

DÉPARTEMENT MATHÉMATIQUE INFORMATIQUE

Big Data : Fondements et Architectures de stockage

Rapport

Examen

Réalisé par :

Ayoub ETOULLALI

Professeur :

Mr. BOUSSELHAM Abdelmajid

2ème année II-BDCC

Filière d'ingénieur : Ingénieur informatique, Big Data et Cloud Computing

SOMMAIRE

Introduction	2
Travail à faire	3
Exercice 1: Manipuler le système de fichiers HDFS	3
Exercice 2:	6
Partie 1 : Spark SQL	6
Partie 2 : Importer et exporter des données avec SQOOP	9
Partie 3: Traitement de données en streaming	10
Conclusion	13



L'analyse des données et la prise de décision basées sur ces données sont devenues des aspects cruciaux pour de nombreuses entreprises. Dans le domaine du Big Data, la capacité à manipuler, traiter et extraire des informations significatives à partir de grands ensembles de données est essentielle. Cet examen pratique en Big Data vise à évaluer vos compétences dans la manipulation du système de fichiers HDFS ainsi que dans l'utilisation des outils de traitement de données distribuées tels que Spark SQL.

Au cours de cet examen, vous serez confronté à différentes tâches liées à la gestion des données à grande échelle. Vous devrez effectuer des opérations telles que la création de répertoires, le chargement de fichiers, l'affichage du contenu des répertoires, la manipulation des fichiers, et bien d'autres encore. Cela vous permettra de mettre en pratique vos connaissances en Big Data et de démontrer votre compréhension des concepts clés.

Exercice 1: Manipuler le système de fichiers HDFS

1. Vérifiez la version Hadoop.

```
ayoub@ACER:~$ hadoop version
Hadoop 3.3.2
Source code repository git@github.com:apache/hadoop.git -r 0bcb014209e219273cb6fd4152df7df713cbac61
Compiled by chao on 2022-02-21T18:39Z
Compiled with protoc 3.7.1
From source with checksum 4b40fff8bb27201ba07b6fa5651217fb
This command was run using /home/ayoub/hadoop/hadoop-3.3.2/share/hadoop/common/hadoop-common-3.3.2.jar
```

2. Démarrez HDFS et vérifiez qu'il est en cours d'exécution.

```
ayoub@ACER:~$ sudo service ssh restart
[sudo] password for ayoub:
* Restarting OpenBSD Secure Shell server sshd
ayoub@ACER:~$
ayoub@ACER:~$ ssh localhost
Welcome to Ubuntu 22.04.2 LTS (GNU/Linux 5.10.102.1-microsoft-standard-WSL2 x86_64)

* Documentation:  https://help.ubuntu.com
* Management:    https://landscape.canonical.com
* Support:       https://ubuntu.com/advantage

Last login: Thu Jun  1 21:15:33 2023 from ::1
ayoub@ACER:~$ jps
_HOME/sbin/start-dfs.sh
jps
$HADOOP_HOME/sbin/start-yarn.sh
jps87 jps
ayoub@ACER:~$ $HADOOP_HOME/sbin/start-dfs.sh

Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [ACER]
ayoub@ACER:~$ jps
228 NameNode
327 DataNode
487 SecondaryNameNode
671 Jps
ayoub@ACER:~$ $HADOOP_HOME/sbin/start-yarn.sh
Starting resourcemanager
Starting nodemanagers
ayoub@ACER:~$ jps
769 ResourceManager
228 NameNode
982 Jps
327 DataNode
487 SecondaryNameNode
875 NodeManager
```

3. Créez deux nouveaux répertoires nommés /enset/bddc et /enset/glsid sur HDFS.

```
ayoub@ACER:~$ hdfs dfs -mkdir /enset
ayoub@ACER:~$ hdfs dfs -mkdir /enset/bddc
ayoub@ACER:~$ hdfs dfs -mkdir /enset/glsid
```

reset

Go!

Show 25 entries

Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	drwxr-xr-x	ayoub	supergroup	0 B	Jun 02 14:35	0	0 B	bddc	<div></div>
<input type="checkbox"/>	drwxr-xr-x	ayoub	supergroup	0 B	Jun 02 14:35	0	0 B	glsid	<div></div>

4. Créez un nouveau fichier java.txt contenant 10 lignes et cpp.txt contenant 10 lignes sur votre système local.

java	02/06/2023 14:37	Document texte	0 Ko
cpp	02/06/2023 14:37	Document texte	0 Ko



4. Charger le fichier java.txt dans /enset/bddc et cpp.txt dans /enset/glsid sur HDFS.





```
ayoub@ACER:~$ hdfs dfs -put ./Tests/java.txt /enset/bddc
ayoub@ACER:~$ hdfs dfs -put ./Tests/cpp.txt /enset/glsid
```

5. Afficher le contenu du répertoire /enset/bddc et /enset/glsid.

`hdfs dfs -ls -R /enset/glsid`


lenset/glsid

Go!



Show 25 entries




Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	-rw-r--r--	ayoub	supergroup	11 B	Jun 02 14:41	1	128 MB	cpp.txt	

`hdfs dfs -ls -R /enset/bddc`


/enset/bdccc

Go!



Show 25 entries

Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	-rw-r--r--	ayoub	supergroup	12 B	Jun 02 14:41	1	128 MB	java.txt	

6. Affichez le contenu du fichier java.txt qui se trouve dans HDFS.

```
ayoub@ACER:~$ hdfs dfs -cat /enset/bddc/java.txt
java : ayoubayoub@ACER:~$
```

7. Déterminez la taille du fichier cpp.txt qui se trouve dans HDFS.

```
ayoub@ACER:~$ hdfs dfs -du -h /enset/glsid/cpp.txt
11 11 /enset/glsid/cpp.txt
```

File information - cpp.txt

Download Head the file (first 32K) Tail the file (last 32K)

Block information - Block 0

Block ID: 1073744674
 Block Pool ID: BP-566145988-127,0.1.1-1684158441564
 Generation Stamp: 3851
 Size: 11
 Availability:
 • ACER.localdomain

8. Déplacez le fichier cpp.txt vers /enset/bdcc et vérifiez si le fichier est bien déplacé.

```
ayoub@ACER:~$ hdfs dfs -mv /enset/glsid/cpp.txt /enset/bdcc
ayoub@ACER:~$ hdfs dfs -ls /enset/bdcc
-rw-r--r--  1 ayoub supergroup      11 2023-06-02 14:41 /enset/bdcc
```

Vider /enset/glsid

/enset/glsid

Show 25 entries

Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
No data available in table							

Remplir /enset/bdcc

/enset/bdcc

Show 25 entries

Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	ayoub	supergroup	11 B	Jun 02 14:53	1	128 MB	cpp.txt
-rw-r--r--	ayoub	supergroup	12 B	Jun 02 14:41	1	128 MB	java.txt

10. Supprimez les fichiers java.txt et cpp.txt dans HDFS.

```
ayoub@ACER:~$ hdfs dfs -rm /enset/bdcc/*
Deleted /enset/bdcc/cpp.txt
Deleted /enset/bdcc/java.txt
```

/enset/bdcc

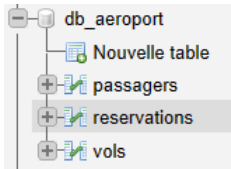
Show 25 entries

Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
No data available in table							

Exercice 2:

Partie 1 : Spark SQL



	ID	NOM	PRENOM	TEL
<input type="checkbox"/> Éditer Copier Supprimer	1	ETOULLALI	ayoub	0658712011
<input type="checkbox"/> Éditer Copier Supprimer	2	ETOULLALI	hayat	0645589714

	ID	DATE_DEPART	DATE_ARRIVEE
<input type="checkbox"/> Éditer Copier Supprimer	1	2023-06-01	2023-06-02
<input type="checkbox"/> Éditer Copier Supprimer	2	2023-06-01	2023-06-03

+ Options

	ID	DATE_RESERVATION	ID_PASSAGER	ID_VOL
<input type="checkbox"/> Éditer Copier Supprimer	1	2023-06-01	1	1
<input type="checkbox"/> Éditer Copier Supprimer	2	2023-06-02	1	2
<input type="checkbox"/> Éditer Copier Supprimer	3	2023-06-01	1	1
<input type="checkbox"/> Éditer Copier Supprimer	4	2023-06-01	1	1

1. Afficher pour chaque vol, le nombre de passagers selon le format d'affichage suivant :

ID_VOL |DATE DEPART| NOMBRE

```
1 SELECT v.ID, v.DATE_DEPART, COUNT(r.ID_PASSAGER) AS NOMBRE
2 FROM VOLS v
3 LEFT JOIN RESERVATIONS r ON v.ID = r.ID_VOL
4 GROUP BY v.ID, v.DATE_DEPART;
5
```

ID	DATE_DEPART	NOMBRE
1	2023-06-01	3
2	2023-06-01	1

2. Afficher la liste des vols en cours selon le format d'affichage suivant :

ID_VOL |DATE DEPART| DATE ARRIVEE

```
1 SELECT ID, DATE_DEPART, DATE_ARRIVEE
2 FROM VOLS;
```

+ Options

<input type="checkbox"/> Éditer Copier Supprimer	ID	DATE_DEPART	DATE_ARRIVEE
<input type="checkbox"/> Éditer Copier Supprimer	1	2023-06-01	2023-06-02
<input type="checkbox"/> Éditer Copier Supprimer	2	2023-06-01	2023-06-03

Problème dans Xampp

Voir le code tel que la vérification des requêtes sont dans MySQL ci-dessous

La réponse dans les deux versions : Dataframes et QuerySql

Dataframes

```
package ma.enset.dataMysql;

import org.apache.log4j.Level;
import org.apache.log4j.Logger;
import org.apache.spark.sql.*;

public class Dataframes {

    public static void main(String[] args) {

        Logger.getLogger("org").setLevel(Level.OFF);

        SparkSession spark = SparkSession.builder()
            .appName("Spark SQL")
            .config("spark.master", "local")
            .getOrCreate();

        Dataset<Row> vols_df = spark.read()
            .format("jdbc")
            .option("url", "jdbc:mysql://localhost:3306/db_aeroport")
            .option("dbtable", "VOLS")
            .option("user", "root")
            .option("password", "")
            .load();

        Dataset<Row> passagers_df = spark.read()
            .format("jdbc")
            .option("url", "jdbc:mysql://localhost:3306/db_aeroport")
            .option("dbtable", "PASSAGERS")
            .option("user", "root")
            .option("password", "")
            .load();

        Dataset<Row> reservations_df = spark.read()
            .format("jdbc")
            .option("url", "jdbc:mysql://localhost:3306/db_aeroport")
            .option("dbtable", "RESERVATIONS")
            .option("user", "root")
            .option("password", "")
            .load();

        // Question 1: Afficher le nombre de passagers par vol
        Dataset<Row> resultat_question1 = vols_df.join(reservations_df,
vols_df.col("ID_VOL").equalTo(reservations_df.col("ID_VOL")), "left")
            .groupBy(vols_df.col("ID_VOL"), vols_df.col("DATE_DEPART"))
            .agg(functions.count(reservations_df.col("ID_PASSAGER")).alias("NOMBRE"));

        resultat_question1.show();

        // Question 2: Afficher la liste des vols en cours
        Dataset<Row> resultat_question2 = vols_df.select(vols_df.col("ID_VOL"),
vols_df.col("DATE_DEPART"), vols_df.col("DATE_ARRIVE"));
        resultat_question2.show();

        // Fermeture de la session Spark
        spark.stop();
    }
}
```



```

package ma.enset.dataMysql;

import org.apache.log4j.Level;
import org.apache.log4j.Logger;
import org.apache.spark.sql.Dataset;
import org.apache.spark.sql.Row;
import org.apache.spark.sql.Session;

import java.util.HashMap;
import java.util.Map;

public class QuerySql {

    public static void main(String[] args) {

        Logger.getLogger("org").setLevel(Level.OFF);
        SparkSession ss = SparkSession
            .builder()
            .master("local[*]")
            .appName("Traitement de données stockées dans Mysql [Query SQL]")
            .getOrCreate();

        Map<String, String> options = new HashMap<>();
        options.put("driver", "com.mysql.cj.jdbc.Driver");
        options.put("url", "jdbc:mysql://localhost:3306/db_aeroport");
        options.put("user", "root");
        options.put("password", "");

        // Question 1: Afficher le nombre de passagers par vol
        Dataset<Row> df1 = ss.read().format("jdbc")
            .options(options)
            .option("query", "SELECT v.ID, v.DATE_DEPART, COUNT(r.ID_PASSAGER) AS
NOMBRE\n" +
                "FROM VOLS v\n" +
                "LEFT JOIN RESERVATIONS r ON v.ID = r.ID_VOL\n" +
                "GROUP BY v.ID, v.DATE_DEPART;\n")
            .load();
        df1.show();
        // Question 2: Afficher la liste des vols en cours
        Dataset<Row> df2 = ss.read().format("jdbc")
            .options(options)
            .option("query", "SELECT ID_VOL, DATE_DEPART, DATE_ARRIVE\n" +
                "FROM VOLS;\n")
            .load();
        df2.show();
    }
}

```

Partie 2 : Importer et exporter des données avec SQOOP


```
ayoub@ACER:~$ sqoop version
Warning: /opt/sqoop/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /opt/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /opt/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /opt/sqoop/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
2023-06-02 18:03:24,710 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Sqoop 1.4.7
git commit id 2328971411f57f0cb683dfb79d19d4d19d185dd8
Compiled by maugli on Thu Dec 21 15:59:58 STD 2017
```

Sqoop ça ne marche pas, j'ai un problème dans MySQL

```
ayoub@ACER:~$ sqoop import --connect jdbc:mysql://localhost:3306/db_aeroport --username "root" --password "" --table VOLS --target-dir /sqoop
Warning: /opt/sqoop/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /opt/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /opt/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /opt/sqoop/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
2023-06-02 18:04:14,362 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
2023-06-02 18:04:14,430 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
2023-06-02 18:04:14,666 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2023-06-02 18:04:14,667 INFO tool.CodeGenTool: Beginning code generation
2023-06-02 18:04:15,104 ERROR manager.SqlManager: Error executing statement: com.mysql.jdbc.exceptions.jdbc4.CommunicationsException: Communications link failure

The last packet sent successfully to the server was 0 milliseconds ago. The driver has not received any packets from the server.
com.mysql.jdbc.exceptions.jdbc4.CommunicationsException: Communications link failure
```

Voici le fichier vols.txt

 vols - Bloc-notes

Fichier Edition Format Affichage Aide

```
1,Vol 1,Avion 1,2023-01-01
2,Vol 2,Avion 2,2023-02-15
3,Vol 3,Avion 1,2023-03-10
|
```

Sqoop ça ne marche pas

Voir le code ci-dessous :

1. Importation des données de la table VOLS dans HDFS :

```
sqoop import --connect jdbc:mysql://localhost:3306/db_aeroport --username "root" --password "" --table VOLS --target-dir /sqoop
```

2. Exportation des données du fichier vols.txt vers la table VOLS en utilisant Sqoop :

```
sqoop export --connect jdbc:mysql://localhost:3306/db_aeroport --username "root" --password "" --table VOLS --export-dir /aeroport/vols.txt
```

Partie 3: Traitement de données en streaming

	A	B	C
1	id,description,nom_avion,date		
2	1,description1,avion1,01/06/2023		
3	2,description2,avion2,01/06/2023		
4	3,description3,avion1,02/06/2023		
5	4,description4,avion1,02/06/2023		
6	5,description5,avion2,02/06/2023		
7			

```
ayoub@ACER:~$ hdfs dfs -put ./Tests/incident1.csv /incidents
ayoub@ACER:~$
```

```
Dataframes x
/usr/lib/jvm/java-1.8.0-openjdk-amd64/bin/java ...
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties

Batch: 0

+---+-----+-----+-----+
| id|  nom_avion|description|    date|
+---+-----+-----+-----+
|  1|description1|  avion1|01/06/2023|
|  2|description2|  avion2|01/06/2023|
|  3|description3|  avion1|02/06/2023|
|  4|description4|  avion1|02/06/2023|
|  5|description5|  avion2|02/06/2023|
+---+-----+-----+-----+
```

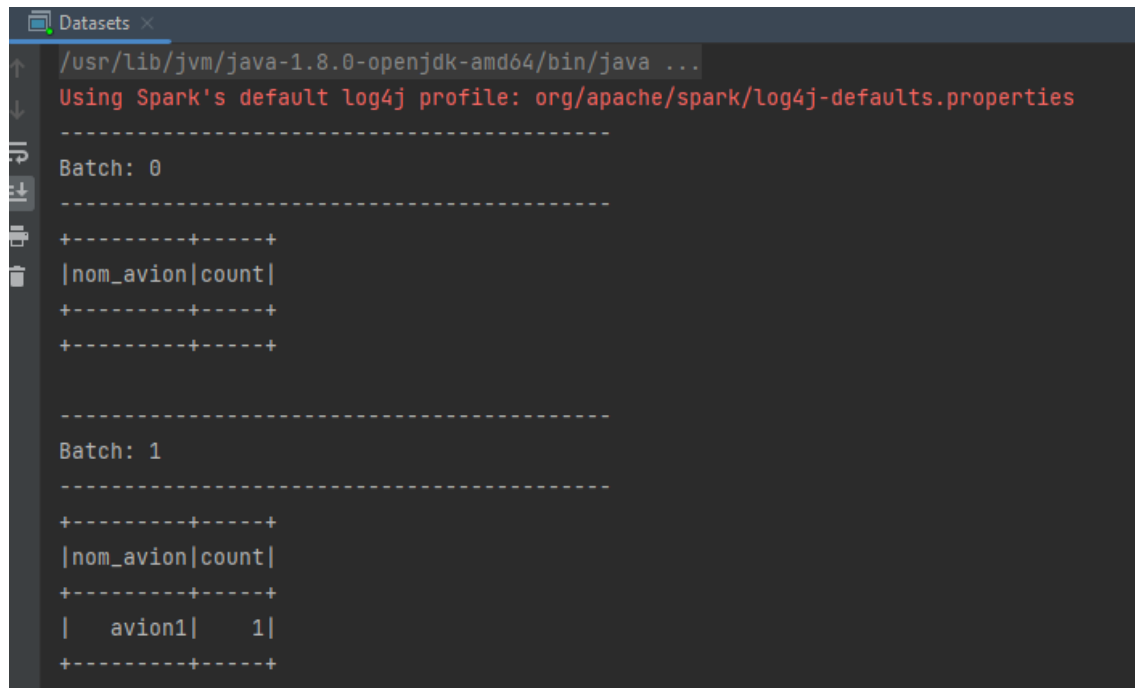
1. Afficher d'une manière continue l'avion ayant plus d'incidents.

Ici c'est l'avion « avion1 » (3 incidents)

```
-----
+-----+
|nom_avion|
+-----+
|  avion1|
+-----+
```

Avec DataStreaming :

```
ayoub@ACER:~$ nc -lk 9090
1,description1,avion1,01/05/2023
2,description2,avion2,01/03/2023
3,description3,avion1,02/04/2023
4,description4,avion1,02/03/2023
```

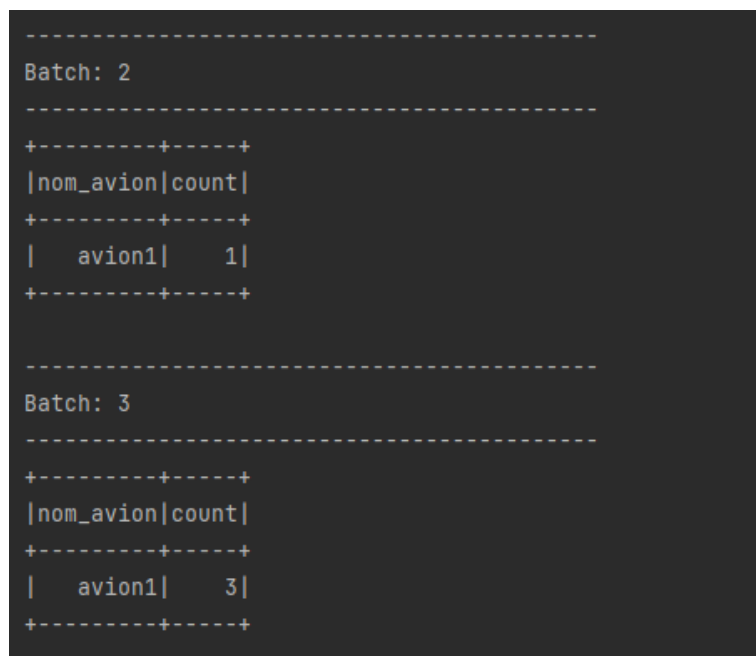


The screenshot shows a terminal window titled "Datasets" with a dark background. It displays the output of a Spark DataStream. The first batch (Batch: 0) shows a header row with columns "nom_avion" and "count". The second batch (Batch: 1) shows a single row with "avion1" and "1".

```
/usr/lib/jvm/java-1.8.0-openjdk-amd64/bin/java ...
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties

Batch: 0
-----
+-----+-----+
|nom_avion|count|
+-----+-----+
+-----+-----+

Batch: 1
-----
+-----+-----+
|nom_avion|count|
+-----+-----+
|  avion1|    1|
+-----+-----+
```



The continuation of the terminal window shows the output for Batches 2 and 3. Batch 2 shows a single row with "avion1" and "1". Batch 3 shows a single row with "avion1" and "3".

```
-----
Batch: 2
-----
+-----+-----+
|nom_avion|count|
+-----+-----+
|  avion1|    1|
+-----+-----+

Batch: 3
-----
+-----+-----+
|nom_avion|count|
+-----+-----+
|  avion1|    3|
+-----+-----+
```

2. Afficher d'une manière continue les deux mois de l'année en cours où il a y avait moins d'incidents.

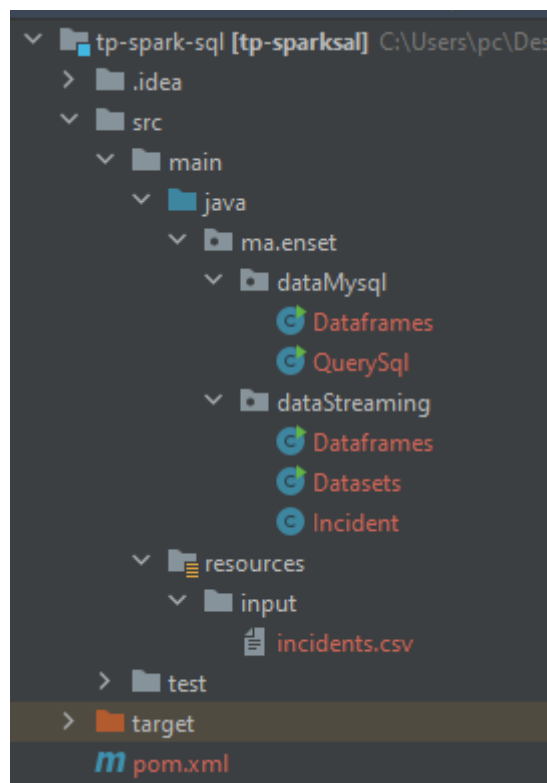
J'ai changé la base de données :

```
id,description,nom_avion,date
1,description1,avion1,01/05/2023
2,description2,avion2,01/03/2023
3,description3,avion1,01/04/2023
4,description4,avion1,01/03/2023
5,description5,avion2,01/06/2023
6,description6,avion2,01/04/2023
```

Alors les deux mois de l'année en cours où il a y avait moins d'incidents sont les mois 05 et 06

```
-----+
+-----+
|      date|
+-----+
|01/06/2023|
|01/05/2023|
+-----+
```

Projet :



Au cours de cet examen, j'ai pu manipuler le système de fichiers HDFS, charger des fichiers, créer des répertoires et effectuer diverses opérations sur les données. J'ai également utilisé Spark SQL pour l'analyse et le traitement des données distribuées, ce qui m'a permis d'extraire des informations utiles et de prendre des décisions éclairées.

Cette expérience m'a permis de mieux comprendre les défis et les opportunités liés à la gestion des données à grande échelle. J'ai réalisé l'importance de pouvoir travailler efficacement sur des ensembles de données massifs et de savoir utiliser des outils tels que Spark pour optimiser les performances et obtenir des résultats précis.