The Who, What and Why of Data Lakehouse Table Formats - Q&A

In the rapidly evolving landscape of big data, Data Lakehouse is heralding a new age of unified analytics, blending the best elements of data lakes and data warehouses. Central to this convergence is the need for advanced table formats that can meet the demands of scalability, performance, and data reliability. This webinar dives deep into the world of Data Lakehouse table formats, specifically focusing on Apache Iceberg, Delta Lake, and Apache Hudi.

Who should attend?

Data engineers, data architects, data analysts, and other professionals interested in modernizing their data platform or seeking deeper insights into the technicalities and advantages of these advanced table formats.

Key Takeaways:

- Introduction to Data Lakehouse: Explore the genesis of the Data Lakehouse paradigm, its significance, and how it's reshaping the way organizations think about big data storage and analytics.
- Demystifying Apache Iceberg, Delta Lake, and Apache Hudi: Understand the intricacies of these popular table formats, their architectural nuances, and how they differ from traditional table structures.
- Features Spotlight: Delve into the unique feature sets that each format brings to the table - from ACID transactions, time-travel queries, to efficient upserts and scalability features.
- The Relevance Quotient: Understand why these table formats matter in today's data-driven world. Learn about their roles in ensuring data consistency, improving query performance, and facilitating near real-time analytics on large datasets.
- Best Practices and Use Cases: Explore real-world scenarios where organizations have leveraged these formats to transform their data analytics operations, and glean best practices for successful implementation and optimization.

Join us to uncover the intricate dance of modern table formats that are at the heart of the Data Lakehouse revolution. Equip yourself with the knowledge to harness their power, ensuring a robust and efficient data infrastructure for your organization.

Alex Merced
Developer Advocate
Dremio

Show's Q&A

1. **How to create indexes for making queries fast?**
   Answer: The metadata for these formats often act as the indexes for efficient queries, but good data management can go a long way in maximizing performance in regular compaction and strategic clustering.

2. **Difference Apache Paimon vs iceberg?**
   Answer: Not as familiar with Apache Paimon, as it is a newer entrant in the space.

3. **Iceberg fans often say that Databricks maintains control over Delta Lake, even though it's open sourced under the Linux foundation. Have you seen any empirical/objective evidence of this?**
   Answer: Read this article

4. **Grafana visualization does not have Dremio datasource. Is Dremio going to build a datasource plugin for the same?**
   Answer: Please reach out to your Dremio Account reps and account executives about integrations you need or want.

5. **In iceberg table format, if there are multiple DML Operations , can I assume the metadata files increase and cause the reads to slow down ?**
   Answer: concurrent transactions shouldn't effect their performance due to snapshot isolation, although large metadata files cause it's a wide column can effect performance and you can configure the metadata collection on the table to only collect on relevant columns.

6. **Once an iceberg table format is created by an engine like Dremio, then only by using the Dremio catalog can we consume data ? other catalogs can not understand?**

Answer: Nope, any engine that supports Iceberg and the particular Iceberg catalog your using can read your tables

7. **Can clustering be done on multiple table columns?**
   Answer: Yes

8. **What is the best way to load incremental data to iceberg tables?**
   Answer: Using Iceberg's CDC feature

9. **Can you please let us know the tools/ Dremio built in options for this?**
   Answer: Once a Change Log is created, any engine can run a merge transaction to merge those changes