

**63-51**

# **PROJET**

**Nerman Muminovic  
Mohamed Abdurahman**



# Sommaire

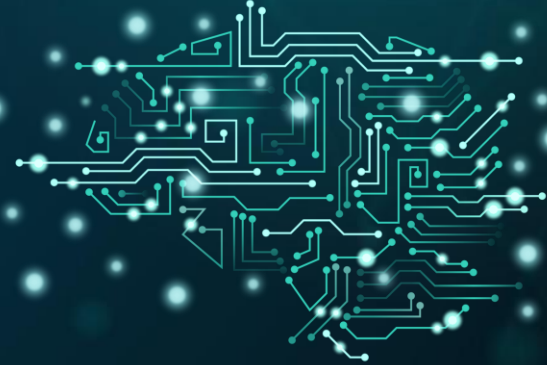
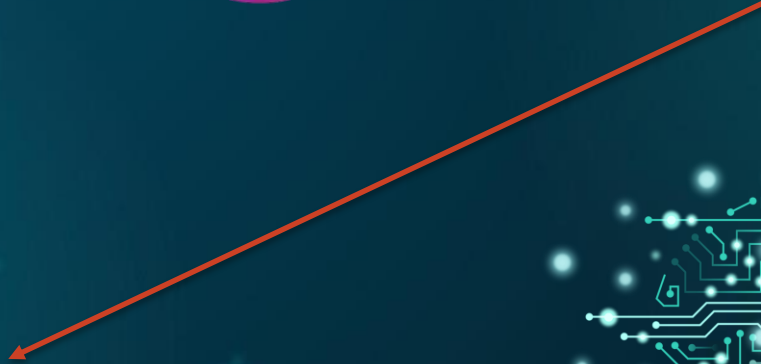
01 | Intégration des données

02 | Outils Cloudera

03 | Code

04 | Résultats

# Intégration données



# Outils Cloudera



- **HDFS (stocker les données)**
- **SPARK (Traiter 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](https://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**



The background is a dark teal color with a complex network of thin, light teal lines connecting various points. Some points are small, bright teal dots, while others are larger, glowing teal spheres. There are also some smaller, dimmer blue and orange dots scattered throughout. The overall effect is a futuristic, digital, or network-like aesthetic.

# Merci !

Avez-vous des questions ?