

Step-by-step instructions:

Part 1 - Data Catalog [8-10 minutes]

Overview: What is Cloudera Data Catalog?

Data Catalog is a service that enables you to understand, manage, secure, and govern data assets across the enterprise, through multiple clusters and multiple CDP environments. You can search to locate relevant data of interest based on various parameters.

Using Data Catalog, you can understand how data is interpreted for use, how it is created and modified, and how data access is secured and protected.

Purpose: Search for a dataset (table) in Data Catalog, called “flights”.

- Find what database(s) the table “flights” is located.
- Find out at least one year that the “flights” table was generated from.
- Find out how many columns the table “flights” contains.

1) Open CDP, using the “admin” user within the Test Drive link.

Your link should look something like (remember click the link in your email not the link below)

http://login.trycdp.com/auth/realms/trycdp-trialxx/protocol/saml/clients/samlclient?tn=trialxx_admin@trycdp.com&p=X

*xx represents the trial user #

*X represents the password

2) Click the “Data Catalog” within the CDP Home Screen



3) Type “flights” in the search box and click “flights” under suggestions

Data Catalog / Search

The screenshot shows the 'Data Catalog / Search' interface. At the top, there is a search bar with the text 'flights'. Below the search bar, there are two sections: 'Entities' and 'Suggestions'. The 'Entities' section lists five items, each with a small icon and a label: 'actualelapsedtime (hive_column)', 'securitydelay (hive_column)', 'year (hive_column)', 'cancelled (hive_column)', and 'weatherdelay (hive_column)'. The 'Suggestions' section shows a single suggestion: 'flights'.

4) Click “Hive Table” under Filters on the left

The screenshot shows the 'Data Catalog / Search' interface with filters applied. On the left, there is a 'Filters' section. Under the 'TYPE' filter, 'Hive Table' is selected. The main table shows search results for 'flights'. The table has three columns: 'Type', 'Name', and 'Location'. There are two rows of results, both of which are 'Hive Table' type and named 'flights'. The first row has a location of '/airlines_new_orc' and the second row has a location of '/airlines_new_parquet'.

*Find what database(s) the table “flights” is located.

5) Click “flights” where the Location = /airlines_new_orc

The screenshot shows the 'Data Catalog / Search' interface with filters applied. On the left, there is a 'Filters' section. Under the 'TYPE' filter, 'Hive Table' is selected. The main table shows search results for 'flights'. The table has three columns: 'Type', 'Name', and 'Location'. There are two rows of results, both of which are 'Hive Table' type and named 'flights'. The first row has a location of '/airlines_new_orc' and the second row has a location of '/airlines_new_parquet'. The first row is highlighted with a red circle.

6) Zoom into the Lineage and scroll over one of the /cdp-lake/data, clicking the “i” for more information

Data Catalog / Asset Details

Name
flights

Type
HIVE TABLE

Data Lake
cdptrialuser31-dl

Dataset
0

Overview

Schema

Metadata Audits

Policy

Access Audits

III 29
Number of Columns

Asset Properties

Owner: **csso_trialuser31**
Qualified Name: **airlines_new_orc.flights@cm**
Created On: **Wed Jan 13 2021 01:10:57 GMT-0600 (Central Standard Time)**
Last Access Time: **Wed Jan 13 2021 01:10:57 GMT-0600 (Central Standard Time)**

Table Type: **MANAGED_TABLE**
Database: **airlines_new_orc**
DB Catalog: **cm**
Parent: **airlines_new_orc**

Managed Classifications | 0

Propagated Tags

Lineage

Legend: → Lineage ⇌ Impact ⇌ Replication ○ Current Entity

/cdp-lake/data/airlines/airlines_new_orc.db/flights/year=1997

Guid: **be5a2094-f572-432e-9fa4-96498f7db650**

Type Name: **aws_s3_pseudo_dir**

Classifications(0): **--**

Owner: **-NA-**

Qualified Name: **s3a://prod-cdptrialuser31-trycdp-com/cdp-lake/data/airlines/airlines_new_orc.db/flights/year=1997@cm**

Created On: **-NA-**

Update Time: **-NA-**

Created By: **csso_trialuser31**

Updated By: **csso_trialuser31**

*Find out at least one year that the “flights” table was generated from.

*Find out how many columns the table “flights” contains.