Monitoring Apache Impala

Date published: 2020-11-30
Date modified: 2021-10-29

CLOUDERA

Legal Notice

Contents

Impala lineage

You can use the Atlas lineage graph to understand the source and impact of data and changes to data over time and across all your data.

Atlas collects metadata from Impala to represent the lineage among data assets. The Atlas lineage graph shows the input and output processes that the current entity participated in. Entities are included if they were inputs to processes that lead to the current entity or they are output from processes for which the current entity was an input. Impala processes follow this pattern.

Note that lineage is not updated between a table and views that the table is a part of when an Impala ALTER TABLE operation runs on the table.

Debug Web UI for Query Timeline

For a detailed report on how a query was executed and to understand the detailed performance characteristics of a query, you can use the built-in web server's UI and look at the Gantt chart.

The Gantt chart is an alternative to the profile command and is a graphical display in the WebUI that renders timing information and dependencies. This chart allows you to determine where queries are spending time instead of looking for timing information in the profile. Once you identify which fragments or operators are consuming the most time you may need to look at the summary profile or full profile to drill down further.

In subsequent releases, the timeline will be enhanced to display additional details so that looking at the profiles will not be necessary.

Here is the screenshot of the Gantt chart that shows the Query timeline of a TPC_DS query.
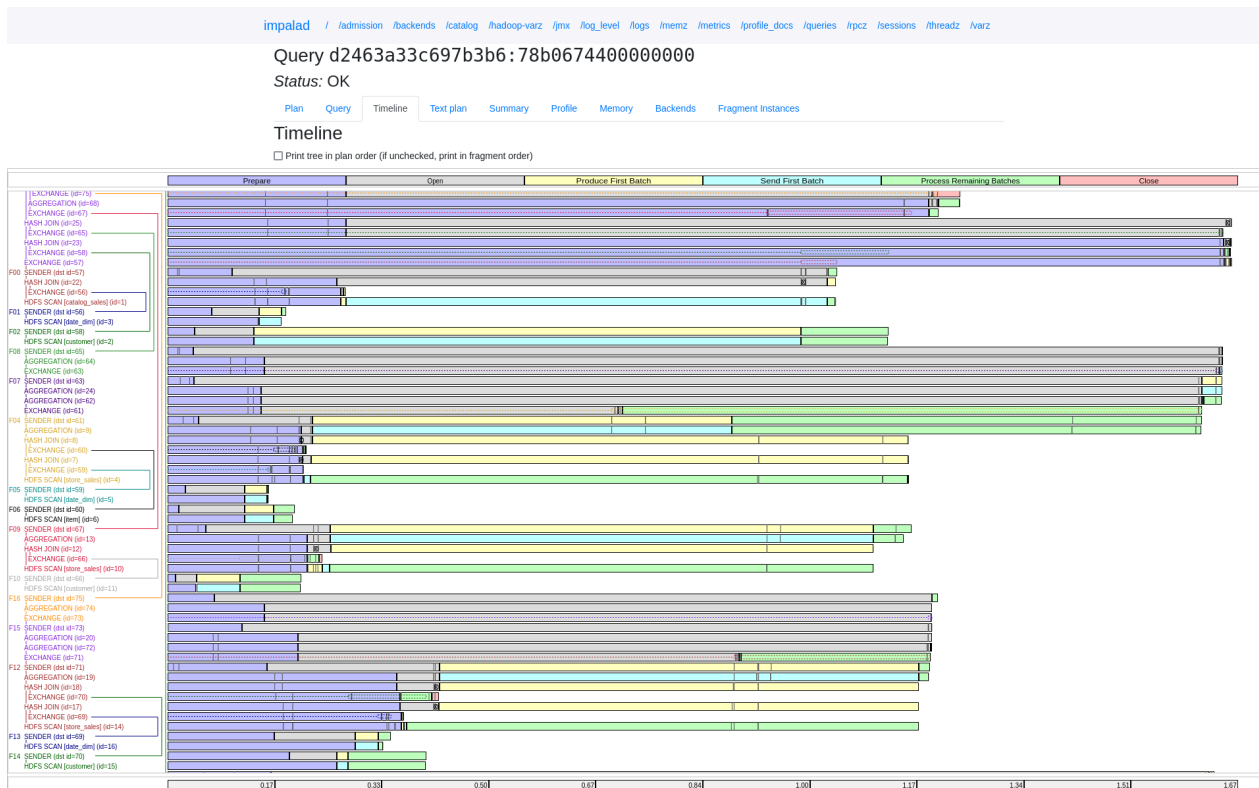
Improvement in Catalog Observability

This release brings significant enhancements to the Impala Catalog Web UI, addressing performance issues related to delays in processing Hive Metastore (HMS) events, which can lead to queries using outdated metadata

These improvements specifically target slow Data Definition Language (DDL) and Data Manipulation Language (DML) operations such as INSERT, LOAD DATA statements, and REFRESH operations. Additionally, it addresses operations stuck in various stages such as startup, among other scenarios. Prior to this release, troubleshooting these issues required analyzing logs, but now the process has been streamlined by enhancing the observability of the catalog across multiple dimensions, including the catalogd WebUI, DDL/DML profiling, and additional log reports on sluggish operations.

Key enhancements include:

- Catalogd WebUI: Utilize the catalogd WebUI to monitor catalog operations.
- DDL/DML Profiling: Monitor DDL/DML profile for detailed execution insights within catalogd.
- Log Reports: Analyze logs for a comprehensive review of slow operations, offering valuable insights into the root causes of performance issues.

These improvements empower users with a more user-friendly and effective means of diagnosing and resolving catalog-related performance challenges.