Домашнее задание №4

Выполнил: **Умертаев Арслан Наушанович БПИ227**

Исходные данные:

Кластер HDFS на 3 применованных нодах (**team-9-nn, team-9-dn-00, team-9-dn-01**)

Username: team

Узел для входа **176.109.91.11** - <global_ip>, JumpNode 192.168.1.38 - <local_ip>

Инициализация

0. Подключение

```
ssh <user_name>@<global_ip>
sudo -i -u hadoop
```

1. Скачаем и разархивируем дистрибутив Spark

```
wget https://archive.apache.org/dist/spark/spark-3.5.3/spark-3.5.3-bin-hadoop3.tgz
tar -xzvf spark-3.5.3-bin-hadoop3.tgz
```

2. Установим pip и venv

```
sudo apt install python3-venv
sudo apt install python3-pip
```

3. Добавим переменные окружения

```
export HADOOP_CONF_DIR="/home/hadoop/hadoop-3.4.0/etc/hadoop"
export HIVE_HOME="/home/hadoop/apache-hive-4.0.1-bin"
export HIVE_CONF_DIR=$HIVE_HOME/conf
export HIVE_AUX_JARS_PATH=$HIVE_HOME/lib/*
export PATH=$PATH:$HIVE_HOME/bin
export SPARK_LOCAL_IP=192.168.1.38
export SPARK_DIST_CLASSPATH="/home/hadoop/spark-3.5.3-bin-
hadoop3/jars/*:/home/hadoop/hadoop-
3.4.0/etc/hadoop:/home/hadoop/hadoop-
3.4.0/share/hadoop/common/lib/*:/home/hadoop/hadoop-
3.4.0/share/hadoop/common/*:/home/hadoop/hadoop-
3.4.0/share/hadoop/hdfs:/home/hadoop/hadoop-
3.4.0/share/hadoop/hdfs/lib/*:/home/hadoop/hadoop-
3.4.0/share/hadoop/hdfs/*:/home/hadoop/hadoop-
3.4.0/share/hadoop/mapreduce/*:/home/hadoop/hadoop-
3.4.0/share/hadoop/yarn:/home/hadoop/hadoop-
3.4.0/share/hadoop/yarn/lib/*:/home/hadoop/hadoop-
3.4.0/share/hadoop/yarn/*:/home/hadoop/apache-hive-4.0.0-alpha-2-
bin/*:/home/hadoop/apache-hive-4.0.0-alpha-2-bin/lib/*"

cd spark-3.5.3-bin-hadoop3/

export SPARK_HOME=`pwd`
export PYTHONPATH=$(ZIPS=("$SPARK_HOME"/python/lib/*.zip); IFS=:;
echo "${ZIPS[*]}"):$PYTHONPATH
export PATH=$SPARK_HOME/bin:$PATH
```

4. Подключим виртуальное окружение

```
cd ../
python3 -m venv venv - создаем ВО
source venv/bin/activate - активируем ВО
```

5. Установим библиотеки связанные с Python

```
pip install -U pip
pip install ipython
pip install onetl[files]
```

6. Подключим оболочку и импортируем модули

```
ipython3
from pyspark.sql import SparkSession
from pyspark.sql import functions as F
from pyspark.sql import functions as F
from onetl.connection import Hive
from onetl.file import FileDFReader
from onetl.file.format import CSV
from onetl.db import DBWriter
```

7. Создадим сессию

```
spark = SparkSession.builder.master("yarn")\
.appName("spark-with-yarn")\
.config("spark.sql.warehouse.dir", "/user/hive/warehouse")\
.config("spark.hive.metastore.uris", "thrift://tmpl-jn:9083")\
.enableHiveSupport().getOrCreate()
```

8. Подключимся к HDFS и считаем данные

```
hdfs = SparkHDFS(host="tean-9-nn", port=9000, spark=spark,
cluster="test")
reader = FileDFReader(connection=hdfs, format=CSV(delimiter=",",
header=True), source_path="/input")
reader.run(["data"])
```

9.