

Version History PRFAQ

Enabling DataCentral users to easily compare dataset versions to self-service issues.

ARLINGTON - INSIDE AMAZON - March 1st, 2024. Troubleshooting datasets just got easier - dataset owners and consumers can now effortlessly see changes between dataset versions in a single location without leaving DataCentral. The comprehensive, user-friendly interface enables users to quickly see a dif between versions at a glance and a detailed log of the history of a version.

Prior to this launch, one of the most highly requested features from DataCentral users was the ability to compare dataset versions. Dataset owners and consumers had to rely on the Andes API and external tools to understand the changes between versions, when they occurred, and by whom. Users would call the Andes API for multiple versions, copy and paste the SDL response into a third party IDE, and use the third party tool to compare the SDLs of the different versions. In the UI, users would need to manually flip back and forth between multiple versions and do “a bit of an eye test” to try to identify the differences.

“Any time there’s a change in description of column or inline schema change, we don’t see it here. In that scenario, I have to go to endpoint and see who created the version, but can’t find out which team deleted it, or when they deleted it. - Data engineer (real user quote)

“We’re consuming data from other teams, adjusted schema updates, suddenly they’re modified. For me to understand exactly what changed and why, it’s a bit of an eye test where I have to toggle back and forth between versions, different tabs...to understand what’s the new column here. I would like to know why there’s a v7, why is there v8, what’s the difference between the two...I would like to know modifications from one version to the next.” - Sr. Data engineer (real user quote)

Now, all DataCentral users can see the differences between versions without needing to use the API, enabling both technical and non-technical users to self-service issues. Troubleshooting dataset issues is streamlined for dataset owners, consumers, and internal BDT team members by being able to quickly access a dataset’s revision history to understand what happened. New Amazonian BIEs and data engineers are able to onboard faster by quickly understanding the history of their team’s datasets.

“The new version history feature has reduced tickets to BDT by X%,” said an Amazon leader. “It shows exactly what changed between versions in a straightforward format so that even non-technical users can self-service.”

Users can navigate to a dataset version and view the associated “History” tab to see at a glance the differences between the selected version and a previous version, such as added attributes, dropped attributes, and updates to primary and partition keys. Inline changes to a dataset and dataset version are also readily available, such as updates to the support level, column descriptions, and more. Users can see what changed, when, and by whom.

“The new version history feature has reduced my troubleshooting time by Y%. Before, I used to have to use the Andes API and a third party tool to figure out what changed between versions, which was a very time consuming process across multiple versions and datasets. Now, I can quickly see in DataCentral exactly what changed!” - BIE

To try out the new version history feature, visit the “History” tab of any dataset in DataCentral at <https://datacentral.a2z.com>.

FAQ

Which fields are supported by version history?

All editable fields are supported by the version history feature. If it can be changed, it will be displayed.

Who can use this feature?

Any DataCentral user can see and use this feature to view the dataset's changes.

Can I see changes across a range of versions, not just adjacent versions?

P0 will be adjacent versions, but long-term we may support changes across a range of versions.

Can I see changes to the load job query?

P0 will only include dataset-specific information, but long-term we may support change history for load job information.

Will subscribers be notified of version changes?

Notifications will not be included in P0, but could be a future release.

Can you restore/rollback to a previous version?

Probably not - needs more technical deep dive.

Will there be any impacts to terminology?

Possibly - "inline" changes could be referred to as "minor" changes, while changes that result in a new version could be referred to as "major" changes.

Where else could this feature be used?

This feature could replace the existing "Review" page UI when creating a new dataset version. It could also be used to show version history of other resources in DataCentral, such as job profiles.

What are the alternatives?

Currently, the only alternatives are the Andes API or manually toggling between versions on the UI.

How will we measure success?

Success will be measured by feature adoption (# of clicks on History tab, # of unique users, time on History tab) and reduced history-related tickets to BDT.