Single-Node Apache Spark + Twitter

Jason B. Hill (@hilljb)
Working with Data & Analytics @ CA Technologies

October 12, 2016

Agenda

1. Light Introduction to the (public/free) Twitter APIs
2. Sports Data (2013 NCAAB championship, Superbowl XLVII, World Cup)
3. Building a Query Engine on top of Apache Spark
4. October 9 US Presidential Debate on Twitter + NLP + Machine Learning

Twitter APIs

Twitter: The (free) public APIs

REST API

- Generally: 15 requests per 15 minutes, each limited in num of responses
- User info, tweets, followers, searches, lots more...

Streaming APIs

- 1% sample stream
- Filter stream

Twitter: The Filtered Streaming API

Some Basic Details:

- Track users or terms
- Rate limited (usually 3K tweets per minute, sometimes more)
- Can stay connected for months
- JSON, gzipped
- Caveat: Many times a line response isn’t an entire JSON entity
- Caveat: Rate limit responses track total undelivered tweets since connection

Twitter: How to Get Access

- Twitter uses OAuth
  - Tokens do not expire
  - Apps are authenticated via a signature in an http request
  - Python packages: requests + requests_oauthlib
  - github.com/bear/python-twitter
- Log in to Twitter and go to apps.twitter.com
- For documentation and resources: dev.twitter.com

Sports Examples

Example 1: NCAA March Hashtag Madness

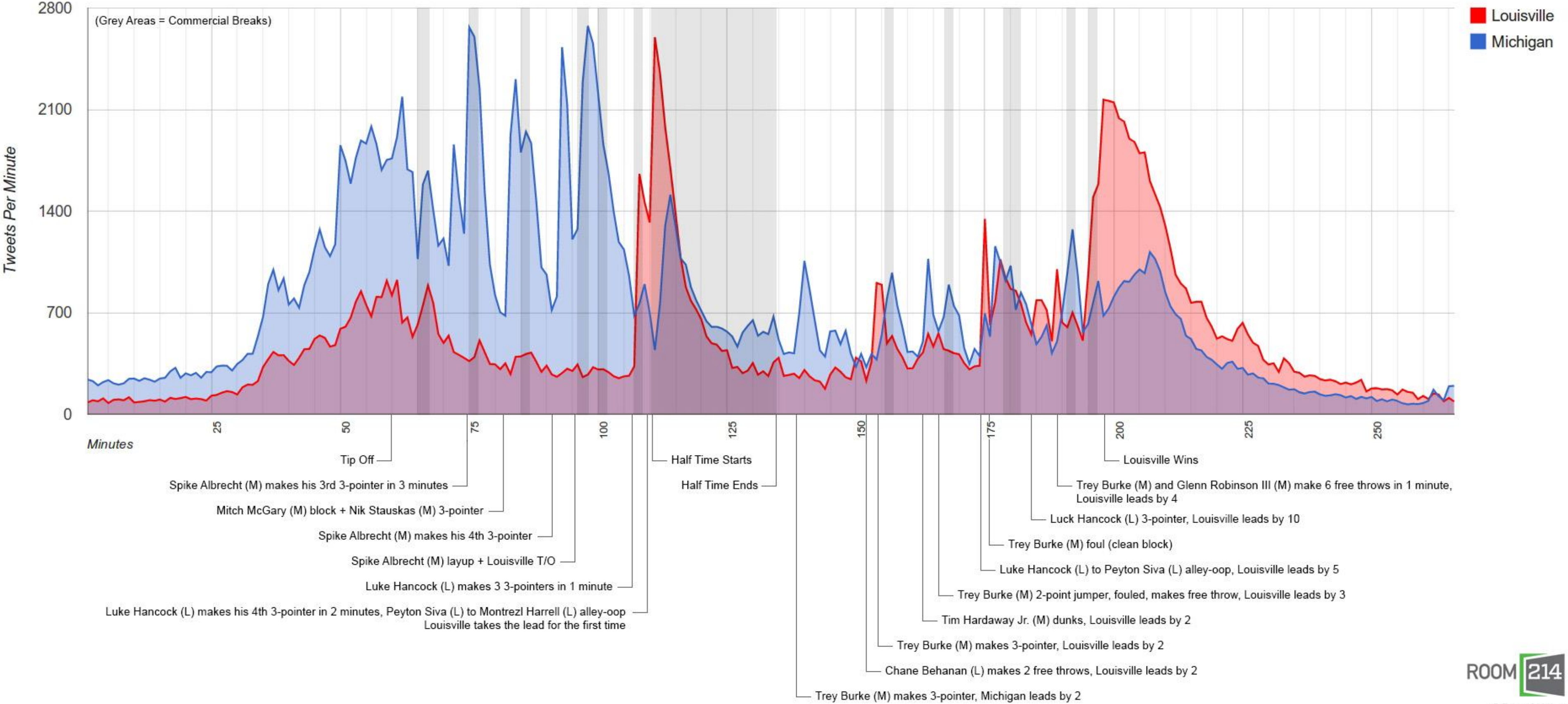
In 2013, a media agency client wanted to know:

- “How much data can we collect for free to analyze ourselves?”
- “Can we get minute-by-minute granularity?”

We decided to analyze the NCAA Tournament

- We used the filtered streaming API
- Every school, every game, every minute

2013 NCAA Championship: #1 Louisville (#l1c4, #uofl, #louisville) vs. #4 Michigan (#goblue, #michigan), Hashtagged Tweets Per Minute



Tip Off

Spike Albrecht (M) makes his 3rd 3-pointer in 3 minutes

Mitch McGary (M) block + Nik Stauskas (M) 3-pointer

Spike Albrecht (M) makes his 4th 3-pointer

Spike Albrecht (M) layup + Louisville T/O

Luke Hancock (L) makes 3 3-pointers in 1 minute

Luke Hancock (L) makes his 4th 3-pointer in 2 minutes, Peyton Siva (L) to Montrezl Harrell (L) alley-oop
Louisville takes the lead for the first time

Half Time Starts

Half Time Ends

Louisville Wins

Trey Burke (M) and Glenn Robinson III (M) make 6 free throws in 1 minute,
Louisville leads by 4

Luck Hancock (L) 3-pointer, Louisville leads by 10

Trey Burke (M) foul (clean block)

Luke Hancock (L) to Peyton Siva (L) alley-oop, Louisville leads by 5

Trey Burke (M) 2-point jumper, fouled, makes free throw, Louisville leads by 3

Tim Hardaway Jr. (M) dunks, Louisville leads by 2

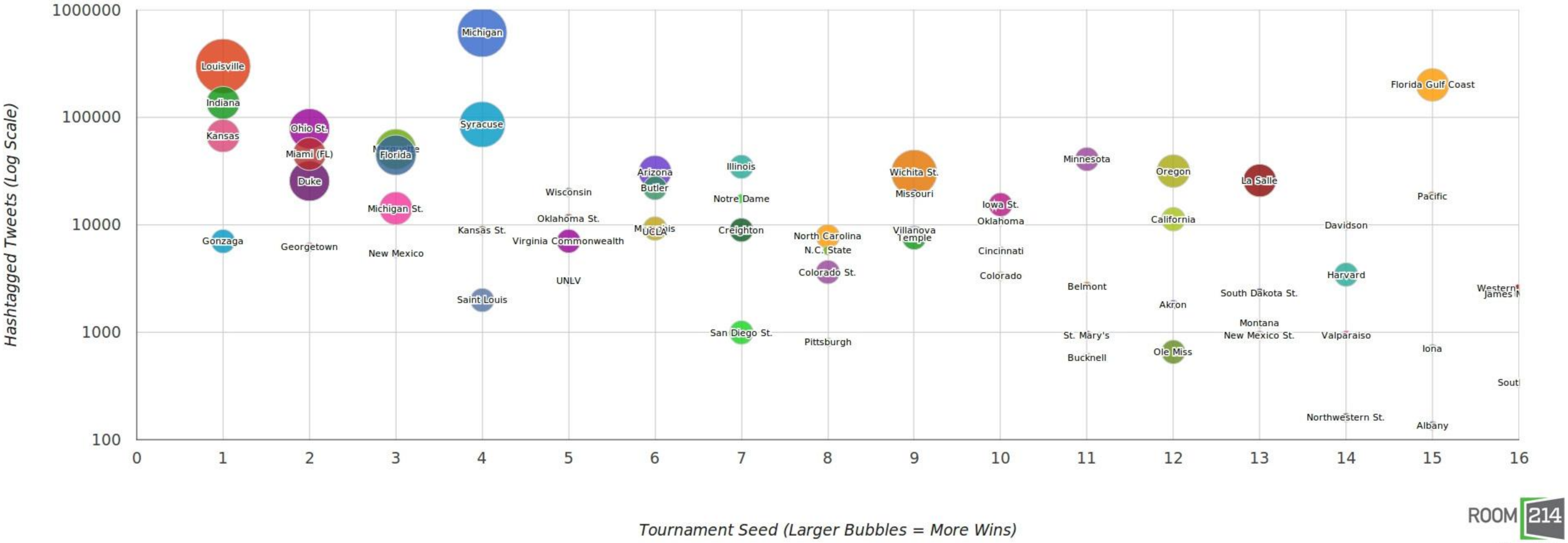
Trey Burke (M) makes 3-pointer, Louisville leads by 2

Chane Behanan (L) makes 2 free throws, Louisville leads by 2

Trey Burke (M) makes 3-pointer, Michigan leads by 2

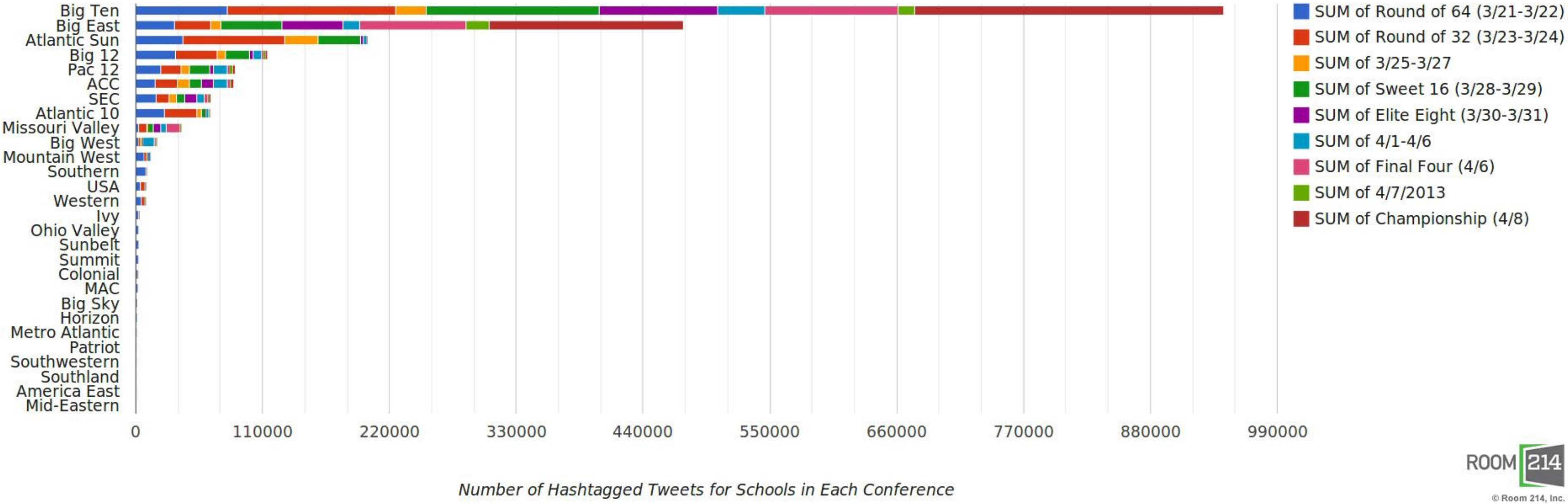
ROOM 214

Tournament Seed vs. Wins vs. Hashtagged Tweets on Twitter During the 2013 NCAA Tournament

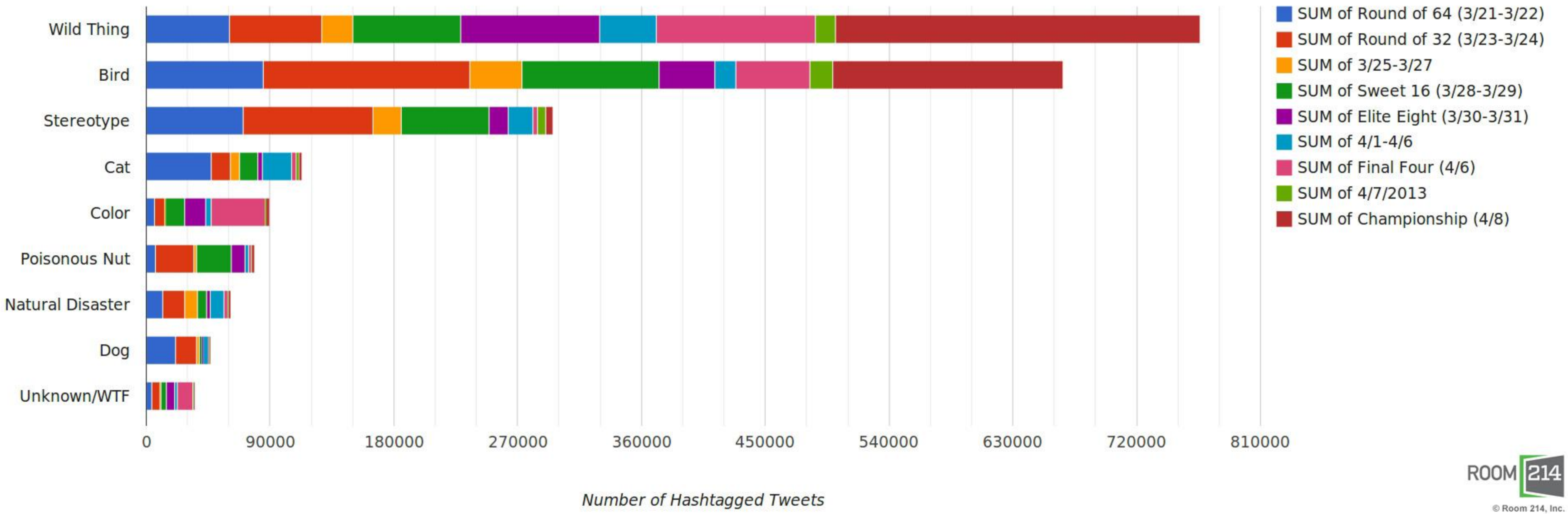


ROOM 214

Hashtag Use on Twitter by Athletic Conference During the 2013 NCAA Tournament



Hashtagged Tournament Tweets vs. Mascot Type During the 2013 NCAA Tournament



ROOM 214

Example 2: Super Bowl XLVII

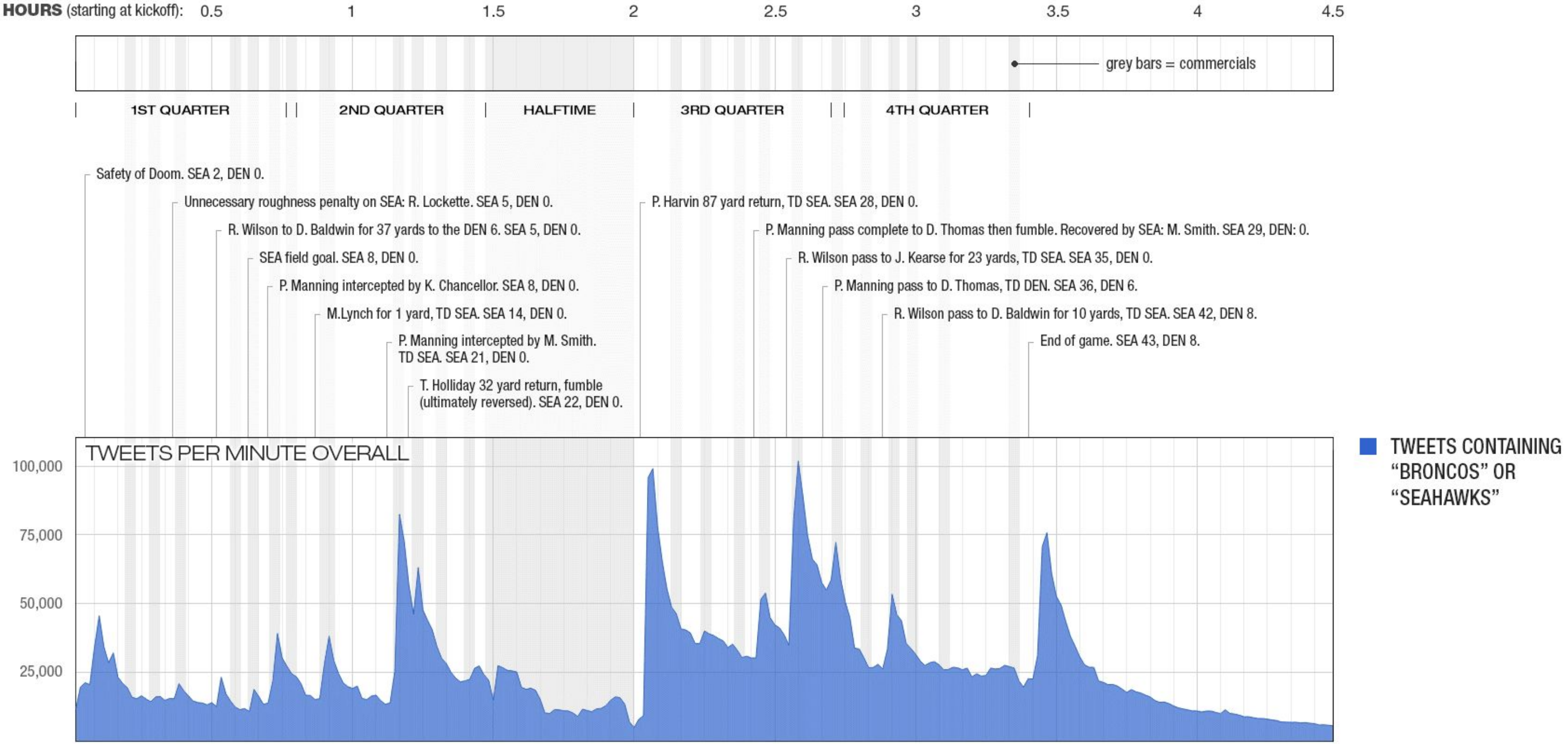
In 2014, the Denver Broncos lost to Seattle:

- 22-0 at halftime, 43-8 overall

We collected tweets mentioning both teams:

- 4,109,946 tweets containing “Broncos”
- 2,502,952 tweets containing “Seahawks”

Twitter During the 2014 Super Bowl



Safety of Doom. SEA 2, DEN 0.

Unnecessary roughness penalty on SEA: R. Lockette. SEA 5, DEN 0.

R. Wilson to D. Baldwin for 37 yards to the DEN 6. SEA 5, DEN 0.

SEA field goal. SEA 8, DEN 0.

P. Manning intercepted by K. Chancellor. SEA 8, DEN 0.

M.Lynch for 1 yard, TD SEA. SEA 14, DEN 0.

P. Manning intercepted by M. Smith. TD SEA. SEA 21, DEN 0.

T. Holliday 32 yard return, fumble (ultimately reversed). SEA 22, DEN 0.

P. Harvin 87 yard return, TD SEA. SEA 28, DEN 0.

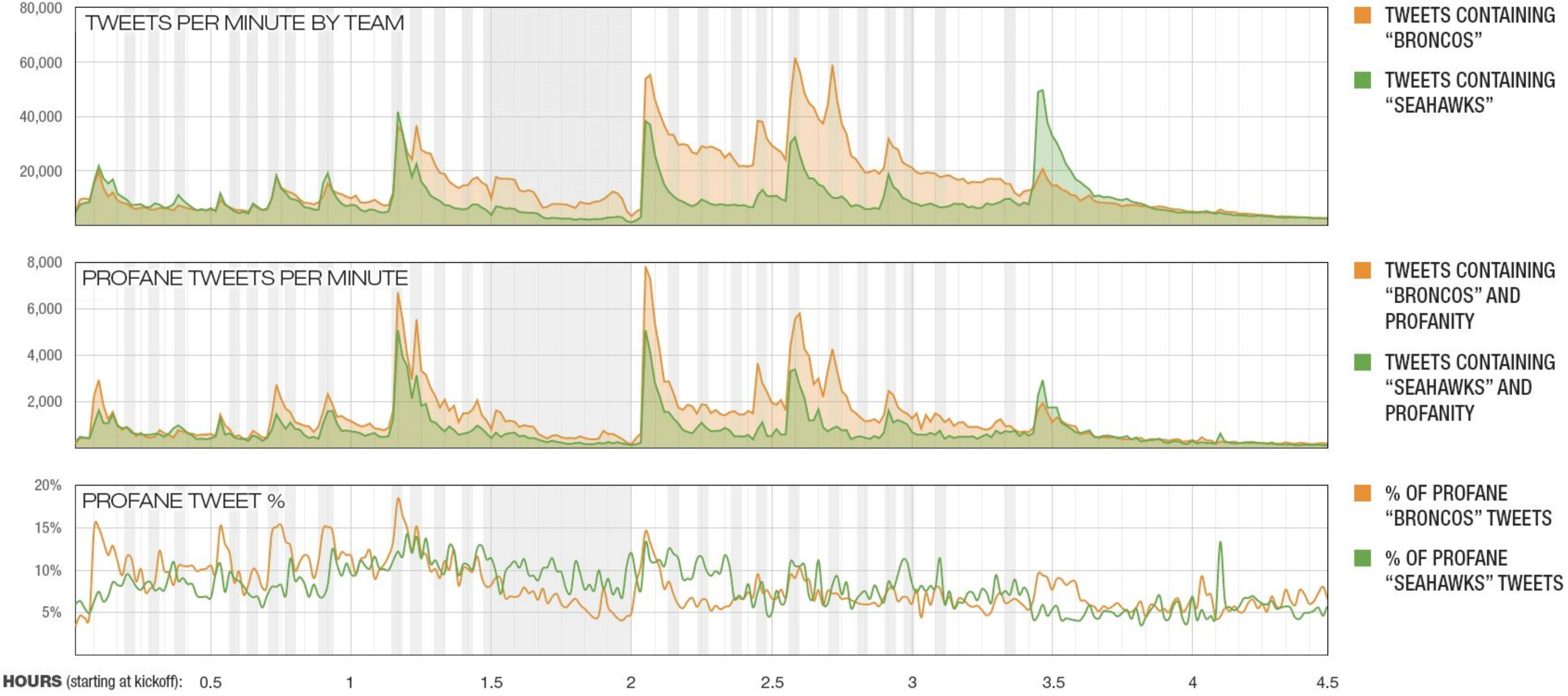
P. Manning pass complete to D. Thomas then fumble. Recovered by SEA: M. Smith. SEA 29, DEN: 0.

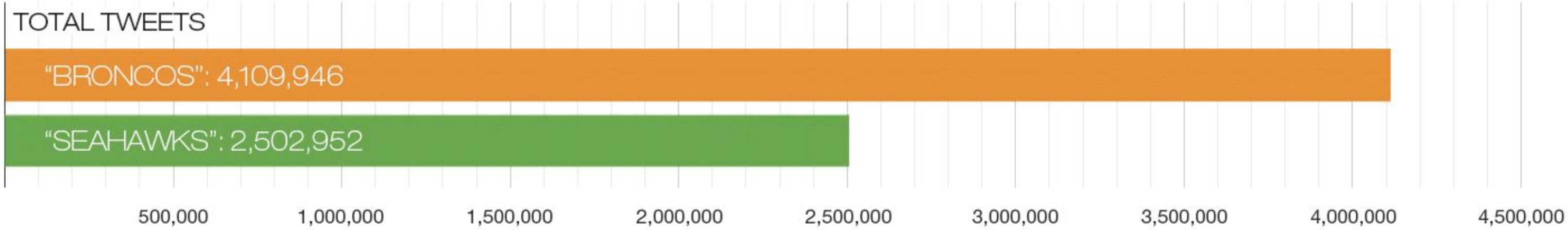
R. Wilson pass to J. Kearse for 23 yards, TD SEA. SEA 35, DEN 0.

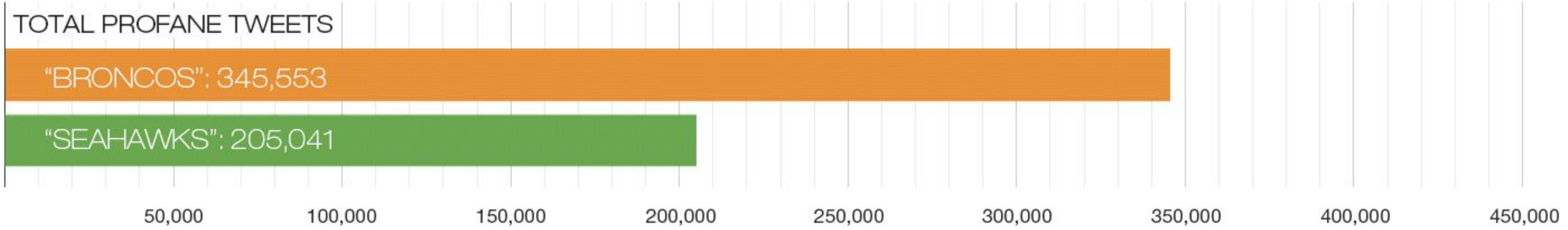
P. Manning pass to D. Thomas, TD DEN. SEA 36, DEN 6.

R. Wilson pass to D. Baldwin for 10 yards, TD SEA. SEA 42, DEN 8.

End of game. SEA 43, DEN 8.


80,000
60,000
40,000
20,000
TWEETS PER MINUTE BY TEAM
🟧 TWEETS CONTAINING "BRONCOS"
🟩 TWEETS CONTAINING "SEAHAWKS"
8,000
6,000
4,000
2,000
PROFANE TWEETS PER MINUTE
🟧 TWEETS CONTAINING "BRONCOS" AND PROFANITY
🟩 TWEETS CONTAINING "SEAHAWKS" AND PROFANITY
20%
15%
10%
5%
PROFANE TWEET %
🟧 % OF PROFANE "BRONCOS" TWEETS
🟩 % OF PROFANE "SEAHAWKS" TWEETS
HOURS (starting at kickoff):
0.5
1
1.5
2
2.5
3
3.5
4
4.5





PERCENTAGE OF PROFANE TWEETS:

"BRONCOS": **7.88%**
"SEAHAWKS": **7.68%**

Example 3: Gooooooaaaaalllll!!!!!!1

The 2014 World Cup on Twitter:

- 400 GiB of gzipped data
- 100 million (hashtagged) tweets
- The streaming API fed at rates over 30K tweets per minute
- Very few rate limit responses

goal

2,352,714 tweets

GOAL

522,353 tweets

Goal

505,513 tweets

GOAL!

125,084 tweets

goal!

52,194 tweets

goall

34,445 tweets

Goal!

16,868 tweets

GOAL!!!

11,034 tweets

Goall

5,947 tweets

goal!!

5,840 tweets

goal!!!

5,399 tweets

gooooaaaallllllll!!!!!

5,127 tweets

How many different spellings of “goal”?

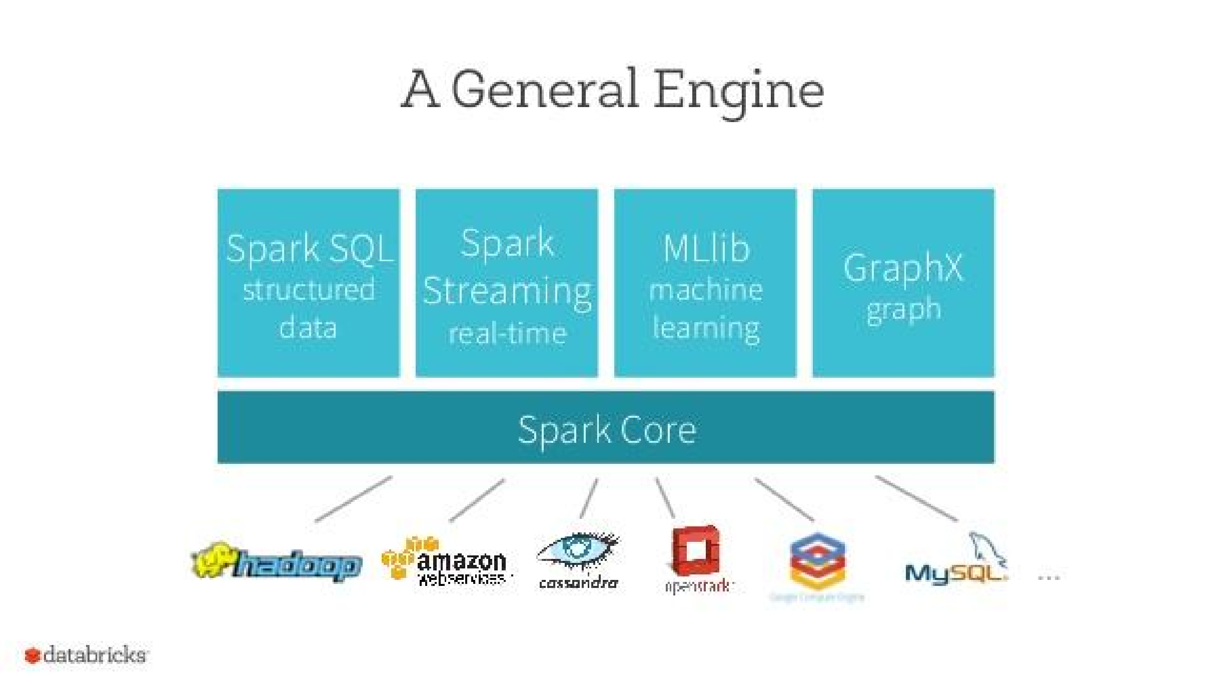
- 3,997,679 tweets contained some variant of “goal”
- 22,430 distinct spellings
- 12,531 (55.9%) spellings were only tweeted once

Building a Query Engine on top of Apache Spark

For data analysis, it would be nice to ...

- Query semi-structured data (e.g., JSON) with SQL.
- Use Python or Scala to access JVM libraries.
- Use Python data modules on top of JVM processes.
- Work in a notebook environment (e.g., Jupyter).
- Do machine learning concurrently.

Apache Spark Components:



A General Engine

Spark SQL
structured data

Spark Streaming
real-time

MLlib
machine learning

GraphX
graph

Spark Core

Hackathon is Amazing!

During a week-long hackathon at Rally/CA, we…

- Used an Anaconda env to run PySpark
- Loaded that Anaconda env in the Jupyter notebook
- Analyzed our JSON and Avro data from Kafka and S3
- Used existing Clojure code as SQL UDFs from Python

Developing this environment became a quarter goal.

October 9 Presidential Debate on Twitter + NLP + ML