CDP Patterns Public Cloud..

Business Intelligence at Scale

Date published: 2022-02-04
Date modified: 2022-02-04

https://docs.cloudera.com/

Legal Notice

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

Introduction to Business Intelligence at Scale

The Business Intelligence (BI) at Scale pattern helps you to build and implement a sustainable BI solution using the Cloudera Data Platform (CDP) and various data services using which you can ingest, stream, explore, optimize, and analyze real-time and enterprise data, and then visualize the results to drive business insights.

Note: The Business Intelligence at Scale pattern is in Technical Preview and is not recommended for production use. Cloudera recommends that you try this pattern in your development or test environments.

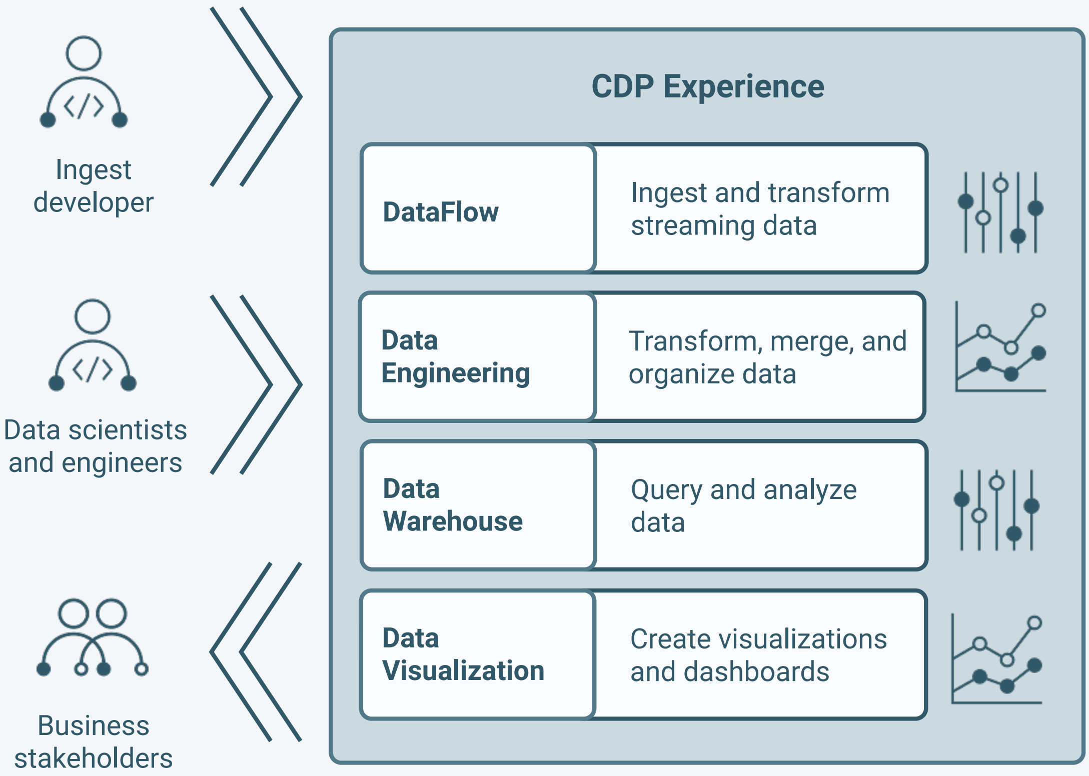
The Business Intelligence at Scale pattern provides prescriptive steps to help you use the Cloudera data pipeline to stream data from a Streams Messaging cluster in CDP Public Cloud to AWS buckets. You can use Cloudera Data Engineering to organize and convert your data, and make the data available to Cloudera Data Warehouse for queries and analysis. Finally, you can use Cloudera Data Visualization tools to create consumable visualizations and dashboards.

Business Intelligence at Scale

Ingest | Merge | Analyze

Use Cloudera's data pipeline to quickly ingest and transform real-time data streams, transform and organize this data and merge with historical data, and then analyze and visualize query results.

Go from data to dashboard in less than a day.



Ingest developer

Data scientists and engineers

Business stakeholders

CDP Experience

DataFlow	Ingest and transform streaming data
Data Engineering	Transform, merge, and organize data
Data Warehouse	Query and analyze data
Data Visualization	Create visualizations and dashboards

Understanding the use case

Learn how the Business Intelligence at Scale pattern empowers Ingest developers, Data engineers, Data Analysts, and Business stakeholders by using various Cloudera data services such as Streams Messaging Manager, DataFlow, Data Engineering, Data Warehouse, and Data Visualisation.

Business Intelligence at Scale is a common design pattern that you can implement in situations where business groups need insight from both enterprise and streaming data to make timely decisions. A traditional analytics solution would struggle with ingesting streaming data due to the technical challenges or experience needed in setting up a streaming ingest. This pattern guides Ingest Engineers to easily ingest streaming data, Data Engineers to model and curate data, and Business Analysts or Data Scientists to analyze, visualize, and share insights through an integrated self-service experience.

The objective is to allow business stake holders to quickly analyze all the data by using the integrated data lifecycle tools built into CDP. This service offers a smooth user experience for line of business decision makers to get answers to common questions through ready to use reports and dashboards, and an enhanced exploratory experience to line of business data analysts and power users to scrutinize data by building pipelines and querying the data sets.

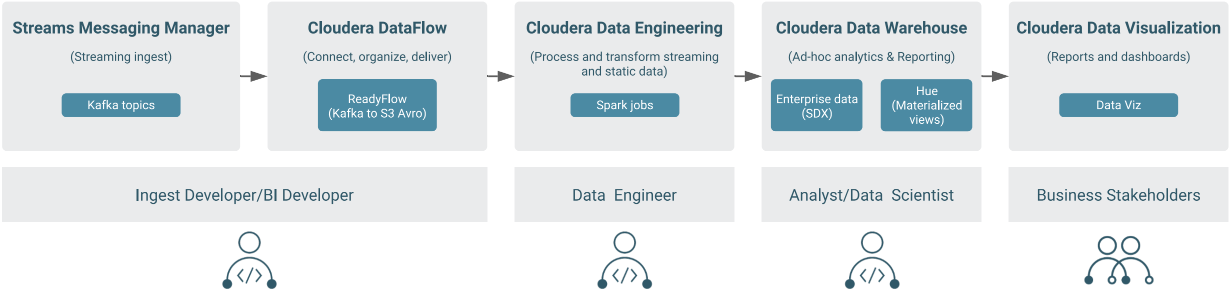
Key problem this pattern solves

Customers can be more agile and better informed when making business decisions or creating ad-hoc reports based on new data as it arrives and becomes available.

Cloudera data lifecycle components

Business Intelligence at Scale pattern uses:

- Cloudera's Streams Messaging Manager (SMM) to ingest external (non-enterprise) data by creating Kafka topics
- Cloudera DataFlow Experience to convert and upload data to Amazon S3 (or any Cloud object store) in Avro format
- Cloudera Data Engineering service to transform data into ORC or Parquet format using Spark jobs
- Cloudera Data Warehouse service to create tables and materialized views from the datasets, run queries, and ad-hoc exploratory analysis
- Cloudera Data Visualization to create reports and dashboards


Streams Messaging Manager
(Streaming ingest)
Kafka topics
Cloudera DataFlow
(Connect, organize, deliver)
ReadyFlow (Kafka to S3 Avro)
Cloudera Data Engineering
(Process and transform streaming and static data)
Spark jobs
Cloudera Data Warehouse
(Ad-hoc analytics & Reporting)
Enterprise data (SDX)
Hue (Materialized views)
Cloudera Data Visualization
(Reports and dashboards)
Data Viz
Ingest Developer/BI Developer
Data Engineer
Analyst/Data Scientist
Business Stakeholders

Using these Cloudera services:

- Ingest developers can quickly develop data flows for real-time data and merge incoming data with existing data sets
- Data engineers can transform, organize, and enrich data
- Data Analysts and Data Scientists can query the data sets using Hue
- Business stakeholders can derive insights using visualizations and dashboards

Business Intelligence at Scale use case

A real-life business narrative enables you to follow and understand the Business Intelligence at Scale 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: Analyzing the impact of dynamic weather conditions on static inventory management for perishable goods

Broger Foods is a mega grocery retail chain, with multiple outlets across the country, selling 100+ perishable goods. Each year, Broger Foods suffers losses ranging between 30 and 45 million dollars due to inventory shrinkage. Damages and spoilage of perishable items in warehouses has been identified as a key contributor to the shrinkage. The challenge Broger faces is the relatively static inventory management system for perishable goods. Stocks of fresh produce and food rot or expire in the warehouses due to intermittent demand and over stocking. Understocking is not a viable option due to the competitive nature of the industry, leading to customers turning to Hole Foods, a rival retail chain gaining popularity.

The CFO of Broger Foods, Roger Walters has tasked their VP of Analytics, Richard, to look into this issue and help solve Broger's supply chain management problem.

Key business questions

- How can Broger Foods optimize its inventory management system to reduce revenue loss and prevent food wastage while not loosing customers to its competitors?
- How can the BI team at Broger Foods deliver insights at scale?

Solving the problem

Richard formulates two hypotheses to analyze the sales trends. His first hypothesis is based on population density of a location and the second one is based on the weather conditions. Upon delving deeper using ad-hoc analytics, Richard figured out that there is a strong correlation between demand for perishable items in each location and weather conditions in that location. To continuously improve Broger’s inventory management process, Richard plans to use the location-wise sales information present in Broger's CDP Public Cloud environment along with real-time weather information from a weather data provider.

To deliver the insights at scale, Richard calls for a meeting with Syd (Streaming App Developer), David (Data Engineer), and Nick (Data Analyst) to explain his model and what he needs to implement this new inventory forecasting model.

To build a streaming pipeline, Richard asks Syd to capture live weather data from OpenWeather and ingest it into CDP. Syd sets up a data flow that ingests data into CDP in real time.

He then asks David to build an unified data pipeline that aggregates data from each stream into two separate tables:

- Current weather data. This table captures temperature, humidity, precipitation, and pressure.
- Historical data for 5 previous days.

David’s pipeline will be triggered to process and aggregate the freshly ingested data every hour and update the Optimized Row Columnar (ORC) or Parquet tables in CDW every hour.

Now, Nick is tasked to write queries that averages out the weather data from both the weather tables and combine it with the location-based sales data for additional analysis. He also has to build a dashboard that showcases the variance between item sales and temperature at a given store location on a daily basis. The dashboard also shows a 7 day forecast of sales. The sales forecast is done on Monday and Friday every week.

Key personas in this use case

- Ingest Developer – Develops and explores the data ingest streams

- Data Engineer – Prepares the data pipelines, optimizes and aggregrates data for use by Analysts
- Data Analyst – Analyzes data using queries and by building reports and dashboards
- Business stakeholders – Consume analyzed data and visualizations

Success criteria

Rapidly develop and build data streams to deliver business insights at scale by enabling self-service DataFlow development, Data Engineering, Data Warehousing, and Business Intelligence reporting and dashboarding capabilities.

Setting up your CDP pattern infrastructure

Before starting, ensure that your central or departmental IT has onboarded to CDP Public Cloud and registered an environment with Cloudera. Business Intelligence at Scale pattern leverages Streams Messaging Manager (SMM), Cloudera DataFlow (CDF), Cloudera Data Engineering (CDE), and Cloudera Data Warehouse (CDW) services. You must set up these services as part of your infrastructure.

Most steps required to set up your CDP pattern infrastructure should be performed by a CDP administrator and require the EnvironmentAdmin CDP role. Since the goal of this pattern is to enable self-service DataFlow development, Data Engineering, and Data Warehousing, make sure that the IT team has created the necessary users and groups, and have granted the required permissions to these users and groups.

When your IT team creates and registers an environment with CDP, it is assumed that they create a medium-duty Data Lake to import and manage streaming data in the Streams Messaging Data Hub cluster, as well as when running other CDP services.

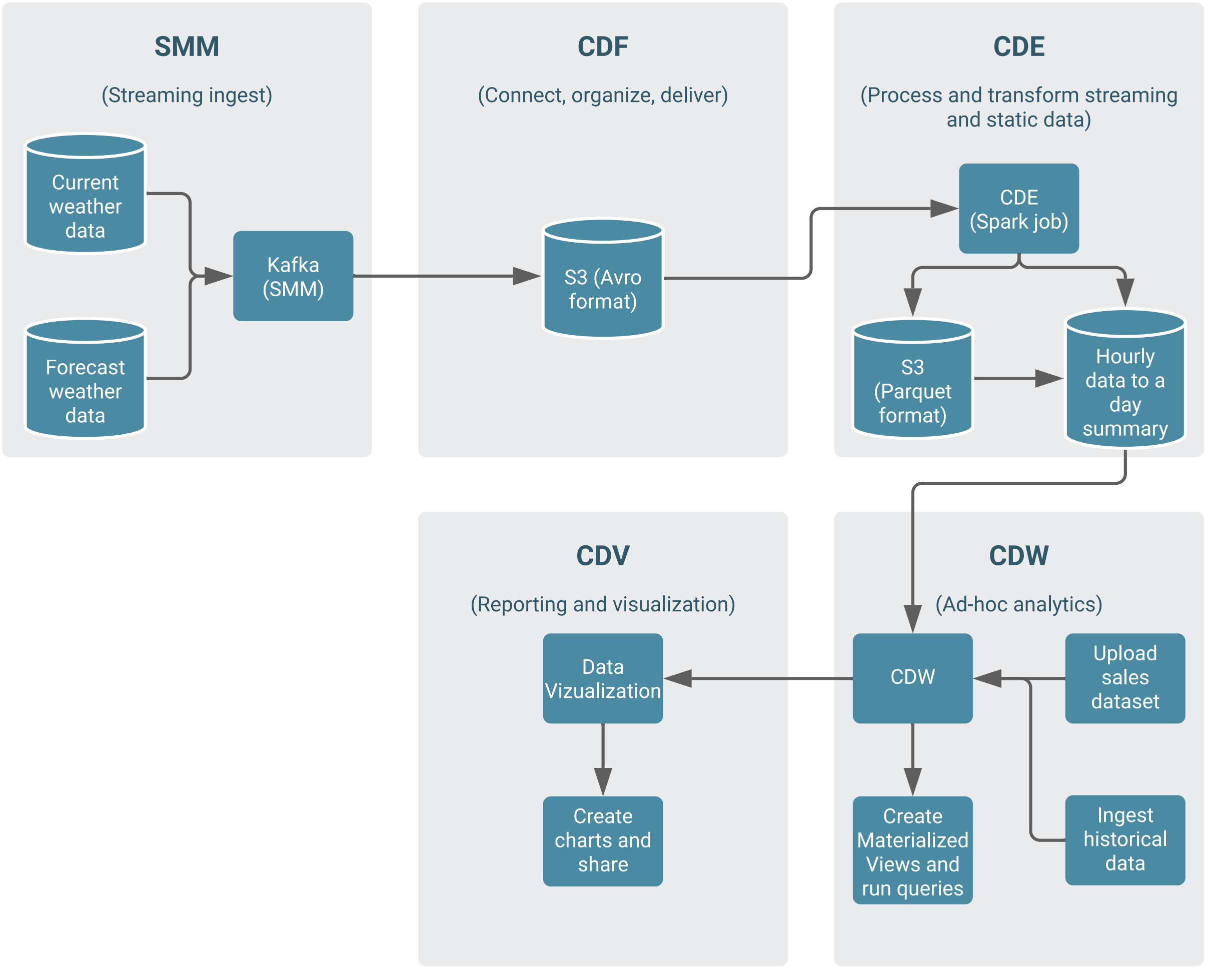
Before you begin self-service development and infrastructure setup:

- Ensure that you have an available CDP environment
- Ensure that you have CDP login credentials
- Ensure that you have a running Data Lake
- Ensure that your CDP user is synchronized to the CDP Public Cloud environment
- Review the AWS environments requirements checklist

For test and development environments, you can set up small to medium sized clusters, and scale up for production use.

After you register your environment with CDP, you must also activate entitlements for using Unified Analytics and Data Visualization in CDW by contacting your Cloudera Account Representative.

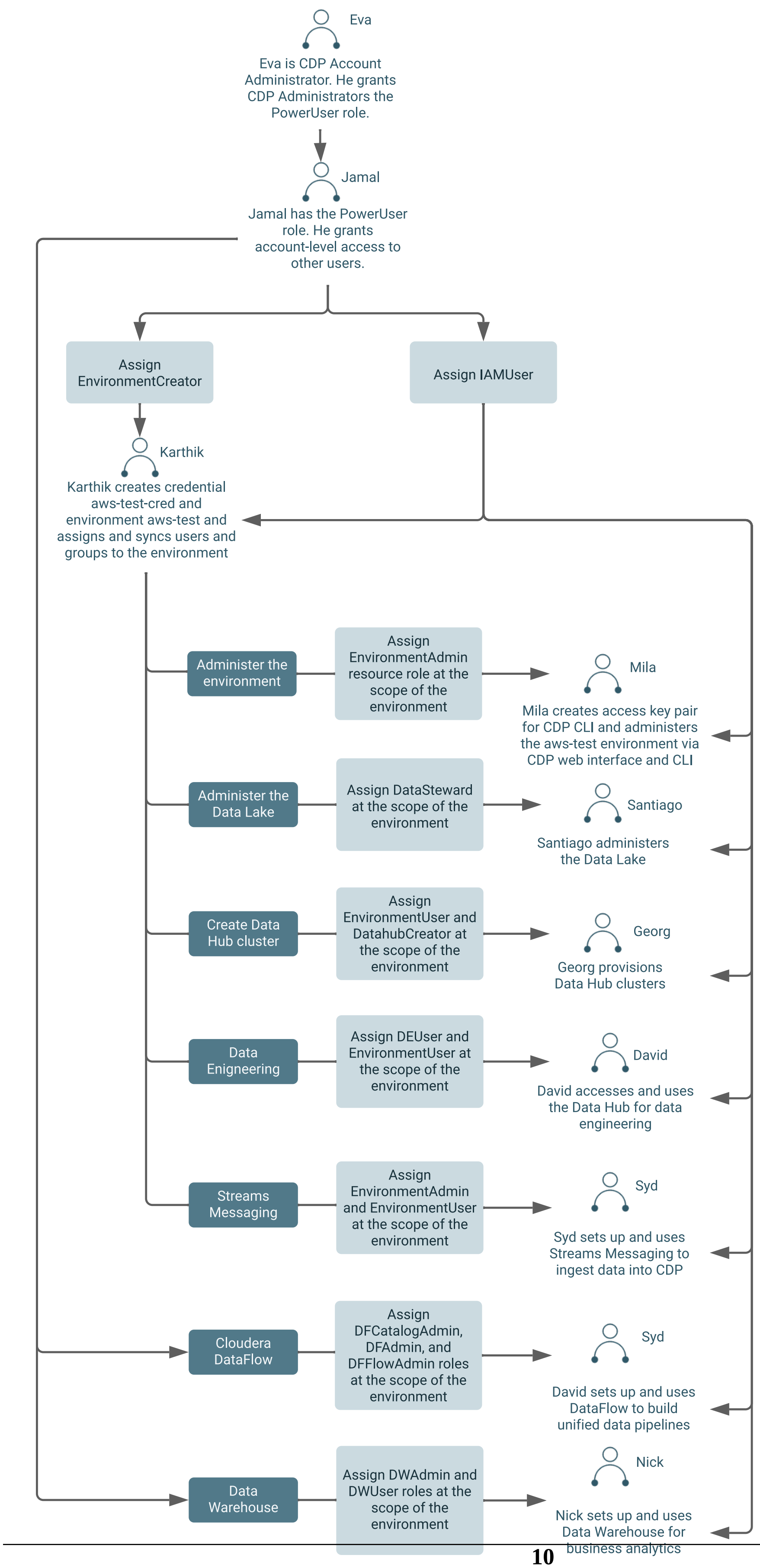
The following diagram shows all the Data Services used to implement the Business Intelligence at Scale pattern:



Giving access to users and groups

To access and use the CDP Data Services, you must have the required roles and permissions. Learn about the various roles that users need to perform specific tasks within the Business Intelligence at Scale pattern.

The following graphic shows the roles required to set up and access various CDP services with this pattern and who must grant those roles:


Eva
Eva is CDP Account Administrator. He grants CDP Administrators the PowerUser role.
Jamal
Jamal has the PowerUser role. He grants account-level access to other users.
Assign EnvironmentCreator
Assign IAMUser
Karthik
Karthik creates credential aws-test-cred and environment aws-test and assigns and syncs users and groups to the environment
Administer the environment
Assign EnvironmentAdmin resource role at the scope of the environment
Mila
Mila creates access key pair for CDP CLI and administers the aws-test environment via CDP web interface and CLI
Administer the Data Lake
Assign DataSteward at the scope of the environment
Santiago
Santiago administers the Data Lake
Create Data Hub cluster
Assign EnvironmentUser and DatahubCreator at the scope of the environment
Georg
Georg provisions Data Hub clusters
Data Enigneering
Assign DEUser and EnvironmentUser at the scope of the environment
David
David accesses and uses the Data Hub for data engineering
Streams Messaging
Assign EnvironmentAdmin and EnvironmentUser at the scope of the environment
Syd
Syd sets up and uses Streams Messaging to ingest data into CDP
Cloudera DataFlow
Assign DFCatalogAdmin, DFAdmin, and DFFlowAdmin roles at the scope of the environment
Syd
David sets up and uses DataFlow to build unified data pipelines
Data Warehouse
Assign DWAdmin and DWUser roles at the scope of the environment
Nick
Nick sets up and uses Data Warehouse for business analytics

Eva is the CDP Account Administrator. Eva is typically the head of IT and purchases the CDP license for her organization.

Eva registers their cloud environment (such as AWS) with CDP and attaches an identity federation to add users to the CDP environment.

Jamal is a Power User. He has full admin access to the CDP tenant. He can assign account roles to other users who need access to CDP.

He assigns IAMUser to all onboarded users so that they can generate access keys, upload SSH keys, and assign resouces to other users.

Jamal assigns EnvironmentCreator to Karthik. This allows Karthik to create a provisioning credential aws-test-cred in CDP and register a cloud provider environment aws-test-env.

Note that since Karthik created the credential and the environment, he automatically gets the EnvironmentAdmin and Owner role over both resources.

Karthik assigns the EnvironmentAdmin resource role to Mila over his aws-test environment. This allows Mila to administer this specific environment.

Karthik assigns the DataSteward resource role to Santiago over his aws-test environment. This allows Santiago to administer the Data Lake running in this specific environment.

Karthik assigns the EnvironmentUser and DatahubCreator resource role to Georg over the aws-test environment.

Georg provisions a Data Hub and automatically gets Owner role at the scope of that Data Hub.

Karthik assigns the EnvironmentUser and DEUser resource roles to David over the aws-test environment so that David can access Data Hubs provisioned by Georg within the environment and import and run Spark jobs.

In addition, Jamal assigns the IAMUser account role to David so that David can access the S3 buckets.

Syd provisions a Streams Messaging cluster, creates Kafka topics, and adds schemas to the Schema Regisrty.

Syd also sets up the CDF service, enables the CDF environment, and builds and deploys flows.

In addition, Jamal assigns IAMUser account to Syd so that Syd can access the S3 buckets.

Nick activates the aws-test environment in CDW , creates Virtual Warehouses, runs queries, and generate reports using Data Visualization.

In addition, Jamal assigns IAMUser account to Nick so that Nick can access the S3 buckets.

Tip: If more than one user needs a certain role or permission to work with a CDP service, Cloudera recommends that you add those users into a group and grant a role or permission to that group.

Related Information

Resource roles

Giving access to Streams Messaging users

An "EnvironmentAdmin" or a "DataSteward" must grant the "EnvironmentUser" role to the user who is tasked to set up the Streams Messaging cluster.

Before you begin

Important: You must have an EnvironmentAdmin or a DataSteward role to perform this task.

About this task

The cluster you have created using the Streams Messaging cluster definition is kerberized and secured with SSL. Users can access cluster UIs and endpoints through a secure gateway powered by Apache Knox. Before you can begin working with Kafka, Schema Registry, Streams Messaging Manager, and Streams Replication Manager, you must give users access to the Streams Messaging cluster components.

Procedure

1. Log in to the CDP web interface.
2. Go to Management Console Environments .
3. Go to Actions Manage Access to display the **Environment Access** page.
4. Find the user to whom you want to grant the EnvironmentUser role, and click Update Roles.
5. Assign the EnvironmentUser role to the users to grant access to the CDP environment and the Streams Messaging cluster.
6. Go to the **Environments** page.
7. Click Actions Synchronize Users to FreeIPA .

   Depending on how many users have access to the environment, this synchronization process can take a few seconds to a few minutes.

Giving access to DataFlow users

Cloudera DataFlow (CDF) restricts who can import flow definitions and deploy them. A "PowerUser" must grant "DFAdmin", "DFFlowAdmin", and "DFCatalogAdmin" roles to users so that they can enable CDF for an environment, import, add, and deploy flow definitions in CDF.

About this task

- The DFAdmin role is required to enable CDF for an environment.
- The DFFlowAdmin role is required deploy flow definitions in CDF.
- The DFCatalogAdmin role is required to import flow definitions to the CDF Catalog.

Before you begin

Important: You must have a PowerUser role to perform this task.

Cloudera recommends that you create a dedicated machine user as a CDP Workload User for CDF. Make sure to set the Workload Password. You need the workload username and password for deploying NiFi flows.

Note: When you create a machine user, then a prefix, "srv", is automatically added to the machine user name. For example, if you create a machine user as "machine-user-1", then the Workload User Name reads "srv_machine-user-1".

You must also assign the EnvironmentUser role to this machine user within your environment.

Procedure

1. Log in to the CDP web interface.
2. To give a user the permission to enable CDF for an environment:
   a) Go to Management Console Environments .

      The **Environment List** page is displayed.
   b) Select the environment in which you want a user to enable CDF.
   c) Go to Actions Manage Access to display the Environment Access page.
   d) Find the user to whom you want to grant the DFAdmin role, and click Update Roles.
   e) Select DFAdmin and click Update Roles.
3. To give a user or group the permission to deploy flow definitions.
   a) Go to Management Console Environments .

      The **Environment List** page is displayed.
   b) Select the environment to which you want a user or group to deploy flow definitions.
   c) Go to Actions Manage Access .

      The **Environment Access** page is displayed.
   d) Find the user or group and click Update Roles.
   e) Select DFFlowAdmin and click Update Roles.
4. To give a user the permission to import flow definitions:
   a) Go to Management Console User Management .
   b) Enter the name of the user or group you wish to authorize in the Search field.
   c) Select the user or group from the list that displays.
   d) Click Roles, then Update Roles.
   e) From Update Roles, select DFCatalogAdmin and click Update.
5. Go to the **Environments** page.
6. Click Actions Synchronize Users to FreeIPA .

   Depending on how many users have access to the environment, this synchronization process can take a few seconds to a few minutes.

Giving access to Data Engineering users

Cloudera Data Engineering (CDE) restricts access to importing and running Spark jobs. An "EnvironmentAdmin" must grant both "DEUser" and "EnvironmentUser" roles to users so that they can run Spark jobs in CDE and access the Data Hub clusters for data engineering.

Before you begin

Important: You must have an EnvironmentAdmin role to perform this task.

Procedure

1. Log in to the CDP web interface.
2. Go to Management Console Environments .
3. Select the environment in which you want a user or group to import and run Spark jobs.

4. Go to Actions Manage Access .

   The **Environment Access** page is displayed.
5. Find the user or group using Search and click Update Roles.
6. Select the DEUser and EnvironmentUser options and click Update Roles.
7. Go to the **Environments** page.
8. Click Actions Synchronize Users to FreeIPA .

   Depending on how many users have access to the environment, this synchronization process can take a few seconds or a few minutes.

Giving access to Data Warehouse users

A "PowerUser" must grant "DWAdmin" and "DWUser" roles to users and groups so that they can activate the Cloudera Data Warehouse (CDW) service, create Database Catalogs and Virtual Warehouses, and manage Hue and other applications within CDW.

Before you begin

Important: You must have a PowerUser role to perform this task.

Procedure

1. Log in to the CDP web interface.
2. Go to Management Console Environments and select your CDP environment.
3. Go to Actions Manage Access .

   The **Environment Access** page is displayed.
4. Find the user or group using Search and click Update Roles.
5. Select the DWAdmin and DWUser options and click Update Roles.

   Note: When you grant access to groups, you must grant the DWAdmin resource role to the group instead of the DWUser role.
6. Go to the **Environments** page.
7. Click Actions Synchronize Users to FreeIPA .

   Depending on how many users have access to the environment, this synchronization process can take a few seconds or a few minutes.

Creating and assiging Ranger policies

Apart from assigning resource roles to users, you need to set up Ranger policies to authorize users (service, machine, or workload users) to perform specific operations and rights to certain resources. Learn which Ranger policies you need for the Business Intelligence at Scale pattern and how to create them.

Before you begin

Important: You must have the EnvironmentUser role to perform this task.

Ensure that you have a list of machine users and workload users you want to add to the specific Ranger policies.

About this task

The following table lists the CDP components for which you need to set up resource-based Ranger policies:

CDP component	Resource-based Ranger policy	Purpose
Kafka (Streams Messaging)	Kafka	Allow the machine user to publish, configure, and consume Kafka topics.
	Kafka	Allow the machine user to consume the Consumer Group IDs.
Schema Registry (Streams Messaging)	Schema-Registry	Allow the machine user to read schema groups from the Schema Registry.
Hue (Cloudera Data Warehouse)	Hadoop SQL	Allow the workload user to perform all database, table, and column operations such as select, update, alter, create, drop, insert, read, and write.

Note: A set of default policies are automatically created for your deployment. You must create a new policy within your main deployment policy.

Procedure

1. Log in to the CDP web interface.
2. Go to Management Console Environments and click on Ranger.

   The Ranger **Service Manager** page is displayed.
3. Create a Kafka policy to allow the machine user to publish, configure, and consume Kafka topics.
   a) Click on your main deployment (SMM) policy under **KAFKA**.
   b) Click Add New Policy on the **List of Policies** page.
   c) Specify a name for your policy in the Policy Name field.
   For this pattern, you can specify bias-smm-kafka-ingest.
   d) Select topic from the drop-down menu.
   e) Enter the names of the Kafka topics to which you want to authorize user access.
   For this pattern, specify weather, weather_forecast or weather*.
   f) Go to the **Allow Conditions** section and enter the machine username under the Select User column.
   g) Click + under the Permissions column and select Publish, Consume, and Configure options and click ✓.
   h) Click Add at the bottom of the page.
4. Create a Kafka policy to allow the machine user to consume, describe, and delete the Consumer Group IDs.
   a) Click on your main deployment (SMM) policy under **KAFKA**.
   b) Click Add New Policy on the **List of Policies** page.
   c) Specify a name for your policy in the Policy Name field.
   For this pattern, you can specify bias-smm-kafka-consumer.
   d) Select consumergroup from the drop-down menu.
   e) Enter the names of the Kafka Consumer Group ID to which you want to authorize user access.
   For this pattern, specify WeatherConsumer, ForecastConsumer.
   f) Go to the **Allow Conditions** section and enter the machine username under the Select User column.
   g) Click + under the Permissions column and select the Consume option and click ✓.
   h) Click Add at the bottom of the page.

5. Create a Schema-Registry policy to allow the machine user to read the schema groups from the Schema Registry.

   a) Click on your main deployment (SMM) policy under **SCHEMA-REGISTRY**.

   b) Click Add New Policy on the **List of Policies** page.

   c) Specify a name for your policy in the Policy Name field.
   For this pattern, you can specify bias-smm-sr-schema-group.

   d) Select schema-group from the drop-down menu and enter weather next to it.

   e) Enter the names of the Schema Registry schemas to which you want to authorize user access in the Schema Name field.
   For this pattern, specify WeatherCurrent, WeatherForecast.

   f) Select schema-branch from the drop-down menu and enter *.

   g) Select schema-version from the drop-down menu and enter *.

   h) Go to the **Allow Conditions** section and enter the machine username under the Select User column.

   i) Click + under the Permissions column and select the Read option and click ✓.

   j) Click Add at the bottom of the page.

6. Create a Hadoop SQL policy to allow the workload users to perform operations on databases, tables, and columns.

   a) Click Hadoop SQL under **HADOOP SQL**.

   b) Click Add New Policy on the **List of Policies** page.

   c) Specify a name for your policy in the Policy Name field.
   For this pattern, you can specify bias-all-database-table-column.

   d) Select database from the drop-down menu and enter *.

   e) Select table from the drop-down menu and enter *.

   f) Select column from the drop-down menu and enter *.

   g) Go to the **Allow Conditions** section and enter the workload username under the Select User column.

   > **Tip:** To provide access to all users, you can specify {USER}.

   h) Click + under the Permissions column and select the select, update, Create, Drop, Alter, Read, and Write option and click ✓.

   i) Click Add at the bottom of the page.

Setting up the Streams Messaging cluster

Streams Messaging provides advanced messaging and real-time processing on streaming data using Apache Kafka, centralized schema management using Schema Registry, as well as management and monitoring capabilities powered by Streams Messaging Manager. Learn how to set up a Streams Messaging cluster in CDP Public Cloud for use with your Business Intelligence at Scale pattern.

For more information about the Streams Messaging cluster, see *Streams Messaging clusters* and *How To: Streams Messaging* in the latest CDF for Data Hub library.

Related Information

Streams Messaging clusters

CDF for Data Hub

Creating a Streams Messaging cluster

Learn how to create a Streams Messaging cluster for the Business Intelligence at Scale pattern. A Data Hub cluster with the Streams Messaging cluster definition hosts the Kafka topics into which you capture real-time streaming data. For this example, you are importing the weather data.

Before you begin

⚠ **Important:** You must have an EnvironmentAdmin role to perform this task.

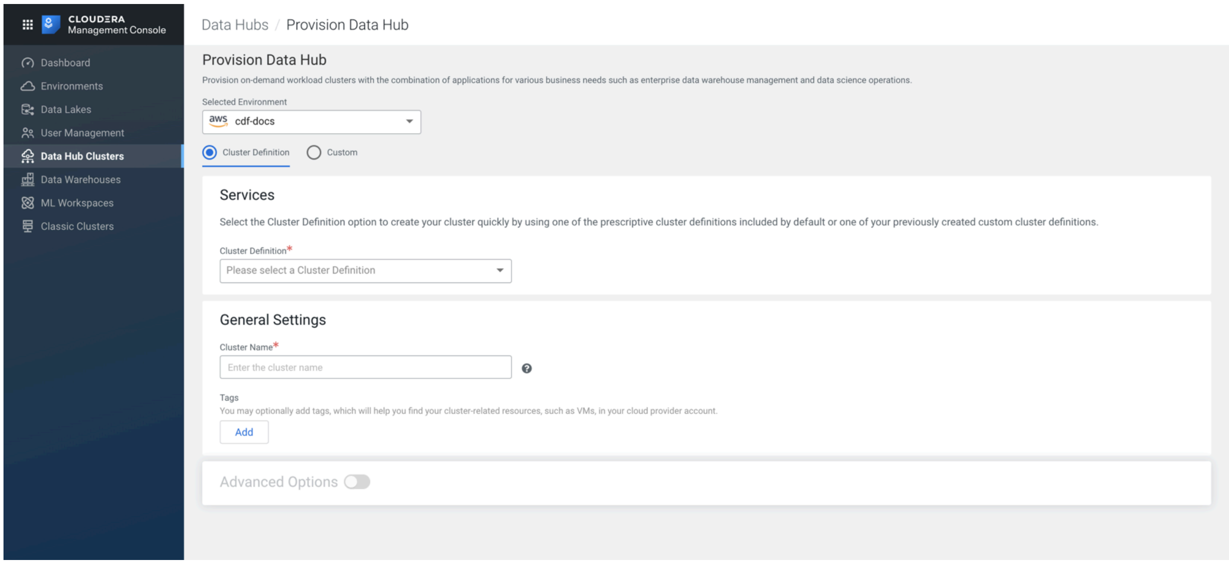
About this task

For the Business Intelligence at Scale pattern proof of concept, use the Streams Messaging Light Duty template for AWS. When you are ready to move your data pipeline to production, you can use either a Light or Heavy duty template, depending on your use case. For this pattern, use the default cluster definition.

Procedure

1. Log into the CDP web interface.
2. Navigate to Management Console Environments , and select the environment where you would like to create a cluster.
3. Click Create Data Hub.

   The following page is displayed:



Provision Data Hub

Provision on-demand workload clusters with the combination of applications for various business needs such as enterprise data warehouse management and data science operations.

Selected Environment: aws cdf-docs

● Cluster Definition ○ Custom

Services

Select the Cluster Definition option to create your cluster quickly by using one of the prescriptive cluster definitions included by default or one of your previously created custom cluster definitions.

Cluster Definition*: Please select a Cluster Definition

General Settings

Cluster Name*: Enter the cluster name

Tags

You may optionally add tags, which will help you find your cluster-related resources, such as VMs, in your cloud provider account.

Add

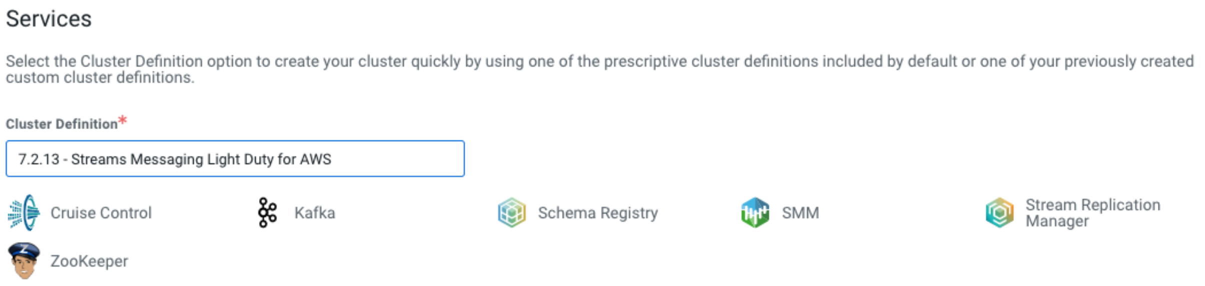
Advanced Options

4. Select the Cluster Definition option.
5. Select the appropriate Streams Messaging cluster definition from the Cluster Definition drop-down menu under the Services section depending on your operational objectives.

   There are two definition options available for each cloud provider:

   - Streams Messaging Heavy Duty for AWS
   - Streams Messaging Light Duty for AWS

   The list of services is automatically shown below the selected cluster definition name as shown in the following image:



Services

Select the Cluster Definition option to create your cluster quickly by using one of the prescriptive cluster definitions included by default or one of your previously created custom cluster definitions.

Cluster Definition*: 7.2.13 - Streams Messaging Light Duty for AWS

6. Specify a name for your cluster and add any tags you might need.
7. Click Provision Cluster.

Results

You are redirected to the Data Hub cluster dashboard, and a new tile representing your cluster appears at the top of the page.

Setting up the Cloudera DataFlow service

In the Business Intelligence at Scale pattern, Cloudera DataFlow (CDF) provides basic flow development and streaming data flow organization and management capabilities.

To set up a CDF cluster, a user with the PowerUser role must first grant you the DFAdmin role. You can then enable the CDF service in your CDP environment.

⚠ **Attention:** Enabling CDF can take upto one hour.

For the Business Intelligence at Scale pattern, you will be using the "Kafka to Avro S3" ReadyFlow flow definition.

For more information about the CDF service, see the *Cloudera DataFlow documentation*.

Related Information

Cloudera DataFlow documentation

Enabling DataFlow for an environment

To deploy flow definitions, you must enable Cloudera DataFlow (CDF) for a CDP Public Cloud environment. Enabling CDF for an environment means that you are preparing an active and healthy CDP environment for use with CDF.

Before you begin

⚠ **Important:** You must have the DFAdmin role to perform this task.

Procedure

1. Log in to the CDP web interface and click on the DataFlow service.
2. Go to Environments and click Enable against the environment you want to enable.
3. On the **Enable Environment** page:
   - Specify the minimum and maximum number of the Kubernetes nodes you want to deploy in the **DataFlow Capacity** section.
   - Under the **Networking** section, select the subnet from the Available Subnets that CDF can use for worker node and load balancer placement.

     If you want to access CDF components from the internet, select the Enable public endpoint option.

     ⚠ **Important:** To specify a public endpoint, you must have at least one public subnet in your environment.
   - A list of source IP address ranges which are allowed to connect to the Kubernetes API server.
4. Click Enable.

   Enabling CDF can take up to one hour.

Setting up the Cloudera Data Engineering service

Set up Cloudera Data Engineering in CDP Public Cloud for use with your Business Intelligence at Scale pattern. In the Business Intelligence at Scale pattern, Data Engineering provides the streaming and static data processing and transformation capabilities.

⚠ **Attention:** Enabling CDE can take upto two hours.

For more information about Cloudera Data Engineering, see the *Data Engineering Overview*.

Related Information

Data Engineering Overview

Creating Data Hub cluster with Data Engineering cluster definition

You must create a Data Hub cluster with the "Data Engineering for AWS" cluster definition to obtain the Hive JDBC URL. You need the Hive JDBC URL while configuring the Spark job.

Before you begin

⚠ **Important:** You must have the EnvironmentUser role to perform this task.

Procedure

1. Log into the CDP web interface.
2. Navigate to Management Console Environments , and select the environment where you would like to create a cluster.
3. Click Create Data Hub.
4. Select the Cluster Definition option.
5. Select Data Engineering for AWS cluster definition from the Cluster Definition drop-down menu under Services.

   The list of services is automatically shown below the selected cluster definition name.
6. Specify a name for your cluster and add any tags you might need.
7. Click Provision Cluster.

Enabling Cloudera Data Engineering

Before you can use the Cloudera Data Engineering (CDE) service, you must enable it on each environment that you want to use CDE on.

Before you begin

⚠ **Important:** You must have the DEAdmin role to perform this task.

Procedure

1. Log in to the CDP web interface and click on the Data Engineering service.
2. In the CDE Services column, click ⊕ or Enable new CDE Service to enable CDE for an environment.
3. On the **Enable CDE Service** page, specify a name for the CDE service.
4. Select your environment from the Environment drop-down menu.

5. Select the Workload Type.

   The workload type corresponds to the instance size that will be deployed to run your submitted Spark jobs. When you select a type, the corresponding cloud provider instance size is displayed in the **Summary** section to the right.

   For proof of concept of this pattern, you can select General - Medium.
6. Accept the default settings for the Autoscale Range.
7. Select the Enable Public Endpoint option under the **Networks** section.
8. Click Enable.

Results

The CDE Overview page displays the status of the environment initialization. You can view logs for the environment by clicking ⋮ on the environment, and then clicking View Logs.

Creating a virtual cluster

In Cloudera Data Engineering (CDE), a virtual cluster is an individual auto-scaling cluster with defined CPU and memory ranges. Jobs are associated with virtual clusters, and virtual clusters are associated with an environment. You can create as many virtual clusters as you need.

Before you begin

Important: You must have the DEAdmin role to perform this task.

Procedure

1. From the CDE Overview page, select the CDE service that you created earlier.
2. In the Virtual Clusters column, click ⊕ to create a new virtual cluster.

   If the environment has no virtual clusters associated with it, the page displays a Create DE Cluster option that launches the same wizard.
3. On the **Create a Cluster** page, enter a Cluster Name.

   Cluster names must:

   - begin with a letter
   - be between 3 and 30 characters (inclusive)
   - contain only alphanumeric characters and hyphens
4. Select the CDE Service that you created earlier from the CDE Service drop-down menu.

   The CDE service you selected before launching the wizard is selected by default, but you can use the wizard to create a virtual cluster in a different CDE service.
5. Accept the default settings for the Autoscale Max Capacity.
6. Select Spark 2.4.7 as the Spark Version to use in the virtual cluster.

   You cannot use Spark 2 and Spark 3 in the same virtual cluster, but you can have separate Spark 2 and Spark 3 virtual clusters within the same CDE service.
7. Click Create.

Results

On the CDE Overview page, select the environment to view the virtual cluster initialization status. You can also click ⋮ for the virtual cluster to view the logs.

Setting up Cloudera Data Warehouse

In the Business Intelligence at Scale pattern, Data Warehouse provides ad-hoc data analytics capabilities and the Data Visualization integration provides visual reporting and dashboarding capabilities. To set up the CDW service, you must activate an environment and create an Impala Virtual Warehouse with the Unified Analytics and Data Visualization options enabled. You can use the default Database Catalog that is created when you activate an environment.

For more information about the Cloudera Data Warehouse service, see the *Data Warehouse Overview*.

Related Information

Cloudera Data Warehouse Data Service overview

Activating features under entitlement

Unified Analytics and Data Visualization features in Cloudera Data Warehouse (CDW) are under entitlement. Contact your Cloudera Account Representative to activate these features for your environment.

Procedure

Activate the following entitlements after you register your cloud environment with CDP:

- CDW_STANDALONE_DATA_VISUALIZATION: To enable Cloudera Data Visualization
- CDW_FENG_IMPALA: To use Unified Analytics in Hue

Activating an AWS Environment for CDW

Activating an environment with Cloudera Data Warehouse service sets up the Kubernetes cluster, which provides the computing resources for the Database Catalog. In addition, activating an environment enables the CDW service to use the existing Data Lake that was set up for the environment, including all data, metadata, and security.

Before you begin

Important: You must have the DWAdmin role to perform this task.

Procedure

1. Log in to the CDP web interface and click Data Warehouse.
2. Expand the Environments column by clicking More… and locate the environment that you want to activate.
3. When you locate the environment, click the activation icon ⚡ to launch the **Activation Settings** dialog.
4. In the **Activation Settings**, select the Private Load Balancer, Private Worker Nodes option from the Deployment Mode drop-down menu.
5. Click ACTIVATE.

Adding an Impala Virtual Warehouse

A Virtual Warehouse is an instance of compute resources that is equivalent to a cluster. You must create an Impala Virtual Warehouse, with the Unified Analytics option selected, to work with materialized views. You must also select the Enable Data Visualization option, so that you can use tables and queries to generate dashboards in Cloudera Data Visualization.

Before you begin

Important: You must have the DWAdmin role to perform this task.

Procedure

1. Log in to the CDP web interface and click the Data Warehouse service.
2. Go to Virtual Warehouses and click ADD NEW.
3. On the **New Virtual Warehouse** dialog:
   a) Specify a Name.

      ⚠ **Important:** The fully qualified domain name of your Virtual Warehouse, which includes the Virtual Warehouse name plus the environment name, must not exceed 64 characters; otherwise, Hue cannot load.

   b) Select IMPALA as the Virtual Warehouse type.
   c) Select the Database Catalog associated with your environment.
   d) Specify the User Groups that can access the endpoints.
   e) Select a size from the Size drop-down menu.
   f) Accept the default configuration for the auto-scaling thresholds.
   g) Click Enable Unified Analytics.
   h) Click Enable Data Visualization .
4. Click CREATE.

Business Intelligence at Scale steps

Learn how to deploy the Broger Foods' inventory management use case in your environment. You can follow the same steps as a guide to deploy your use case leveraging the Business Intelligence at Scale pattern.

Downloading the use case artifacts

To get you started with the Broger Foods' Business Intelligence at Scale use case, Cloudera has made the required use case artifacts available to you. You can use these artifacts as a starting point to build your own use case.

Schema Registry schemas for ingesting data into your Streams Messaging cluster

- weather-current-schema
- weather-forecast-schema

Custom Cloudera DataFlow flow definition for ingesting data into Kafka in your Streams Messaging cluster

- weather-data-ingest-into-kafka.json

Historical weather data and sales data ingest scripts to run in Impala Virtual Warehouse

- historical-weather-data-ingest.txt
- sale-data-ingest.txt
- bi-at-scale-historic-weather-data.zip
- bi-at-scale-historic-sales-data.zip

Spark jar file for running the Spark job in Cloudera Data Engineering

- bi-workflow_2.11-0.1.jar

Data Visualization dashboard to visualize the analytic findings

- weather-impact-on-sales.json

Meeting the prerequisites

The Business Intelligence at Scale pattern is designed to run on CDP and leverages various Cloudera Data Services. You must meet the individual prerequisites for Cloudera DataFlow (CDF), Cloudera Data Engineering (CDE), and Cloudera Data Warehouse (CDW) Data Services. All your users must have the necessary roles and permissions to perform their respective tasks.

Prerequisites for using the Business Intelligence at Scale pattern

- You have downloaded the pattern artifacts for use throughout the Business Intelligence at Scale pattern steps.
- You have signed up with openweathermap.org and have obtained their API key for real time weather data. This is so that you can stream weather data into Kafka.
- You have set up the necessary Ranger policies across your data pipeline. You have granted a user or group access to:
  - Kafka
  - Schema Registry
  - Cloudera DataFlow
  - Cloudera Data Engineering
  - Cloudera Data Warehouse

Prerequisites for using CDF

- You have the CDF service enabled and running.
- You have set up a Streams Messaging cluster in CDP Public Cloud with Kafka installed.
- You have created the required Kafka topics.
- You have the required schemas managed by Schema Registry.
- You have set up IDBroker mapping for CDF and CDE to access S3 buckets.

Prerequisites for using CDE

- You have created a Data Hub cluster with the Data Engineering cluster definition.
- You have the CDE experience enabled.
- You have one Virtual Cluster.
- You have access control set up for the team.

Prerequisites for using CDW

- You have the CDW experience enabled.
- You have activated the Unified Analytics and Data Visualization entitlement features with help of your Cloudera Account Representative.
- You have created an Impala Virtual Warehouse with Enable Unified Analytics and Enable Data Visualization options.
- You have the Hive Warehouse Connector installed.

Ingesting data using Streams Messaging

You must configure access to the streaming data from a third-party data provider, create Kafka topics, and add schemas to the Schema Registry in the Streams Messaging cluster.

Setting up the streaming data source

For the Business Intelligence at Scale pattern, you need to stream weather data from an external data provider into CDP. To obtain data from OpenWeather, you must sign up for an account with them and generate API keys.

Procedure

1. Go to https://openweathermap.org/api and create your account.
2. Go to your profile and click My API keys.
3. Specify an API key name and click Generate.

   A new API key is created. You will need this while deploying flows in CDF.

Adding schemas to the Schema Registry

Learn how to add new schemas to Schema Registry so that you ensure data consistency across your data pipeline.

About this task

Schema Registry provides a shared repository for your schemas to ensure that your data is accessible across the data pipeline. For the Business Intelligence at Scale Pattern, you must create the following two schemas using the Schema Regisrty files that you downloaded as part of the pattern artifacts:

- WeatherCurrent
- WeatherForecast

Before you begin

⚠ **Important:** You must have the EnvironmentAdmin role to perform this task.

⚠ **Attention:** Ensure that you understand compatibility policies. Once selected, you cannot change the compatibility policy for a schema.

Procedure

1. Log in to the CDP web interface.
2. Go to Management Console Data Hub Clusters and select the CMM cluster that you created earlier.
3. Click Schema Registry under the **Services** section.

   The **All Schemas** page is displayed.
4. Click + .

   The **Add New Schema** page is displayed.
5. Add the schema metadata as follows:

Field name	Description and values
NAME	Used as a key to look up schemas. Specify WeatherCurrent and WeatherForecast if you are trying the Business Intelligence at Scale pattern.
DESCRIPTION	A short description of the schema. For example, "Schema for getting the current weather data".
TYPE	The schema format. Select Avro schema provider if you are trying the Business Intelligence at Scale pattern.
SCHEMA GROUP	To group schemas in any logical order. Specify Kafka if you are trying the Business Intelligence at Scale pattern.
COMPATIBILITY	To set the compatibility policy for the schema. Once set, this cannot be changed. Select BACKWARD

Field name	Description and values
	if you are trying the Business Intelligence at Scale pattern.

6. To allow schema to evolve over time by creating multiple versions, select the Evolve option.

   Note: If you deselect the Evolve option, then you can only have one version of the schema.

7. Upload the current and forecast weather schemas that you downloaded as part of the pattern artifacts by clicking Browse in the SCHEMA TEXT area.
8. Click SAVE.

Creating Kafka Topics

For Business Intelligence at Scale, you need to create the following two Kafka Topics: "weather" and "weather_forecast". The "weather" Topic is used for streaming current weather information and the "weather_forecast" Topic is used for the weather forecast information.

Before you begin

Important: You must have an EnvironmentUser role to perform this task.

Procedure

1. Log in to the CDP web interface.
2. Go to Management Console Data Hub Clusters and select the CMM cluster that you created earlier.
3. Click Streams Messaging Manager under the **Services** section.
4. Go to Topics and click Add New.
5. Provide the following required information if you are trying the Business Intelligence at Scale pattern:

Field name	Description and values
TOPIC NAME	A name for your Topic. Specify weather and weather_forecast.
PARTITIONS	Specify 2.
Availability	The level of availability. Select MAXIMUM.
CLEANUP.POLICY	The operation to cleanup the data. Select delete. .

6. Click Save.

Building data pipelines using Cloudera DataFlow

You must deploy a ReadyFlow in the Cloudera DataFlow cluster, develop your flow definition, collect flow information, and deploy Nifi flows.

Adding a ReadyFlow to the Catalog

ReadyFlows are out-of-the-box flow definitions designed to help you get started with Cloudera DataFlow (CDF) quickly and easyily. For the Business Intelligence at Scale pattern, select the "Kafka to S3 Avro" ReadyFlow definition.

About this task

To get started with the data ingest portion of your Business Intelligence at Scale pattern, start by adding the necessary data flows to your CDF Catalog. You must add a custom NiFi data flow (in JSON format) which is available in your pattern artifacts downloads package, and the Kafka to S3 Avro ReadyFlow to your CDF Catalog. The custom NiFi data flow moves data from your streaming data source into the Kafka topics that you have created. The Kafka to S3 Avro ReadyFlow moves data from Kafka into an AWS S3 bucket so that it can be consumed by Cloudera Data Engineering.

Before you begin

⚠ **Important:** You must have the DFCatalogAdmin role to perform this task.

Procedure

1. From the ReadyFlow Gallery page, identify the ReadyFlow you want use.
2. Review the ReadyFlow details by clicking View Details.
3. Click Add To Catalog to add the ReadyFlow into the CDF Catalog and make it ready for deployment.
4. Click Add to confirm that you want to add the ReadyFlow to the Catalog.

Results

The ReadyFlow is added to the CDF Catalog and is ready for deployment.

What to do next

Import the custom NiFi data flow.

Importing the custom NiFi flow definition

Before you can deploy a flow, you must import a flow definition that contains the data flow logic you want to deploy to a Cloudera DataFlow (CDF) environment.

About this task

Import the custom NiFi flow definition available in your pattern artifacts downloads package. This flow definition moves streaming weather data into Kafka.

Before you begin

⚠ **Important:** You must have the DFCatalogAdmin role to perform this task.

Procedure

1. From the CDF left navigation pane, click Catalog.
2. Click Import Flow Definition and provide the following information:

Field name	Description
Flow Name	Specify an unique name for the flow definition you are importing. For example, Weather-data-to-kafka.
Flow Description	(Optional) Provide a description of your flow definition. Use this value to help you differentiate between the other flows imported to your catalog.
NiFi Flow Configuration	Upload the flow definition JSON file.

3. Click Import.

What to do next

Now that you have imported your custom NiFi flow definition, you can move on to collecting the flow configuration that you need to deploy the ReadyFlow and this custom NiFi flow definition for your Business Intelligence at Scale pattern.

Collecting configuration information for ReadyFlow

Learn how to collect the information you need to configure the "Kafka to S3 Avro" ReadyFlow.

The following sections help you to gather configuration information such as Schema Registry host name, Kafka brokers, and other details before you deploy NiFi flows:

For your ReadyFlow data ingest source

- You have the Schema Registry Host Name.

  1. From the Management Console, go to Data Hub Clusters and select the Streams Messaging cluster you are using.
  2. Navigate to the **Hardware** tab to locate the Master Node FQDN. Schema Registry is always running on the Master node, so copy the Master node FQDN.
- You have the Kafka broker end points.

  1. From the Management Console, click Data Hub Clusters.
  2. Select the Streams Messaging cluster from which you want to ingest data.
  3. Click the Hardware tab.
  4. Note the Kafka Broker FQDNs for each node in your cluster.
  5. Construct your Kafka Broker Endpoints by using the FQDN and Port number 9093 separated by a colon. Separate endpoints by a comma. For example:

     ```
     broker1.fqdn:9093,broker2.fqdn:9093,broker3.fqdn:9093
     ```

     Kafka broker FQDNs are listed under the **Core_broker** section.
- You have the Kafka Consumer Group ID.

  This ID is defined by the user. Pick an ID and then create a Ranger policy for it. Use the ID when deploying the flow in DataFlow.

For your ReadyFlow data ingest target

- You have your S3 path and bucket into which you want to ingest data.

- Perform one of the following to configure access to S3 buckets:
    - You have configured access to S3 buckets with a RAZ enabled environment.

      It is a best practice to enable RAZ to control access to your object store buckets. This allows you to use your CDP credentials to access S3 buckets, increases auditability, and makes object store data ingest workflows portable across cloud providers.

      1. Ensure that Fine-grained access control is enabled for your DataFlow environment.
      2. From the Ranger UI, navigate to the S3 repository.
      3. Create a policy to govern access to the S3 bucket and path used in your ingest workflow. For example: kafka-to-s3-avro-ingest

         Tip:

         The Path field must begin with a forward slash (/).
      4. Add the machine user that you have created for your ingest workflow to ingest the policy you just created.

      For more information, see *Creating Ranger policy to use in RAZ-enabled AWS environment*.
    - You have configured access to S3 buckets using ID Broker mapping.

      If your environment is not RAZ-enabled, you can configure access to S3 buckets using ID Broker mapping.

      1. Access IDBroker mappings.

         a. To access IDBroker mappings in your environment, click Actions Manage Access .
         b. Choose the IDBroker Mappings tab where you can provide mappings for users or groups and click Edit.
      2. Add your CDP Workload User and the corresponding AWS role that provides write access to your folder in your S3 bucket to the Current Mappings section by clicking the blue + sign.

         Note: You can get the AWS IAM role ARN from the Roles Summary page in AWS and can copy it into the IDBroker role field. The selected AWS IAM role must have a trust policy allowing IDBroker to assume this role.
      3. Click Save and Sync.

Deploying the NiFi flows

Learn about the steps to deploy flows (custom flow definitions or ReadyFlows) so that you can start running your NiFi flows with Cloudera DataFlow.

Before you begin

Important: You must have the DFFlowAdmin role to perform this task.

About this task

Now that you have imported the custom NiFi flow definition and added the ReadyFlow to the catalog you may proceed by deploying your Business Intelligence at Scale data flows. You must deploy the following three flows:

- weather_forecast_to_kafka: This flow gets both current and forecast weather data and sends it to the corresponding Kafka topics.
- Current_Weather_Data_To_S3: This deployment moves current weather data from the Kafka topic into an S3 bucket.
- Forecast_Weather_Data_To_S3: This deployment moves forecast weather data from the Kafka topic into an S3 bucket.

Procedure

1. Log in to the CDP web interface.

2. To create the custom flow deployment "weather_forecast_to_kafka", perform the following steps:

a) Go to DataFlow Catalog and click the "Kafka to S3 Avro" ReadyFlow definition that you added earlier.

The version information about the flow is displayed.

b) Click Deploy #.

c) Select your CDP environment from the Target Environment drop-down menu on the **New Deployment** dialog and click Continue.

d) Provide a unique name to your deployment on the **Overview** page.

For the Business Intelligence at Scale pattern, specify weather_forecast_to_kafka.

e) Accept the default settings on the **NiFi Configuration** page.

f) Specify the following on the **Parameters** page:

Field name	**Description and value**
CDP Workload User	Specify your CDP workload username.
CDP Workload User Password	Specify your CDP workload password.
Current Weather Topic Name	Specify weather.
Kafka Brokers	Provide a comma separated list of Kafka FQDNs that you collected from the Data Hub cluster earlier as follows: [****broker1.fqdn****]:9093,[****broker2.fqdn****]:9093,[****broker3.fqdn****]:9093
Schema Name	Specify WeatherForecast. This should match with the schema name that you specified in Schema Registry.
Schema Registry URL	Specify [****master-node-fqdn****]:7790/api/v1 that you collected earlier from the Data Hub cluster.
Topic Name	Specify weather_forecastThis should match with the Kafka topics that you created in Streams Messaging Manager (SMM).
Weather Forecast Topic Name	Specify weather_forecast.
api.key	Specify the API key that you generated on the OpenWeather portal.
api.key.current	Specify the API key that you generated on the OpenWeather portal.

g) In KPIs, identify key performance indicators, the metrics to track those KPIs, and when and how to receive alerts about the KPI metrics tracking.

Tip: You can reorder the KPIs by dragging them up or down the list. The order you configure here is reflected in the detailed monitoring view of the resulting deployment.

You can skip this step for this pattern.

h) Accept the default settings on the **Sizing & Scaling** page for the Business Intelligence at Scale pattern.

For production use cases, you can specify your initial flow deployment size, whether to enable auto-scaling, and up to how many nodes you want in your cluster.

Note: Flow deployments scale horizontally by adding or removing NiFi nodes to a deployment. Deployments do not scale vertically, and the node size cannot be changed.

i) Review the configuration summary and click Deploy.

You will be redirected to the Alerts tab in the detail view for the deployment where you can track its progress.

3. To create the "Current_Weather_Data_To_S3" deployment:

a) Go to DataFlow Catalog and click the "Kafka to S3 Avro" ReadyFlow definition that you added earlier.

The version information about the flow is displayed.

b) Click Deploy #.

c) Select your CDP environment from the Target Environment drop-down menu on the **New Deployment** dialog and click Continue.

d) Provide a unique name to your deployment on the **Overview** page.

For the Business Intelligence at Scale pattern, specify Current_Weather_Data_To_S3.

e) Accept the default settings on the **NiFi Configuration** page.

f) Specify the following on the **Parameters** page:

Field name	**Description and value**
CDP Workload User	Specify your CDP workload username.
CDP Workload User Password	Specify your CDP workload password.
CSV Delimiter	Specify , (a comma).
Data Input Format	Specify JSON. Ensure to input this is in All-Caps.
Kafka Broker Endpoint	Provide a comma separated list of Kafka FQDNs that you collected from the Data Hub cluster earlier as follows: [****broker1.fqdn****]:9093,[****broker2.fqdn****]:9093,[****broker3.fqdn****]:9093
Kafka Consumer Group ID	Specify WeatherConsumer.
Kafka Source Topic	Specify weather.
S3 Path	Specify the path to your S3 bucket upto the root.
S3 Bucket	Specify the name of your S3 bucket.
Schema Name	Specify WeatherCurrent. This should match with the schema name that you specified in Schema Registry.
Schema Registry Hostname	Specify [****master-node-fqdn****] that you collected earlier from the Data Hub cluster.

g) In KPIs, identify key performance indicators, the metrics to track those KPIs, and when and how to receive alerts about the KPI metrics tracking.

Tip: You can reorder the KPIs by dragging them up or down the list. The order you configure here is reflected in the detailed monitoring view of the resulting deployment.

You can skip this step for this pattern.

h) Accept the default settings on the **Sizing & Scaling** page for the Business Intelligence at Scale pattern.

For production use cases, you can specify your initial flow deployment size, whether to enable auto-scaling, and up to how many nodes you want in your cluster.

Note: Flow deployments scale horizontally by adding or removing NiFi nodes to a deployment. Deployments do not scale vertically, and the node size cannot be changed.

i) Review the configuration summary and click Deploy.

You will be redirected to the Alerts tab in the detail view for the deployment where you can track its progress.

4. To create the "Forecast_Weather_Data_To_S3" deployment:
   a) Go to DataFlow Catalog and click the "Kafka to S3 Avro" ReadyFlow definition that you added earlier.

      The version information about the flow is displayed.
   b) Click Deploy #.
   c) Select your CDP environment from the Target Environment drop-down menu on the **New Deployment** dialog and click Continue.
   d) Provide a unique name to your deployment on the **Overview** page.

      For the Business Intelligence at Scale pattern, specify Forecast_Weather_Data_To_S3.
   e) Accept the default settings on the **NiFi Configuration** page.
   f) Specify the following on the **Parameters** page:

Field name	**Description and value**
CDP Workload User	Specify your CDP workload username.
CDP Workload User Password	Specify your CDP workload password.
CSV Delimiter	Specify , (a comma).
Data Input Format	Specify JSON. Ensure to input this is in All-Caps.
Kafka Broker Endpoint	Provide a comma separated list of Kafka FQDNs that you collected from the Data Hub cluster earlier as follows: [****broker1.fqdn****]:9093,[****broker2.fqdn****]:9093,[****broker3.fqdn****]:9093
Kafka Consumer Group ID	Specify ForecastConsumer.
Kafka Source Topic	Specify weather_forecast.
S3 Path	Specify the path to your S3 bucket upto the root.
S3 Bucket	Specify the name of your S3 bucket.
Schema Name	Specify WeatherForecast. This should match with the schema name that you specified in Schema Registry.
Schema Registry Hostname	Specify [****master-node-fqdn****] that you collected earlier from the Data Hub cluster.

   g) In KPIs, identify key performance indicators, the metrics to track those KPIs, and when and how to receive alerts about the KPI metrics tracking.

      Tip: You can reorder the KPIs by dragging them up or down the list. The order you configure here is reflected in the detailed monitoring view of the resulting deployment.

      You can skip this step for this pattern.
   h) Accept the default settings on the **Sizing & Scaling** page for the Business Intelligence at Scale pattern.

      For production use cases, you can specify your initial flow deployment size, whether to enable auto-scaling, and up to how many nodes you want in your cluster.

      Note: Flow deployments scale horizontally by adding or removing NiFi nodes to a deployment. Deployments do not scale vertically, and the node size cannot be changed.
   i) Review the configuration summary and click Deploy.

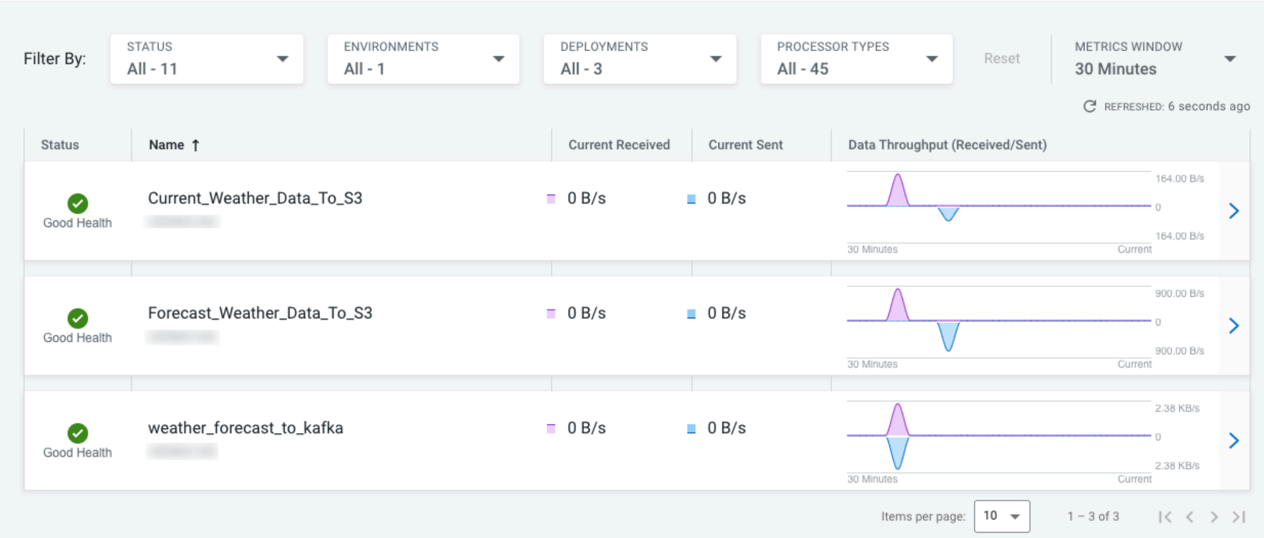
      You will be redirected to the Alerts tab in the detail view for the deployment where you can track its progress.

Results

When you have deployed all three flows associated with the Business Intelligence at Scale pattern you will have 3 flow deployments, such as:

- Current_Weather_Data_To_S3
- Forecast_Weather_Data_To_S3
- weather_forecast_to_kafka

Dashboard



Filter By: STATUS All - 11 | ENVIRONMENTS All - 1 | DEPLOYMENTS All - 3 | PROCESSOR TYPES All - 45 | Reset | METRICS WINDOW 30 Minutes

REFRESHED: 6 seconds ago

Status	Name ↑	Current Received	Current Sent	Data Throughput (Received/Sent)
Good Health	Current_Weather_Data_To_S3 [redacted]	0 B/s	0 B/s	164.00 B/s / 0 / 164.00 B/s, 30 Minutes – Current
Good Health	Forecast_Weather_Data_To_S3 [redacted]	0 B/s	0 B/s	900.00 B/s / 0 / 900.00 B/s, 30 Minutes – Current
Good Health	weather_forecast_to_kafka [redacted]	0 B/s	0 B/s	2.38 KB/s / 0 / 2.38 KB/s, 30 Minutes – Current

Items per page: 10 1 – 3 of 3

Transforming and organizing data using Cloudera Data Engineering

You must define the Spark job configurations, resources, and schedule in the Cloudera Data Engineering (CDE) cluster.

About this task

In CDE, you perform the organization and transformation piece of the Business Intelligence at Scale pattern. A Spark jar file (bi-workflow_2.11-0.1.jar) has been created for you and is available as part of your pattern artifacts download package to create a Spark job. The Spark job converts the weather data that you put into the AWS S3 bucket in Avro format to Parquet, and makes the Parquet data available to Cloudera Data Warehouse.

Before you begin

⚠ **Important:** You must have the DEUser role to perform this task.

⚠ **Attention:** You must load data into CDW and then run the Spark job. Perform this task after completing About adding data to Data Warehouse and Creating tables and inserting data.

Procedure

1. Log in to your AWS console and create a tmp sub-folder in your S3 bucket within the path that you are using for this pattern.
2. Log in to the CDP web interface.
3. Go to Data Engineering.
4. In the Environments column, select the environment containing the virtual cluster where you want to create the job.

5. In the Virtual Clusters column on the right, click the View Jobs icon on the virtual cluster where you want to create the application.
6. Go to Jobs Create Job .
7. Provide the job details:
   a) Select Spark 2.4.7 as the Job Type.
   b) Specify the Name.
   c) Select File and upload the JAR file that you downloaded earlier.
   d) Specify the following in the Main Class field:

   ```
   com.cloudera.cde.workflow.BiWorkflow
   ```

   e) Specify the following in the Arguments field:

Argument
--weatherTable
weather
--weatherForecastTable
weather_forecast
--weatherDataDirRoot
s3a://[****FULL-PATH-TO-YOUR-S3-BUCKET****]

   Click ⊕ (Add Argument) to add multiple arguments as necessary.

   weather and weather_forecast sub-folders are automatically created in the S3 bucket by the Spark job.

   f) Enter the following in the Configurations field:

Configuration	Value
spark.datasource.hive.warehouse.load.staging.dir	[***PATH-TO-S3-BUCKET***]/tmp
spark.driver.extraJavaOptions	-Duser.timezone=UTC
spark.executor.extraJavaOptions	-Duser.timezone=UTC
spark.sql.hive.hiveserver2.jdbc.url	jdbc:hive2://[****HIVE-SERVER-2-FQDN****]-fsio.int.cldr.work:2181/default;httpPath=cliservice;serviceDiscoveryMode=zooKeeper;ssl=t Obtain the [****HIVE-SERVER-2-FQDN****] from going to CDP Home Management Console Data Hub Clusters [***DATA-HUB-WITH-DATA-ENGINEERING-CLUSTER-DEFINITION***] Cloudera Manager Clusters Hive service Instances and copying the FQDN corresponding to HiveServer2.
spark.sql.session.timeZone	UTC

   Click ⊕ (Add Configuration) to add multiple configuration parameters as necessary.

   g) Click Schedule to display scheduling options.

8. Schedule the application to run periodically using the Basic controls as follows:

   Every hour at 30 minute(s) past the hour.

   Select a Start Date and an End Date.
9. Click Schedule to create the job.

Analyzing data and visualizing results using Cloudera Data Warehouse

To run queries and draw insights, you must add data to the Data Warehouse, create materialized view, create tables to generate reports and dashboards in Cloudera Data Visualization, and render and share dashboards from Cloudera Data Visualization with your stakeholders.

About adding data to Data Warehouse

To run queries and understand the correlation between the historical sales and historical weather data, you must make this data available in Cloudera Data Warehouse tables.

In a real-world scenario, the historical sales data and weather data could be available in your CDP Data Lake. For the Business Intelligence at Scale pattern, you can download the sample historical sales and weather data set provided by Cloudera from your Business Intelligence at Scale download package on your computer and then upload it into your S3 bucket in individual folders. These two packages contain historical data in Parquet format.

You will fetch this data in Hue directly by specifying the complete path of the S3 directory in the LOCATION clause of a query.

Important: You must have access to the S3 bucket to perform this task.

Creating tables and inserting data

After making the sales and weather data available in S3, you first create external tables, insert data into them, and then create managed tables using the external tables.

Before you begin

Important: You must have the DWUser role to perform this task.

Procedure

1. Log in to the CDP web interface.
2. Go to Data Warehouse.
3. Locate your Impala Virtual Warehouse and click on Hue.
4. Open the Unified Analytics editor from the left assist panel.
5. Run the following query to create an external table "sales":

```
CREATE EXTERNAL TABLE `sales_external`(
  `store_nbr` bigint,
  `item_nbr` bigint,
  `local_date` date,
  `city_name` string,
  `item_name` string,
  `units` double)
ROW FORMAT SERDE
  'org.apache.hadoop.hive.ql.io.parquet.serde.ParquetHiveSerDe'
STORED AS INPUTFORMAT
  'org.apache.hadoop.hive.ql.io.parquet.MapredParquetInputFormat'
OUTPUTFORMAT
```

```
  'org.apache.hadoop.hive.ql.io.parquet.MapredParquetOutputFormat'
LOCATION
  's3a://[***COMPLETE-PATH-OF-S3-BUCKET***]/bi-at-scale-historic-sales-d
ata/';
```

bi-at-scale-historic-sales-data is the name of the directory containing Parquet files that you uploaded to S3.

6. Verify that the historical sales data is available by running the following query:

```
SELECT count(*) FROM sales_external;
```

7. Run the following query to create a managed table that can receive sales data from the external table:

```
CREATE TABLE sales (
    store_nbr bigint,
    item_nbr bigint,
    local_date date,
    item_name string,
    units double
                   )
PARTITIONED BY (city_name string);
```

8. Run the following query to insert data into the managed "sales" table:

```
INSERT INTO sales PARTITION (city_name)
SELECT  store_nbr,
        item_nbr,
        local_date,
        item_name,
        units,
        city_name
FROM sales_external;
```

9. Verify that the historical sales data is available by running the following query:

```
SELECT count(*) FROM sales;
```

10. Run the following query to create an external table "weather_hist_external":

```
CREATE external TABLE weather_hist_external(
  city string,
  local_date date,
  avg_temperature DOUBLE,
  min_temperature DOUBLE,
  max_temperature DOUBLE,
  avg_feels_like_temperature DOUBLE,
  min_feels_like_temperature DOUBLE,
  max_feels_like_temperature DOUBLE,
  avg_humidity DOUBLE,
  min_humidity DOUBLE,
  max_humidity DOUBLE,
  avg_pressure DOUBLE,
  min_pressure DOUBLE,
  max_pressure DOUBLE,
  avg_wind_speed DOUBLE,
  min_wind_speed DOUBLE,
  max_wind_speed DOUBLE,
  avg_snow_1h DOUBLE,
  min_snow_1h DOUBLE,
  max_snow_1h DOUBLE,
  avg_rain_1h DOUBLE,
  min_rain_1h DOUBLE,
  max_rain_1h DOUBLE,
  creation_timestamp timestamp
  )
  location 's3a://[***COMPLETE-PATH-OF-S3-BUCKET***]/bi-at-scale-historic-
weather-data/';
```

bi-at-scale-historic-weather-data is the name of the directory containing Parquet files that you uploaded to S3.

11. Verify that the historical weather data is available by running the following query:

```
SELECT count(*) FROM weather_hist_external;
```

12. Run the following query to insert data into the "weather_forecast" table:

```
INSERT INTO weather_forecast --PARTITION (city) -- CHECK FOR PARTITION E
RROR ----
SELECT  local_date,
        avg_temperature as avg_temp_k,
        min_temperature as min_temp_k,
        max_temperature as max_temp_k,
        (avg_temperature - 273.15) * 9/5 + 32 as avg_temp_f,
        (min_temperature - 273.15) * 9/5 + 32 as min_temp_f,
        (max_temperature - 273.15) * 9/5 + 32 as max_temp_f,
        avg_feels_like_temperature as avg_feels_like_temp_k,
        min_feels_like_temperature as min_feels_like_temp_k,
        max_feels_like_temperature as max_feels_like_temp_k,
        (avg_feels_like_temperature - 273.15) * 9/5 + 32 as avg_feels_li
ke_temp_f,
        (min_feels_like_temperature - 273.15) * 9/5 + 32 as min_feels_li
ke_temp_f,
        (max_feels_like_temperature - 273.15) * 9/5 + 32 as max_feels_li
ke_temp_f,
        avg_humidity,
        min_humidity,
        max_humidity,
        avg_pressure,
        min_pressure,
        max_pressure,
        avg_wind_speed,
        min_wind_speed,
        max_wind_speed,
        avg_snow_1h,
        min_snow_1h,
        max_snow_1h,
        avg_rain_1h,
        min_rain_1h,
        max_rain_1h,
        creation_timestamp,
        city
FROM    weather_hist_external;
```

Note: "weather" and "weather_forecast" tables are created by the Spark job.

13. Verify that the forecast weather data is available by running the following query:

```
SELECT count(*) FROM weather_forecast;
```

Creating a materialized view

After creating managed tables and inserting the data, you can join the tables using a materialized view. You run the query, and the optimizer takes advantage of the precomputation performed by the materialized view for a faster response time.

Before you begin

Important: You must have the DWUser role to perform this task.

Procedure

1. Log in to the CDP web interface.
2. Go to Data Warehouse.
3. Locate your Impala Virtual Warehouse and click on Hue.
4. Open the Unified Analytics editor from the left assist panel.
5. Run the following query to create materialized view:

```
CREATE MATERIALIZED VIEW sales_weather_join
STORED AS PARQUET AS
SELECT  w.local_date,
        w.min_temp_f,
        w.min_temp_k,
        w.max_temp_f,
        w.max_temp_k,
        w.avg_temp_k,
        w.avg_temp_f,
        w.min_rain_1h,
        w.max_rain_1h,
        w.avg_rain_1h,
        w.min_snow_1h,
        w.max_snow_1h,
        w.avg_snow_1h,
        w.city,
        s.item_name,
        s.units,
        s.city_name AS city_name
FROM    weather_forecast w INNER JOIN sales s
ON      w.city = s.city_name
AND     w.local_date = s.local_date;

ANALYZE TABLE weather COMPUTE STATISTICS;

ANALYZE TABLE weather_forecast COMPUTE STATISTICS;

ANALYZE TABLE sales COMPUTE STATISTICS;

ALTER MATERIALIZED VIEW sales_weather_join REBUILD;
```

Creating a Data Visualization dashboard by importing visual artifacts

To visualize the correlation and share the insights with your stakeholders, you can manually create a dashboard in Cloudera Data Visualization (CDV) or use a visual artifact in the form of a JSON file and import it in CDV.

About this task

For the Business Intelligence at Scale pattern, use the "cdv-dashboard-Effect-of-Weather-on-Sales.json" script provided as part of the download package to create the tables required for generating dashboard.

Before you begin

Important: You must have the DWUser role to perform this task.

Procedure

1. Log in to the CDP web interface.
2. Go to Data Warehouse.
3. Locate your Impala Virtual Warehouse and click on Data VIZ.
4. Click DATA on the main navigation bar.

   All Connections and the available Datasets are displayed.
5. Click on Default Impala VW from the left navigation pane.
6. Click ••• at the top of the page and click Import Visual Artifacts.
7. On the **Import Visual Artifacts** pop-up, click Choose File and select the file you want to import.
8. Specify the destination workspace in the Import to Workspace drop-down menu.

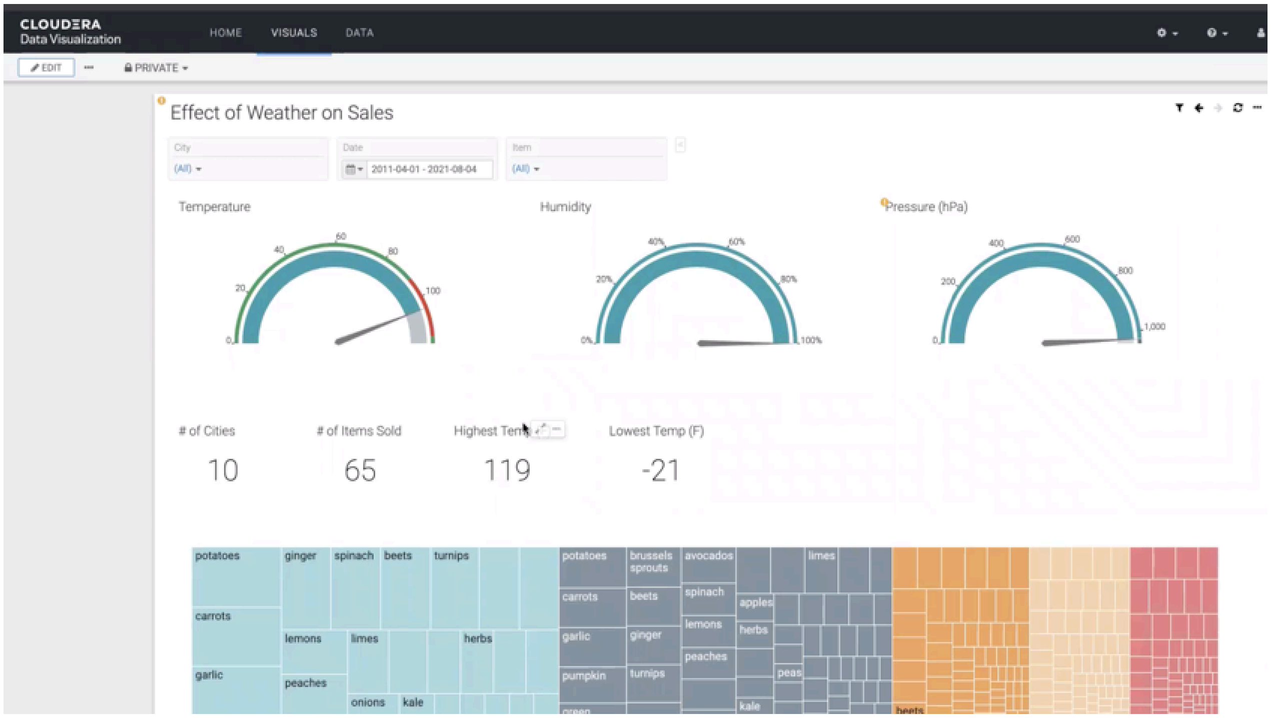
   By default, Data Visualization imports the visual artifacts into your “Private” workspace. However, you can choose the “Public” workspace, or another workspace.
9. Select Import from the file from the Thumbnail Action drop-down menu.
10. Click IMPORT and confirm by clicking ACCEPT AND IMPORT.

Results

After the import completes, a success message appears on the Import interface. All the artifacts have an assigned ID, which are generated by the system, sequentially. Visuals/Dashboards and datasets have separate ID queues.

Example

Completing this task for your Business Intelligence at Scale pattern results in a dashboard that looks similar to the following example:



What to do next

Now that you have created your Business Intelligence at Scale pattern dashboard, share it with other stakeholders.

1. Open your dashboard from the VISUAL interface.
2. Click the more option and select Get URL.
3. Copy the URL by right-clicking it and share it with your stakeholders.