Cloudera Runtime 7.2.11

DAS: Core Concepts

Date published: 2020-02-28
Date modified: 2021-09-09

Legal Notice

Contents

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 run 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 run 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

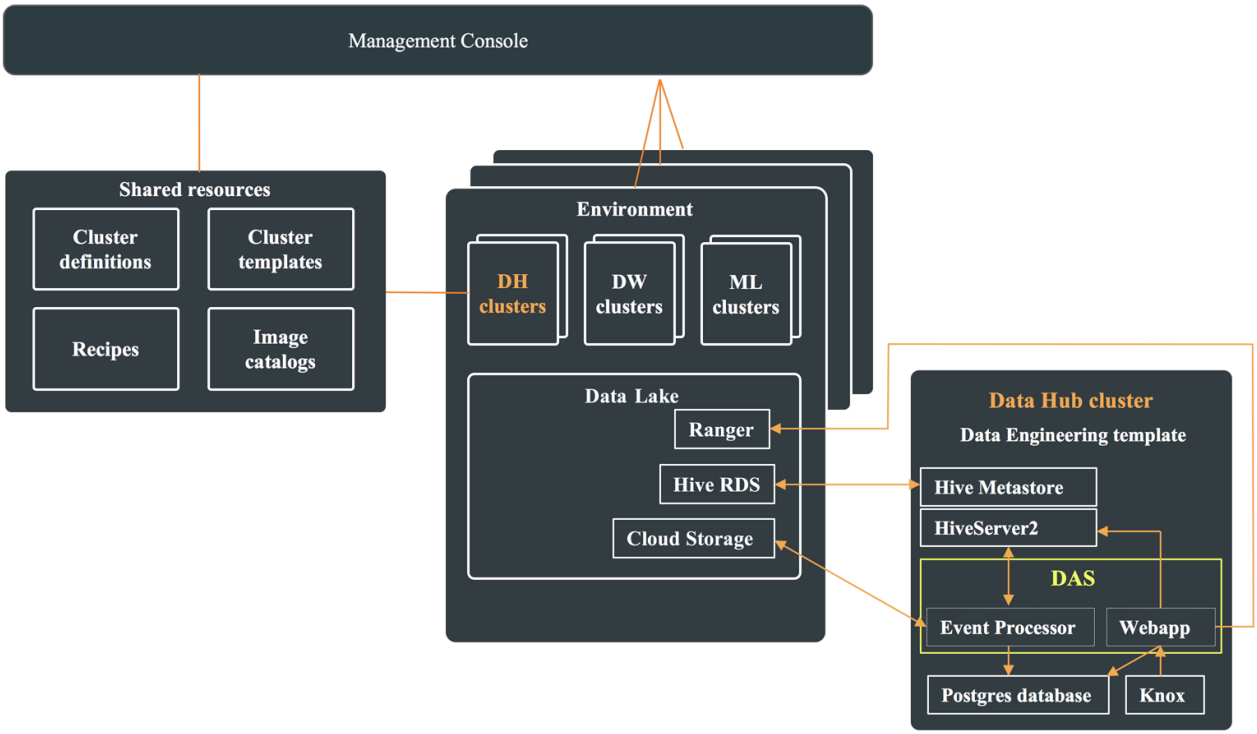
In a CDP public cloud deployment, DAS is available as one of the many Cloudera Runtime services within the Data Engineering template. To use DAS, create a Data Hub cluster from the Management Console, and select the Data Engineering template from the Cluster Definition dropdown menu.

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



Difference between Tez UI and DAS

Both Tez UI and Data Analytics Studio (DAS) provide a framework to analyze Hive and Hive on Tez queries by providing granular data about query execution, such as Directed Acyclic Graph (DAG) information, DAG swimlanes, visual explain plan, task and task attempt break-down, and so on. Additionally, DAS provides a UI to compose and run Hive queries, compare two queries, query optimization recommendation, and also provides Read/Write and Join reports to understand database usage.

The following table provides a comparison between Tez UI and DAS:

Table 1: Difference between Tez UI and DAS

Comparison factor	Tez UI	DAS
Philosophy	Provides all available data abou a Tez job required to debug a query.	A UI for running and debugging Hive queries.
	Part of the Apache Tez project.	Part of the CDP platform made available under the terms of the GNU Affero General Public License (GNU AGPLv3).
Backend	No dedicated backend for data processing. Depends on YARN Application Timeline Server (ATS) for the data and obtains the progress information from the Tez Application Master (AM).	DAS Event Processor is used for processing data, in addition to the DAS Web App (Java backend) and PostgreSQL database.
Hive	Cannot run or terminate a job.	Can compose, run, and terminate Hive queries.
	Cannot manage a database schema.	Can manage a Hive database schema and can visualize database usage using Read/Write and Join reports.

Comparison factor	Tez UI	DAS
Available data	Following entities are displayed: • Application, Hive query, DAG, vertex, task, task attempts, DAG counters • Hyperlinks interconnecting all these entities for easy drill-down • Detailed tables for all entities • Dedicated **details + stats** page for all these entities • All counters of each entity are displayed in the respective tables	Following entities are displayed: • All Hive query details • Limited DAG and vertex details • No details of Application, task and task attempt • Table only for Hive queries • **Details** page only for Hive queries • Tables are available only for queries and information about counters is not visible
Search and sort	• Supports limited filtering across all query and all DAG tables • Supports substring search and sort on DAGs (under a query), vertex, task and task attempt data tables • Does not support sorting on all queries and all DAG tables	• Supports full text search, but only on all queries table • Supports sorting on query data
Progress information	• Provides real-time progress information for all entities • All pages refresh automatically	• No progress information available • Pages do not refresh automatically
Data download	You can download a zip file containing the following data in JSON format: • Application details • Query details • DAG, vertex, task, and task attempt details You can also download entity-level logs in text format.	You can download a zip file containing the following data in JSON format: • Query details • DAG and vertex details aggregated by DAS • DAG, vertex, task, and task attempt events from Tez • Application attemps information You can also download container logs in text format.