**Sharepoint Rest API – Change Request**

**Glossary:**

Pandas/PySpark – Library within Python to read Excel files

DataFrame – A data structure constructed with rows and columns similar to a excel spreadsheet

CSV – File type (Comma Separated Values)

JSON – File type (JavaScript Object Notation)

XLSX – File type (Microsoft Excel Spreadsheet)

REST API – Representational State Transfer API, is an application programming interface (API)

Directory – Location/path

ID – Identification

**Description of change and why**

We can input our data into a SharePoint site however we currently are unable to get the data from SharePoint and automate the data that sits in a directory. With the Finance team, we are collaborating on a project that incorporates data from SAP 4 Hana however there are some look up codes that are added to over time that are stored in an excel spreadsheet. This is where the task has occurred from. We want to connect a SharePoint site and connect directly to a file. This keeps the responsibility of the data upload with Finance as opposed to other interventions where Data Engineers must support at the data import stage.

**Risks involved**

|  |  |  |  |
| --- | --- | --- | --- |
| Risk | Impact (1-10) 1 - Low Risk, 10 – High Risk | Outages | Mitigation |
| If SharePoint is not available or resource goes down | 6 | This would be ran daily/weekly therefore we can wait until resource is back up and running | Manually can be uploaded into Databricks to specified table |
| If Databricks goes down | 8 | Unable to load anything manually through Databricks – suggestion is to go to PowerBI and load through there | Use PowerBI to load data |
| If Azure App Registration is removed | 6 | The app would not be available to connect to | Manually can be uploaded into Databricks to specified table |

**Expected outcome**

I would expect to see the REST API to see the files that sit within a specific file path for the desired SharePoint Site. Moreover, I would expect to see the csv/xlsx/json file to be viewable in a Pandas/PySpark DataFrame that can then be saved as table within our Database and be used for Dashboarding and other data uses.

Important to note that any IDs will be securely accessed using our Azure Key Vault that only the Data Engineers can access.

**Testing**

Testing is being done to ensure of a successful connection to the SharePoint site. Further, the Finance team will be informed of the process ensuring that they are happy with the process. Further monitoring and managing will be completed by the Data Engineers using the scheduled jobs ensuring the jobs are running successfully or seeing if the job has failed. Data Engineers will validate the REST API to optimize and ensure successful connection.