# **Risk and Issues Documentation**

Group Number – **08**

Project - **Medium Data Analytics Leaderboard System**

Team Members – **Piyush Kumar Sultania**

**Gauri Narendra Bhale**

**Nameera Mohsin**

## **Risk/Issue - 01: Unavailability of Source Metadata**

**Statistics:**

|  |  |
| --- | --- |
| **Issue Date** | 03/20/2022 |
| **Type** | Issue |
| **Date Assigned** | 03/21/2022 |
| **Assigned To** | Sultania, Piyush |
| **Impact Level** | Medium |
| **Issue Category** | Profiling |
| **Status** | Closed |

**Risk/Issue:**

**Unavailability of Source Metadata** : Understanding assessments of consistency, accuracy, completeness, and precision

-- Based on the dataset profiling and analysis, it was figured that the metadata for the source dataset is not available at Kaggle where the dataset was retrieved from. It is quite evident that performing data wrangling and analytics without the awareness and visibility on metadata is highly difficult and the reports and visualizations presented won’t be highly efficient

**Mitigation:**

To tackle the issue with unavailability of metadata information, we reached out to the data owner and connected to get the information on the metadata. Although we were not able to fully get the visibility on the data but it was good enough to make sense of the domain and start analyzing and cleansing activities. Additionally, we did some research on the KBEs by the column headers of the source file and also some of the data values (realistic references) and figured the domain of the fields.

## **Risk/Issue - 02: Collaboration**

**Statistics:**

|  |  |
| --- | --- |
| **Issue Date** | 03/15/2022 |
| **Type** | Risk |
| **Date Assigned** | 03/17/2022 |
| **Assigned To** | Mohsin, Nameera |
| **Impact Level** | Low |
| **Issue Category** | Team |
| **Status** | Closed |

**Risk/Issue:**

Collaboration - The project has multiple action items to be accomplished before the deployment like planning, design, execution, testing, documentation, operation and deployment. In order to accomplish high efficiency and accuracy of the project, the team needs to collaborate on assigned tasks adhering the deadlines for each action item. If not collaborated by project team members, the project timeline gets disturbed leading to issues with untimely and inefficient deliverables.

**Mitigation:**

Efficient collaboration of each individual in the team and taking equal accountability of the assigned tasks to complete project in given deadline.

## **Risk/Issue - 03: Python (Jupyter) Memory Allocation Overflow Error**

**Statistics:**

|  |  |
| --- | --- |
| **Issue Date** | 03/25/2022 |
| **Type** | Issue |
| **Date Assigned** | 03/25/2022 |
| **Assigned To** | Sultania, Piyush |
| **Impact Level** | Medium |
| **Issue Category** | Development |
| **Status** | Closed |

**Risk/Issue:**

Python (Jupyter) Memory Allocation Overflow Error: During the data wrangling and cleansing activity, developers came across an issue while using the Pivot/Melt function to pivot the Tag\_\* columns in order to store them as records, where the Jupyter notebook throw memory overflow error due to such significant volume of data. Our dataset contains 1.4 million records with 108 columns out of which we had to use 13 columns as Dimension and Metric (Measures) and the other derived column called 'Category' which will have each active Indicator (all tags which has value =1). So, each record get get multiplicated by ~90 leading to such a significant volume of data not being able to processed by Python and thus getting memory overflow error for all the developers.

**Mitigation:**

The resolution to this issue was limiting the number of records and processing it for the graph database and visualization. We processed around ~700K or 60% of the records, but all random data picked using .random() by python and building the work on the dataset. We were able to process this volume of data. We were able to transform the data and loaded into the eneo4j graph database and performed Exploratory Data Analysis over the database (EDA).

## **Risk/Issue - 04: Python to Neo4j Integration**

**Statistics:**

|  |  |
| --- | --- |
| **Issue Date** | 04/21/2022 |
| **Type** | Issue |
| **Date Assigned** | 04/22/2022 |
| **Assigned To** | Sultania, Piyush |
| **Impact Level** | High |
| **Issue Category** | Development |
| **Status** | Closed |

**Risk/Issue:**

Python to Neo4j Integration:

While integrating to Neo4j, we faced challenges in compatibility of Python version to Neo4j database. Due to incompatible version of Python or Neo4j we were unable to establish the connection between neo4j and python kernel. There were certain libraries which were not imported from py2neo and neo4j which are required to integrate neo4j with python and because we were unable to query the data into neo4j from python notebook

**Mitigation:**

As a resolution, we tried multiple options including upgrading versions of Python to 3.8 and 3.9 with latest version installation of py2neo and neo4j libraries, the final successful integration from python to neo4j completed with installation of Python 3.10 with latest libraries of py2neo and neo4j. Finally, we were able to establish connection between Jupyter and neo4j database, created datasets and visualizations using python libraries.

## **Risk/Issue - 05: Access Provisioning**

**Statistics:**

|  |  |
| --- | --- |
| **Issue Date** | 04/03/2022 |
| **Type** | Risk |
| **Date Assigned** | 04/04/2022 |
| **Assigned To** | Bhale, Gauri |
| **Impact Level** | Low |
| **Issue Category** | Admin |
| **Status** | Closed |

**Risk/Issue:**

Resources need to be provisioned with adequate level of access to various systems and applications. Access to hardware systems, software applications such as Neo4j database, Tableau desktop, MS SQL server is very much important to keep the project on track. If access are not provided at the project initial level it can lead to timeline disruption for deliverables.

**Mitigation:**

All the required clearance to hardware and systems should be provided well in advance.

## **Risk/Issue - 06: Gaps in Data Profiling and Wrangling**

**Statistics:**

|  |  |
| --- | --- |
| **Issue Date** | 03/24/2022 |
| **Type** | Risk |
| **Date Assigned** | 03/25/2022 |
| **Assigned To** | Sultania, Piyush |
| **Impact Level** | Medium |
| **Issue Category** | Development |
| **Status** | Closed |

**Risk/Issue:**

Data profiling and wrangling is one crucial factor in leveraging the data-driven analytics. In order to perform deep data level insights, data need to be profile extensively and wrangled/cleaned in each frame identified during profiling. If the data is not properly cleaned and customized in the analytical format for OLAP, the results can see high deviations and user might lose the value that could have been derived from the data.

**Mitigation:**

Data profiling done using Python libraries (Pandas, NumPy) and based on the profiled data certain corrections were made in the sourced data like (missing values, duplicate data, data casing etc.) and cleaned data was sourced into Neo4j for data analysis and visualization processes.