erwin Data Intelligence – erwin Data Quality
Best Practices Guide

Legal Notices

This Documentation, which includes embedded help systems and electronically distributed materials (hereinafter referred to as the Documentation), is for your informational purposes only and is subject to change or withdrawal by Quest Software, Inc and/or its affiliates at any time. This Documentation is proprietary information of Quest Software, Inc and/or its affiliates and may not be copied, transferred, reproduced, disclosed, modified or duplicated, in whole or in part, without the prior written consent of Quest Software, Inc and/or its affiliates.

If you are a licensed user of the software product(s) addressed in the Documentation, you may print or otherwise make available a reasonable number of copies of the Documentation for internal use by you and your employees in connection with that software, provided that all Quest Software, Inc and/or its affiliates copyright notices and legends are affixed to each reproduced copy.

The right to print or otherwise make available copies of the Documentation is limited to the period during which the applicable license for such software remains in full force and effect. Should the license terminate for any reason, it is your responsibility to certify in writing to Quest Software, Inc and/or its affiliates that all copies and partial copies of the Documentation have been returned to Quest Software, Inc and/or its affiliates or destroyed.

TO THE EXTENT PERMITTED BY APPLICABLE LAW, QUEST SOFTWARE, INC. PROVIDES THIS DOCUMENTATION AS IS WITHOUT WARRANTY OF ANY KIND, INCLUDING WITHOUT LIMITATION, ANY IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, OR NONINFRINGEMENT. IN NO EVENT WILL QUEST SOFTWARE, INC. BE LIABLE TO YOU OR ANY THIRD PARTY FOR ANY LOSS OR DAMAGE, DIRECT OR INDIRECT, FROM THE USE OF THIS DOCUMENTATION, INCLUDING WITHOUT LIMITATION, LOST PROFITS, LOST INVESTMENT, BUSINESS INTERRUPTION, GOODWILL, OR LOST DATA, EVEN IF QUEST SOFTWARE, INC. IS EXPRESSLY ADVISED IN ADVANCE OF THE POSSIBILITY OF SUCH LOSS OR DAMAGE.

The use of any software product referenced in the Documentation is governed by the applicable license agreement and such license agreement is not modified in any way by the terms of this notice.

The manufacturer of this Documentation is Quest Software, Inc and/or its affiliates Provided with Restricted Rights. Use, duplication or disclosure by the United States Government is subject to the restrictions set forth in FAR Sections 12.212, 52.227-14, and 52.227-19(c)(1) - (2) and DFARS Section 252.227-7014(b)(3), as applicable, or their successors.

Contact erwin

Understanding your Support

Review support maintenance programs and offerings.

Registering for Support

Access the erwin support site and click Sign in to register for product support.

Accessing Technical Support

For your convenience, erwin provides easy access to "One Stop" support for erwin Data Intelligence (erwin DI), and includes the following:

- Online and telephone contact information for technical assistance and customer services
- Information about user communities and forums
- Product and documentation downloads
- erwin Support policies and guidelines
- Other helpful resources appropriate for your product

For information about other erwin products, visit http://erwin.com/.

Provide Feedback

If you have comments or questions, or feedback about erwin product documentation, you can send a message to distechpubs@erwin.com.

erwin Data Modeler News and Events

Visit www.erwin.com to get up-to-date news, announcements, and events. View video demos and read up on customer success stories and articles by industry experts.

Contents

Introduction

This document is intended to serve as a best practice guide for the installation, setup, and configuration of the DQLabs platform that has been performed.

Modules Involved

The document covers the best practices in the following sections of the portal

- Installation
- Platform Configuration
- Integration and Sensing
- Profile and Monitoring
- Curation and Push
- Integration via API
- Troubleshooting

Installation

Please follow the below items while doing a Windows or Linux install

Windows Installation

Checks to make while installing DQLabs on the Windows environment

- Make sure that the system and software requirements are as per the installation guide
- The Installation of the DQLabs platform in a separate server is recommended
- Always run Windows PowerShell ISE with admin privileges to install the DQLabs platform
- While installing DQLabs, please install in a drive other than "C" and verify the version number to be installed from the Erwin Quest team
- Ensure that the environmental variables are properly setup as per the installation guide
- For a fresh install, it is recommended to have the machine clean and not to manually install any software that is mentioned in the installation guide
- Always check if the portal is accessible after the installation

Linux Installation

Checks to make while installing DQLabs on the Linux environment

- Make sure that the system and software requirements are as per the installation guide
- The Installation of the DQLabs platform in a separate server is recommended
- For a fresh install, it is recommended to have the machine clean and not to manually install any software that is mentioned in the installation guide
- Always check if the portal is accessible after the install
- Ensure to check if the Livy server is functioning and the execution status is "OK"
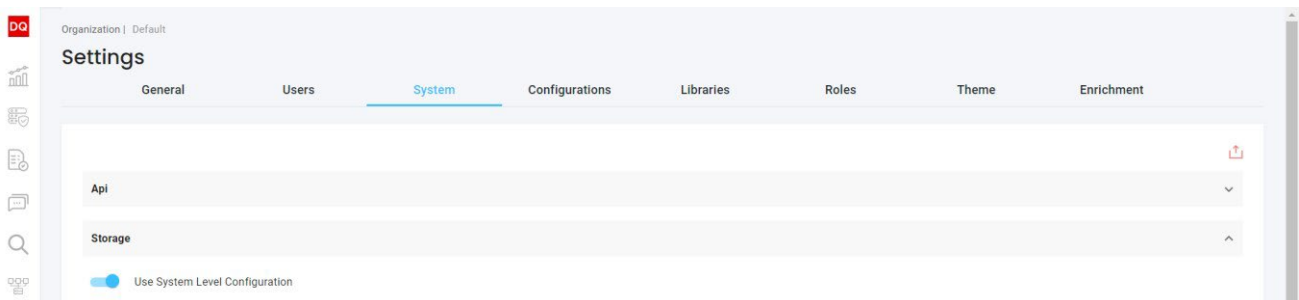
First Run

Checks to make once the installation is complete

- Use the IP address to verify that the installation is done correctly
- Login to the platform using the email and password set up during installation and activate the license
- Configure a demo dataset to see if all the jobs are running properly
- Check the Livy /Spark job server to see the status of the jobs while running
- Once all jobs in the demo dataset have been completed, check the profile, attributes, and properties page to see if all the data is populated
- The above checks confirm that the portal is installed properly

Platform Configuration

- Ensure that the license provided is a "Single Tenant" license
- If the License is a "Single Tenant Single User" ensure that the storage is set to "Use System Level Configuration" in the system page of the organization

- Set the "Dataset Percentage" under the configuration page to 10 to pull only a snapshot of the data unless the customer wants to profile the whole data
- Set the "Preview Count" as 20 in the configuration's page
- Set the Max Concurrent jobs as 5 for better queuing

- With the Max Concurrent job at 2 and the system configuration of 16 core 32 GB RAM, DQLabs supports upto 13 GB file (Example: 75 Million rows and 18 columns).
- With the Max Concurrent job at 1 and the system configuration of 16 core 32 GB RAM, DQLabs supports upto 8.7 GB file (Example: 50 Million rows and 18 columns).

- Follow the settings mentioned below for the "System" and "Configuration" recommended settings

Note: (*) indicated module/sub-module is applicable for the professional-level license.

S.NO	MODULE	SUB MODULE	FEATURE	YES	NO	VALUE	COMMENTS
1	Properties	Preview count				20	
2		Dataset Percentage				10	
3		Drift days count(*)				3	
4		Max Concurrent Jobs				5	
5		Skip Drift Date(*)			NO		
6		Push Permission(*)			NO		
7		Host Address				Application URL Eg: https://dqlabs-qadock.qaserverix.co	
8		DNS Name				Application URL Eg: https://dqlabs-qadock.qaserverix.co	
9		Max Time Limit (Mins)				60	
10		Check Job Status (Mins)				45	
11		Include Organisation and Domain Score (*)		YES			
12		Enable event based Scheduling (*)		YES			
13		Enable Custom Export Metrics (*)		YES			

14	Sense	Percentage Change		YES		0	
15		Schema Change			NO		
16		Fieldtype Change			NO		
17		Duplicate Count			NO		
18	Profile	Known Field Types		YES			
19		Drift Enabled(*)		YES			
20		Custom Rules Configured		YES			
21		Known Data Types		YES			
22		Sense Enabled Scoring			NO		
23		Valid Records Email Notification			NO		
24		Enable Data Source Level Rule Configuration			NO		DEPENDS ON THE CUSTOMER REQUIREMENT
25		Count Statistical Rule in Invalid Records			NO		
27	Rules		Null Count		NO		
		Count					DEPENDS ON THE CUSTOMER REQUIREMENT
28			Blank		NO		
29			Digits		NO		
30			Distinct Count		NO		
31			Duplicate		NO		

32			Unique Count		NO	
33			Characters		NO	
34			White Space		NO	
35			Special Characters		NO	
36			Zero Values		NO	
38		Statistical	Descriptive	YES		
39			Quantile	YES		
41			Length	YES		
42		Basic	Pattern	YES		
43			Range (Extreme Values)		NO	
44			Values		NO	
46		Outlier	Normal Distribution		NO	
47			K-means		NO	
48			Box Plot		NO	
49			Scatter Plot		NO	
51		Sampling	Histogram		NO	
53		Text	Soundex		NO	
54			Case sensitivity		NO	
55		Custom		YES		
56		Field type Sense		YES		

57		Data type sense		YES		
58		Invalid Records Save			NO	DEPENDS ON THE CUSTOMER REQUIREMENT
59		Lookup		YES		
60	Curation	Basic		YES		
61		Reference		YES		
62		Advanced		Yes		
63	Semantics Learning(*)	Enable Semantic learning			NO	
64	Drift Rules(*)	Correlation			NO	DEPENDS ON THE CUSTOMER REQUIREMENT
65		Duplicate Count			NO	
66		Primary Duplicate Count		YES	NO	
67		Row Count			NO	
68		Multicolumn Duplicate Count			NO	
69		Add Column			NO	

70		Remove Column			NO	
71		Null Count		YES		
72		Unique Count		YES		
73		Blank		YES		
74		Digits		YES		
75		Special Characters		YES		
76		Distinct Count		YES		
77		Infinite		YES		
78		Duplicate		YES		
79		Missing		YES		
80		Characters		YES		
81		White Space		YES		
82		Length		YES		
83		Pattern		YES		
84		Valid Values		YES		
85		Time series Analysis		YES		
86		Behaviour Analysis		YES		
87		DataType Change			NO	
88		FieldType Change			NO	
89		Sum			NO	
90		Q1			NO	

91		Q2			NO	
92		Q3			NO	
93		Mean			NO	
94		Kurtosis			NO	
95		Skewness			NO	
96		Variance			NO	
97		Standard Deviation			NO	
98		Coefficient of Variation			NO	
99		Median Absolute Deviation			NO	
100		IQR			NO	
101		Range			NO	
102		Max			NO	
103		min			NO	
104		Median			NO	
105		Zero Values			NO	
106		Custom Rules			NO	
107		Incremental Row Count			NO	
108	Notification	Alert Limit			NO	If Tured ON, then provide the number of alrets that are to received for a given period of time
109		Push Notification		YES		

110			Drift(*)	YES		
111			Monitoring	YES		
112			Action Item	YES		
85		Time series Analysis		YES		
86		Behaviour Analysis		YES		
87		DataType Change			NO	
88		FieldType Change			NO	
89		Sum			NO	
90		Q1			NO	
91		Q2			NO	
92		Q3			NO	
93		Mean			NO	
94		Kurtosis			NO	
95		Skewness			NO	
96		Variance			NO	
97		Standard Deviation			NO	
98		Coefficient of Variation			NO	
99		Median Absolute Deviation			NO	
100		IQR			NO	
101		Range			NO	
102		Max			NO	

103		min			NO	
104		Median			NO	
105		Zero Values			NO	
106		Custom Rules			NO	
107		Incremental Row Count			NO	
108	Notification	Alert Limit			NO	If Tured ON, then provide the number of alrets that are to received for a given period of time
109		Push Notification		YES		
110			Drift(*)	YES		
111			Monitoring	YES		
112			Action Item	YES		

- Make sure the roles are set up as per the organization's requirements
- For notifications, it is recommended to limit the number of alerts sent to the user
- Under notifications, please configure SMTP server to receive user invite mail while adding users in the users tab and email notifications.
- For storage, ensure the inbound and the outbound storage details are provided properly

Integration And Sensing

This section details the check and best practices in pull and sense jobs

Do's

- While connecting to a dataset, ensure that all the details are correctly provided
- Connect with "FULL LOAD" or "INCREMENTAL LOAD" based on the data in the DB
- Wait till all the PULL and sense job to be completed to add profile rules
- For field type sensing make sure the respective semantics is added
- If the field type is not mapped to the semantic field, check the contains and equals in the semantic field and change it based on the attribute name

Don'ts

- Don't pull the whole data from DB instead pull a snapshot of the data to start profiling

- Don't re-run the PULL and SENSE job when the job is already running
- Don't delete the dataset when a job is running
- Do not enable all the rules in the properties page, enable only the required rules

Profile And Monitoring

This section details the check and best practices in profile and drift jobs

Do's

- While adding custom rules for the profile, ensure the rule has valid and invalid data to cross verify
- While adding custom queries for profile rules, check the following ○ PULL mode supports spark SQL queries irrespective to the connected DB
- Once a profile rule added, re- run only profile from the execution log to check the custom rule chart in profile page
- Turn OFF/ON the required rules for an attribute to calculate the DQ score from the rules section at the bottom of the profile page

Rules

Rule	Description	Accepted Count	Unaccepted Count	Total Count	Accepted %	Actions
NULL COUNT		172	836	1008	82.9%	
BLANK		0	1008	1008	100%	
DIGITS		0	1008	1008	NA	
DISTINCT COUNT		825	183	1008	81.8%	
DUPLICATE		183	825	1008	81.8%	

- Enable only the required drift rules for the drift analysis from the settings page

Drift Rules

Correlation	Duplicate Count	Primary Duplicate Count	Row Count
Multicolumn Duplicate Count	Add Column	Remove Column	Null Count
Unique Count	Blank	Digits	Special Characters
Distinct Count	Infinite	Duplicate	Missing
Characters	White Space	Length	Pattern
Valid Values	Timeseries Analysis	Behaviour Analysis	DataType Change
FieldType Change	Sum	Q1	Q2
Q3	Mean	Kurtosis	Skewness
Variance	Standard Deviation	Coefficient of Variation	Median Absolute Deviation
IQR	Range	Max	min
Median	Zero Values	Custom Rules	Incremental Row Count

- For behavioral analysis use any of the following combinations ○ Numeric Date/Time - NT ○ Character Numeric Date/Time - CNT
    - ○ Character Character Numeric Date/Time – CCNT
- Use the alert limit feature on the settings page to restrict number of alerts sent to the users if you have a very large number of drift alerts
- Set up the drift days as 3 from the settings page to calculate the drift alerts based on the values of the past three days.

Properties						^
Preview Count	Dataset Percentage	Max Concurrent Jobs	Drift Days Count		Skip Drift Date	Push Permission
100	100	5	3			

Don'ts

- Do not use the "WITH" statement while using custom queries for profiling
- Do not use native database SQL functions while using custom queries
- Do not use more than 2 joins while using custom queries for profiling
- Do not provide a high number of concurrent jobs in the properties page

Curate and Push

This section details the check and best practices in curate and push jobs

Do's

- For reference curation, make sure the reference file is added in the correct format
- Ensure to accept or reject the curated values in the curation page for the advanced curation only which the values will be changed/curated
- Click on the preview icon next to curate on the execution log to view the curated records
- Ensure to provide the correct details for the outbound connector where the curated records are to be pushed

Don'ts

- Don't use incorrect reference file for reference curation
- Don't delete the target table in the DB where the curated records are pushed

Integration via API

This section details the check and best practices for the API integration

Do's

- Always ensure that the "Client_id" and "Client_Secret" key is valid
- Once a data source is created in DQLabs from the DI tool, please ensure the profile the dataset in DQLabs to get the DQ Score
- Check if all the request parameters in the API has been sent to get a response
- While update the data source configuration ensure that correct details is passed based on the connector

Don'ts

- Don't use incorrect "Client_id" and "Client_Secret" key
- Don't use different parameters for different database connectors while connecting

Troubleshooting

This section provides the basic checks to be done if any features are not working as expected

PULL QUERY MODE IS FAILING

- Check if the query is being executed successfully on the source DB
- Check if there are any duplicate column names in the select statement
- Remove all the comments in the query
- Check if the table is authorized in the source table for the user used in DQLabs

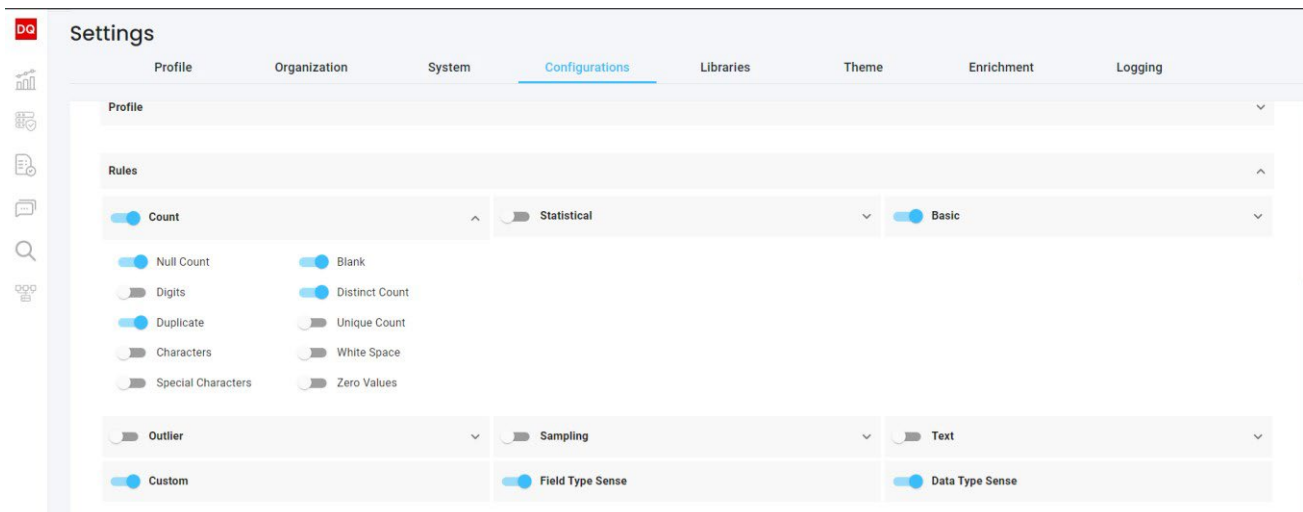## ATTRIBUTE PAGE NOT LOADING

- Remove the ";" at the end of the query if the datasource is connected by query method
- If the attribute page takes more time to load on a PULL query method, check the execution time of the query in the source DB to see if the actual query itself takes more time to execute and update the query accordingly

FIELD TYPES ARE NOT MAPPED PROPERLY

- Check if the "Field Type Sense" is enabled in the profile settings
- Check if the semantics field has the attribute name in "Equals" and "Contains" section of the semantic field

SOME CHARTS FOR DEFAULT PROFILE RULE IS NOT SHOWING PROFILE PAGE

- Check if the profile rule is enabled in the profile rule section under settings

- Ensure that the sense/profile has run successfully

CUSTOM RULE CHART IS NOT SHOWING IN PROFILE PAGE

- Make sure that the custom rule is valid and returns results
- Make sure the syntax of the custom rule query is correct
- Make the custom rule is enabled in profile rules

DRIFT ALERT CHART IS NOT SHOWING UP

- Check the drift day count setup in the settings page

- Check if learning is run at least two times
- For behavioral analysis, make sure there is a data time column

CURATION IS NOT PICKING AFTER PROFILE

- Check if curation is turned on in the settings page

PUSH IS NOT PICKING AFTER CURATION

- The push normally takes 1-2 mins to pick up after curation
- Make sure the outbound connect details are provided
- Make sure the user has write permission for the DB