Cloudera Runtime 7.2.11

Atlas Lineage

Date published: 2019-09-23
Date modified: 2021-09-09

CLOUDERA

Legal Notice

Contents

Lineage overview

Atlas lineage helps you understand the source and impact of data and changes to data over time and across all your data.

Lineage information helps you understand the origin of data and the transformations it may have gone through before arriving in a file or table. In Atlas, if transformations occurred in services that provide process metadata, a lineage graph shows how data in a given column was generated. When a column appears in the output of a process, Atlas reads the content of the process and links the input column or columns to the output asset. This relationship is stored as a vertex in Atlas's graph database. It is displayed as a lineage graph in the details of each entity.
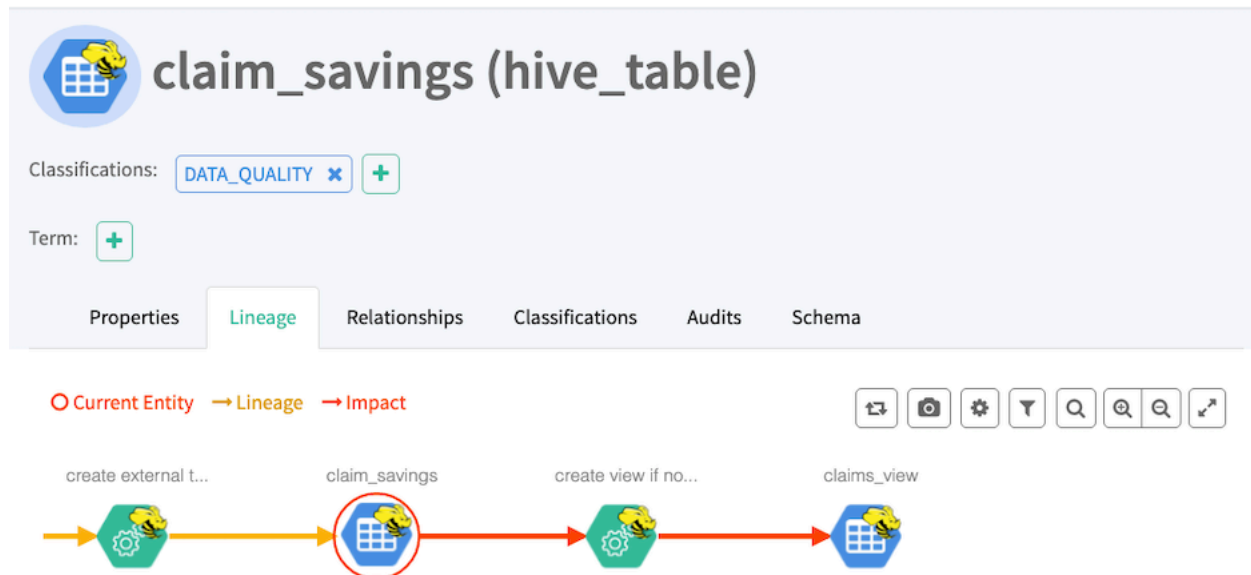
By default, Atlas can collect lineage information from the following sources:

- HiveServer
- Impala
- Spark

The lineage metadata produced by these sources may refer to additional asset entities. For example, when a Hive operation moves data from a Hive table to an HBase table or an HDFS file, Atlas includes an entity to represent the HBase table or HDFS file, and the entity is connected to the Hive table through lineage. The following sources may appear in lineage graphs when referenced:

- HBase
- HDFS
- S3

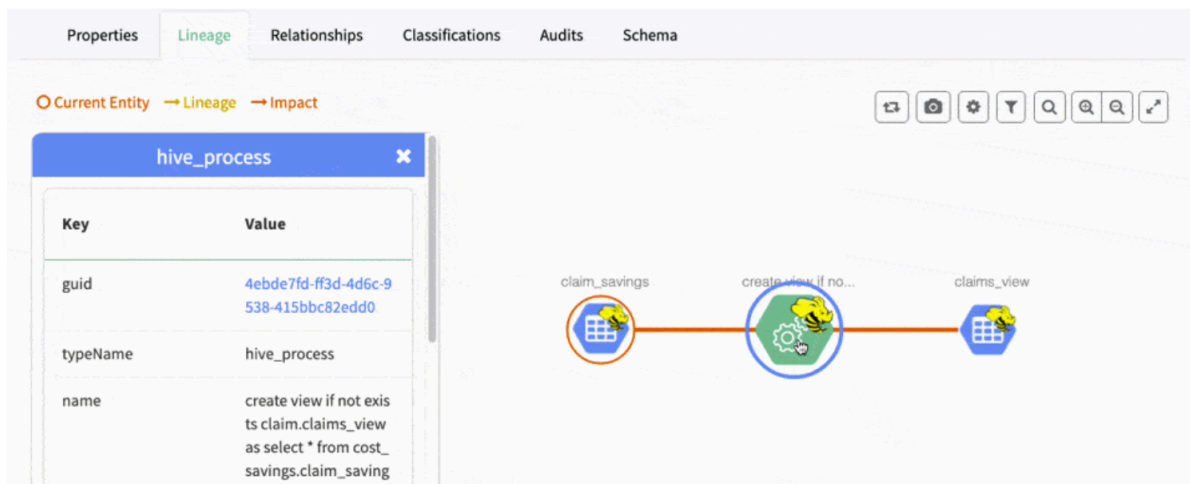Data flow lineage from Cloudera Flow Management (NiFi) can be included as well by configuring the appropriate reporting task.
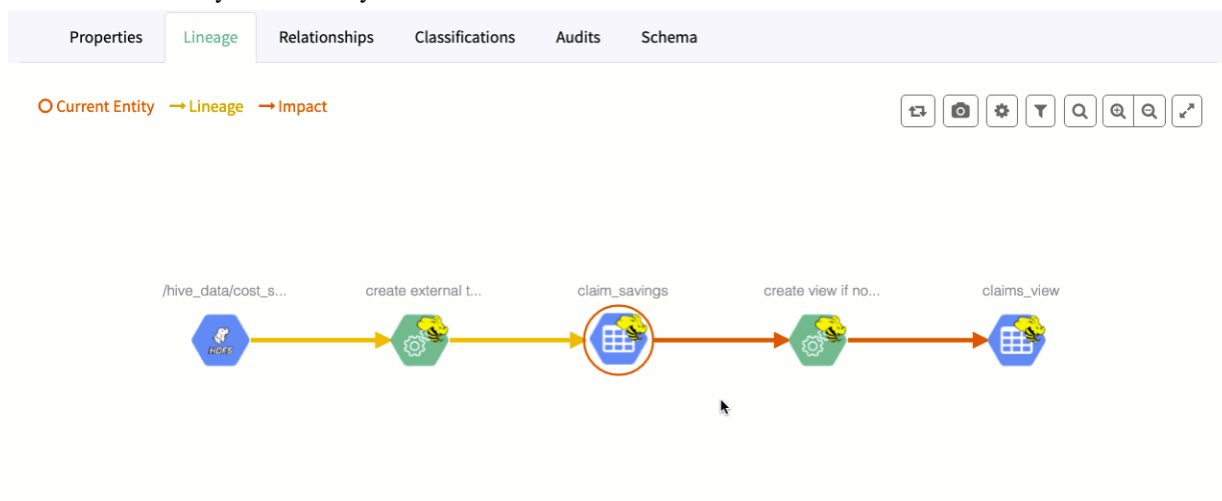
Viewing lineage

Atlas lineage graphs include lots of detail that you can reveal and configure.

Use your cursor to explore a lineage graph:

- Click to show details for an entity

- Hover over an entity to show only one ancestor and descendant

The following symbols can appear in lineage graphs:

Symbols	Name	Description and Actions
	Data Set entity	Represents a column, table, view, file, database, or other physical or logical data asset. While all data set entities are shown as a blue hexagon, the icon may vary based on the type of data asset and the source system.
	Process entity	Represents a query, job, or other process that applied to data assets. While all process entities are shown as a green hexagon, the icon may vary based on the type of process and the source system.

Symbols	Name	Description and Actions
	Current entity	A red circle indicates the current entity in the lineage graph.
	Selected entity	A blue circle indicates the entity selected in the lineage graph.
	Lineage	Connects the current entity with entities that provided input data. Entities connected with yellow lines are ancestors of the current entity.
	Impact	Connects the current entity with entities that could have an impact if data or schema information is changed in the current entity. Entities connected with red lines are descendants of the current entity.

The upper right corner of the lineage picture includes the following controls:

Realign lineage: reset the image to the default scale with the current entity in the center of the tab.

Export to PNG: creates a PNG file of just the lineage picture as it currently appears. Uses the browser file download settings for the file location.

Settings: set cursor display actions.

Filter: hide processes or deleted entities; set the number of lineage steps to show.

Search: find entities in the lineage graph by name.

Zoom In / Zoom Out: scale the lineage graph smaller or larger.

Full Screen: fill the current window with the lineage tab content. Toggle this mode with

Related Information

Propagating classifications through lineage

Lineage lifecycle

Tables are dropped, schemas change, views are created: lineage tracks these changes over time.

Atlas reads the content of the metadata it collects to build relationships among data assets. When Atlas receives query information, it notes the input and output of the query at the column level: Atlas generates a lineage map that traces how data is used and transformed over time. This visualization of data transformations allows governance teams to quickly identify the source of data and to understand the impact of data changes.

Atlas processes contain lineage info; data assets by themselves do not. Impala queries are represented as processes and have lineage information; the data asset affected by Impala queries appear as Hive entities.

HDFS, S3, ADLS files appear when they are referenced by Hive, Impala, or Spark queries; operations that occur on the file system are not reflected in Atlas lineage.

The contents of a lineage graph are determined by what metadata is collected from services. If a process refers to a data asset but Atlas doesn't have an entity for that data asset, Atlas isn't able to create an entity for the process and the lineage defined by that process won't appear in Atlas.

Deleted data assets

Entities that represent data assets that have been deleted (such as after a DROP TABLE command in Hive) are marked as deleted. They show up in search results only if the checkbox to Show historical entities is checked. Deleted entities appear in lineage graph dimmed-out.

Historical entities are never automatically removed or archived from Atlas' metadata. If you find you need to remove specific deleted entities, you can purge specific entities by their GUIDs through REST API calls.

Temporary data assets

Sometimes operations include data assets that are created and then deleted as part of the operation (or as part of a series of operations that occur close together in time). Atlas collects metadata for these temporary objects. The technical metadata for the operation, such as query text, includes a reference to the temporary object; the object itself will show in the Atlas lineage diagrams.

For example, consider a Hive pipeline that writes data to a table, transforms the data and writes it to a second table, then removes the first table. The lineage graph shows the source file, the process that creates the first table, the first table, the process that transforms the data and loads it into the second table, and the second table. Atlas also collects the process where the first table is dropped. When you view the lineage graph, you can choose to show the first table or to exclude it by setting the filter option Hide Deleted Entity.