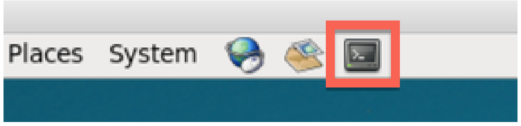
[Download a PDF of this hands-on tutorial]

ExploringtheSemistructuredDataModelofJSONdata.pdf

By the end of this activity, you will be able to:

1. Display the nested structure of a JSON file.
2. Extract data from a JSON file.

Step 1. Open a terminal shell. Open a terminal shell by clicking on the square black box on the top left of the screen.

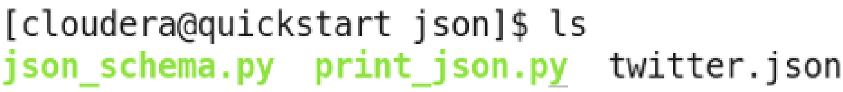


Change into the JSON directory:

```
cd Downloads/big-data-2/json
```

Run *ls* to see the JSON file and scripts:

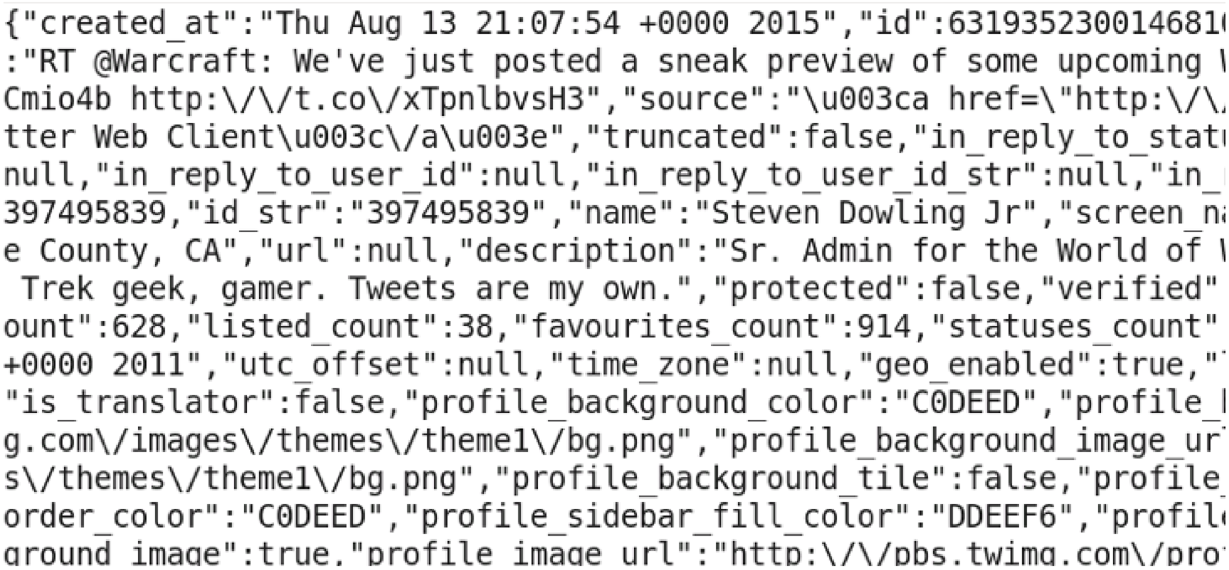
```
ls

```



```
[cloudera@quickstart json]$ ls
json_schema.py  print_json.py  twitter.json
```

Step 2. Look at JSON file. Let's look at the contents of the JSON file:

```
more twitter.json
```



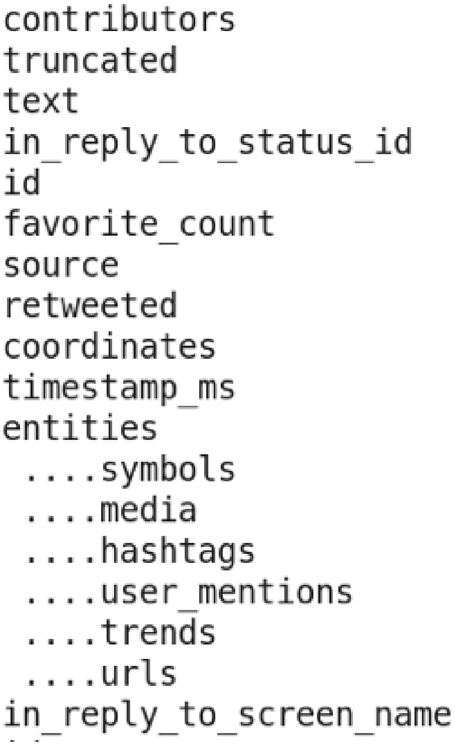
```
{"created_at":"Thu Aug 13 21:07:54 +0000 2015","id":63193523001468l
:"RT @Warcraft: We've just posted a sneak preview of some upcoming \
Cmio4b http:\/\/t.co\/xTpnlbvsH3","source":"\u003ca href=\"http:\/\
tter Web Client\u003c\/a\u003e","truncated":false,"in_reply_to_stat
null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_
397495839,"id_str":"397495839","name":"Steven Dowling Jr","screen_n
e County, CA","url":null,"description":"Sr. Admin for the World of \
 Trek geek, gamer. Tweets are my own.","protected":false,"verified"
ount":628,"listed_count":38,"favourites_count":914,"statuses_count"
+0000 2011","utc_offset":null,"time_zone":null,"geo_enabled":true,"
"is_translator":false,"profile_background_color":"C0DEED","profile_
g.com\/images\/themes\/theme1\/bg.png","profile_background_image_ur
s\/themes\/theme1\/bg.png","profile_background_tile":false,"profile_
order_color":"C0DEED","profile_sidebar_fill_color":"DDEEF6","profil
ground_image":true,"profile_image_url":"http:\/\/pbs.twimg.com\/pro
```

Press the spacebar to go down and *q* to quit more.

The contents of the file is difficult to understand since it is packed together.

Step 3. View JSON schema. We can view the schema of the JSON file by running *schema.py*:

```
./json_schema.py twitter.json | more
```



```
contributors
truncated
text
in_reply_to_status_id
id
favorite_count
source
retweeted
coordinates
timestamp_ms
entities
 ....symbols
 ....media
 ....hashtags
 ....user_mentions
 ....trends
 ....urls
in_reply_to_screen_name
```

The top-level fields are contributes, trucated, text, etc. Some fields have nested fields, such as entities, which contains symbols, media, hashtags, etc. If go you down (press spacebar), you will see multiple levels of nesting.

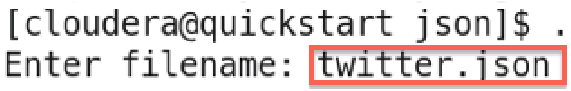
Enter *q* to quit more.

Step 4. Extract values in JSON data. We can extract individual values from fields within the JSON data by running print_json.py:

```
./print_json.py
```

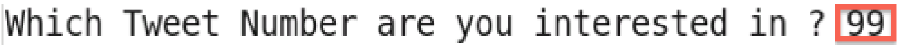
The print_json.py asks for the file name, tweet number, and path to extract. The path is the path to the field in the schema.

Let's look at the value for the *text* field in the 99th tweet. First, enter *twitter.json* for the filename:



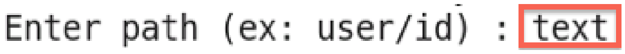
```
[cloudera@quickstart json]$ ./print_json.py
Enter filename: twitter.json
```

Next, enter *99* for the number:



```
Which Tweet Number are you interested in ? 99
```

Next, enter *text* for the path:

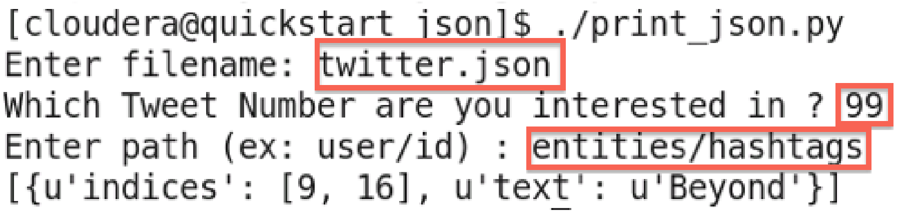


```
Enter path (ex: user/id) : text
```

The result is:

```
RT @IGN: #Beyond: Sony's game plan for the rest of 2015 http://t.co/AIzUexSjrM http://t.co/o7jVHjnVuw
```

Now let's find the value for *entities.hashtags* in the 99th tweet. The *hashtags* field is nested in the *entities* field, so we enter *entities/hashtags* for the path:



```
[cloudera@quickstart json]$ ./print_json.py
Enter filename: twitter.json
Which Tweet Number are you interested in ? 99
Enter path (ex: user/id) : entities/hashtags
[{u'indices': [9, 16], u'text': u'Beyond'}]
```

Mark as completed