Build your open data lakehouse on Iceberg with Fivetran and Dremio - Q&A

The data lakehouse is quickly emerging as the ideal data architecture because it combines the flexibility and scalability of data lakes with the data management, data governance, and data analytics capabilities of data warehouses. Table formats bring many of the "house" features to the data lakehouse. Apache Iceberg is a truly open table format that is built for easy management and high performance analytics on the largest data volumes in the world.

We'll discuss:

- Why open table formats are fundamental to building a data lakehouse
- How Fivetran automates data movement and helps organizations easily move data from various sources to their Amazon S3 data lake in Apache Iceberg tables.
- How Dremio & Fivetran simplify your data lakehouse architecture while providing high performance and ease of use.

Coral Trivedi
Destination
Product Manager
Fivetran

Alex Merced
Developer
Advocate
Dremio

Brett Roberts
Principal Partner
Solutions Architect
Dremio

Webinar Q&A

1. **Can Fivetran handle transformation after the data lands in s3….say from raw to cleansed?**
   Answer:
   Coral - Yes, actually have a whole Transformations feature. We work with DBT and transformations is something that we offer as a completely standalone feature and functionality. We can do that with an S3 as well as any of the other destinations that we landed in.

Brett - We also work with DBT as well so it's kind of awesome because wherever you are within the pipe you can leverage DBT for those Transformations whether it's when it's landed in S3 with Dremio or with Fivetran.

2. **Is Fivetran adding support for Dremio Arctic catalogs?**
   Answer:
   Coral - At the moment, we aren't offering support right now. What we've built out is support for the open source format so Iceberg and on our roadmap for this year, we're going to be supporting Delta format as well. We're keeping kind of Partners in mind, while we do those but right now our focus is those two open source formats.

3. **Does Fivetran support the Iceberg REST catalog?**
   Answer:
   Coral - We do but we're leveraging uh I think what we're doing is we're leveraging AWS glue specifically so we're not relying on the Iceberg REST catalog.

4. **How does fivetran handle incremental transfer of data from new data from the source?**
   Answer:
   Coral - You can basically set your sync frequency. We discover and do data at the source, we'll sync that in conjunction with your incremental data. Initially the way that it works, you'll plug-in your destination, identify the source, we'll do what's called a historical sync, we'll bring all the data over and then based on the increment that is set, we'll sync new data into the destination. That includes smaller chunks of data as well as new tables but basically it happens at set increments.

5. **How does Fivetran avoid the “small files” problem in Iceberg tables?**
   Answer:
   Coral - it's related to kind of how we write the data. Specifically, kind of what we decide to do on our end in terms of reading data at the destination. What happens with an RS3 Pipeline and then how we landed so that's fundamentally kind of the inner workings of Fivetran and how we land data into S3.

6. **My question is how do you identify new data in sources, including deletes at the source? Versioning required at the source? Reading logs?**
   Answer:
   Coral - We like to compare. Basically, to really simplify - we're looking at the data, at the source. We know what we've transferred over to the destination and regardless of the connector the destination we're doing that sort of comparison and so it really depends on the connector.

The source connector that you're using to land data in S3, the mechanism will vary slightly but fundamentally we're either keeping track of the primary keys and making sure that we identify both new data and remove data or we're doing something you know on our own to kind of figure out how data has changed as it moves through the pipeline. We're monitoring and keeping track of what's happening at the source. At the destination, it's temporary tables and that data is deleted after 30 days so we're not continuously monitoring but for our purposes to make sure we're syncing the right data and landing the right data, we're keeping track.

7. **How do you handle "bought" data? Like Dow Jones**

   Answer:

   Coral - We have integrations with some of our partners. Typically, those Partners will leverage Fivetran, like you can purchase this data and land it directly into a destination. If you have for example like Dow Jones data and you're putting it into big query or something like that you just leverage our big query connector. It's not necessarily where a situation where we will have a Dow Jones connector specifically for data from There. I imagine that even the Dow Jones is like landing data within a destination and then that's where you would be pulling that data from. So there's like an integration that might exist there but just an example.

8. **Where do I see what "bought" data is handled?**

   Answer:

   Coral - folks selling the data themself so in this example would be the Dow Jones. Fivetran isn't necessarily selling the data. What we're doing is just facilitating the data movement so if you are a partner who's looking to sell data to customers, we have solutions for that. If you're a customer who's looking to move data within an ecosystem, that's also kind of where you would leverage Fivetran. In terms of purchasing external sets of data that would be a different.

9. **How do you handle 10's PB of data from Hadoop sources?**

   Answer:

   Coral - we have our incremental and our historical sync and so initially to sync that amount of data it would take some time. That effects is obviously very large volume but fundamentally it would work in the same way where we would sync the data historically uh once we have all the historical data into the source um and we would think incrementally as well uh I don't believe we have Hadoop as a source Sprinter.

   Brett - one of the things that you guys can do is those sources that would traditionally feed into hdfs or into Hadoop those those sources can now be moved to your S3 destination instead of Hadoop um so that could be you know whatever the origin of

source is could be pointed to S3 instead of Hadoop and then obviously there's probably a data migration that needs to be done for historical data as well

Coral - if that is the case, please definitely reach out to Brenner and I. We're starting to think about as well the kind of like how to facilitate these exact types of migrations in a way that's like painless. What we went through is if you have S3 and you don't have a ton of complex pipelines, it's really easy to kind of start to leverage Fivetran and AWS. Obviously, if you have like a very mature system, it's going to be a little bit more thought and planning required and so that's something that we'd be happy to help.