SOCIAL MEDIA SEMINAR

Ryan Wesslen / Project Mosaic

Part 2: April 27, 2017

OVERVIEW

- Recap & Json Query Example: 10 minutes
- Gnip Job Acceptance Process: 20 minutes
  - SOPHI-Gnip ingestion process
- Data Processing: 20 minutes
  - Json-to-CSV, SOPHI Twitter Datasets
- Potential Data Analysis Methods: 10 minutes

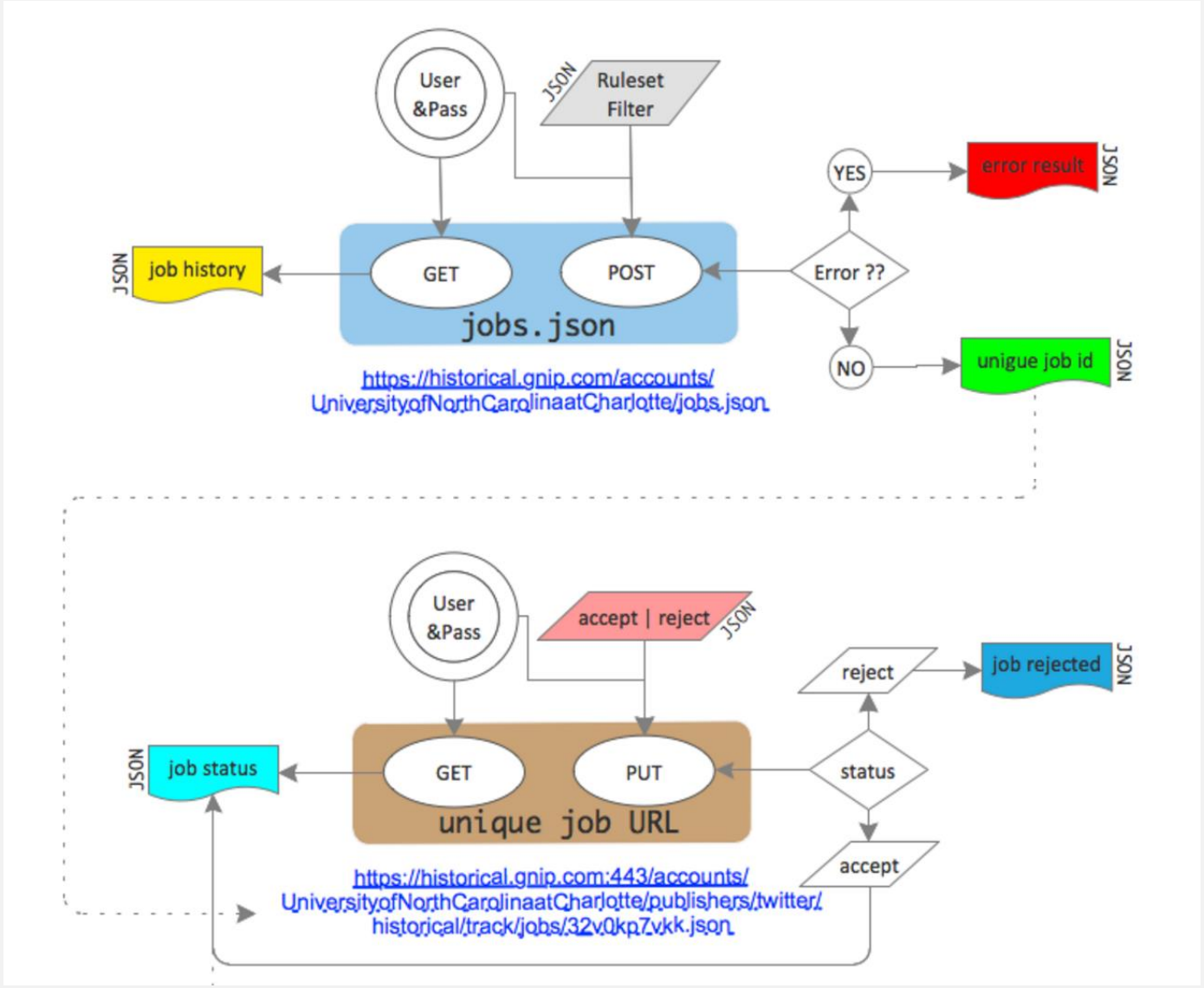
Download these slides: http://webpages.uncc.edu/rwesslen/social-media-part2.pptx

JSON EXAMPLE: MARCH FOR SCIENCE


```
{
  "title": "MarchforScience",
  "publisher": "twitter",
  "streamType": "track_v2",
  "dataFormat": "activity_streams",
  "fromDate": "201704080000",
  "toDate": "201704260000",
  "serviceUsername": "twitterUsername",
  "rules": [
    {
      "value": "((March4Science) OR (March for Science) OR (Science Matters) OR #ScienceMatters OR
#ScienceMarch OR (Science March) OR (Because Science) OR #EverydayScience OR
#ScienceEmpowers OR @ScienceMarchDC OR #NoSidesInScience)",
      "tag": "marchscience"
    }
  ]
}
```

GNIP HISTORICAL POWERTRACK PROCESS

Step 1: Submit json filter rules through command line (cURL command)



Step 2: Was the job created?

Step 3: Estimate is created

Step 4: Accept or reject the job.

JSON RESPONSE


```
{
  "title": "MarchforScience",
  "account": "UniversityofNorthCarolinaatCharlotte",
  "publisher": "twitter",
  "streamType": "track_v2",
  "format": "activity_streams",
  "fromDate": "201704080000",
  "toDate": "201704260000",
  "requestedBy": "rwesslen@uncc.edu",
  "requestedAt": "2017-04-26T11:50:31Z",
  "status": "quoted",
  "statusMessage": "Job quoted and awaiting customer acceptance.",
  "jobURL": "https:\/\/gnip-
api.gnip.com:443\/historical\/powertrack\/accounts\/UniversityofNorthCarolinaatCharlotte\/publishers\/twitter\/jobs\/rgk4xg54c0.json"
,
  "quote": {
    "estimatedActivityCount": 1113000,
    "estimatedDurationHours": "7.0",
    "estimatedFileSizeMb": "748.39",
    "expiresAt": "2017-05-03T12:57:41Z"
  },
  "percentComplete": 0
}
```

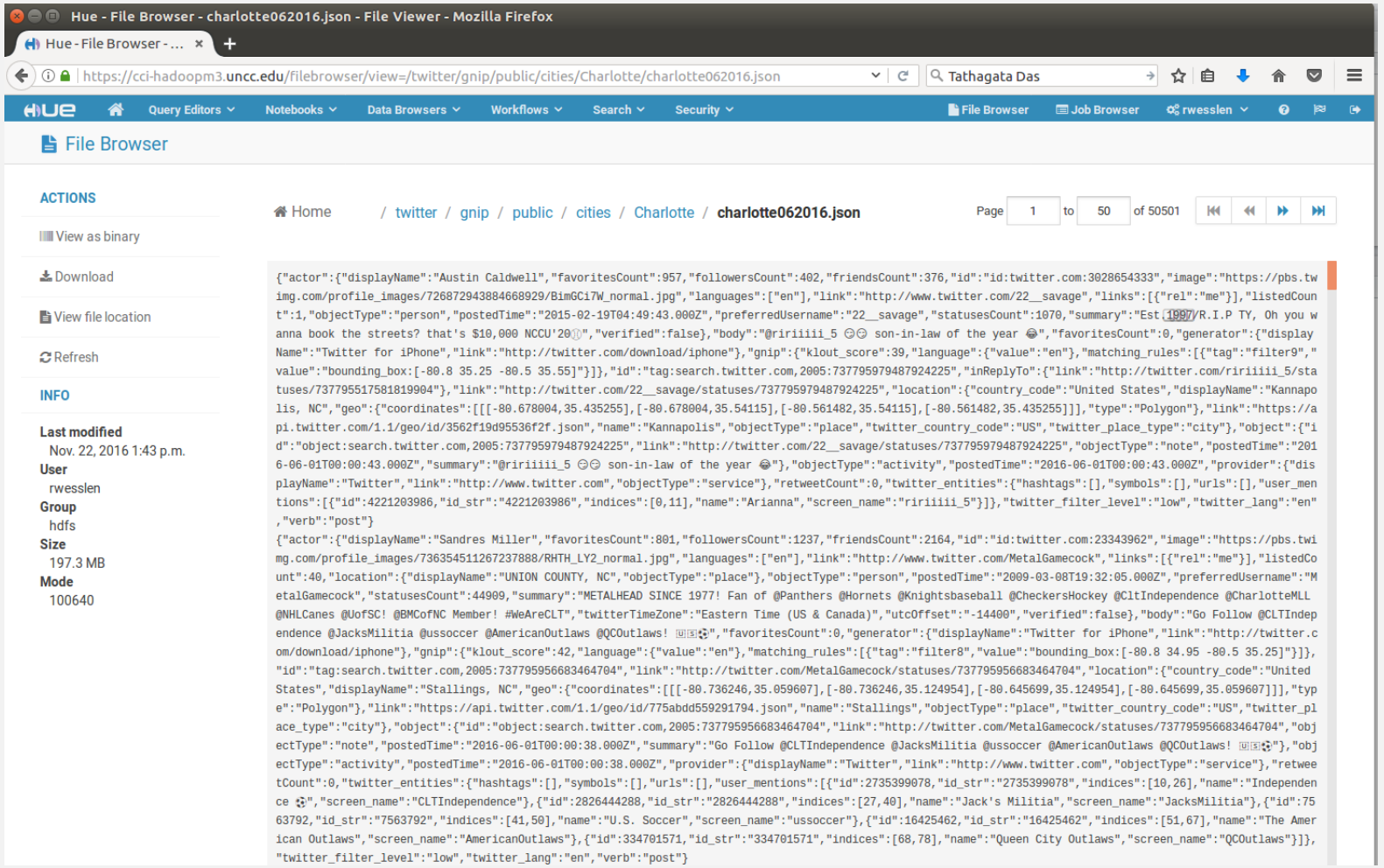
SOPHI INGESTION PROCESS

- Once completed, a URL (endpoint) is provided to download the files.
  - The files come as **distributed** 10 min files

- Instead of users pulling these files, we have an automated process that will download the files, archive on SOPHI, and concatenate the files into one large file.

http://support.gnip.com/apis/historical_api/overview.html

JSON DATASET FILE



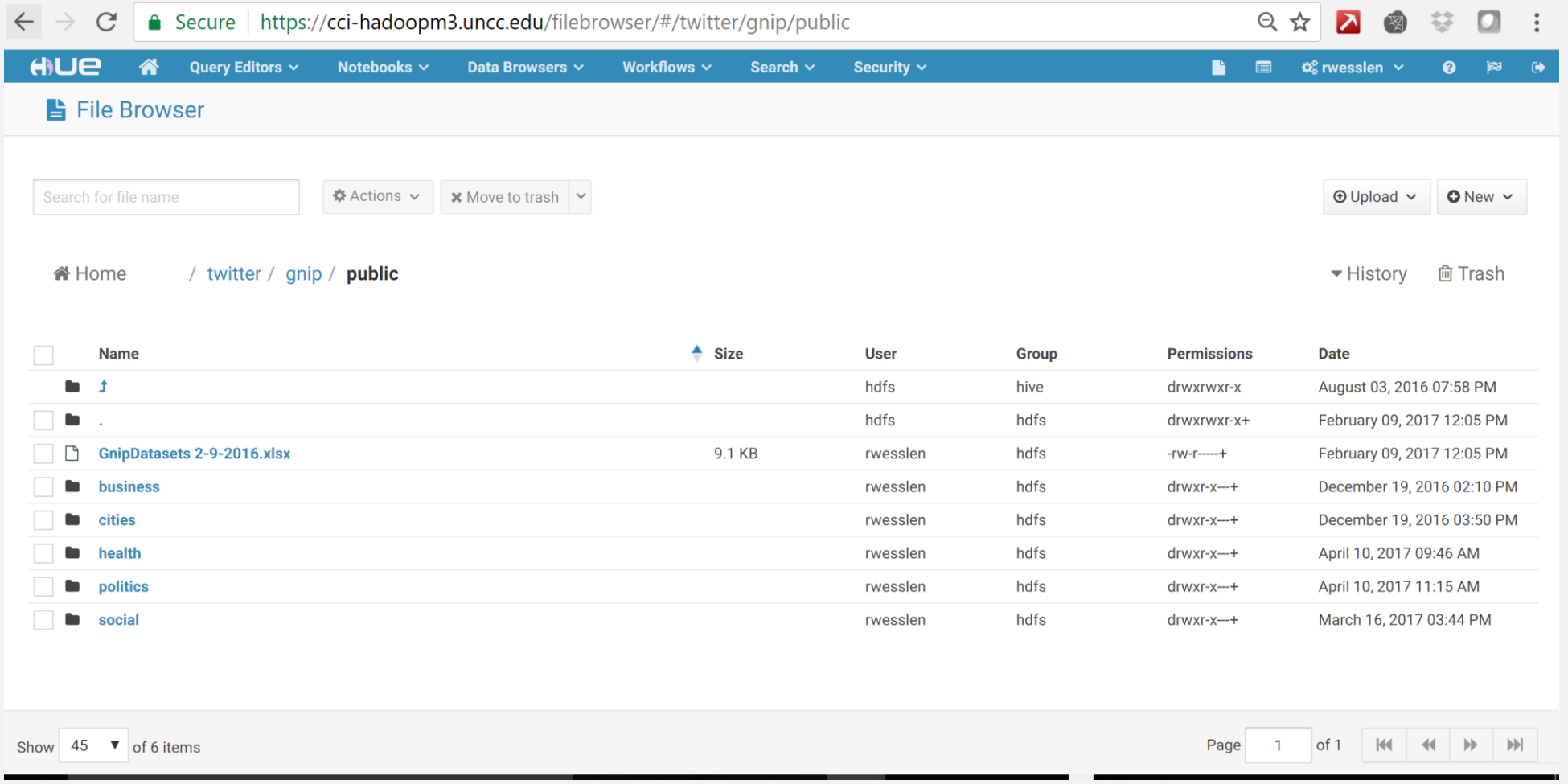
Example (Field Names Differ)

Tweets are stored in JSON format:


```
{ "created_at": "Wed Nov 07 04:16:18 +0000 2012",
  "id": 266031293945503744,
  "text": "Four more years. http://t.co/bAJE6Vom",
  "source": "web",
  "user": {
    "id": 813286,
    "name": "Barack Obama",
    "screen_name": "BarackObama",
    "location": "Washington, DC",
    "description": "This account is run by Organizing for Action staff.
       Tweets from the President are signed -bo.",
    "url": "http://t.co/8aJ56Jcemr",
    "protected": false,
    "followers_count": 54873124,
    "friends_count": 654580,
    "listed_count": 202495,
    "created_at": "Mon Mar 05 22:08:25 +0000 2007",
    "time_zone": "Eastern Time (US & Canada)",
    "statuses_count": 10687,
    "lang": "en" },
  "coordinates": null,
  "retweet_count": 756411,
  "favorite_count": 288867,
  "lang": "en"
}
```


https://github.com/jimmoffitt/json2csv/blob/master/templates/tweet_standard.json

SOPHI DEMO



Sophi main link: https://sophi.uncc.edu/hue [must be connected to UNCC network or VPN]
Twitter Datasets: https://cci-hadoopm3.uncc.edu/filebrowser/#/twitter/gnip/public
Twitter-Sophi Datasets Workshop: https://webpages.uncc.edu/rwesslen/sparktwitter.pptx

JSON2CSV CONVERSION

Converting Data from JSON to CSV

json2csv

Customers often ask us about converting Tweet JSON into comma-separated values (CSV). These customers have received Twitter data from a Gnip Product such as Historical PowerTrack, 30-Day Search or Full-Archive Search, which all encode Tweets in JSON. The reasons for this are numerous, from wanting to work with a sample of the data in a spreadsheet, or needing to import the data into a relational database or legacy system. The CSV format's value is in its simplicity, and most software systems are able to import it.

To help with this process, the json2csv app is available. This tool manages the conversion of Gnip Activity Stream (AS) JSON to the comma-separated values (CSV) format.

Tweet attributes of interest are indicated by referencing a Tweet Template of choice. If the Tweet Template has an attribute it will be written to the output CSV files. If the Template does not have the attribute, it is dropped and not written. You can design your own Tweet Template, or use one of the provided example Templates.

This tool works with an input folder and attempts to convert all *.json and *.json.gz files it finds there, writing the resulting CSV files to an output folder. This tool will work with Tweet JSON produced with both Gnip Historical PowerTrack and Search API products. Thus, this tool was designed to convert JSON Tweets in bulk.

Before deciding to perform this type of conversion, you should consider the following trade-offs:

1. JSON data are multi-dimensional, with multiple levels of nested data. However, CSVs are two dimensional. Converting from JSON to CSV means that you are sacrificing detail and flexibility in the data by either flattening it, or discarding some fields from the data.

http://support.gnip.com/articles/json2csv.html
Annotated instructions: Mac or Windows
Example Dataset: https://www.dropbox.com/s/cbu1kwz2b1j1c9z/CharlotteTweets20Sample.csv?dl=0

HELPFUL LINKS ON GNIP DATA

- Gnip Twitter Metadata Dictionary
- Identifying and understanding Retweets
- Filtering By Location
  - Geo Metadata
- Converting Json to CSV (Ruby Code)
- Consuming, Parsing and Processing Tweets with Python

DATA ANALYSIS: WORKSHOPS

- Social Media (Twitter, Facebook) Data Acquisition in R
- Twitter Text Analysis with R Workshop
- R Interactive Visualizations Tutorials (Uses Gnip Data)
- Beer in Charlotte on Twitter Tutorial
  - GitHub Repository

PARTING WORDS

- Start simple!
  - Social media is **unlike** traditional data sources.
  - Avoid data deluge. More data is not always better!!
- Practice, practice, practice!
  - Twitter Public API or sample datasets.
  - If you're serious, need programming skills (at least R, some Python).

Project Mosaic R Faculty Workshop coming in June!

- Why Twitter?
  - Avoid single platform dependency and model organism problem (Tufekci, 2012).