# Début du projet et pre-processings

# INF729 – Spark – TP 2 (début du projet)

Xavier Bracquart

24/10/2019

Sujet: <https://github.com/Flooorent/cours-spark-telecom>

# Etape 1: Setup

Setup: [https://github.com/Flooorent/cours-spark-telecom/blob/master/setup.md#setup-tp-1-mac-et-linux-installation-de-spark](https://github.com/Flooorent/cours-spark-telecom/blob/master/setup.md" \l "setup-tp-1-mac-et-linux-installation-de-spark)

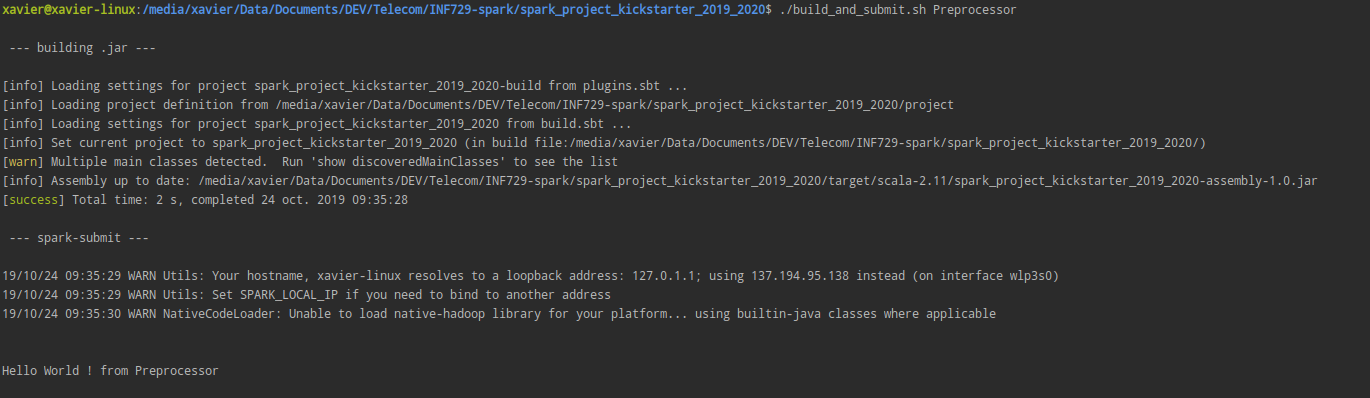
SBT et IntelliJ ont été installés à la séance précédente.

On télécharge et lance le projet sur IntelliJ.

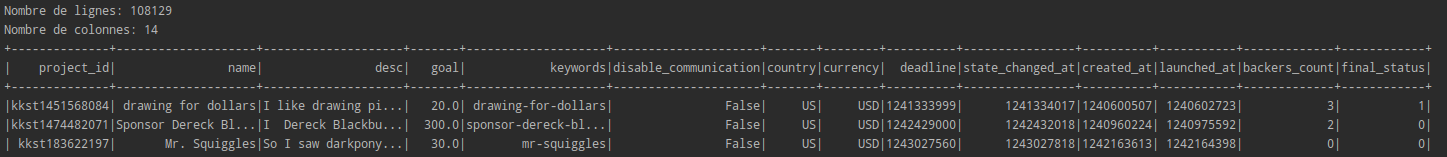
On vérifie que le script build\_and\_submit.sh et exécutable et on y modifie le chemin d’accès de Spark.

On exécute le script (dans un terminal classique, ou celui intégré à Intellij) avec :

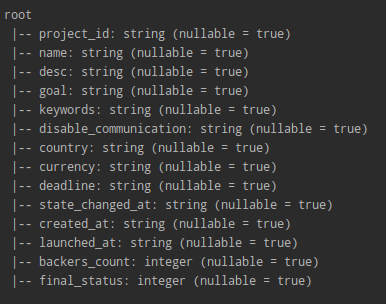
> ./build\_and\_submit.sh Preprocessor



# Etape 2 : Chargement des données

On affiche le nombre de lignes, de colonnes et un extrait du dataframe.

On affiche le schema du dataframe :

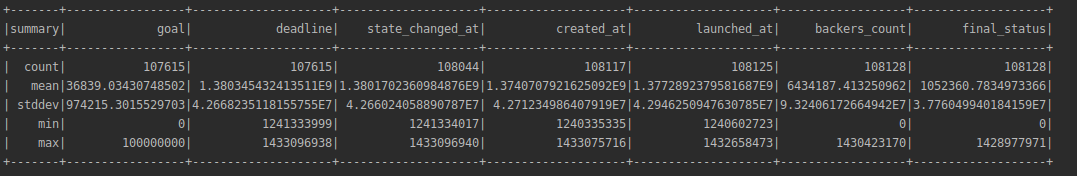


On change ensuite le type de plusieurs colonnes en int.

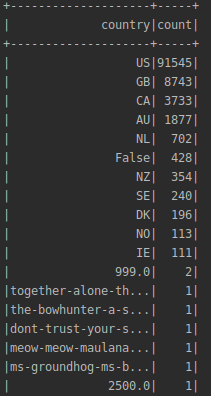
## Etape 3 : Cleaning

On analyse le contenu du dataframe, colonne par colonne, à l’aide des fonctions describe et show notamment.

Describe des colonnes de type int :

On remarque que le nombre de lignes avec une deadline et un goal est différent du nombre de lignes avec une date de lance (environ 500 lignes d’écart). On remarque que la valeur maximale dans les colonnes backers\_count et final\_status est très élevé (alors qu’il devrait être de 1 pour final\_status par exemple). Ces valeurs semblent être des timestamp.

On affiche les colonnes avec les fonctions proposées sur le sujet. Certaines colonnes sont incohérentes et semblent contenir parfois des données venant d’autres colonnes. Ex :

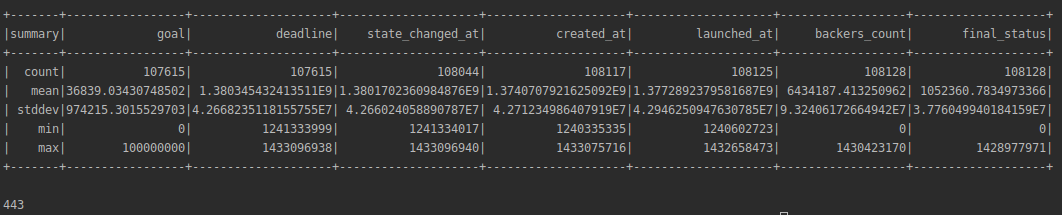


On enlève les colonnes disable\_communication (trop peu de valeurs) et backers\_count et state\_changed\_at (fuites du futur).

Certaines lignes du csv, qui contiennent des virgules dans la description (la virgule est aussi le délimiteur du csv) ont mal été lues. On ajoute donc ces deux options à la lecture du csv :

.option("quote", "\"")

.option("escape", "\"")



Seules 443 lignes étaient concernées (cf les valeurs incohérentes dans final\_status par exemple (la valeur maximale devrait être 1). C’est peu par rapport aux 108 000 lignes de départ, mais le fait de les enlever facilitera le data cleaning.