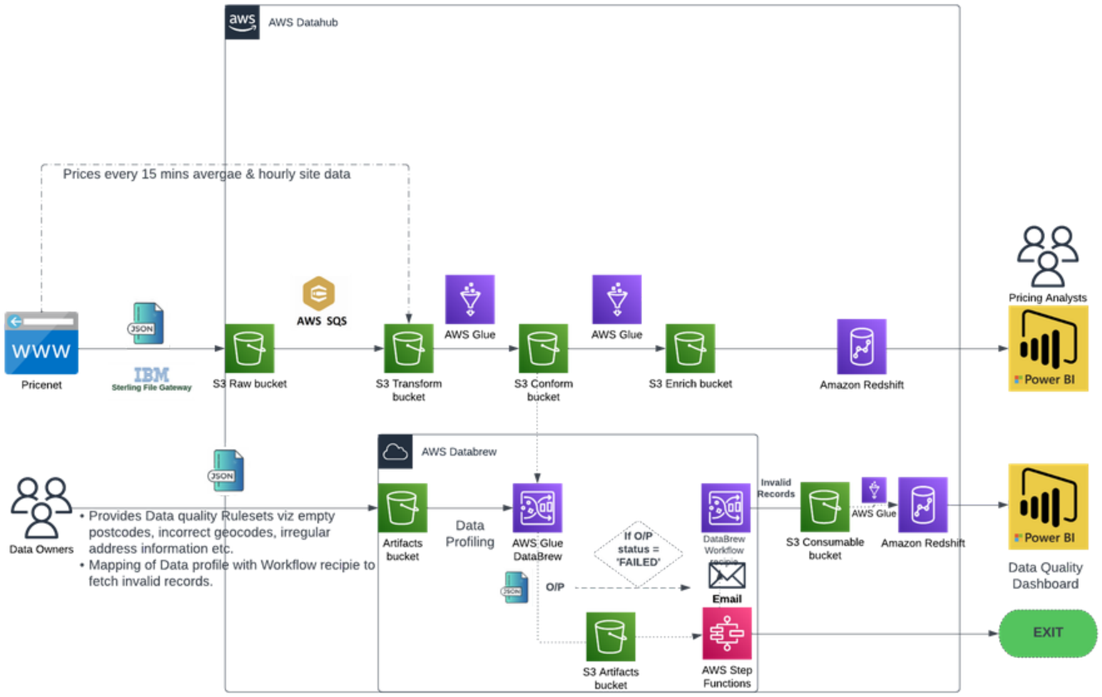
LTA - Data Quality - DataBrew Profiling and Recipe Job Orchestration

Status	IN PROGRESS
Version	1.0
Publish Date	20 April 2023
Author	@Ayhan Onder (Deactivated) @Agnibesh Banerjee
Actions	

Architecture

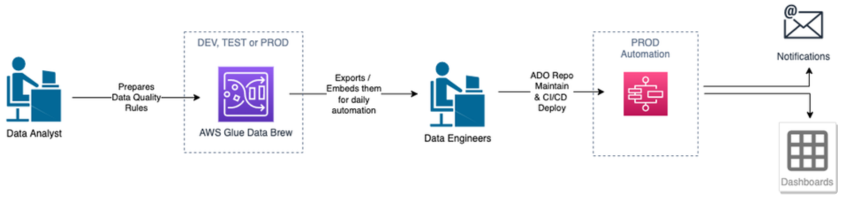


- Provides Data quality Rulesets viz empty postcodes, incorrect geocodes, irregular address information etc.
- Mapping of Data profile with Workflow recipie to fetch invalid records.

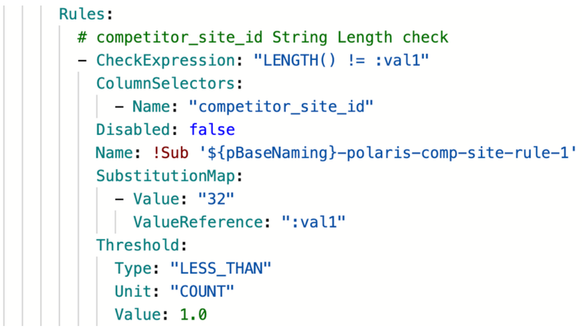
Data Quality Rule Definition Flow

Typically Data Analyst or Data Engineer prepares the data quality rules and relevant recipes using AWS Data Brew either in Dev or Prod environment. Once, template rule definitions are ready, they are taken into the infrastructure as code into the SDLF repository to deploy and maintain via CI/CD pipelines.

To do this; you can simply use YAML export functionalities from AWS Console and AWS CLI commands such as `aws databrew describe-ruleset` You can also easily copy/paste their definitions for multiple countries or leverage the common rules for multiple data sets.

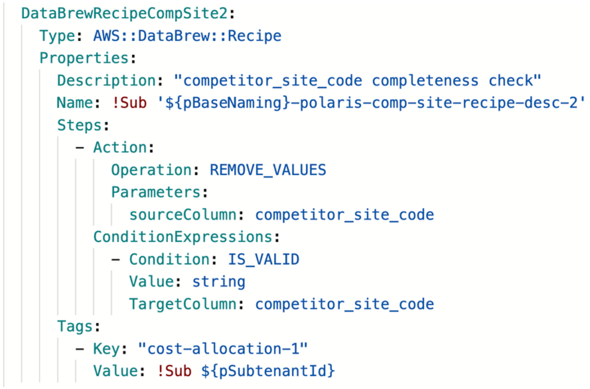

Data Analyst
Prepares Data Quality Rules
DEV, TEST or PROD
AWS Glue Data Brew
Exports / Embeds them for daily automation
Data Engineers
ADO Repo Maintain & CI/CD Deploy
PROD Automation
Notifications
Dashboards

Sample YAML rule definition for cloud formation ;



```
Rules:
  # competitor_site_id String Length check
  - CheckExpression: "LENGTH() != :val1"
    ColumnSelectors:
      - Name: "competitor_site_id"
    Disabled: false
    Name: !Sub '${pBaseNaming}-polaris-comp-site-rule-1'
    SubstitutionMap:
      - Value: "32"
        ValueReference: ":val1"
    Threshold:
      Type: "LESS_THAN"
      Unit: "COUNT"
      Value: 1.0
```

Sample YAML recipe definition for cloud formation ;



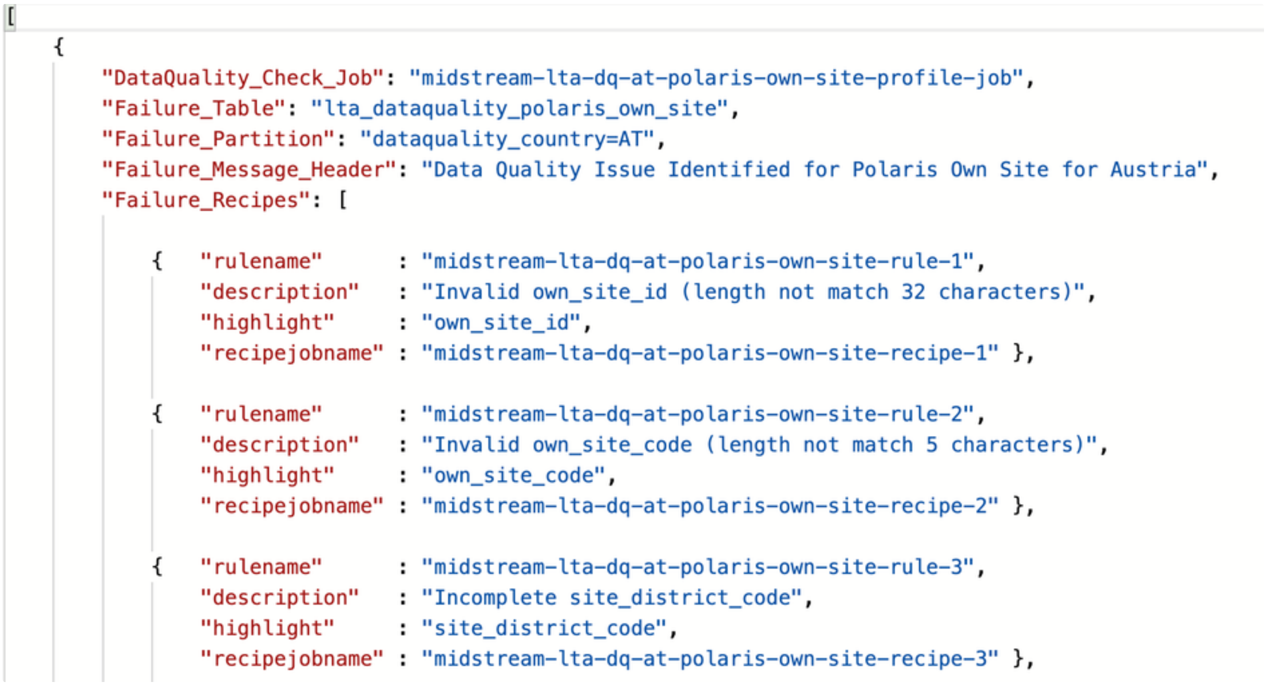
```
DataBrewRecipeCompSite2:
  Type: AWS::DataBrew::Recipe
  Properties:
    Description: "competitor_site_code completeness check"
    Name: !Sub '${pBaseNaming}-polaris-comp-site-recipe-desc-2'
    Steps:
      - Action:
          Operation: REMOVE_VALUES
          Parameters:
            sourceColumn: competitor_site_code
        ConditionExpressions:
          - Condition: IS_VALID
            Value: string
            TargetColumn: competitor_site_code
    Tags:
      - Key: "cost-allocation-1"
        Value: !Sub ${pSubtenantId}
```

Scheduled Orchestration of Data Quality Rules

Orchestration is handled via a state machine / AWS Step Functions which loops through the configuration file.

Configuration file defines the data profiling jobs and their relevant recipes in case any data quality issues is identified. However, recipes in this project, is mostly used for persisting the rows having data quality issues into historical archive/reporting tables, they are not used for fixing the data quality issues. Intention here is to report the issues to be corrected at the source.

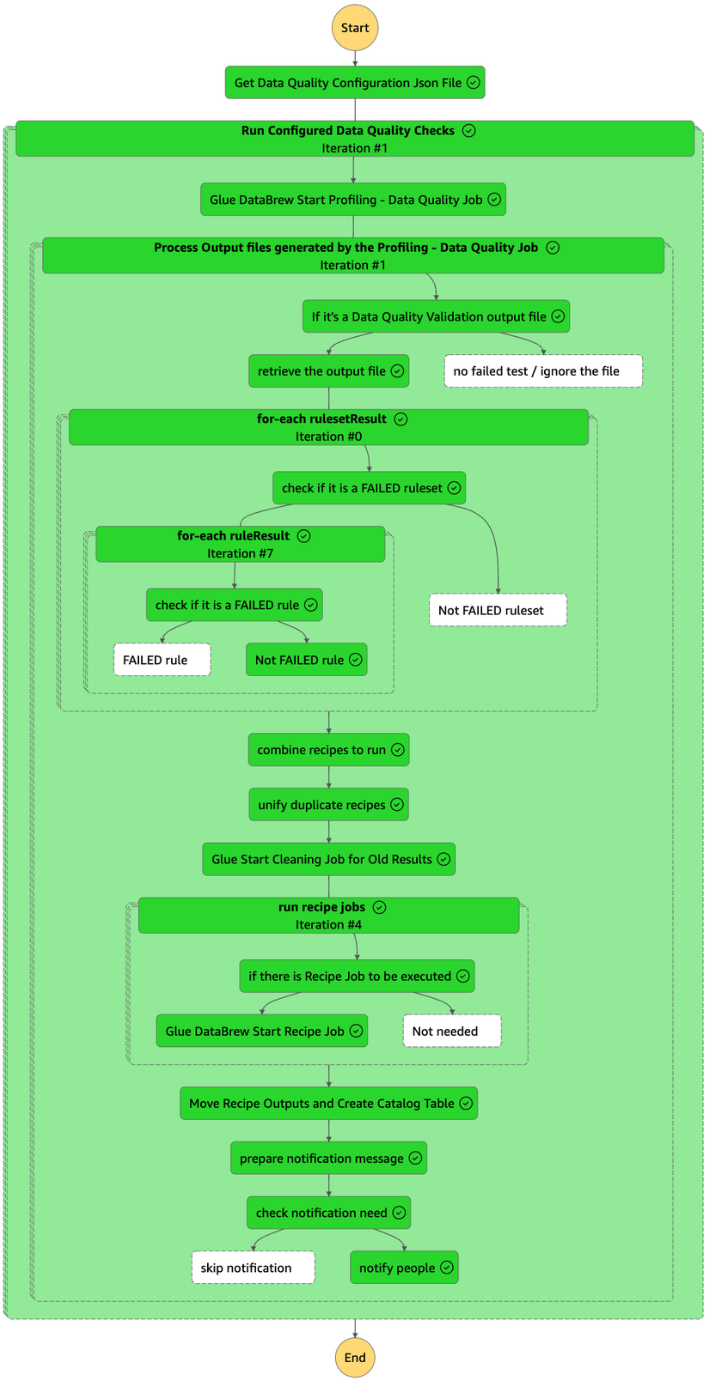
Configuration file is kept in SDLF repository, **data_quality.json** ;



```
[
    {
        "DataQuality_Check_Job": "midstream-lta-dq-at-polaris-own-site-profile-job",
        "Failure_Table": "lta_dataquality_polaris_own_site",
        "Failure_Partition": "dataquality_country=AT",
        "Failure_Message_Header": "Data Quality Issue Identified for Polaris Own Site for Austria",
        "Failure_Recipes": [

            {   "rulename"      : "midstream-lta-dq-at-polaris-own-site-rule-1",
                "description"   : "Invalid own_site_id (length not match 32 characters)",
                "highlight"     : "own_site_id",
                "recipejobname" : "midstream-lta-dq-at-polaris-own-site-recipe-1" },

            {   "rulename"      : "midstream-lta-dq-at-polaris-own-site-rule-2",
                "description"   : "Invalid own_site_code (length not match 5 characters)",
                "highlight"     : "own_site_code",
                "recipejobname" : "midstream-lta-dq-at-polaris-own-site-recipe-2" },

            {   "rulename"      : "midstream-lta-dq-at-polaris-own-site-rule-3",
                "description"   : "Incomplete site_district_code",
                "highlight"     : "site_district_code",
                "recipejobname" : "midstream-lta-dq-at-polaris-own-site-recipe-3" },
```

Jobs and recipes are orchestrated in controlled parallelism using parallel flow task of the AWS Step Functions, main idea here is to ;

- Run the configured profiling jobs, therefore validating all the data quality rules associated with the profiling job
- For each result file produced by profiling job ;
    - If they are Data Quality Validation output file
        - Check their content for any failed validation rule
        - Consolidate list of failed rules, and their respective recipes to be executed
    - Clean old data quality issues from history tables and persist statistics (we keep last 14 days of issues in history tables)
    - If there are data quality issues identified & recipes to be executed ;
        - Run recipes that needs to be triggered
        - Move results to the history table
        - Raise notification via logger notification

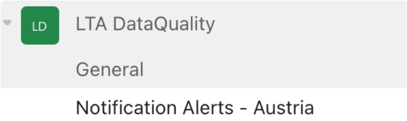

Start
Get Data Quality Configuration Json File
Run Configured Data Quality Checks
Iteration #1
Glue DataBrew Start Profiling - Data Quality Job
Process Output files generated by the Profiling - Data Quality Job
Iteration #1
If it's a Data Quality Validation output file
retrieve the output file
no failed test / ignore the file
for-each rulesetResult
Iteration #0
check if it is a FAILED ruleset
for-each ruleResult
Iteration #7
check if it is a FAILED rule
FAILED rule
Not FAILED rule
Not FAILED ruleset
combine recipes to run
unify duplicate recipes
Glue Start Cleaning Job for Old Results
run recipe jobs
Iteration #4
if there is Recipe Job to be executed
Glue DataBrew Start Recipe Job
Not needed
Move Recipe Outputs and Create Catalog Table
prepare notification message
check notification need
skip notification
notify people
End

Notification Email / Teams Notification

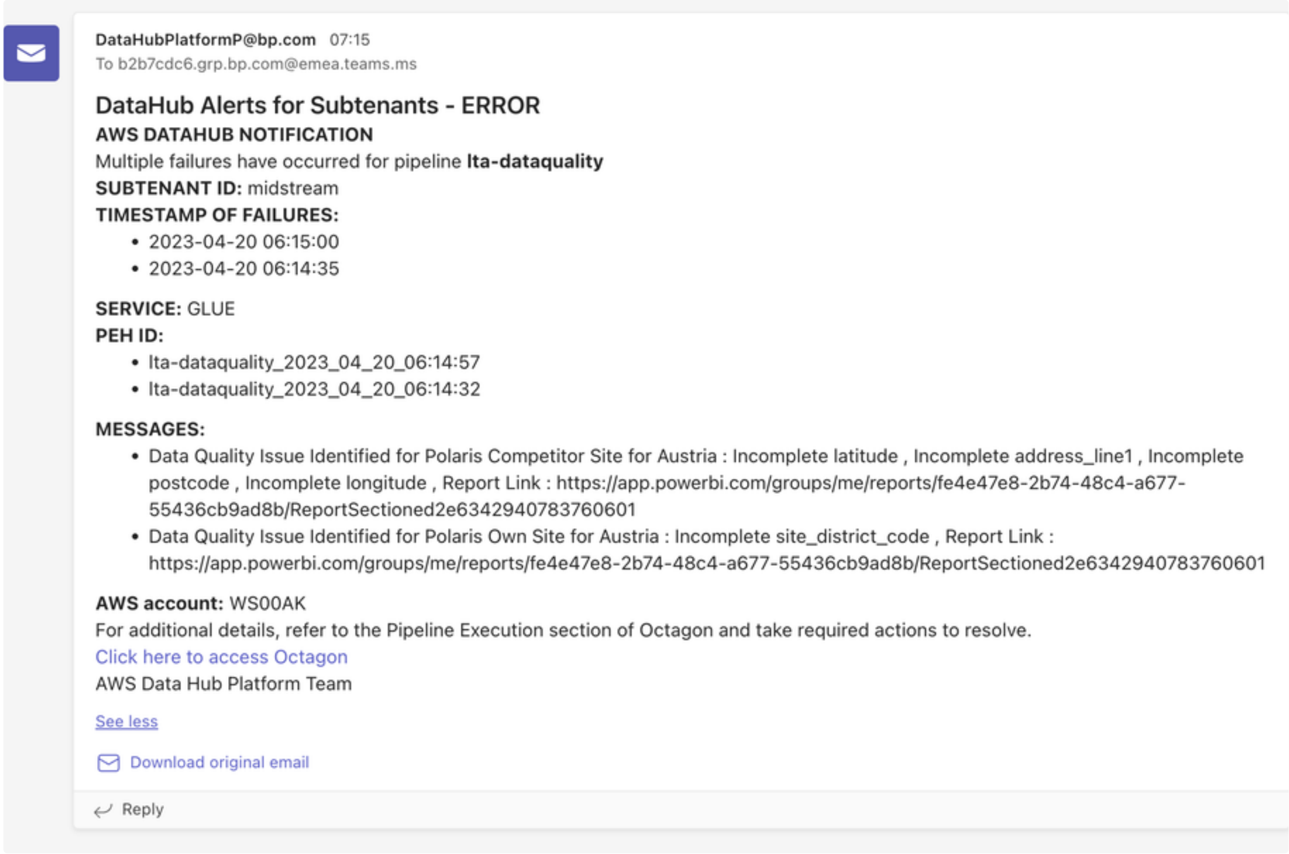
Notifications are sent via email, using the DataHub / SDLF logger notification feature as shown below ;

```
    "prod": {
        "use_case_list": [{
            "name": "lta-dataquality",
            "sdlf_version": "2",
            "custom_logger_notification": {
                "email_target": "b2b7cdc6.grp.bp.com@emea.teams.ms"
            },
            "use_case_name": "b2cpricingreports"
        }]
    }
```

Currently, email messages are targeting the Teams channel email address. The plan is to have specific channel per country and redirect their messages to their dedicated channel using dedicated data quality pipelines.



Notification messages, includes description of issues identified in the message section ;



DataHubPlatformP@bp.com 07:15
To b2b7cdc6.grp.bp.com@emea.teams.ms

DataHub Alerts for Subtenants - ERROR
AWS DATAHUB NOTIFICATION
Multiple failures have occurred for pipeline **lta-dataquality**
SUBTENANT ID: midstream
TIMESTAMP OF FAILURES:

- 2023-04-20 06:15:00
- 2023-04-20 06:14:35

SERVICE: GLUE
PEH ID:

- lta-dataquality_2023_04_20_06:14:57
- lta-dataquality_2023_04_20_06:14:32

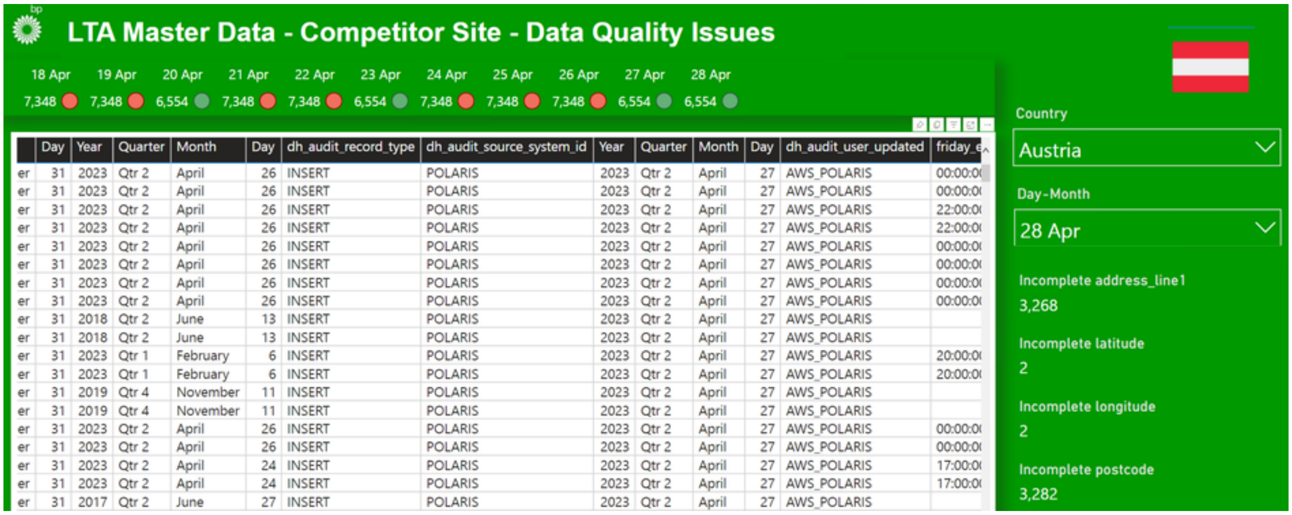
MESSAGES:

- Data Quality Issue Identified for Polaris Competitor Site for Austria : Incomplete latitude , Incomplete address_line1 , Incomplete postcode , Incomplete longitude , Report Link : https://app.powerbi.com/groups/me/reports/fe4e47e8-2b74-48c4-a677-55436cb9ad8b/ReportSectioned2e6342940783760601
- Data Quality Issue Identified for Polaris Own Site for Austria : Incomplete site_district_code , Report Link : https://app.powerbi.com/groups/me/reports/fe4e47e8-2b74-48c4-a677-55436cb9ad8b/ReportSectioned2e6342940783760601

AWS account: WS00AK
For additional details, refer to the Pipeline Execution section of Octagon and take required actions to resolve.
Click here to access Octagon
AWS Data Hub Platform Team

Data Quality Issue Reporting

Issues and statistics are persisted to conform zone tables. Sample PowerBI reports have been created for end user access.



LTA Master Data - Competitor Site - Data Quality Issues

18 Apr	19 Apr	20 Apr	21 Apr	22 Apr	23 Apr	24 Apr	25 Apr	26 Apr	27 Apr	28 Apr
7,348 🔴	7,348 🔴	6,554 🟢	7,348 🔴	7,348 🔴	6,554 🟢	7,348 🔴	7,348 🔴	7,348 🔴	6,554 🟢	6,554 🟢

	Day	Year	Quarter	Month	Day	dh_audit_record_type	dh_audit_source_system_id	Year	Quarter	Month	Day	dh_audit_user_updated	friday_e
er	31	2023	Qtr 2	April	26	INSERT	POLARIS	2023	Qtr 2	April	27	AWS_POLARIS	00:00:0
er	31	2023	Qtr 2	April	26	INSERT	POLARIS	2023	Qtr 2	April	27	AWS_POLARIS	00:00:0
er	31	2023	Qtr 2	April	26	INSERT	POLARIS	2023	Qtr 2	April	27	AWS_POLARIS	22:00:0
er	31	2023	Qtr 2	April	26	INSERT	POLARIS	2023	Qtr 2	April	27	AWS_POLARIS	22:00:0
er	31	2023	Qtr 2	April	26	INSERT	POLARIS	2023	Qtr 2	April	27	AWS_POLARIS	00:00:0
er	31	2023	Qtr 2	April	26	INSERT	POLARIS	2023	Qtr 2	April	27	AWS_POLARIS	00:00:0
er	31	2023	Qtr 2	April	26	INSERT	POLARIS	2023	Qtr 2	April	27	AWS_POLARIS	00:00:0
er	31	2023	Qtr 2	April	26	INSERT	POLARIS	2023	Qtr 2	April	27	AWS_POLARIS	00:00:0
er	31	2018	Qtr 2	June	13	INSERT	POLARIS	2023	Qtr 2	April	27	AWS_POLARIS	
er	31	2018	Qtr 2	June	13	INSERT	POLARIS	2023	Qtr 2	April	27	AWS_POLARIS	
er	31	2023	Qtr 1	February	6	INSERT	POLARIS	2023	Qtr 2	April	27	AWS_POLARIS	20:00:0
er	31	2023	Qtr 1	February	6	INSERT	POLARIS	2023	Qtr 2	April	27	AWS_POLARIS	20:00:0
er	31	2019	Qtr 4	November	11	INSERT	POLARIS	2023	Qtr 2	April	27	AWS_POLARIS	
er	31	2019	Qtr 4	November	11	INSERT	POLARIS	2023	Qtr 2	April	27	AWS_POLARIS	
er	31	2023	Qtr 2	April	26	INSERT	POLARIS	2023	Qtr 2	April	27	AWS_POLARIS	00:00:0
er	31	2023	Qtr 2	April	26	INSERT	POLARIS	2023	Qtr 2	April	27	AWS_POLARIS	00:00:0
er	31	2023	Qtr 2	April	24	INSERT	POLARIS	2023	Qtr 2	April	27	AWS_POLARIS	17:00:0
er	31	2023	Qtr 2	April	24	INSERT	POLARIS	2023	Qtr 2	April	27	AWS_POLARIS	17:00:0
er	31	2017	Qtr 2	June	27	INSERT	POLARIS	2023	Qtr 2	April	27	AWS_POLARIS	

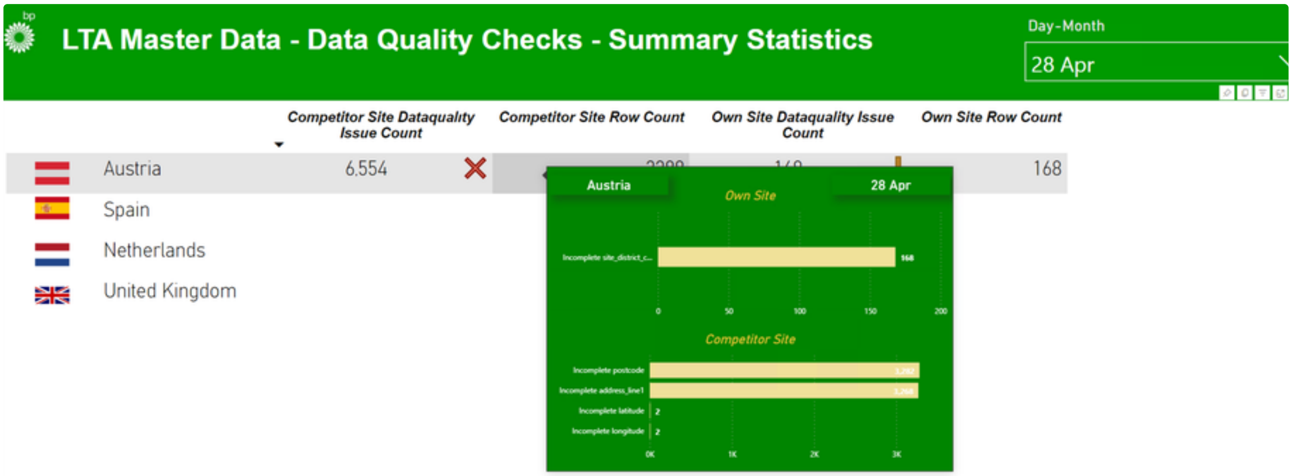
Country: Austria
Day-Month: 28 Apr

Incomplete address_line1
3,268

Incomplete latitude
2

Incomplete longitude
2

Incomplete postcode
3,282



LTA Master Data - Data Quality Checks - Summary Statistics

Day-Month: 28 Apr

	Competitor Site Dataquality Issue Count	Competitor Site Row Count	Own Site Dataquality Issue Count	Own Site Row Count
Austria	6.554 ✗	[illegible]	[illegible]	168
Spain				
Netherlands				
United Kingdom				

PowerBI Sample Report Location

The origin pbix file of published report can be found in documents folder of Agnibesh’s VDI. Another backup copy is available at the ADO repository below;

Repo

Pipeline Source Code Location

lta-dataquality