..

Cloud Bursting

Date published: 2025-11-15
Date modified:

https://docs.cloudera.com/

Legal Notice

Contents

Creating Data Hub Clusters in Hybrid Environments

Learn how to create data hub clusters in hybrid environments.

About this task

After registering your hybrid environment, you can create Cloudera Data Hub clusters that enable cloud bursting from your on-premises environment.

Before you begin

- **DataHubCreator**
- **EnvironmentAdmin** at the scope of the environment where the Cloudera Data Hub cluster is running, or
- Owner of the environment

Procedure

1. Navigate to your hybrid environment.
2. Select the Data Hubs tab on the environment details page.
3. Click Create Data Hub.
4. Select the Hybrid Data Engineering: HA: Apache Spark3, Apache Hive definition. The definition will automatically match the on-premises base cluster version.
5. Provide the Cluster Name.

   The name must be 5 to 40 characters, start with a letter, and can only include lowercase letters, numbers, and hyphens.
6. Optionally, add tags that the data hub should use to tag your Cloud resources.

   Click Add to add a tag, and then enter a key and value for each tag. Repeat the steps if you would like to add more tags. For more information about tags, refer to Tags in our documentation.
7. Optionally, click Advanced Options to modify advanced cluster settings. For more information on these options, refer to Advanced cluster options.
8. Optionally, you can enable Autoscaling on this cluster to leverage cloud elasticity.
9. Click Provision Cluster.

Results

You will be redirected to the Cloudera Data Hub cluster dashboard. When your cluster is ready, its status will change to Running.

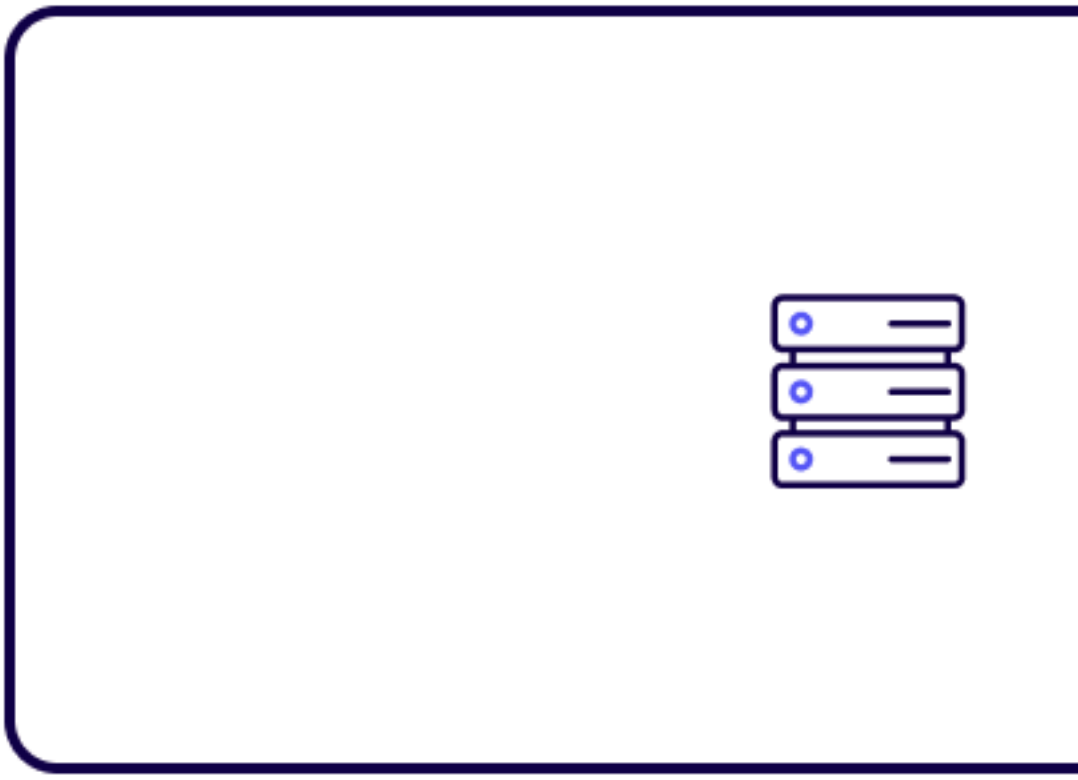
Using Hybrid Data Hubs to burst Cloudera on premises workloads

Learn how to burst Cloudera on premises to Cloudera Hybrid Data Hubs.

Hybrid Data Hubs extend on-premise capabilities by providing autoscaling cloud compute resources. Workloads submitted to cloud-based engines (such as Spark and Hive) access on-premises datasets in-place, strictly adhering to existing local security policies.

Because Hybrid Data Hubs maintain runtime parity with the on-premise cluster, workloads do not require re-engineering to migrate between platforms. Consequently, on-premise workloads can be burst to the cloud simply by submitting them to the corresponding engine within the Hybrid Data Hub.

The following diagram represents the lifecycle of such a Spark workload.



When the on premises Spark workload is submitted to Hybrid Data Hub, its Spark engine generates an execution plan that reads only the necessary data subsets, such as specific partitions or columns, from the on-premise HDFS.

Job processing occurs within the cloud infrastructure and the final results are written back to the on-premise storage (if there is a persist stage). All intermediate data generated during execution is ephemeral and is cleaned up upon job completion.

Note: You can track workloads burst to Hybrid Data Hubs via the Hybrid Data Hub Resource Manager UI.

Additionally, on-premise workload schedules can be optimized for **compute redundancy**. By integrating Hybrid Data Hub endpoints (such as Spark and Hive) into the scheduler logic, jobs can be automatically submitted to the hybrid environment whenever on-premise YARN utilization approaches capacity.