Yet Another Twitter Sentiment Analyzer (YATSA)

Paul Durkin, Ye (Kevin) Pang, Laura Williams, Walt Burge, Matt Proetsch

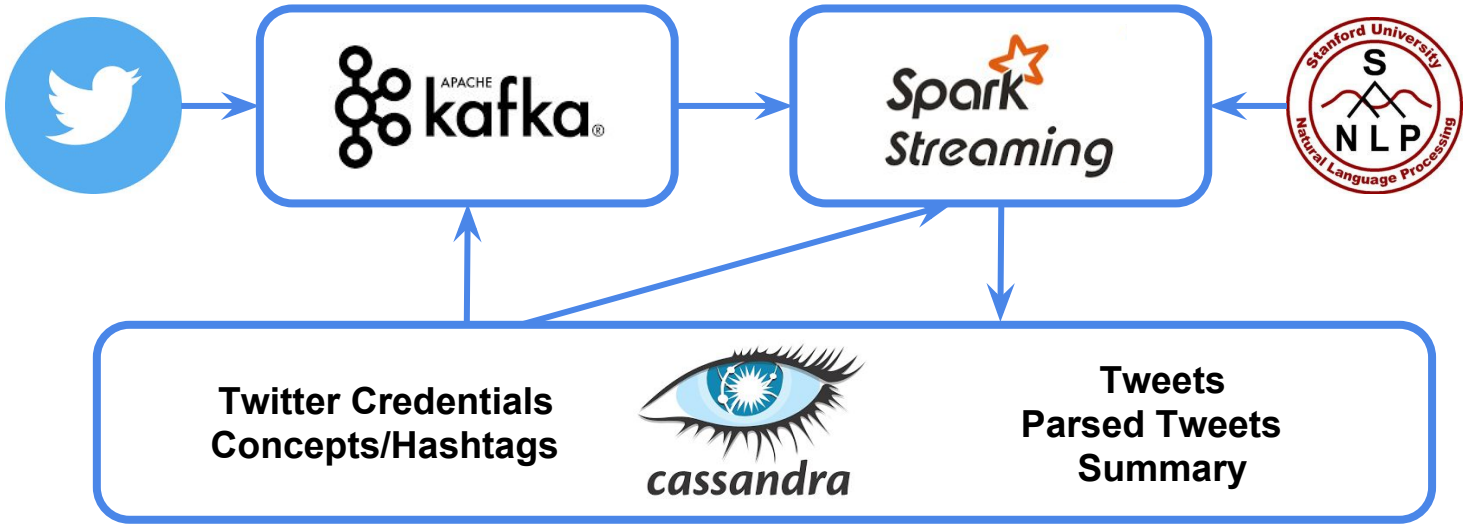
Introduction

- Problem : Need for insight on public response
  - specific, topical (e.g. current movies)
  - dynamic, fluid
  - varied perspective

- Solution : Yet Another Twitter Sentiment Analysis (YATSA)
  - Twitter messages: raw stream of expression
  - positive/negative statements
  - specific references

- Requirement: targeted sentiment analysis
  - Robust, scalable capture
  - trends over time
  - measuring sentiment for specific concepts (e.g. movies, products, politicians)

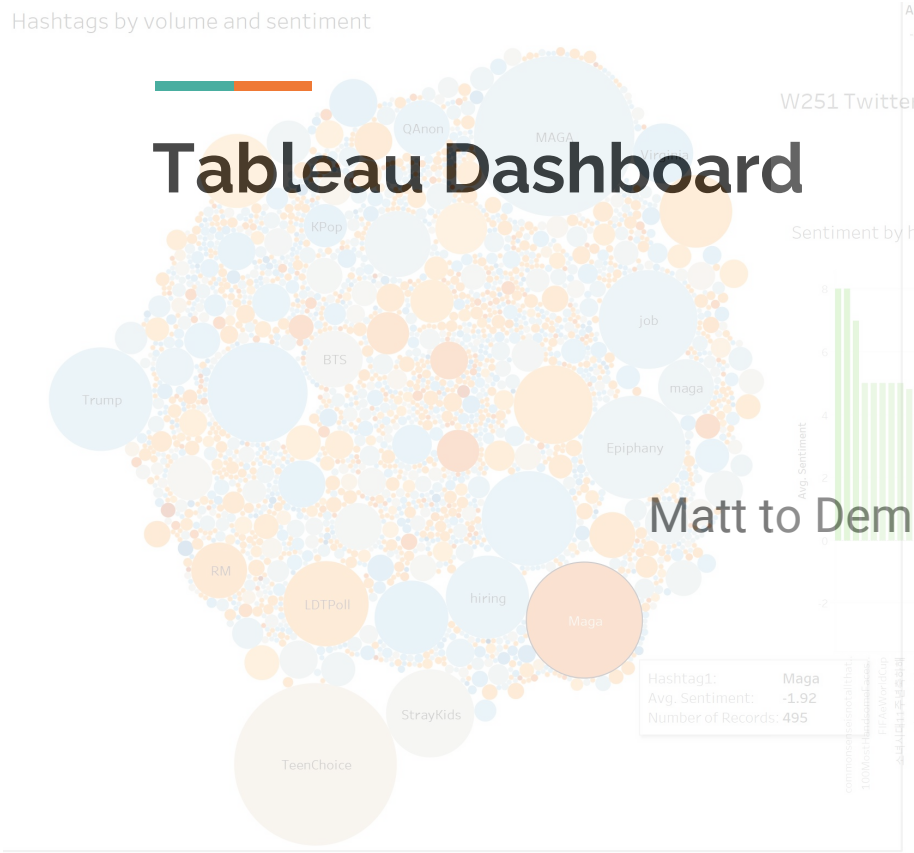
Data Flow


APACHE kafka
Spark Streaming
Stanford University
S
NLP
Natural Language Processing
Twitter Credentials
Concepts/Hashtags
cassandra
Tweets
Parsed Tweets
Summary

Sentiment Analysis

- Stanford CoreNLP - complex, recurrent Neural Net
  - Freely-available, self-contained software package providing a variety of NLP services
  - Sentiment scores: Positive/Negative, 0 <= score <= 3
  - REST API
  - Built in, extensible language models providing additional control

Tableau Dashboard



Matt to Demo During Class!

Backend: Kafka, Spark, Cassandra

- Kafka (0.8.2)
  - decoupled message queue
  - replay messages according to retention settings
  - zookeeper (coordinator) + kafka brokers (topic pub/sub service)
  - additional brokers can be added per demand

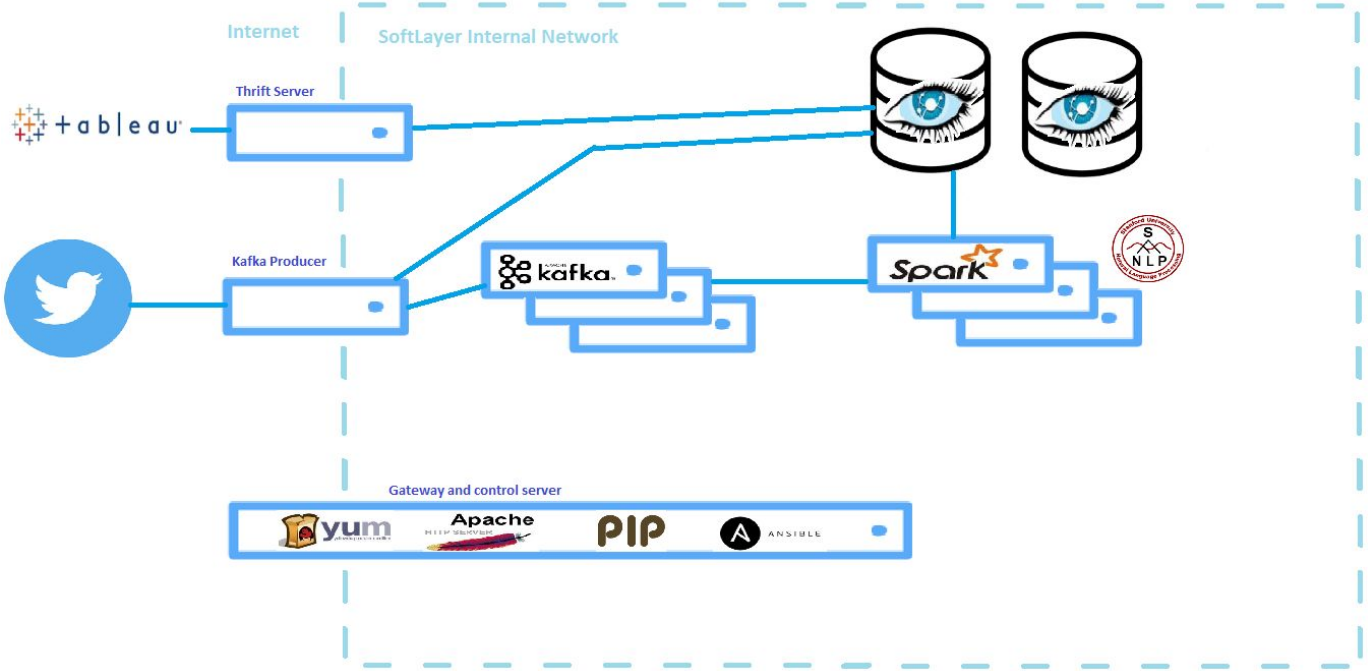
- Spark (2.3.1)
  - continuous micro batch processing, a.k.a. stream processing
  - standalone mode (no resource manager)
  - dynamic allocation and the external shuffle; workers can be added and removed dynamically

- Cassandra (3.11.2)
  - availability and partition tolerance
  - 3 nodes, each node needs 2 disks server (OS + Data Storage); add additional nodes when necessary
  - Simple Replication Strategy (use next highest token value node)

Backend: Additional Mentions

- Tableau is a killer for creating visual dashboards – easy, fast, and self-service.
- Apache Thrift is the technology underlying a middleware layer handling communication between BI tools like Tableau and our backing datastore written for Cassandra.
  - Uses Hive metastore to expose data to external clients
  - When a SQL query is sent to the Thrift server requesting data from a table in the metastore, a Spark job is launched to query the backing datastore in order to return the data to the client.
- Stanford CoreNLP the server will be installed on each Spark worker

Infrastructure



Configuration with Ansible

- Popular configuration management tool
- Agentless works over ssh
- Has PlayBooks and Roles
- Also using to update all hosts files when servers added or removed
- Supported by:
  - Local Yum repo of Epel
  - Local Pip repo
  - Local Apache http webserver for software packages
- Using hostname aliases so that servers can be reprovisioned without needing to reconfigure, only host entries updated
- Using playbooks to decommission Spark and Cassandra

Questions?