DATA PREPROCESSING USING SPARK AND DATA VISUALIZATION USING SPARK DATABASE

Problem#2

Sr284605@dal.ca
Srikrishnan Sengottai Kasi

Task 1

1)Algorithm:

Method using RDD in Spark: [Considering the entire Tweet file]

1) BEGIN
2) Load the contents of extracted tweet text file into a Resilient Distributed Datasets(RDD) in Apache Spark.
3) Apply flatMap() over the contents of RDD. The function receives the content of file string and reads it line by line and creates a Sequence Array of words and it is stored in another RDD.
4) Then flattened RDD is filtered with the provided key words in the problem statement.
5) Create a Mapper with word as key and initialize value for all keys as 1. So, the output of (K,v) is created with K as Words and V as 1.
6) Now the output of Mapper is passed to the reducer. The reducer makes use of add function. The add function is performed for same keys and the resultant is provided as (K, sum(V)).
7) END.

Method Using DataFrame in Spark: [Considered the Tweet text content for better accuracy]

1) BEGIN
2) Read the content from the twitter files into a json content and store it in the form of a DataFrame.
3) Since the tweets are structured, retrieve only the "text" filed in each tweet to do the MapReducer algorithm and store that in a DataFrame. Punctuations are removed for better accuracy.
4) Since the tweets are array, explode function is used to flatten the tweet texts and store it in another Dataframe.
5) Tweet text sentence is tokenized into words array by utilizing the white character between them and storing it back in a dataframe.
6) The word array type is further exploded to keep each word as arow in a separate dataframe and completing the Mapper module.
7) Then SQL functions of dataframe is used and the words are grouped, and count is stored as a separate column in another dataframe.
8) Once the reducer is implemented the Result is filtered with the keywords given in the problem statement.
9) END.

2) Frequencies and Observations:

	Highest Frequency	Lowest Frequency
Method 1	"emergency" - 137	"flu" - 58
Method 2	"emergency" - 111	"flu" - 55

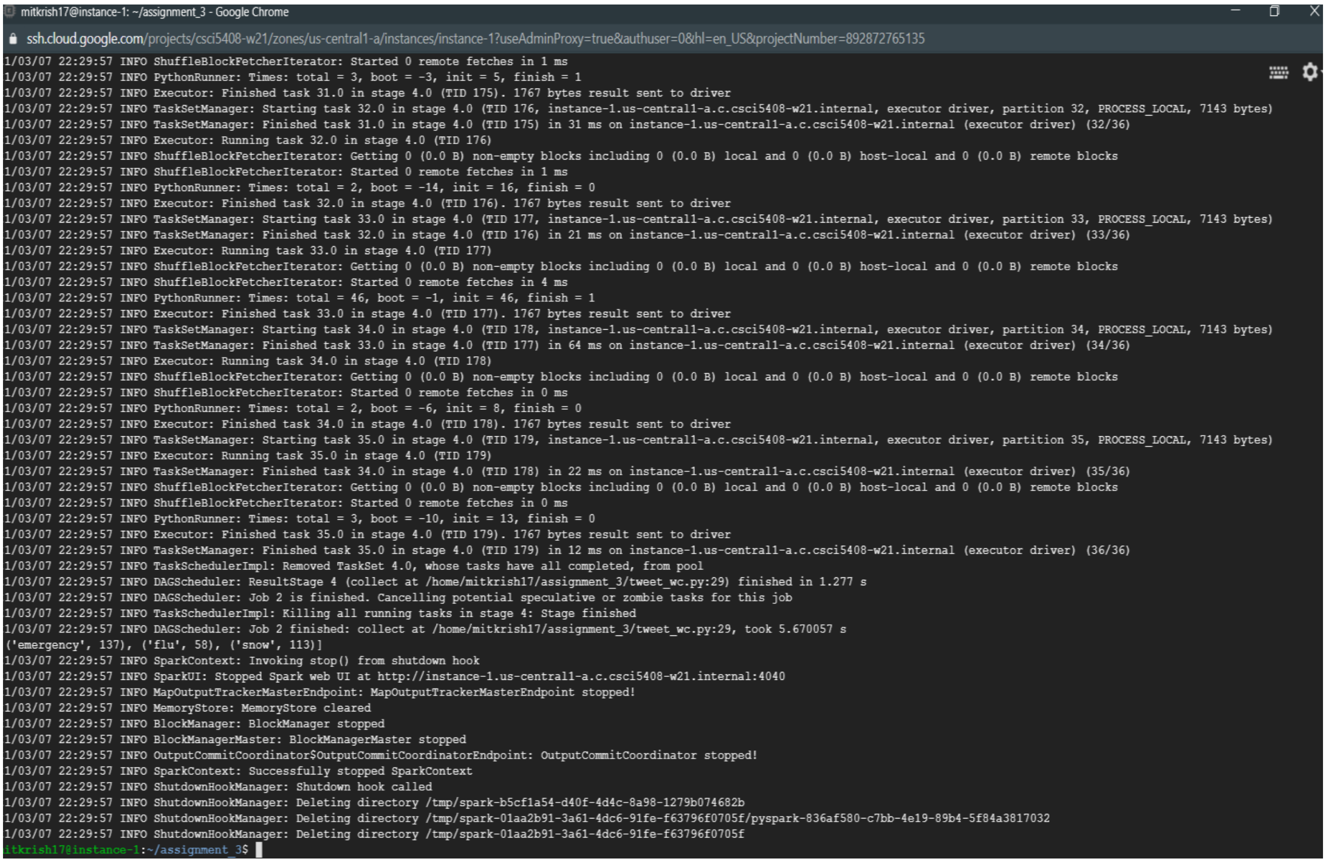
The difference in Method 1 and Method 2 is because of the input data. First used the entire text file but method 2 used only the tweet data text to perform the MapReducer approach.

When the word count is performed over entire word list, it is better to remove the stop words and punctuations, emoji's rigorously to get more accuracy and context to the MapReducer performance.

Also, while using RDD, parallelize can be used to run at various clusters. This quickly computes the required operation than single instance.

Thus, RDD can be used for processing MapReducer over the tweet.txt file as such. While dataframe can be used to process the MapReducer on a structured json file.

Output using RDD to get word count over the complete tweet file.

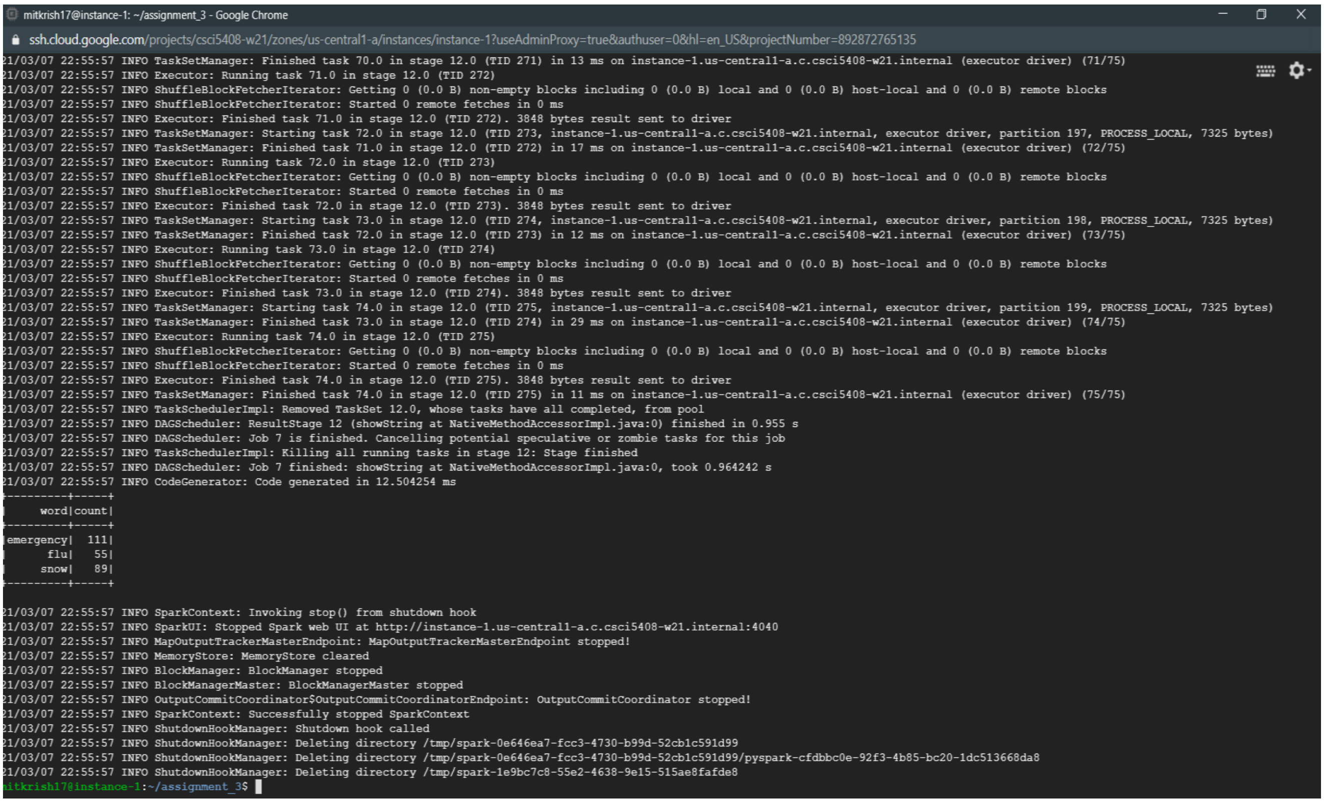


mitkrish17@instance-1: ~/assignment_3 - Google Chrome

ssh.cloud.google.com/projects/csci5408-w21/zones/us-central1-a/instances/instance-1?useAdminProxy=true&authuser=0&hl=en_US&projectNumber=892872765135

```
1/03/07 22:29:57 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
1/03/07 22:29:57 INFO PythonRunner: Times: total = 3, boot = -3, init = 5, finish = 1
1/03/07 22:29:57 INFO Executor: Finished task 31.0 in stage 4.0 (TID 175). 1767 bytes result sent to driver
1/03/07 22:29:57 INFO TaskSetManager: Starting task 32.0 in stage 4.0 (TID 176, instance-1.us-central1-a.c.csci5408-w21.internal, executor driver, partition 32, PROCESS_LOCAL, 7143 bytes)
1/03/07 22:29:57 INFO TaskSetManager: Finished task 31.0 in stage 4.0 (TID 175) in 31 ms on instance-1.us-central1-a.c.csci5408-w21.internal (executor driver) (32/36)
1/03/07 22:29:57 INFO Executor: Running task 32.0 in stage 4.0 (TID 176)
1/03/07 22:29:57 INFO ShuffleBlockFetcherIterator: Getting 0 (0.0 B) non-empty blocks including 0 (0.0 B) local and 0 (0.0 B) host-local and 0 (0.0 B) remote blocks
1/03/07 22:29:57 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
1/03/07 22:29:57 INFO PythonRunner: Times: total = 2, boot = -14, init = 16, finish = 0
1/03/07 22:29:57 INFO Executor: Finished task 32.0 in stage 4.0 (TID 176). 1767 bytes result sent to driver
1/03/07 22:29:57 INFO TaskSetManager: Starting task 33.0 in stage 4.0 (TID 177, instance-1.us-central1-a.c.csci5408-w21.internal, executor driver, partition 33, PROCESS_LOCAL, 7143 bytes)
1/03/07 22:29:57 INFO TaskSetManager: Finished task 32.0 in stage 4.0 (TID 176) in 21 ms on instance-1.us-central1-a.c.csci5408-w21.internal (executor driver) (33/36)
1/03/07 22:29:57 INFO Executor: Running task 33.0 in stage 4.0 (TID 177)
1/03/07 22:29:57 INFO ShuffleBlockFetcherIterator: Getting 0 (0.0 B) non-empty blocks including 0 (0.0 B) local and 0 (0.0 B) host-local and 0 (0.0 B) remote blocks
1/03/07 22:29:57 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 4 ms
1/03/07 22:29:57 INFO PythonRunner: Times: total = 46, boot = -1, init = 46, finish = 1
1/03/07 22:29:57 INFO Executor: Finished task 33.0 in stage 4.0 (TID 177). 1767 bytes result sent to driver
1/03/07 22:29:57 INFO TaskSetManager: Starting task 34.0 in stage 4.0 (TID 178, instance-1.us-central1-a.c.csci5408-w21.internal, executor driver, partition 34, PROCESS_LOCAL, 7143 bytes)
1/03/07 22:29:57 INFO TaskSetManager: Finished task 33.0 in stage 4.0 (TID 177) in 64 ms on instance-1.us-central1-a.c.csci5408-w21.internal (executor driver) (34/36)
1/03/07 22:29:57 INFO Executor: Running task 34.0 in stage 4.0 (TID 178)
1/03/07 22:29:57 INFO ShuffleBlockFetcherIterator: Getting 0 (0.0 B) non-empty blocks including 0 (0.0 B) local and 0 (0.0 B) host-local and 0 (0.0 B) remote blocks
1/03/07 22:29:57 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
1/03/07 22:29:57 INFO PythonRunner: Times: total = 2, boot = -6, init = 8, finish = 0
1/03/07 22:29:57 INFO Executor: Finished task 34.0 in stage 4.0 (TID 178). 1767 bytes result sent to driver
1/03/07 22:29:57 INFO TaskSetManager: Starting task 35.0 in stage 4.0 (TID 179, instance-1.us-central1-a.c.csci5408-w21.internal, executor driver, partition 35, PROCESS_LOCAL, 7143 bytes)
1/03/07 22:29:57 INFO Executor: Running task 35.0 in stage 4.0 (TID 179)
1/03/07 22:29:57 INFO TaskSetManager: Finished task 34.0 in stage 4.0 (TID 178) in 22 ms on instance-1.us-central1-a.c.csci5408-w21.internal (executor driver) (35/36)
1/03/07 22:29:57 INFO ShuffleBlockFetcherIterator: Getting 0 (0.0 B) non-empty blocks including 0 (0.0 B) local and 0 (0.0 B) host-local and 0 (0.0 B) remote blocks
1/03/07 22:29:57 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
1/03/07 22:29:57 INFO PythonRunner: Times: total = 3, boot = -10, init = 13, finish = 0
1/03/07 22:29:57 INFO Executor: Finished task 35.0 in stage 4.0 (TID 179). 1767 bytes result sent to driver
1/03/07 22:29:57 INFO TaskSetManager: Finished task 35.0 in stage 4.0 (TID 179) in 12 ms on instance-1.us-central1-a.c.csci5408-w21.internal (executor driver) (36/36)
1/03/07 22:29:57 INFO TaskSchedulerImpl: Removed TaskSet 4.0, whose tasks have all completed, from pool
1/03/07 22:29:57 INFO DAGScheduler: ResultStage 4 (collect at /home/mitkrish17/assignment_3/tweet_wc.py:29) finished in 1.277 s
1/03/07 22:29:57 INFO DAGScheduler: Job 2 is finished. Cancelling potential speculative or zombie tasks for this job
1/03/07 22:29:57 INFO TaskSchedulerImpl: Killing all running tasks in stage 4: Stage finished
1/03/07 22:29:57 INFO DAGScheduler: Job 2 finished: collect at /home/mitkrish17/assignment_3/tweet_wc.py:29, took 5.670057 s
('emergency', 137), ('flu', 58), ('snow', 113)]
1/03/07 22:29:57 INFO SparkContext: Invoking stop() from shutdown hook
1/03/07 22:29:57 INFO SparkUI: Stopped Spark web UI at http://instance-1.us-central1-a.c.csci5408-w21.internal:4040
1/03/07 22:29:57 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
1/03/07 22:29:57 INFO MemoryStore: MemoryStore cleared
1/03/07 22:29:57 INFO BlockManager: BlockManager stopped
1/03/07 22:29:57 INFO BlockManagerMaster: BlockManagerMaster stopped
1/03/07 22:29:57 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
1/03/07 22:29:57 INFO SparkContext: Successfully stopped SparkContext
1/03/07 22:29:57 INFO ShutdownHookManager: Shutdown hook called
1/03/07 22:29:57 INFO ShutdownHookManager: Deleting directory /tmp/spark-b5cf1a54-d40f-4d4c-8a98-1279b074682b
1/03/07 22:29:57 INFO ShutdownHookManager: Deleting directory /tmp/spark-01aa2b91-3a61-4dc6-91fe-f63796f0705f/pyspark-836af580-c7bb-4e19-89b4-5f84a3817032
1/03/07 22:29:57 INFO ShutdownHookManager: Deleting directory /tmp/spark-01aa2b91-3a61-4dc6-91fe-f63796f0705f
itkrish17@instance-1:~/assignment_3$ 
```

Output after cleaning and using DataFrame focusing on the Tweets texts content.



mitkrish17@instance-1: ~/assignment_3 - Google Chrome

ssh.cloud.google.com/projects/csci5408-w21/zones/us-central1-a/instances/instance-1?useAdminProxy=true&authuser=0&hl=en_US&projectNumber=892872765135

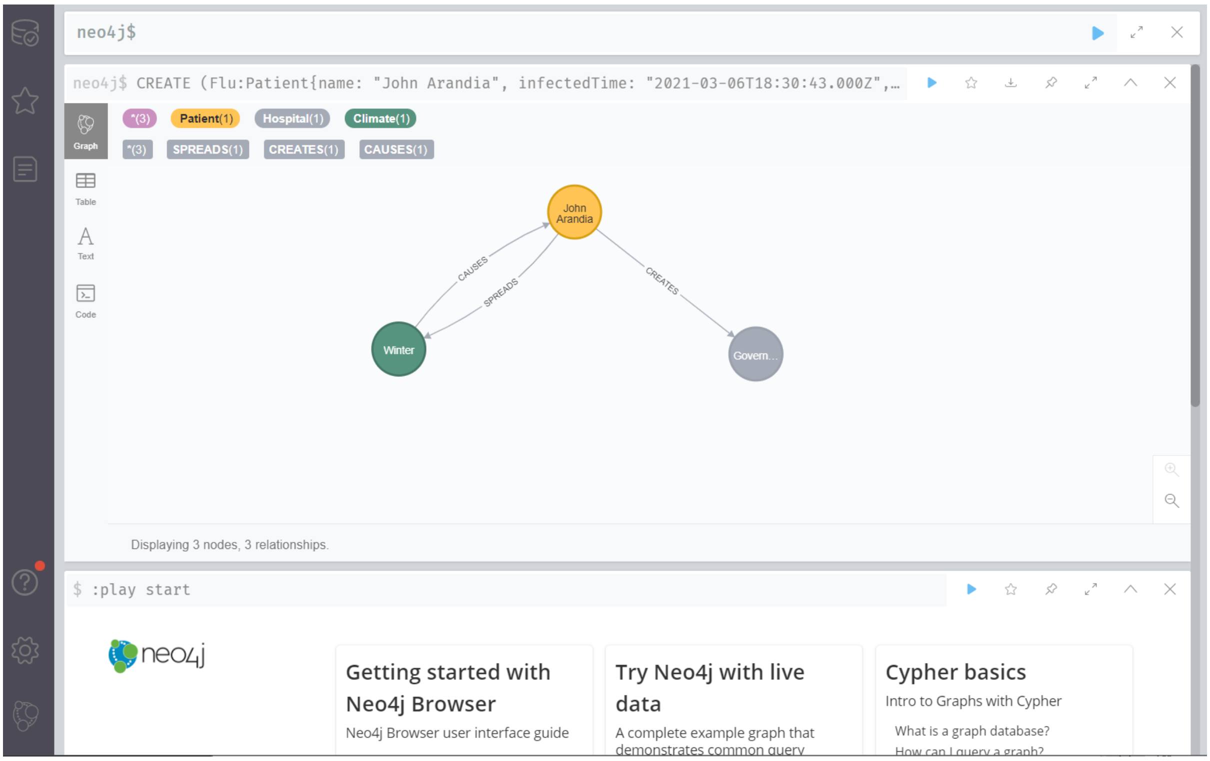
```
21/03/07 22:55:57 INFO TaskSetManager: Finished task 70.0 in stage 12.0 (TID 271) in 13 ms on instance-1.us-central1-a.c.csci5408-w21.internal (executor driver) (71/75)
21/03/07 22:55:57 INFO Executor: Running task 71.0 in stage 12.0 (TID 272)
21/03/07 22:55:57 INFO ShuffleBlockFetcherIterator: Getting 0 (0.0 B) non-empty blocks including 0 (0.0 B) local and 0 (0.0 B) host-local and 0 (0.0 B) remote blocks
21/03/07 22:55:57 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
21/03/07 22:55:57 INFO Executor: Finished task 71.0 in stage 12.0 (TID 272). 3848 bytes result sent to driver
21/03/07 22:55:57 INFO TaskSetManager: Starting task 72.0 in stage 12.0 (TID 273, instance-1.us-central1-a.c.csci5408-w21.internal, executor driver, partition 197, PROCESS_LOCAL, 7325 bytes)
21/03/07 22:55:57 INFO TaskSetManager: Finished task 71.0 in stage 12.0 (TID 272) in 17 ms on instance-1.us-central1-a.c.csci5408-w21.internal (executor driver) (72/75)
21/03/07 22:55:57 INFO Executor: Running task 72.0 in stage 12.0 (TID 273)
21/03/07 22:55:57 INFO ShuffleBlockFetcherIterator: Getting 0 (0.0 B) non-empty blocks including 0 (0.0 B) local and 0 (0.0 B) host-local and 0 (0.0 B) remote blocks
21/03/07 22:55:57 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
21/03/07 22:55:57 INFO Executor: Finished task 72.0 in stage 12.0 (TID 273). 3848 bytes result sent to driver
21/03/07 22:55:57 INFO TaskSetManager: Starting task 73.0 in stage 12.0 (TID 274, instance-1.us-central1-a.c.csci5408-w21.internal, executor driver, partition 198, PROCESS_LOCAL, 7325 bytes)
21/03/07 22:55:57 INFO TaskSetManager: Finished task 72.0 in stage 12.0 (TID 273) in 12 ms on instance-1.us-central1-a.c.csci5408-w21.internal (executor driver) (73/75)
21/03/07 22:55:57 INFO Executor: Running task 73.0 in stage 12.0 (TID 274)
21/03/07 22:55:57 INFO ShuffleBlockFetcherIterator: Getting 0 (0.0 B) non-empty blocks including 0 (0.0 B) local and 0 (0.0 B) host-local and 0 (0.0 B) remote blocks
21/03/07 22:55:57 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
21/03/07 22:55:57 INFO Executor: Finished task 73.0 in stage 12.0 (TID 274). 3848 bytes result sent to driver
21/03/07 22:55:57 INFO TaskSetManager: Starting task 74.0 in stage 12.0 (TID 275, instance-1.us-central1-a.c.csci5408-w21.internal, executor driver, partition 199, PROCESS_LOCAL, 7325 bytes)
21/03/07 22:55:57 INFO TaskSetManager: Finished task 73.0 in stage 12.0 (TID 274) in 29 ms on instance-1.us-central1-a.c.csci5408-w21.internal (executor driver) (74/75)
21/03/07 22:55:57 INFO Executor: Running task 74.0 in stage 12.0 (TID 275)
21/03/07 22:55:57 INFO ShuffleBlockFetcherIterator: Getting 0 (0.0 B) non-empty blocks including 0 (0.0 B) local and 0 (0.0 B) host-local and 0 (0.0 B) remote blocks
21/03/07 22:55:57 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
21/03/07 22:55:57 INFO Executor: Finished task 74.0 in stage 12.0 (TID 275). 3848 bytes result sent to driver
21/03/07 22:55:57 INFO TaskSetManager: Finished task 74.0 in stage 12.0 (TID 275) in 11 ms on instance-1.us-central1-a.c.csci5408-w21.internal (executor driver) (75/75)
21/03/07 22:55:57 INFO TaskSchedulerImpl: Removed TaskSet 12.0, whose tasks have all completed, from pool
21/03/07 22:55:57 INFO DAGScheduler: ResultStage 12 (showString at NativeMethodAccessorImpl.java:0) finished in 0.955 s
21/03/07 22:55:57 INFO DAGScheduler: Job 7 is finished. Cancelling potential speculative or zombie tasks for this job
21/03/07 22:55:57 INFO TaskSchedulerImpl: Killing all running tasks in stage 12: Stage finished
21/03/07 22:55:57 INFO DAGScheduler: Job 7 finished: showString at NativeMethodAccessorImpl.java:0, took 0.964242 s
21/03/07 22:55:57 INFO CodeGenerator: Code generated in 12.504254 ms
+---------+-----+
|     word|count|
+---------+-----+
|emergency|  111|
|      flu|   55|
|     snow|   89|
+---------+-----+

21/03/07 22:55:57 INFO SparkContext: Invoking stop() from shutdown hook
21/03/07 22:55:57 INFO SparkUI: Stopped Spark web UI at http://instance-1.us-central1-a.c.csci5408-w21.internal:4040
21/03/07 22:55:57 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
21/03/07 22:55:57 INFO MemoryStore: MemoryStore cleared
21/03/07 22:55:57 INFO BlockManager: BlockManager stopped
21/03/07 22:55:57 INFO BlockManagerMaster: BlockManagerMaster stopped
21/03/07 22:55:57 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
21/03/07 22:55:57 INFO SparkContext: Successfully stopped SparkContext
21/03/07 22:55:57 INFO ShutdownHookManager: Shutdown hook called
21/03/07 22:55:57 INFO ShutdownHookManager: Deleting directory /tmp/spark-0e646ea7-fcc3-4730-b99d-52cb1c591d99
21/03/07 22:55:57 INFO ShutdownHookManager: Deleting directory /tmp/spark-0e646ea7-fcc3-4730-b99d-52cb1c591d99/pyspark-cfdbbc0e-92f3-4b85-bc20-1dc513668da8
21/03/07 22:55:57 INFO ShutdownHookManager: Deleting directory /tmp/spark-1e9bc7c8-55e2-4638-9e15-515ae8fafde8
mitkrish17@instance-1:~/assignment_3$ 
```

Task 2

Cypher Query Language:

```
1) CREATE (Flu:Patient{name: "John Arandia", infectedTime: "2021-03-06T18:30:43.000Z", condition:"risk", ambulanceNeeded:"yes", isCovid:"yes"})

CREATE (Emergency:Hospital{name: "Government Facility", calledTime: "2021-03-05T20:30:43.000Z", type:"health", department:"hospital"})

CREATE (Snow:Climate{name:"Winter", temperature: "-5", date: "2021-03-05", type:"freezing"})

CREATE (Snow)-[r1:CAUSES]->(Flu)
CREATE (Flu)-[r2:CREATES]->(Emergency)
CREATE (Flu)-[r3:SPREADS]-> (Snow)

Return Flu, Emergency, Snow
```

Output of above Cypher Query Language



```
neo4j$
```

```
neo4j$ CREATE (Flu:Patient{name: "John Arandia", infectedTime: "2021-03-06T18:30:43.000Z",…
```

Displaying 3 nodes, 3 relationships.

```
$ :play start
```

Getting started with Neo4j Browser

Neo4j Browser user interface guide

Try Neo4j with live data

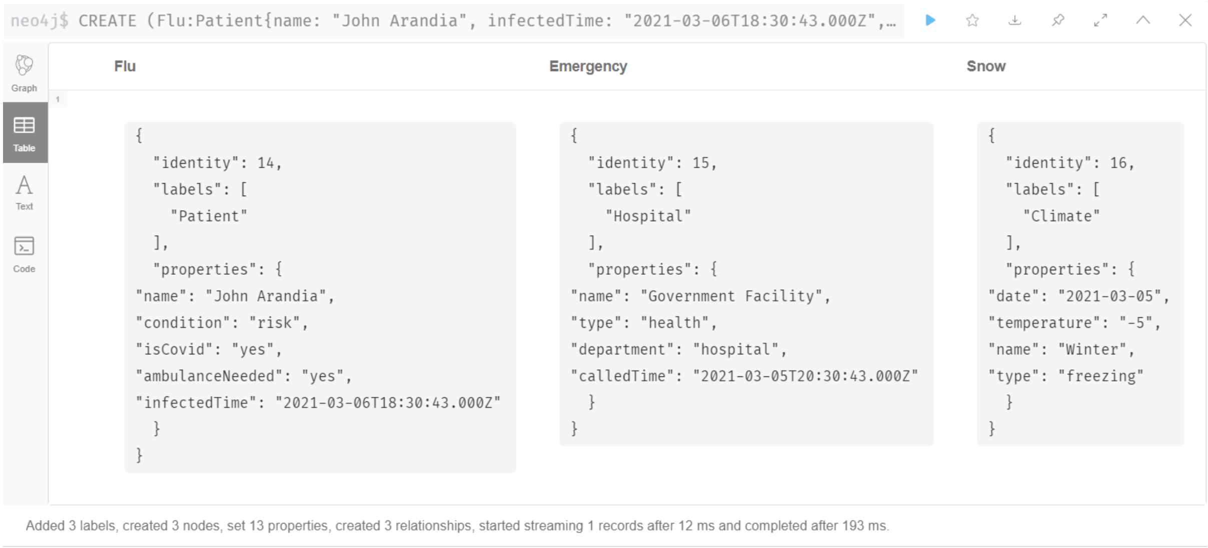
A complete example graph that demonstrates common query

Cypher basics

Intro to Graphs with Cypher

What is a graph database?

How can I query a graph?



```
neo4j$ CREATE (Flu:Patient{name: "John Arandia", infectedTime: "2021-03-06T18:30:43.000Z",…
```

Flu

```
{
  "identity": 14,
  "labels": [
    "Patient"
  ],
  "properties": {
"name": "John Arandia",
"condition": "risk",
"isCovid": "yes",
"ambulanceNeeded": "yes",
"infectedTime": "2021-03-06T18:30:43.000Z"
  }
}
```

Emergency

```
{
  "identity": 15,
  "labels": [
    "Hospital"
  ],
  "properties": {
"name": "Government Facility",
"type": "health",
"department": "hospital",
"calledTime": "2021-03-05T20:30:43.000Z"
  }
}
```

Snow

```
{
  "identity": 16,
  "labels": [
    "Climate"
  ],
  "properties": {
"date": "2021-03-05",
"temperature": "-5",
"name": "Winter",
"type": "freezing"
  }
}
```

Added 3 labels, created 3 nodes, set 13 properties, created 3 relationships, started streaming 1 records after 12 ms and completed after 193 ms.

References

1) https://spark.apache.org/docs/latest/rdd-programming-guide.html#parallelized-collections
2) http://spark.apache.org/examples.html
3) https://data-flair.training/blogs/apache-spark-rdd-vs-dataframe-vs-dataset/
4) https://spark.apache.org/docs/latest/sql-ref-functions-builtin.html
5) https://www.tutorialspoint.com/neo4j/index.htm
6) https://neo4j.com/developer/get-started/