# Principles of Big Data Management

# Project Report – Phase 2

**Anil, Suravarapu**

**Naga Nidesh, Doppalapudi**

**Praveen, Poluri**

# Goal

The theme of the project is to collect tweets on **Cricket data** andto develop spark SQL queries to retrieve data from the collected tweets and visualize them using Apache Zeppelin.

# Tools and Frameworks Used

**Languages**: Java, Scala

**Framework**: Spark

**File System** : Hadoop Distributed File System

**Visualization Platform** : Zeppelin

Query 1 :

The following spark sql command is used to examine the most tweeted twitter account:

**A screenshot of a cell phone

Description automatically generated**

A picture containing screenshot

Description automatically generated

Query2:

The following spark sql command is used to examine the the most discussed cricket topic:

A close up of a logo

Description automatically generated

A picture containing screenshot

Description automatically generated

Query3:

The following spark sql command is used to examine the number of tweets tweeted in a particular type of place (city, country etc)

A screenshot of a cell phone

Description automatically generated

A picture containing screenshot

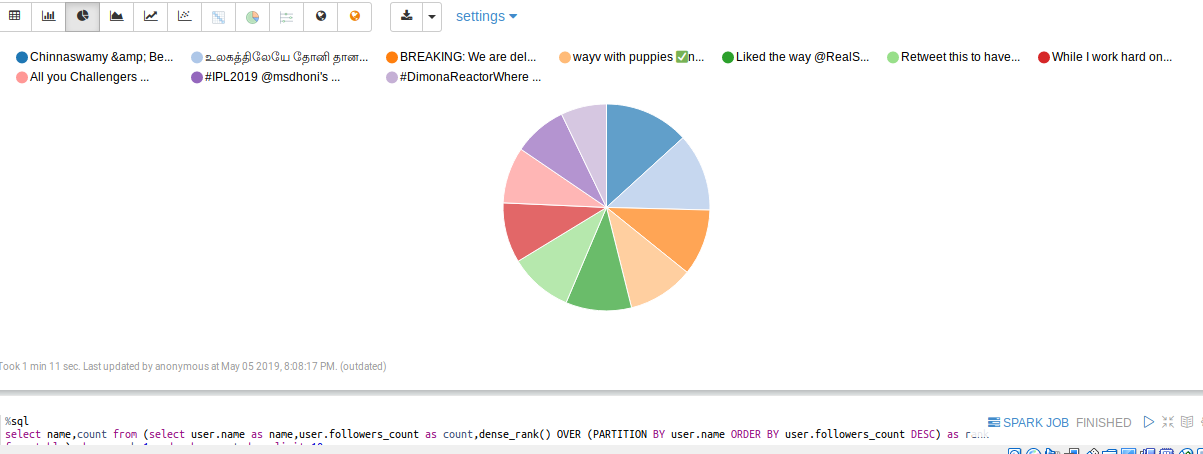
Description automatically generated

Query4:

The following spark sql command is used to examine the most discussed tweet:

A picture containing screenshot

Description automatically generated



Query5:

The following spark sql command is used to examine which twitter account has more followers:



A picture containing screenshot

Description automatically generated

Query6:

The following spark sql command is used to examine in which language people tweeted more:

A screenshot of text

Description automatically generated

A picture containing screenshot

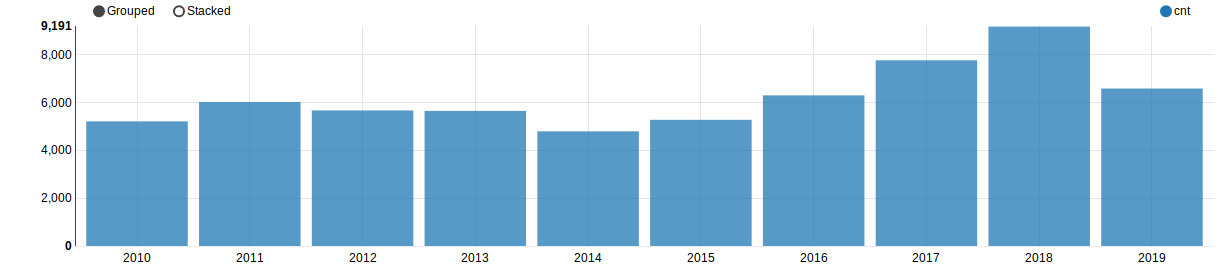
Description automatically generated

Query7:

The following spark sql command is used to examine in which year more accounts are created:

A screenshot of a cell phone

Description automatically generated



Query8:

The following spark sql command is used to examine which cricket player is more discussed:



A picture containing screenshot

Description automatically generated

Query9:

The following spark sql command is used to examine which cricket player is more discussed countrywise:

A screenshot of a social media post

Description automatically generated

A screenshot of a social media post

Description automatically generated

Query 10:

The following spark sql command is used to examine the most discussed tweet in cricket about a topic:

A screenshot of a cell phone

Description automatically generated

A screenshot of a computer

Description automatically generated

**val df=spark.sqlContext.read.json("hdfs://localhost:9000/project/Cricket\_Tweets.json")**

**df.registerTempTable("table");**

**val query1=spark.sqlContext(“select name,TweetCount from(select user.name as name,count(\*) TweetCount from table group by user.name) order by TweetCount DESC LIMIT 10”)query1.coalesce(1).write.format("com.databricks.spark.csv").option("header","true").save("/home/praveen/Downloads/project/query1");**

The above query is used to run the spark SQL query in Spark and storing the output in a csv format. With this file we manually tested the output generated by the zeppelin is correct or not

Output generated are stored in GitHub: <https://github.com/praveenpoluri/principles-of-big-data>

Reference:

https://zeppelin.apache.org/docs/0.7.0/interpreter/spark.html