Twitter Big Data Analysis Using Spark

Principles of Big Data Management (CS5540)

Sri Harsha Chennavajjala, Teja Garidepally, Raj Kiran Reddy Munnangi

Introduction

Big Data:

Big data is the buzzing word in the present software industry. Huge amounts of data is being generated daily from various sources.

Apache Spark:

Apache Spark is an open source cluster computing framework originally developed in the AMPLab at University of California, Berkeley but was later donated to the Apache Software Foundation where it remains today.

Spark SQL is a component on top of Spark Core that introduces a new data abstraction called DataFrames, which provides support for structured and semi-structured data.

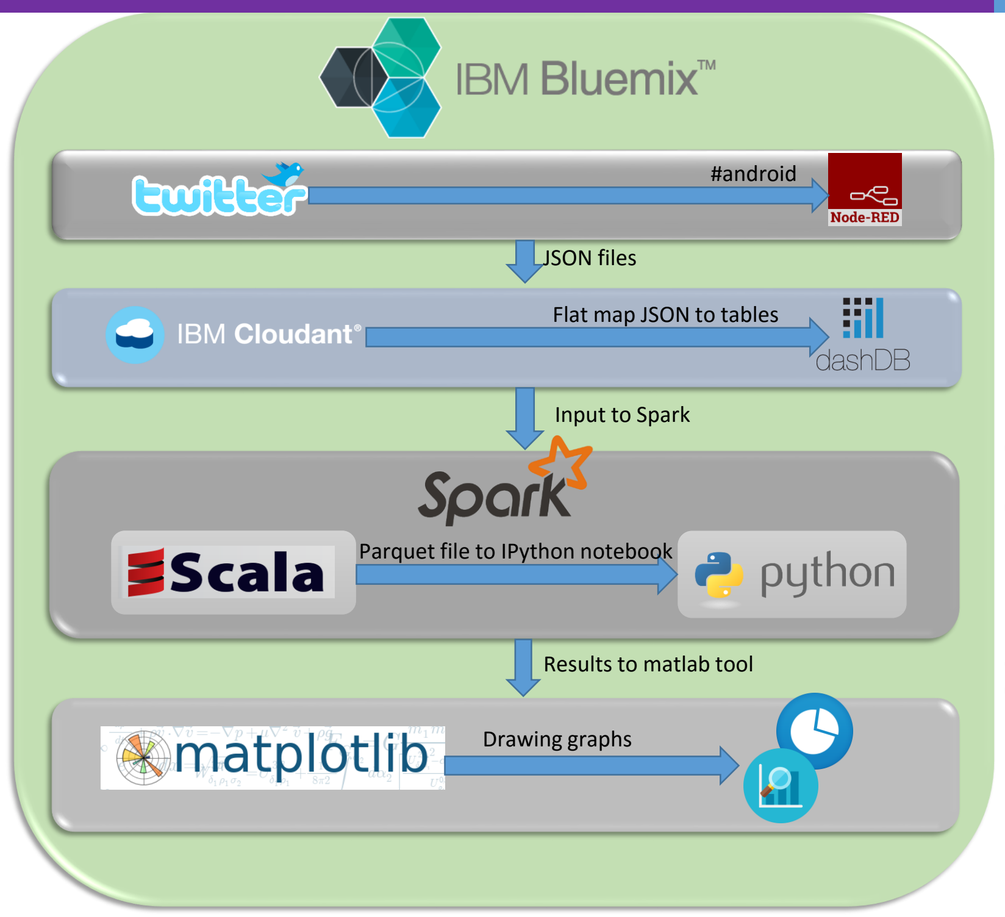
Introduction to IBM Bluemix:

IBM Bluemix is a cloud platform as a service (PaaS) developed by IBM. It supports several programming languages and services as well as integrated DevOps to build, run, deploy and manage applications on the cloud.

Bluemix provides the following features:

- A range of services that enable you to build and extend web and mobile apps fast.
- Processing power for you to deliver application changes continuously.
- Fit-for-purpose programming models and services.
- Manageability of services and apps.
- Optimized and elastic workloads.

Architecture



Services we used in Bluemix:

➢ **IBM Cloudant:**

IBM Cloudant is a fully managed JSON document DBaaS that's optimized for data availability, durability, and mobility…perfect for fast-growing mobile & web apps.

➢ **dashDB:**

dashDB offers massive scalability and performance through its MPP architecture, and is compatible with a wide range of business intelligence toolsets and analytics

➢ **Node-RED:**

Node-RED provides a browser-based UI for creating flows of events and deploying them to its light-weight runtime. With built in node.js, it can be run at the edge of the network or in the cloud.

➢ **Object Storage:**

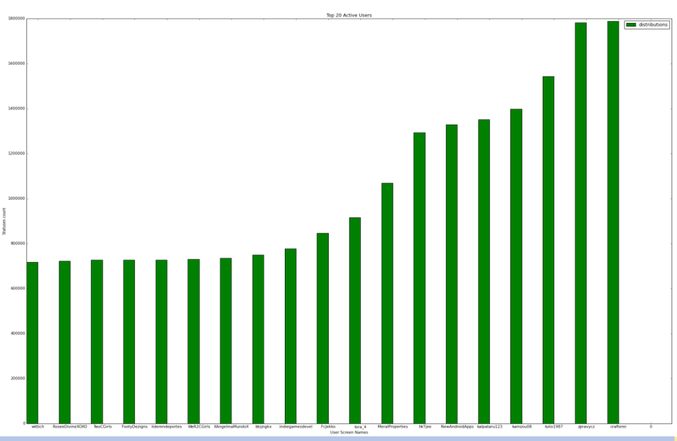
IBM Object Storage for Bluemix provides us with access to a fully provisioned Swift Object Storage account to manage our data.

Query 1:

Description: To find top active users in dataset

➢ **SELECT** TWEET_USER_SCREEN_NAME**, MAX(**TWEET_USER_STATUSES_COUNT**) AS** TWEET_USER_STATUSES_COUNT **FROM** tweetdata **group by** TWEET_USER_SCREEN_NAME **order by** TWEET_USER_STATUSES_COUNT **DESC LIMIT** 20

Visualization:

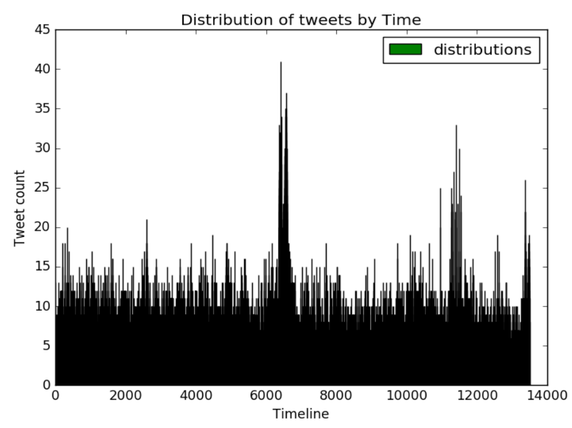


Query 2:

Description: To find tweets per second

➢ val tweetCreatedDt = tweets.filter(_.nonEmpty).map(x => (extractTweetDate(x), 1)) val tweetCreatedDtCnt = tweetCreatedDt.reduceByKey((a, b) => a + b) tweetCreatedDtCnt.repartition(1).saveAsTextFile("D:/UMKC/Docs/Subjects/PBDM/PB_Project/TweetsPerTime")

Visualization:

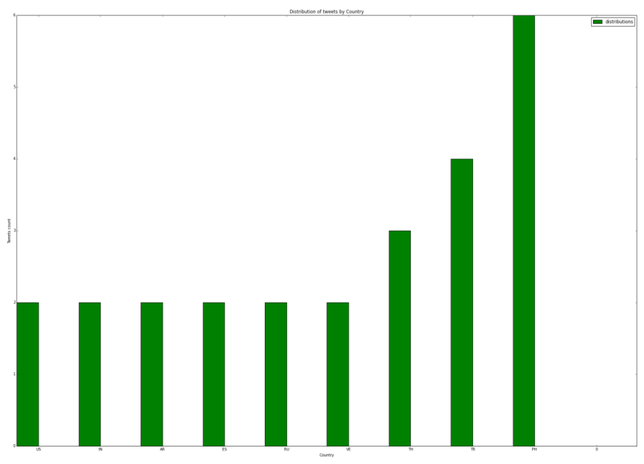


Query 3:

Description: To find tweets per country

➢ **SELECT** TWEET_PLACE_COUNTRY_CODE**, COUNT(**1**) AS** TWEETS_PER_COUNTRY **FROM** tweetdata **group by** TWEET_PLACE_COUNTRY_CODE **order by** TWEETS_PER_COUNTRY **DESC LIMIT** 10

Visualization:

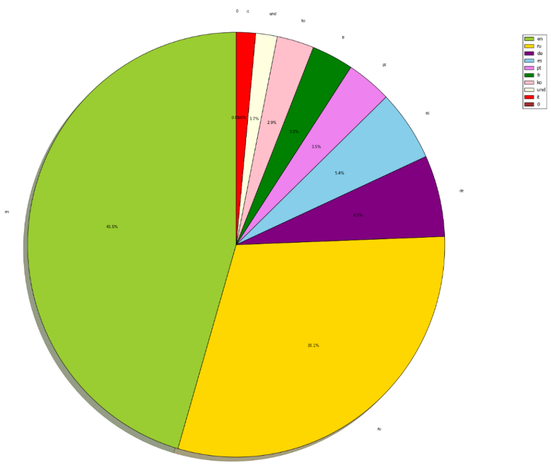


Query 4:

Description: To find number of tweets per language

➢ **SELECT** TWEET_LANG**, count(**1**) as** totTweets **from** tweetdata **group by** TWEET_LANG **order by** totTweets

Visualization:

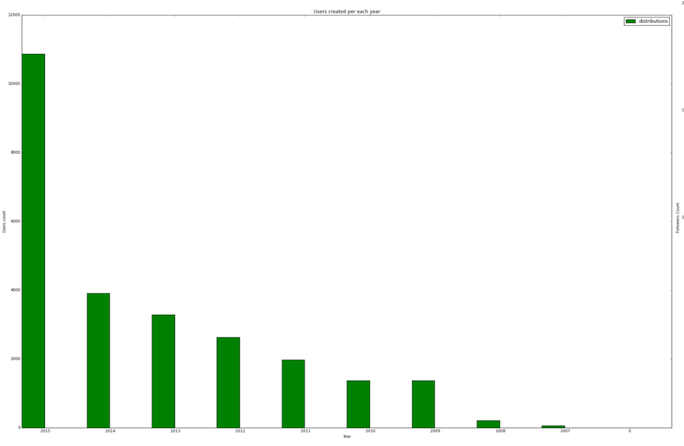


Query 5:

Description: To find user accounts created per year

➢ **SELECT** t1.YR **as** YR**, count(**1**) as** CNT **FROM (select DISTINCT** TWEET_USER_ID**, substring(**TWEET_USER_CREATED_AT,26**) AS** YR **FROM** tweetdata **where** TWEET_USER_CREATED_AT **IS NOT NULL)** t1 **group by** YR **order by** YR

Visualization:

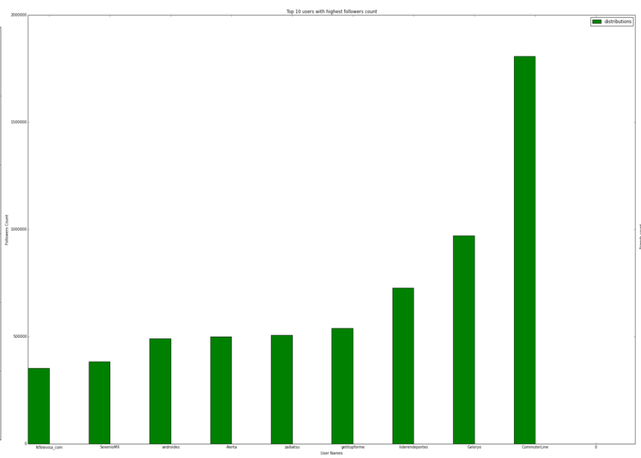


Query 6:

Description: To find Top Users with highest Followers Count

➢ **SELECT** TWEET_USER_SCREEN_NAME**, MAX(**TWEET_USER_FOLLOWERS_COUNT**) AS** TWEET_USER_FOLLOWERS_COUNT **FROM** tweetdata **group by** TWEET_USER_SCREEN_NAME **order by** TWEET_USER_FOLLOWERS_COUNT **DESC LIMIT** 10

Visualization:

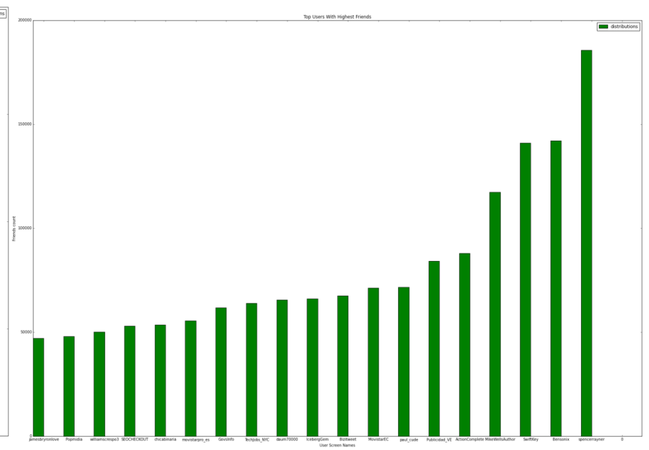


Query 7:

Description: To find Top Users with highest Friends Count

SELECT MAX(TWEET_USER_FRIENDS_COUNT) AS FRIENDS_COUNT, TWEET_USER_SCREEN_NAME FROM tweetdata GROUP BY TWEET_USER_SCREEN_NAME ORDER BY FRIENDS_COUNT DESC LIMIT 20

Visualization:



Conclusion:

➢ Social media provides valuable datasets, but the real challenge is in collecting and analyzing the live streaming data. In this project we've analyzed and visualized twitter data on a #android keyword.

➢ In future work, we would like to perform domain specific live streaming analysis and try to capture valuable insights from data.

Acknowledgements:

➢ We would like to thank Dr.Rao , Anas Katib and Venu Kolla for their extended support in successful execution of this project.