

TwitterViz

Jonathan Cobian & Maribeth Rauh

Objective & End Goal

- Analyze the content of tweets about major events over time
 - Hashtags, keywords, images, links
- Spread of location over time
 - How tweets about an event spread and differ over time
- Allow user to analyze the tweets about an event at a deeper level

Twitter Data

- Apollo Project, University of Notre Dame
- Collect Tweets from Twitter in real time that match certain keywords or location preferences



Create new task

Keyword 1 or

Keyword 2 or

Keyword 3 or from

Latitude Longitude Radius (miles)

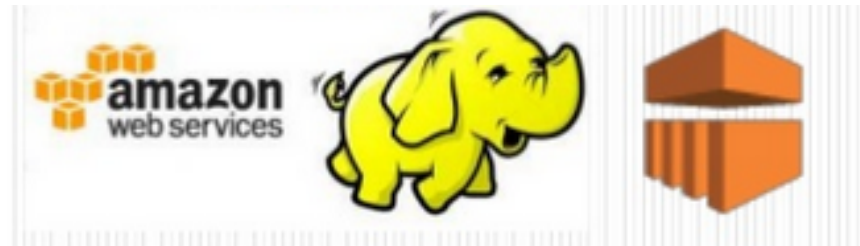
The map shows the Middle East and surrounding regions, including Europe, Africa, and Asia. A red location pin is placed over Egypt, with a blue circle around it. The map is labeled with various countries and regions, including France, Spain, Italy, Greece, Turkey, Iraq, Iran, Afghanistan, Pakistan, India, and others. The map is titled "Timeline Recovery of Real-World Events" and includes a copyright notice for 2011.

Timeline Recovery of Real-World Events

Map data ©2011 Europa Technologies, Google, DeLorme, INEGI, MapLink, Tele Atlas - Terms of Use

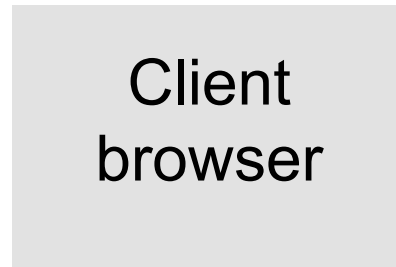
Map-Reduce -> MySQL

- Run MR jobs on AWS Elastic Map Reduce
- Generate aggregate data
 - Ex: Top hashtags segmented by time range
- Place into MySQL Database

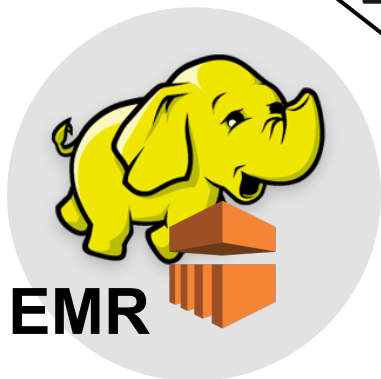


Top #s Table Example

Hashtag	Start Range	End Range	Rank	NumHashtags
#ukraine	2014-02-19 00:00:00	2014-02-25 23:59:59	1	4981
#euromaidan	2014-02-19 00:00:00	2014-02-25 23:59:59	2	2718
#help	2014-03-06 00:00:00	2014-03-12 23:59:59	1	1832
#uarevolution	2014-03-06 00:00:00	2014-03-12 23:59:59	2	1621



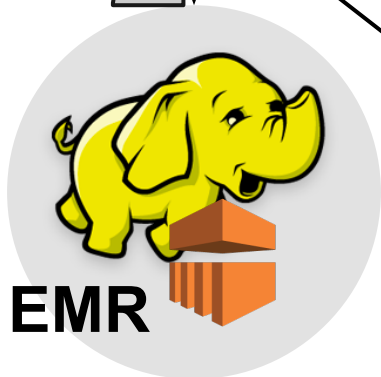
1



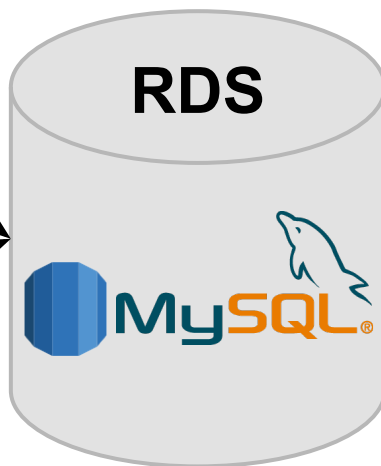
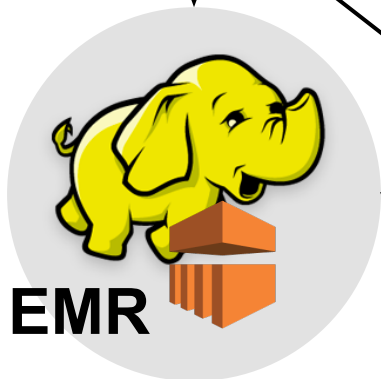
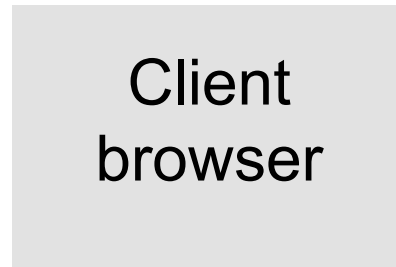
Data Flow



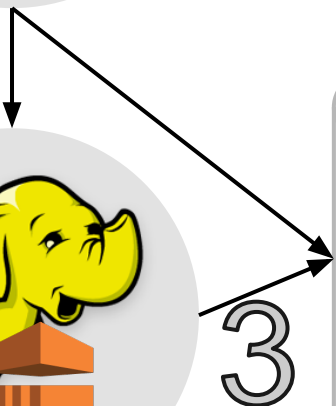
2

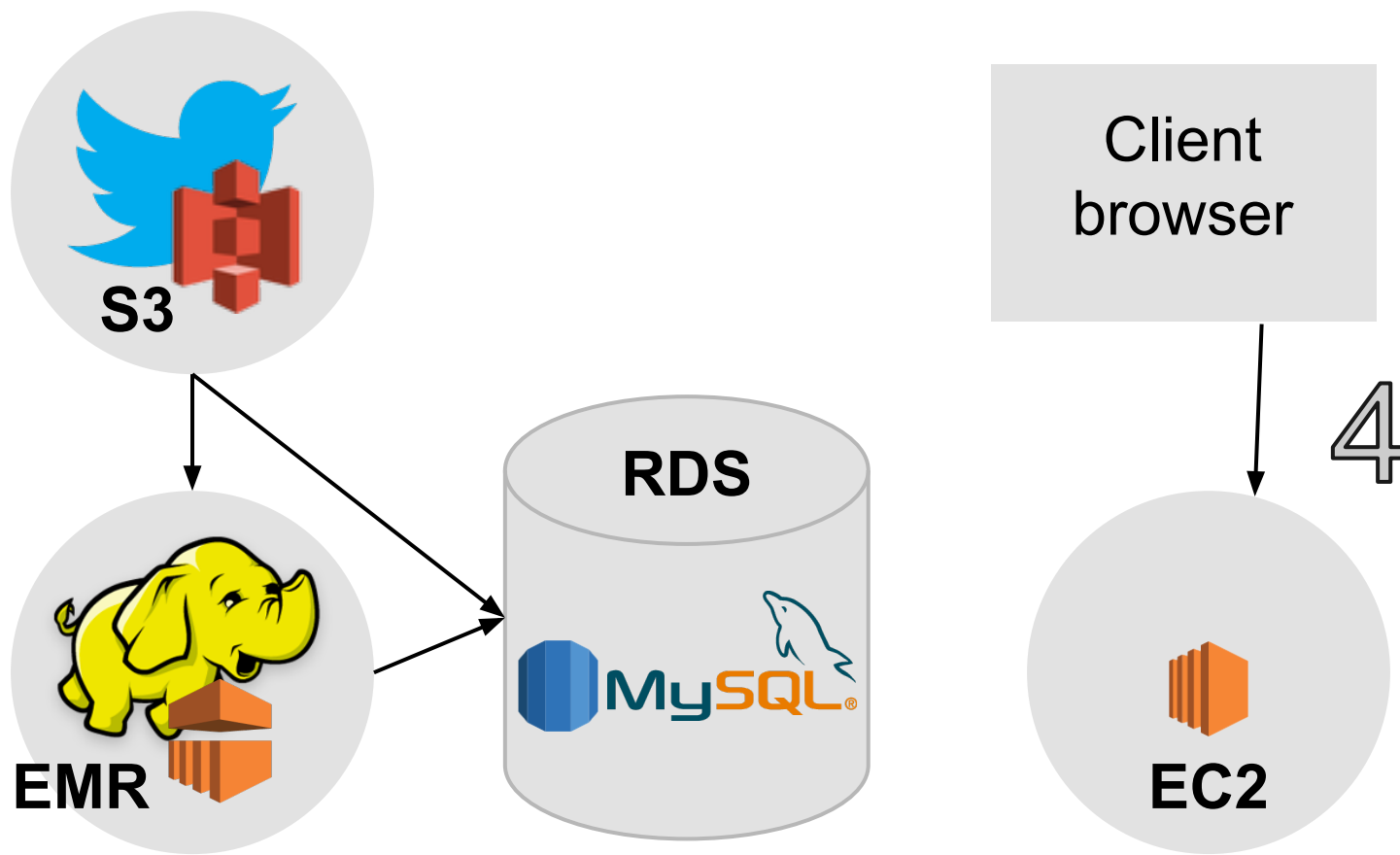


Data Flow

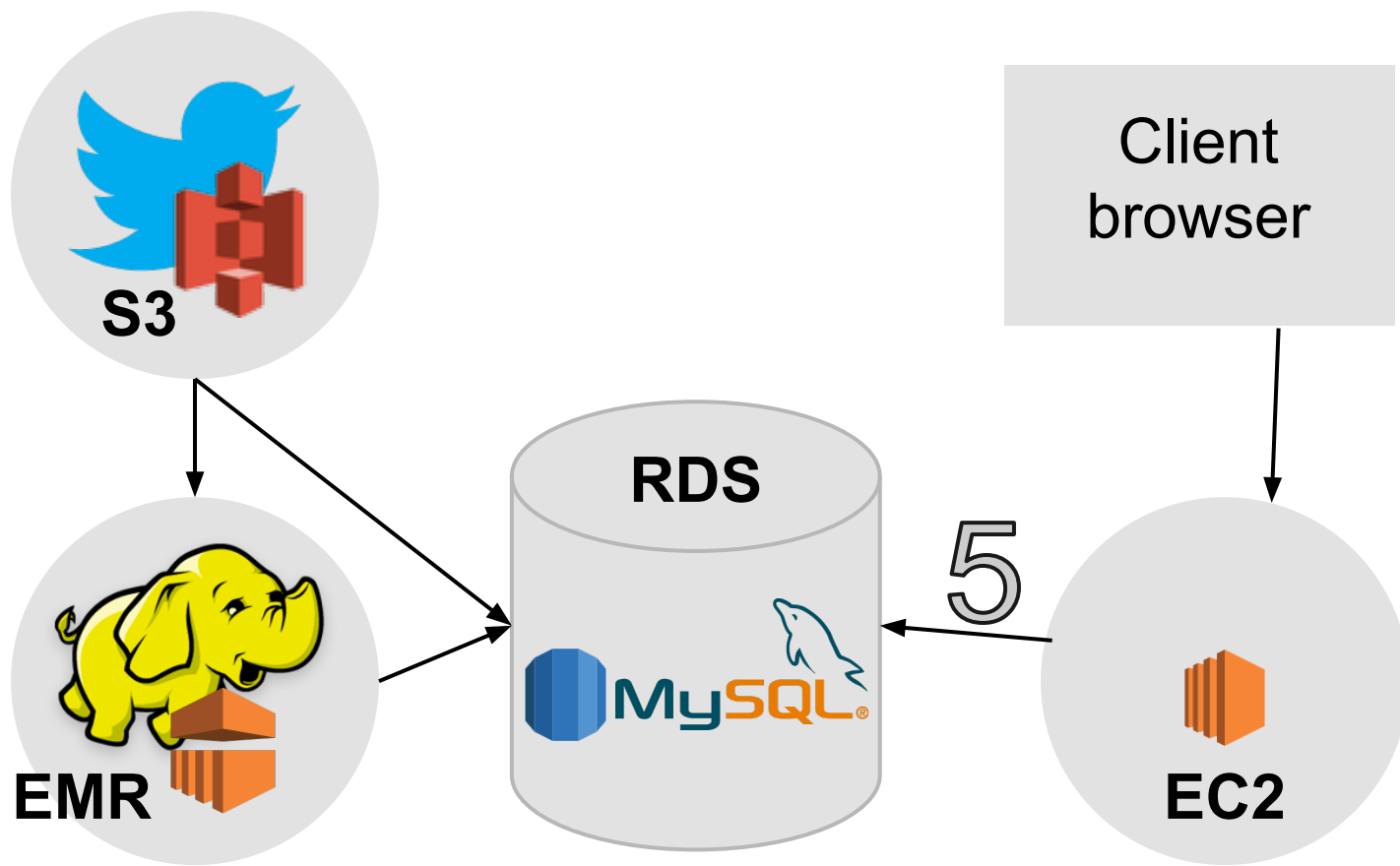


Data Flow

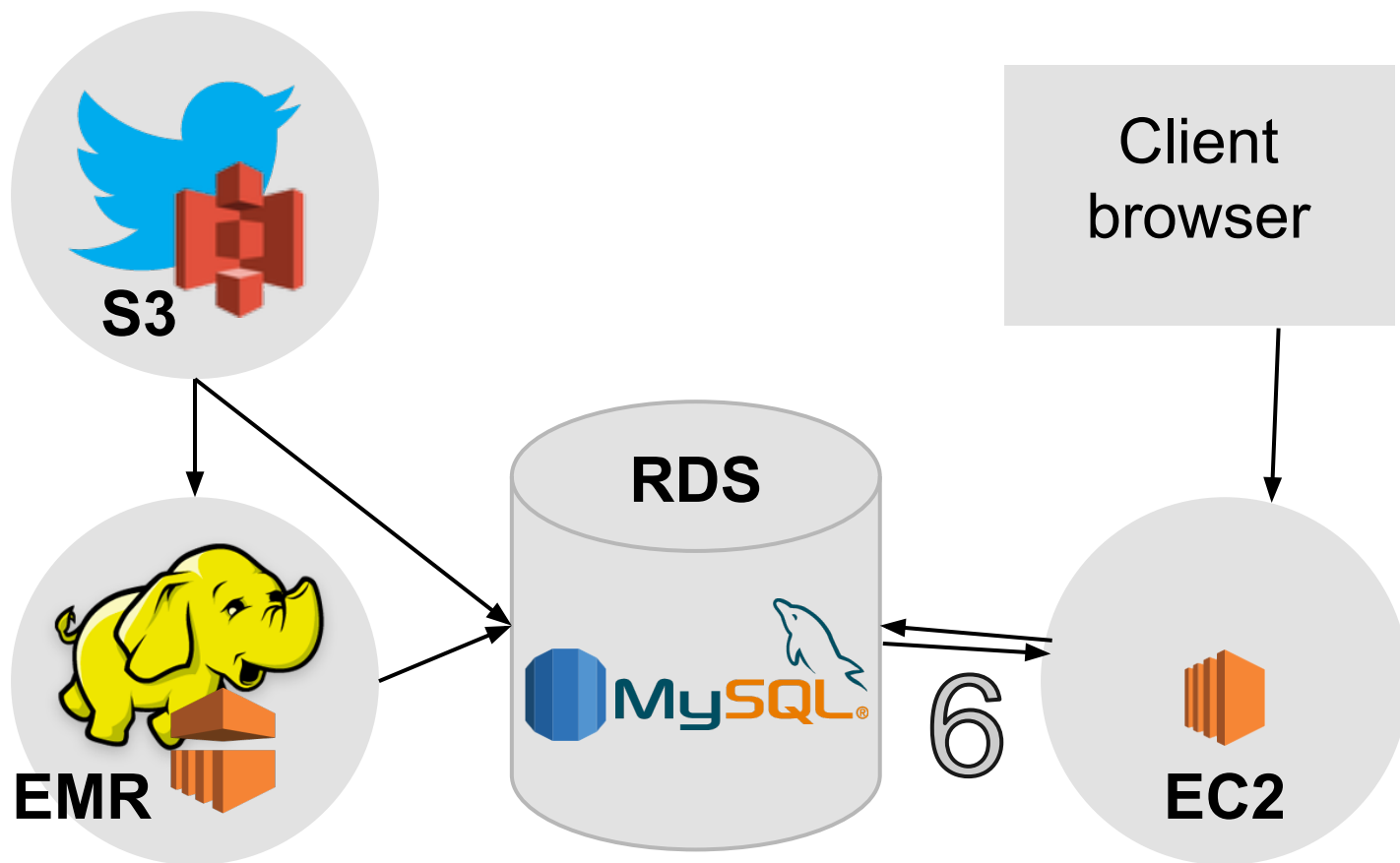




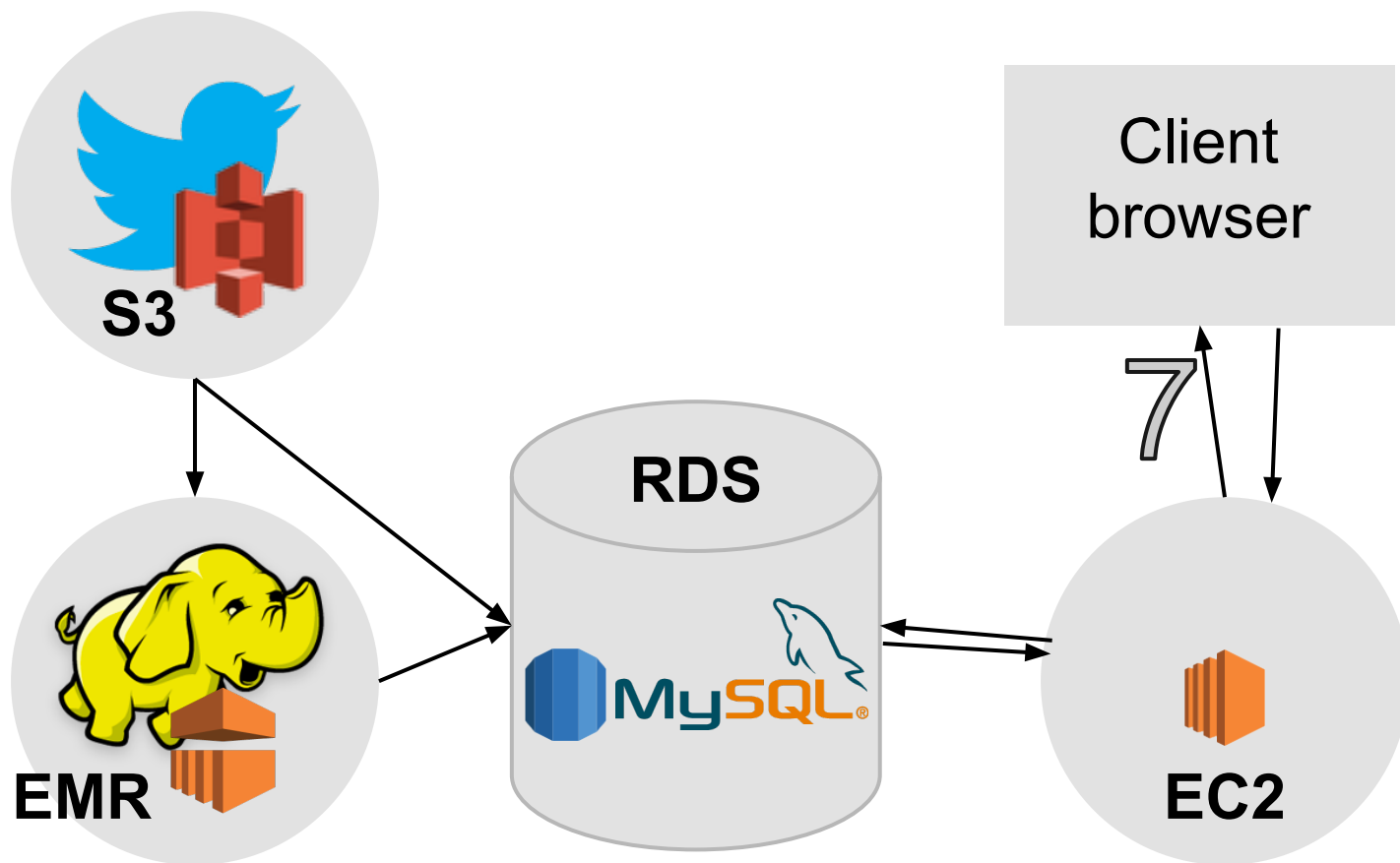
Data Flow



Data Flow



Data Flow



Data Flow



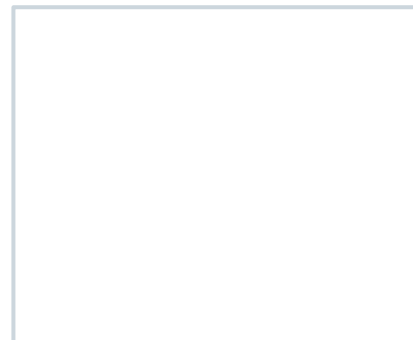
Top Images



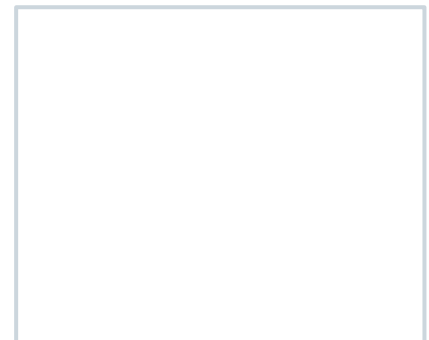
Top Hashtags

#Ukraine
#Crimea
#Euromaiden
#help

Top Keywords



Top Links



Web Interface

Current Progress

- Twitter data and Map-Reduce jobs and computation complete
- Data loaded into MySQL Database
 - Top Hashtags, Links, etc. aggregated by time range
- Working on web server and data visualization
- Measure performance and scalability

Measuring Performance & Scalability

- Python Map-Reduce Streaming on EMR vs standalone EC2 instance with pipes
- Reading Apollo Data on EC2 c3.xlarge instance
 - 4 vCPUs, 7.5 GiB

Challenges

- Picking the right time ranges to then find interesting shifts in trends of hashtags, media, keywords, etc. is limited
- Utilizing the small percentage of tweets that have location data to make geographical inferences
- Data cleaning