**COMPTE RENDU PROJET DATA INTEGRATION**

**LAM STEPHANE / MARVIN MEZOUANI / VICTOIRE BORLETSIS**

DEFINIR LES PATHS APRES AVOIR TOUT INSTALLER (voir README)

**Configuration de Hadoop :**

* Modifiez le fichier core-site.xml (dans le dossier etc/hadoop) pour indiquer où stocker vos fichiers.
* Modifiez hdfs-site.xml pour définir le dossier où les données seront stockées.

**Vérifiez que HDFS fonctionne :**

* Ligne de commande :

./bin/hdfs namenode -format

./sbin/start-dfs.sh

* Accédez à <http://localhost:9870>, on voit bien le site en local.

**Vérifiez que Spark fonctionne :**

* Ligne de commande : spark-shell

**Vérifier que Kafka fonctionne :**

* Démarrer dans un terminal Ubuntu :

bin/zookeeper-server-start.sh config/zookeeper.properties

* Démarrer dans un autre terminal Ubuntu :

bin/kafka-server-start.sh config/server.properties

* Création d’un sujet Kafka :

bin/kafka-topics.sh --create --topic student\_loan\_data--bootstrap-server localhost:9092 --partitions 1 --replication-factor 1

**Enrichier et charger les fichiers sur HDFS :**

* Démarrez HDFS : start-dfs.sh
* Pour verifier l’état de HDFS : jps (on verra tout les Nodes)
* hdfs dfs -mkdir /student\_loan\_data
* hdfs dfs -put FL\_Dashboard\* /student\_loan\_data
* Pour verifier si nos fichiers sont bien dans HDFS : hdfs dfs -ls /student\_loan\_data

**AVANT DE LANCER UN SCRIPT KAFKA :**

* pip install kafka-python

**Lancer un producer Kafka :**

* Le script kafka\_producer diffuse les données local du dataset et grâce au zookeeper on peux voir les données qu’envoie le producer : python3 kafka\_producer.py

**Lancer un consumer Kafka avec Spark :**

* spark-submit --packages com.crealytics:spark-excel\_2.12:0.13.5 kafka\_consumer.py
* spark-submit --packages org.apache.spark:spark-sql-kafka-0-10\_2.12:3.3.0 scripts/spark\_streaming.py (Spark Streaming sera utilisé pour lire les messages de Kafka et les intégrer dans le système)

Problème : le format des fichiers à prendre en compte pour pouvoir les lire.

**Aborescence :**

DataInteg : contient le projet et tout les outils de data integration pour le projet (Hadoop / Spark / Kafka)

Dans le fichier student\_loan\_project :

* raw\_data : contient les fichiers de données brutes
* script : contient les script python du projet (j’ai tester plusieurs type de script différent pour mener à bien mon projet)

**Le fichier script contient :**

* kafka\_producer : qui n’est autre que le producer du zookeeper Kafka
* kafka\_consumer : qui va consommer les données envoyer
* load\_and\_merge\_dashboard : qui va charger les données depuis la base de données HDFS en testant la connexion à HDFS. Il va aussi faire les jointures des fichiers puis les stocker dans un nouveau fichier sur HDFS
* read\_parquet\_hdfs : qui va lire les fichiers aux formats parquets pour une meilleure compréhension de la base de données où les fichiers sont séparer/diviser en plusieurs fichier (format parquet)
* spark\_streaming : pour tester si le streaming marchait et respectait bien le schéma JSON donnée