Trending YouTube Video Statistics Visualization - Design Document



Submitted To,

Professor Weider D.Yu

Submitted By,

**Team - B**

Mahesh Reddy (013823095)

Vanditt Sama (014210521)

Pardis Tashakori (011548199)

Vignesh Thyagarajan (009584861)

Vinay Kovuri (013710684)

Sai Krishna Reddy Jali (013752440)

Sandhya Ramachandraiah (013312949)

Naveen Ravipati (013756028)

**TABLE OF CONTENTS**

[**1. Introduction**](#_dezhvh49cvjl) **1**

[**2. Abstract**](#_twrekgw4cjt2) **2**

[**3. Project Architecture**](#_betz2h2yjixt) **2**

[**3. Project Data**](#_mtye6gs9b3x0) **3**

[**4. Cassandra**](#_w88bosrcwbq6) **4**

[Setup](#_jf04r1o7eo34) 4

[Implementation](#_bv8h9dgbxfhz) 5

[**5. Tableau and Project Functions**](#_yla4pv8zg0ua) **6**

[Tableau Introduction:](#_v92xo3mm7een) 6

[**6. Visualizations using R studio**](#_uyux8v5pz1by) **9**

[**7. Performance Measure**](#_77kjq27z1dve) **10**

[Business Performance Measure:](#_z8yli7sk6lzy) 10

[Scorecard:](#_agdqdvuzmfob) 11

[Balanced Scorecard:](#_pqrnamuai2de) 11

[Descriptive Analysis:](#_a77eddz66xw4) 11

[**8. Expected Results**](#_p4jkb8kqrqw2) **12**

# 

# 1. Introduction

YouTube has been the most used video platform in the world today. It has a list of trending videos that are updated constantly. We analyzed the csv files in YouTube to analyze the data store. The data set consists of trending videos of that day. We will analyze this data to get insight into YouTube trending videos and to see what is common between the videos. The goal of the analysis is to determine the number of views of trending videos, to find out how views, likes, dislikes, and comment counts of videos are related to one another. We also find the most common words in trending video titles.

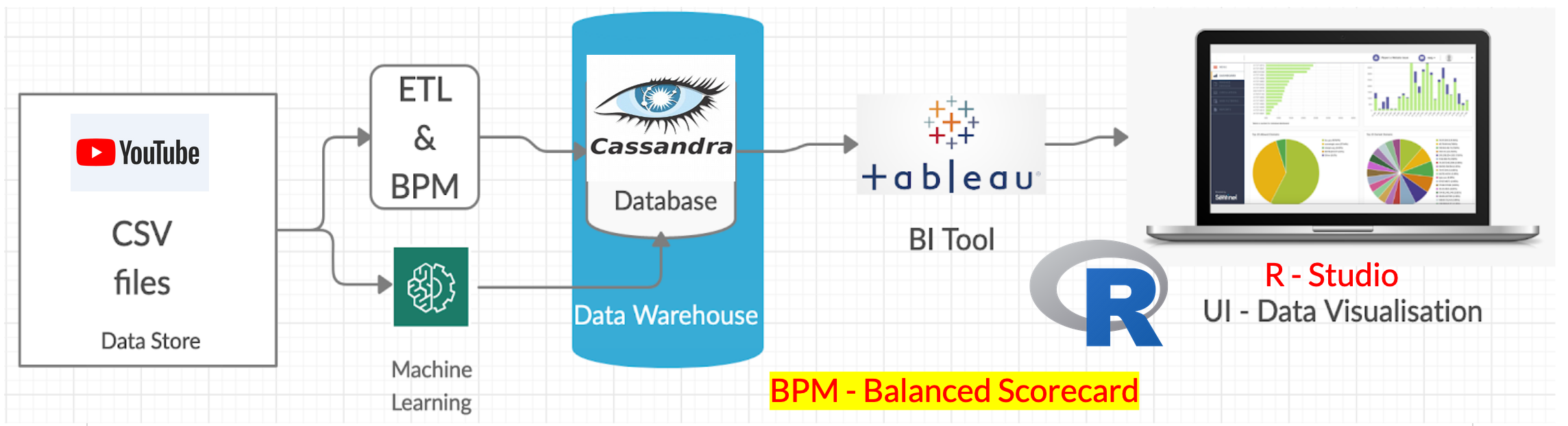
# 2. Abstract

Youtube trending videos represent content that targets viewers attention.In this project we present our findings for measuring, analysing key aspects of youtube trending videos.Trending videos and channels are compared to other youtube content which has not been labeled as trending.Better understanding of youtube trending videos and their statistics can be helpful in strategies of marketing,target advertising , recommendation systems and search engines .Our analysis sheds some insights on some of the factors that might be influencing determination and popularity of trending videos.

# 3. Project Architecture

This section discusses the technology choices and deployment strategies along with their justification.

* Dataset: The data utilized for this project was obtained from a Kaggle dataset and arranged in a CSV (Comma Separated Values) file. This is a popular format which is readily parsable using simple code and all popular databases support direct import of data from CSV files.
* Database: The dataset will be loaded onto the Cassandra database after performing necessary ETL operations. Since the provided dataset is structurally refined, only minimal business related operations will be required on data at this stage. Detailed description in section 4.
* RStudio: R with Rstudio will be used to perform statistical analysis. Data from the Cassandra will be accessed using R and relevant functions will be applied to plot graphs and visualizations. Detailed description in section 6.
* Tableau: A user facing Tableau dashboard will be created to visualize the results of statistical analysis. Intuitive Tableau widgets with graphs will allow businesses to draw inferences from the given data.

 **Fig: Architecture Diagram**

# 

# 3. Project Data

[Trending YouTube Statistics](https://www.kaggle.com/datasnaek/youtube-new)

The project data we will be using comes from Kaggle, and contains data on trending YouTube video statistics.

Our dataset is 514 MB in size. Columns include: *video\_id*, *trending\_date*, *title*, *channel\_title*, *category\_id*, *publish\_time*, *tags*, *views*, *likes*, *dislikes*, *comment\_count*, *thumbnail\_link*, *comments\_disabled*, *ratings\_disabled*, *video\_error\_or\_removed*, and *description*.

While all important, we will hone in on: **trending\_date**, **title**, **channel\_title**, **views**, **likes**, **dislikes**, **comment\_count**, **comments\_disabled** and **ratings\_disabled**.

# 4. Cassandra

Cassandra is a distributed database management system which is open source with a wide column store, NoSQL database to handle large amounts of data across many commodity servers which provides high availability with no single point of failure.

Listed below are some points of Apache Cassandra:

* It is scalable, fault-tolerant, and consistent.
* It is a column-oriented database.
* Its distributed design is based on Amazon’s Dynamo and its data model on Google’s Big table.
* It is Created at Facebook and it differs sharply from relational database management systems.

## Setup

1. Cassandra needs JDK. Install JDK on the PC.

* Go To [Oracle](https://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html) (SignUp required) or [Filehippo](https://filehippo.com/download_java_development_kit_64/86378/) to download JDK 1.8 from which you find suitable.
* Run the Install as it is.
* Copy the JDK where it is installed it’s bin directory. Mine is 64-bit found inside “C:\Program Files\Java\jdk1.8.0\_181\bin” and placed on Environment Variable as new Env\_name : “JAVA\_HOME” later we use for cassandra.
* Note: Progra~1 = ‘Program Files’ for 64-bit. Progra~2 = ‘Program Files(x86)’ for 32-bit in Environment path.

1. Go to [Apache Cassandra](http://cassandra.apache.org/download/) Download Page. And Download the latest version. The latest version at that time is cassandra-3.11.4.

* Unzip it and place all files inside sub folder into ……………….. “C:\Program Files\apache-cassandra-3.11.4”
* Open CMD inside “C:\Program Files\apache-cassandra-3.11.4\bin” and type cassandra then the output as shown below.
* If you got above error then Edit “cassandra.bat” Add JAVA\_HOME
* Save the cassandra.bat then run “cassandra.bat -f” with CMD Run as Administrator. This time, error should resolve if your edit is successful.

1. Need Python2.7 to run Cassandra Query shell cqlsh. Download [Python2.7](https://www.python.org/download/releases/2.7/) latest version and extract inside the “C:\Program Files\apache-cassandra-3.11.4\bin” during installation. Or simply copy after installing python2.7 all files inside the “bin” directory of cassandra-3.11.4. This is an easy way to go.
2. Finally run the Cassandra Server as “cassandra.bat -f” with CMD Run as Administrator from the bin directory. Following screen should show a successful start. Don’t close it, keep running. Then, open another CMD go over the bin directory of cassandra installed. To run cqlsh by type **cqlsh .** Following output should get.

## 

## Implementation

1. **Create Keyspace in Cassandra**

A keyspace is an object that holds the column families, user defined types. In Cassandra, Keyspace is similar to RDBMS Database. Keyspace holds column families, indexes, user defined types, data center awareness, strategy used in keyspace, replication factor, etc.

**Syntax**: Create keyspace KeyspaceName with replication={'class':strategy name, 'replication\_factor': No of replications on different nodes};

1. **Create Table in Cassandra**

Column family in Cassandra is similar to the RDBMS table. Column family is used to store data. Command 'Create Table' is used to create column families in Cassandra.

**Syntax**: Create table KeyspaceName.TableName

(ColumnName DataType, ColumnName DataType, ColumnName DataType

Primary key(ColumnName)) with PropertyName=PropertyValue;

1. **Insert Data into Cassandra**

Command 'Insert into' writes data in Cassandra columns in row form. It will store only those columns that are given by the user. You have to necessarily specify just the primary key column. It will not take any space for not given values. No results are returned after insertion.

**Syntax:** Insert into KeyspaceName.TableName(ColumnName1, ColumnName2, ColumnName3 . . . .)values (Column1 Value, Column2Value, Column Value . . . .)

# 5. Tableau and Project Functions

## Tableau Introduction:

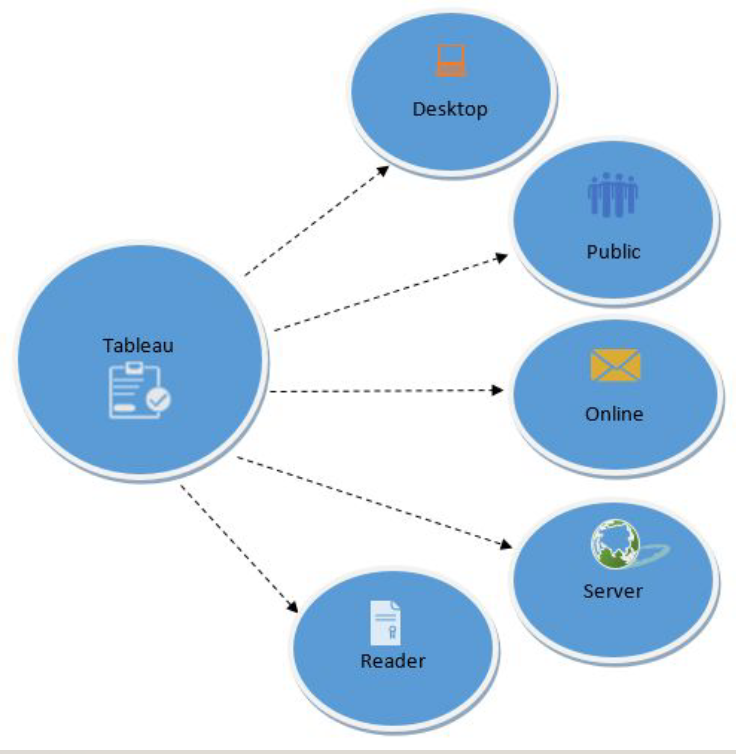
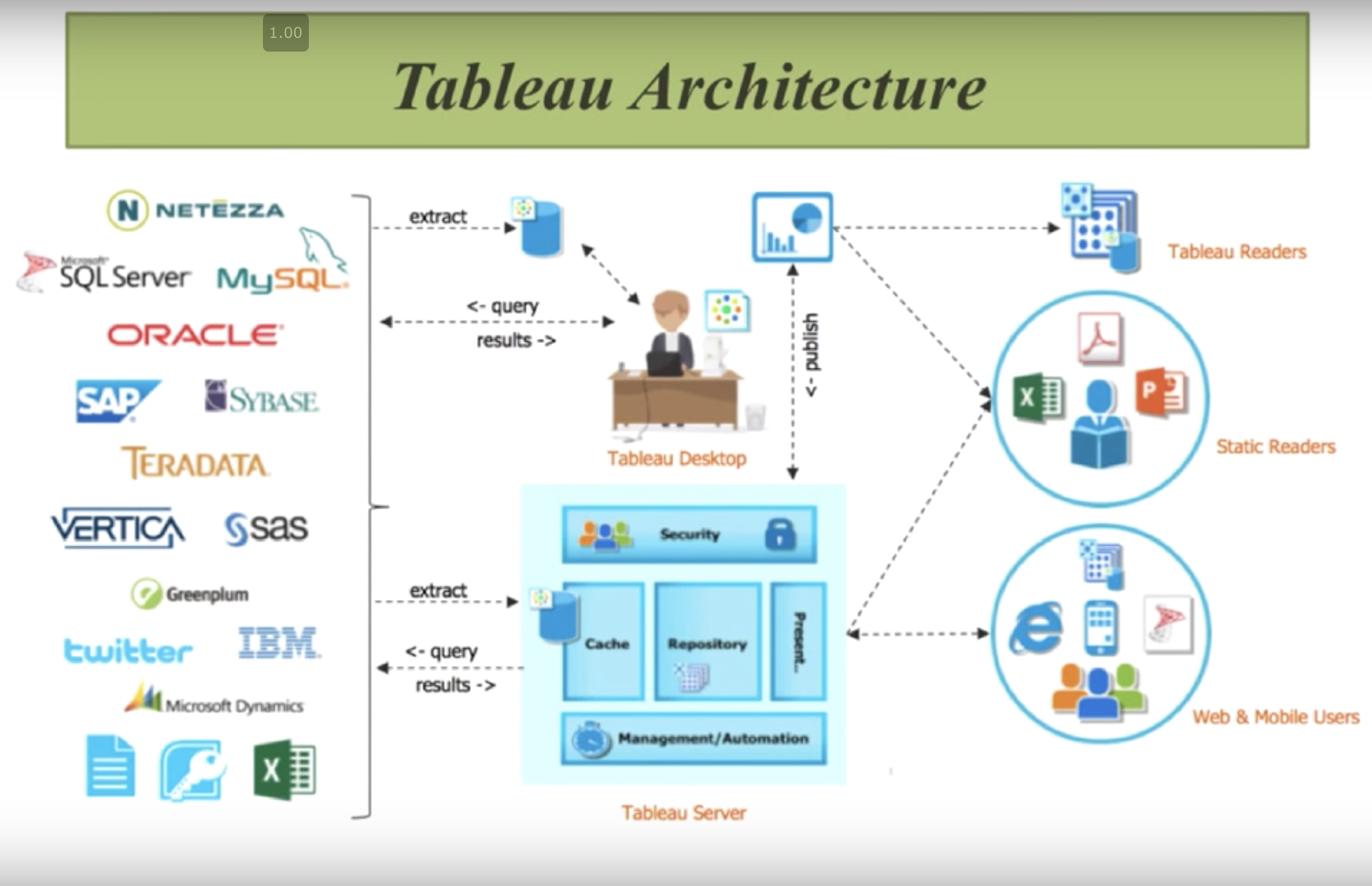
****

Tableau is a powerful and fastest growing data visualization tool used in the Business Intelligence Industry. It helps in simplifying raw data into the very easily understandable format.

Data analysis is much faster with Tableau and the visualizations created are in the form of worksheets and dashboards. Data that is created by using Tableau can be easily understood by persons of all levels in every organization. It also allows the non-technical users to design the customized dashboard.



**Tableau Workspace**

The workspace of Tableau has menus, the Data pane, a toolbar, cards and shelves, and one or more sheets. Sheets can be worksheets, dashboards, or stories.

**Types of Tableau:**

Tableau Desktop, Tableau Public, and Tableau Online, all offer Data Visual Creation and choice depends upon the type of work.

**Types of representations in Tableau:**

There are many representations available in Tableau. Some of the popular ones are,

Line graph, Bar graph, Heat map, Pie chart, Scatter plot, Box and Whisker plot, Packed bubbles.



So, to convey the insights effectively and efficiently, We are using horizontal bar graphs.

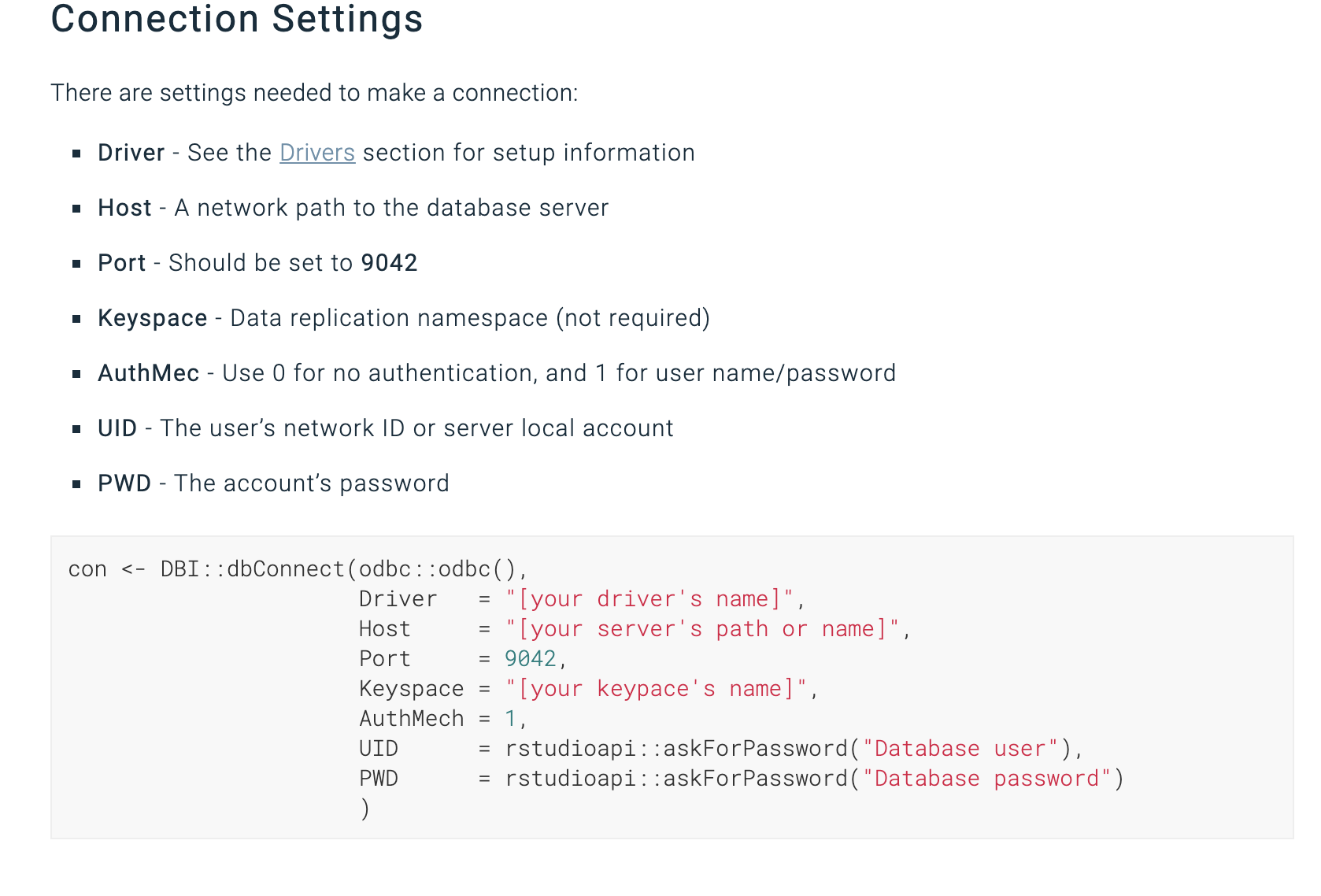
Project Functions that can possibly be used to help visualize and analyze our data fall under the “Aggregate Function” category of Tableau’s possible functions. Aggregate functions allow us to summarize or change the granularity of our data. Since the data we are looking at involves trends, the data is statistical in nature and we will need functions that suit the data. Some functions are ATTR, AVG, COLLECT, CORR, COUNT, COUNTD, COVAR, COVARP, MAX, MEDIAN, MIN, PERCENTILE, STDEV, STDEVP, SUM, VAR, and VARP. More information on these functions can be found [here](https://help.tableau.com/current/pro/desktop/en-us/functions_all_categories.htm) on the Tableau Help site.

Additional Project Functions include functions for R packages and Cassandra. These functions are Dplyr and Tidyr for R, and CREATE\_TABLE, COPY & FROM, COPY & TO. Dplyr is a package function that allows for data manipulation in R, while Tidyr is used to tidy the columns where each variable is present in the column. The CREATE\_TABLE function creates a table with all the specified columns as the primary key. The remaining columns in the importing csv will be set to null. The COPY & FROM function specifies the path from where data can be imported, and the COPY & TO function specifies the path to which data can be exported. These functions can be useful

# 6. Visualizations using R studio

R language combined with Cassandra database is a great tool for visualisation and plotting of youTube trending statistical data. Since R is a high level scripting language just like Python and Javascript, there are a lot of IDEs available out in the market. R studio is a very famous environment where we can plot a lot of graphs and charts for visualisation very easily. Jupyter notebook can also be used to run the R programming language with some basic settings modification. There are several options and steps available to set up R with Cassandra database. They are explained below.

**ODBC Package**



**RJDBC Package**

**Step 1:**

library(RJDBC)

**Step 2:**

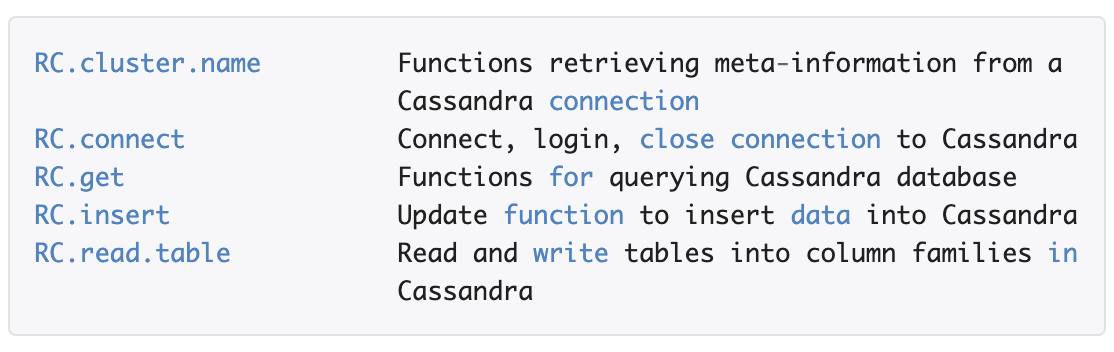
driver <- JDBC(driverClass = "cdata.jdbc.cassandra.CassandraDriver", classPath = "MyInstallationDir\lib\cdata.jdbc.cassandra.jar", identifier.quote = "'")

**Step 3:**

conn <-

dbConnect(driver,"jdbc:cassandra:Database=MyCassandraDB;Port=7000;Server=127.0.0.1;")

**RCassandra Package**

****

# 7. Performance Measure

## Business Performance Measure:

During the past few years, many companies have recognized the significance of enforcing and reaching the goals defined by the strategy through metrics-driven management. Organizations have vast technologies that help them monitor and report.

Business performance measure(BPM) is a key business strategy that allows businesses to manage their performance and get better decisions and action plans. It is used to design to link with top-level planning, monitoring, and planning. Business management performance is more than the technology that helps in delivering the insights that are needed to improve the overall business performance. Business performance technology is not a single entity but it has many other technologies that are merged into a single entity. It analyses the data from the past until now and this information is crucial in planning the goals.

## Scorecard:

Scorecarding allows you to estimate how you are delivering upon those planned goals. It is a snapshot of the performance of the corporates when related to their goals. This technique is important to manage the performance and take the appropriate decision on the current performance. The scorecard is a static view.

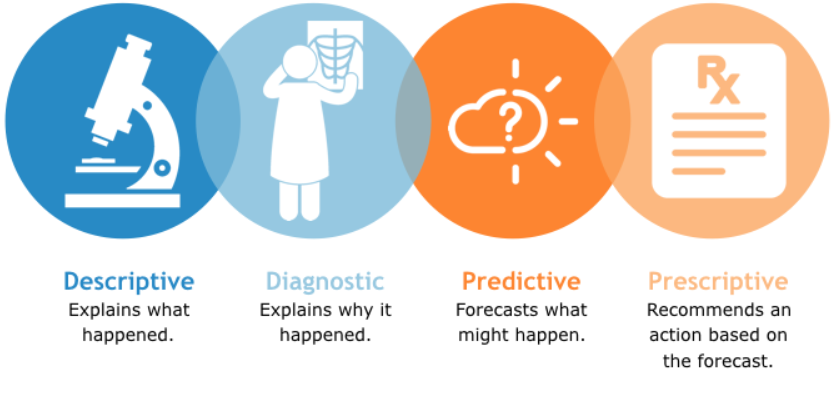
## Balanced Scorecard:

A performance measurement and management methodology that helps translate an organization’s financial, customer, internal process, and learning and growth objectives and targets into a set of actionable initiatives.

## Descriptive Analysis:

Descriptive Analysis tells you what happened in the past. This type of analysis helps in understanding the past patterns.

Here in this project, we are performing the descriptive analysis. We are using the data and performing various functions like ATTR, AVG, COLLECT, CORR, COUNT, COUNTD, COVAR, COVARP, MAX, MEDIAN, MIN, PERCENTILE, STDEV, STDEVP, SUM, VAR, and VARP. Each of these functions is used to analyze the past data from the dataset which is stored in the Cassandra database. Using these tableau functions we can easily anticipate the data patterns and help us to clearly understand the patterns. These functions help in understanding the data and make an easy picture out of it.

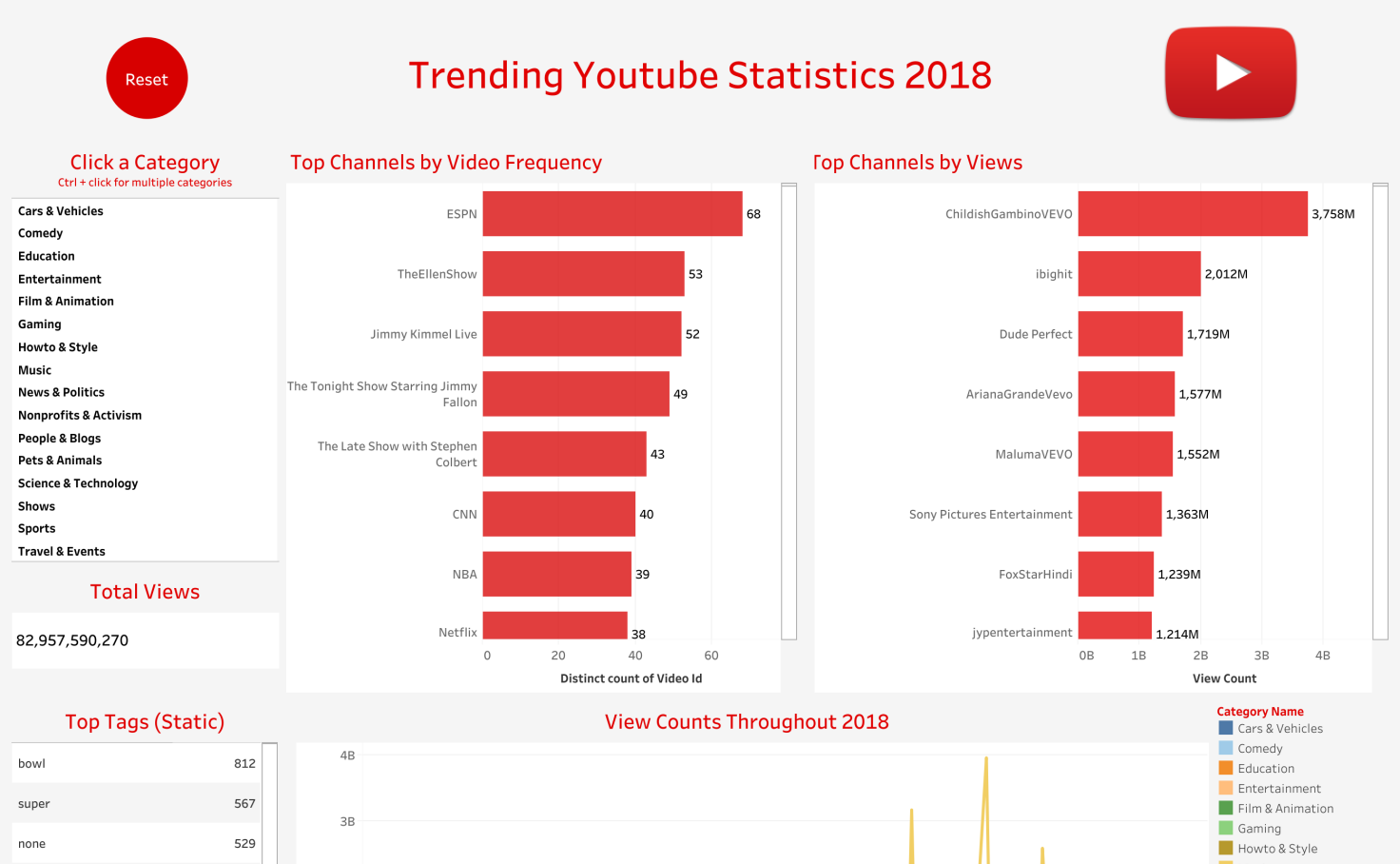
****

Here in this project, YouTube data is used to analyze the trending statistics. The data is all about the daily trending videos. To determine the overall year trending videos the youtube platform uses not only the like but also uses likes, comments data of the videos. A descriptive analysis of these data helps us to find the trending videos.

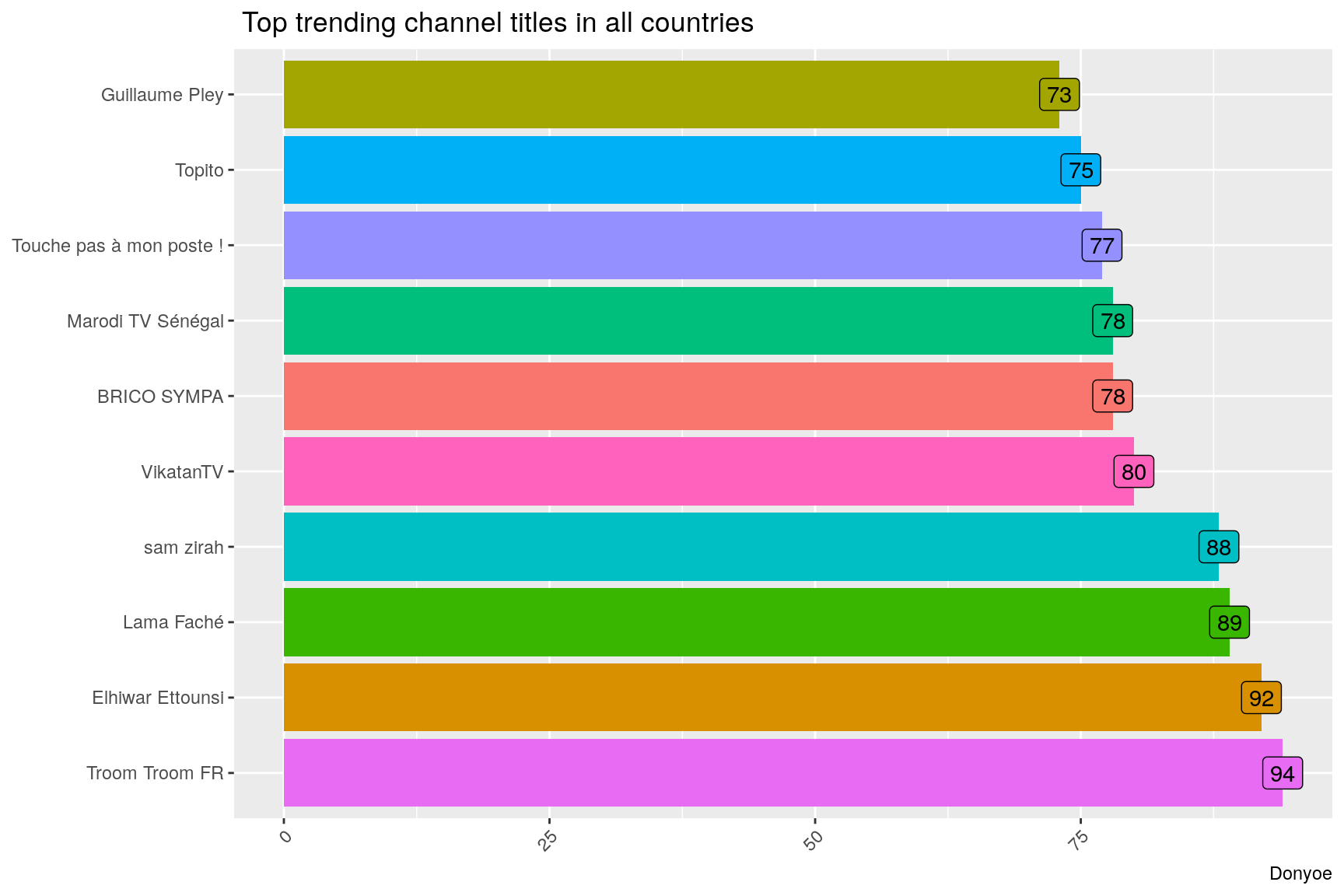
# 

# 8. Expected Results

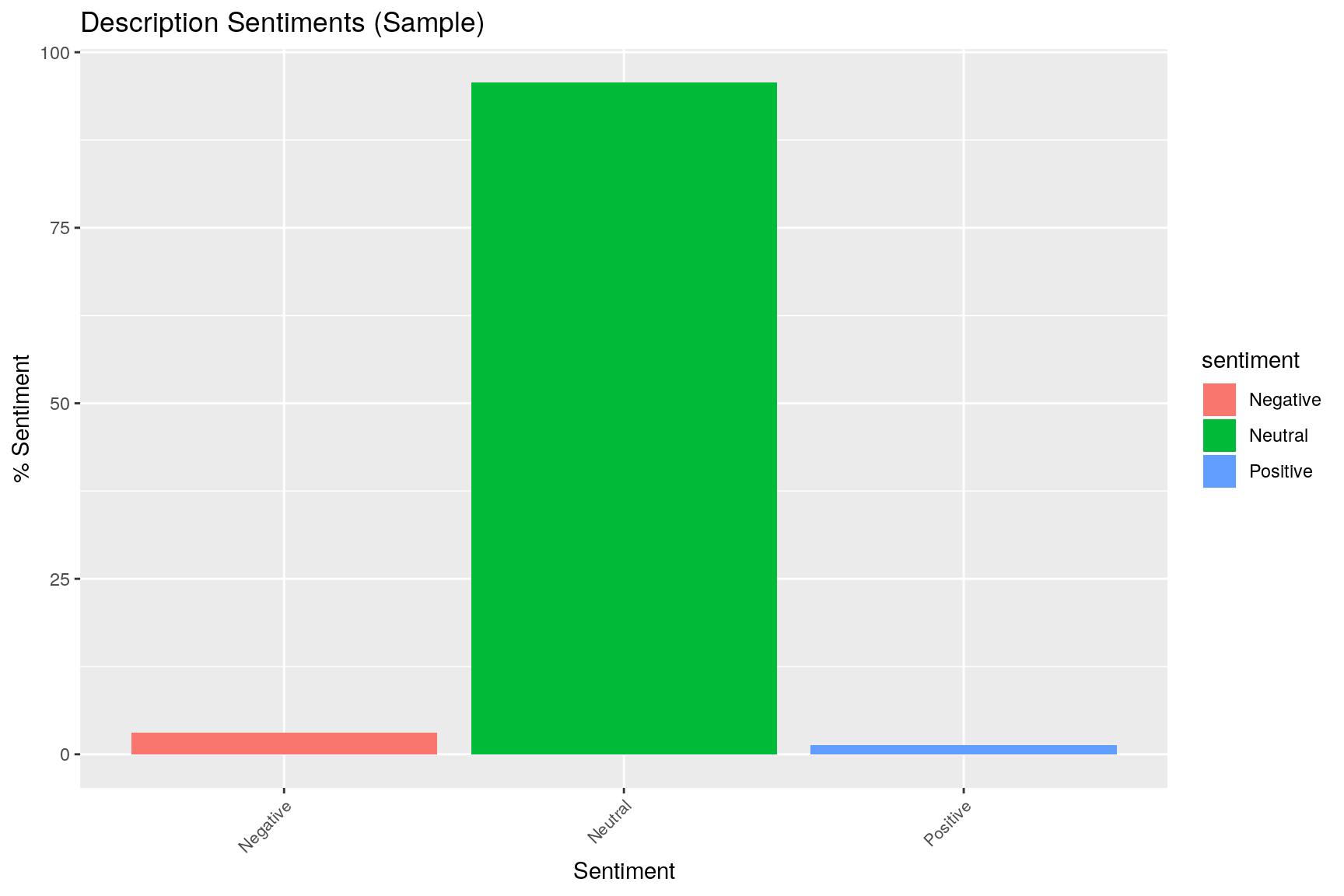
We anticipate analyzing and understanding the data well enough to create proper visualizations using Tableau. Moreover, our goal is to create a proper dashboard of visualizations that take into account proper user interface and dashboard principles. An example can be seen below:



Top trending Channels in all countries



Sentiment Analysis Description field (Sample)



A vital dashboard principle is avoiding scrollbars in individual graphs and the entire dashboard as a whole. While the figure above offers an idea as to what our goal is, we will make sure to keep all proper principles in mind and correctly implement them.

**9. Conclusion**

Overall, the important parts of our project include the analysis of data through R and Tableau, with Cassandra allowing us to efficiently manage our YouTube statistic data as needed. Firstly we extracted the youTube trending analytics data from kaggle website and imported that into a cassandra database. After that we perform ETL to load in cassandra database and leverage the Tableau tool for visualisations, R studio is connected using RJDBC with cassandra database to plot and display the visualisations.