**PROJECT PHASE II**

**Raghavendra Rao Yentrapathi (16188744)**

**Hiren Shah Bhai (16211493)**

**Pavana Krishna Chunduru (16209370)**

**Project Description:**

Design and implement your ideas using Apache Spark

* Figure out how to store the tweets
* Write interesting analytic queries to explore and understand the data
* Develop interesting visualizations

**Steps to Implement the Project:**

Below are the steps we followed to complete project phase-II

1) We collected the tweets in the JSON file as we did for the Phase I by taking the topic of

“Aamir Khan Intolerance” by using Python

2) We stored the collected tweets into spark table using Spark SQL

3) We wrote the below mentioned queries on the created table with different criteria. We stored the data queried into the MY SQL database

4) We visualized the data that was stored into MYSQL database by plotting the graphs using ASP.NET where ever they are necessary through the designed User Interface.

5) Description of each Queries and analysis on the data they retrieved

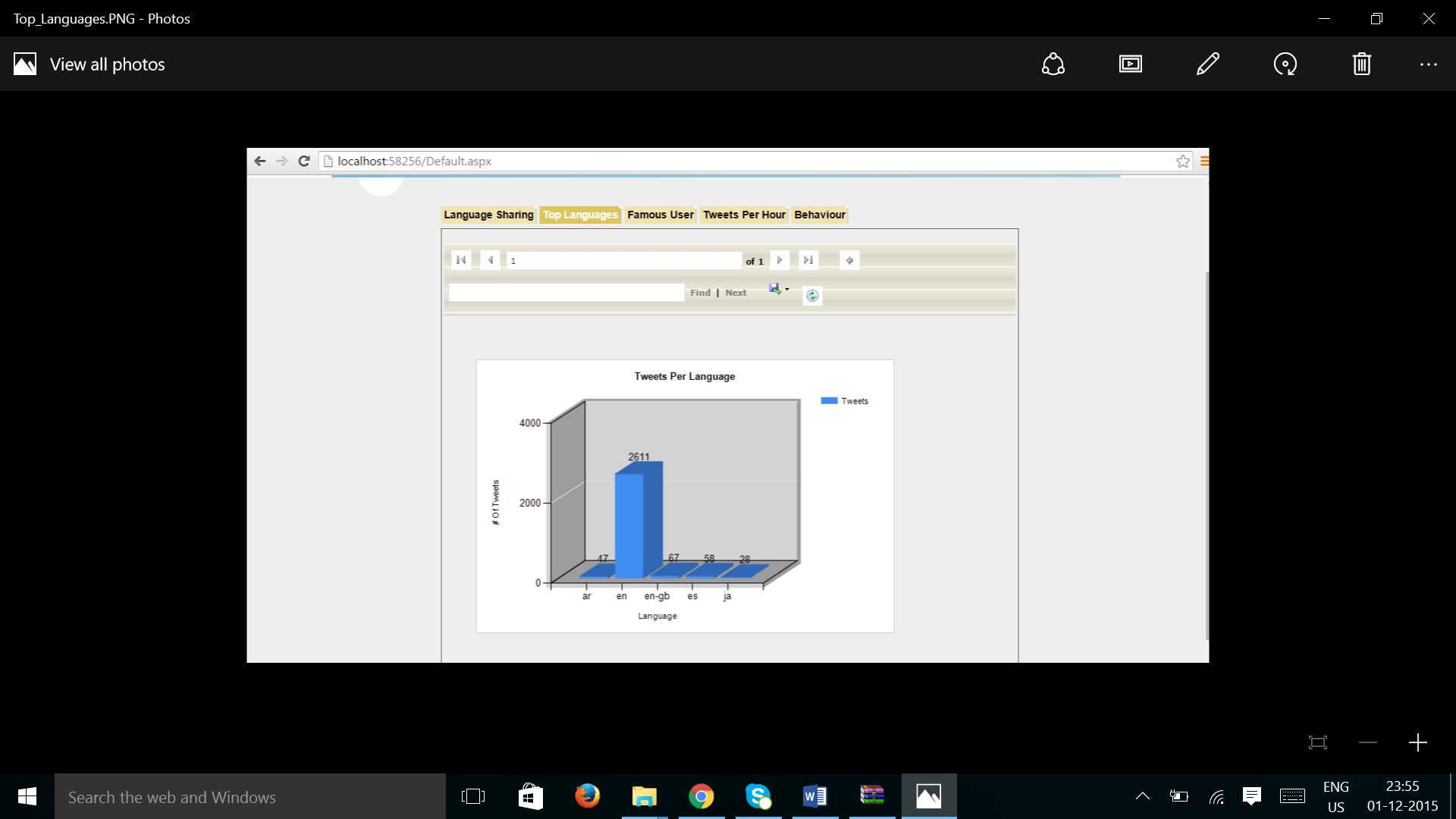
*A. First Query:*

Description:

We wrote the query to get the total number of tweets per language

Analysis:

We know which language comprises the number of tweets i.e., The Most used language for this stream of data

Visualization: 

*B. Second Query:*

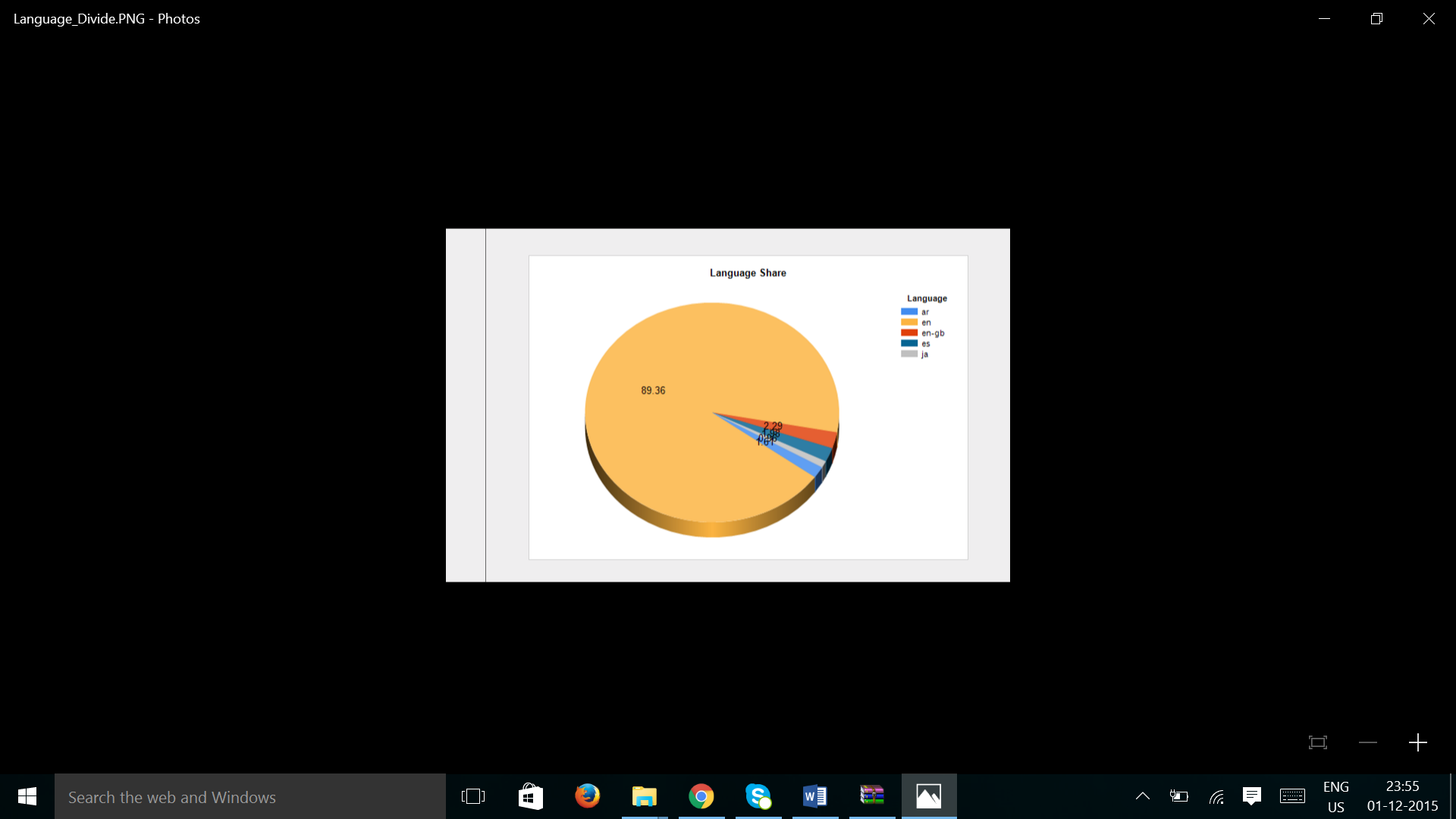
Description:

We wrote the query to get the Tweets Percentage of each language among the total languages. By using the existing column we have calculated the percentage per language in spark itself

Analysis:

We got the share of the tweets per each language among the total 100%

Visualization:



*C. Third Query:*

Description:

We implemented the query to get Users whose tweets are retweeted maximum number of times

Analysis:

We found the most influential personalities whose tweets are retweeted mostly in the discussion of Aamir Khan’s Intolerance topic

Visualization:

*D. Fourth Query:*

Description:

We implemented the query to get the total number of Hash Tags

Analysis:

We found all the Hash Tags that are involved in the tweets and how many times they are used in the discussion of Aamir Khan’s Intolerance topic

*E. Fifth Query:*

Description:

We implemented the query to get the top 20 Hashtags

Analysis:

We found the most interactive and used top 20 Hashtags as per the number of tweets

F. Sixth Query:

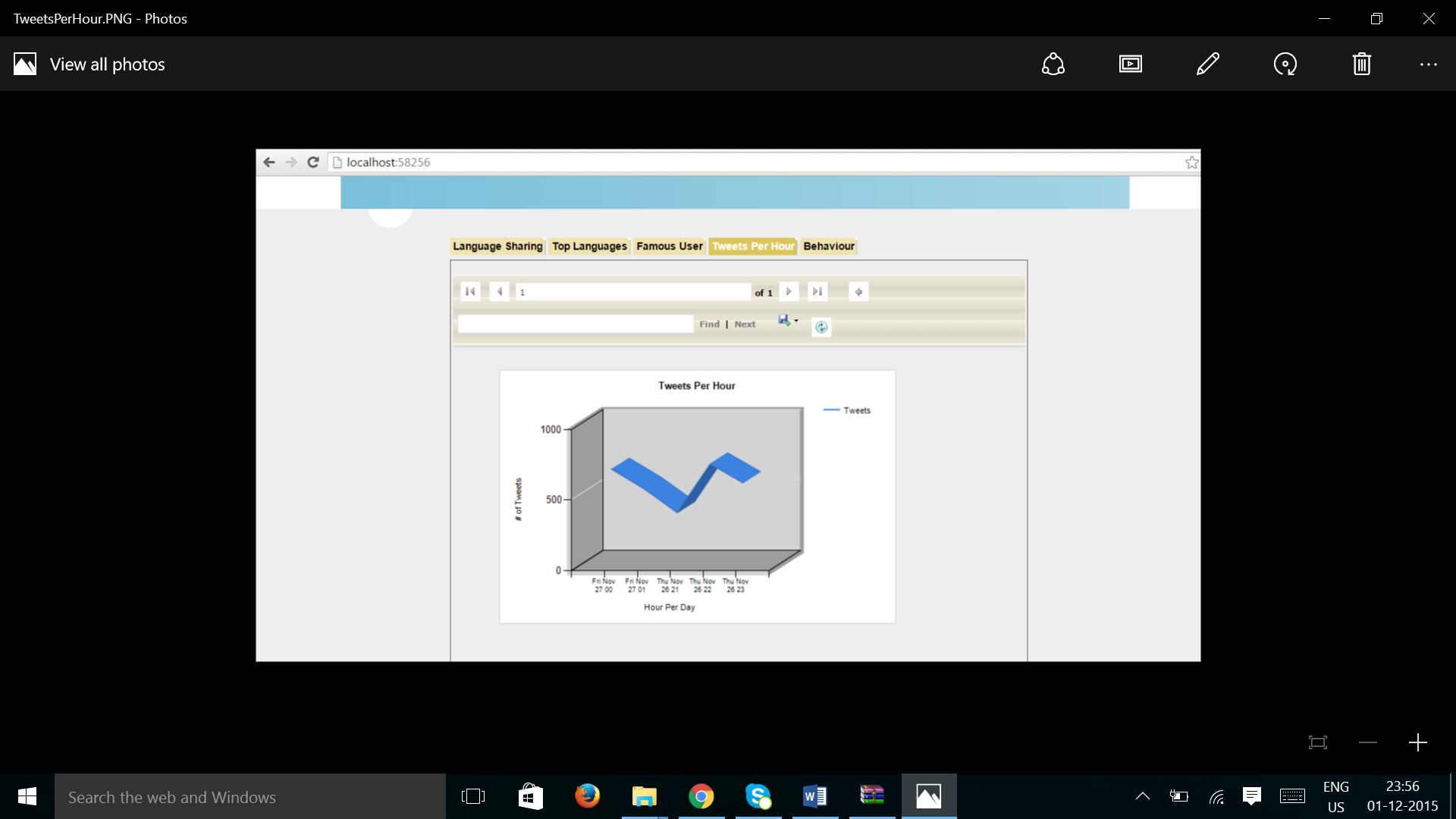
Description:

We implemented the query to represent the total number of tweets that are tweeted per hour

Analysis:

We come to know at what time the discussion on the topic of Aamir Khan’s Intolerance is at its prime/peak by observing the visualization

Visualization:



*G. Seventh Query:*

Description:

We implemented the query to represent the number of tweets per latitude and longitude

Analysis:

We found the tweets number from different latitudes and longitudes which are presented in different Geographic locations from the data

*H. Eighth Query:*

Description:

We implemented the query to represent how many tweets are supporting Aamir Khan, Opposing Aamir Khan, boot the snap deal app.

Analysis:

We found the total number of tweets that oppose, support Aamir Khan and how many are supporting to uninstall the Snap deal application. It is nothing but the reaction of the public for the Aamir Khan’s Behavior

Visualization:

**References:**

<http://spark.apache.org/sql/>

<http://www.scala-lang.org/api/current/#package>

<http://www.asp.net/>

<https://www.python.org/>