CDP Public Cloud / Data Catalog

# Profilers in VM-Based Environments

**Date published: 2019-11-14**
**Date modified: 2026-03-31**

CLOUDERA

**https://docs.cloudera.com/**

# Legal Notice

© Cloudera Inc. 2026. 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.....5
The Hive Column Profiler.....30
The Ranger Audit Profiler.....32

**Profiler data testing.....33**

# 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.

**Warning:** Delete launched profilers before upgrading to  on  1.5.5 SP2.

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 might 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 .

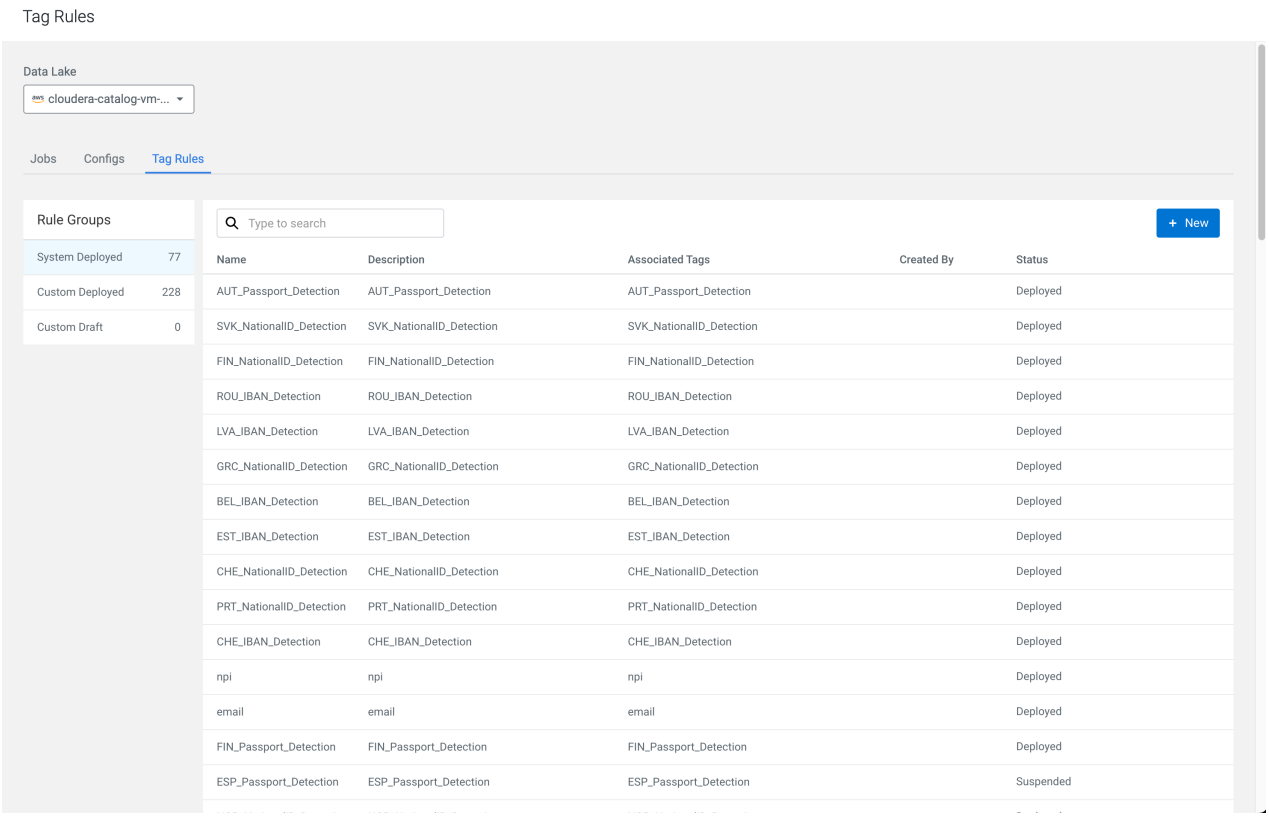
  **Note:** Is your infrastructure is deployed completely in AWS, profilers might hang if the underlying AWS cloud provider environment cannot provide the necessary memory for the executor instances. In this case, reconfigure your executors with 4-5 GB memory in  Profiler Details Configuration .

# 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.

## Tag Rules

The Cluster Sensitivity Profiler have system-defined and user-defined tags that identify data types based on regular expressions .




## Auto-detected data types and system-defined regular expressions

The system-defined tag rules identify the following data types:

## Bank account, credit card, email

- Bank account
- Credit card
- Driver number (UK)
- Email
- Name: **bankaccount**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:
    - ```
      regex("\\b(AD)\\d{2}\\d{8}[a-zA-Z0-9]{12}|(AE|IL)\\d{2}\\d{19}|(AL|CY)\\d{2}\\d{8}[a-zA-Z0-9]{16}|(PL)\\d{2}\\d{24}|(AT|BA|EE|LT)\\d{2}\\d{16}|(BE)\\d{2}\\d{12}|(BG)\\d{2}[A-Z]{4}\\d{6}[a-zA-Z0-9]{8}|(BH)
      ```
      ```

```
\\d{2}[A-Z]{4}[a-zA-Z0-9]{14}|(CH)\\d{2}\\d{5}[a-zA-Z0-9]{12}|(CZ|ES
|SE|SK|TN)\\d{2}\\d{20}|(DE|ME|RS)\\d{2}\\d{18}|(DK|FI|FO|GL)\\d{2}\
\d{14}|(DO)\\d{2}[A-Z]{4}\\d{20}|(FR|MC)\\d{2}\\d{10}[a-zA-Z0-9]{11}
\\d\\d|(GB)\\d{2}[A-Z]{4}\\d{14}|(IE)\\d{2}[a-zA-Z0-9]{4}\\d{14}|(GE
)\\d{2}[a-zA-Z0-9]{2}\\d{16}|(GI)\\d{2}[A-Z]{4}[a-zA-Z0-9]{15}|(GR)\
\d{2}\\d{7}[a-zA-Z0-9]{16}|(HR)\\d{2}\\d{17}|(HU)\\d{2}\\d{24}|(IS)\
\d{2}\\d{22}|(IT|SM)\\d{2}[A-Z]\\d{10}[a-zA-Z0-9]{12}|(KW)\\d{2}[A-Z
]{4}[a-zA-Z0-9]{22}|(KZ)\\d{2}\\d{3}[a-zA-Z0-9]{3}\\d{10}|(LU)\\d{2}\
\d{3}[a-zA-Z0-9]{13}|(LB)\\d{2}\\d{4}[a-zA-Z0-9]{20}|(LI)\\d{2}\\d{5}
[a-zA-Z0-9]{12}|(LV)\\d{2}[A-Z]{4}[a-zA-Z0-9]{13}|(MK)\\d{2}\\d{3}[a-
zA-Z0-9]{10}\\d\\d|(MR)\\d{2}\\d{23}|(MT)\\d{2}[A-Z]{4}\\d{5}[a-zA-Z0
-9]{18}|(MU)\\d{2}[A-Z]{4}\\d{19}[A-Z]{3}|(NL)\\d{2}[A-Z]{4}\\d{10}|
(NO)\\d{2}\\d{11}|(PT)\\d{2}\\d{21}|(RO)\\d{2}[A-Z]{4}[a-zA-Z0-9]{16
}|(SA)\\d{2}\\d{2}[a-zA-Z0-9]{18}|(SI)\\d{2}\\d{15}|(TR)\\d{2}\\d{5}
[a-zA-Z0-9]{17}\\b")
```

- Match Type: **Column name**
  - Confidence: 15
  - Regular expression:

```
is_in_ci("bankaccount","account","accountnumber")
```

- Name: **creditcard**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("\\b((5[1-5][0-9]{2}[0-9]{4}[0-9]{4}[0-9]{4})|((633(3|4)|67(67
|59)|49(36|11|05|03))[0-9]{4}[0-9]{4}[0-9]{4}(([0-9]{3}|[0-9]{2}))?)
|(4[0-9]{3}[0-9]{4}[0-9]{4}[0-9]{4}|4[0-9]{3}[0-9]{4}[0-9]{4}[0-9]{1
})|((6011)[0-9]{4}[0-9]{4}[0-9]{4})|((622)[0-9]{1}[0-9]{4}[0-9]{4}[0
-9]{4})|((65)[0-9]{2}[0-9]{4}[0-9]{4}[0-9]{4})|((64)[0-9]{2}[0-9]{4}
[0-9]{4}[0-9]{4})|(((3[47][0-9]{2})|2014|2149)[0-9]{4}[0-9]{4}[0-9]{
3})|(3(0[0-5]|[68][0-9])[0-9][0-9]{4}[0-9]{4}[0-9]{2})|((3528)[0-9]{
4}[0-9]{4}[0-9]{4})|((3529)[0-9]{4}[0-9]{4}[0-9]{4})|((35)[3-8][0-9]
[0-9]{4}[0-9]{4}[0-9]{4})|(63[789][0-9][0-9]{4}[0-9]{4}[0-9]{4})|(56
10[0-9]{4}[0-9]{4}[0-9]{4})|(56022[1-5][0-9]{2}[0-9]{4}[0-9]{4}))\\b
|\\b((5893|0604|6304|6759|676[1-3]|50(18|20|38))[0-9]{4}[0-9]{4}([0-
9]{0,4})?([0-9]{0,3})?)\\b|\\b(67(06|71|09)[0-9]{4}[0-9]{4}[0-9]{4}(
[0-9]{0,3})?)\\b|\\b((633110|564182)[0-9]{2}[0-9]{4}[0-9]{4}([0-9]{3
}|[0-9]{2})?)\\b","\\b((5[1-5][0-9]{2}[\\][0-9]{4}[\[ ][0-9]{4}[\\ ]
[0-9]{4})|((633(3|4)|67(67|59)|49(36|11|05|03))[\\][0-9]{4}[\\][0-9]
{4}[\[ ][0-9]{4}([\\]([0-9]{3}|[0-9]{2}))?)|(4[0-9]{3}[\\ ][0-9]{4}[
\[ ][0-9]{4}[\\][0-9]{4}|4[0-9]{3}[\\ ][0-9]{4}[\\ ][0-9]{4}[\[ ][0-
9]{1})|((6011)[\\][0-9]{4}[\\][0-9]{4}[\\ ][0-9]{4})|((622)[0-9]{1}[
\[ ][0-9]{4}[\\][0-9]{4}[\\][0-9]{4})|((65)[0-9]{2}[\\ ][0-9]{4}[\[
][0-9]{4}[\\][0-9]{4})|((64)[0-9]{2}[\\][0-9]{4}[\[ ][0-9]{4}[\\][0-
9]{4})|(((3[47][0-9]{2})|2014|2149)[\\ ][0-9]{4}[\\][0-9]{4}[\[ ][0-
9]{3})|((3[47][0-9]{2})[\\ ][0-9]{6}[\\][0-9]{5})|(3(0[0-5]|[68][0-9
])[0-9][\[ ][0-9]{4}[\\ ][0-9]{4}[\\][0-9]{2})|(3(0[0-5]|[68][0-9])[
0-9][\\ ][0-9]{6}[\[ ][0-9]{4})|((3528)[\\][0-9]{4}[\\][0-9]{4}[\\ ]
[0-9]{4})|((3529)[\\ ][0-9]{4}[\[ ][0-9]{4}[\\][0-9]{4})|((35)[3-8][
0-9][\\][0-9]{4}[\\ ][0-9]{4}[\[ ][0-9]{4})|(63[789][0-9][\\][0-9]{4
}[\\ ][0-9]{4}[\\][0-9]{4})|(5610[\\ ][0-9]{4}[\[ ][0-9]{4}[\\][0-9]
{4})|(5602[\[ ]2[1-5][0-9]{2}[\\][0-9]{4}[\\][0-9]{4}))|((5893|0604|
6304|6759|676[1-3]|50(18|20|38))[\[ ][0-9]{4}[\\][0-9]{4}([\\ ][0-9]
{0,4})?([\\ ][0-9]{0,3})?)|(67(06|71|09)[\[ ][0-9]{4}[\\][0-9]{4}[\\
][0-9]{4}([\\ ][0-9]{0,3})?)|((6331[\\ ]10|5641[\\]82)[0-9]{2}[\[ ][
0-9]{4}[\\][0-9]{4}([\[ ][0-9]{3}|[\\][0-9]{2})?)\\b","\\b((5[1-5][0
-9]{2}-[0-9]{4}-[0-9]{4}-[0-9]{4})|((633(3|4)|67(67|59)|49(36|11|05|
03))-[0-9]{4}-[0-9]{4}-[0-9]{4}(-([0-9]{3}|[0-9]{2}))?)|(4[0-9]{3}-[
```

```
0-9]{4}-[0-9]{4}-[0-9]{4}|4[0-9]{3}-[0-9]{4}-[0-9]{4}-[0-9]{1})|((60
11)-[0-9]{4}-[0-9]{4}-[0-9]{4})|((622)[0-9]{1}-[0-9]{4}-[0-9]{4}-[0-
9]{4})|((65)[0-9]{2}-[0-9]{4}-[0-9]{4}-[0-9]{4})|((64)[0-9]{2}-[0-9]
{4}-[0-9]{4}-[0-9]{4})|(((3[47][0-9]{2})|2014|2149)-[0-9]{4}-[0-9]{4
}-[0-9]{3})|((3[47][0-9]{2})-[0-9]{6}-[0-9]{5})|(3(0[0-5]|[68][0-9])
[0-9]-[0-9]{4}-[0-9]{4}-[0-9]{2})|(3(0[0-5]|[68][0-9])[0-9]-[0-9]{6}
-[0-9]{4})|((3528)-[0-9]{4}-[0-9]{4}-[0-9]{4})|((3529)-[0-9]{4}-[0-9
]{4}-[0-9]{4})|((35)[3-8][0-9]-[0-9]{4}-[0-9]{4}-[0-9]{4})|(63[789][
0-9]-[0-9]{4}-[0-9]{4}-[0-9]{4})|(5610-[0-9]{4}-[0-9]{4}-[0-9]{4})|(
5602-2[1-5][0-9]{2}-[0-9]{4}-[0-9]{4}))|((5893|0604|6304|6759|676[1-
3]|50(18|20|38))-[0-9]{4}-[0-9]{4}(-[0-9]{0,4})?(-[0-9]{0,3})?)|(67(
06|71|09)-[0-9]{4}-[0-9]{4}-[0-9]{4}(-[0-9]{0,3})?)|((6331-10|5641-8
2)[0-9]{2}-[0-9]{4}-[0-9]{4}(-[0-9]{3}|-[0-9]{2})?)\\b")
```

- Match Type: **Column name**
  - Confidence: 15
  - Regular expression:

```
is_in_ci("creditcard","ccnumber")
```

- Name: **ukdrivernumber**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("\\b([A-Z0-9]{5}[0-9][0156][0-9]([0][1-9]|[1-2][0-9]|[3][01])[
0-9][A-Z0-9]{3}[A-Z]{2})\\b")
```

- Name: **email**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("\\b(([a-zA-Z0-9_\\-\\.]+)@((\\[[0-9]{1,3}\\.[0-9]{1,3}\\.[0-9
]{1,3}\\.)|(([a-zA-Z0-9\\-]+\\.)+))([a-zA-Z]{2,4}|[0-9]{1,3})(\\]?))\
\b")
```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

```
is_in_ci("emailaddress","emailaddr","email")
```

  - Match Type: **Table name**
    - Confidence: 100
    - Regular expression:

```
(?i)\w*email\w*
```

### IBAN numbers

- 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)

- Name: **LVA_IBAN_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("(\\s|,|;|\\A)(LV\\d{2}([ ]?)[A-Z]{4}\\3[A-Za-z0-9]{4}\\3[A-Za-z0-9]{4}\\3[A-Za-z0-9]{4}\\3[A-Za-z0-9])(\\s|,|;|\\Z)")
```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

```
is_in_ci("Latvia IBAN","Latvia International Bank Account Number")
```

- Name: **ROU_IBAN_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("(\\s|,|;|\\A)(RO\\d{2}(?)[A-Z]{4}\\3[A-Za-z0-9]{4}\\3[A-Za-z0-9]{4}\\3[A-Za-z0-9]{4}\\3[A-Za-z0-9]{4})(\\s|,|;|\\Z)")
```

  - Match Type: **Column name**
    - Confidence: 15

    - Regular expression:

```
is_in_ci("Romania IBAN","Romania International Bank Account Number")
```

- Name: **FRA_IBAN_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("(\\s|,|;|\\A)(FR\\d{2}([ ]?)\\d{4}\\3\\d{4}\\3\\d{2}[0-9A-Z]{
2}\\3[0-9A-Z]{4}\\3[0-9A-Z]{4}\\3[0-9A-Z](9[0-7]|0[1-9]|[1-8]\\d))(\
\s|,|;|\\Z)")
```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

```
is_in_ci("France IBAN","France International Bank Account Number")
```

- Name: **DEU_IBAN_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("(\\s|,|;|\\A)(DE\\d{2}([ ]?)\\d{4}\\3\\d{4}\\3\\d{4}\\3\\d{4}\
\3\\d{2})(\\s|,|;|\\Z)")
```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

```
is_in_ci("Germany IBAN","Germany International Bank Account Number")
```

- Name: **CYP_IBAN_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("(\\s|,|;|\\A)(CY\\d{2}([ ]?)\\d{4}\\3\\d{4}\\3\\d{4}\\3\\d{4}\
\3\\d{4}\\3\\d{4})(\\s|,|;|\\Z)")
```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

```
is_in_ci("Cyprus IBAN","Cyprus International Bank Account Number")
```

- Name: **AUT_IBAN_Detection**
  - Match Type: **Column value**
    - Confidence: 85

        - Regular expression:

```
regex("(\\s|,|;|\\A)(AT\\d{2}([ ]?)\\d{4}\\3\\d{4}\\3\\d{4}\\3\\d{4})
(\\s|,|;|\\Z)")
```

    - Match Type: **Column name**
        - Confidence: 15
        - Regular expression:

```
is_in_ci("Austria IBAN","Austria International Bank Account Number")
```

- Name: **BEL_IBAN_Detection**
    - Match Type: **Column value**
        - Confidence: 85
        - Regular expression:

```
regex("(\\s|,|;|\\A)(BE\\d{2}([ ]?)\\d{4}\\3\\d{4}\\3\\d{4})(\\s|,|;|
\\Z)")
```

    - Match Type: **Column name**
        - Confidence: 15
        - Regular expression:

```
is_in_ci("Belgium IBAN","Belgium International Bank Account Number")
```

- Name: **EST_IBAN_Detection**
    - Match Type: **Column value**
        - Confidence: 85
        - Regular expression:

```
regex("(\\s|,|;|\\A)(EE\\d{2}([ ]?)\\d{4}\\3\\d{4}\\3\\d{4}\\3\\d{4})
(\\s|,|;|\\Z)")
```

    - Match Type: **Column name**
        - Confidence: 15
        - Regular expression:

```
is_in_ci("Estonia IBAN","Estonia International Bank Account Number")
```

- Name: **CHE_IBAN_Detection**
    - Match Type: **Column value**
        - Confidence: 85
        - Regular expression:

```
regex("(\\s|,|;|\\A)(CH\\d{2}([ ]?)\\d{4}\\3\\d{4}\\3\\d{4}\\3\\d{4}\
\3\\d)(\\s|,|;|\\Z)")
```

    - Match Type: **Column name**
        - Confidence: 15
        - Regular expression:

```
is_in_ci("Switzerland IBAN","Switzerland International Bank Account
Number")
```

- Name: **ESP_IBAN_Detection**

- Match Type: **Column value**
  - Confidence: 85
  - Regular expression:

```
regex("(\\s|,|;|\\A)(ES\\d{2}([ ]?)\\d{4}\\3\\d{4}\\3\\d{4}\\3\\d{4}\
\3\\d{4})(\\s|,|;|\\Z)")
```

- Match Type: **Column name**
  - Confidence: 15
  - Regular expression:

```
is_in_ci("Spain IBAN","Spain International Bank Account Number")
```

- Name: **ITA_IBAN_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("(\\s|,|;|\\A)(IT\\d{2}([ ]?)[A-Z]{1}\\3\\d{10}\\3[A-Z0-9]{12})
(\\s|,|;|\\Z)")
```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

```
is_in_ci("Italy IBAN","Italy International Bank Account Number")
```

- Name: **SVK_IBAN_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("(\\s|,|;|\\A)(SK\\d{2}([ ]?)\\d{4}\\3\\d{4}\\3\\d{4}\\3\\d{4}\
\3\\d{4})(\\s|,|;|\\Z)")
```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

```
is_in_ci("Slovakia IBAN","Slovakia International Bank Account Number
")
```

- Name: **SWE_IBAN_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("(\\s|,|;|\\A)(SE\\d{2}([ ]?)\\d{4}\\3\\d{4}\\3\\d{4}\\3\\d{4}\
\3\\d{4})(\\s|,|;|\\Z)")
```

  - Match Type: **Column name**
    - Confidence: 15

        - Regular expression:

          ```
          is_in_ci("Sweden IBAN","Sweden International Bank Account Number")
          ```

- Name: **NOR_IBAN_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)(NO\\d{2}([ ]?)\\d{4}\\3\\d{4}\\3\\d{3})(\\s|,|;|
      \\Z)")
      ```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

      ```
      is_in_ci("Norway IBAN","Norway International Bank Account Number")
      ```

- Name: **GRC_IBAN_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)(GR\\d{2}([ ]?)\\d{3}\\3\\d{4}\\3[A-Z0-9]{4}\\3[
      A-Z0-9]{4}\\3[A-Z0-9]{4}\\3[A-Z0-9]{3})(\\s|,|;|\\Z)")
      ```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

      ```
      is_in_ci("Greece
       IBAN","Greece International Bank Account Number")
      ```

- Name: **IBAN_Detection**
  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

      ```
      is_in_ci("IBAN","International Bank Account Number")
      ```

- Name: **GBR_IBAN_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)(GB\\d{2}([ ]?)[A-Z]{4}\\3\\d{4}\\3\\d{4}\\3\\d{
      4})(\\s|,|;|\\Z)")
      ```

  - Match Type: **Column name**
    - Confidence: 15

        - Regular expression:

```
is_in_ci("United Kingdom IBAN","United Kingdom International Bank Ac
count Number")
```

- Name: **IRL_IBAN_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("(\\s|,|;|\\A)(IE\\d{2}([ ]?)[A-Z]{4}\\3\\d{4}\\3\\d{4}\\3\\d{
4})(\\s|,|;|\\Z)")
```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

```
is_in_ci("Ireland IBAN","Ireland International Bank Account Number")
```

- Name: **FIN_IBAN_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("(\\s|,|;|\\A)(FI\\d{2}([ ]?)\\d{4}\\3\\d{4}\\3\\d{4}\\3\\d{2})
(\\s|,|;|\\Z)")
```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

```
is_in_ci("Finland IBAN","Finland International Bank Account Number")
```

- Name: **NLD_IBAN_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("(\\s|,|;|\\A)(NL\\d{2}([ ]?)[A-Z]{4}\\3\\d{4}\\3\\d{4}\\3\\d{
2})(\\s|,|;|\\Z)")
```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

```
is_in_ci("Netherlands IBAN","Netherlands International Bank Account
Number")
```

- Name: **BGR_IBAN_Detection**
  - Match Type: **Column value**
    - Confidence: 85

    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)(BG\\d{2}([ ]?)[A-Z]{4}\\3\\d{4}\\3[A-Z0-9]{4}\\
      3[A-Z0-9]{4}\\3[A-Z0-9])(\\s|,|;|\\Z)")
      ```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

      ```
      is_in_ci("Bulgaria IBAN","Bulgaria International Bank Account Number
      ")
      ```

- Name: **SVN_IBAN_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)(SI\\d{2}([ ]?)\\d{4}\\3\\d{4}\\3\\d{4}\\3\\d{3})
      (\\s|,|;|\\Z)")
      ```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

      ```
      is_in_ci("Slovenia IBAN","Slovenia International Bank Account Number
      ")
      ```

- Name: **LIE_IBAN_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)(LI\\d{2}([ ]?)\\d{4}\\3\\d{1}[A-Z0-9]{3}\\3[A-Z
      0-9]{4}\\3[A-Z0-9]{4}\\3[A-Z0-9]{1})(\\s|,|;|\\Z)")
      ```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

      ```
      is_in_ci("Liechtenstein IBAN","Liechtenstein International Bank Acco
      unt Number")
      ```

- Name: **POL_IBAN_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)(PL\\d{2}([ ]?)\\d{4}\\3\\d{4}\\3\\d{4}\\3\\d{4}\
      \3\\d{4}\\3\\d{4})(\\s|,|;|\\Z)")
      ```

  - Match Type: **Column name**
    - Confidence: 15

        - Regular expression:

```
is_in_ci("Poland IBAN","Poland International Bank Account Number")
```

- Name: **PRT_IBAN_Detection**
    - Match Type: **Column value**
        - Confidence: 85
        - Regular expression:

```
regex("(\\s|,|;|\\A)(PT\\d{2}([n]?)\\d{4}\\3\\d{4}\\3\\d{4}\\3\\d{4}\
\3\\d{4}\\3\\d{1})(\\s|,|;|\\Z)")
```

    - Match Type: **Column name**
        - Confidence: 15
        - Regular expression:

```
is_in_ci("Portugal IBAN","Portugal International Bank Account Number
")
```

- Name: **LTU_IBAN_Detection**
    - Match Type: **Column value**
        - Confidence: 85
        - Regular expression:

```
regex("(\\s|,|;|\\A)(LT\\d{2}([ ]?)\\d{4}\\3\\d{4}\\3\\d{4}\\3\\d{4})
(\\s|,|;|\\Z)")
```

    - Match Type: **Column name**
        - Confidence: 15
        - Regular expression:

```
is_in_ci("Lithuanian IBAN","Lithuania International Bank Account Num
ber")
```

- Name: **HRV_IBAN_Detection**
    - Match Type: **Column value**
        - Confidence: 85
        - Regular expression:

```
regex("(\\s|,|;|\\A)(HR\\d{2}([ ]?)\\d{4}\\3\\d{4}\\3\\d{4}\\3\\d{4}\
\3\\d{1})(\\s|,|;|\\Z)")
```

    - Match Type: **Column name**
        - Confidence: 15
        - Regular expression:

```
is_in_ci("Croatian IBAN","Croatia International Bank Account Number")
```

- Name: **LUX_IBAN_Detection**
    - Match Type: **Column value**
        - Confidence: 85

    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)(LU\\d{2}([ ]?)\\d{3}[A-Z0-9]{1}\\3[A-Z0-9]{4}\\
      3[A-Z0-9]{4}\\3[A-Z0-9]{4})(\\s|,|;|\\Z)")
      ```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

      ```
      is_in_ci("Luxembourg IBAN","Luxembourg International Bank Account Nu
      mber")
      ```

- Name: **DNK_IBAN_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)(DK\\d{2}([ ]?)\\d{4}\\3\\d{4}\\3\\d{4}\\3\\d{2})
      (\\s|,|;|\\Z)")
      ```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

      ```
      is_in_ci("Denmark IBAN","Denmark International Bank Account Number")
      ```

- Name: **HUN_IBAN_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)(HU\\d{2}([ ]?)\\d{4}\\3\\d{4}\\3\\d{4}\\3\\d{4}\
      \3\\d{4}\\3\\d{4})(\\s|,|;|\\Z)")
      ```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

      ```
      is_in_ci("Hungary IBAN","Hungary International Bank Account Number")
      ```

- Name: **CZE_IBAN_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)(CZ\\d{2}([ ]?)\\d{4}\\3\\d{4}\\3\\d{4}\\3\\d{4}\
      \3\\d{4})(\\s|,|;|\\Z)")
      ```

  - Match Type: **Column name**
    - Confidence: 15

    - Regular expression:

      ```
      is_in_ci("Czech Republic IBAN","Czechia IBAN","Czech Republic Intern
      ational Bank Account Number","Czechia International Bank Account Num
      ber")
      ```

- Name: **ISL_IBAN_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)(IS\\d{2}([ ]?)\\d{4}\\3\\d{4}\\3\\d{4}\\3\\d{4}\
      \3\\d{4}\\3\\d{2})(\\s|,|;|\\Z)")
      ```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

      ```
      is_in_ci("Iceland IBAN","Iceland International Bank Account Number")
      ```

- Name: **MLT_IBAN_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

      ```
      regex("\\A(MT\\d{2}([ ]?)[A-Z]{4}\\3\\d{4}\\3\\d{1}[A-Z0-9]{3}\\3[A-
      Z0-9]{4}\\3[A-Z0-9]{4}\\3[A-Z0-9]{4}\\3[A-Z0-9]{3})\\Z")
      ```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

      ```
      is_in_ci("Malta IBAN","Malta International Bank Account Number")
      ```

### IP addresses, NPI, names

- IP address
- NPI
- Name

Regular expressions:

- Name: **ipaddress**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)(((25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\\.(25[0
      -5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\\.(25[0-5]|2[0-4][0-9]|[01]?[0-9][
      0-9]?)\\.(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)))(\\s|,|;|\\Z)")
      ```

  - Match Type: **Column name**
    - Confidence: 15

        - Regular expression:

          ```
          is_in_ci("ipaddress","ipaddr","ip")
          ```

- Name: **npi**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)((([12]\\d{9}))|(((80840)[12]\\d{9})))(\\s|,|;|\
      \Z)")
      ```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

      ```
      is_in_ci("npi")
      ```

### National ID numbers

- 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)

Regular expressions:

- Name: **NOR_NationalID_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)((0[1-9]|[12][0-9]|3[01])([0][1-9]|[1][012])\\d{
      2}\\d{3}\\d{2})(\\s|,|;|\\Z)")
      ```

  - Match Type: **Column name**
    - Confidence: 15

        - Regular expression:

```
is_in_ci("Norway National ID","Norway National identification number
")
```

- Name: **FIN_NationalID_Detection**
    - Match Type: **Column value**
        - Confidence: 85
        - Regular expression:

```
regex("(\\s|,|;|\\A)((0[1-9]|[12][0-9]|3[01])([0][1-9]|[1][012])\\d{
2}[+\\-A](00[2-9]|[0][1-9]\\d|[1-9]\\d{2})[A-Za-z0-9])(\\s|,|;|\\Z)")
```

    - Match Type: **Column name**
        - Confidence: 15
        - Regular expression:

```
is_in_ci("Finland National ID","Finland National identification numb
er")
```

- Name: **GRC_NationalID_Detection**
    - Match Type: **Column value**
        - Confidence: 85
        - Regular expression:

```
regex("(\\s|,|;|\\A)([ABEZHIKMNOPTYX]{2}\\-\\d{6})(\\s|,|;|\\Z)")
```

    - Match Type: **Column name**
        - Confidence: 15
        - Regular expression:

```
is_in_ci("Greece National ID","Greece National identification number
")
```

- Name: **CHE_NationalID_Detection**
    - Match Type: **Column value**
        - Confidence: 85
        - Regular expression:

```
regex("(\\s|,|;|\\A)(756([ -.]?)\\d{4}\\3\\d{4}\\3\\d{2})(\\s|,|;|\\
Z)")
```

    - Match Type: **Column name**
        - Confidence: 15
        - Regular expression:

```
is_in_ci("Switzerland National ID","Switzerland National identificat
ion number")
```

- Name: **PRT_NationalID_Detection**
    - Match Type: **Column value**
        - Confidence: 85

    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)(\\d{8}([ -]?)\\d[0-9a-zA-Z]{2}\\d)(\\s|,|;|\\Z)
      ")
      ```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

      ```
      is_in_ci("Portugal
       National ID","Portugal National identification number")
      ```

- Name: **LTU_NationalID_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)([3-6]\\d{2}([0][1-9]|[1][012])(0[1-9]|[12][0-9]|
      3[01])\\d{3}\\d)(\\s|,|;|\\Z)")
      ```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

      ```
      is_in_ci("Lithuania
       National ID","Lithuania National identification number")
      ```

- Name: **POL_NationalID_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)(\\d{2}([0][1-9]|[1][012])(0[1-9]|[12][0-9]|3[01
      ])\\d{3}\\d{1}\\d{1})(\\s|,|;|\\Z)")
      ```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

      ```
      is_in_ci("Poland National ID","Poland National identification number
      ")
      ```

- Name: **LVA_NationalID_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)((0[1-9]|[12][0-9]|3[01])([0][1-9]|[1][012])\\d{
      2}\\-?[012]\\d{3}\\d)(\\s|,|;|\\Z)")
      ```

  - Match Type: **Column name**
    - Confidence: 15

    - Regular expression:

```
is_in_ci("Latvia National ID","Latvia National identification number
")
```

- Name: **DNK_NationalID_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("(\\s|,|;|\\A)(([0][1-9]|[12][0-9]|[3][01])([0][1-9]|[1][012])\
\d{2}[ -]?\\d{4})(\\s|,|;|\\Z)")
```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

```
is_in_ci("Denmark National ID","Denmark National identification numb
er")
```

- Name: **CZE_NationalID_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("(\\s|,|;|\\A)(\\d{4}(0[1-9]|[12][0-9]|3[01])\\/?\\d{4})(\\s|,
|;|\\Z)")
```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

```
is_in_ci("Czech Republic National ID","Czechia National ID","Czech R
epublic National identification
 number","Czechia National identification number")
```

- Name: **SWE_NationalID_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("(\\s|,|;|\\A)((19|20)?\\d{2}(0[1-9]|1[012])([06][1-9]|[12]\\d|
3[01])[-+]\\d{4})(\\s|,|;|\\Z)")
```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

```
is_in_ci("Sweden
 National ID","Sweden National identification number")
```

- Name: **ISL_NationalID_Detection**
  - Match Type: **Column value**
    - Confidence: 85

    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)((0[1-9]|[12][0-9]|3[01])([0][1-9]|[1][012])\\d{
      2}\\-\\d{2}\\d[09])(\\s|,|;|\\Z)")
      ```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

      ```
      is_in_ci("Iceland National ID","Iceland National identification numb
      er")
      ```

- Name: **ESP_NationalID_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)(\\d{8}([ -]?)[A-Z])(\\s|,|;|\\Z)")
      ```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

      ```
      is_in_ci("Spain National ID","Spain National identification number")
      ```

- Name: **EST_NationalID_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)([1-6]\\d{2}([0][1-9]|[1][012])(0[1-9]|[12][0-9]|
      3[01])\\d{3}\\d)(\\s|,|;|\\Z)")
      ```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

      ```
      is_in_ci("Estonia National ID","Estonia National identification numb
      er")
      ```

- Name: **ITA_NationalID_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)([A-Z]{3}[A-Z]{3}\\d{2}[A-EHLMPRST]([04][1-9]|[1
      256]\\d|[37][01])[A-Z]\\d{3}[A-Z])(\\s|,|;|\\Z)")
      ```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

      ```
      is_in_ci("Italy National ID","Italy National identification number")
      ```

- Name: **NationalID_Detection**

    - Match Type: **Column name**
        - Confidence: 15
        - Regular expression:

```
is_in_ci("National ID","National identification number")
```

- Name: **ROU_NationalID_Detection**
    - Match Type: **Column value**
        - Confidence: 85
        - Regular expression:

```
regex("(\\s|,|;|\\A)([1-9]\\d{2}([0][1-9]|[1][012])(0[1-9]|[12][0-9]|
3[01])([01234]\\d|5[012]|99)\\d{4})(\\s|,|;|\\Z)")
```

    - Match Type: **Column name**
        - Confidence: 15
        - Regular expression:

```
is_in_ci("Romania National ID","Romania National identification numb
er")
```

- Name: **BGR_NationalID_Detection**
    - Match Type: **Column value**
        - Confidence: 85
        - Regular expression:

```
regex("(\\s|,|;|\\A)(\\d{2}([024][1-9]|[135][012])(0[1-9]|[12][0-9]|
3[01])\\d{3}\\d{1})(\\s|,|;|\\Z)")
```

    - Match Type: **Column name**
        - Confidence: 15
        - Regular expression:

```
is_in_ci("Bulgaria National ID","Bulgaria National identification nu
mber")
```

- Name: **IRL_NationalID_Detection**
    - Match Type: **Column value**
        - Confidence: 85
        - Regular expression:

```
regex("(\\s|,|;|\\A)(\\d{7}[A-Za-z]{1,2})(\\s|,|;|\\Z)")
```

    - Match Type: **Column name**
        - Confidence: 15
        - Regular expression:

```
is_in_ci("Ireland National ID","Ireland National identification numb
er")
```

- Name: **SVK_NationalID_Detection**
    - Match Type: **Column value**
        - Confidence: 85

- Regular expression:

```
regex("(\\s|,|;|\\A)(\\d{4}(0[1-9]|[12][0-9]|3[01])\\/\\d{4})(\\s|,|
;|\\Z)")
```

- Match Type: **Column name**
  - Confidence: 15
  - Regular expression:

```
is_in_ci("Slovakia National ID","Slovakia National identification nu
mber")
```

### National insurance number (UK)

Regular expressions:

- Name: **uknationalinsurancenumber**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("\\b[A-CEGHJ-NOPR-TW-Z]{1}[A-CEGHJ-NPR-TW-Z]{1}[0-9]{6}[A-D\\s]
{1}\\b")
```

### Passport numbers

- 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)

Regular expressions:

- Name: **AUT_Passport_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("(\\s|,|;|\\A)([A-Z][ ]?\\d{7})(\\s|,|;|\\Z)")
```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

```
is_in_ci("Austria Passport Number")
```

- Name: **ESP_Passport_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)([A-Z]{2}[ -]?\\d{6})(\\s|,|;|\\Z)")
      ```
  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

      ```
      is_in_ci("Spain Passport Number")
      ```
- Name: **DEU_Passport_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)([0-9CF-HJ-NPRTV-Z]{9})(\\s|,|;|\\Z)")
      ```
  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

      ```
      is_in_ci("Germany Passport Number")
      ```
- Name: **FIN_Passport_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)([A-Za-z]{2}\\d{7})(\\s|,|;|\\Z)")
      ```
  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

      ```
      is_in_ci("Finland Passport Number")
      ```
- Name: **POL_Passport_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

      ```
      regex("(\\s|,|;|\\A)([A-Za-z]{2}\\d{7})(\\s|,|;|\\Z)")
      ```
  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

      ```
      is_in_ci("Poland Passport Number")
      ```
- Name: **CHE_Passport_Detection**

- Match Type: **Column value**
  - Confidence: 85
  - Regular expression:

```
regex("(\\s|,|;|\\A)([A-Z]\\d{7})(\\s|,|;|\\Z)")
```

- Match Type: **Column name**
  - Confidence: 15
  - Regular expression:

```
is_in_ci("Switzerland Passport Number")
```

- Name: **Passport_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("(\\s|,|;|\\A)([A-Z][A-Z0-9][0-9]{7})(\\s|,|;|\\A)","(\\s|,|;|
\\A)([A-Za-z]{2}\\d{7})(\\s|,|;|\\Z)","(\\s|,|;|\\A)([A-Z][ ]?\\d{7}
)(\\s|,|;|\\Z)","(\\s|,|;|\\A)([A-Z]{2}[0-9]{7})(\\s|,|;|\\Z)","(\\s
|,|;|\\A)([A-Za-z]{2}\\d{7})(\\s|,|;|\\Z)","(\\s|,|;|\\A)([A-Z]{2}[0
-9]{7})(\\s|,|;|\\Z)","(\\s|,|;|\\A)([0-9CF-HJ-NPRTV-Z]{9})(\\s|,|;|
\\Z)","(\\s|,|;|\\A)([A-Za-z]{2}\\d{6})(\\s|,|;|\\Z)","(\\s|,|;|\\A)
([A-Z]{2}[ -]?\\d{6})(\\s|,|;|\\Z)","(\\s|,|;|\\A)([A-Z]\\d{7})(\\s|
,|;|\\Z)","(\\s|,|;|\\A)(\\d{2}[A-Z]{2}\\d{5})(\\s|,|;|\\Z)")
```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

```
is_in_ci("Passport Number")
```

- Name: **GRC_Passport_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("(\\s|,|;|\\A)([A-Z]{2}[0-9]{7})(\\s|,|;|\\Z)")
```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

```
is_in_ci("Greece Passport Number")
```

- Name: **BEL_Passport_Detection**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("(\\s|,|;|\\A)([A-Za-z]{2}\\d{6})(\\s|,|;|\\Z)")
```

  - Match Type: **Column name**

        - Confidence: 15
        - Regular expression:

```
is_in_ci("Belgium Passport Number")
```

- Name: **FRA_Passport_Detection**
    - Match Type: **Column value**
        - Confidence: 85
        - Regular expression:

```
regex("(\\s|,|;|\\A)(\\d{2}[A-Z]{2}\\d{5})(\\s|,|;|\\Z)")
```

    - Match Type: **Column name**
        - Confidence: 15
        - Regular expression:

```
is_in_ci("France Passport Number")
```

- Name: **ITA_Passport_Detection**
    - Match Type: **Column value**
        - Confidence: 85
        - Regular expression:

```
regex("(\\s|,|;|\\A)([A-Z]{2}[0-9]{7})(\\s|,|;|\\Z)")
```

    - Match Type: **Column name**
        - Confidence: 15
        - Regular expression:

```
is_in_ci("Italy Passport Number")
```

- Name: **IRL_Passport_Detection**
    - Match Type: **Column value**
        - Confidence: 85
        - Regular expression:

```
regex("(\\s|,|;|\\A)([A-Z][A-Z0-9][0-9]{7})(\\s|,|;|\\Z)")
```

    - Match Type: **Column name**
        - Confidence: 15
        - Regular expression:

```
is_in_ci("Ireland Passport Number")
```

### Miscellaneous identifications: bank, passport, social security

- SWIFT code
- UK passport number
- Bank Routing Number
- US Social Security Number
- Society for Worldwide Interbank Financial Telecommunication (SWIFT)
- Telephone

Regular expressions:

- Name: **swiftcode**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("\\b[a-zA-Z]{4}(AF|AX|AL|DZ|AS|AD|AO|AI|AQ|AG|AR|AM|AW|AU|AT|A
Z|BS|BH|BD|BB|BY|BE|BZ|BJ|BM|BT|BO|BQ|BA|BW|BV|BR|IO|BN|BG|BF|BI|KH|
CM|CA|CV|KY|CF|TD|CL|CN|CX|CC|CO|KM|CG|CD|CK|CR|CI|HR|CU|CW|CY|CZ|DK|
DJ|DM|DO|EC|EG|SV|GQ|ER|EE|ET|FK|FO|FJ|FI|FR|GF|PF|TF|GA|GM|GE|DE|GH|
GI|GR|GL|GD|GP|GU|GT|GG|GN|GW|GY|HT|HM|VA|HN|HK|HU|IS|IN|ID|IR|IQ|IE|
IM|IL|IT|JM|JP|JE|JO|KZ|KE|KI|KP|KR|KW|KG|LA|LV|LB|LS|LR|LY|LI|LT|LU|
MO|MK|MG|MW|MY|MV|ML|MT|MH|MQ|MR|MU|YT|MX|FM|MD|MC|MN|ME|MS|MA|MZ|MM|
NA|NR|NP|NL|NC|NZ|NI|NE|NG|NU|NF|MP|NO|OM|PK|PW|PS|PA|PG|PY|PE|PH|PN|
PL|PT|PR|QA|RE|RO|RU|RW|BL|SH|KN|LC|MF|PM|VC|WS|SM|ST|SA|SN|RS|SC|SL|
SG|SX|SK|SI|SB|SO|ZA|GS|ES|LK|SD|SR|SJ|SZ|SE|CH|SY|TW|TJ|TZ|TH|TL|TG|
TK|TO|TT|TN|TR|TM|TC|TV|UG|UA|AE|GB|US|UM|UY|UZ|VU|VE|VN|VG|VI|WF|EH|
YE|ZM|ZW)[a-zA-Z0-9]{2}([a-zA-Z0-9]{3})?\\b")
```

  - Match Type: **Column name**
    - Confidence: 15
    - Regular expression:

```
is_in_ci("swiftcode")
```

- Name: **ukpassportnumber**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("\\b([A-Z0-9]{1}[0-9]{4}[A-Z0-9]{1})\\b")
```

- Name: **telephone**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("\\b(((\\+)?(1|7|2([07]|1[12368]|[2-6][0-9]|9[01789])|3([57][0
-9]|8[0125679]|[0123469])|4([013456789]|2[013])|5([1-8]|39(9)?|[09][
0-9])|6([0-6]|[78][0-9]|9[0-2])|(758)|8([1246]|0[089]|5[02356]|7(8|[
0-4])|8(1[01236789]|21[36]|[016]))|9(6[0-8]|7[0-7]|[0123458]|9(8|[2-
6]))|(1-(68|2(4[16]|6[48])|(34[05])|4(41|73)|6(49|64|7[01]|84)|7(58|
67|8[47])|8(0|6[89]|[24]9|76)|(939)))|(7-7))(\[ )?)?([2-9][0-9]{9}))
\\b","\\b(((\\+)?(1|7|2([07]|1[12368]|[2-6][0-9]|9[01789])|3([57][0-
9]|8[0125679]|[0123469])|4([013456789]|2[013])|5([1-8]|39(9)?|[09][0
-9])|6([0-6]|[78][0-9]|9[0-2])|(758)|8([1246]|0[089]|5[02356]|7(8|[0
-4])|8(1[01236789]|21[36]|[016]))|9(6[0-8]|7[0-7]|[0123458]|9(8|[2-6
]))|(1-(68|2(4[16]|6[48])|(34[05])|4(41|73)|6(49|64|7[01]|84)|7(58|6
7|8[47])|8(0|6[89]|[24]9|76)|(939)))|(7-7))(\[ )?([2-9][0-9]{2}[\[
][0-9]{3}[\\][0-9]{4}))\\b","\\b(((\\+)?(1|7|2([07]|1[12368]|[2-6][0
-9]|9[01789])|3([57][0-9]|8[0125679]|[0123469])|4([013456789]|2[013]
)|5([1-8]|39(9)?|[09][0-9])|6([0-6]|[78][0-9]|9[0-2])|(758)|8([1246]
|0[089]|5[02356]|7(8|[0-4])|8(1[01236789]|21[36]|[016]))|9(6[0-8]|7[
0-7]|[0123458]|9(8|[2-6]))|(1-(68|2(4[16]|6[48])|(34[05])|4(41|73)|6
(49|64|7[01]|84)|7(58|67|8[47])|8(0|6[89]|[24]9|76)|(939)))|(7-7))[-
\[ ])?([2-9][0-9]{2}-[0-9]{3}-[0-9]{4}))\\b","\\b(((\\+)?(1|7|2([07]
|1[12368]|[2-6][0-9]|9[01789])|3([57][0-9]|8[0125679]|[0123469])|4([
```

```
013456789]|2[013])|5([1-8]|39(9)?|[09][0-9])|6([0-6]|[78][0-9]|9[0-2
])|(758)|8([1246]|0[089]|5[02356]|7(8|[0-4])|8(1[01236789]|21[36]|[0
16]))|9(6[0-8]|7[0-7]|[0123458]|9(8|[2-6]))|(1-(68|2(4[16]|6[48])|(3
4[05])|4(41|73)|6(49|64|7[01]|84)|7(58|67|8[47])|8(0|6[89]|[24]9|76)
|(939))))|(7-7))[\\.\[ ])?([2-9][0-9]{2}\\.[0-9]{3}\\.[0-9]{4}))\\b",
"((\\+)?(1|7|2([07]|1[12368]|[2-6][0-9]|9[01789])|3([57][0-9]|8[0125
679]|[0123469])|4([013456789]|2[013])|5([1-8]|39(9)?|[09][0-9])|6([0
-6]|[78][0-9]|9[0-2])|(758)|8([1246]|0[089]|5[02356]|7(8|[0-4])|8(1[
01236789]|21[36]|[016]))|9(6[0-8]|7[0-7]|[0123458]|9(8|[2-6]))|(1-(6
8|2(4[16]|6[48])|(34[05])|4(41|73)|6(49|64|7[01]|84)|7(58|67|8[47])|
8(0|6[89]|[24]9|76)|(939))))|(7-7))[\[ ]?)?(\\([2-9][0-9]{2}\\)[\[ ][
0-9]{3}-[0-9]{4})","((\\+)?(1|7|2([07]|1[12368]|[2-6][0-9]|9[01789])
|3([0123469]|[57][0-9]|8[0125679])|4(2[013]|[013456789])|5([1-8]|399
|[09][0-9]|39)|6([0-6]|[78][0-9]|9[0-2])|(758)|8(0[089]|5[02356]|7([
0-4]|8)|8([016]|1[01236789]|21[36])|[1246])|9(6[0-8]|7[0-7]|9([2-6]|
8)|[0123458])|(1-(68|2(4[16]|6[48]|84)|(34[05])|4(41|73)|6(49|64|7[0
1]|84)|7(58|67|8[47])|8(0[89]|[24]9|6[89]|76)|(939))))|(7-7))\\.)?(\\
([2-9][0-9]{2}\\)[\\.][0-9]{3}\\.[0-9]{4})","((\\+)?(1|7|2([07]|1[12
368]|[2-6][0-9]|9[01789])|3([57][0-9]|8[0125679]|[0123469])|4([01345
6789]|2[013])|5([1-8]|39(9)?|[09][0-9])|6([0-6]|[78][0-9]|9[0-2])|(7
58)|8([1246]|0[089]|5[02356]|7(8|[0-4])|8(1[01236789]|21[36])|[016])
)|9(6[0-8]|7[0-7]|[0123458]|9(8|[2-6]))|(1-(68|2(4[16]|6[48])|(34[05
])|4(41|73)|6(49|64|7[01]|84)|7(58|67|8[47])|8(0|6[89]|[24]9|76)|(93
9))))|(7-7)))?(\\([2-9][0-9]{2}\\)[0-9]{3}-[0-9]{4})","((\\+)?(1|7|2(
[07]|1[12368]|[2-6][0-9]|9[01789])|3([0123469]|[57][0-9]|8[0125679])
|4(2[013]|[013456789])|5([1-8]|399|[09][0-9]|39)|6([0-6]|[78][0-9]|9
[0-2])|(758)|8(0[089]|5[02356]|7([0-4]|8)|8([016]|1[01236789]|21[36]
)|[1246])|9(6[0-8]|7[0-7]|9([2-6]|8)|[0123458])|(1-(68|2(4[16]|6[48]
|84)|(34[05])|4(41|73)|6(49|64|7[01]|84)|7(58|67|8[47])|8(0[89]|[24]
9|6[89]|76)|(939))))|(7-7))\\.)?(\\([2-9][0-9]{2}\\)[\\.][0-9]{3}\\.[
0-9]{4})")
```

- Match Type: **Column name**
  - Confidence: 15
  - Regular expression:

```
is_in_ci("telephone","tel")
```

- Name: **routingnumber**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("\\b[^459][0-9]{8}\\b")
```

- Name: **ssn**
  - Match Type: **Column value**
    - Confidence: 85
    - Regular expression:

```
regex("\\b((((?!(000|666))([0-8][0-9][0-9]))((?!00)([0-9]{2}))((?!00
00)([0-9]{4}))))|(98765432[0-9]))\\b","\\b((((?!(000|666))([0-8][0-9]
[0-9]))[\\]((?!00)([0-9]{2}))[\\]((?!0000)([0-9]{4}))))|((987)-(65)-(
432)[0-9]))\\b","\\b((((?!(000|666))([0-8][0-9][0-9]))-((?!00)([0-9]
{2}))-((?!0000)([0-9]{4}))))|((987)-(65)-(432)[0-9]))\\b")
```

  - Match Type: **Column name**
    - Confidence: 15

- Regular expression:

```
is_in_ci("ssn","ssnnumber","ssnno")
```

## 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.

### Categorical Data

The Hive Column Profiler can categorize your Hive tables according to their data type. Datatypes are received from Atlas. These include the following:

- array
- bigint
- binary
- boolean
- char
- date
- decimal
- double
- float
- int[1]
- map
- smallint
- string
- struct
- timestamp
- tinyint
- varchar

**Note:** The Hive Column Profiler can profile Iceberg tables.

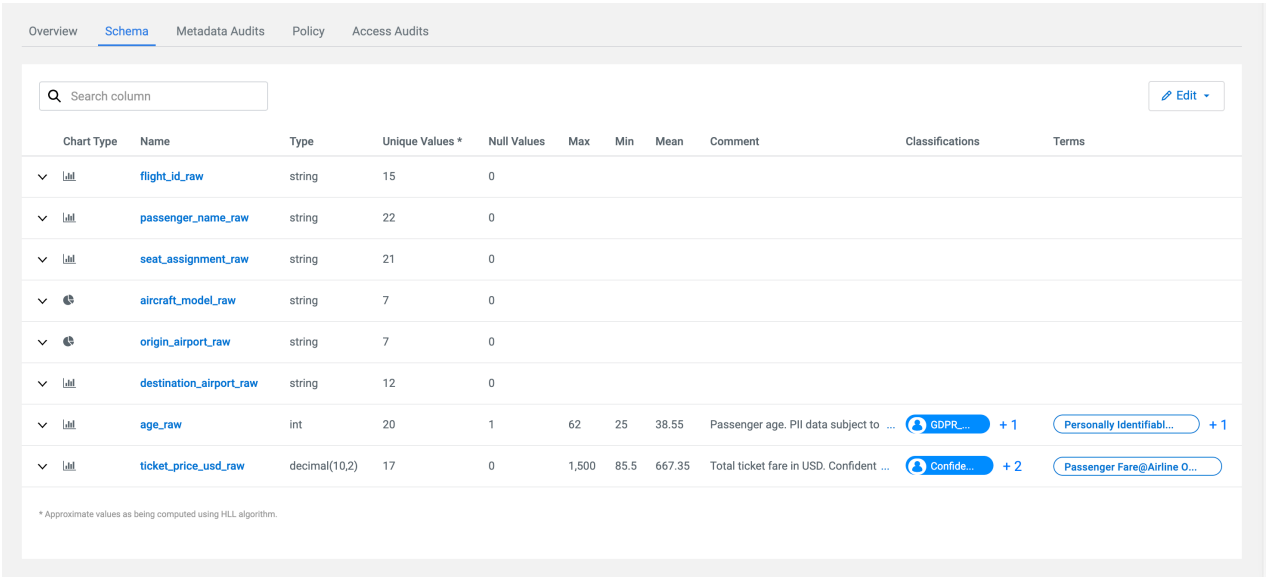
### Numerical Data

**Note:** Non-primitive data types are not supported. Only basic data types such as numbers, string, boolean, floating point numbers and so on are supported

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:

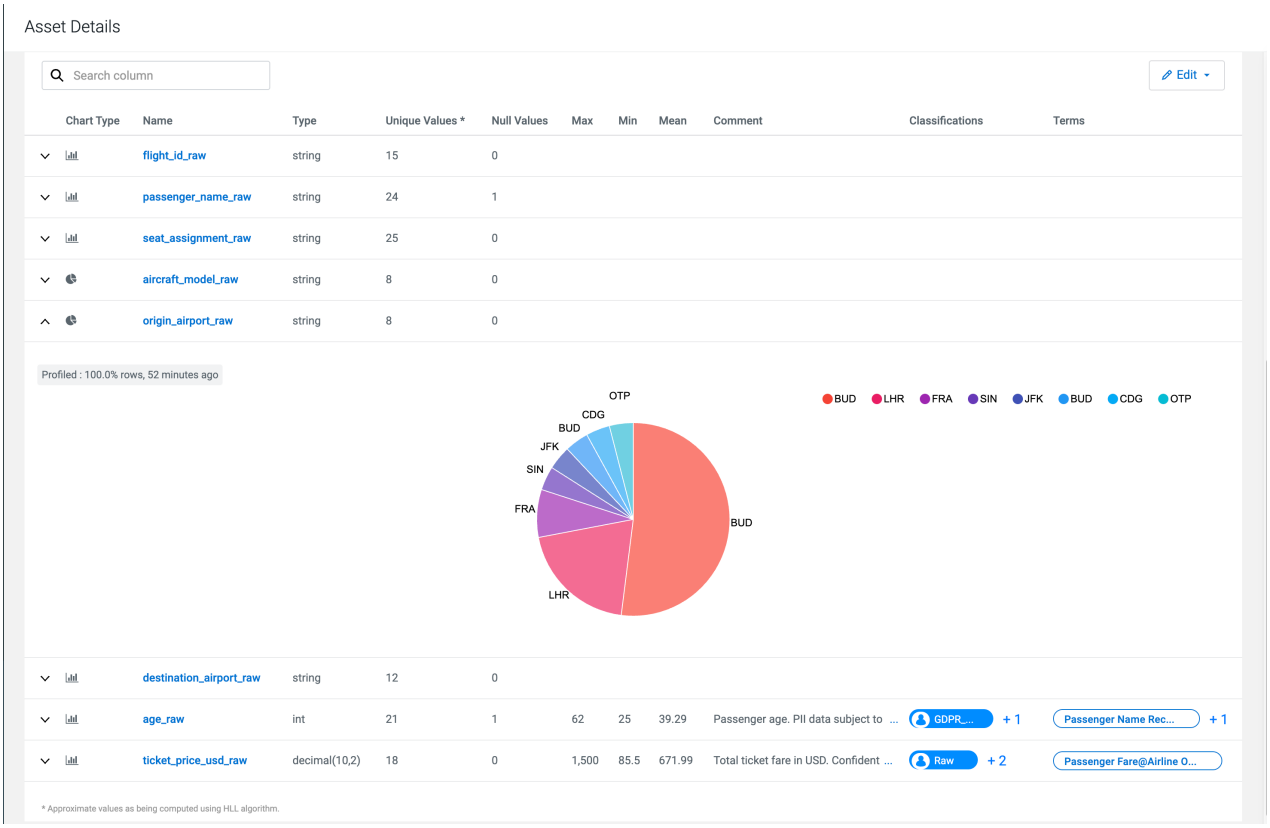
- Unique Values
- Null Values
- Max
- Min
- Mean

[1] These rows will receive the boxplot graphs and the min, max mean values after profiling.




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).



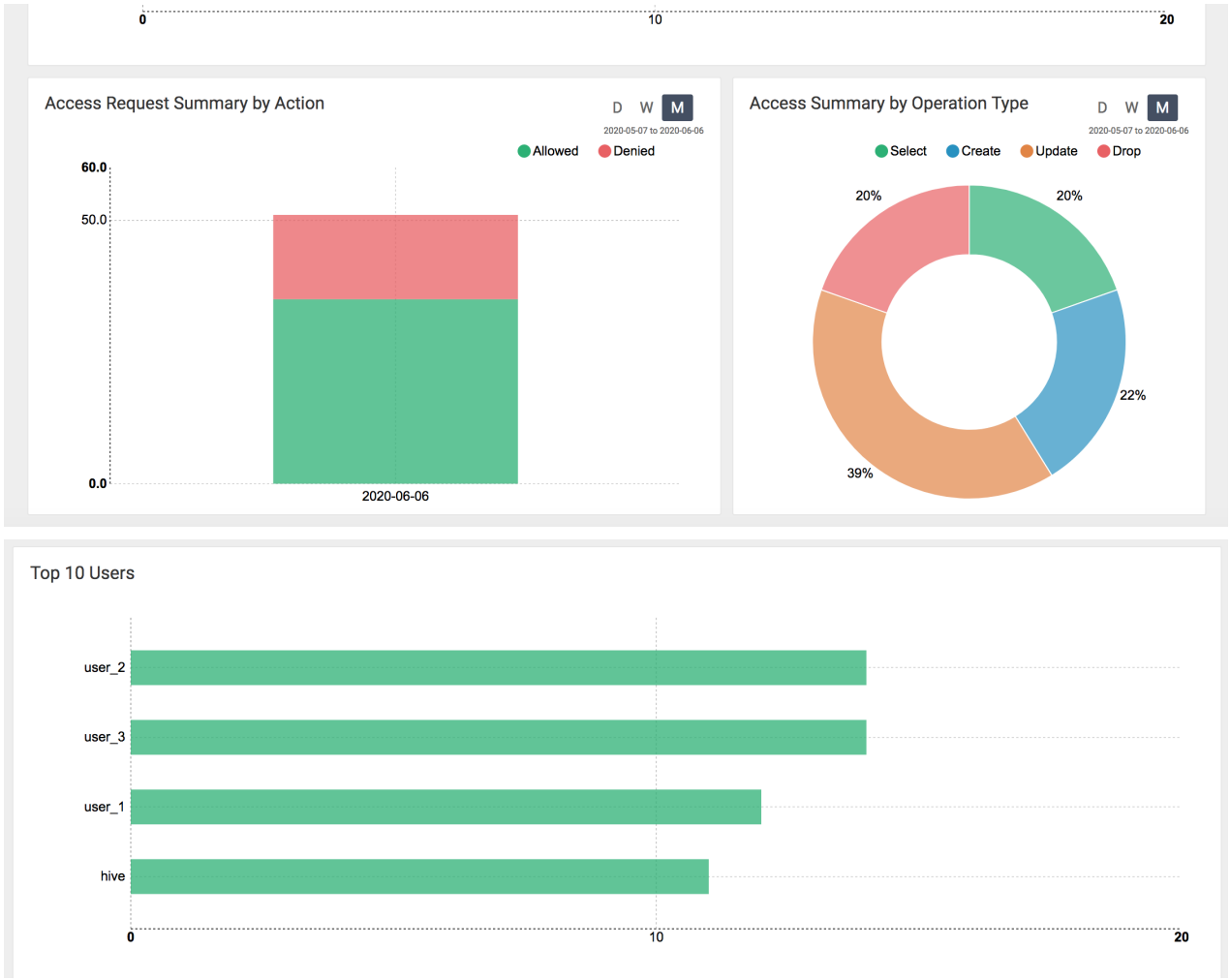

When the data within columns is numeric, a histogram of the distribution of values organized into 10 groups (decile frequency histogram) and a box plot with a five-number summary (mean, median, quartiles, maximum, and minimum values) are shown for the column, if there are more than 10 elements available.




## 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