# Storage Account Explorer

## Methodology

To extract some information about csv data in the storage account, we pursued the following approach.

* Using an SAS token, we connected to the *hackathonstoragewe* storage account through the Azure Storage Account Python SDK.
* We implemented the functionality to recursively find all CSVs in a container.
* We implemented a *get\_sa\_csv\_erd* function which:
  + Gets all columns from all CSVs (or from a specified list of CSVs);
  + Formats the result into a string that is then passed to the GPT 4 model;
  + Prompts the llm to match columns from different tables with the same or similar names (this is done in the *get\_column\_dependencies\_using\_llm* function). A sample output of the llm is shown in Annex 1.
  + Formats the llm output so that it can be passed to the front-end to render the ERD in the UI. (Note that all table relationships were hardcoded as one-to-one.) A sample llm output is shown in Annex 2. In this case, it only found relationships between the *Webpages-Classification* and *Webpages-Classification-Large* datasets.
* We implemented a *get\_sa\_csv\_descriptions* function which:
  + For a single file, gets each column and 3 sample row values for each column in a standardized formatted way for llm input.
  + Through the *get\_table\_description*  function, prompts the llm for a detailed description of the table based on the column names and sample data. A sample llm output is shown in Annex 3.
* We also implemented the *get\_sa\_erd* and *get\_sa\_csv\_descriptions* endpoints in the *function\_app.py* file (although these endpoints were not integrated with the front-end in time for the final demo).

# Annex 1

{

    "dependencies": [

        {

            "sourceTableName": "Webpages-Classification",

            "sourceColumnName": "url",

            "targetTableName": "Webpages-Classification-Large",

            "targetColumnName": "url"

        },

        {

            "sourceTableName": "Webpages-Classification",

            "sourceColumnName": "url\_len",

            "targetTableName": "Webpages-Classification-Large",

            "targetColumnName": "url\_len"

        },

        {

            "sourceTableName": "Webpages-Classification",

            "sourceColumnName": "ip\_add",

            "targetTableName": "Webpages-Classification-Large",

            "targetColumnName": "ip\_add"

        },

        {

            "sourceTableName": "Webpages-Classification",

            "sourceColumnName": "geo\_loc",

            "targetTableName": "Webpages-Classification-Large",

            "targetColumnName": "geo\_loc"

        },

        {

            "sourceTableName": "Webpages-Classification",

            "sourceColumnName": "tld",

            "targetTableName": "Webpages-Classification-Large",

            "targetColumnName": "tld"

        },

        {

            "sourceTableName": "Webpages-Classification",

            "sourceColumnName": "who\_is",

            "targetTableName": "Webpages-Classification-Large",

            "targetColumnName": "who\_is"

        },

        {

            "sourceTableName": "Webpages-Classification",

            "sourceColumnName": "https",

            "targetTableName": "Webpages-Classification-Large",

            "targetColumnName": "https"

        },

        {

            "sourceTableName": "Webpages-Classification",

            "sourceColumnName": "js\_len",

            "targetTableName": "Webpages-Classification-Large",

            "targetColumnName": "js\_len"

        },

        {

            "sourceTableName": "Webpages-Classification",

            "sourceColumnName": "js\_obf\_len",

            "targetTableName": "Webpages-Classification-Large",

            "targetColumnName": "js\_obf\_len"

        },

        {

            "sourceTableName": "Webpages-Classification",

            "sourceColumnName": "content",

            "targetTableName": "Webpages-Classification-Large",

            "targetColumnName": "content"

        },

        {

            "sourceTableName": "Webpages-Classification",

            "sourceColumnName": "label",

            "targetTableName": "Webpages-Classification-Large",

            "targetColumnName": "label"

        }

    ]

}

# Annex 2

"{\"tables\": [{\"tableName\": \"Historical-Active-Sales\", \"tableComments\": \"\", \"x\": 0, \"y\": 0}, {\"tableName\": \"Webpages-Classification-Large\", \"tableComments\": \"\", \"x\": 0, \"y\": 0}, {\"tableName\": \"Webpages-Classification\", \"tableComments\": \"\", \"x\": 0, \"y\": 0}], \"tableCols\": [{\"tableName\": \"Historical-Active-Sales\", \"columnName\": \"Order\", \"dataType\": \"str\"}, {\"tableName\": \"Historical-Active-Sales\", \"columnName\": \"File\_Type\", \"dataType\": \"str\"}, {\"tableName\": \"Historical-Active-Sales\", \"columnName\": \"SKU\_number\", \"dataType\": \"str\"}, {\"tableName\": \"Historical-Active-Sales\", \"columnName\": \"SoldFlag\", \"dataType\": \"str\"}, {\"tableName\": \"Historical-Active-Sales\", \"columnName\": \"SoldCount\", \"dataType\": \"str\"}, {\"tableName\": \"Historical-Active-Sales\", \"columnName\": \"MarketingType\", \"dataType\": \"str\"}, {\"tableName\": \"Historical-Active-Sales\", \"columnName\": \"ReleaseNumber\", \"dataType\": \"str\"}, {\"tableName\": \"Historical-Active-Sales\", \"columnName\": \"New\_Release\_Flag\", \"dataType\": \"str\"}, {\"tableName\": \"Historical-Active-Sales\", \"columnName\": \"PriceReg\", \"dataType\": \"str\"}, {\"tableName\": \"Historical-Active-Sales\", \"columnName\": \"ReleaseYear\", \"dataType\": \"str\"}, {\"tableName\": \"Historical-Active-Sales\", \"columnName\": \"ItemCount\", \"dataType\": \"str\"}, {\"tableName\": \"Historical-Active-Sales\", \"columnName\": \"LowUserPrice\", \"dataType\": \"str\"}, {\"tableName\": \"Historical-Active-Sales\", \"columnName\": \"LowNetPrice\", \"dataType\": \"str\"}, {\"tableName\": \"Webpages-Classification-Large\", \"columnName\": \"url\", \"dataType\": \"str\"}, {\"tableName\": \"Webpages-Classification-Large\", \"columnName\": \"url\_len\", \"dataType\": \"str\"}, {\"tableName\": \"Webpages-Classification-Large\", \"columnName\": \"ip\_add\", \"dataType\": \"str\"}, {\"tableName\": \"Webpages-Classification-Large\", \"columnName\": \"geo\_loc\", \"dataType\": \"str\"}, {\"tableName\": \"Webpages-Classification-Large\", \"columnName\": \"tld\", \"dataType\": \"str\"}, {\"tableName\": \"Webpages-Classification-Large\", \"columnName\": \"who\_is\", \"dataType\": \"str\"}, {\"tableName\": \"Webpages-Classification-Large\", \"columnName\": \"https\", \"dataType\": \"str\"}, {\"tableName\": \"Webpages-Classification-Large\", \"columnName\": \"js\_len\", \"dataType\": \"str\"}, {\"tableName\": \"Webpages-Classification-Large\", \"columnName\": \"js\_obf\_len\", \"dataType\": \"str\"}, {\"tableName\": \"Webpages-Classification-Large\", \"columnName\": \"content\", \"dataType\": \"str\"}, {\"tableName\": \"Webpages-Classification-Large\", \"columnName\": \"label\", \"dataType\": \"str\"}, {\"tableName\": \"Webpages-Classification\", \"columnName\": \"url\", \"dataType\": \"str\"}, {\"tableName\": \"Webpages-Classification\", \"columnName\": \"url\_len\", \"dataType\": \"str\"}, {\"tableName\": \"Webpages-Classification\", \"columnName\": \"ip\_add\", \"dataType\": \"str\"}, {\"tableName\": \"Webpages-Classification\", \"columnName\": \"geo\_loc\", \"dataType\": \"str\"}, {\"tableName\": \"Webpages-Classification\", \"columnName\": \"tld\", \"dataType\": \"str\"}, {\"tableName\": \"Webpages-Classification\", \"columnName\": \"who\_is\", \"dataType\": \"str\"}, {\"tableName\": \"Webpages-Classification\", \"columnName\": \"https\", \"dataType\": \"str\"}, {\"tableName\": \"Webpages-Classification\", \"columnName\": \"js\_len\", \"dataType\": \"str\"}, {\"tableName\": \"Webpages-Classification\", \"columnName\": \"js\_obf\_len\", \"dataType\": \"str\"}, {\"tableName\": \"Webpages-Classification\", \"columnName\": \"content\", \"dataType\": \"str\"}, {\"tableName\": \"Webpages-Classification\", \"columnName\": \"label\", \"dataType\": \"str\"}], \"columnRelations\": [{\"sourceTableName\": \"Webpages-Classification\", \"sourceColumnName\": \"url\", \"targetTableName\": \"Webpages-Classification-Large\", \"targetColumnName\": \"url\", \"type\": \"ONE\_TO\_ONE\"}, {\"sourceTableName\": \"Webpages-Classification\", \"sourceColumnName\": \"url\_len\", \"targetTableName\": \"Webpages-Classification-Large\", \"targetColumnName\": \"url\_len\", \"type\": \"ONE\_TO\_ONE\"}, {\"sourceTableName\": \"Webpages-Classification\", \"sourceColumnName\": \"ip\_add\", \"targetTableName\": \"Webpages-Classification-Large\", \"targetColumnName\": \"ip\_add\", \"type\": \"ONE\_TO\_ONE\"}, {\"sourceTableName\": \"Webpages-Classification\", \"sourceColumnName\": \"geo\_loc\", \"targetTableName\": \"Webpages-Classification-Large\", \"targetColumnName\": \"geo\_loc\", \"type\": \"ONE\_TO\_ONE\"}, {\"sourceTableName\": \"Webpages-Classification\", \"sourceColumnName\": \"tld\", \"targetTableName\": \"Webpages-Classification-Large\", \"targetColumnName\": \"tld\", \"type\": \"ONE\_TO\_ONE\"}, {\"sourceTableName\": \"Webpages-Classification\", \"sourceColumnName\": \"who\_is\", \"targetTableName\": \"Webpages-Classification-Large\", \"targetColumnName\": \"who\_is\", \"type\": \"ONE\_TO\_ONE\"}, {\"sourceTableName\": \"Webpages-Classification\", \"sourceColumnName\": \"https\", \"targetTableName\": \"Webpages-Classification-Large\", \"targetColumnName\": \"https\", \"type\": \"ONE\_TO\_ONE\"}, {\"sourceTableName\": \"Webpages-Classification\", \"sourceColumnName\": \"js\_len\", \"targetTableName\": \"Webpages-Classification-Large\", \"targetColumnName\": \"js\_len\", \"type\": \"ONE\_TO\_ONE\"}, {\"sourceTableName\": \"Webpages-Classification\", \"sourceColumnName\": \"js\_obf\_len\", \"targetTableName\": \"Webpages-Classification-Large\", \"targetColumnName\": \"js\_obf\_len\", \"type\": \"ONE\_TO\_ONE\"}, {\"sourceTableName\": \"Webpages-Classification\", \"sourceColumnName\": \"content\", \"targetTableName\": \"Webpages-Classification-Large\", \"targetColumnName\": \"content\", \"type\": \"ONE\_TO\_ONE\"}, {\"sourceTableName\": \"Webpages-Classification\", \"sourceColumnName\": \"label\", \"targetTableName\": \"Webpages-Classification-Large\", \"targetColumnName\": \"label\", \"type\": \"ONE\_TO\_ONE\"}]}"

# Annex 3

### Table Description

\*\*Brief Description:\*\*

The table "Webpages-Classification-Large" contains data related to the classification of various webpages. It includes information such as URLs, their lengths, associated IP addresses, geographical locations, top-level domains, WHOIS data, HTTPS status, JavaScript lengths, and content. Each entry is labeled as "good."

\*\*PII:\*\*

- \*\*IP Addresses:\*\* `ip\_add`

- \*\*Geographical Locations:\*\* `geo\_loc`

\*\*Data Granularity:\*\*

The data is at the granularity of individual webpages, with each row representing a unique webpage and its associated metadata.

\*\*Performance Considerations for Calculating KPIs:\*\*

- \*\*URL Length (`url\_len`):\*\* Calculating average or distribution of URL lengths.

- \*\*JavaScript Length (`js\_len` and `js\_obf\_len`):\*\* Summarizing or averaging JavaScript lengths.

- \*\*HTTPS Status (`https`):\*\* Counting the number of secure (HTTPS) vs. non-secure (HTTP) webpages.

- \*\*Geographical Distribution (`geo\_loc`):\*\* Aggregating data based on geographical locations.

\*\*Potential Performance Bottlenecks:\*\*

- \*\*Large Text Fields:\*\* The `content` column contains extensive text data, which can slow down queries that involve text processing or full-text search.

- \*\*IP Address Processing:\*\* Operations involving IP addresses (`ip\_add`) might require specialized indexing or conversion for efficient querying.

- \*\*Geographical Data:\*\* Aggregating and analyzing geographical data (`geo\_loc`) can be computationally intensive, especially if the dataset is large.

- \*\*JavaScript Length Calculations:\*\* Summarizing or processing JavaScript lengths (`js\_len` and `js\_obf\_len`) might require optimization if these fields are frequently queried.

\*\*Areas for Optimization:\*\*

- \*\*Indexing:\*\* Proper indexing on frequently queried columns like `url`, `ip\_add`, and `geo\_loc` can improve query performance.

- \*\*Text Search Optimization:\*\* Implementing full-text search indexes for the `content` column can enhance performance for text-based queries.

- \*\*Data Partitioning:\*\* Partitioning the table based on geographical location or other relevant columns can help in managing large datasets more efficiently.

- \*\*Caching:\*\* Frequently accessed data or results of common queries can be cached to reduce load on the database.