**Mini Project**

Technical Design

Top 10 values

(Extract and Rank top 10 values from data files)

Developed By: Bharat Shailesh Chandora

Creation Date: 01/27/2022

**Introduction**

Having right values are the important to part to live an authentic life. Values such as Trust, Creativity, Curiosity, Honesty etc. defines the personality of person. Working in teams means working with people with different mindsets on a common goal. If a teammate does not understand others members he/she is working with, then this may lead to chaos. A team that share’s more values in common can bond well which then leads to efficient working pattern.

# Document Scope

This document will cover the technical design details related to creating mappings for given subject area.

This document also includes

* the structure of the target table(s)
* the data source(s) for the target table(s)
* join logic between source tables
* the ETL maps that will extract the data from the **EXTRACT**  staging environment to the **LOAD** staging environment for final load.

This document does not include details of the actual code used to extract data from each source table to load the target. Details on each mapping built is added in the xml\_files folder.

**Scenario**

A batch of 45 members, where asked to choose the top 10 values in which they believe in the most. The chosen values were sent to the mentor via e-mail.

All the collected data is given to me via another email from mentor.

**Data Files**

The collected data is consisted of 21 files out which 20 files are simple text file and 1 is excel document.

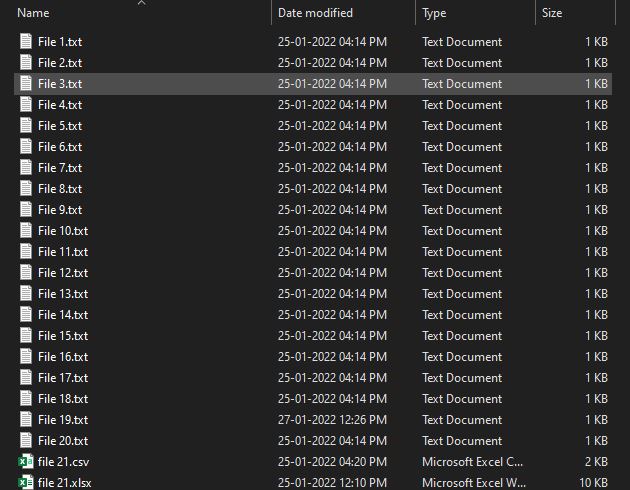


Figure 1: Data Files

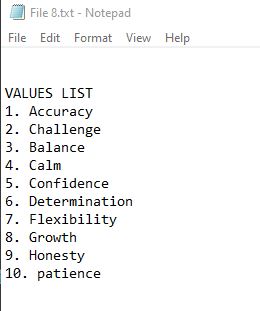
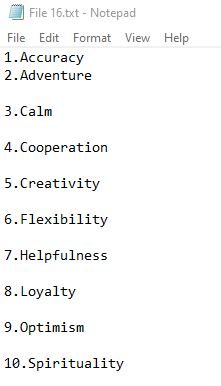
 

Figure 2: Data File Contents

Above shows the content of the files.

**Design** **Logic**

The goal here is to extract top 10 values from the given data. For that I am assigning weights to the values based on the position as shown below:

|  |  |  |
| --- | --- | --- |
| Position | Value | Weights |
| 1 | Calm | 100 |
| 2 | Trust | 90 |
| 3 | Curiosity | 80 |
| 4 | Thankfulness | . |
| 5 | .. | .. |
| 6 | … | So on |

The assignment of weights will allows us to retain the prefrences as given by the members.

Also, I will count the occurance of the values throught out the dataset which will what are the strength of members that believe in partical value.

Combining this two method will provied the top 10 values that most member from batch believe in and also the their preferences.

To solve the problem we first divide the problem in these segments

Data cleaning => Extracting => Tranformation => Loading => Dashboarding

**Data Cleaning**

The process of structing the data from raw data is Data Cleaning. In the project, the raw data has three issue

1. The data in text file is not structred as it contains lots of unwanted information.

2. Some values are repeated.

3. It contains spelling mistakes

The first two issues will be taken care of by JAVA TRANSFORMATION.

Talking about spelling mistakes. It requires to use of spell checking algorithms of a machine learning model to correct the spelling. Which is out of the scope of my knwoledge as of now.

For the time being I have take care of this issue by writing a python scrip findandreplace.py which help me to rectify some common mistakes that I have observed while exploring the data.

**Step 1:**

Store the path of all the text file in single file (sources/output\_file.csv).

I did it using python script file\_path\_name\_list.py

Store the path of the sources/output\_file.csv in sources/pathfile.csv.

The pathfile.csv will act as dummy source file for java transformation as the actual data will be extracted from the files listed in the output\_file.csv file.

**Step 2:**

Import source file (pathfile.csv) in source analyser.

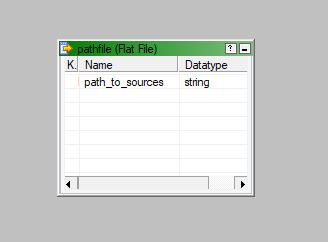


Figure 3: Dummy source file

**Step 3:**

Use Java tranformation to structure the data as per requirement.

The Code of Java transforamtion is available in separate file.

JavaTestCode.txt modifiy this code to use it JavaTransformation

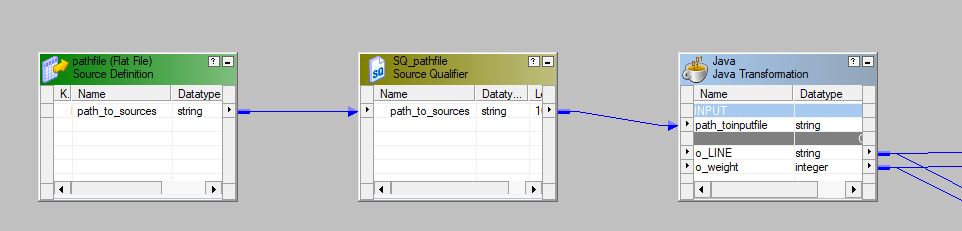


Figure 4: Java Transformation 1

**Step 4:**

Similarly, The file 21.csv which we made by converting the file 21.xlsx to csv file. It was achieved with python script excel\_to\_csv\_converter.py

Import the converted file as source file.

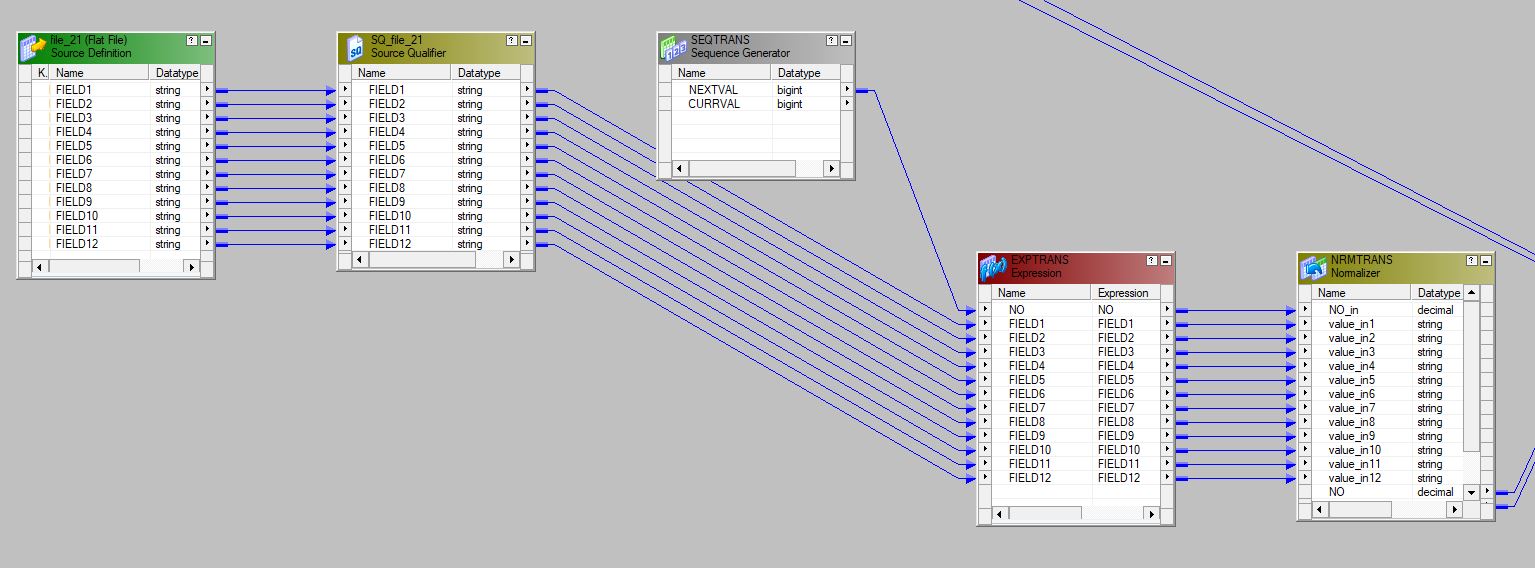


Figure 5: second file source

The above mapping is use to convert the file in required format:

|  |  |
| --- | --- |
| Values | Weights |

**Extraction**

Union Transformation is used to combine all the data from both the types of file formats as one.

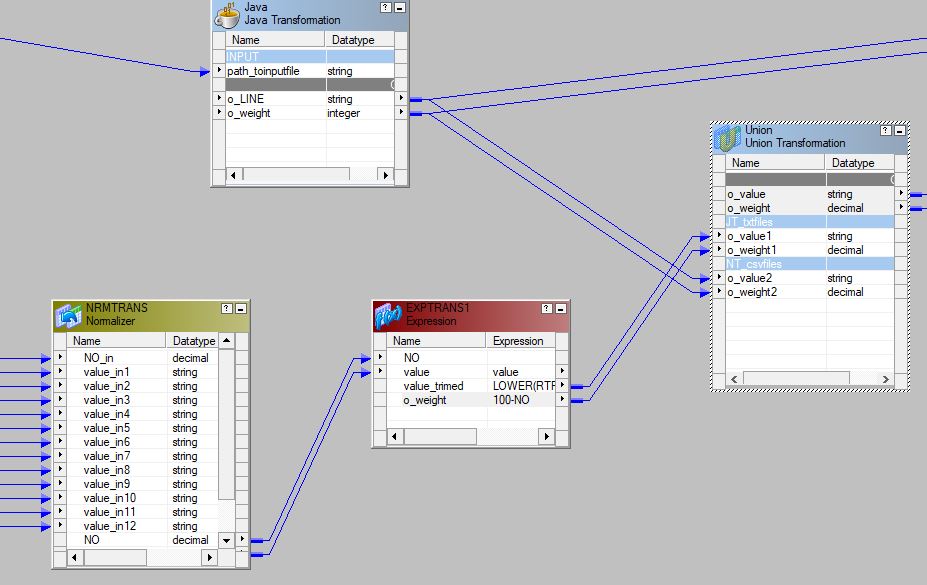


Figure 6: Combining data from files

Here the expression transformation shown accounts for the weights.

**Transformation**

Till here we have extracted all the data in require format. Here we first aggregate the weights for each values and also count the corresponding repetition of the values.

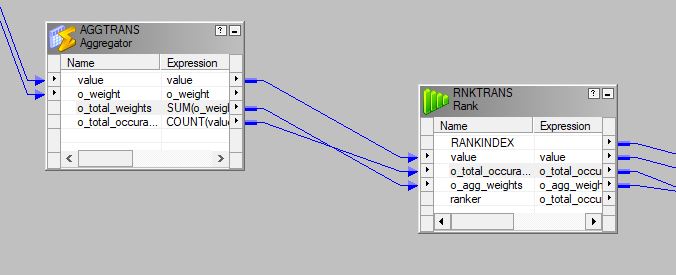


Figure 7: Transformation

The Data is Ranked using rank transformation. The ranking is done based of

(o\_total\_occurance \* o\_agg\_weights)

**Loading**

Data is then loaded in target flat file listing the top 10 core values of team.

