliste



### Anthropologie marxiste



### Dennō Senshi Porigon

