TWITTER ANALYTIC WITH SPARK

HPC final project proposal

Voyager

Yuxing Wang
Zhenying Tao
Jing Cui

Twitter analytic with Spark

Motivation:

- Stream processing framework need to be:
  - ✓Scalable to large clusters
  - ✓Achieves second-scale latencies
  - ✓Integrates with batch & interactive workloads
  - ✓Ensures efficient & fault-tolerance
  - ✓Has simple programming model

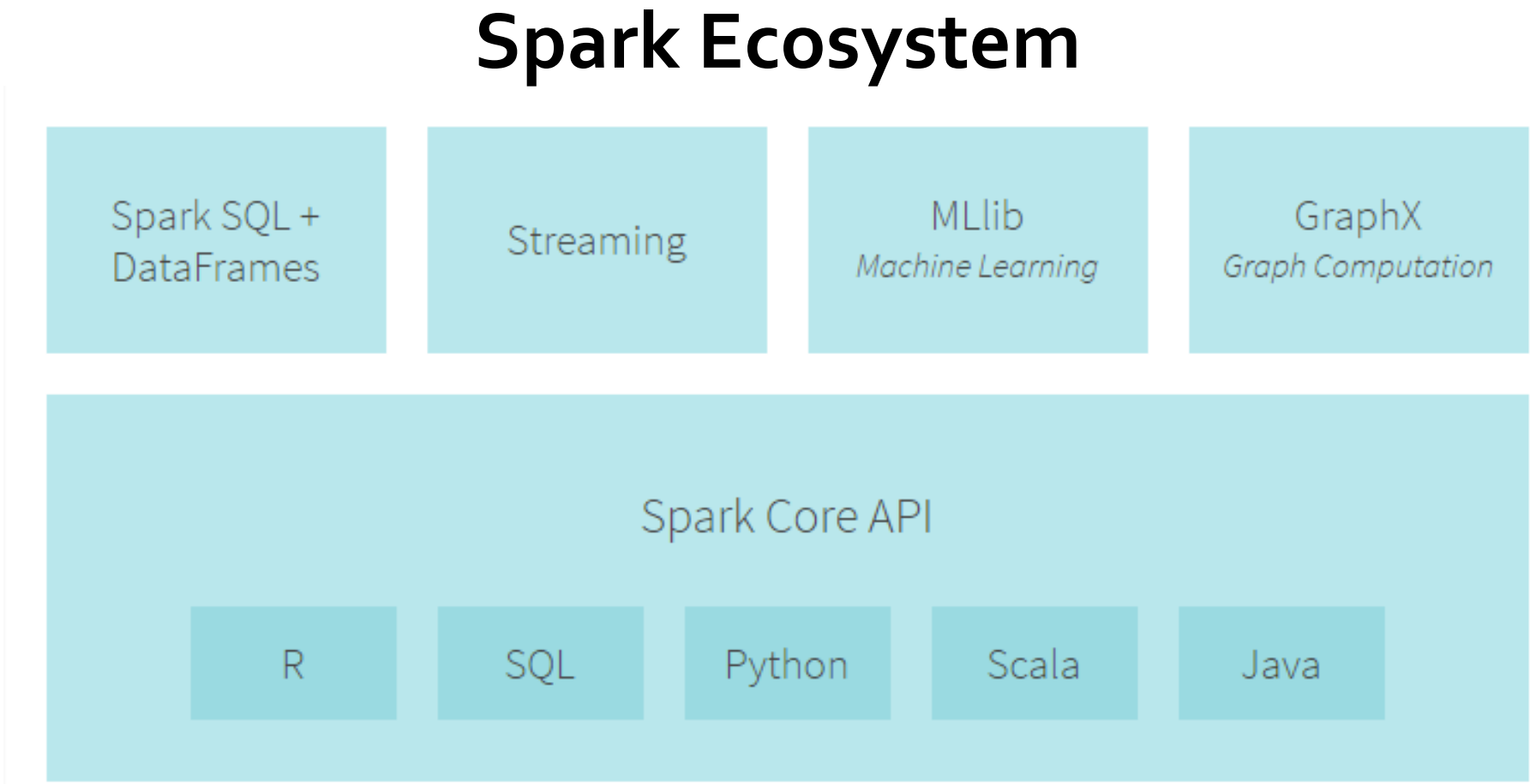
Twitter analytic with Spark

Latest version: 1.6.1

- Apache Spark is an open source cluster computing **framework**
- Fast, expressive cluster computing system compatible with Apache Hadoop
  - Works with any Hadoop-supported storage system (HDFS, S3, Avro, …)
- Improves **efficiency** through:
  - In-memory computing primitives
  - General computation graphs

  → Up to 100× faster
- Improves **usability** through:
  - Rich APIs in Java, Scala, Python
  - Interactive shell in Scala and Python

Twitter analytic with Spark



How to get familiar with Spark API

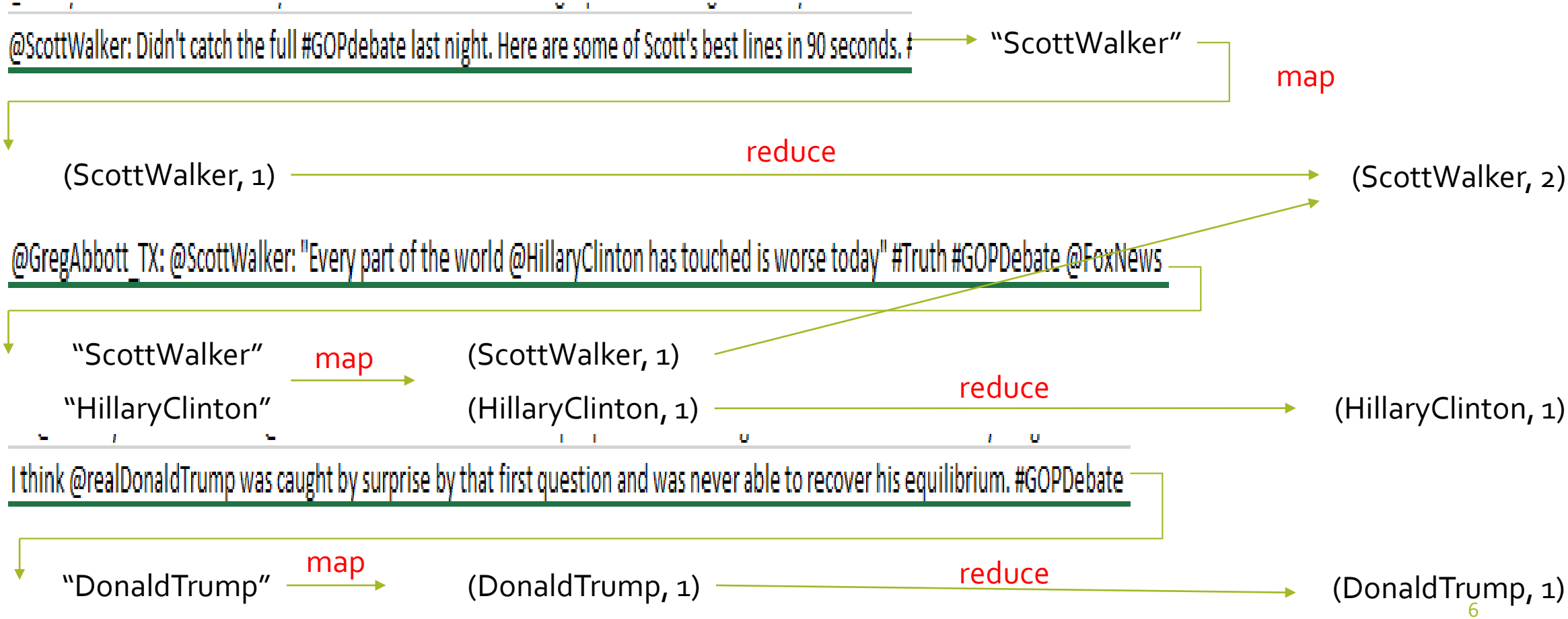
- **First GOP debate sentiment analysis dataset**

candidate	candidate:confidence	relevant_yn	relevant_yn:confidence	sentiment	sentiment:confidence	subject_matter	subject_matter:confidence
No candidate mentioned	1	yes	1	Neutral	0.6578	None of the above	1
Scott Walker	1	yes	1	Positive	0.6333	None of the above	1
No candidate mentioned	1	yes	1	Neutral	0.6629	None of the above	0.6629

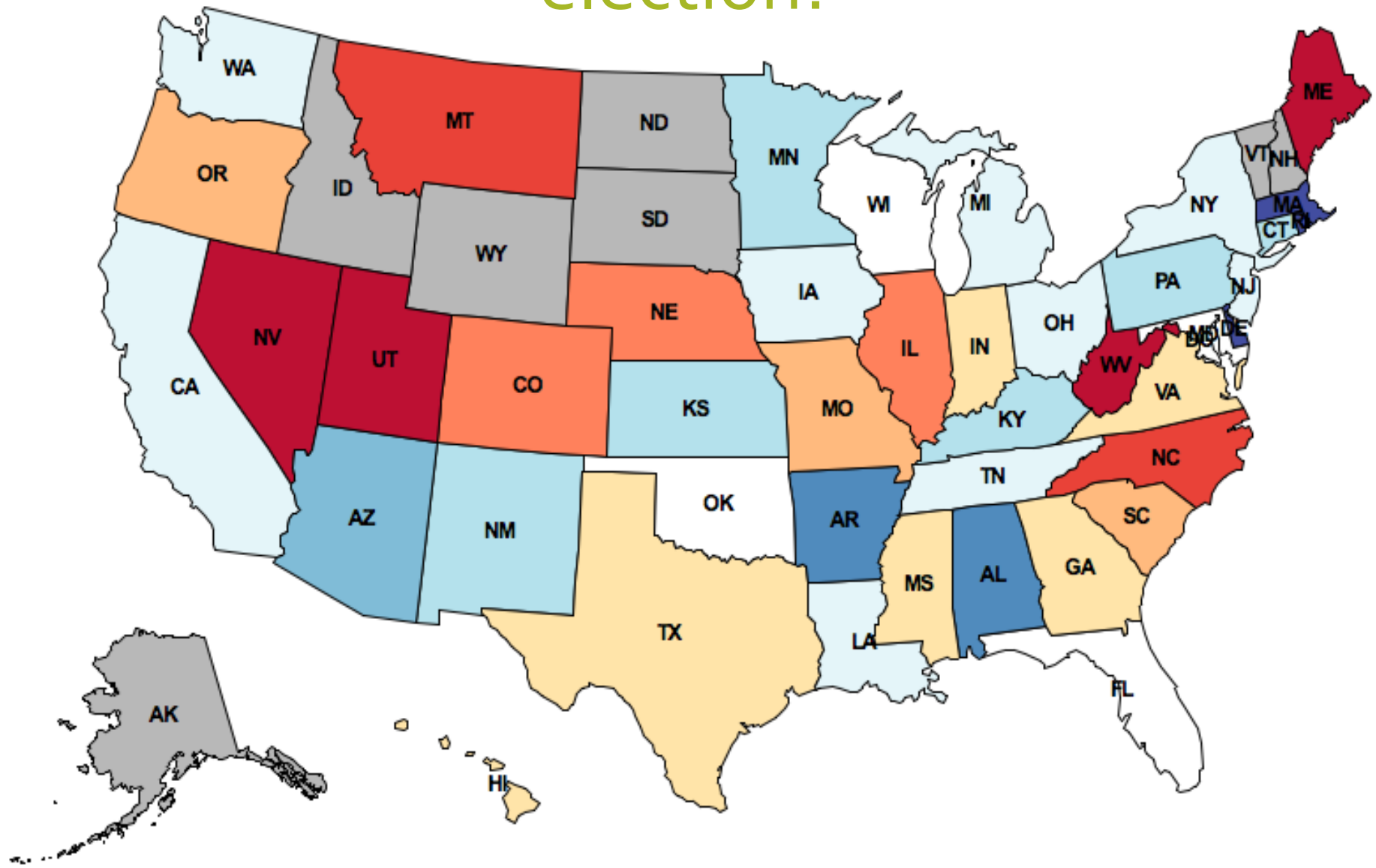
#tag

retweet_count	sentimen	subject_m	text
5			RT @NancyLeeGrahn: How did everyone feel about the Climate Change question last night? Exactly #GOPDebate
26			RT @ScottWalker: Didn't catch the full #GOPdebate last night. Here are some of Scott's best lines in 90 seconds. #Walker16 http://t.co/ZSfFâ€¦
27			RT @TJMShow: No mention of Tamir Rice and the #GOPDebate was held in Cleveland? Wow.
138			RT @RobGeorge: That Carly Fiorina is trending -- hours after HER debate -- above any of the men in just-completed #GOPdebate says she's on â€¦

Eample: Candidate Mentions



Twitter Trends: Who would win the election?



- Use Twitter Streaming API to get real time tweets.
- Use Spark Streaming to analyze the real time tweets
  - What is the percentage of tweets are talking about “Hillary Clinton”, “Donald Trump”, “Ted Cruz”;
  - Where are these tweets from;
  - Word cloud;
- Assigning a sentiment (positive or negative) to each tweet, based on all of the words it contains (MLlib)
- Aggregating tweets by the state and coloring the graph
- Other interesting results found during data mining

Goals for this project

- Play with Twitter API
- Learn Spark Ecosystem
- Play with spark streaming, Mllib, GraphX

Time line

- Setup Spark working environment (Cloudera? AWS? Azure?) now - 3.25
- Play with basic Spark API, implement basic functions like word count. 3.26 – 4. 15
- Twitter trends. 4.16 – end of semester

Thank You