Big Data & BI

Analyse Big Data avec Apache Spark

Pour l’analyse des données, nous utilisons le framework Spark. C’est une solution de la fondation Apache pour accélérer le traitement des systèmes [Hadoop](https://fr.wikipedia.org/wiki/Hadoop" \o "). Il est rapide en termes d'exécution des tâches par rapport à [MapReduce](https://fr.wikipedia.org/wiki/MapReduce" \o ").

1. **Installer Apache Spark:**

Visiter le site web et suivez les étapes d’installation pour votre système: <https://docs.mongodb.com/v3.2/administration/install-community/>

1. **Installer PySpark:**

Afin de développer des scripts d’analyse en Spark, installez pyspark en suivant les étapes des liens suivants :

<https://pypi.org/project/pyspark/>

<https://www.sicara.ai/blog/2017-05-02-get-started-pyspark-jupyter-notebook-3-minutes>

pour connecter pyspark avec mongodb et manipuler les données stockées dans les collections de la BD suivez les étapes suivantes:

<https://docs.mongodb.com/spark-connector/v1.1/python-api/>

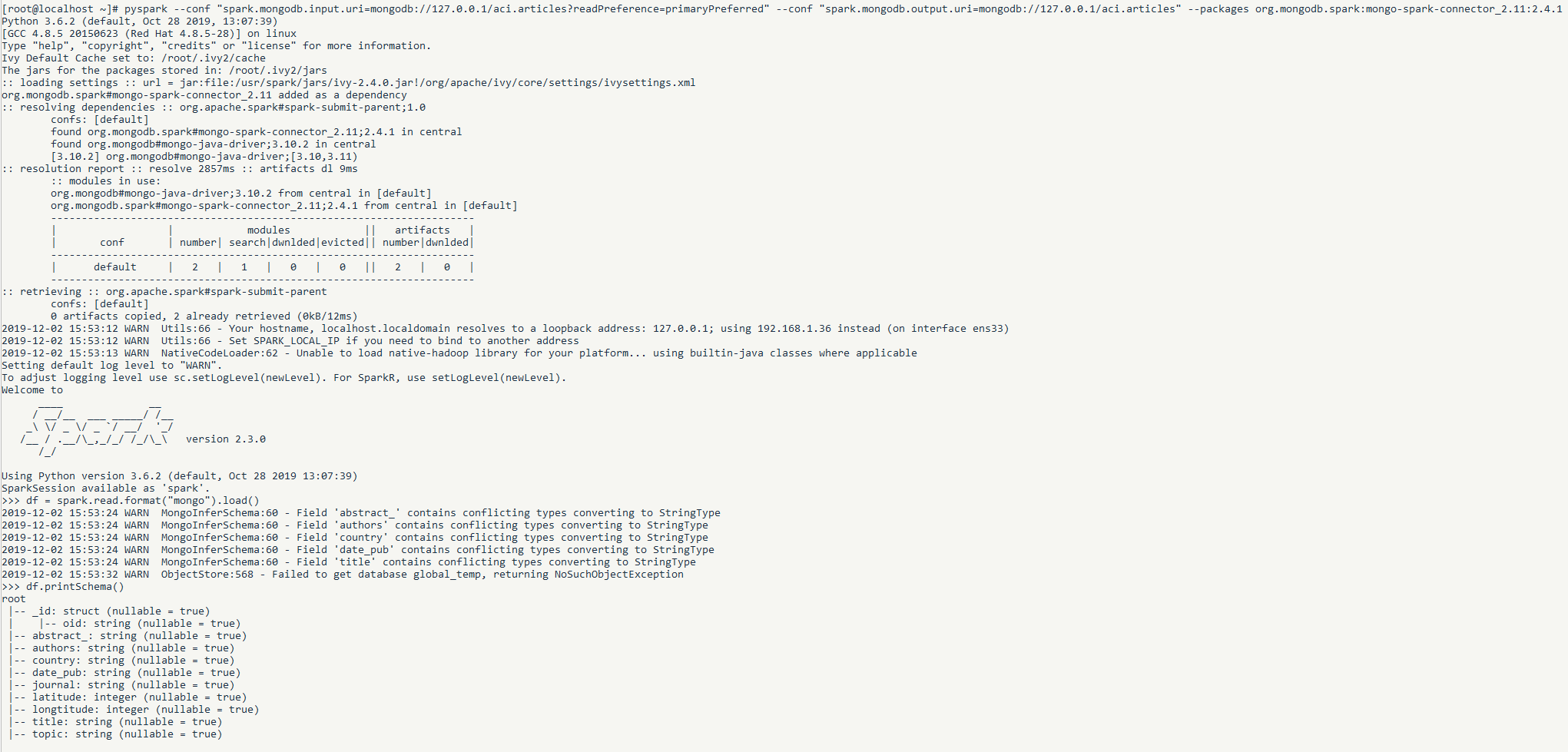
1. **Exmples d’analyse:**

L’objectif de ces exemples est de charger les données à partir du mongoDB et lancer des traitments statistiques pour récupérer les informations implicites contenu dans le corpus collecté.

Lancer la commande suivante pour lancer le shell pyspark :

* pyspark --conf "spark.mongodb.input.uri=mongodb://127.0.0.1/aci.articles?readPreference=primaryPreferred" --conf "spark.mongodb.output.uri=mongodb://127.0.0.1/aci.articles" --packages org.mongodb.spark:mongo-spark-connector\_2.11:2.4.1

Note : remplacer aci.articles par “le nom de la base de données”.”le nom de la collection des articles”

****

Sur le Shell Spark, Lancer la commande suivante pour récupérer les données :

* df = spark.read.format("mongo").load()

Pour vérifier le schéma et la structure de nos items :

* df.printSchema()

Afficher les premiers 5 documents :

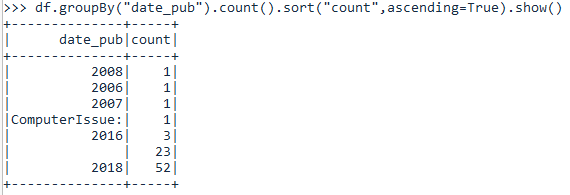
* df.show(5, truncate = False)

Sélectioner des champs précis du document :

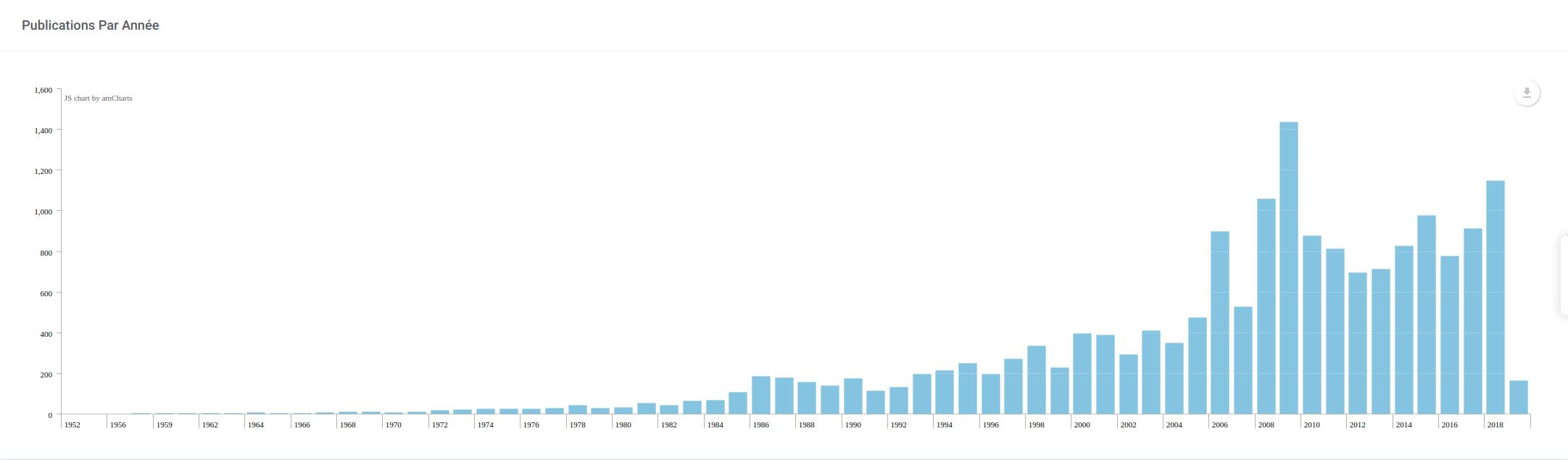
* df.select('title','country').show(5)
  1. **Statistiques des publications par année:**

Pour compter le nombre des articles par date de publication :

* df.groupBy("date\_pub").count().sort("count",ascending=True).show()

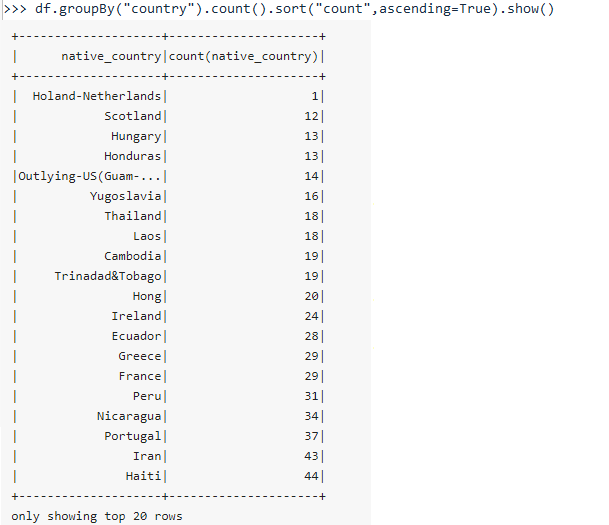
****

En utilisant une bibliothèque de votre choix, afficher les résultats obtenues dans un graphe.

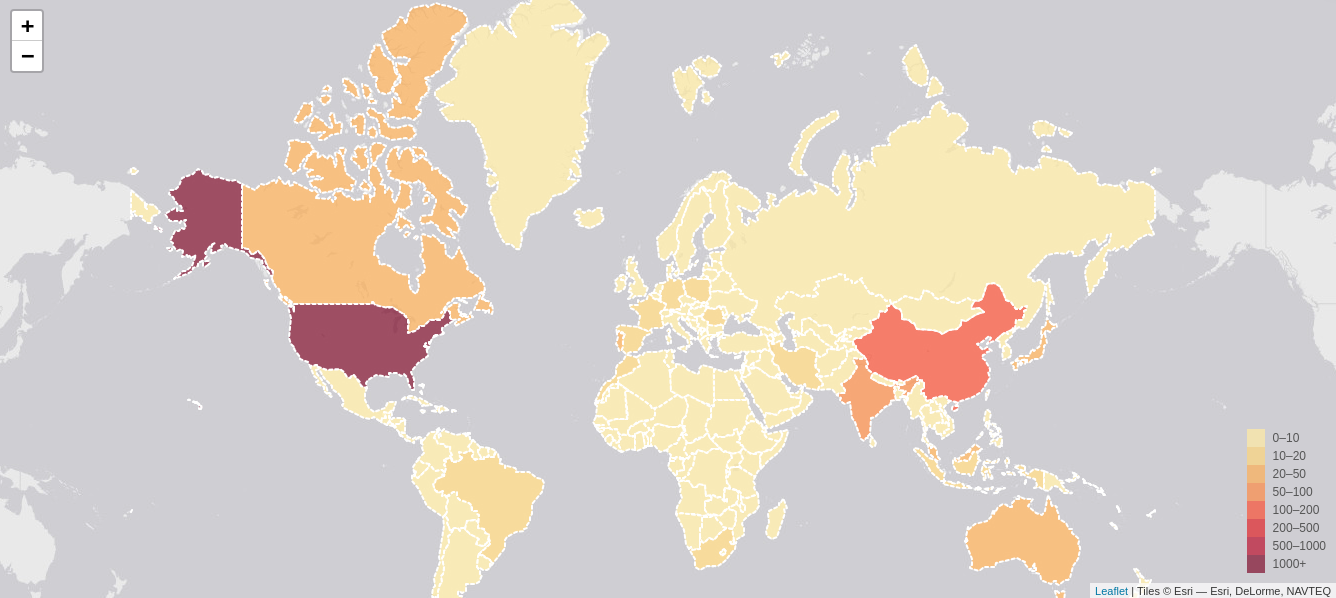
****

* 1. **Statistiques des publications par pays:**

Pour compter le nombre des publications par pays :



En utilisant une bibliothèque de votre choix, afficher les résultats obtenues dans un carte.

****

1. **Travail à faire :** 
   1. **Développer une API d’analyse BI.**
   2. **Explorer les données et extraire les autres indicateurs que vous jugez intéressants.**
   3. **Développer des visualisations adéquates.**

**Webographie :**