6.5.24 Summary: Airflow Monthly Virtual Town Hall

Event Details

- Date/Time: June 5th 2024, 8 AM PST/11 AM EST
- **Recording**
  - Passcode: C3keB??6
- Deck

Introduction

The meeting began with Briana introducing the agenda for the event:

- **Airflow 3 Discussion w/ Constance Martineau**

- **PR Demos**

  - **Daniel Standish**:

    - PR #39336: Scheduler to handle incrementing of try_number

  - **Pankaj Koti**:

    - PR #39275: Remove unused index idx_last_scheduling_decision on dag_run table

    - PR #39638: Add indexes on dag_id column in referencing tables to speed up deletion of dag records

- **What we learned after running Airflow on Kubernetes for 2 years with Alexandre Magno Lima Martins**

- **Project Spotlight: Cosmos with Tatiana Al-Chueyr Martins & Daniel Reeves**

- **Community Spotlight with Briana Okyere**

Airflow 3 Update with Constance Martineau

Airflow 3 Update

Constance provided an update on Airflow 3, discussing its current status and the challenges it faced in the past. She highlighted the significant increase in the number of managed services and the

evolving usage patterns of Airflow, which now powers millions of tasks per month and is being used in ways that were not imagined in 2020. She emphasized that the current scale and industry advances made it a suitable time to revisit Airflow 3.

Airflow 3 Development Principles and Goals

Constance outlined the principles and goals for the development of Airflow 3. The main principle is to solidify Airflow as a modern orchestrator with state-of-the-art support for data, AI, and ML workloads, while also ensuring a good security story and a simplified learning curve for new users. The goals include minimizing Dag code changes required for migration, implementing least privilege by default, and aiming for a shorter release cycle, with a target release date in March 2023. The team is also considering whether to release one more feature for Airflow 2 before moving to Airflow 3. Constance encouraged participation in the development calls to provide feedback and suggestions.

Development Calls, Prioritized Features, and Airflow Improvements

Constance encouraged everyone to participate in the development calls to voice their opinions on the prioritized features for the 3.0 version, emphasizing that the team is particularly interested in foundational or break-changing features. She also promised to publish the meeting agendas and notes in Confluence for those unable to attend the calls. Briana, the QA lead, merged QA sessions with the end of each call. Daniel Standish then presented on a problem with the "try number" system in Airflow, explaining the previous issues with incrementing and decrementing values, and noted that future improvements could include a separate record for each task instance.

PR Demos

Daniel Standish on PR #39336: [Scheduler to handle incrementing of try_number](#)

Daniel Standish then presented on a problem with the "try number" system in Airflow, explaining the previous issues with incrementing and decrementing values, and noted that future improvements could include a separate record for each task instance. Daniel Standish discussed the concept of DAG versioning in relation to maintaining stable try numbers.

Pankaj Koti:

- **PR #39275: [Remove unused index idx_last_scheduling_decision on dag_run table](#)**
- **PR #39638: [Add indexes on dag_id column in referencing tables to speed up deletion of dag records](#)**

Briana then invited Pankaj to speak about his two PRs. Pankaj, part of the open source engineering team at Astronomer, discussed a project aimed at optimizing database performance by identifying and removing unused indexes and adding missing ones. This PR involved removing an unused index, which not only saved space but also improved operation speeds.

[What we learned after running Airflow on Kubernetes for 2 years](#) with Alexandre Magno Lima Martins

Alexandre then presented on the experience of running Airflow on Kubernetes for over two years, emphasizing its challenges and learnings. He highlighted the platform's ability to scale, its use across various teams, and the company's increased pull request activity in Q1 and Q2. However, he also noted issues encountered due to the choice of the Q&A executor in Airflow, particularly with task scheduling as the number of tasks grew, leading to a peak in scheduled tasks.

Alexandre discussed the team's challenges with using Kubernetes for their Dbt jobs, focusing on issues with scaling nodes and short-lived pods. They switched to using Docker Salary as their executor, which initially resolved their problems but led to new issues with Salary workers dying due to out-of-memory issues. Alexandre identified the root cause of this problem as high concurrency levels resulting in memory leakage and proposed solutions to mitigate this, such as limiting the number of tasks a worker process can execute and controlling the maximum amount of memory it can use. He emphasized the importance of the termination grace period in Kubernetes in preventing long-living Salary workers from being terminated during node rotation.

Project Spotlight: Cosmos with Tatiana Al-Chueyr Martins & Daniel Reeves

Tatiana and Daniel, primary contributors to the project, introduced themselves and the benefits of using Airflow and Vivica. Daniel then detailed the new features and improvements in version 1.4 of the Cosmos software, hinting at further enhancements in version 1.5. The team also emphasized the importance of community feedback and encouraged participation in the upcoming Airflow Summit and various related groups. Briana emphasized that any administrative questions should be posted in the Town Hall chat or directly to her.

Resources/Links
- [Join the Dev list](#)
- [Join Airflow Slack](#)
- [Subscribe to the Newsletter](#)
- [Astronomer Champions Program for Apache Airflow](#)
- [Join Airflow Meetup](#)
- [Airflow Summit 2024](#)
- [Join Airflow 3 devlist calls](#)
- [Scheduler to handle incrementing of try_number](#)

- [Remove unused index idx_last_scheduling_decision on dag_run table](#)
- [Add indexes on dag_id column in referencing tables to speed up deletion of dag records](#)
- [What we learned after running Airflow on Kubernetes for 2 years](#)
- [dbt partial parsing](#)
- [InvocationMode.DBT_RUNNER](#)
- https://github.com/astronomer/astronomer-cosmos
- [Survey on dbt in Cosmos](#)
- https://astronomer.github.io/astronomer-cosmos/