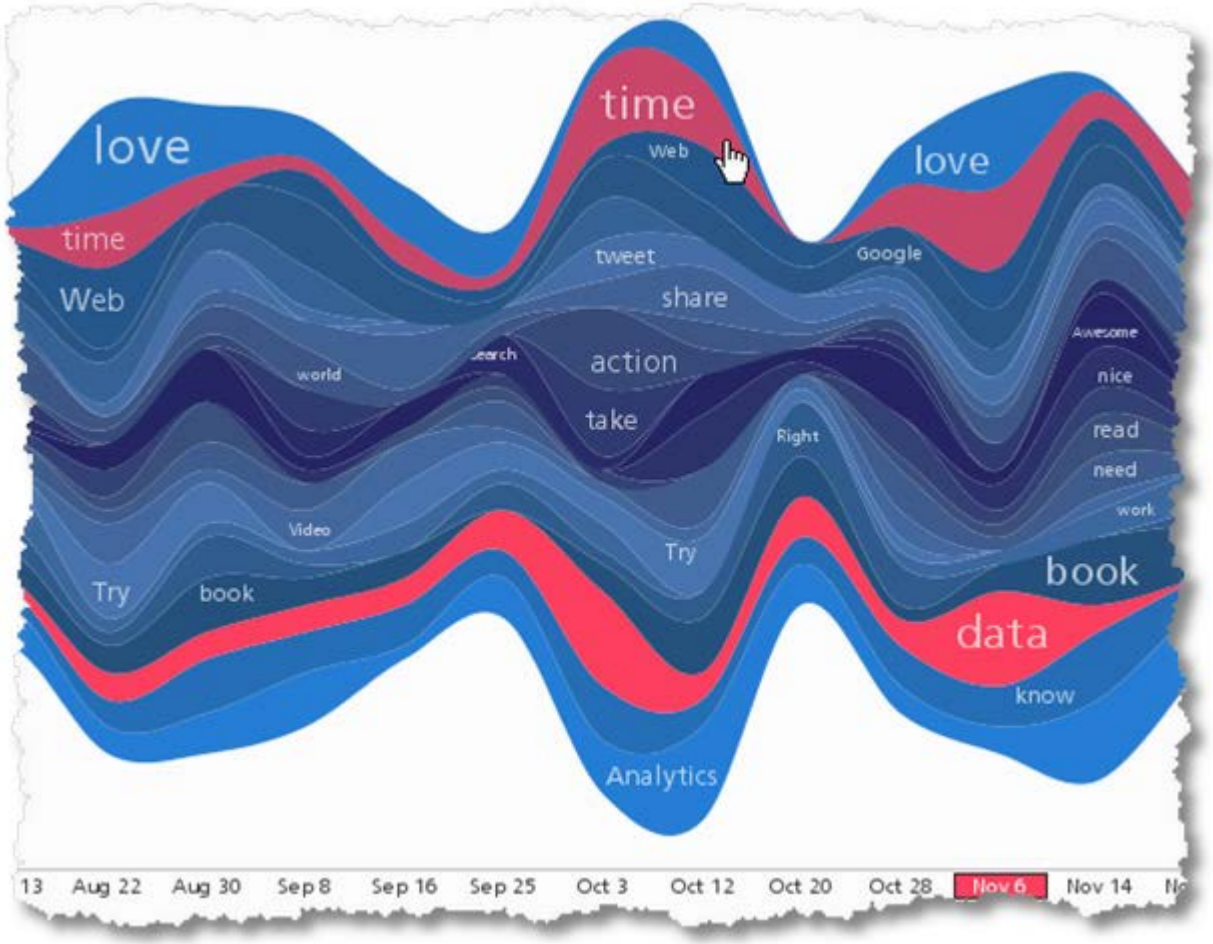
W205 Spring 2016 – Exercise #2

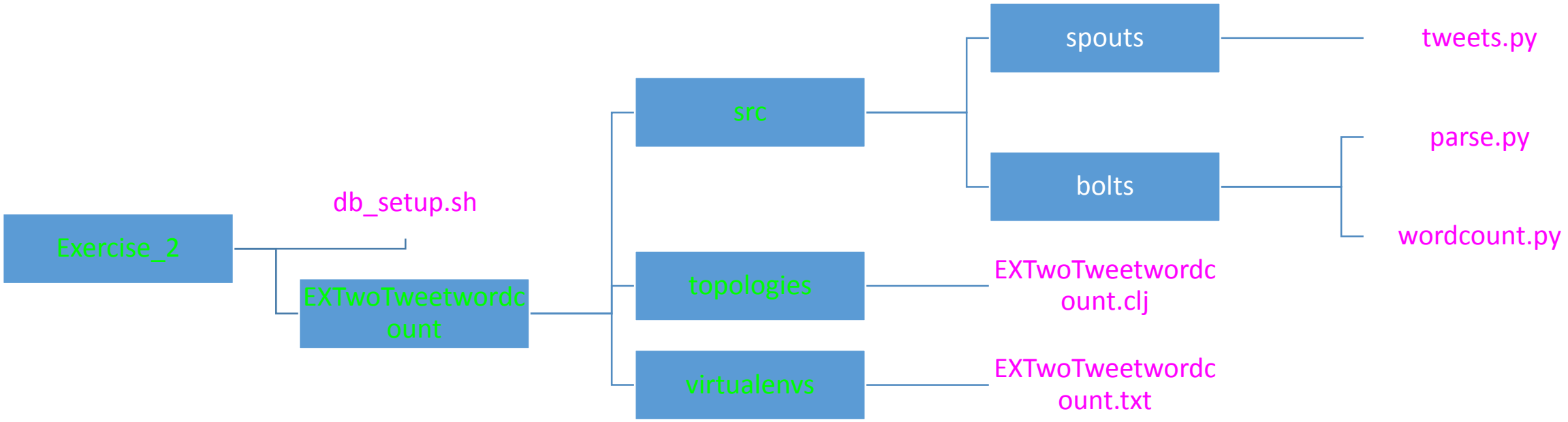
Beth Partridge

Application Idea

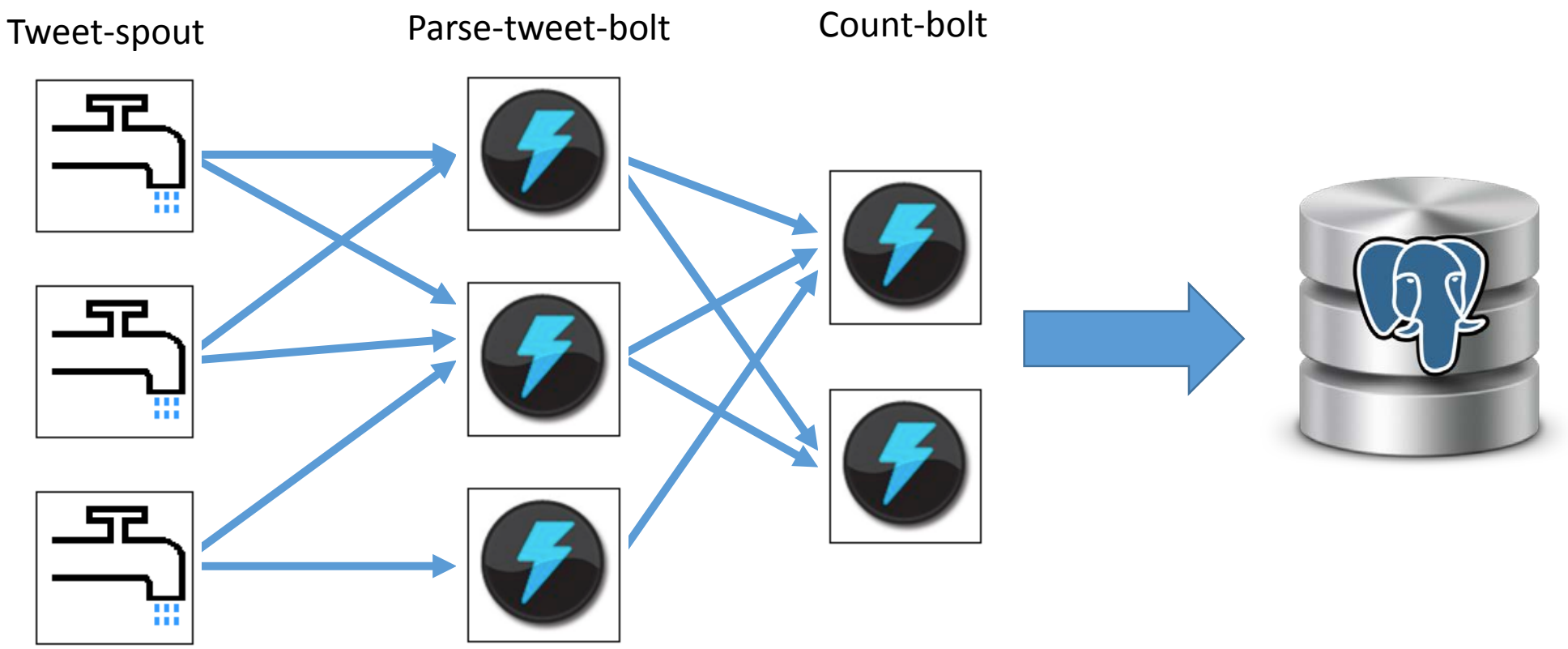
- Capture and analyze live Twitter data for a deeper understanding of social trends and demands using Apache Storm
- Use Tweepy library to read the live Twitter stream
- Parse the stream down to individual ascii words
- Store and update cumulative word counts in a Postgres database



Directory & File Structure



Architecture Block Diagram



Assumptions & Dependencies

- Assumptions
  - Python 2.7
  - PostgreSQL 8.4.20
- Required packages:
  - Tweepy
  - Psycopg2
  - Streamparse
  - Matplotlib
  - Re
- Other
  - Twitter application and credentials

Instructions to Run Application

1. Launch AWS EC2 instance using UCB W205 Spring Ex 2 Image (ami-4cf9f826)
2. Change to python 2.7 environment: source /opt/py27environment/bin/activate
3. Attach a volume already prepared from previous labs and exercises
4. Clone or copy submitted Exercise_2 repository from Github
5. Change to Exercise_2 directory: cd Exercise_2
6. Make the db_setup script executable: chmod a+x db_setup.sh
7. Execute the setup script: bash db_setup.sh
8. Change to the EXTwoTweetwordcount directory: cd EXTwoTweetwordcount
9. Run the application: sparse run
10. When you see a word count reach 10 stop the application with ^C
11. Run finalresults.py with no argument: python finalresults.py
12. Run finalresults.py with a common word argument: python finalresults.py the
13. Run histogram.py to see all the words with counts between 6-10: python histogram.py 6,10