

Spark advanced report.

The algorithm that I apply to this problem is the following:

- 1) Read new data from kafka.
- 2) Partition it by date, hour, hash tag and user id.
- 3) Find intersecting partitions – partitions that are already present on the HDFS such that they are also present in the new kafka data – and read them from the HDFS to later be able to merge them with the new data.
- 4) Remove the partitions that we identified in the previous step from the HDFS.
- 5) Merge the offloaded HDFS data with the new data and write it back to the HDFS.

Here are the stages:

Completed Stages (6)

Stage id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
6	save at SparkExtensions.scala:190	2019/01/16 02:19:12	7 s	200/200		18.0 B	747.0 B	
5	save at SparkExtensions.scala:190	2019/01/16 02:19:08	4 s	203/203	9.1 KB			747.0 B
3	count at DataFrameExtensions.scala:46	2019/01/16 02:19:08	87 ms	1/1			177.0 B	
2	count at DataFrameExtensions.scala:46	2019/01/16 02:19:06	2 s	3/3	18.0 B			177.0 B
1	collect at SparkExtensions.scala:173	2019/01/16 02:18:59	6 s	200/200			369.0 B	
0	cache at TopicDataIngestor.scala:46	2019/01/16 02:18:56	3 s	2/2				369.0 B

Next I provide screenshots of each stage's DAG.

Details for Stage 6 (Attempt 0)

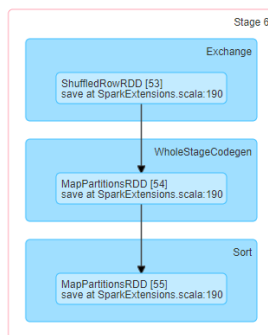
Total Time Across All Tasks: 3 s

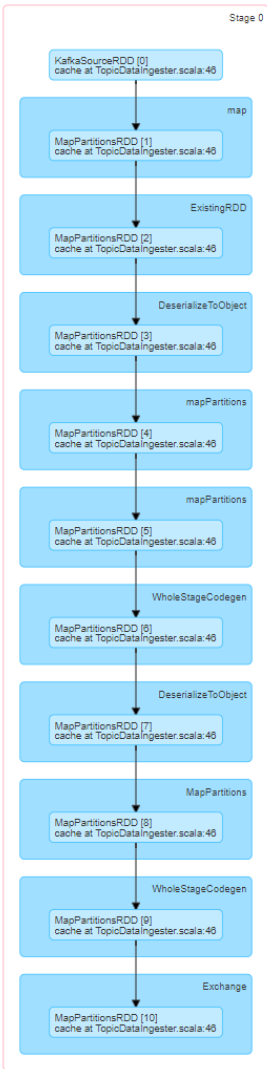
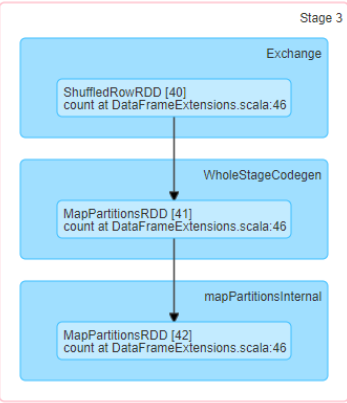
Locality Level Summary: Node local: 3; Process local: 197

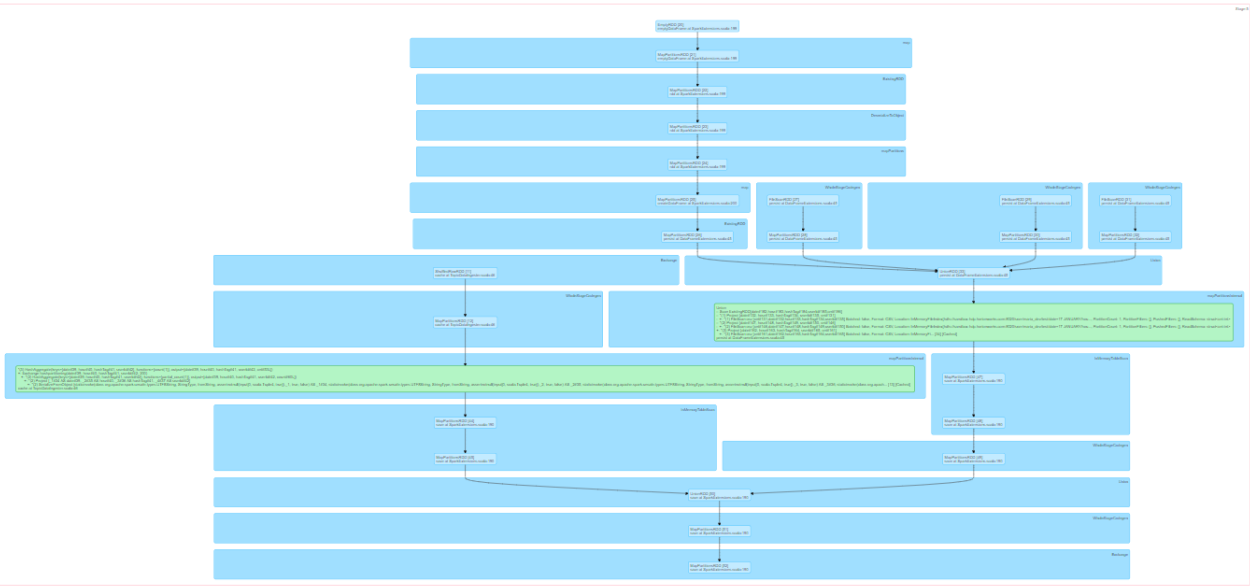
Output: 18.0 B / 3

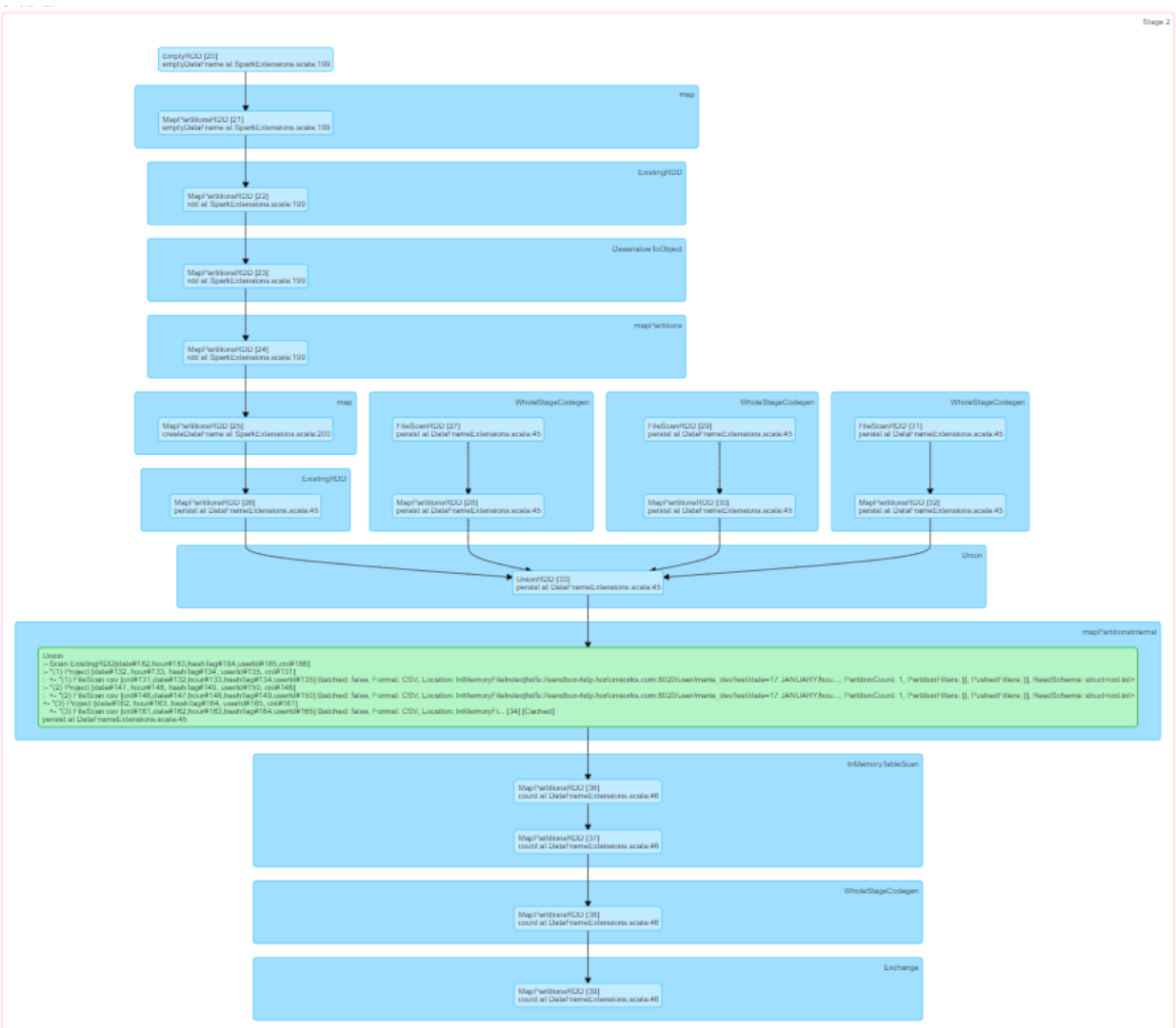
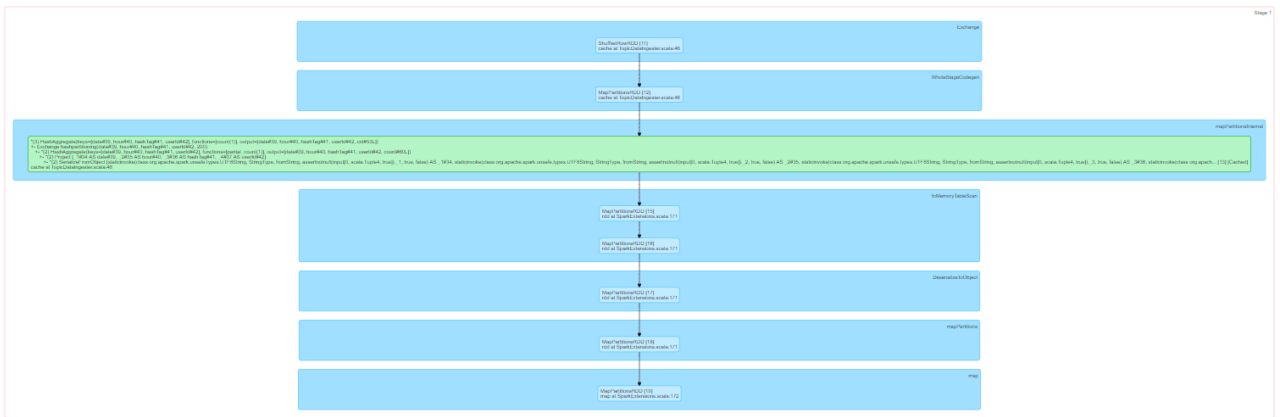
Shuffle Read: 747.0 B / 6

▼ DAG Visualization

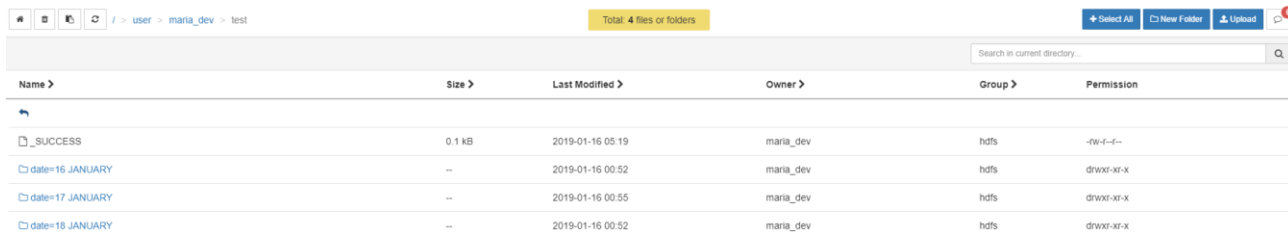








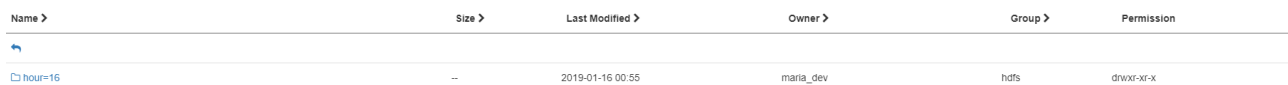
After the program is finished running this is what we have in our HDFS data folder:



The image shows a web-based HDFS file browser interface. At the top, there's a breadcrumb path: / > user > maria_dev > test. A yellow box indicates 'Total: 4 files or folders'. On the right, there are buttons for 'Select All', 'New Folder', and 'Upload'. Below this is a search bar labeled 'Search in current directory...'. The main area is a table with columns: Name, Size, Last Modified, Owner, Group, and Permission. The table lists four items: a folder named '_SUCCESS' (0.1 kB, 2019-01-16 05:19, maria_dev, hdfs, -rw-r--r--), and three folders named 'date=16 JANUARY', 'date=17 JANUARY', and 'date=18 JANUARY' (all 0.1 kB, 2019-01-16 00:52, maria_dev, hdfs, drwxr-xr-x).

Name	Size	Last Modified	Owner	Group	Permission
_SUCCESS	0.1 kB	2019-01-16 05:19	maria_dev	hdfs	-rw-r--r--
date=16 JANUARY	--	2019-01-16 00:52	maria_dev	hdfs	drwxr-xr-x
date=17 JANUARY	--	2019-01-16 00:55	maria_dev	hdfs	drwxr-xr-x
date=18 JANUARY	--	2019-01-16 00:52	maria_dev	hdfs	drwxr-xr-x

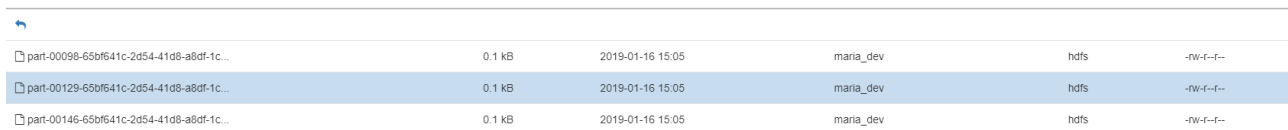
This is what “date=17 JANUARY” folder has:



The image shows a web-based HDFS file browser interface displaying the contents of the 'date=17 JANUARY' folder. The table has columns: Name, Size, Last Modified, Owner, Group, and Permission. It lists a single folder named 'hour=16' (0.1 kB, 2019-01-16 00:55, maria_dev, hdfs, drwxr-xr-x).

Name	Size	Last Modified	Owner	Group	Permission
hour=16	--	2019-01-16 00:55	maria_dev	hdfs	drwxr-xr-x

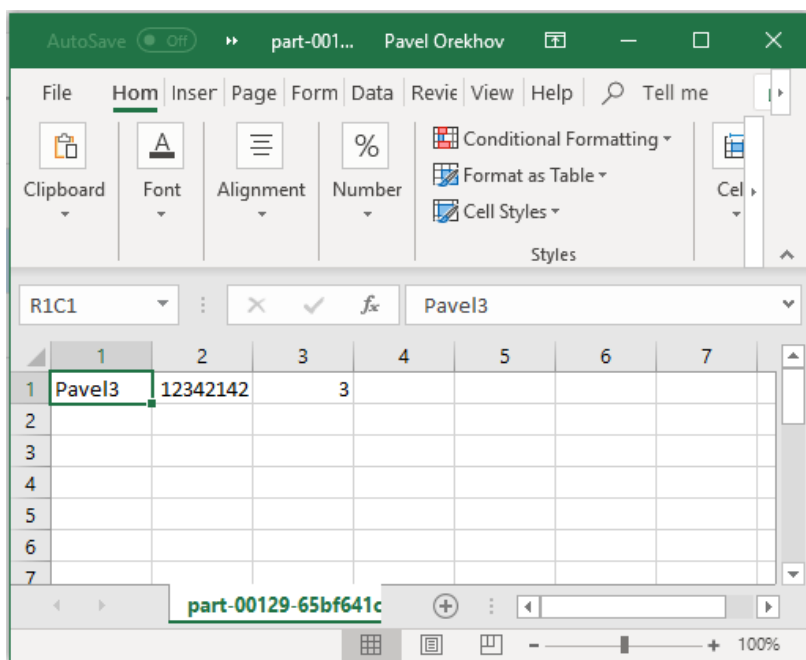
And finally here are some of the result files:



The image shows a web-based HDFS file browser interface displaying a list of result files. The table has columns: Name, Size, Last Modified, Owner, Group, and Permission. It lists three files, all 0.1 kB, 2019-01-16 15:05, maria_dev, hdfs, -rw-r--r--: 'part-00098-65bf641c-2d54-41d8-a8df-1c...', 'part-00129-65bf641c-2d54-41d8-a8df-1c...' (highlighted), and 'part-00146-65bf641c-2d54-41d8-a8df-1c...'.

Name	Size	Last Modified	Owner	Group	Permission
part-00098-65bf641c-2d54-41d8-a8df-1c...	0.1 kB	2019-01-16 15:05	maria_dev	hdfs	-rw-r--r--
part-00129-65bf641c-2d54-41d8-a8df-1c...	0.1 kB	2019-01-16 15:05	maria_dev	hdfs	-rw-r--r--
part-00146-65bf641c-2d54-41d8-a8df-1c...	0.1 kB	2019-01-16 15:05	maria_dev	hdfs	-rw-r--r--

If we open it in excel, we'll see the hashtag, user id and count data (by which we did not partition):



The image is a screenshot of the Microsoft Excel application. The title bar shows 'part-001...' and 'Pavel Orekhov'. The ribbon includes 'File', 'Home', 'Insert', 'Page Layout', 'Formulas', 'Data', 'Review', 'View', and 'Help'. The 'Home' ribbon is active, showing 'Clipboard', 'Font', 'Alignment', 'Number', and 'Styles' groups. The formula bar shows 'Pavel3'. The worksheet grid shows columns 1 through 7 and rows 1 through 7. Cell A1 (row 1, column 1) contains 'Pavel3' and is highlighted with a green border. Cell B1 (row 1, column 2) contains '12342142'. Cell C1 (row 1, column 3) contains '3'. The status bar at the bottom shows 'part-00129-65bf641c' and '100%' zoom.

	1	2	3	4	5	6	7
1	Pavel3	12342142	3				
2							
3							
4							
5							
6							
7							