Tech preview: 
ML based data rule definition generator 
IBM® Cloud Private for Data 3.0

Note: This is a technology preview and is not yet supported for use in production environments.

Overview

For data in tabular structures, relationship exists at column level. Some of the relationships are straightforward and can easily be represented via simple rules or non-complex mathematical expressions. Other relationships between the columns might be complex and not easy to represent with simple expression.

Data rules and quality rules are used to validate and check certain user-defined conditions in your data sets. For straightforward relationships, users can define simple conditions in the rules and apply such rules on relevant data sets. For more complex relationships, no simple conditions can be defined. Thus, violations of such relationships are difficult to identify and thus do not play a part in data quality evaluation.

With the ML based Data Rule Definitions feature, you can create and train an ML model to identify the relationship between the columns and generate a data rule definition based on this model. You can then create data rules to bind that definition to data in similar data assets when you run the rule in data quality evaluation. Thus, complex relationships are considered in data quality evaluation and contribute to determining the overall quality score of the data asset.

For more information about rules, see Using rules in the IBM Cloud Pak for Data documentation.
Generate the data rule definition

For proper training of your model, make sure to use a data asset with sufficient representative data. Training the model on a small amount of data will work, but the resulting model will be less robust and reliable.

Prerequisite tasks:

Generate a bearer token

Generate the access token by running the following command. Replace IIS_URL, IIS_PORT, IIS_USERNAME, and IIS_PASSWORD with your values.

```bash
```

Where:
- **IIS_URL**: The hostname or IP address of the server where Cloud Pak for Data is installed.
- **IIS_PORT**: The port number to be used on the server where Cloud Pak for Data is installed
- **IIS_USERNAME**: Cloud Pak for Data username
- **IIS_PASSWORD**: Cloud Pak for Data password

Add the user `isadmin` to the data quality project

1. Go to Organize > Data quality > `project_name` > Settings > Users and groups.
2. Click Add users and groups.
3. On the Users tab, search for the user `isadmin`.
4. Select the entry and click Add.

Required role:

Administrator or Data Quality Analyst (Analyze data quality permission)
Submit a POST request to the /v3/insights/data_rule_definitions endpoint. You must pass the Authorization header with the Bearer token in the request. The request body must be a JSON object containing the following information:

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>workspace_rid</td>
<td>The ID of the data quality project in which you want to create the data rule definition. The project ID is part of the URL when you access a project: https://host_name/ibm/iis/quality/workspace/&lt;workspace_rid&gt;</td>
</tr>
<tr>
<td>drd_name</td>
<td>The name you want the data rule definition to have.</td>
</tr>
<tr>
<td>drd_description</td>
<td>A description of the data rule definition.</td>
</tr>
<tr>
<td>Host (optional)</td>
<td>The hostname or IP address of the database housing the data asset.</td>
</tr>
<tr>
<td>Port (optional)</td>
<td>Database port number.</td>
</tr>
<tr>
<td>username (optional)</td>
<td>The database username for accessing the data asset.</td>
</tr>
<tr>
<td>password (optional)</td>
<td>Database password.</td>
</tr>
<tr>
<td>asset_id</td>
<td>The ID of the data asset that serves as training data. The asset ID is part of the URL when you access a data asset under Information assets: https://host_name/ibm/iis/igcui/assets/&lt;asset_id&gt;</td>
</tr>
<tr>
<td>columns</td>
<td>The columns for which the relationship is to be learned and translated into a data rule definition.</td>
</tr>
</tbody>
</table>

**cURL example:**

curl -X POST --header 'Content-Type: application/json' --header 'Accept: application/json' --header 'Authorization: Bearer <bearer_token>' -d '{
  "workspace_rid": "ec1481df.64b1b87d.bc070msl2.vbpioa5.cwui26.i6ir7c517oqlm2u3f9ut6",
  "drd_name": "salary_constraint",
  "drd_description": "Data rule to learn salary constraint",
  "data_set": {
    "asset_id": "b1c497ce.54bd3a08.bc070n0fi.48ublkq.niiffh.Bbk8q7vejscti10gk3c3b",
    "columns": [
      "EXPERIENCE>Total",
      "SALARY",
      "SUBCATEGORY"
    ],
    "host": "198.51.100.0",
    "port": "50000"
  }
}' 'https://<wkc_server>/v3/insights/data_rule_definitions'

If you omit any optional parameters, these are fetched from the catalog if present. If the information is not available in the catalog, an error is returned. Resubmit your request providing the required information.

© Copyright IBM Corporation 2020.
US Government Users Restricted Rights – Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.
Check the status of training

On successful submission, the POST request to the /v3/insights/data_rule_definitions endpoint returns a task ID along with the data rule definition ID in JSON format:

```json
{
   "drd_id": "<drd_guid>",
   "task_id": "<task_guid>
}
```

To check the status of the training, submit a GET request to the /v3/insights/data_rule_definitions/status/{drd_id} endpoint or the /v3/insights/tasks/{task_id} endpoint. You must pass the Authorization header with the Bearer token in the request to retrieve the status of the submitted training.

**cURL examples:**

```
```

Get the list of generated data rules

To obtain the list of generated data rule definitions, submit a GET request to the /v3/insights/data_rule_definitions endpoint. You must pass the Authorization header with the Bearer token in the request.

**cURL example:**

```
```
View the data rule definition

The generated data rule definition is available in the Data rules tab of the project for which you created the definition.

To check the definition:
1. Go to Organize > Data quality > project name > Data rules.
2. Find the definition in the All folder in the project and select View from the menu.
3. In the Rule logic tab, check the definition. It should look similar to this example:

   ![Rule Logic Example](image)

   The rule expression returns 1 if the relationship between the specified columns is intact per the learning of the ML algorithm on the training data asset. Otherwise, 0 is returned.
4. Leave the definition.

Retrain the ML model on new data

To invoke the routine to retrain the model associated with the previously generated data rule definition, submit a POST request to the /v3/insights/data_rule_definitions/{drd_id} endpoint with the same data asset information that you specified in the training request. You must pass the Authorization header with the Bearer token with the request.

If you want the training to happen on a different data asset, change the request body so that it looks like this:

```json
{
  "data_asset": {
    "username": "username",
    "password": "password",
    "host": "hostname",
    "port": "port",
    "asset_id": "alternative_asset_ID",
    "columns": [
      "column"
    ]
  }
}
```
cURL example:

curl -X POST --header 'Content-Type: application/json' --header 'Accept: application/json' --header 'Authorization: dqsds' -d '{
  "data_asset": {
    "username": "dqa1",
    "password": "pwd_string",
    "host": "198.51.100.0",
    "port": "50000",
    "asset_id": "ec1481df.0e6c3ac.01a70prio.p1v84so.5m03uu.sf4mdv05sdf6svpe42l57",
    "columns": ["EXPERIENCE_TOTAL", "SALARY", "SUBCATEGORY"
  ]
}
}' 'https://<wkc_server>/v3/insights/data_rule_definitions/{drd_id}'
Add a data or quality rule for data quality evaluation

Add a data or quality rule that uses the generated definition.

Complete these steps:
1. Go to Organize > Data quality > project name > Data assets > data asset name. Select a data asset similar to the one you ran the training on.
2. Click Add rule.
3. From the list of rule types to create, select Data rule or Quality rule and click Next.
4. If you selected to add a data rule, provide a name for the rule and click Next.
5. Bind the columns of the data rule definition to the respective columns in your test data. Click Next.
6. If you are adding a data rule, specify the required settings for joins, output content, and output settings.
7. Test the rule. The test results show the number of rows evaluated, the number of rows that met the criteria, and the number of rows with violations.
8. Save the rule.

Delete the generated data rule definition

To delete the generated data rule definition, submit a DELETE request to the /v3/insights/data_rule_definitions/{drd_id} endpoint. You must pass the Authorization header with the Bearer token in the request.