Cloudera Runtime 7.2.6

DAS: Core Concepts

Date published: 2020-02-28
Date modified: 2020-12-11

https://docs.cloudera.com/

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

Data Analytics Studio overview..... 4

DAS architecture.....4

Data Analytics Studio overview

Data Analytics Studio (DAS) is an application that provides diagnostic tools and intelligent recommendations to make the business analysts self-sufficient and productive with Hive.

DAS helps you to perform operations on Hive tables and provides recommendations for optimizing the performance of your queries. You can use DAS to:

- Search queries: You can search for queries executed on Hive tables in a database. You can further refine your search results using filters. DAS provides recommendations to optimize the performance of your queries on Hive tables. You can view the recommendations and edit your queries.
- Compose and execute queries: You can compose queries using the intuitive query composer. It has context based auto-complete feature that helps you to edit a query faster. You can also view the visual explain of the query after executing it. You can save queries to view them later and edit them. You can edit the existing, saved queries and then save them as new queries. When you try to edit a query, you can use the query composer to easily create and execute your queries.
- Compare queries: You can compare two queries to know how each query is performing in terms of speed and cost effectiveness. DAS compares various aspects of the two queries, based on which you can identify what changed between the execution of those two queries, and you can also debug performance-related issues between different runs of the same query.
- Manage databases: Using the Database Explorer, you (the admin user) can manage existing databases by creating new tables, editing existing tables, and deleting tables. You can also create new database and add tables to it. You can manage existing tables by editing them to modify existing columns or add new columns. You can create new tables in DAS or upload existing tables available in CSV, JSON, and XML formats. You can edit columns in tables and also view suggestions for partitions and implement these recommendations.
- View reports: You can view which columns and tables are used for joins and make changes to the data layout to optimize the performance of the query with different search criteria.

DAS architecture

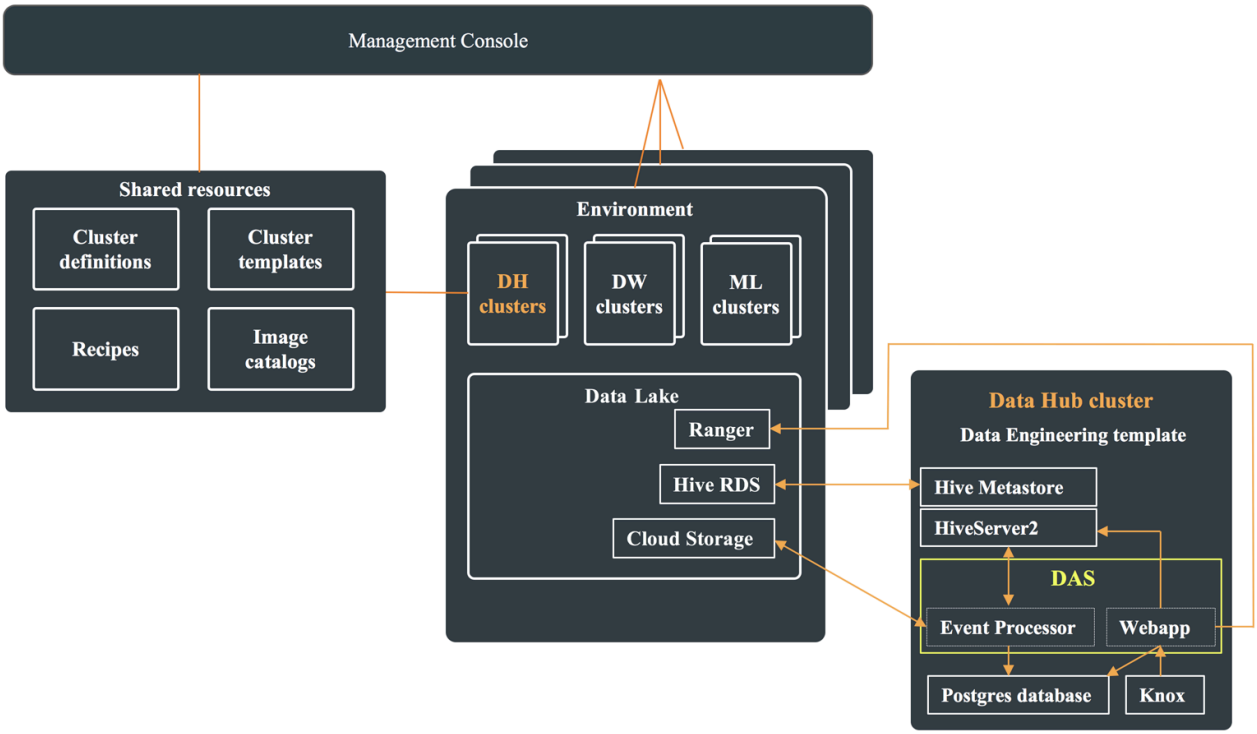
In a CDP public cloud deployment, DAS is available as one of the many Cloudera Runtime services within the Data Engineering template. To use DAS, create a Data Hub cluster from the Management Console, and select the Data Engineering template from the Cluster Definition dropdown menu.

Each Data Hub cluster that is deployed using the Data Engineering template has an instance of Hive Metastore (HMS), HiveServer2 (HS2), and DAS. The HMS communicates with the Hive RDS in the Data Lake for storing the metadata information. It also communicates with Ranger in the Data Lake for authorization. Hive is configured to use the HMS in the Data Hub cluster, and Ranger in the Data Lake cluster.

DAS comprise of an Event Processor and a Webapp. The Event Processor processes Hive and Tez events, and also replicates the metadata. The Webapp serves as the DAS UI. It serves data from the local Postgres instance which is populated by the event processor, and uses the HiveServer2 to run Hive queries triggered through the **Query Compose** page. The Event Processor communicates with the Data Hub HS2 instance for the repl dump operation, and reads the events published by Hive and Tez from the cloud storage (S3 or ADLS).

The following diagram shows the components inside the Data Hub cluster which use the shared resources within the Data Lake, such as Ranger, Hive RDS, and the cloud storage:

Figure 1: DAS Architecture in CDP Public Cloud


Management Console
Shared resources
Cluster definitions
Cluster templates
Recipes
Image catalogs
Environment
DH clusters
DW clusters
ML clusters
Data Lake
Ranger
Hive RDS
Cloud Storage
Data Hub cluster
Data Engineering template
Hive Metastore
HiveServer2
DAS
Event Processor
Webapp
Postgres database
Knox