“Catch the Pink Flamingo” Data Exploration with Splunk

Some of the simulated game data files generated by the “Catch the Pink Flamingo” python scripts lend themselves to analysis with Splunk. You were first introduced to Splunk in one of the previous courses in the specialization, “Introduction to Big Data Analytics”. We include some of those readings here, as well as links to the original video lectures.

First, assuming you already have Splunk installed, we need to upload our data to Splunk.

Upload the Data to Splunk

You can upload your data in its entirety to Splunk with a compressed zip file. However, for some tasks you may need to upload the files individually. If you generated the data yourself, then you should create a zip archive of the following eight files:

- users.csv
- user-session.csv
- teams.csv
- team-assignments.csv
- ad-clicks.csv
- buy-clicks.csv
- game-clicks.csv
- level-events.csv

If you downloaded a zip file of the data, you should be able to upload that same zip file to Splunk, as instructed in Week 3 of the earlier course in this Specialization on ‘Introduction to Big Data Analytics’.

QUERY 1: What is the Distribution of Operating Systems Used by Users?

The data file “user-session.csv” contains a column of data for “platformType”. This column is ‘enumerated’ (or ‘enum’ for short) with five values: windows, mac, android, ios, linux. We want to know what the relative distribution of these operating systems is for our dataset.

To do this, we begin by running Splunk and accessing the list of Data Sources:

Click the link for the “user-session.csv” file. This will load the data into Splunk and prepare it for searching. The command we will use is the “count by” command, as follows:

source="user-session.csv" | stats count by platformType

```
source="user-session.csv" | stats count by platformType
```

This should return the following:

platformType	count
android	3274
iphone	3874
linux	504
mac	358
windows	1240

Creating a Chart in Splunk

To generate a chart of the above results, you simply click the “Visualization” tab and select the type of chart you want from the Type menu.

QUERY 2: What are the two most Commonly-Clicked Ads?

A similar query may be performed in order to determine the ads which are clicked most frequently. The file “ad-clicks.csv” contains data on what users clicked what ads.

To begin, in the Data Sources panel, select “ad-clicks.csv”.

In the Search box, enter the following Splunk query:

source=“ad-clicks.csv” | stats count by adCategory | sort 2 -num(count)

```
source=“ad-clicks.csv” | stats count by adCategory | sort 2 -num(count)
```

The results should be:

computers: 2638

games: 2601

QUERY 3: What are the two most Commonly-Purchased Products?

The file “buy-clicks.csv” contains data on which users purchased which products on which dates. We can perform yet another similar search query in Splunk to determine which products are purchased most frequently.

To begin, in the Data Sources panel, select “buy-clicks.csv”.

In the Search box, enter the following Splunk query:

source="buy-clicks.csv" | stats count by buyId | sort 2 -num(count)

```
source="buy-clicks.csv" | stats count by buyId | sort 2 -num(count)
```

The results should be:

buyId	count
2	714
5	610

QUERY 4: What is the Average Team Size?

To calculate the average team size requires a somewhat more complex query. The “team-assignments.csv” file contains data on which users are part of which teams.

Our first step is to select the “team-assignments.csv” data file from the Data Sources. Next, we need to generate a distribution of team sizes using the same “count by” command we used in the previous step. Once we have that distribution, we need to calculate the average.

Here is the final query:

```
source="team-assignments.csv" | stats count by team | stats avg(count)
```

This should result in 77.984127.