

..

Self-service Exploratory Analytics

Date published: 2021-07-15

Date modified:

CLOUDERA

Legal Notice

© Cloudera Inc. 2024. 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

Self-service exploratory analytics.....	4
Understanding the use case.....	5
Self-service exploratory analytics use case.....	6
Self-service exploratory analytics steps.....	8
Meeting the prerequisites.....	8
Downloading sample data.....	9
Exploring and querying data in Hue.....	11
Generating and sharing dashboards.....	17
Troubleshooting.....	19
Unable to see the S3 File Browser on the Hue web interface.....	19
External File option is unavailable.....	19
Error while accessing a directory in S3.....	20
Incorrect region error while uploading a file to S3.....	20

Self-service exploratory analytics

The self-service exploratory analytics pattern provides exact steps that you can follow to quickly explore, query, and analyze data present in CDP Data Lakes, public cloud storage, or on your computers in a file, and generate visual dashboards—all using Cloudera Data Warehouse.

Figure 1: Self-service exploratory analytics in a nutshell

SELF-SERVICE EXPLORATORY ANALYTICS

EXPLORE | QUERY | VISUALIZE



Explore stored data sets in CDP Data Lakes, public cloud storage, or upload data sets from your computer and analyze them in BI visualizations on a self-service basis without direct support from central IT to quickly provide one-time business insights

Go from data to dashboard in 15 minutes



Data Analyst



Business stakeholders



Cloudera Data Warehouse

Hue

Explore, query, and analyze data



Cloudera Data Visualization

Generate and share visual dashboards



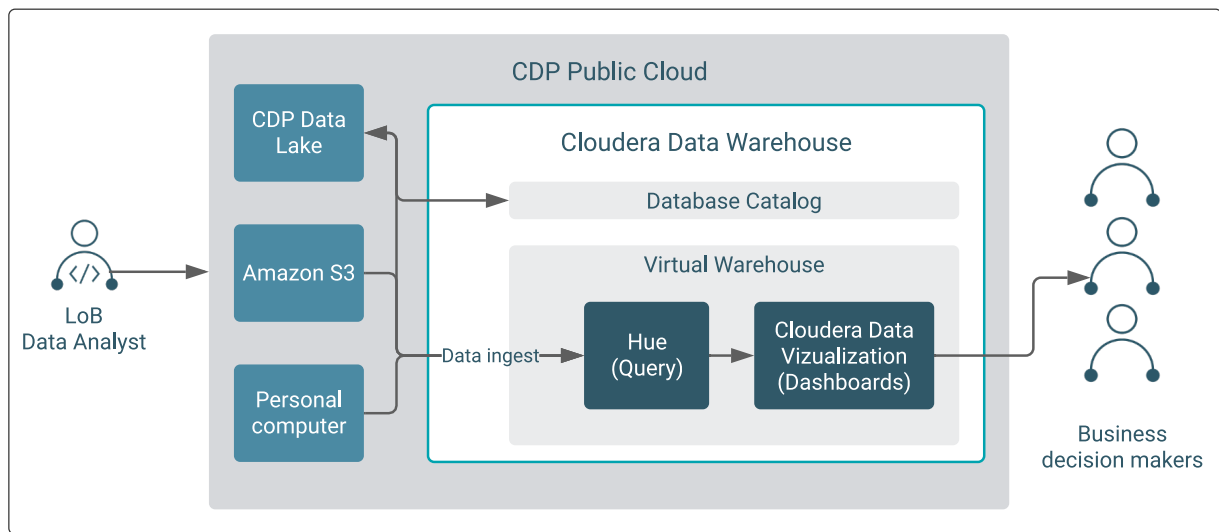
Understanding the use case

Learn about the self-service exploratory analytics pattern.

In modern data-driven businesses, Line of Business (LoB) Data Analysts are often tasked with exploring and analyzing new data sets and providing business insights on the fly, without waiting for the curated and prepared data or depending on the central IT.

Today, it is hard for LoB Data Analysts, and the Data Engineers that support them, to respond in a timely manner to LoB's needs to update existing dashboards and reports for new insights or to create new reports to respond to an emergent need. Often, these requests need integration of various data and application silos that require high levels of support from central IT services to make resources available, to set up ingest and curation processes to bring usable data to the report or dashboard consumer or to safely import and process the data available outside the current infrastructure (flat files on a user computer). This process is time-consuming, expensive, and often quite rigid. As a result, many LoB decisions are delayed, costing businesses opportunities, revenue, and customers.

With Cloudera Data Warehouse (CDW), LoB Data Analysts and Business Analysts can explore data sets stored in Cloudera Data Platform (CDP) Data Lakes, cloud object stores such as Amazon S3, or uploaded data sets from their computers and analyze them in new or existing BI visualizations on a self-service basis without direct support from central IT resources, as shown in the following diagram:



CDW enables you to go from data to dashboard in 15 minutes.

Self-service exploratory analytics use case

A real-life business narrative enables you to follow and understand the Self-service exploratory analytics CDP pattern. The narrative introduces the key personas at play, establishes burning business questions, and sets parameters for the success criteria of the pattern.

This is a sample use case. The steps in this pattern can be applied to similar use cases across other industries.

The use case: Sales promotion at duty-free stores for airline customers with longer layovers or flight delays

You are a marketing data aggregator (Special Marketing Group (SMG)) company, helping a group of duty-free shops in airports to answer some questions around a new marketing campaign they plan to run. The duty-free shop owners want to partner with specific airlines to offer discounts through the airlines to their passengers who spend more time in an international terminal than warranted for their flight. In this partnership, the airlines can improve customer satisfaction during extended wait times, and duty-free shops can increase sales.

A representative of a number of duty-free shops has requested help to select airlines to partner with. The shops will offer discounts to passengers of those airlines experiencing delays.

To decide which airlines to partner with, the marketing managers from the duty-free shops need to know which airlines have the most customers who experience delays and base the size of the discount on a range of dissatisfaction levels.

This short-lived sales promotion does not warrant a permanent dashboard or report, and you do not expect more than 100 people to access the data resulting from this pattern.

Key personas of this use case

The primary personas have access to the data and perform analytics on it. The secondary personas draft sales promotion policies from business insights.

Primary personas:

- **Data Analyst:** The Data Analyst from the SMG company is responsible for providing the duty-free shop owners with a visual tool to choose airline partners for offering duty-free discounts.
- **LoB Marketing Representative:** The LoB Marketing Representative for the duty-free shops is globally responsible for supporting a number of duty-free shop owners in sales and marketing activities, and requests decision support services from the SMG company.

Secondary personas:

- **Marketing Data Analysts:** The Marketing Data Analysts from each duty-free organization would want to access the dashboards for insight. The assumption is that there are no more than 10 duty-free shop management firms with no more than 10 analysts interested in evaluating data that is published based on this request.
- **LoB Managers:** Managers from each duty-free organization use the insight and draft the promotional policies.

Criteria and conditions for the sales promotion

The LoB Marketing Representative has identified ideal airline partners as ones that have the most customers that experience the following inconveniences:

- Long layovers before an international flight
- Elongated layovers caused by delayed international flights (connection)
- Missed international connections due to a delay on the pre-connection flights

The LoB Marketing Representative wants to offer a discount program in partnership with their ideal airlines, which partially subsidize the discount percentage based on passenger inconveniences as follows:

- 10% discount at the duty-free shop for passengers with long layovers not caused by delays. A long layover is any layover greater than 90 minutes.
- 20% discount for passengers that have an elongated layover due to a delay in their next leg. An elongated delay is a delay of the departing flight greater than 60 minutes.
- 30% discount for passengers displaced by missing their connecting flight due to a delay in their previous leg. A missed connection is a delay of 45 minutes, or more, between arrival and the departure of the connecting international flight. The minimum time between flights to safely make a connection is 45 minutes.

Burning business questions

The marketing data coordinator from the SMG company has data about the airlines, flights, airports, and passengers in a number of places. The SMG Data Analyst needs to address the following business questions using data they have aggregated from airlines and airports tables:

- Which airlines have the most passengers who have long layovers built into their tickets?
- Which airlines have the most passengers who have delayed international legs in their itinerary causing an extended layover in their flight itinerary?
- Which airlines have the most passengers who are displaced by a missed international connection, caused by a delay in their previous leg?

Success criteria

Going from data to dashboard in 15 minutes.

The SMG Data Analyst is able to explore and analyze data using Hue and can share a dashboard in CDP Data Visualization with the Marketing Data Analysts and LoB Managers. The Marketing Data Analysts and LoB Managers should be able to use this dashboard and the insights to frame the promotional policies for their duty-free outlets.

Self-service exploratory analytics steps

For this CDP pattern, you must be onboard with CDP Public Cloud and your IT team must have set up the Cloudera Data Warehouse service with other necessary prerequisites listed in the *Meeting the prerequisites* section. As a Data Analyst or a Line of Business user, you can then explore and query data, import new or missing data, create data sets for the business questions that you are trying to answer, and generate visual dashboards to share with your stakeholders.

Meeting the prerequisites

Before the Data Analysts can explore and query data, your central or departmental IT must have onboarded to CDP Public Cloud and must meet the requirements listed in this section.

To do in the CDP Management Console

1. Make sure that the Cloudera Data Lake is created and running. For more information, see [Creating an AWS environment with a medium duty data lake using the CLI](#).
2. Grant EnvironmentUser role to the Data Analyst user and synchronize the user to FreeIPA. For more information, see [Assigning account roles to users](#) and [Performing user sync](#).

To do in Cloudera Data Warehouse (CDW)

1. Make sure that the CDW environment is activated. For more information, see [Activating AWS environments](#).
2. Add an S3 bucket as an external bucket in the CDW environment with read-only access.



Important: You must add your own S3 bucket to the CDW environment.

- a. Go to CDW service Environments and click its edit icon.
 - b. On the Environment Details page, type the name of the AWS bucket you want to configure access to in the Add External S3 Bucket text box and select read-only access mode.
 - c. Click Add Bucket and then click Apply to update the CDW environment.
3. Create a non-default Database Catalog with the Load Demo Data option enabled.
 - a. Go to CDW service Database Catalogs and click Add New.
 - b. Specify a name, select an environment, select SDX from the Datalake drop-down list and enable the Load Demo Data option, and click CREATE.
 4. Create an Impala-based Virtual Warehouse with the Enable Data Visualization option and make sure it is in the running state.
 - a. Go to CDW service Virtual Warehouses and click Add New.
 - b. Select the type as IMPALA, select the Database Catalog associated with the Virtual Warehouse, select the size, click Enable Data Visualization, and then click CREATE.

5. Enable the S3 File Browser for Hue.

- a. Go to CDW service Virtual Warehouses , select your Virtual Warehouse and click edit.
- b. On the **Virtual Warehouses** page, go to CONFIGURATIONS Hue and select hue-safety-valve from the drop-down list.
- c. Add the following configuration for Hive or Impala Virtual Warehouse in the space provided and click APPLY:

```
[desktop]
# Remove the file browser from the blocked list of apps.
# Tweak the app_blacklist property to suit your app configuration.
app_blacklist=spark,zookeeper,hive,hbase,search,oozie,jobsup,pig,sqoop
,security

[aws]
[[aws_accounts]]
[[[default]]]
access_key_id=[ ***AWS-ACCESS-KEY*** ]
secret_access_key=[ ***SECRET-ACCESS-KEY*** ]
region=[ ***AWS-REGION*** ]
```

6. Enable a link to the Data VIZ application (Cloudera Data Visualization) from the Hue web interface or provide the Data VIZ application URL to the Data Analyst.
 - a. Go to CDW service Virtual Warehouses , select your Virtual Warehouse and click edit.
 - b. On the Virtual Warehouses details page, go to CONFIGURATIONS Hue and select hue-safety-valve from the drop-down list.
 - c. Add the following lines in the safety valve and click APPLY:

```
[desktop]
custom_dashboard_url=[ ***DATA-VIZ-URL*** ]
```

To do in Ranger

Grant the required DDL and DML Hadoop SQL policies to the Data Analyst user in Ranger.

1. Go to CDW service Database Catalogs , click the more option, and click Open Ranger.
2. On the **Ranger Service Manager** page, click Hadoop SQL.
3. Select the all - url policy.

The **Edit Policy** page is displayed.

4. Under the Add Conditions section, add the users under the Select User column and add permissions such as Create, Alter, Drop, Select, and so on from the Permissions column.
5. Scroll to the bottom of the page and click Save.

Other prerequisites

- Data in a file to be ingested is in a structured format, such as CSV, with headers.
- The data file is available to the Data Analyst on their computer or accessible on a shared drive with permissions already granted.
- Hue application URL is shared with the Data Analyst, if a link to the Data Visualization application is not enabled in Hue.

Downloading sample data

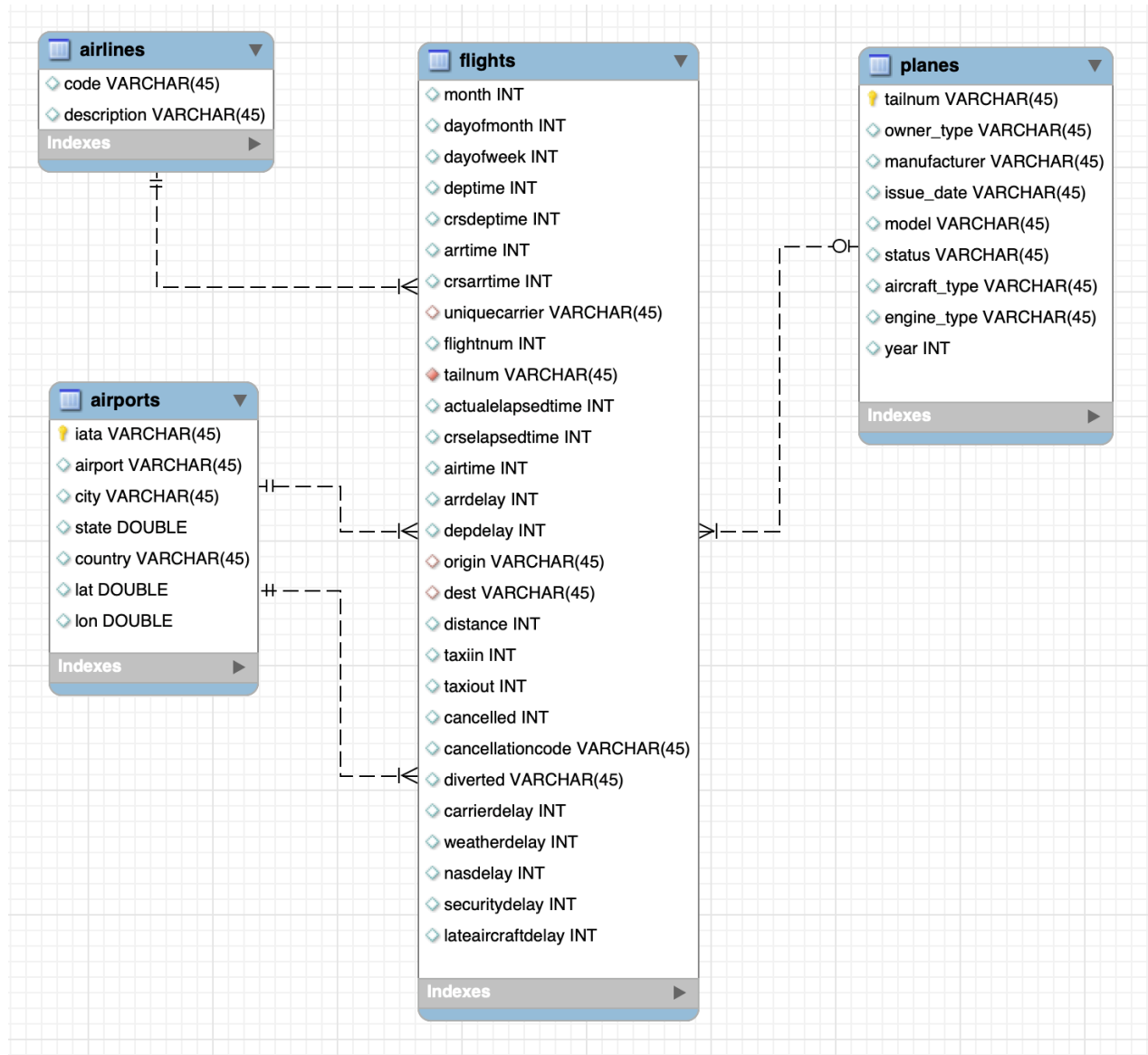
The airline's data is already available in CDW (Hue), but the passenger data is not. Download the sample data for this exercise provided with this pattern.

Airlines data (available in CDP)

The airline data consists of the following four tables:

- airlines
- airports
- flights
- planes

The airline data looks as follows:



Passenger data (to be downloaded)

The passenger data is present in the following CSV file: [passengertickets.csv](#). You will create a “unique_tickets” table from the CSV file by importing it in Hue.

SQL queries for this exercise (to be downloaded)

The following file contains the SQL queries that you will use for this exercise: [sample-queries.txt](#)

Exploring and querying data in Hue

As an SMG Data Analyst, you follow logical steps to explore and query the airline data to create data sets and then use those data sets to draw insight and create dashboards.

Exploring the databases and airline tables

Use Hue to look up existing databases, tables, and data that is available in your data warehouse.

Procedure

1. Log in to Hue with your credentials.
2. Explore the databases and airline tables from the left-assist panel.
3. Run the following query to get a list of international flights:

```
-- Query to find all international flights: flights where destination ai
rport country is not the same as origin airport country
SELECT DISTINCT
    flightnum,
    uniquecarrier,
    origin,
    dest,
    month,
    dayofmonth,
    `dayofweek`
FROM
    `airline_ontime_orc`.flights f,
    `airline_ontime_orc`.airports oa,
    `airline_ontime_orc`.airports da
WHERE
    f.origin = oa.iata
    and f.dest = da.iata
    And oa.country <> da.country
ORDER BY
    month ASC,
    dayofmonth ASC
;
```

Importing the passenger data

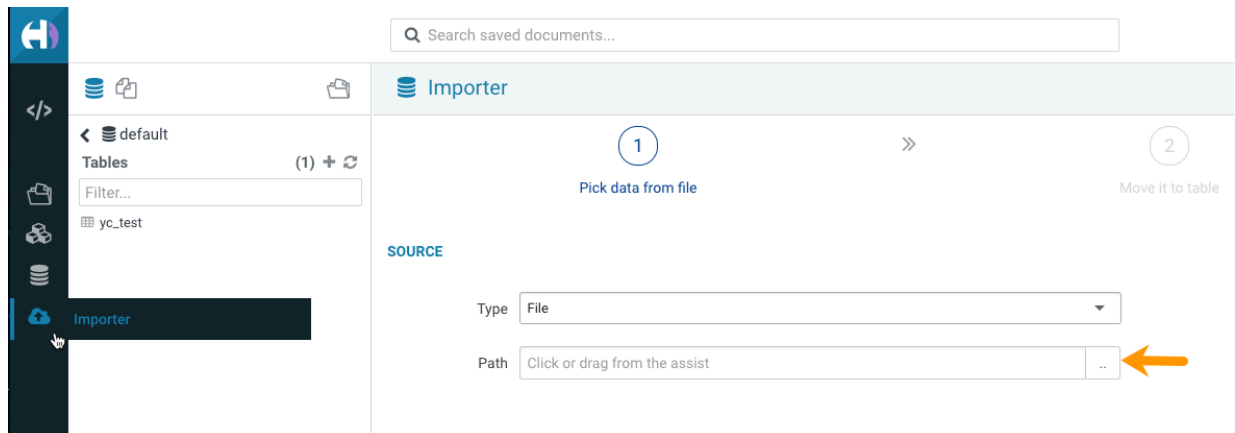
Use Hue to import data which is not present in your data warehouse, but is available on your computer or cloud storage such as Amazon S3.

About this task

In this exercise, the passenger data is not already available in Hue or in CDP but instead is available on your computer in a CSV file. Import the passenger data into Hue from your computer as follows:

Procedure

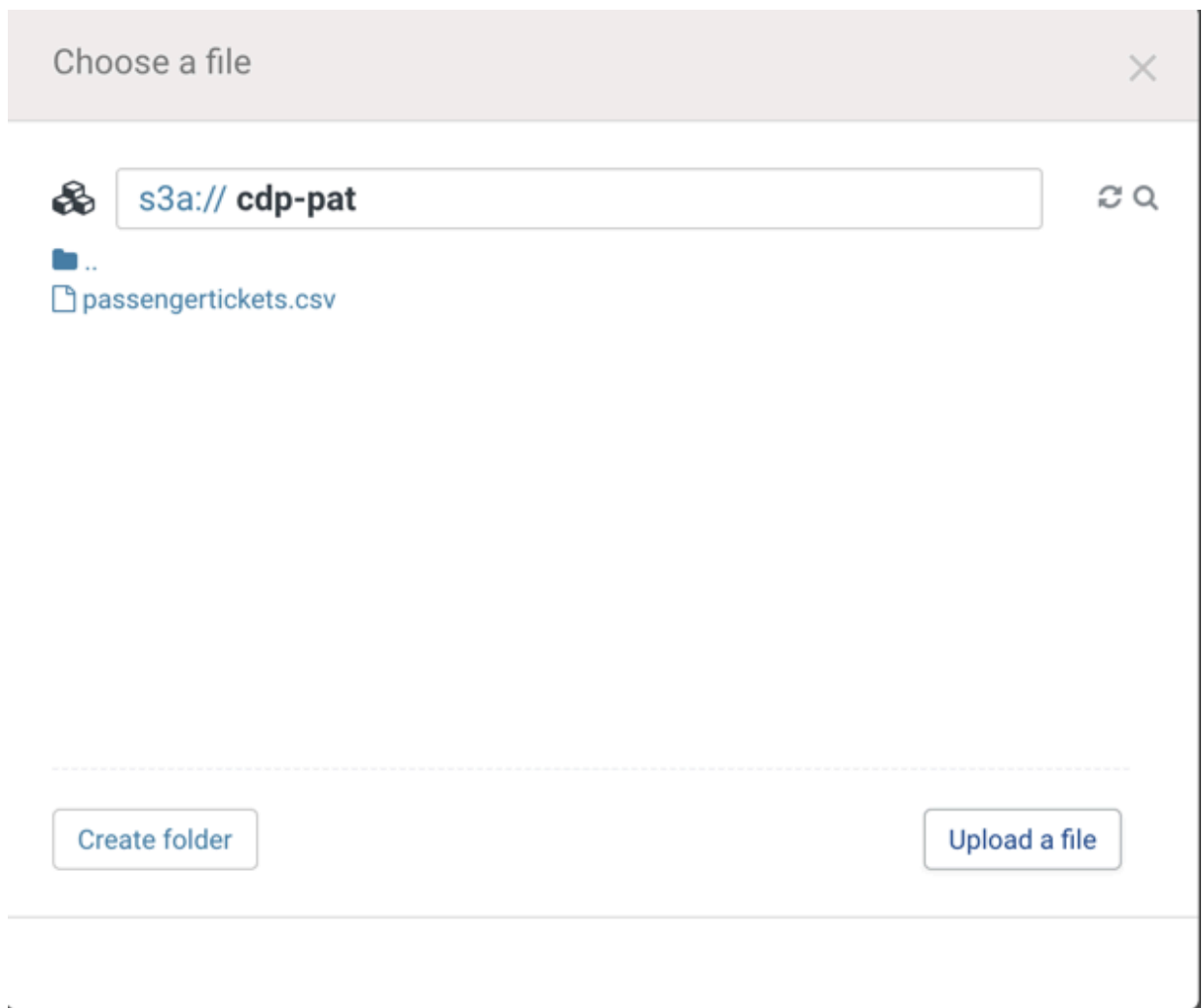
1. Go to Importer and click .. at the end of the Path field.



A pop-up window is displayed where you can choose a file.

2. Type s3a:// in the address text box and press enter.

The S3 buckets associated with the CDW environment are displayed. You can narrow down the list of results using the search option.



3. Navigate to the bucket in which you want to upload the file and click Upload a file.

- Select the CSV file that you want to import into Hue.

Hue displays the preview of the table along with the format.

Hue automatically detects the field separator, record separator, and the quote character from the CSV file. If you want to override a specific setting, select a different value from the drop-down list.

Type Remote File

Path s3a://cdp-pat/passengertickets.csv

FORMAT

File Type CSV File

Field Separator Comma (,) Record Separator New line Quote Character Double Quote

☒ Has Header

PREVIEW

ticketnumber	leg1flightnum	leg1uniquecarrier	leg1origin	leg1dest	leg1month	leg1dayofmonth
6857528732333	662	NW	PHX	MEM	10	10
2331153269614	550	NW	LAX	MEM	10	10
4191763179365	282	NW	SAT	MEM	10	10
7435945889948	1036	NW	MCI	MEM	10	10
6142750610063	1084	NW	IAH	MEM	10	10

- Click Next.

On this page under DESTINATION, enter `airline_ontime_orc.unique_tickets` in the Name field.

The `unique_tickets` table is created in the `airline_ontime_orc` database.

- Select Format as Text.


7. Expand Extras and select the Store in Default location option.

Type Table

Name airline_ontime_orc.unique_tickets

PROPERTIES

Format Text

Extras 

☒ Store in Default location

☒ Import data

Description Description

☒ Use first row as header

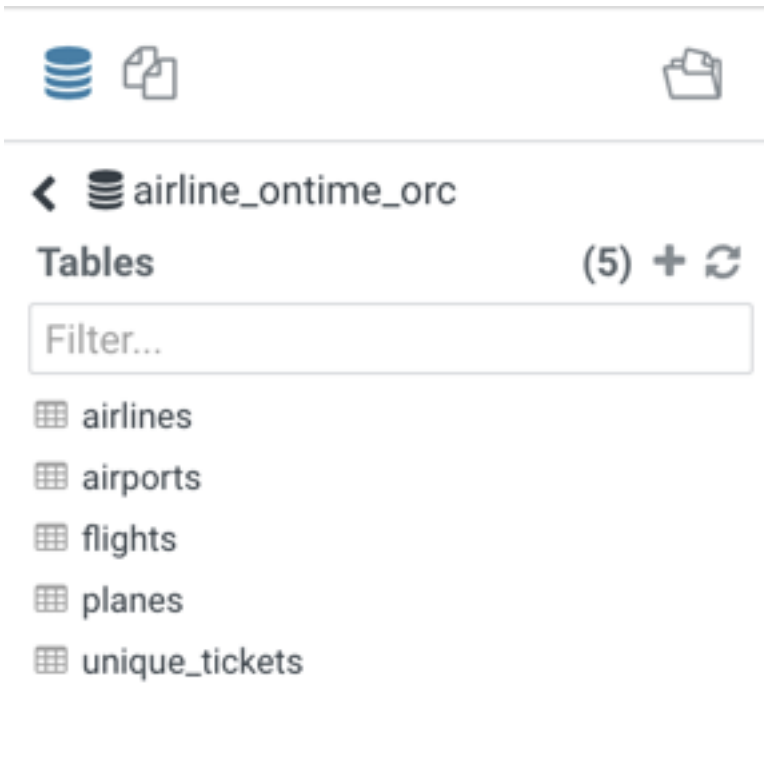
☐ Custom char delimiters

Partitions [+ Add partition](#)

8. Change all the “bigint” column types to “int” under FIELDS and click Submit.
The CREATE TABLE query is triggered.



Hue displays the logs and opens the **Table Browser** from which you can view the newly created table when the operation completes successfully.



9. Run the following query to find passengers with international connecting flights:

```
-- Query to explore passenger manifest data: do we have international c
onnecting flights?
SELECT
*
FROM
  `airline_ontime_orc`.unique_tickets a,
  `airline_ontime_orc`.flights o,
  `airline_ontime_orc`.flights d,
  `airline_ontime_orc`.airports oa,
  `airline_ontime_orc`.airports da
WHERE
  a.leg1flightnum = o.flightnum
  AND a.leg1uniquecarrier = o.uniquecarrier
  AND a.leg1origin = o.origin
  AND a.leg1dest = o.dest
  AND a.leg1month = o.month
```

```

AND a.leg1dayofmonth = o.dayofmonth
AND a.leg1dayofweek = o.`dayofweek`
AND a.leg2flightnum = d.flightnum
AND a.leg2uniquecarrier = d.uniquecarrier
AND a.leg2origin = d.origin
AND a.leg2dest = d.dest
AND a.leg2month = d.month
AND a.leg2dayofmonth = d.dayofmonth
AND a.leg2dayofweek = d.`dayofweek`
AND d.origin = oa.iata
AND d.dest = da.iata
AND oa.country <> da.country
;

```

Creating data sets as per the business criteria

Run queries to create data sets that help answer burning business questions.

Procedure

1. Run the following query to get the number of passengers on an airline with long layovers:

```

-- Number of passengers on the airline that have long, planned layovers
for an international flight
SELECT
    a.leg1uniquecarrier as carrier,
    count(a.leg1uniquecarrier) as passengers
FROM
    `airline_ontime_orc`.unique_tickets a
where
    a.leg2deptime - a.leg1arrtime > 90
group by
    a.leg1uniquecarrier
;

```

2. Run the following query to get the number of passengers on an airline with delayed international leg:

```

-- Number of passengers on airlines that have elongated layovers for an
international flight caused by delayed connection
SELECT
    a.leg1uniquecarrier as carrier,
    count(a.leg1uniquecarrier) as passengers
FROM
    `airline_ontime_orc`.unique_tickets a,
    `airline_ontime_orc`.flights o,
    `airline_ontime_orc`.flights d
where
    a.leg1flightnum = o.flightnum
    AND a.leg1uniquecarrier = o.uniquecarrier
    AND a.leg1origin = o.origin
    AND a.leg1dest = o.dest
    AND a.leg1month = o.month
    AND a.leg1dayofmonth = o.dayofmonth
    AND a.leg1dayofweek = o.`dayofweek`
    AND a.leg2flightnum = d.flightnum
    AND a.leg2uniquecarrier = d.uniquecarrier
    AND a.leg2origin = d.origin
    AND a.leg2dest = d.dest
    AND a.leg2month = d.month
    AND a.leg2dayofmonth = d.dayofmonth
    AND a.leg2dayofweek = d.`dayofweek`
    AND o.depdelay > 60
group by

```



```

    a.legluniquecarrier
;

```

3. Run the following query to get the number of passengers on an airline with missed international connections:

```

-- Number of passengers on airlines that have elongated layovers for an
international flight caused by missed connection
SELECT
    a.legluniquecarrier as carrier,
    count(a.legluniquecarrier) as passengers
--    o.arrdelay as delay
FROM
    `airline_ontime_orc`.unique_tickets a,
    `airline_ontime_orc`.flights o,
    `airline_ontime_orc`.flights d
where
    a.leg1flightnum = o.flightnum
    AND a.legluniquecarrier = o.uniquecarrier
    AND a.leglorigin = o.origin
    AND a.legldest = o.dest
    AND a.leg1month = o.month
    AND a.leg1dayofmonth = o.dayofmonth
    AND a.leg1dayofweek = o.`dayofweek`
    AND a.leg2flightnum = d.flightnum
    AND a.leg2uniquecarrier = d.uniquecarrier
    AND a.leg2origin = d.origin
    AND a.leg2dest = d.dest
    AND a.leg2month = d.month
    AND a.leg2dayofmonth = d.dayofmonth
    AND a.leg2dayofweek = d.`dayofweek`
    AND d.deptime-o.arrtime < o.arrdelay-45
group by
    a.legluniquecarrier
;

```

Generating and sharing dashboards

Use Cloudera Data Visualization to create dashboards from queries and share them with your stakeholders.

Procedure

1. Log in to the CDW service.
2. Go to Virtual Warehouses, click the options menu and select Open Viz.
3. Go to the DATA interface and click NEW DATASET.
4. On the **New Dataset** pop-up, specify a Dataset title.

Select From SQL from the Dataset Source drop-down list.

Copy and paste the following query in the Enter SQL below field and click CREATE:

```

-- Number of passengers on the airline that have long, planned layovers
for an international flight
SELECT
    a.legluniquecarrier as carrier,
    count(a.legluniquecarrier) as passengers
FROM
    `airline_ontime_orc`.unique_tickets a
where
    a.leg2deptime - a.leglarrrtime > 90
group by
    a.legluniquecarrier

```

;

New Dataset

Create a dataset from data on this connection. You need to create a dataset before you can create dashboards or apps.

Dataset title *

Duty-free-analysis-delays

Dataset Source

From SQL

Enter SQL below

```
SELECT
  a.legluniquecarrier as carrier,
  count(a.legluniquecarrier) as passengers
FROM
  airlinedata.unique_tickets_orc a
where
  a.leg2deptime - a.leglarrtime > 90
group by
  a.legluniquecarrier
;
```

☒ Autocomplete on

CANCEL CREATE

A new data set is created based on which you can create the visualization.

5. Go to the HOME interface and click NEW DASHBOARD.

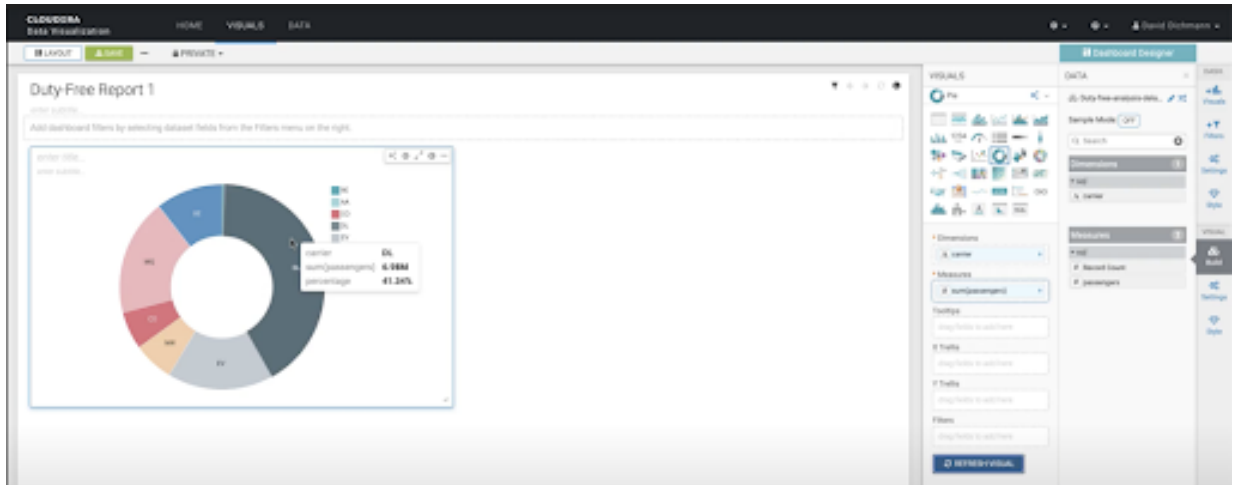
The **Dashboard Designer** opens on an untitled dashboard.

6. Enter a name for your dashboard, select the dataset that you just created from the ADD VISUALS menu, and click NEW VISUAL.

The data is rendered in a tabular format from the SQL query.

7. Select a visual type from the VISUALS menu, such as a pie chart. Select the appropriate Dimensions and Measures from the DATA menu, and click REFRESH VISUAL.

The pie chart shows the airlines having the most passengers with long layovers on international flights.



8. Click SAVE.
9. To share the dashboard:
 - a) Open your dashboard from the VISUAL interface.
 - b) Click the more option and select Get URL.
A pop-up containing the URL is displayed.
 - c) Copy the URL by right-clicking it and share it with your stakeholders.

Troubleshooting

This section helps you to narrow down a cause for an issue and provides a possible direction to resolve it.

Unable to see the S3 File Browser on the Hue web interface

Condition

Unable to see the S3 File Browser on the Hue web interface.

Cause

You may not have configured your Cloudera Data Warehouse (CDW) environment to have an S3 bucket that you can use to save data as part of the import process.

Solution

Add an S3 bucket in the CDW environment that you have access to. For more information, see the prerequisites section.

External File option is unavailable

Condition

External File option is unavailable on the Hue Importer dropdown list.

Cause

You may not have configured Hue to access the S3 bucket that you added in the environment.

Solution

Add the AWS access key and secret access key for your S3 bucket in the hue-safety-valve property. For more information, see the prerequisites section.

Error while accessing a directory in S3**Condition**

Error stating "cannot access directory" while importing a file in Hue.

Cause

You may not have specified the right credentials in the hue-safety-valve property correctly.

Solution

Ensure that the account that owns the S3 bucket that you have added to your CDW environment has been added to the hue-safety-valve correctly.

Incorrect region error while uploading a file to S3**Condition**

"Incorrect region" error while importing and uploading a file to S3 from your computer using Hue.

Cause

You may not have specified the AWS region where your S3 bucket is stored in the hue-safety-valve.

Solution

Ensure that the AWS region that you have specified in the hue-safety-valve is the same as the AWS region where your S3 bucket that you have added to your environment is managed.