Департамент образования города Москвы

Государственное автономное образовательное учреждение
высшего образования города Москвы
«Московский городской педагогический университет»

Институт цифрового образования
Департамент информатики, управления и технологий

Распределенные системы
Установка и настройка распределенной системы. Простейшие операции и знакомство с функциональностью системы.
Лабораторная работа №1

Выполнила:
Студентка группы АДЭУ-221
Муханова Анна Игоревна
_____
Проверил:
Босенко Тимур Муртазович
_____

Москва
2024

Вариант 9.

1. Создаем директорию и загружаем наш файл (данные биржевых акций «Полюс» PLZL)

```
devops@devopsvm:~$ ls
Desktop                                 plzl                        thinclient_drives
Documents                               Public                      Untitled1.ipynb
Downloads                               snap                        Untitled.ipynb
google-chrome-stable_current_amd64.deb  spark-3.4.3-bin-hadoop3.tgz Videos
Music                                   spark_data                  work_with_data_2024.ipynb
Pictures                                Templates
devops@devopsvm:~$ cd ~/plzl
devops@devopsvm:~/plzl$ ls
PLZL.csv
devops@devopsvm:~/plzl$
```

2. Загружаем данные

```
hadoop@devopsvm: ~/o...   hadoop@devopsvm: ~   devops@devopsvm: ~/plzl   devops@devopsvm: ~/plzl
|25.10.2024|14.594,0|14.850,0|14.978,5|14.412,0|404,06K|-1,49%|
|24.10.2024|14.815,0|14.540,0|14.825,0|14.465,0|255,35K| 2,17%|
|23.10.2024|14.500,0|14.633,5|14.900,0|14.368,0|957,71K|-0,70%|
+----------+--------+--------+--------+--------+-------+------+
only showing top 5 rows

>>> df = spark.read.csv("file:///home/devops/plzl/PLZL.csv",header=True,inferSchema=True)
>>> df.show(5)
+----------+-------+-------+-------+-------+-----+------+
|      Дата|   Цена|  Откр.|  Макс.|   Мин.|Объём|Изм. %|
+----------+-------+-------+-------+-------+-----+------+
|01.12.2019|7.103,5|6.882,0|7.265,0|6.720,0|1,80M| 3,08%|
|01.11.2019|6.891,0|7.470,0|7.621,0|6.666,0|2,48M|-7,74%|
|01.10.2019|7.469,0|7.502,0|7.739,0|7.026,0|1,86M|-1,01%|
|01.09.2019|7.545,0|7.670,0|7.833,0|6.527,0|2,16M|-1,57%|
|01.08.2019|7.665,0|6.433,0|7.665,0|6.374,5|3,36M|18,65%|
+----------+-------+-------+-------+-------+-----+------+
only showing top 5 rows

>>>
```

```
>>> df.printSchema()
root
 |-- Дата: string (nullable = true)
 |-- Цена: string (nullable = true)
 |-- Откр.: string (nullable = true)
 |-- Макс.: string (nullable = true)
 |-- Мин.: string (nullable = true)
 |-- Объём: string (nullable = true)
 |-- Изм. %: string (nullable = true)
```

3. Фильтруем данные

```
>>> df_filtrd = df.filter(df["Дата"]>="01-01-2019")
>>> df_filtrd.show(5)
+----------+-------+-------+-------+-------+-----+------+
|      Дата|   Цена|  Откр.|  Макс.|   Мин.|Объём|Изм. %|
+----------+-------+-------+-------+-------+-----+------+
|01.12.2019|7.103,5|6.882,0|7.265,0|6.720,0|1,80M| 3,08%|
|01.11.2019|6.891,0|7.470,0|7.621,0|6.666,0|2,48M|-7,74%|
|01.10.2019|7.469,0|7.502,0|7.739,0|7.026,0|1,86M|-1,01%|
|01.09.2019|7.545,0|7.670,0|7.833,0|6.527,0|2,16M|-1,57%|
|01.08.2019|7.665,0|6.433,0|7.665,0|6.374,5|3,36M|18,65%|
+----------+-------+-------+-------+-------+-----+------+
only showing top 5 rows
```

Spark 3.4.3 | Jobs | Stages | Storage | Environment | Executors | SQL / DataFrame | PySparkShell application UI

Details for Stage 2 (Attempt 0)

Resource Profile Id: 0
Total Time Across All Tasks: 0.3 s
Locality Level Summary: Process local: 1
Input Size / Records: 1734.0 B / 6
Associated Job Ids: 2

- DAG Visualization
- Show Additional Metrics
- Event Timeline

Summary Metrics for 1 Completed Tasks

Metric	Min	25th percentile	Median	75th percentile	Max
Duration	0.3 s	0.3 s	0.3 s	0.3 s	0.3 s
GC Time	0.0 ms	0.0 ms	0.0 ms	0.0 ms	0.0 ms
Input Size / Records	1.7 KiB / 6	1.7 KiB / 6	1.7 KiB / 6	1.7 KiB / 6	1.7 KiB / 6

- Aggregated Metrics by Executor

Tasks (1)

4. Переходим в Hadoop

```
hadoop@devopsvm:~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [devopsvm]
2024-10-30 01:46:51,700 WARN util.NativeCodeLo
latform... using builtin-java classes where ap
hadoop@devopsvm:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hadoop@devopsvm:~$ jps
29542 NameNode
30633 Jps
29753 DataNode
30313 ResourceManager
30442 NodeManager
29947 SecondaryNameNode
hadoop@devopsvm:~$
```

5. Создаю директории в hdfs

```
hadoop@devopsvm:~$ hadoop fs -mkdir /userind
2024-10-31 19:44:42,711 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your p
latform... using builtin-java classes where applicable
```

```
hadoop@devopsvm:~$ hadoop fs -mkdir /userind/hadoop
2024-10-31 19:45:21,734 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your p
latform... using builtin-java classes where applicable
```

```
hadoop@devopsvm:~$ hadoop fs -mkdir /userind/hadoop/input
2024-10-31 19:46:28,167 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your p
latform... using builtin-java classes where applicable
```

6. Загрузим данные в HDFS

```
hadoop@devopsvm:~$ wget https://raw.githubusercontent.com/cucann/5_semester/refs/heads/main/PLZL.cs
v
--2024-10-31 19:49:58--  https://raw.githubusercontent.com/cucann/5_semester/refs/heads/main/PLZL.c
sv
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.109.133
, 185.199.110.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.133|:443... connect
ed.
HTTP request sent, awaiting response... 200 OK
Length: 922 [text/plain]
Saving to: ‘PLZL.csv’

PLZL.csv                 100%[===================================>]     922  --.-KB/s    in 0.001s

2024-10-31 19:49:59 (1.23 MB/s) - ‘PLZL.csv’ saved [922/922]

hadoop@devopsvm:~$ hadoop fs -mkdir /userind/hadoop/polus_data
2024-10-31 19:51:56,536 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your p
latform... using builtin-java classes where applicable
hadoop@devopsvm:~$
```

```
hadoop@devopsvm:~$ hadoop fs -put PLZL.csv /userind/hadoop/polus_data/
2024-10-31 19:53:26,563 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your p
latform... using builtin-java classes where applicable
hadoop@devopsvm:~$
```

```
scala> val data = spark.read.option("header", "true").csv("file:///home/hadoop/PLZL.csv")
data: org.apache.spark.sql.DataFrame = [Дата: string, Цена: string ... 5 more fields]

scala> data.printSchema()
root
 |-- Дата: string (nullable = true)
 |-- Цена: string (nullable = true)
 |-- Откр.: string (nullable = true)
 |-- Макс.: string (nullable = true)
 |-- Мин.: string (nullable = true)
 |-- Объём: string (nullable = true)
 |-- Изм. %: string (nullable = true)
```

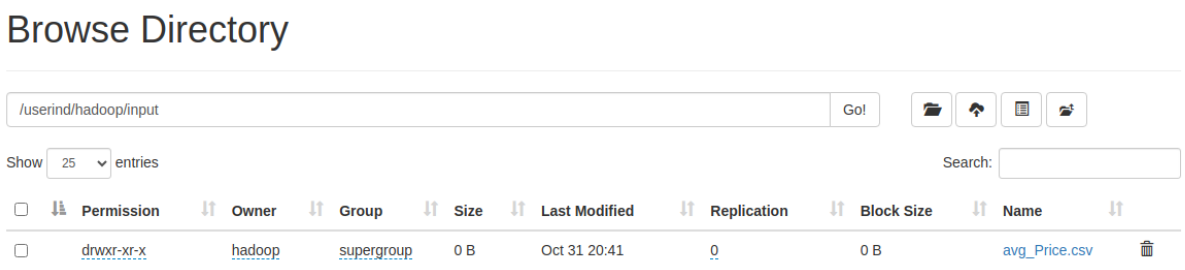
7. Считаю среднюю цену

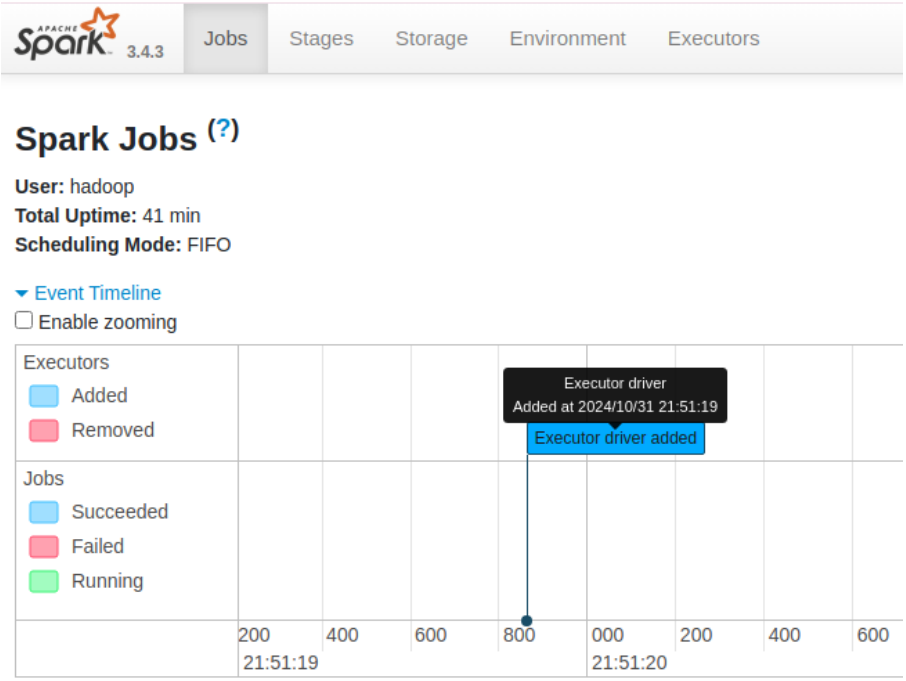
```
scala> val result = data.selectExpr("avg(`Цена`) as avg_Price")
result: org.apache.spark.sql.DataFrame = [avg_Price: double]
```

8. Сохраняем в CSV

```
scala> result.write.option("header", "true").csv("/home/hadoop/output/avg_Price.csv")
```

Проверка записи данных





Nodes в Hadoop

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0

Show 20 entries

Node Labels	Rack	Node State	Node Address	Node HTTP Address	Last health-update	Health-report	Containers	Allocation Tags	Mem Used	Mem Avail	Phys Mem Used %
	/default-rack	RUNNING	devopsvm:34327	devopsvm:8042	Thu Oct 31 22:30:18 +0300 2024		0		0 B	8 GB	72

Showing 1 to 1 of 1 entries