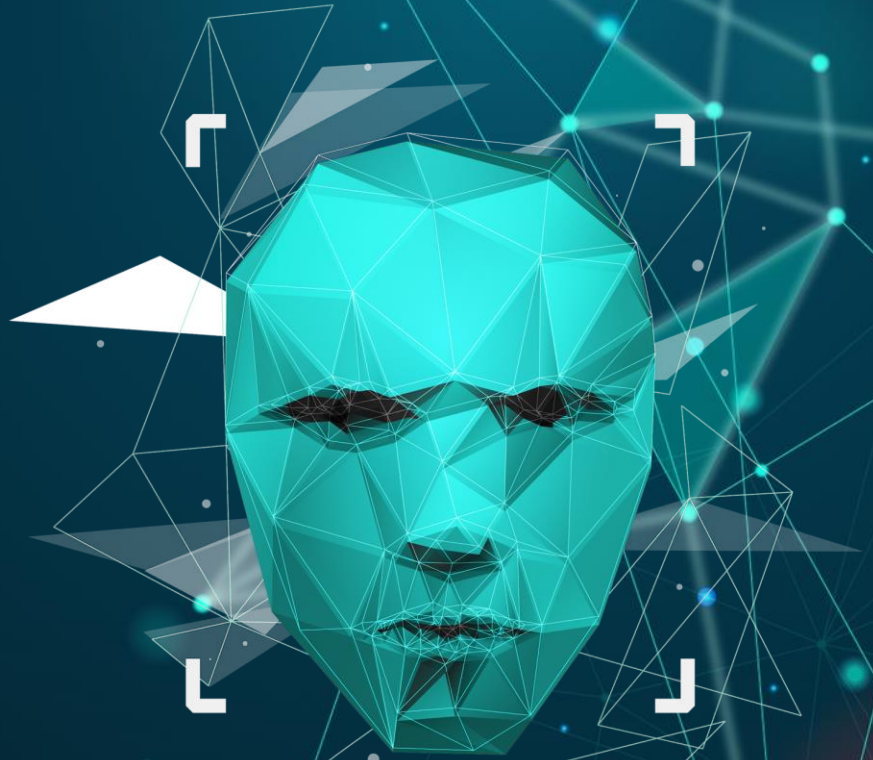
63-51 PROJET

Nerman Muminovic
Mohamed Abdurahman



Sommaire

01 Intégration des données

02 Outils Cloudera

03 Code

04 Résultats

Intégration données

Outils Cloudera

- HDFS (stocker les données)
- SPARK (Traîter les données)
- KAFKA (Récupérer les données)
- HIVE (Créer table pour requêtes)

Code

```
def save_to_hadoop(t, rdd):
    print("=====Pull from Stream=====")
    if not rdd.isEmpty():
        now = datetime.now()
        current_time = now.strftime("%H:%M:%S")
        hour = now.strftime("%H")
        print("=some records=")
        rdd=rdd.map(lambda x: (x[0],x[1],str(current_time),str(hour)))
        print(str(t))
        print(rdd.collect())
        df = rdd.toDF().withColumnRenamed("_1", "theme").withColumnRenamed("_2", "count").withColumnRenamed("_3", "time").withColumnRenamed(
"_4", "hour")
        df.printSchema()
        spark = SparkSession.builder.getOrCreate
        df.write.format("parquet").mode("append").save("/user/spark/test")
```

- Changement de port dans twitter_reader.py et twitter_spark.py
- Remplacer le localhost dans le code par sandbox-hdp.hortonworks.com
- Création de la fonction save_to_hadoop(t, rdd) pour utilisation de HDFS (ci-dessus)
- Suppression de HBASE et remplacer par HDFS
- Changement les tokens toutes les 24h afin de pouvoir lancer le script spark

Résultats

DEMO

Merci !

Avez-vous des questions ?