**Preparation/Analysis Big Data - BIA 6305**

**Leonardo Ji**

**Assignment 3**

**Part 1**

****

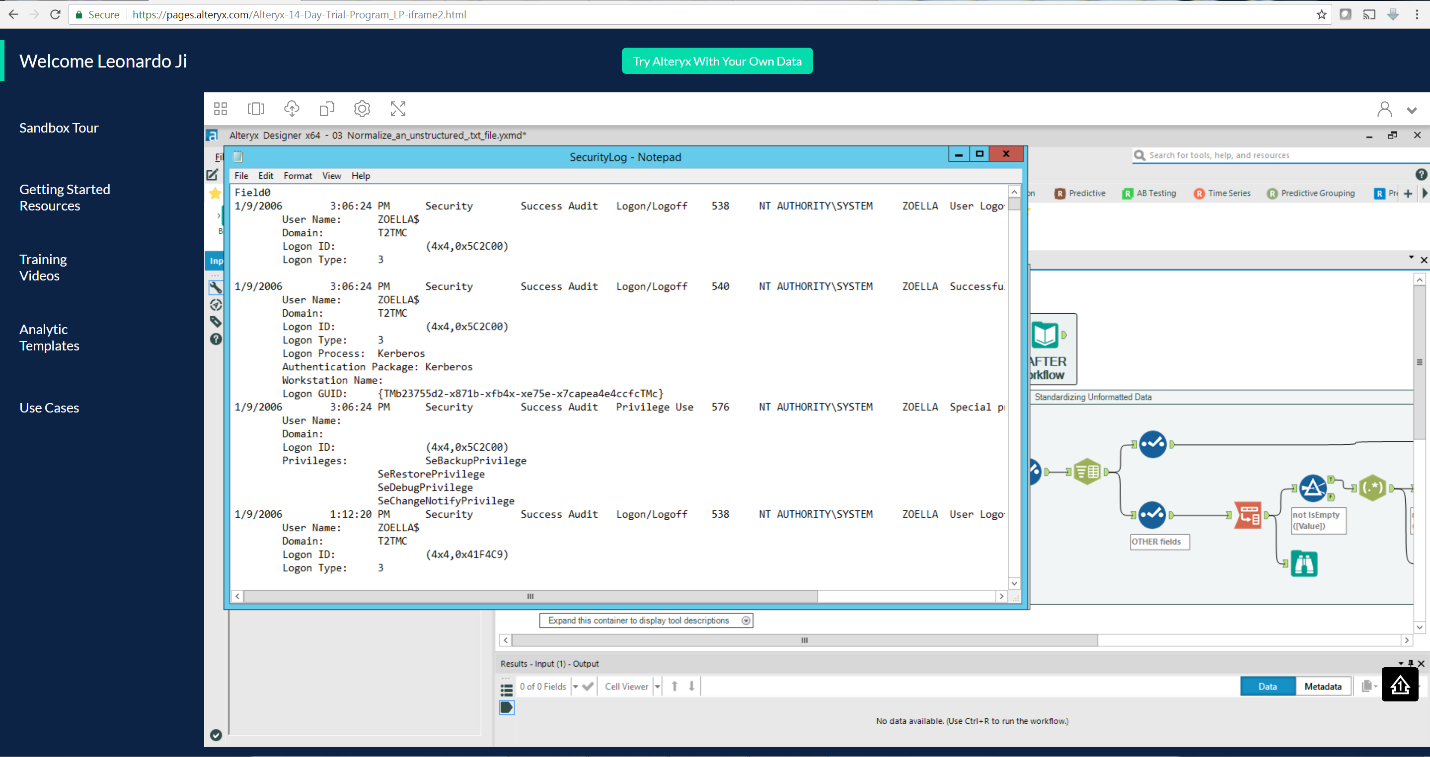
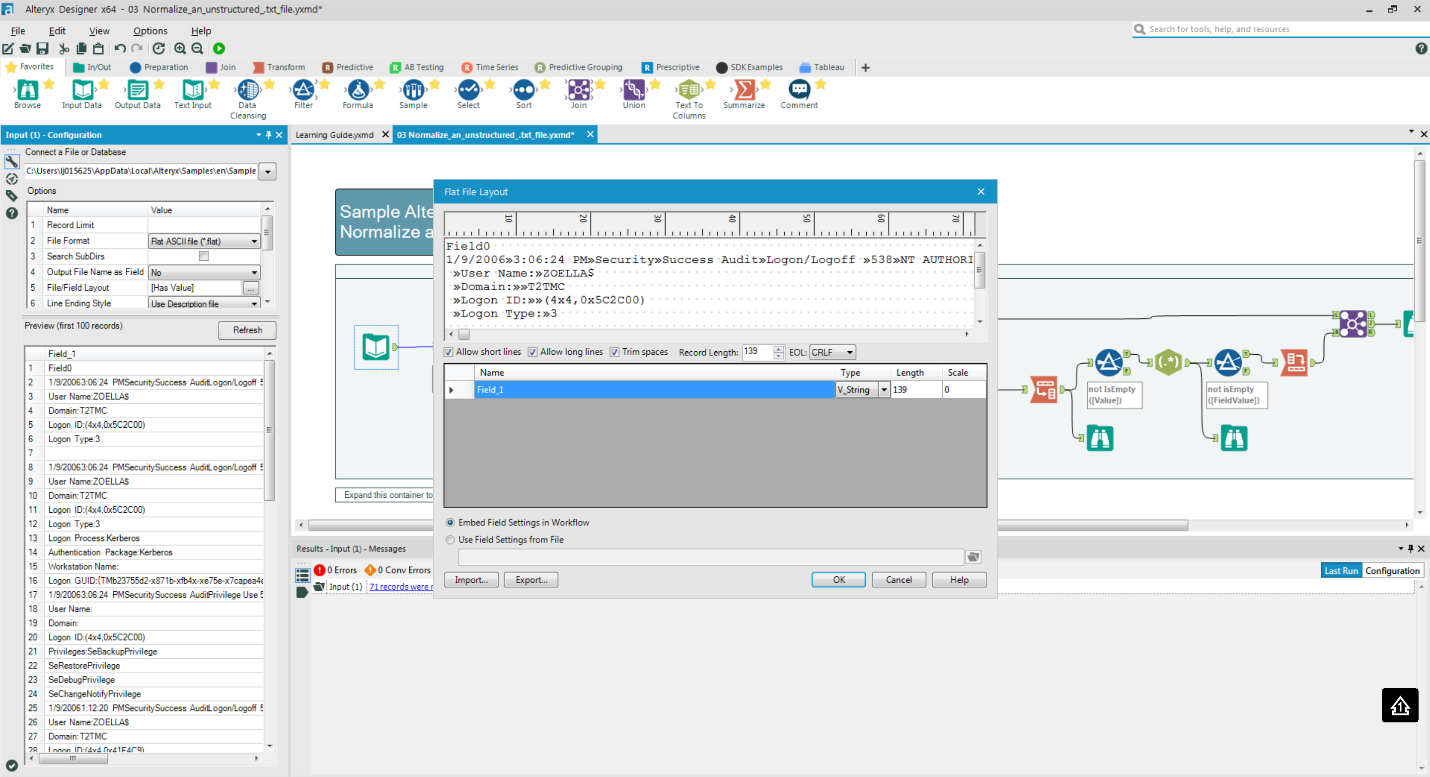
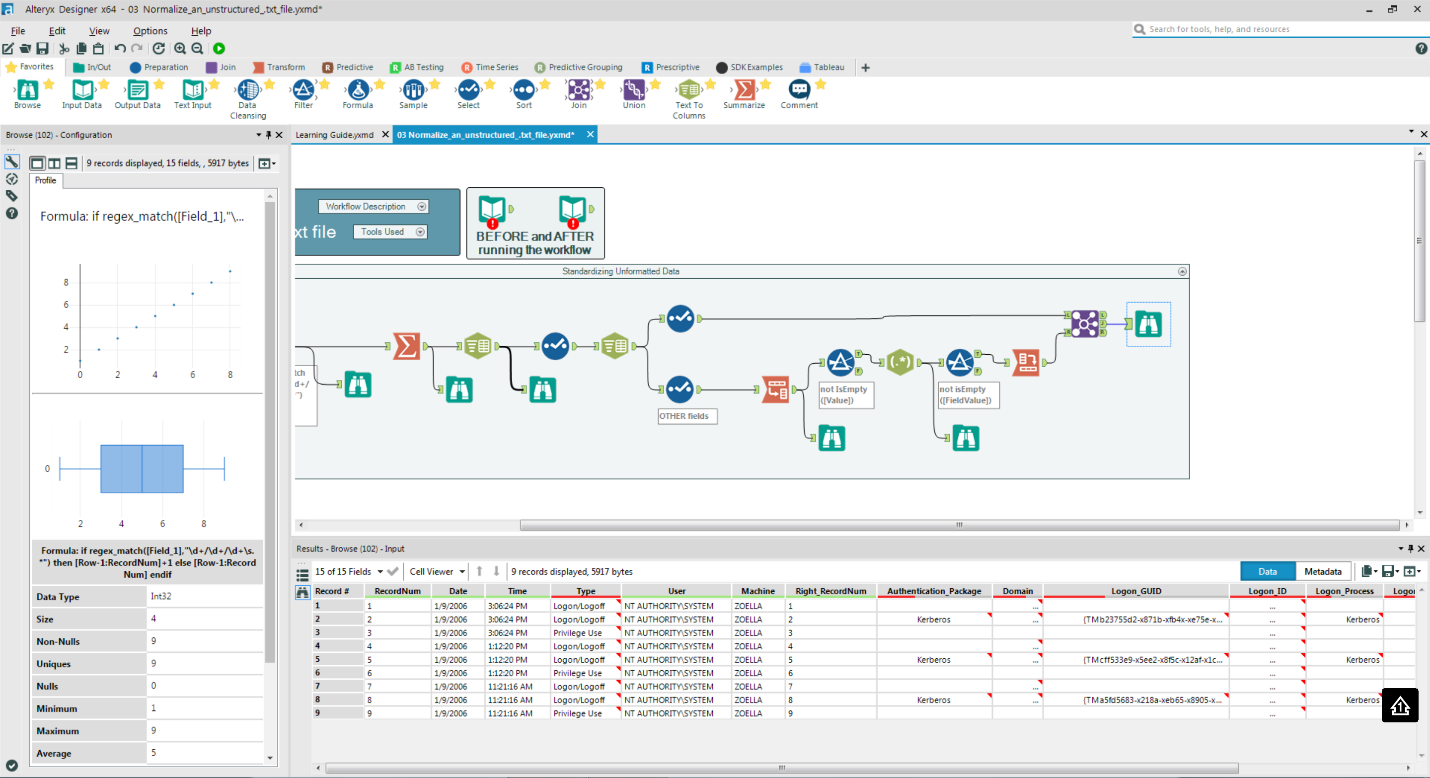
**Part 2**

****

**Part 3** <https://www.alteryx.com/analytics/cloud-data-analytics>

Alteryx is a commercial software that help Data Scientist to do data cleaning, build models without writing Python or R code. Alteryx allow drag and drop component into its workflow. The component is based on R code. The Alteryx installation requires install R version 3.3.2. I downloaded and installed Alteryx desktop version. The example is a machine generated securityLog.txt file. This file is in a flat file format with equal spacing between columns. Alteryx workflow can parse and cleaned the file and output a nice table. We can then save the file to a csv format to be read by any machine learning models.

|  |  |  |
| --- | --- | --- |
| Tools | Pros | Cons |
| Alteryx | * Drag and Drop interface without writing code * Based on Open Source R libraries * Lots of examples * Faster building models to test hypothesis | * Costly commercial license only 14 days free trial * Difficult to navigate * Difficult to customize * The Predictive feature allows someone create model without tune the model |
| Azure ML Studio | * Drag and Drop interface without writing any code * Lots of examples * Good online documentation and tutorials * Cloud version without install software * Faster building models to test hypothesis * Free for testing * Integration with web service for input and output * Can add Python or R code snippets | * Cost money for larger datasets or move to production * It also allows someone create model without tune the model * Where is Grid Search * Difficult to customize * Other libraries are not supported (Tensorflow, MXNet, Kreks…etc.) |
| Open Refine | * Free * Easy interface to navigate * Based on fast Key Collision Clustering or KNN algorithm to find matching | * Can give wrong results because Clustering or KNN algorithms are not semantically-aware * Does not work for other character set languages |

**Part 4**

**Explain what MapReduce does and give an example in your own words or a diagram.**

MapReduce are programs that runs as jobs on Hadoop. MapReduce executes the work in parallel on many nodes in a Hadoop cluster. MapReduce program has two parts: Mapper and Reducer. Mapper part distributes original works into many smaller data. Then the smaller data are passed to the Reducer on each worker node(slaves). The worker node performs works on the smaller data and returns the results. The result is arrived by combining all Reducer results from slaves.

The most famous MapReduce job is word counts on large documents. Many data science algorithms like K-Means clustering job can run in parallel be programmed using a MapReduce job.

**Explain the Importance of YARN.**

YARN (Yet Another Resource Negotiator) is a Hadoop’s resource manager and node manager.

Resource manager has a scheduler and an application manager. Scheduler schedules job based on available resource. Application manager stays open to receive requests to run jobs.

Node manager is like a task tracker. It monitors resources on the node, coordinates the tasks on the node, inform scheduler and application manager on the available resources on the node.