Contextual and Text Analysis of Nepal Tweets

Invicta
Analyse the world

Keerthi Yanda, Sandeep Ballu, Rakesh Vistarakula | Praveen Rao, Anas Katib | School of Computing And Engineering, University of Missouri – Kansas City

Objective

- Using the various tools in Hadoop framework, implement ETL process and visualize the extracted data in graphical format
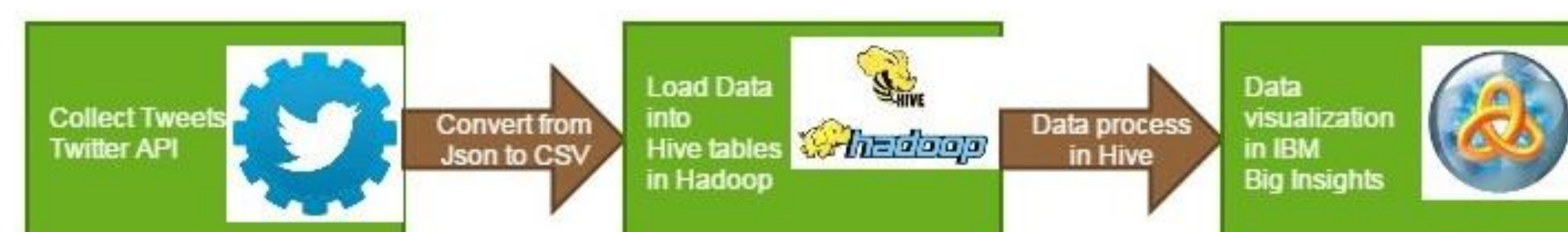
Motivation

- Usage of the unstructured data and implementing the big data tools to analyze the visualize the large amounts of structured and unstructured data is popularly in demand
- Social media data could be used to extract information
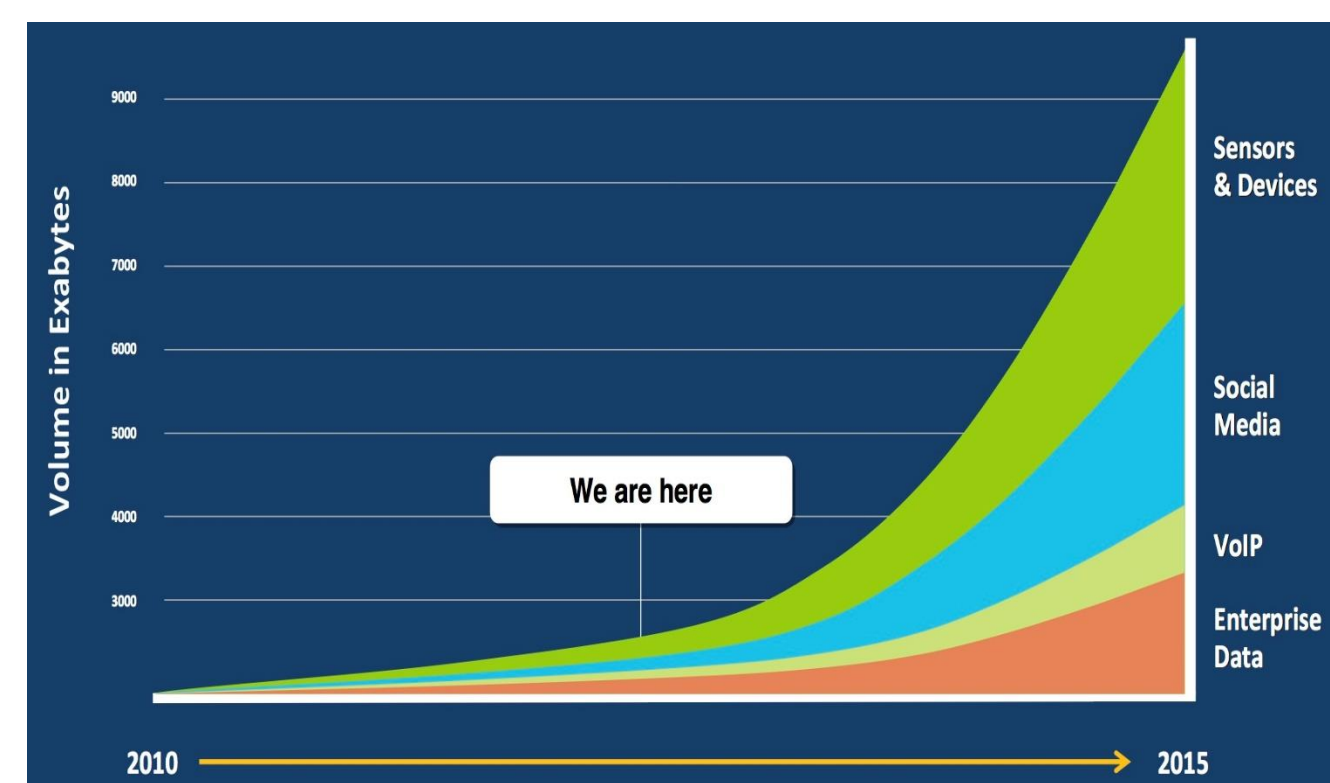
Project Overview

- Twitter is one of the social media. We collected tweets related to recent Nepal Earthquake using twitter API and used hive warehouse for ETL processing and IBM Big sheets for data visualization

Key Challenges

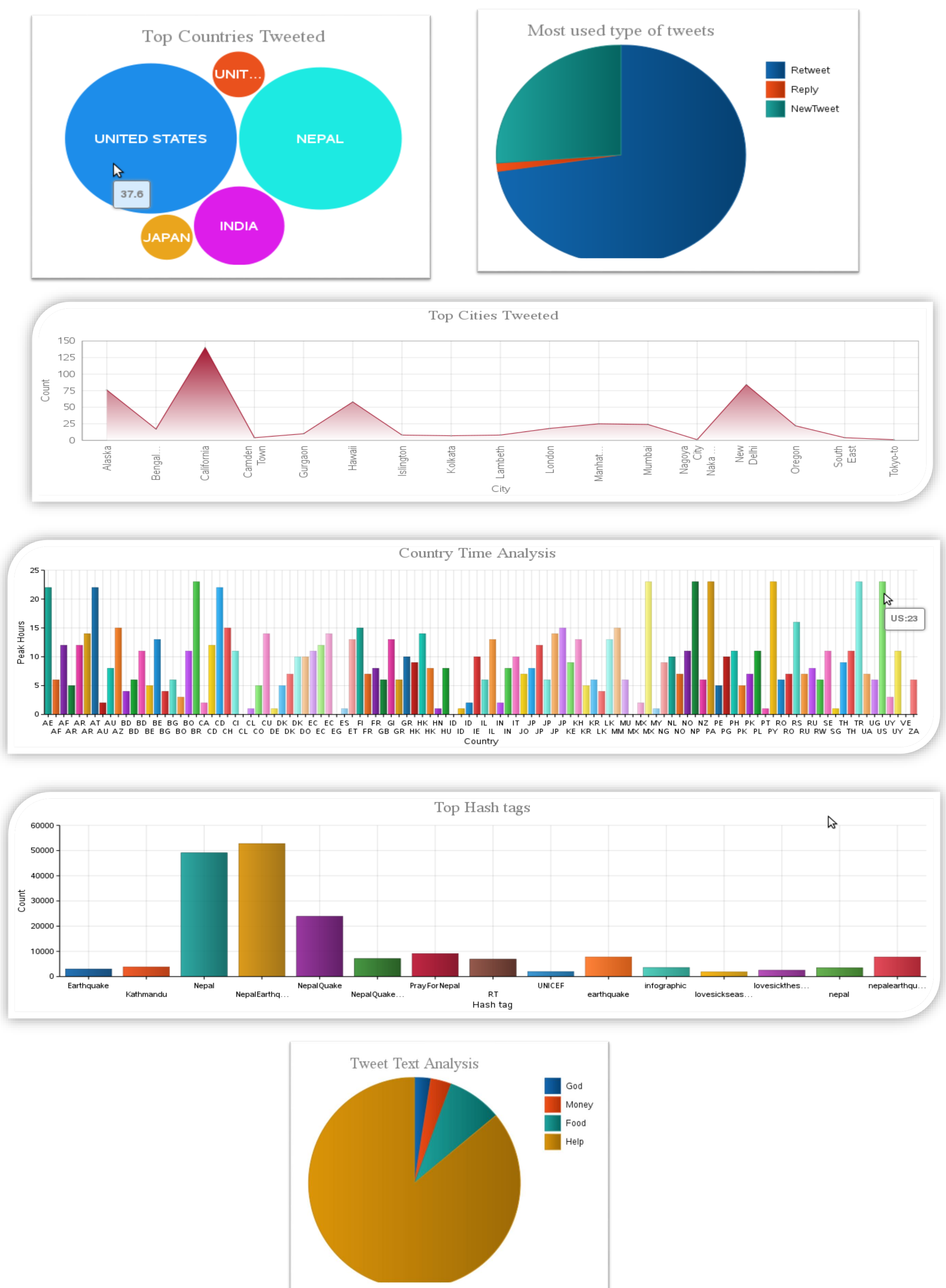| Structured to unstructured data	Data extraction	Visualization and deployment
• Data collected from twitter is in JSON, which cannot be directly loaded into database for feasible analysis	• Extracting the data from the Hive warehouse using the complex joins was time taking and challenging task	• Visualizing the analyzed data was challenging task. IBM Bigsheets made it easier • Deploying the entire project on cloud was challenging

Tools and Technologies

Tool	Purpose
Eclipse	Java programming
Twitter API	Twitter authorization
IBM Big Insights	Hadoop Framework
IBM Big Insights	Hive
IBM Big Sheets	Data visualization
JSP	User Interface

Data / Observations

- 1 GB of data is collected and analyzed
- **Contextual analysis includes**
  - Top 5 tweeted countries are United States, Nepal, India, UK and Japan. USA tweeted the most
  - Top cities tweeted are extracted. California being the top city
  - Most Active time of the users is extracted city based on city
  - Type of tweet is extracted, which included – retweet, reply, new tweet
- **Text analysis includes:**
  - Based on text, tweets are categorized into 4 types – god, money, food, help
  - Top Hashtags used for Nepal tweets #NepalEarthquake and #Nepal being the most used

Results

Conclusion

- contextual and text analysis of Tweets is done and visualized

Works Cited

- https://apps.twitter.com/app/7883650/show
- http://www-01.ibm.com/software/data/infosphere/hadoop/enterprise.html https://hadoop.apache.org/
- https://hive.apache.org/