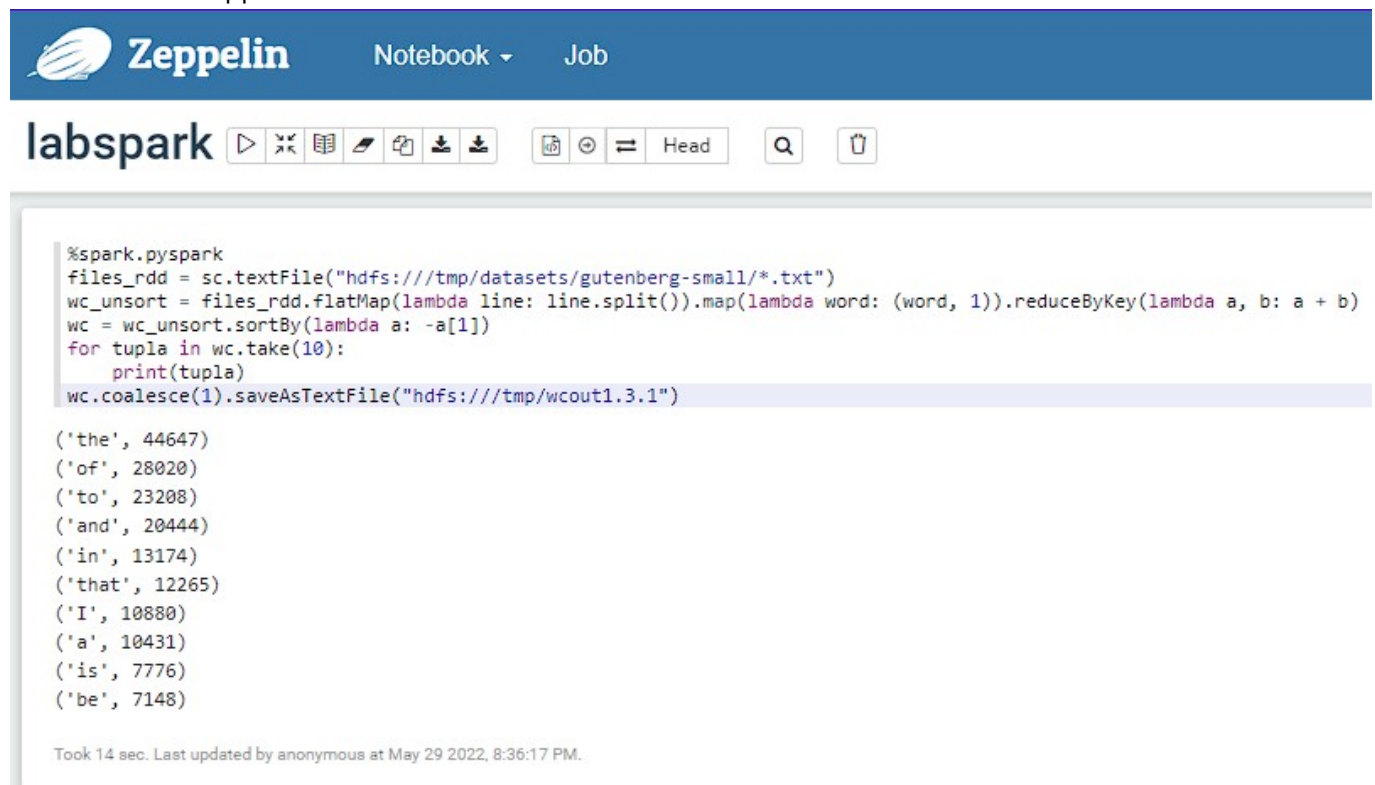


Word count in Zeppelin:



The screenshot shows the Zeppelin Notebook interface. At the top, there's a blue header with the Zeppelin logo, a 'Notebook' dropdown, and a 'Job' button. Below the header, the 'labspark' logo is visible on the left, and a toolbar with various icons (play, stop, refresh, etc.) is on the right. The main area contains a code block with a Scala script for word counting. The script reads text files from HDFS, calculates word counts, sorts them, and prints the top 10 words. The output of the script is displayed below the code, showing the top 10 words and their counts. At the bottom, a status bar indicates the execution took 14 seconds and was last updated by an anonymous user on May 29, 2022, at 8:36:17 PM.

```
%spark.pyspark
files_rdd = sc.textFile("hdfs:///tmp/datasets/gutenberg-small/*.txt")
wc_unsort = files_rdd.flatMap(lambda line: line.split()).map(lambda word: (word, 1)).reduceByKey(lambda a, b: a + b)
wc = wc_unsort.sortBy(lambda a: -a[1])
for tupla in wc.take(10):
    print(tupla)
wc.coalesce(1).saveAsTextFile("hdfs:///tmp/wcout1.3.1")

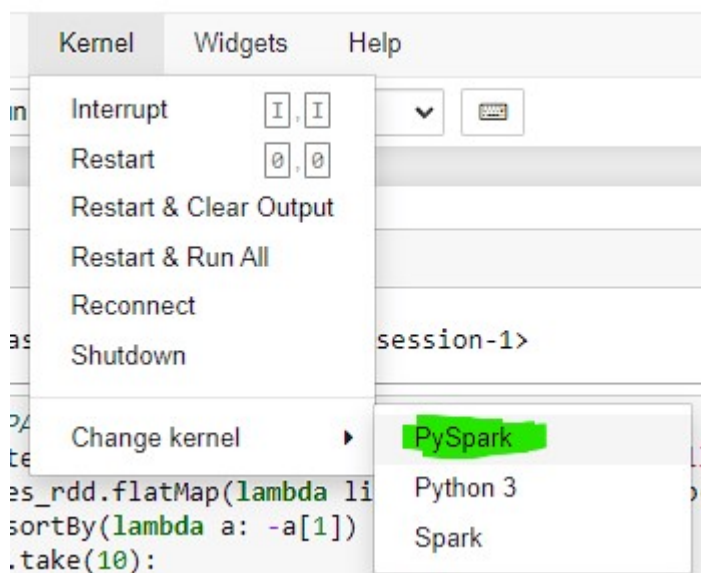
('the', 44647)
('of', 28020)
('to', 23208)
('and', 20444)
('in', 13174)
('that', 12265)
('I', 10880)
('a', 10431)
('is', 7776)
('be', 7148)
```

Took 14 sec. Last updated by anonymous at May 29 2022, 8:36:17 PM.

JupyterHub

Upload [wordcount-spark.ipynb](#).

Change kernel to **PySpark**:



The screenshot shows the Jupyter Notebook interface. At the top, there's a header with 'Kernel', 'Widgets', and 'Help' tabs. Below the header, a dropdown menu is open, showing options for kernel management: 'Interrupt', 'Restart', 'Restart & Clear Output', 'Restart & Run All', 'Reconnect', 'Shutdown', and 'Change kernel'. The 'Change kernel' option is selected, and a sub-menu is displayed below it, showing three available kernels: 'PySpark' (highlighted in green), 'Python 3', and 'Spark'. The background shows a code editor with a snippet of code related to word counting.

```
files_rdd.flatMap(lambda line: line.split()).map(lambda word: (word, 1)).reduceByKey(lambda a, b: a + b)
wc = wc_unsort.sortBy(lambda a: -a[1])
wc.coalesce(1).saveAsTextFile("hdfs:///tmp/wcout1.3.1")
```

Run:

jupyterhub

wordcount-spark (autosaved)

Logout

Control Panel

File

Edit

View

Insert

Cell

Kernel

Widgets

Help

Trusted

PySpark

+

↶

↷

↻

↺

↻

↻

Code

▼

In [2]:

sc

<SparkContext master=yarn appName=livy-session-1>

In [1]:

```
# WORDCOUNT COMPACTO
files_rdd = sc.textFile("hdfs:///tmp/datasets/gutenberg-small/*.txt")
wc_unsort = files_rdd.flatMap(lambda line: line.split()).map(lambda word: (word, 1)).reduceByKey(lambda a, b: a + b)
wc = wc_unsort.sortBy(lambda a: -a[1])
for tupla in wc.take(10):
    print(tupla)
```

Starting Spark application

ID	YARN Application ID	Kind	State	Spark UI	Driver log	Current session?
1	application_1653869584941_0010	pyspark	idle	Link	Link	✓

SparkSession available as 'spark'.

```
('the', 44647)
('of', 28020)
('to', 23208)
('and', 20444)
('in', 13174)
('that', 12265)
('I', 10880)
('a', 10431)
('is', 7776)
('be', 7148)
```

Hive

Create database, create table, store HDI data in table in hive directly.

1 use sflorezsidb;

2

3 CREATE TABLE HDI (id INT, country STRING, hdi FLOAT, lifeex INT, mysch INT, eysch INT, gni INT)

4 ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

5 STORED AS TEXTFILE;

6

7 LOAD DATA INPATH 'hdfs:///tmp/datasets/onu/hdi-data.csv' INTO TABLE HDI;

INFO : Loading data to table default.hdi from hdfs:///ap-172-31-01-74-002.x-internal2.0020/tmp/datasets/onu/hdi-data.csv

INFO : Starting task [Stage-1:STATS] in serial mode

INFO : Completed executing command(queryId=hive_20220530020907_47ab3013-7f4c-439b-9bce-a088bc6ccfb0); Time taken: 2.098 seconds

INFO : OK

INFO : Concurrency mode is disabled, not creating a lock manager

✓ Success.

Query History

Saved Queries

a few seconds ago

✓

LOAD DATA INPATH 'hdfs:///tmp/datasets/onu/hdi-data.csv' INTO TABLE HDI

a few seconds ago

⌘

CREATE TABLE HDI (id INT, country STRING, hdi FLOAT, lifeex INT, mysch INT, eysch INT, gni INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE

2 minutes ago

✓

create database sflorezsidb

3 / 8

Show tables.

2
3 show tables;
4 describe hdi;

INFO : Executing Command(queryId=hive_2022030021224_349acaa7-71ef-41a6-a731-5d904d30a834), show tables
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_2022030021224_349acaa7-71ef-41a6-a731-5d904d30a834); Time taken: 0.306 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager

Query History

Saved Queries

Results (1)

tab_name

1 hdi

Describe HDI table.

1 use sfloresz1db;
2
3 show tables;
4 describe hdi;

INFO : Executing Command(queryId=hive_2022030021250_fddd45ce-0f17-4607-86d8-0c5d84b038d9), describe hdi
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_2022030021250_fddd45ce-0f17-4607-86d8-0c5d84b038d9); Time taken: 0.178 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager

Query History

Saved Queries

Results (7)

col_name

data_type

comment

1 id int
2 country string
3 hdi float
4 lifeex int
5 mysch int
6 eysch int
7 gni int

Select data from said table.

1 use sfloreszldb;
2
3|select * from HDI;|

select * from HDI
INFO : Completed executing command(queryId=hive_20220530021049_50ac87b5-36ad-4151-a9d6-03b69613c8ba); Time taken: 0.0 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager

Query History

Saved Queries

Results (100+)

	hdi.id	hdi.country	hdi.hdi	hdi.lifeex	hdi.mysch	hdi.eyesch	hdi.gni
1	NULL	country	NULL	NULL	NULL	NULL	NULL
2	1	Norway	0.943	81	12	17	47557
3	2	Australia	0.929	81	12	18	34431
4	3	Netherlands	0.91	80	11	16	36402
5	4	United States	0.91	78	12	16	43017
6	5	New Zealand	0.908	80	12	18	23737
7	6	Canada	0.908	81	12	16	35166
8	7	Ireland	0.908	80	11	18	29322
9	8	Liechtenstein	0.905	79	10	14	83717
10	9	Germany	0.905	80	12	15	34854
11	10	Sweden	0.904	81	11	15	35837
12	11	Switzerland	0.903	82	11	15	39924
13	12	Japan	0.901	83	11	15	32295
14	13	Hong Kong China (SAR)	0.898	82	10	15	44805

Select all countries that have a gni greater than 2000 and show their gni.

1 use sfloreszldb;
2
3|select country, gni from hdi where gni > 2000;

select country, gni from hdi where gni > 2000
INFO : Completed executing command(queryId=hive_20220530021459_6f4cadff-a23b-44f1-9ba3-9923b83e91ef); Time taken: 0.001 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager

Query History

Saved Queries

Results (100+)

	country	gni
1	Norway	47557
2	Australia	34431
3	Netherlands	36402
4	United States	43017
5	New Zealand	23737
6	Canada	35166
7	Ireland	29322
8	Liechtenstein	83717
9	Germany	34854
10	Sweden	35837
11	Switzerland	39924
12	Japan	32295
13	Hong Kong China (SAR)	44805

Create new table for JOIN in external s3 storage.

5 / 8

1 use sfloresz1db;
2
3 CREATE EXTERNAL TABLE EXPO (country STRING, expct FLOAT)
4 ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
5 STORED AS TEXTFILE
6 LOCATION 'hdfs:///tmp/datasets/onu/export';
7

INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20220530022809_95db61ca-dadb-4a7c-b3cb-3a9973f6dd4f): select * from expo
INFO : Completed executing command(queryId=hive_20220530022809_95db61ca-dadb-4a7c-b3cb-3a9973f6dd4f); Time taken: 0.001 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager

Query History

Saved Queries

Results (100+)

a few seconds ago

select * from expo

a few seconds ago

CREATE EXTERNAL TABLE EXPO (country STRING, expct FLOAT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION 'hdfs:///tmp/datasets/onu/export'

a few seconds ago

use sfloresz1db

3 minutes ago

CREATE EXTERNAL TABLE EXPO (country STRING, expct FLOAT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION 'hdfs:///tmp/datasets/onu/export'

Join expo and hdi tables.

1 SELECT h.country, gni, expct FROM default.HDI h JOIN sfloresz1db.EXPO e ON (h.country = e.country) WHERE gni > 2000;

INFO : Map 1: 1/1 Map 2: 0(*)/1
INFO : Map 1: 1/1 Map 2: 1/1
INFO : Completed executing command(queryId=hive_20220530024005_7e5d4416-8df1-4c32-adc5-3605b1ed3d23); Time taken: 12.837 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager

Query History

Saved Queries

Results (100+)

h.country

gni

expct

1 Albania 7803 29.77231

2 Algeria 7658 30.830406

3 Andorra 36095 NULL

4 Angola 4874 56.835884

5 Antigua and Barbuda 15521 44.08267

6 Argentina 14527 21.706469

7 Armenia 5188 20.58361

8 Australia 34431 19.780243

9 Austria 35719 53.971355

10 Azerbaijan 8666 55.125755

11 Bahrain 28169 NULL

12 Barbados 17966 47.34396

Word count in Hive

Using alternative 1.


```
1 use sflores1db;
2
3 CREATE EXTERNAL TABLE docs (line STRING)
4 STORED AS TEXTFILE
5 LOCATION 'hdfs:///tmp/datasets/gutenberg-small/';
6
```

select * from docs

INFO : Completed executing command(queryId=hive_20220530024503_1a455f05-3936-458e-9a46-df5e123fcf7a); Time taken: 0.0 seconds

INFO : OK

INFO : Concurrency mode is disabled, not creating a lock manager

Query History

Saved Queries

Results (100+)

a few seconds ago

select * from docs

a minute ago

use sflores1db; CREATE EXTERNAL TABLE docs (line STRING) STORED AS TEXTFILE LOCATION 'hdfs:///tmp/datasets/gutenberg-small/';

Sort by word.

```
1 use sflores1db;
2
3 SELECT word, count(1) AS count FROM (SELECT explode(split(line, ' ')) AS word FROM docs) w
4 GROUP BY word
5 ORDER BY word DESC LIMIT 10;
```

INFO : Map 1: 2/2 Reducer 2: 2/2 Reducer 3: 0/1/1

INFO : Map 1: 2/2 Reducer 2: 2/2 Reducer 3: 1/1

INFO : Completed executing command(queryId=hive_20220530024754_428d31b3-ada5-48be-80df-c252dc727ec9); Time taken: 15.675 seconds

INFO : OK

INFO : Concurrency mode is disabled, not creating a lock manager

Query History

Saved Queries

Results (10)

	word	count
1	Æschines,	1
2	zigzag	1
3	zest	1
4	zenith	1
5	zealously	1
6	zealous,	1
7	zealous	5
8	zeal,	3
9	zeal	8
10	youthful	2

Sort by word descending.

1 use sfloresz1db;
2
3 SELECT word, count(1) AS count FROM (SELECT explode(split(line, ' ')) AS word FROM docs) w
4 GROUP BY word
5 ORDER BY count DESC LIMIT 10;

INFO : Map 1: 2/2 Reducer 2: 1(1)/2 Reducer 3: 0(1)/1
INFO : Map 1: 2/2 Reducer 2: 2/2 Reducer 3: 1/1
INFO : Completed executing command(queryId=hive_20220530024958_3c42bc97-3962-431e-b56c-26a4a5ad9291); Time taken: 10.796 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager

application_1653869584941_0014

Query History

Saved Queries

Results (10)

	word	count
1	the	44647
2	of	28020
3		27298
4	to	23208
5	and	20444
6	in	13174
7	that	12265
8	I	10880
9	a	10431
10	is	7776

Insert the last result into another table.

1 use sfloresz1db;
2
3 create table count_dict (word string, quantity int);
4
5 INSERT INTO COUNT_DICT SELECT word, count(1) AS count FROM (SELECT explode(split(line, ' ')) AS word FROM docs) w
6 GROUP BY word
7 ORDER BY count DESC;
8
9 select * from count_dict;

INFO : Completed executing command(queryId=hive_20220530025754_bb55a297-1316-4c12-a898-774a095f95bf); Time taken: 0.001 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager

Query History

Saved Queries

Results (100+)

	count_dict.word	count_dict.quantity
1	the	44647
2	of	28020
3		27298
4	to	23208
5	and	20444
6	in	13174
7	that	12265
8	I	10880
9	a	10431
10	is	7776
11	be	7148
12	it	6899
13	as	6473