

Profilers in VM-Based Environments

Date published: 2019-11-14

Date modified: 2025-10-17



Legal Notice

© Cloudera Inc. 2025. All rights reserved.

The documentation is and contains Cloudera proprietary information protected by copyright and other intellectual property rights. No license under copyright or any other intellectual property right is granted herein.

Unless otherwise noted, scripts and sample code are licensed under the Apache License, Version 2.0.

Copyright information for Cloudera software may be found within the documentation accompanying each component in a particular release.

Cloudera software includes software from various open source or other third party projects, and may be released under the Apache Software License 2.0 (“ASLv2”), the Affero General Public License version 3 (AGPLv3), or other license terms. Other software included may be released under the terms of alternative open source licenses. Please review the license and notice files accompanying the software for additional licensing information.

Please visit the Cloudera software product page for more information on Cloudera software. For more information on Cloudera support services, please visit either the Support or Sales page. Feel free to contact us directly to discuss your specific needs.

Cloudera reserves the right to change any products at any time, and without notice. Cloudera assumes no responsibility nor liability arising from the use of products, except as expressly agreed to in writing by Cloudera.

Cloudera, Cloudera Altus, HUE, Impala, Cloudera Impala, and other Cloudera marks are registered or unregistered trademarks in the United States and other countries. All other trademarks are the property of their respective owners.

Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES NOT WARRANT THAT CLOUDERA PRODUCTS NOR SOFTWARE WILL OPERATE UNINTERRUPTED NOR THAT IT WILL BE FREE FROM DEFECTS NOR ERRORS, THAT IT WILL PROTECT YOUR DATA FROM LOSS, CORRUPTION NOR UNAVAILABILITY, NOR THAT IT WILL MEET ALL OF CUSTOMER’S BUSINESS REQUIREMENTS. WITHOUT LIMITING THE FOREGOING, AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, CLOUDERA EXPRESSLY DISCLAIMS ANY AND ALL IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, QUALITY, NON-INFRINGEMENT, TITLE, AND FITNESS FOR A PARTICULAR PURPOSE AND ANY REPRESENTATION, WARRANTY, OR COVENANT BASED ON COURSE OF DEALING OR USAGE IN TRADE.

Contents

Cloudera Data Catalog Profilers.....	4
The Cluster Sensitivity Profiler.....	4
The Hive Column Profiler.....	6
The Ranger Audit Profiler.....	8
 Profiler data testing.....	 8

Cloudera Data Catalog Profilers

Profilers create metadata annotations that summarize the content and shape characteristics of the data assets (such as distribution of values in a box plot or histogram).



Note:

The VM-based architecture (using the Cluster) is deprecated from the 3.0.0 release but remains available until 7.2.18 is supported (Sept 2025). Therefore, based profilers will also not be available in versions after 7.2.18. Only Compute Cluster enabled environment will be able to run profilers after version 7.2.18.

For more information, see [Cloudera Support lifecycle policy](#).

Cloudera Data Catalog supports different environments:

- When using the VM-based environment, you can create a Cloudera Data Hub cluster for a profiler engine to run data profiling operations as a pipeline on data located in one of your data lakes. You can install the profiler agent in a data lake and set up a specific schedule to generate various types of data profiles.
- When using the Compute Cluster enabled environment, after launching a profiler cluster, an internal service provisions new Kubernetes pods, scheduling and running profiler jobs on-demand.



Note:

- In a VM-based Cloudera Data Hub cluster environment, you can launch the profiler engine (on the Cloudera Data Hub cluster) for a data lake after selecting the data lake and clicking Launch Profilers. Note that, if the selected data lake already has the profiler engine attached, then the Launch Profilers option is not displayed.
- Also, the Cloudera Data Hub for the profiler can be set up only by accessing the Cloudera Data Catalog UI and is not supported through the Cloudera Management Console.

Profiler Name	Description
Cluster Sensitivity Profiler	The profiler automatically classifies your data with preconfigured tags, such as, PII, PCI, HIPAA and others.
Ranger Audit Profiler	A Ranger audit log summarizer.
Hive Column Profiler	Provides summary statistics like Maximum, Minimum, Mean, Unique, and Null values at the Hive column level.

Limitations

- In VM-based environments (with Cloudera Data Hub workflows), profilers do not support Iceberg tables, however, they are discoverable. In Compute Cluster enabled environments, Iceberg tables can be profiled.
- Supported file formats:
 - VM-based environments:
 - CSV
 - Avro
- Compute Cluster based profilers may hang if the underlying AWS cloud provider environment cannot provide the necessary memory for the executor instances. In this case, reconfigure your executors with 4-5 GB memory in Profiler Details Configuration .

The Cluster Sensitivity Profiler

The Cluster Sensitivity Profiler automatically performs context and content inspection to detect various types of sensitive data and suggest suitable classifications or tags based on the type of sensitive content detected or discovered.

Auto-detected data types

Type of data

- Bank account
- Credit card
- Driver number (UK)
- Email
- IBAN number
 - Austria (AUT)
 - Belgium (BEL)
 - Bulgaria (BGR)
 - Switzerland (CHE)
 - Cyprus (CYP)
 - Czech Republic (CZE)
 - Germany (DEU)
 - Denmark (DNK)
 - Spain (ESP)
 - Estonia (EST)
 - Finland (FIN)
 - France (FRA)
 - United Kingdom (GBR)
 - Greece (GRC)
 - Croatia (HRV)
 - Hungary (HUN)
 - Ireland (IRL)
 - Iceland (ISL)
 - Italy (ITA)
 - Liechtenstein (LIE)
 - Lithuania (LTU)
 - Latvia (LVA)
 - Luxembourg (LUX)
 - Malta (MLT)
 - Netherlands (NLD)
 - Norway (NOR)
 - Poland (POL)
 - Portugal (PRT)
 - Romania (ROU)
 - Slovakia (SVK)
 - Slovenia (SVN)
 - Sweden (SWE)
- IP address
- NPI
- Name

- National ID number
 - Bulgaria (BGR)
 - Switzerland (CHE)
 - Czech Republic (CZE)
 - Denmark (DNK)
 - Spain (ESP)
 - Estonia (EST)
 - Finland (FIN)
 - Greece (GRC)
 - Ireland (IRL)
 - Iceland (ISL)
 - Italy (ITA)
 - Lithuania (LTU)
 - Latvia (LVA)
 - Norway (NOR)
 - Poland (POL)
 - Portugal (PRT)
 - Romania (ROU)
 - Slovakia (SVK)
 - Sweden (SWE)
- National insurance number (UK)
- Passport number
 - Austria (AUT)
 - Belgium (BEL)
 - Switzerland (CHE)
 - Germany (DEU)
 - Spain (ESP)
 - Finland (FIN)
 - France (FRA)
 - Greece (GRC)
 - Ireland (IRL)
 - Italy (ITA)
 - Poland (POL)
 - United Kingdom (UK)
- Bank Routing Number
- US Social Security Number
- Society for Worldwide Interbank Financial Telecommunication (SWIFT)
- Telephone

The Hive Column Profiler

You can view the shape or distribution characteristics of the columnar data within a Hive table based on the Hive Column Profiler.

There are different charts available to help visualize the shape and distribution of the data within the column as well as summary statistics for a column with numerical data. These include the following:

- Unique Values
- Null Values
- Max

- Min
- Mean

OverviewSchemaMetadata AuditsPolicyAccess Audits

Q Search column

Edit

Chart Type	Name	Type	Unique Values *	Null Values	Max	Min	Mean	Comment	Classifications	Terms
▼	flight_id_raw	string	15	0						
▼	passenger_name_raw	string	22	0						
▼	seat_assignment_raw	string	21	0						
▼	aircraft_model_raw	string	7	0						
▼	origin_airport_raw	string	7	0						
▼	destination_airport_raw	string	12	0						
▼	age_raw	int	20	1	62	25	38.55	Passenger age. PII data subject to ...	GDPR... + 1	Personally Identifiabl... + 1
▼	ticket_price_usd_raw	decimal(10,2)	17	0	1,500	85.5	667.35	Total ticket fare in USD. Confident ...	Confide... + 2	Passenger Fare@Airline 0...

* Approximate values as being computed using HLL algorithm.

The profiler computes column univariate statistics that are displayed using an appropriate chart in the **Schema** tab.

Pie charts are presented for categorical data with limited number of categories or classes. Examples include data such as eye colors that only have a fixed list of values (categories or labels).

Asset Details

Q Search column

Edit

Chart Type	Name	Type	Unique Values *	Null Values	Max	Min	Mean	Comment	Classifications	Terms
▼	flight_id_raw	string	15	0						
▼	passenger_name_raw	string	24	1						
▼	seat_assignment_raw	string	25	0						
▼	aircraft_model_raw	string	8	0						
▲	origin_airport_raw	string	8	0						

Profiled : 100.0% rows, 52 minutes ago

BUD

LHR

FRA

SIN

JFK

CDG

OTP

▼	destination_airport_raw	string	12	0						
▼	age_raw	int	21	1	62	25	39.29	Passenger age. PII data subject to ...	GDPR... + 1	Passenger Name Rec... + 1
▼	ticket_price_usd_raw	decimal(10,2)	18	0	1,500	85.5	671.99	Total ticket fare in USD. Confident ...	Raw + 2	Passenger Fare@Airline 0...

* Approximate values as being computed using HLL algorithm.

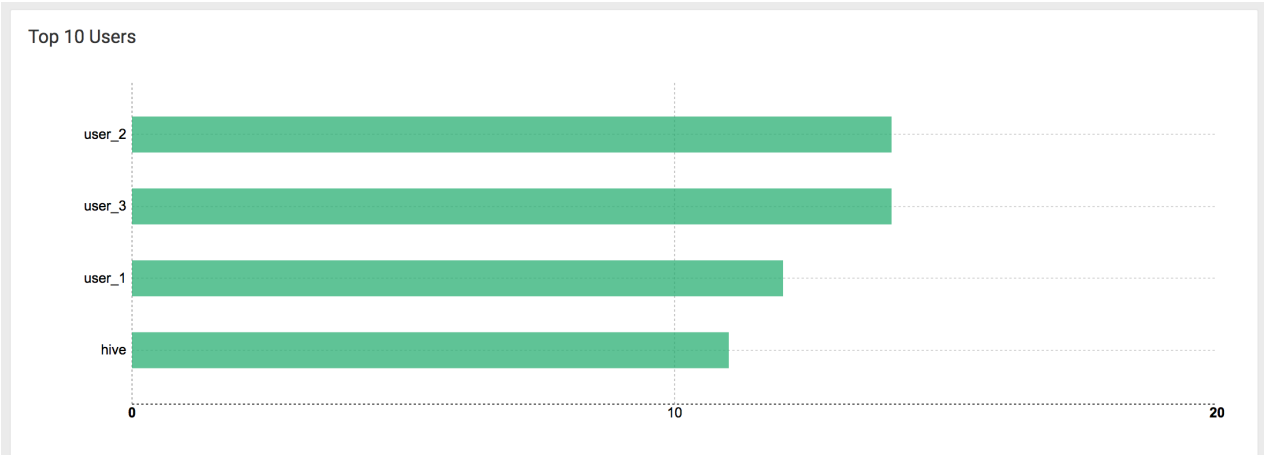
When the data within columns is numeric, a histogram of the distribution of values organized into 10 groups (decile frequency histogram) and a box plot with a five-number summary (mean, median, quartiles, maximum, and minimum values) are shown for the column.



The Ranger Audit Profiler


You can view who has accessed which data from a forensic audit or compliance perspective, visualize access patterns, and identify anomalies in access patterns using the Ranger Audit Profiler.

The Ranger Audit Profiler uses the Apache Ranger logs to show the most recent raw event data as well as summarized views of event by type of access and access outcomes (allowed/denied). These views are obtained by profiling the records in the data lake.



Profiler data testing

You must note the important information about profiler services.

 **Note:** The Cloudera Data Catalog profilers are not tested at par with the Hive scale.

Test data for VM-based environments

The following dataset has been validated and works as expected:

- DataHub Master: m5.4xlarge
- Hive tables: 3000 Hive assets
- Total Number of assets (including Hive columns, tables, databases): 1,000,000
- Total Data Size = 1 GB
- Partitions on Hive tables: Around 5000 partitions spread across five tables