LAB2: DATA AGGREGATION, BIG DATA ANALYSIS AND VISUALIZATION: CSE- 587

SUBMITTED BY:

ABHISHEK MUNI
UB ID: 50291410

SHWETASREE CHOWDHURY
UB ID: 50296995

TOPIC:

NBA

Data aggregation from Twitter, New York Times and Common Crawl and their analysis in the following sequence-

1. Data aggregation from more than one source using the APIs (Application programming interface) exposed by data sources.
2. Applying classical big data analytic method of MapReduce to the unstructured data collected.
3. Store the data collected on WORM infrastructure Hadoop using S3.
4. Building a visualization data product

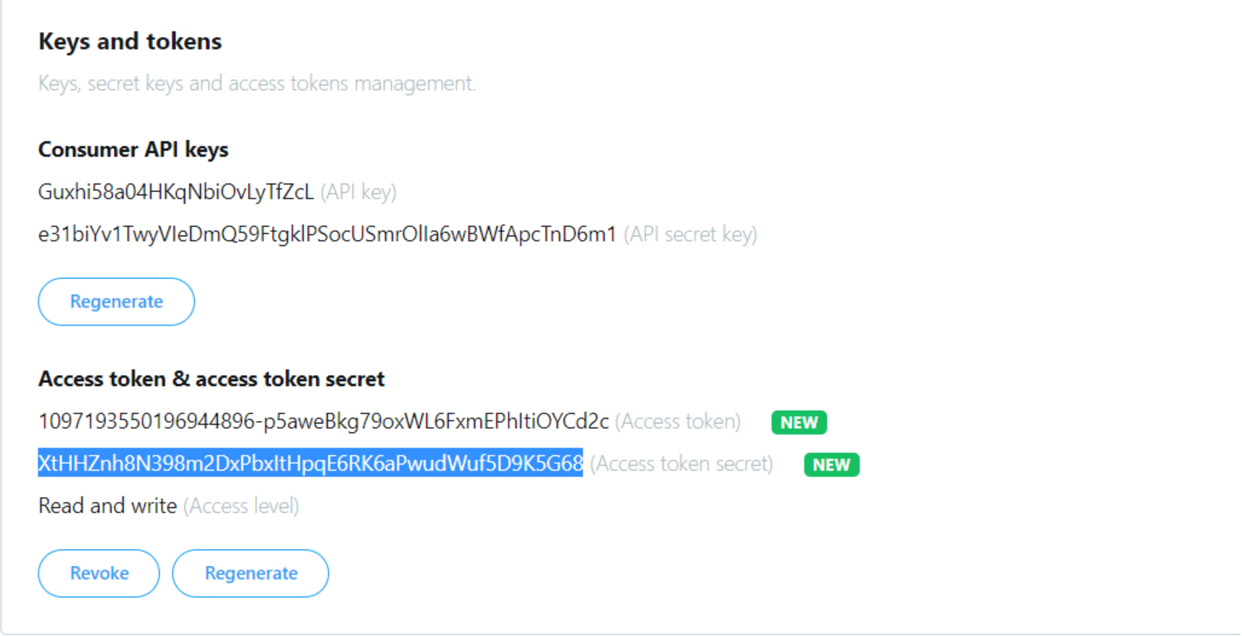
IMPLEMENTATION:

1. Data aggregation:

a. Data aggregation from tweets:

For collecting data from twitter, API search was used-

- A twitter developer account was created and an APP was created within it from which the consumer_ key,consumer_token, acess_key and access token were taken-



Keys and tokens

Keys, secret keys and access tokens management.

Consumer API keys

Guxhi58a04HKqNbiOvLyTfZcL (API key)

e31biYv1TwyVleDmQ59FtgklPSocUSmrOlla6wBWfApcTnD6m1 (API secret key)

Access token & access token secret

1097193550196944896-p5aweBkg79oxWL6FxmEPhltiOYCd2c (Access token) NEW

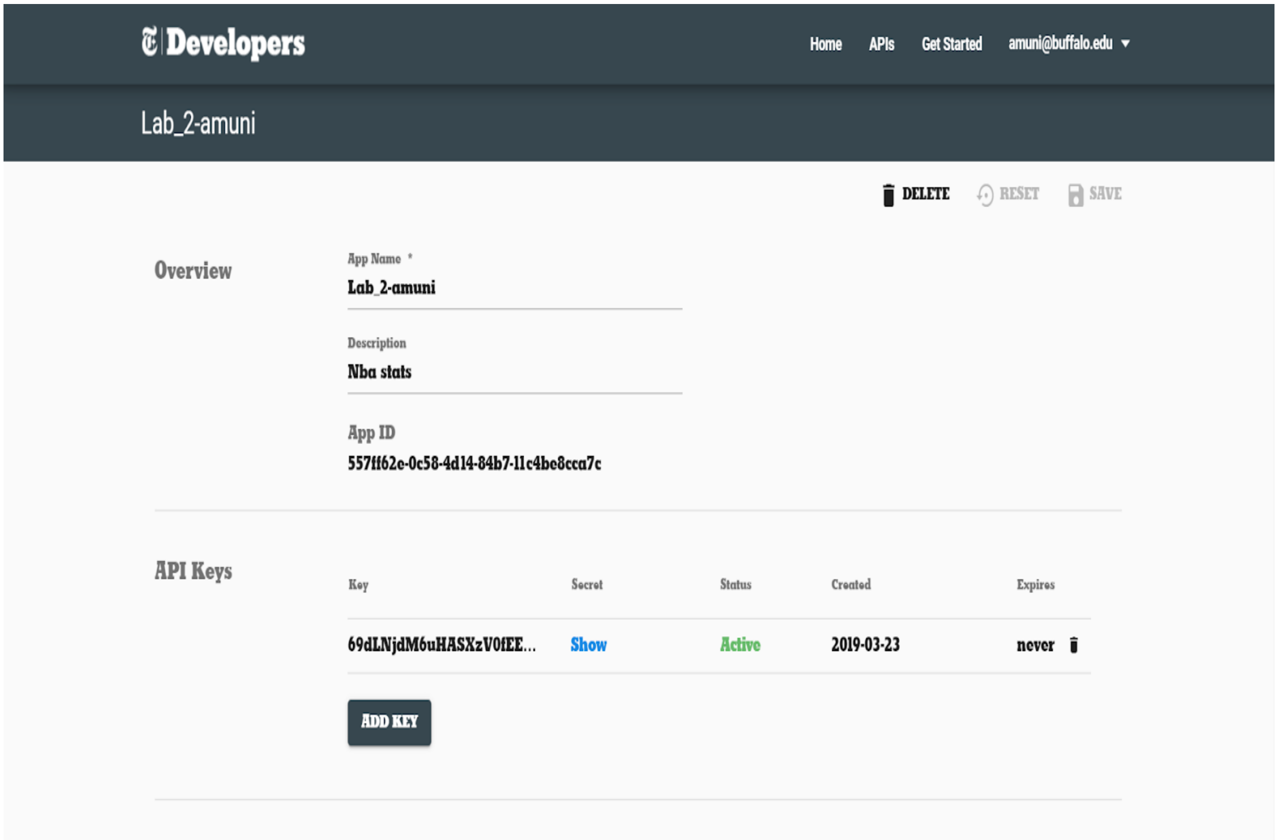
XtHHZnh8N398m2DxPbxltHpqE6RK6aPwudWuf5D9K5G68 (Access token secret) NEW

Read and write (Access level)

- These were used in Rtweet package and the tweets were collected using the keyword ‘NBA’ -
- The duplicate tweets were removed and were written into a csv file.

b) Data aggregation from NewYork Times:

- For collection of data from New york times, we used the article search API in the NYT app.
- The access token was taken from above and the URLs related to the keyword 'NBA' was obtained using the NYT Article search package in python.
- Then using Beautiful Soup package the meta-data was obtained from the URLs .
- Also using the above package,HTML parsing was done and the output was written in a text file.Below given is the screenshot of the NYT app from which the article search API was obtained.



Lab_2-amuni

Overview

App Name *
Lab_2-amuni

Description
Nba stats

App ID
557ff62e-0c58-4d14-84b7-11c4be8cca7c

API Keys

Key	Secret	Status	Created	Expires
69dLNjdM6uHASXzV0fEE...	Show	Active	2019-03-23	never

c) Data aggregration from common crawl :

- We have used python to mine common crawl data.
- Common Crawl is a gigantic dataset that is created by crawling the web. They provide the data in both downloadable format (gigantic) or we can query against their indices and only retrieve back the information you are after.
- We access the compressed archive files stored on Amazon S3 and pull out the actual content.

d) After data was obtained from three different sources some amount of pre-processing was done before sending it for Big data analysis. The following operations were performed-

- **Stemming operation**: Stemming operation is performed to produce different morphological variants of the stem word. This is used to put together the words that are different in grammatical perspective, but are same for the purpose of data analysis.
- **Stop words:** These are a set of words like grammatical articles('the','is','in') which are redundant for data analysis. These were removed before the file was sent for big-data processing.
- The above implementations were used in a python script to filter out the same and provide the results which were then fed into the MR framework.

2) Application of classical big data analytic method of MapReduce to the unstructured data collected-

We have used the AWS Elastic Map Reduce framework to achieve the implementation.

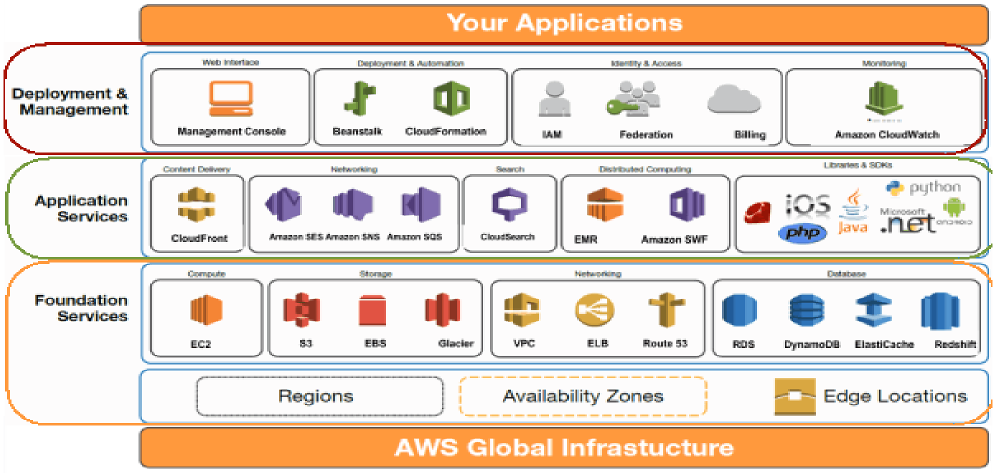
AWS EMR is a flexible and scalable approach and provides an immense enterprise level flexibility when it comes to working on Big Data.

Setting up AWS:

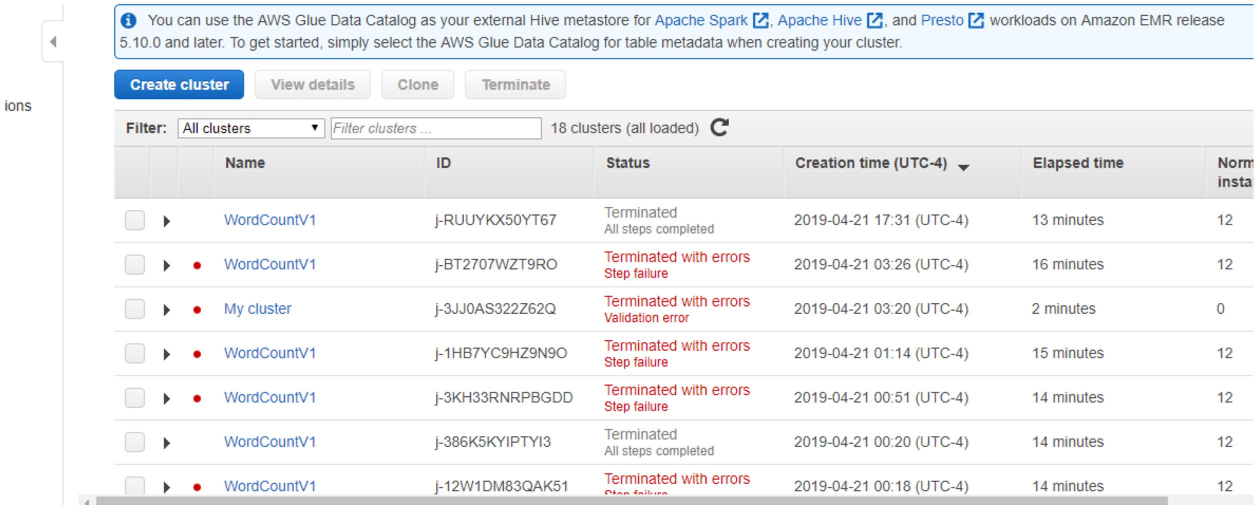
We have largely followed the traditional AWS documentation to create an EMR cluster, create a key pair, add steps and so on.

https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-gs-launch-sample-cluster.html

AWS Architecture:



Cluster Dashboard:



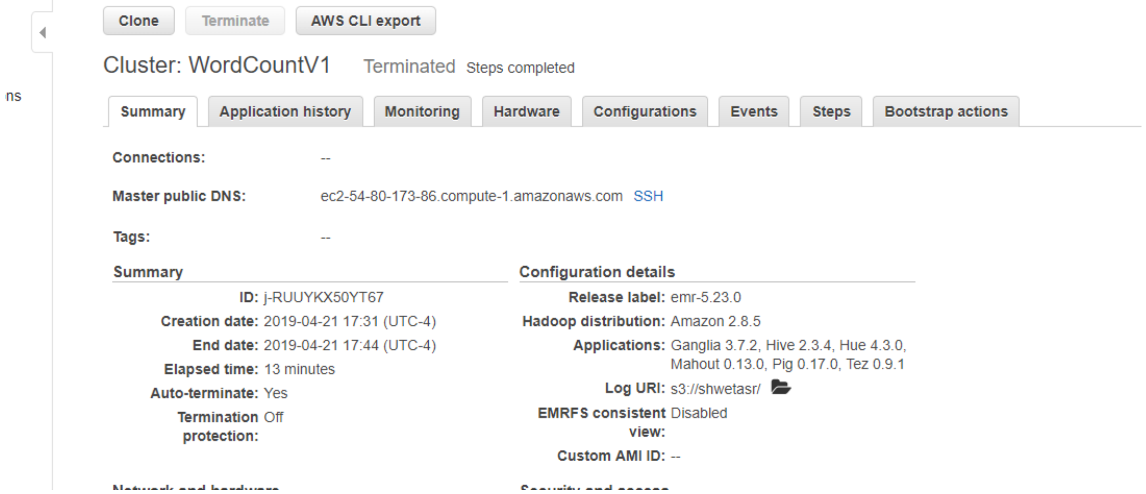
ions

You can use the AWS Glue Data Catalog as your external Hive metastore for Apache Spark, Apache Hive, and Presto workloads on Amazon EMR release 5.10.0 and later. To get started, simply select the AWS Glue Data Catalog for table metadata when creating your cluster.

Filter: All clusters ▾ Filter clusters ... 18 clusters (all loaded)

			Name	ID	Status	Creation time (UTC-4)	Elapsed time	Norm insta
□	▸		WordCountV1	j-RUUYKX50YT67	Terminated All steps completed	2019-04-21 17:31 (UTC-4)	13 minutes	12
□	▸	●	WordCountV1	j-BT2707WZT9RO	Terminated with errors Step failure	2019-04-21 03:26 (UTC-4)	16 minutes	12
□	▸	●	My cluster	j-3JJ0AS322Z62Q	Terminated with errors Validation error	2019-04-21 03:20 (UTC-4)	2 minutes	0
□	▸	●	WordCountV1	j-1HB7YC9HZ9N9O	Terminated with errors Step failure	2019-04-21 01:14 (UTC-4)	15 minutes	12
□	▸	●	WordCountV1	j-3KH33RNRPBGDD	Terminated with errors Step failure	2019-04-21 00:51 (UTC-4)	14 minutes	12
□	▸		WordCountV1	j-386K5KYIPTYI3	Terminated All steps completed	2019-04-21 00:20 (UTC-4)	14 minutes	12
□	▸	●	WordCountV1	j-12W1DM83QAK51	Terminated with errors Step failure	2019-04-21 00:18 (UTC-4)	14 minutes	12

Properties of a cluster



ns

Cluster: WordCountV1 Terminated Steps completed

Connections: --

Master public DNS: ec2-54-80-173-86.compute-1.amazonaws.com SSH

Tags: --

Summary

ID: j-RUUYKX50YT67
Creation date: 2019-04-21 17:31 (UTC-4)
End date: 2019-04-21 17:44 (UTC-4)
Elapsed time: 13 minutes
Auto-terminate: Yes
Termination protection: Off

Configuration details

Release label: emr-5.23.0
Hadoop distribution: Amazon 2.8.5
Applications: Ganglia 3.7.2, Hive 2.3.4, Hue 4.3.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.9.1
Log URI: s3://shwetasr/
EMRFS consistent view: Disabled
Custom AMI ID: --

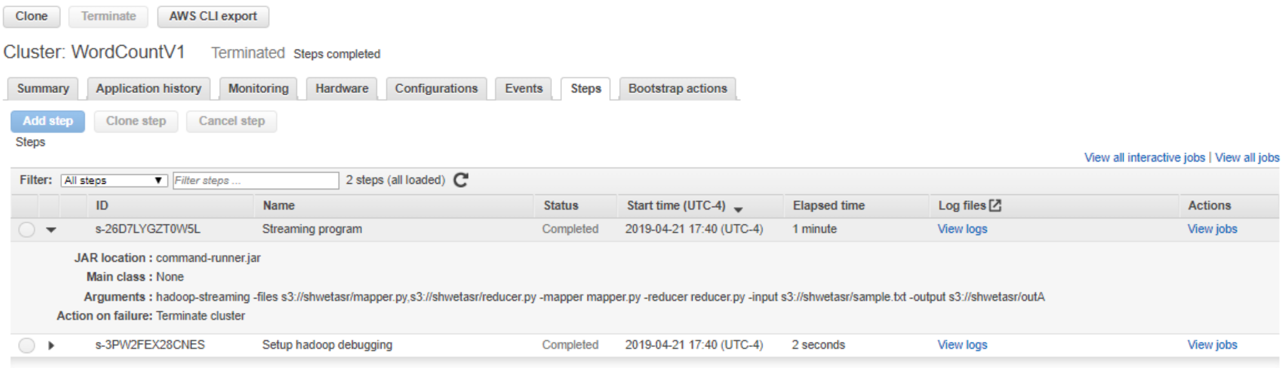
Network and hardware

Security and access

Properties of an MR Streamline step:

*We use custom jar option if we are to run the MR job on the basis of a jar file.

*We use streamlined operation if we are to use .py approach to separately use python files as mapper, reducer, locations of input and output folder.



Cluster: WordCountV1 Terminated Steps completed

2 steps (all loaded)

ID	Name	Status	Start time (UTC-4)	Elapsed time	Log files	Actions
s-26D7LYGZT0W5L	Streaming program	Completed	2019-04-21 17:40 (UTC-4)	1 minute	View logs	View jobs
s-3PW2FEX28CNES	Setup hadoop debugging	Completed	2019-04-21 17:40 (UTC-4)	2 seconds	View logs	View jobs

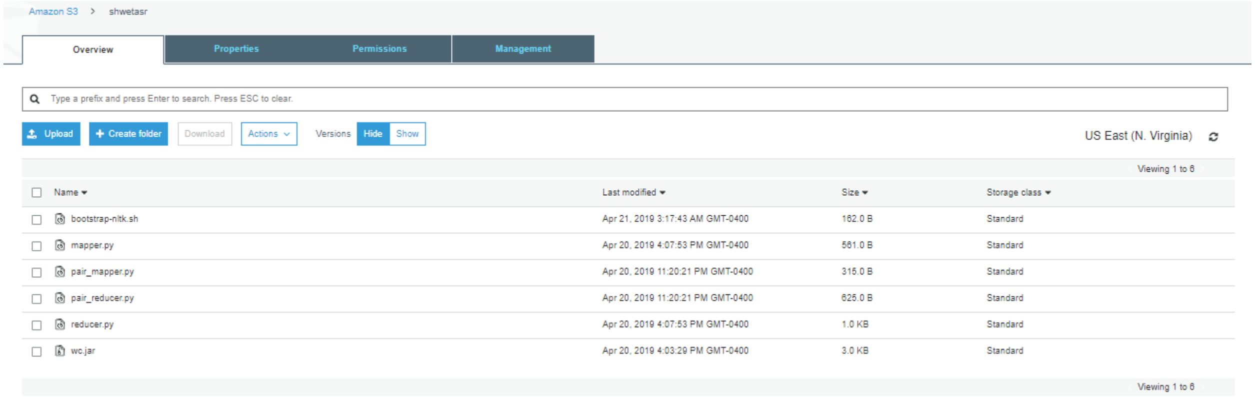
JAR location : command-runner.jar
Main class : None
Arguments : hadoop-streaming -files s3://shwetasr/mapper.py,s3://shwetasr/reducer.py -mapper mapper.py -reducer reducer.py -input s3://shwetasr/sample.txt -output s3://shwetasr/outA
Action on failure: Terminate cluster

Setting up the S3 bucket:

Amazon S3 is cloud storage for the internet. To upload your data (photos, videos, documents etc.), you first create a bucket in one of the AWS Regions. You can then upload any number of objects to the bucket.

In terms of implementation, buckets and objects are resources, and Amazon S3 provides APIs for you to manage them. For example, you can create a bucket and upload objects using the Amazon S3 API. You can also use the Amazon S3 console to perform these operations. The console uses the Amazon S3 APIs to send requests to Amazon S3.

Files can be uploaded on a drag and drop basis.

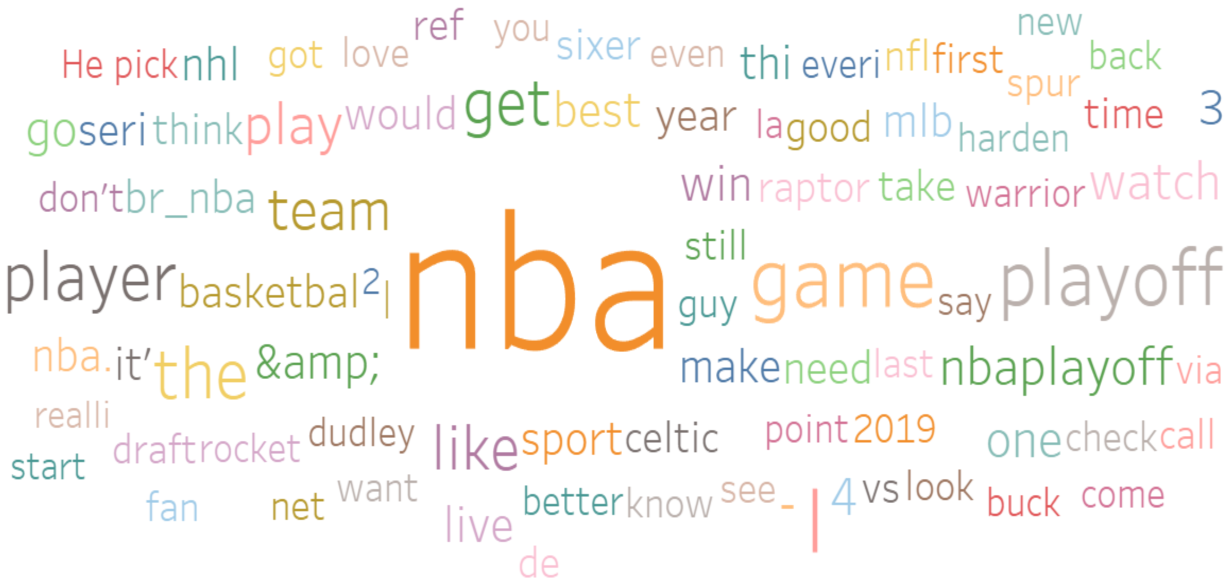


Name	Last modified	Size	Storage class
bootstrap-nltk.sh	Apr 21, 2019 3:17:43 AM GMT-0400	162.0 B	Standard
mapper.py	Apr 20, 2019 4:07:53 PM GMT-0400	561.0 B	Standard
pair_mapper.py	Apr 20, 2019 11:20:21 PM GMT-0400	315.0 B	Standard
pair_reducer.py	Apr 20, 2019 11:20:21 PM GMT-0400	625.0 B	Standard
reducer.py	Apr 20, 2019 4:07:53 PM GMT-0400	1.0 KB	Standard
wc.jar	Apr 20, 2019 4:03:29 PM GMT-0400	3.0 KB	Standard

4) Data visualization using Tableau-

Finally the outputs obtained after the big data analysis was visualized using Tableau. The outputs obtained after running word-count on three different data-bases were as follows-

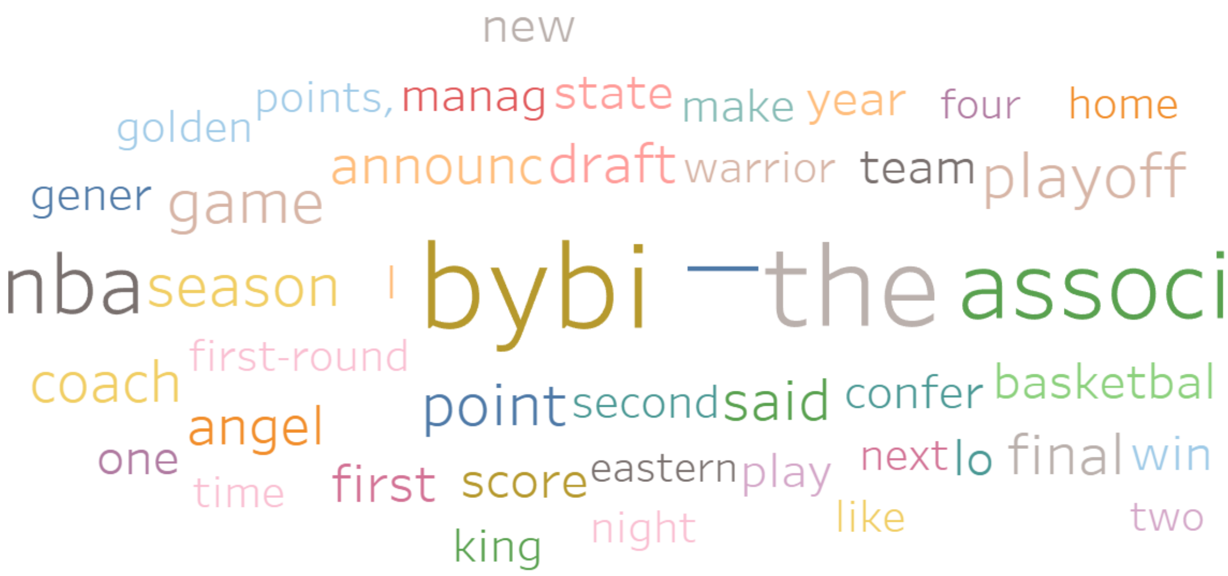
1. **Output for running word-count on twitter dataset:**



Interpretations:

- As we were doing data twitter data analysis on the key-word nba we could see highest number of “**nba**” words in the related tweets.
- As the **playoff** season of nba-2019 was onset in mid-march, we also find considerable amount of playoff related tweets during this time of the year.
- As “James harden” is one of the leading scores in this season, we get some considerable number of tweets related to him.
- Also we could see many tweets related to the team that are table-toppers in both the conferences like Boston Celtics, mlb, Toronto Raptors, Golden states warriors

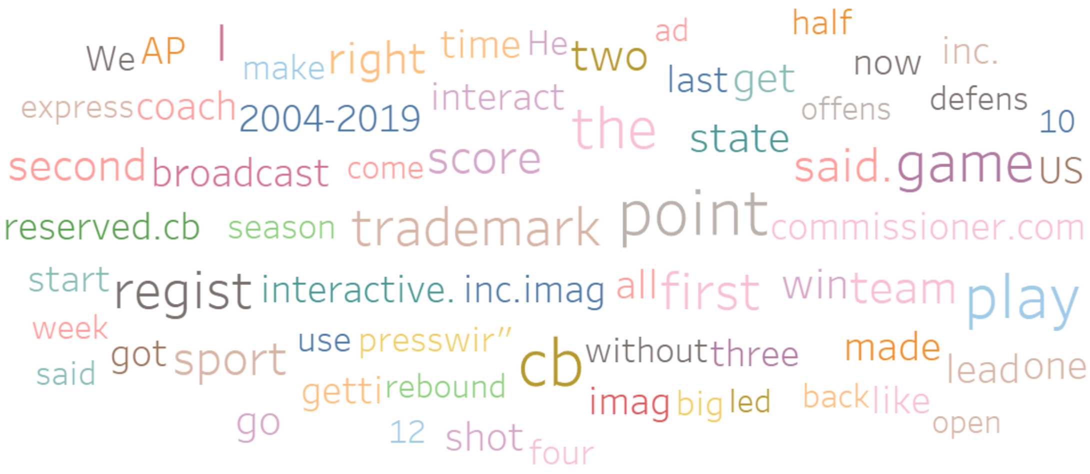
2)Output for running word-count on NewYork Times data:



Interpretation:

- As we searched for nba keyword for article search we could find considerable **nba** words inside the articles.
- As **golden state warrior** has been the league champion for past couple of seasons we could find many analyses on the same team.
- As nba is a basketball league we could find many **basketball** words in the article analysis.
- As nba involves a lot of matches we could find many analyses related to winning of team.
- As recently we had summer **draft** for the nba we could find some of the related articles as it is crucial for the upcoming season.

Output generated for running word-count on common-crawl data:



Data visualization using Tableau:

https://public.tableau.com/views/lab2_15559316767240/nyt_worsdcouccurence?:embed=y&:display_count=yes&publish=yes