1. Get Started with Kibana

Windows command prompt basics

1. Using Logstash to upload CSV data into Elastisearch
2. Import CSV data into ES using Pandas & Eland, Bulk Import

Export data from ES into Python using Eland

1. Bulk Import single CSV file into ES
2. Bulk Import multiple files into ES at once using looping
3. Get Started with Python.

* Make summary report using Pandas, Grouping & Joining concept
* Dask is good for big data than Pandas, but less support and flexibility

1. Using Joining concept, Removing Duplicate & Filtering method in handling big data

* Remove duplicates
* Filtering out the unnecessary data based on the conditions
* Make partition into smaller dataset and join back into the original big data in order to remove unmatched items. Outer left join is used here
* Add unique key to the data manually if no unique column is found
* No Looping involved in all operations. Higher efficiency, no time-consuming
* Partition big data into smaller sets of data, performing operations & append back into a big data frame.

1. Using basic Regular Expression in masking customer sensitive information
2. Learning how to run Python in Linux
3. Automate repetitive tasks using function in Python
4. Selection of delimiter & different file format extension.
5. Joining data into a big dataset from various sources.

* Selection of unique keys to join

1. Python programming in writing efficient loop using nested loop.
2. Visualizing data using Plotly & Python
3. Moving file from Folder A to Folder B in Windows & Linux
4. Automate the moving file process and let it run continuously in Linux using Crontab & VI editor
5. Learning how PowerBI integrate with Python in data visualization
6. Direct query approach in getting data into PowerBI without loading the data using Dremio and DB
7. Learning QlikView basics functionality
8. Filebeat import CSV into ES directly using PowerShell, JS script
9. Filebeat import CSV/JSON into Logstash and then ES.
10. CSV append JSON
11. Import JSON into ES
12. JSON append JSON
13. Read large JSON file using Python external library *ijson*
14. Transform application log data into DF, which cannot be read by Python in the beginning

* The format is inconsistent, XML-like but not in XML format

1. Transform XML data into DF with key-value format, without key-value format
2. From Redis to ES using Hash type, List type (Redis)

* Using Redis Basics commands

1. Learning difference between Redis & Kafka
2. Learning Redis Publish/Subscribe architecture

* Single app publish data and single client receive data via channel
* Multiple apps publish data into the same channel and single client receive the data

1. Come up pipeline in Clickstream Analysis on PMO website

* Using Divolte to collect data from PMO website
* How to simulate PMO website and modify HTML file to use Divolte
* Groovy programming language used in Divolte .groovy file
* Clickstream Analysis used widely in e-commerce website recommendation function

1. Using curl to by-pass in installing software
2. Data Engineer Learning Path

* <https://www.datacamp.com/community/blog/the-path-to-becoming-a-data-engineer>

1. Free certification by AnalyticsVidya

* [https://www.analyticsvidhya.com/blog/2020/10/4-free-certificate-courses-data-science-- machine-learning-analytics-vidhya/](https://www.analyticsvidhya.com/blog/2020/10/4-free-certificate-courses-data-science--%20machine-learning-analytics-vidhya/)

1. Batch Processing VS Real-Time Processing

Bath Processing handles collected data at once.

Real-time processing handles data whenever there is input (in real-time streaming)

<https://data-flair.training/blogs/batch-processing-vs-real-time-processing/>

1. Apache Flink vs Apache Spark

Flink is the 4G of big data

Spark is the 3g of big data

1. Spark terminology & PySpark SQL

**Basics**

* <https://towardsdatascience.com/a-neanderthals-guide-to-apache-spark-in-python-9ef1f156d427>
* <https://s3.amazonaws.com/assets.datacamp.com/blog_assets/PySpark_SQL_Cheat_Sheet_Python.pdf>

**PySpark\_SQL**

* <https://s3.amazonaws.com/assets.datacamp.com/blog_assets/PySpark_SQL_Cheat_Sheet_Python.pdf>