Big Data Linguistics

W205.1 Group Project

Week 15 Presentation: 12/18/2015

Brandon Shurick & Jared Maslin

User Case & Motivation

- Traveling to a foreign country soon?
- How will you communicate with native citizens?
- What tools are available to help you?
  - Options available today fall into a few basic buckets:
    - Language software (i.e., Dragon, Rosetta Stone)
    - Coursework from an online institution or a university
    - Educational books
  - Common themes -- Time-consuming and often expensive
- Are there any other alternatives?

Our Solution

- Objective:
  - To offer an approachable platform for self-driven learning of a new language, directly from native writers/speakers
  - For our prototype, we chose to focus on **Spanish**
- Traditional learning methods often spend too much time on conjugation/syntax, which is often unimportant
- Real-time social media data can quickly and effectively identify common words and phrases that an average person can digest in a reasonable period of time
- Gain relevant information about the language and its native speakers in an easily accessible manner

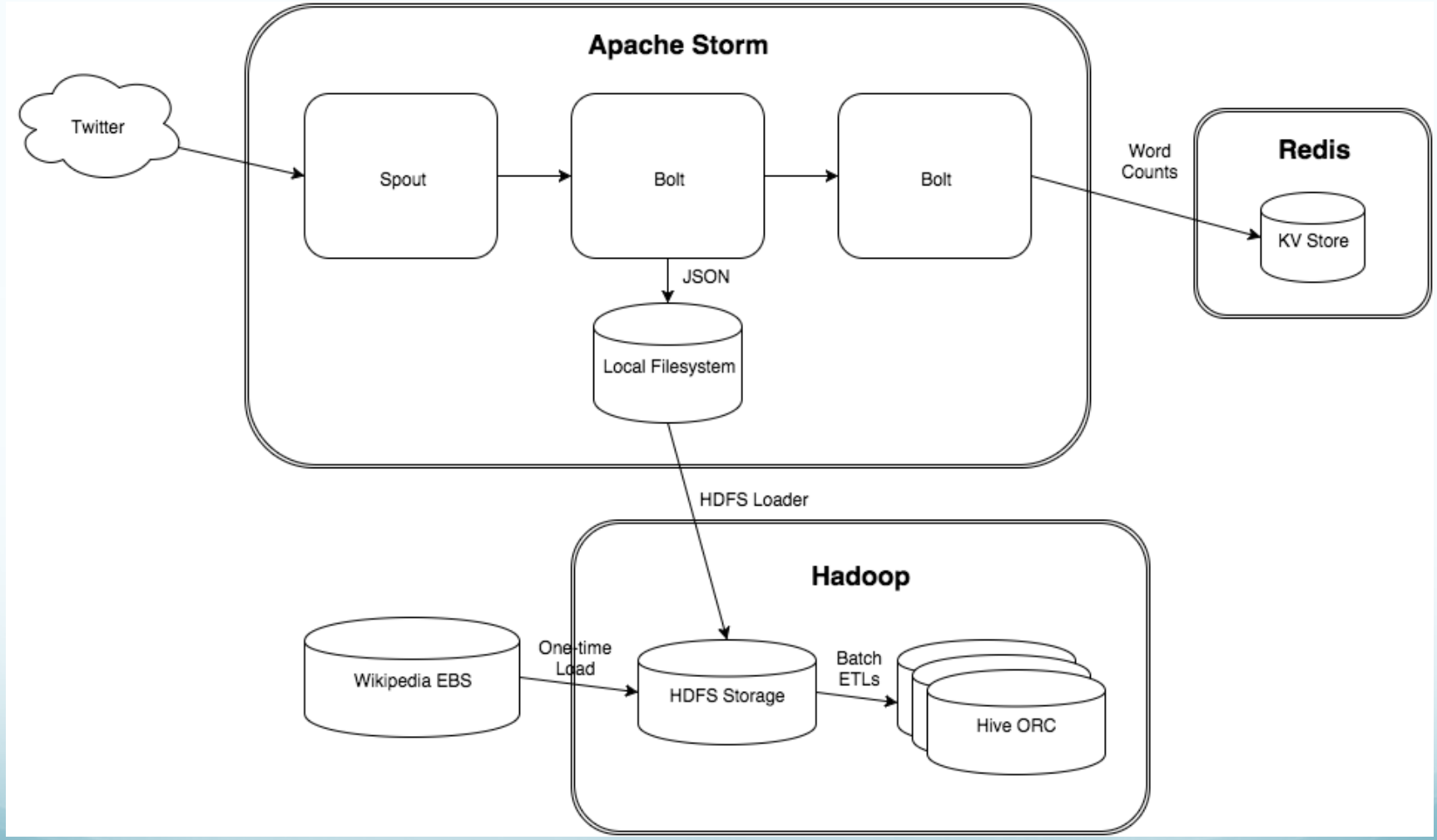
The Approach

- In order to address this gap in language education, we chose two data sources to start with:
  - Twitter API
    - A proxy for informal / conversational speech
    - Store and retrieve vast amounts of data based on the language of each tweet
    - Analyze and summarize data for prioritizing to the user
  - Wikipedia public data (available via EBS volume)
    - More formal language in Wikipedia documents
    - Public datasets available to summarize popular searches by language/location

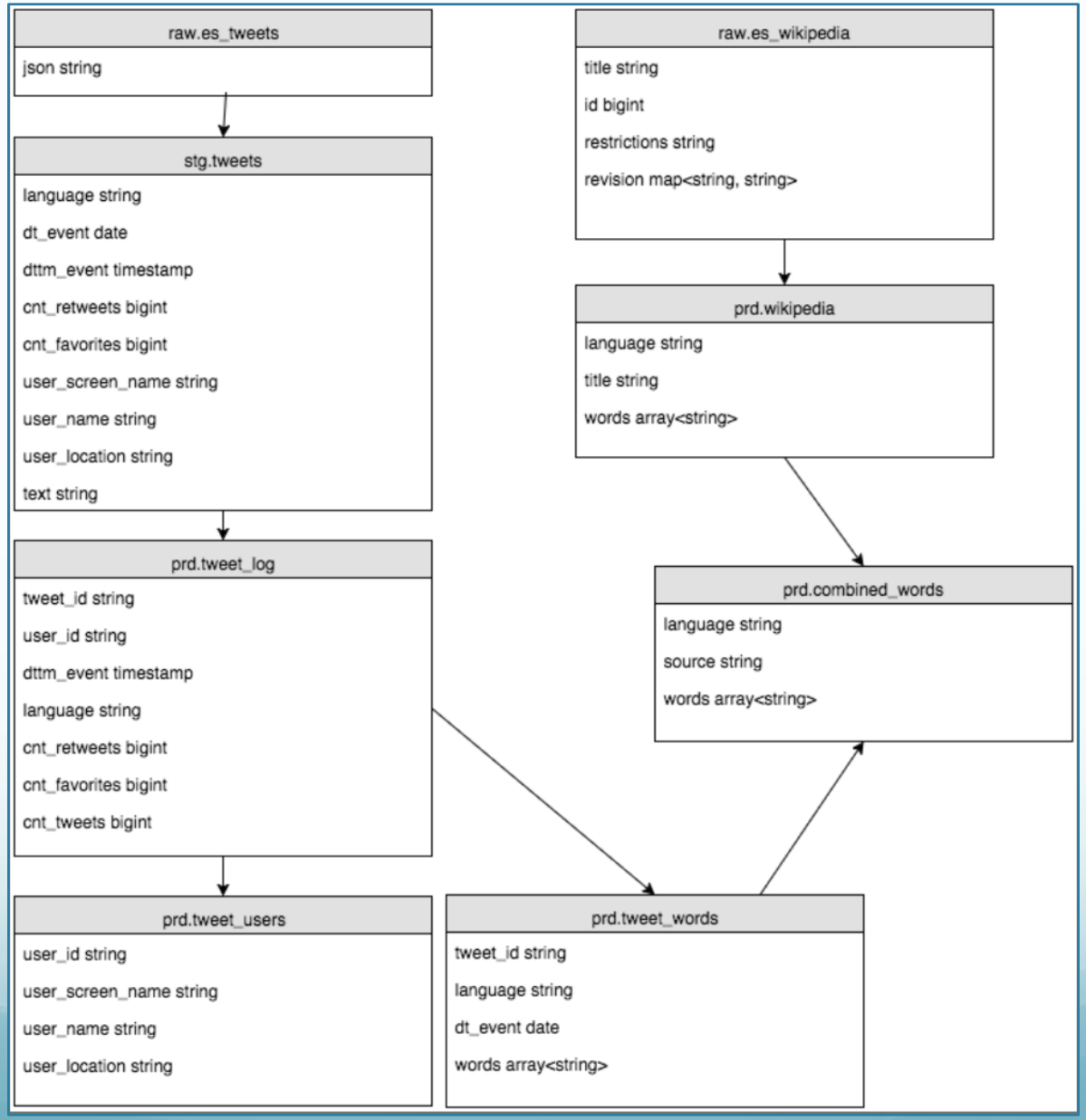
Laying the Foundation

- Created a streaming Twitter application (**27+ million tweets**)
- Loaded a stream of Spanish-language tweets into local files, then compiling and loading data to HDFS every hour
- Increment real-time word counts in Redis
- Attach EBS volume with **787k Spanish Wikipedia documents**
- ETLs to load tweet and Wikipedia data from raw formats
- Combined table with words from documents across all data sources and languages

Architecture



Our ERD At a Glance



Main Challenges

- Handling unicode data from the Spanish-language tweets
- Parsing out words & characters that shouldn't be included
  - URLs, hashtags, usernames (tweets)
  - Markup language code (wikipedia)
- Parsing different formats (JSON and XML)
- Running on a single node & minimizing cost

Next Steps

- Linguistics analysis and presentation layer with emphasis on analysis for developing language instruction
- Consider other languages to include going forward
  - Current model was designed to shift easily from one language to the next
- Extend to other data sources
  - Model is also extendible to additional data sources (ex: Facebook, web as a corpus, Google news, academic journals)

Questions?