Resources

1. Demos Link
2. Ingesting CSV into Elastic Search Link
3. Spark and Elasticsearch Link

Lab Exercise 1: Ingesting data into Elastic-search

At the end of this lab you would have:

1. Ingested gdelt events data using the logstash application.
2. Created multiple logstash configuration files to ingest various csv file.
3. Use logstash to transform data before ingestion.
4. Have created the architecture shown in the figure below:

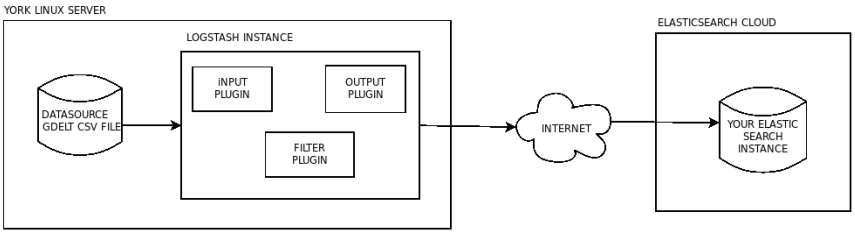

Figure 1: ingestion architecture

In this lab we will be using logstash to ingest our cleaned gdelt dataset: You will reuse your spark code to cleaned the data set and then ingest that clean dataset into elasticsearch.

Part 1: Setting up your elastic cloud instance

1. Signup for a 14 day elastic search trial here: 14 Day Trial Link
   - Enter your email address and press start free trial.
   - Then go to inbox and verify your account
   - This will take you back to the create cluster page. Go ahead and create a cluster.
2. After creating the cluster be sure to save your cluster.
   - url
   - user name
   - password

You now have completed setting up your elastic cloud instance.

Part 2:

Before proceeding, read the following introduction to understand what is logstash: Introduction to logstash Link

In this section we will be using logstash to ingest data into the elastic cloud instance you configure in part 1.

1. Follow the following instructions to login into the york server and open a terminal (see instructions I posted on moodle yesterday)
2. Once you have logged into the server and open a terminal. Enter the following command at the terminal prompt.
3. Create a directory to store the files you will be working with. By entering the following command at the terminal prompt:

   ```
   >> mkdir gdelt_ingestion
   ```

4. Go into the directory by entering the command and the terminal prompt:

   ```
   >> cd gdelt_ingestion
   ```

5. Now that you are in the gdelt_ingestion folder, lets create a data directory for your gdelt events files. Enter the following command at the bash prompt to create the directory:

   ```
   >> mkdir -p  data/events
   ```

6. Enter the follow command to download the events data file from github into the data/events folder

   ```
   >> wget  https://raw.githubusercontent.com/LeotisBuchanan/
   csd1020_fulltime/master/data/events/20180730151500.events.csv
   -P data/events
   ```

   Check if the file was copied by running the following command:

   ```
   >> ls  data/events/20180730151500.events.csv
   ```

7. You can view the first 20 rows of the data by entering the following command `console >> head -20 data/events/20180730151500.events.csv`
8. Now that you have downloaded the data file, enter the follow command at the terminal prompt to download the events-logstash.config file located here: events-logstash-config Link

   First let create a directory to store the logstash configuration file. Enter the following command that bash prompt to do that:

   ```
   >> mkdir config
   ```

```
>> wget -c  https://raw.githubusercontent.com/LeotisBuchanan/csd1020_fulltime/
master/week2/logstash_configs/events-logstash.config.template -P  config
```

Run the following command to copy events-logstash.config.template to events-logstash.config

```
>> cp configs/events-logstash.config.template configs/events-logstash.config
```

This will copy the events-logstash.config file to the current config directory.Go ahead and check that the file as been downloaded correctly.

9. The events-logstash.config file will tell logstash how to ingest your data file into your elasticsearch instance. Since your elasticsearch instance is secure you will have to edit to file to provide the following:

   - The absolute path to the data file that you want to load into elasticsearch.
   - the url for your elasticsearch cluster
   - the password for your elasticsearch cluster

   I asked you to save these in part 1 of this Document.

10. You will need the absolute path to the data file. Run the following command to get it:

    ```
    >> ls "`pwd`data/events/20180730151500.events.csv"
    ```

    Save the file path. for the next steps.

11. Open the file using the following command:

    ```
    >> nano config/events-logstash.config
    ```

12. Find the following text in the file: **path => "/ENTER/EVENTS/FILE/HERE/*.csv"** enter the absolute path to the events data file(you got it in step 10)

13. Find the following text in the file: **hosts => "ENTER THE URL FOR YOUR ELASTIC INSTANCE HERE"**

    Enter your cluster url here

14. Repeat for the password.

15. Close the nano editor by pressing Crtl X key on your keyboard. When ask to save the file type Y. This will save the file.

16. We are now ready to ingest the gdelt data into your elastic. To do this we use logstash. Logstash as already installed this for you :-) .

    Start logstash and tell it to use the config you configured in steps 11 - 15 Type the following command at the terminal: change the "/path/to/config" to the path to your config i.e config/events-logstash.config

```
logstash -f /path/to/config --path.data ~/Data/logstash/
```

17. If every thing went well your data file should be uploaded to your elasticsearch instance.

Practice Exercise

Now that you have ingested the gdelt events data into elasticsearch. Go ahead and ingest the gdelt mentions data and the cameevents data into elasticsearch.