

Truata Calibrate



System Requirements Guide Calibrate v2.2

The screenshot displays the Truata Calibrate dashboard interface. At the top, there is a navigation bar with the Truata logo and menu items: Dashboard, Jobs, Data profiles, Admin, and Help. Below the navigation bar, the dashboard title "Dashboard" is followed by a "Scope" filter set to "Data profile" and a date range of "11/12/2021 - 11/03/2022".

The dashboard features three key performance indicators (KPIs) in a row:

- 53 data assets published:** 1,048 columns, ~310 million rows.
- 99 avg risk detected:** 40 risky columns, 78 columns with fix.
- 90 avg risk for mitigated assets:** 10 risky columns, 28 columns with fix.

Below the KPIs is an "Insights" section with a line chart titled "Average risk identified in all data assets of all data profiles over a 11/12/2021 - 11/03/2022 period". The chart shows a fluctuating risk score over time, with a red dashed line indicating a target risk level at 90. The current risk score is shown as 90.00/90.

The "Data assets" section includes a "PUBLISH DATA ASSET" button and a table of data assets. The table has columns for "DATA ASSET NAME", "NUMBER OF COLUMNS", "ROWS", "PERSONAL DATA TYPE COUNT", "RISK SCORE", "CREATED ON", "TRANSFORMED", "LAST ACTION DATE", and "LAST ACTION".

DATA ASSET NAME	NUMBER OF COLUMNS	ROWS	PERSONAL DATA TYPE COUNT	RISK SCORE	CREATED ON	TRANSFORMED	LAST ACTION DATE	LAST ACTION
cy_asset_parquet 1.Default Profile1	13	1000	9		13/01/2022, 10:04		13/01/2022, 10:12	Identify
payroll1 1.Public	16	591054	8		17/01/2022, 18:22		17/01/2022, 19:29	Identify
Cryptic_Payroll_Data_New 1.Public	16	591054	7		12/01/2022, 11:43		12/01/2022, 12:09	Identify

Contents

Change History	2
1 Introduction	3
1.1 Deployment.....	3
2 Prerequisites	3
2.1 Calibrate Application	4
2.2 The Calibrate engine.....	4
3 Minimum Databricks instance requirements	5
3.1 High-level guidance on cluster size for Databricks	5
3.2 Recommendation on Databricks cluster specifications.....	5

Change History

Version	Release Date	Author	Change reason
0.1	19-March-2021	Truata	Draft version
0.2	24-March-2021	Truata	Reviewed
1.0	01-May-2021	Truata	Release version
2.0	12-Aug-2021	Truata	Updated for Calibrate Release 2.0
2.1	01-Nov-2021	Truata	Updated for Calibrate Release 2.1
2.2	03-May-2022	Truata	Updated for Calibrate 2.2

System Requirements

1 Introduction

This System Requirements guide is intended for IT users responsible for provisioning the runtime environment for the Calibrate application - typically the IT team. The document will help you understand the prerequisites required before the application can be installed.

This guide also outlines recommendations on execution environment sizing that can be used as guidance.

Calibrate V2.0 is a product built to run on the Azure environment. Calibrate can be deployed on Azure public cloud and Azure on-premises.

Calibrate product has two main deployment components:

- Calibrate Application
- Calibrate Engine

1.1 Deployment

A high-level view of how Calibrate fits into your enterprise is depicted below.

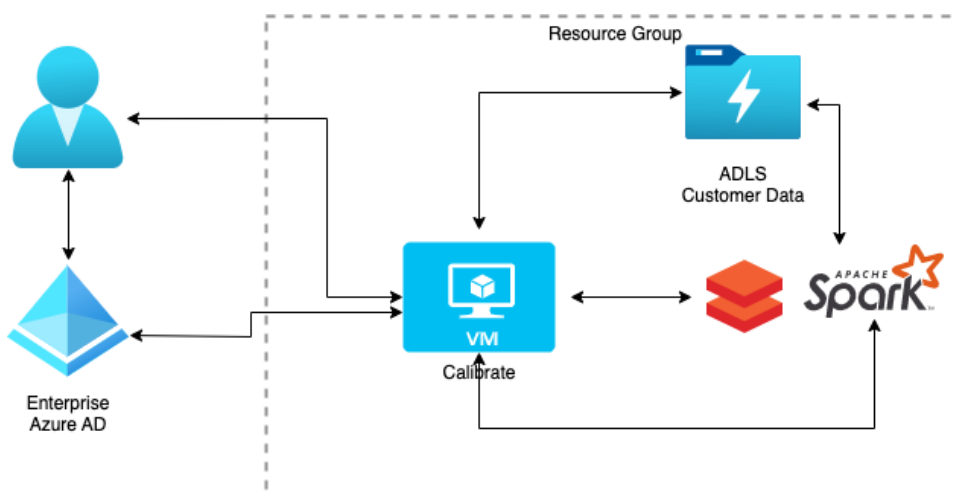


Figure 1-1: Logon screen

2 Prerequisites

In order to deploy and run Calibrate, the following are required to be provisioned on your Azure tenancy.

- Data Lake storage (ADLS Gen 2 or Blob Storage)
- Databricks
- VM (for calibrate application)

Note: All of the above should preferably be in one resource group. If the VM is placed outside of above resource group, bi-directional communication between the resource group and VM needs to be accommodated.

2.1 Calibrate Application

Calibrate is a containerized application that will run on Linux OS.

Virtual machine specifications:

- OS: Linux *
- Minimum RAM of 8GB
- Minimum disk space of 32GB

The VM should be pre-installed with:

- Docker (version 19 or newer)
- Docker-compose (version 1.27.4 or newer)
- Jq, bash, curl
- System user that can run Docker (user added to Docker group)

And VM should be able to access below components in Azure:

- Data lake (ADLS Gen 2 or Blob Storage)
- Databricks

Authentication to these components is through service principals/access tokens as mentioned in the Installation guide.

Note: *: Application is tested with CentOS and RHEL. However, the application is built to run on other modern Linux flavors as well.

2.2 The Calibrate engine

The Calibrate engine is a Spark based application and requires the following to execute:

- Databricks resource instance provisioned.

With access to:

- Data lake (ADLS Gen 2 or Blob Storage)
- Azure KeyVault

3 Minimum Databricks instance requirements

The Calibrate engine can run on a Job-Cluster type of cluster on Databricks.

Job-Cluster

Job clusters are created when the job run starts and terminate upon completion of job run. Databricks recommends job-clusters for production and repeated workloads and has the benefits of isolation for cluster resources, debugging etc. It will take few minutes to start the cluster depending on number of nodes and node type. This start time can be reduced using Databricks pools. See - <https://docs.microsoft.com/en-us/azure/databricks/clusters/instance-pools/> for more details.

3.1 High-level guidance on cluster size for Databricks

The following details serve as a guideline for the sizing of the Databricks cluster for input file size and estimated execution time. Appropriate cluster size should be provided to Databricks configuration in Configuration/Installation guide.

The size of the job cluster depends on size and complexity of data, main factors are:

- Length – row count
- Breadth – column count
- Number of Categorical Columns (For Fingerprint)

3.2 Recommendation on Databricks cluster specifications

The following are different cluster sizes categorized into small, medium and large clusters.

Small:

Worker Node: Standard_DS4_v2, 28.0 GB Memory, 8 Cores, 1.5 DBU,

Number of workers: 5

Driver node: Standard_DS5_v2, 56.0 GB Memory, 16 Cores, 3.0 DBU

cost/per hour: $(5 * € 1.141) + € 2.28 = € 7.985$

Medium:

Worker Node: Standard_DS4_v2, 28.0 GB Memory, 8 Cores, 1.5 DBU,

Number of workers: 10

Driver node: Standard_DS5_v2, 56.0 GB Memory, 16 Cores, 3.0 DBU

cost/per hour: $(10 * € 1.141) + € 2.28 = € 13.96$

Large:

Worker Node: Standard_DS5_v2, 56.0 GB Memory, 16 Cores, 3.0 DBU,

System Requirements

Number of workers: 40

Driver node: Standard_DS5_v2, 56.0 GB Memory, 16 Cores, 3.0 DBU

cost/per hour: $(40 + 1) * € 2.28 = € 93.48$

Note: The above prices are based on Microsoft's pay-as-you-go plans. The pricing above is for guidance only and may have changed since the time of writing. Up-to-date Azure pricing details can be obtained from <https://azure.microsoft.com/en-us/pricing/details/databricks/>.

The following table is a reference for cluster size based on our benchmarking:

Rows	Columns	Cluster size	Execution time*
100M	100	Large	9 hrs
10M	50	Medium	4 hrs
1M	25	Small	2 hrs

Figure 3-1: Identify benchmark.

Rows	Columns	Categorical cols	Cluster size	Execution time*
100M	100	70	Large	7 hrs
10M	50	25	Medium	3.5 hrs
1M	50	25	Small	2.5 hrs

Figure 3-2: Fingerprint benchmark.

Rows	Columns	Cluster size	Execution time*
1000M	100	Large	35 min
100M	100	Medium	20 min
10M	100	Small	5 min

Figure 2-3: Transform benchmark.

System Requirements

Note: Transform actions are applied on 24 columns (4 columns for each transformation - Suppression, Masking, Rounding, Tokenization, FPE and Noise Addition)

Note: These are approximated times and are purely recommendations based on performance tests on synthetic data generated by Truata. This may vary for other datasets.

**The above figures are for Apache Parquet formatted data assets.*

Cluster sizes can be selected separately for Identify, Fingerprint and Transform.

System Requirements

Truata Limited reserves the right to revise or amend this documentation in whole or in part in its sole discretion from time to time, without notice. The latest version of this documentation can be found at <https://peap.truata.com>. Truata Limited does not assume any liability for defects and damage which may result through use of the information contained herein. This content does not form part of any contract or of business relations between the parties, nor does it change the contract or business relations as entered into between the parties. All obligations of Truata Limited are stated in the relevant contractual agreements as entered into by and between the parties. No part of this document may be reproduced or transmitted in any form or by any means, electronically or mechanical, for any purpose, without the express written permission of Truata.