GNARLY Data_Waves

PRESENTED BY dremio

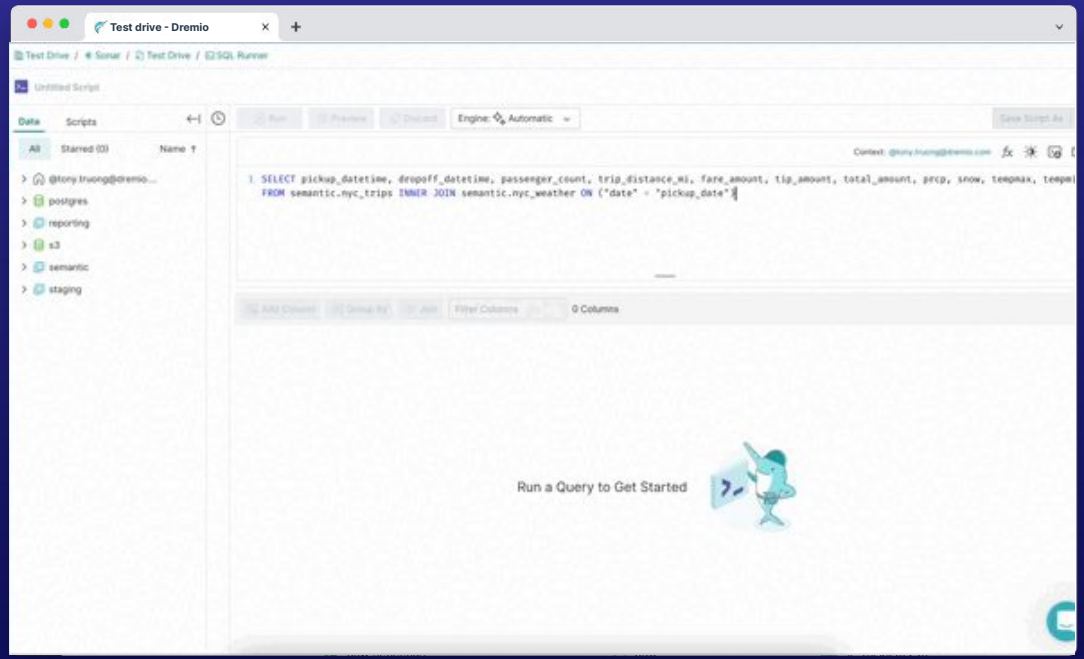
EPISODE 31

ETL, ELT and the Dremio Lakhouse

Experience the data lakehouse with Dremio Test Drive

- Sub-second query on 1 billion rows of data joining Amazon S3 with a Postgres database
- Connect to Tableau or Power BI and build a dashboard with this dataset
- Everything hosted by Dremio - 100% free for you

Start Test Drive



GNARLY Data_Waves

BIGDATA & AI by Xcorp PARIS

TIME TO ACCELERATE

September 25 & 26, 2023

Paris Convention Center

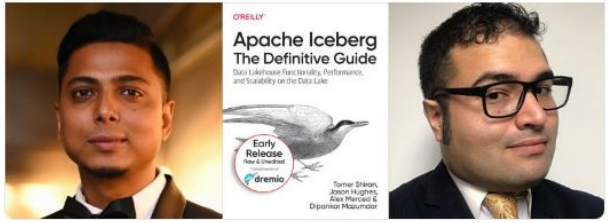
dremio

Coalesce by dbt

Oct 16-20, 2023
Hilton Bayfront San Diego

Data Day Texas
930 followers
13h ·

Yes! O'Reilly co-authors Dipankar Mazumdar and Alex Merced will be co-hosting an Apache Iceberg: Ask Me Anything session at the upcoming Data Day Texas. They'll holding office hours as well. It's free consulting!
https://lnkd.in/g3bQcfxx
This is the year you don't want to miss. Early Bird tickets still available.
#iceberg #datalake #dataengineering Dremio



ETL, ELT and the Dremio Lakhouse

Presented by Alex Merced

Alex Merced

Developer Advocate, Dremio

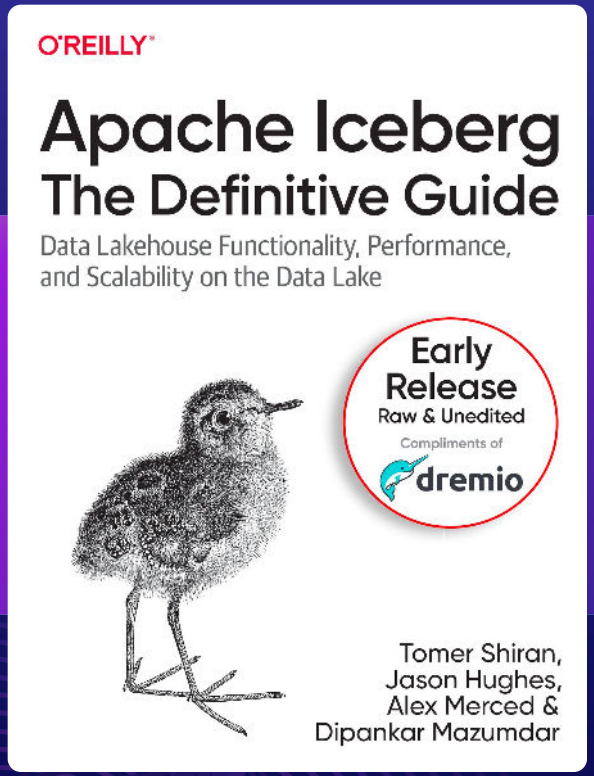
Alex Merced is a developer advocate for Dremio, a developer, and a seasoned instructor with a rich professional background. Having worked with companies like GenEd Systems, Crossfield Digital, CampusGuard, and General Assembly.

Alex is a co-author of the O'Reilly Book "Apache Iceberg: The Definitive Guide." With a deep understanding of the subject matter, Alex has shared his insights as a speaker at events including Data Day Texas, OSA Con, P99Conf and Data Council.

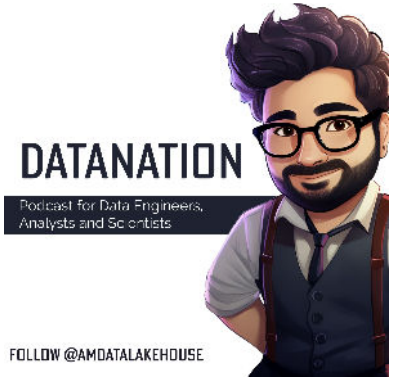
Driven by a profound passion for technology, Alex has been instrumental in disseminating his knowledge through various platforms. His tech content can be found in blogs, videos, and his podcasts, Datanation and Web Dev 101.

Moreover, Alex Merced has made contributions to the JavaScript and Python communities by developing a range of libraries. Notable examples include SencilloDB, CoquitoJS, and dremio-simple-query, among others.

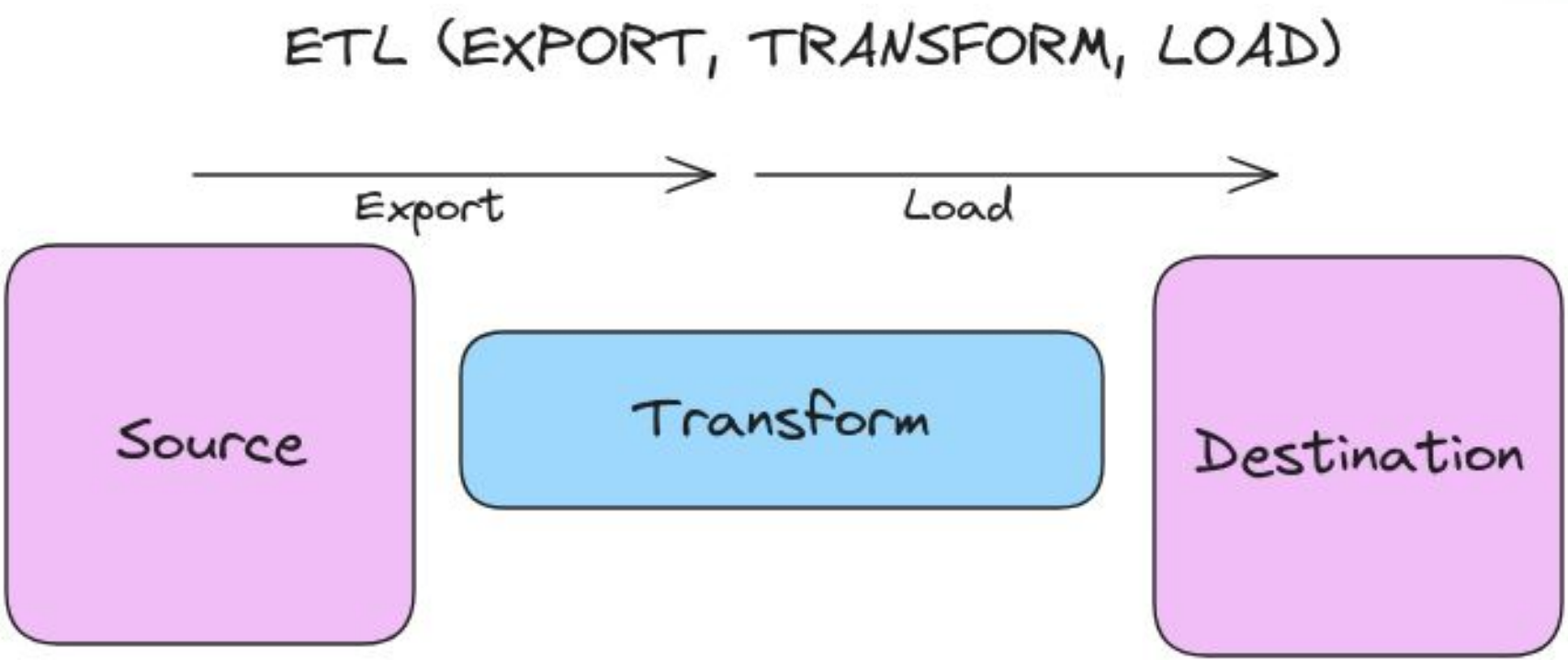
Apache Iceberg: The Definitive Guide



Podcasts



Subscribe on Spotify/iTunes


ETL (EXPORT, TRANSFORM, LOAD)
Export
Load
Source
Transform
Destination

Pros of ETL:

- **Data Quality Control:** ETL processes often include data cleaning and validation, ensuring higher data quality before it's loaded into the warehouse.
- **Performance:** Data transformation occurs before loading, which can lead to better query performance, especially when dealing with large datasets.
- **Structured Data:** ETL is well-suited for transforming data into structured, consistent formats, making it easier for downstream applications.

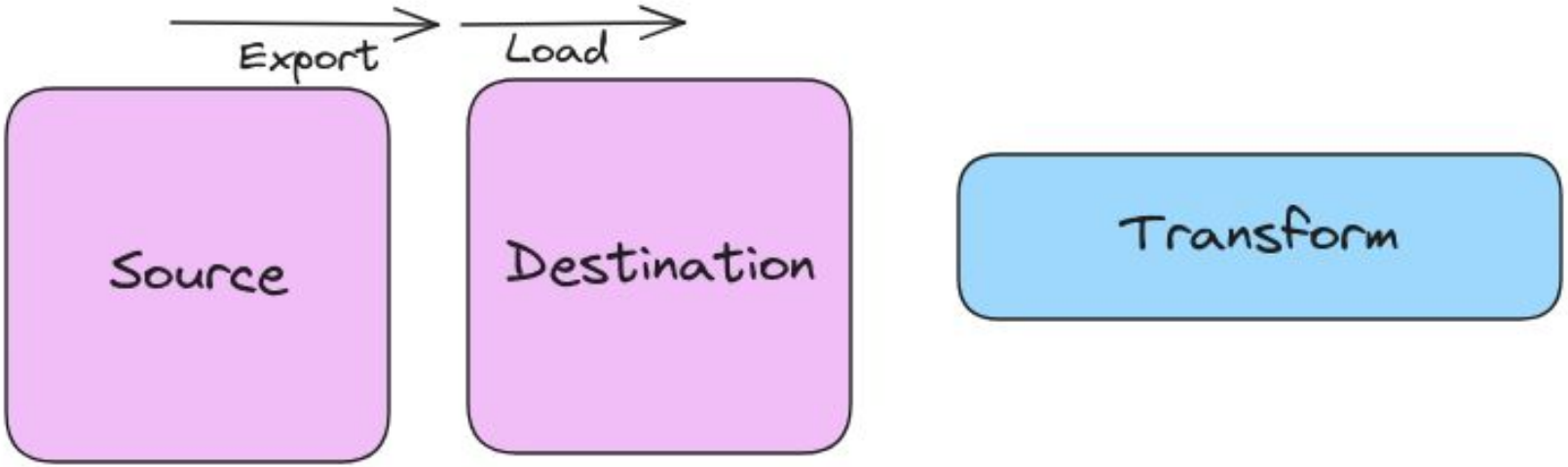
Cons of ETL:

- **Latency:** ETL processes can introduce latency since data transformation occurs before loading. Real-time data analysis is challenging.
- **Complexity:** Designing and maintaining ETL pipelines can be complex and time-consuming, especially as data sources evolve.
- **Scalability:** Scaling ETL pipelines can be expensive, as the infrastructure must handle both data extraction and transformation simultaneously.

Best Practices for ETL:

- **Data Profiling:** Profile source data to understand its quality and structure before designing ETL processes.
- **Incremental Loading:** Implement mechanisms for incremental data loading to minimize latency.
- **Monitoring and Logging:** Establish robust monitoring and logging to detect and troubleshoot ETL failures.
- **Data Lineage:** Document data lineage to understand the flow of data through ETL pipelines.

ELT (EXPORT, TRANSFORM, LOAD)



Pros of ELT:

- **Flexibility:** ELT allows you to store raw data, providing more flexibility for future analyses.
- **Scalability:** ELT can easily scale by adding more compute resources to the data warehouse.
- **Real-Time Analysis:** Data can be available for analysis immediately after extraction.

Cons of ELT:

- **Data Quality:** Data quality issues may persist since transformation occurs after loading.
- **Performance:** Query performance can be slower for complex transformations within the data warehouse.

Best Practices for ELT:

- **Schema-on-Read:** Implement a schema-on-read approach, where data is structured as needed during analysis.
- **Metadata Management:** Maintain metadata to track data transformations and ensure data lineage.
- **Data Governance:** Establish data governance practices to monitor and improve data quality post-loading.


Cloud Costs
Storage
Compute
Access

If Destination is a data warehouse

- Storage of raw data and transformed data
= more premium warehouse storage costs

- Cost of Transform compute using more expensive warehouse compute

If Destination is a data lakehouse

- raw data and transformed data exist on lower cost data lake
- Use lower cost lakehouse compute to transform when necessary

If Destination is a Dremio

- Data movement and duplication may not be necessary
- Transforms can be done via logical views
- Data reflections can be turned on when more performance is needed

Summary of ELT & ETL with Dremio

- Zero-copy logical view first architecture
- Less data movement, lower compute and storage costs
- Columnar Cloud Cache, lower cloud access costs
- Data reflections create reusable synced physical representations when needed
- Can denormalize among disparate sources
- Semantic Layer allows you to document, govern and monitor data curation
- Dremio Arctic catalog enables zero-copy cloning via branching

Alex Merced
follow @amdatalakehouse