PRINCIPLES OF BIG DATA

MANAGEMENT

PHASE-1 PROJECT



BY

Naga Srividya Varanasi-16269844

Rishitha Chowdary Bobba-16274445

Naga Bindu Madamanchi-16273070

1.Objective

The main goal of the project is to collect the real time twitter data in Json format and to analyze this big data using Apache Hadoop and Apache Spark. The twitter data we collected here is about Bitcoin. Then we extracted the hashtags and URL’s from the data and performed word count operation on the extracted hashtags and URL’s using Hadoop and Spark.

We have used Hadoop in order to perform the Map Reduce framework. Map Reduce exclusively works on key value pair. It works on large data sets.

We have used Spark here as it mainly works on Hadoop. It does faster computation than Hadoop and we can write applications in Java, Scala, Python and R language. Spark comes in rescue when we have to perform complex analytics on vast data.

2.System Requirements

We need few languages and software that has to be installed in the system before collecting the data.

Languages: Java, Python, Sql, Scala.

Software requirements: Apache Hadoop (version 2.8.1)

Apache Spark (version 2.3.1)

JDK 8

Scala (version 2.12.6)

3.Collecting the tweets from twitter

1.First we have to create the twitter developer account using <https://apps.twitter.com/>

2.Then we collect the API keys as well as the Access token keys :

twitterAccessToken = "1039581348267216896-X356h5IuiIWuiHNIR4sgMbIoe6Czq4"

twitterAccessTokenSecret = "wlUWfSuJnLGlczTaKN1HBlDr6EMrBHIniYSteegTG7hKB"

myConsumerKey = "Y3ndR0ynAmxZhe39U7bFORvNM"

myConsumerSecret = "gIKPFvZKJk0EXMyoFSn3TZOzaptkktcG8LUGAuhrOxOwDRSZZV"

3.Then we have written a code in python that is used to fetch the code in JSON format.

4.The data has all the details about when the file is created, id’s , text,URL’s, and the hashtags.

5.From the data we extract the hashtags and URL’s and store them in a file.

4. Wordcount example on the extracted hashtags and URL’s is done using Apache Hadoop and Apache Spark.

We have written a code in python to extract the hashtags and URL’s from the collected data.

**APACHE SPARK**

2018-09-27 19:02:33 INFO Executor:54 - Finished task 0.0 in stage 1.0 (TID 1). 95206 bytes result sent to driver

2018-09-27 19:02:33 INFO TaskSetManager:54 - Finished task 0.0 in stage 1.0 (TID 1) in 203 ms on localhost (executor driver) (1/1)

2018-09-27 19:02:33 INFO TaskSchedulerImpl:54 - Removed TaskSet 1.0, whose tasks have all completed, from pool

2018-09-27 19:02:33 INFO DAGScheduler:54 - ResultStage 1 (collect at JavaWordCount.java:53) finished in 0.219 s

2018-09-27 19:02:33 INFO DAGScheduler:54 - Job 0 finished: collect at JavaWordCount.java:53, took 1.941214 s

bitcoinfiyati: 1

占い: 3

RaysNetwork: 1

Rothchilds: 1

взаимнаяподписка: 3

BlockChain: 62

ToolTales: 1

avcilarescort: 1

三遊亭楽天: 1

inversion: 7

mainnet: 1

augmentedreality: 1

bearmarket: 1

BtcMarkets: 1

Feathercoin: 2

itunes: 2

merge: 1

StartupAsylum: 1

Bitsend: 8

kiskin\_house: 1

IIoT: 3

liquideos: 1

sample wordcount output of hashtags in spark.

2018-09-27 19:03:34 INFO TaskSetManager:54 - Finished task 0.0 in stage 1.0 (TID 1) in 284 ms on localhost (executor driver) (1/1)

2018-09-27 19:03:34 INFO TaskSchedulerImpl:54 - Removed TaskSet 1.0, whose tasks have all completed, from pool

2018-09-27 19:03:34 INFO DAGScheduler:54 - ResultStage 1 (collect at JavaWordCount.java:53) finished in 0.297 s

2018-09-27 19:03:34 INFO DAGScheduler:54 - Job 0 finished: collect at JavaWordCount.java:53, took 1.444884 s

https://t.co/HeZbnArgPa: 2

https://t.co/n8sSvvPzPy: 1

https://t.co/sQjXf1XUyX: 1

https://t.co/CWtHcdAGCe: 1

https://t.co/A1NUvkIS7G: 1

https://t.co/dmiA34fGiH: 2

https://t.co/F4pKxTvShe: 1

https://t.co/i4FLs4QSi4: 1

https://t.co/eEszGLYAGo: 1

https://t.co/z4WV1LAtDs: 1

https://t.co/LfWe2ZeR9P: 1

https://t.co/mhVpZTumZn: 1

https://t.co/TI7hGpvrJ6: 1

sample wordcount output of URL’s in spark

**APACHE HADOOP**

First we have to create a input file for the MapReduce program as:

$ bin/hdfs dfs –mkdir /pbphase1

Now folder is created in HDFS and we have to move it to the HDFS from local using:

$ bin/hdfs dfs –copyFromLocal ht\_output.txt /pbphase1/

To view the file in HDFS use:

$ bin/hdfs dfs –cat /pbphase1/ht\_output.txt

The MapReduce was run in Hadoop using the below command:

hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.8.1.jar wordcount /pbphase1/ht\_output.txt /pbphase1/out

To view the output file :

hdfs dfs -ls /pbphase1/out/

the output is viewed in console by using :

$ hdfs dfs -cat /yourname/out/part-r-00000

cryptocurrencies

Bitcoin

Ethereum

XRP

xrp

Seguridad

DASH

BTC

USD

dash

bitcoin

blockchain

cryptonews

cryptolife

bitcoin

HourlyCryptoStatus

BTC

Bitcoin

ETH

Ethereum

BTC

Sample wordcount output of hashtags in Hadoop

https://t.co/0Dm3Yp8MLL

https://t.co/csfRKKVXVv

https://t.co/IQBx3eJrDC

https://t.co/rtkaDblX7b

https://t.co/Z3NoAlyII7

https://t.co/dVi7ysvA2Y

https://t.co/l8VdwbwHAP

https://t.co/eAYSVkZ1bZ

https://t.co/Zx67Lrlkky

https://t.co/uX5ZwMwKIn

https://t.co/jZ87y4xna3

https://t.co/Ndc96jM4kP

https://t.co/ArupJMQmvB

https://t.co/NpKTM5Q8Ir

https://t.co/qc4rNdPhky

https://t.co/l1QrMEj8og

https://t.co/cOVWubuVDH

https://t.co/lgGcbFCgg0

https://t.co/LL16drQrwj

https://t.co/czJUNUOz3l

sample wordcount output of URL’s in Hadoop