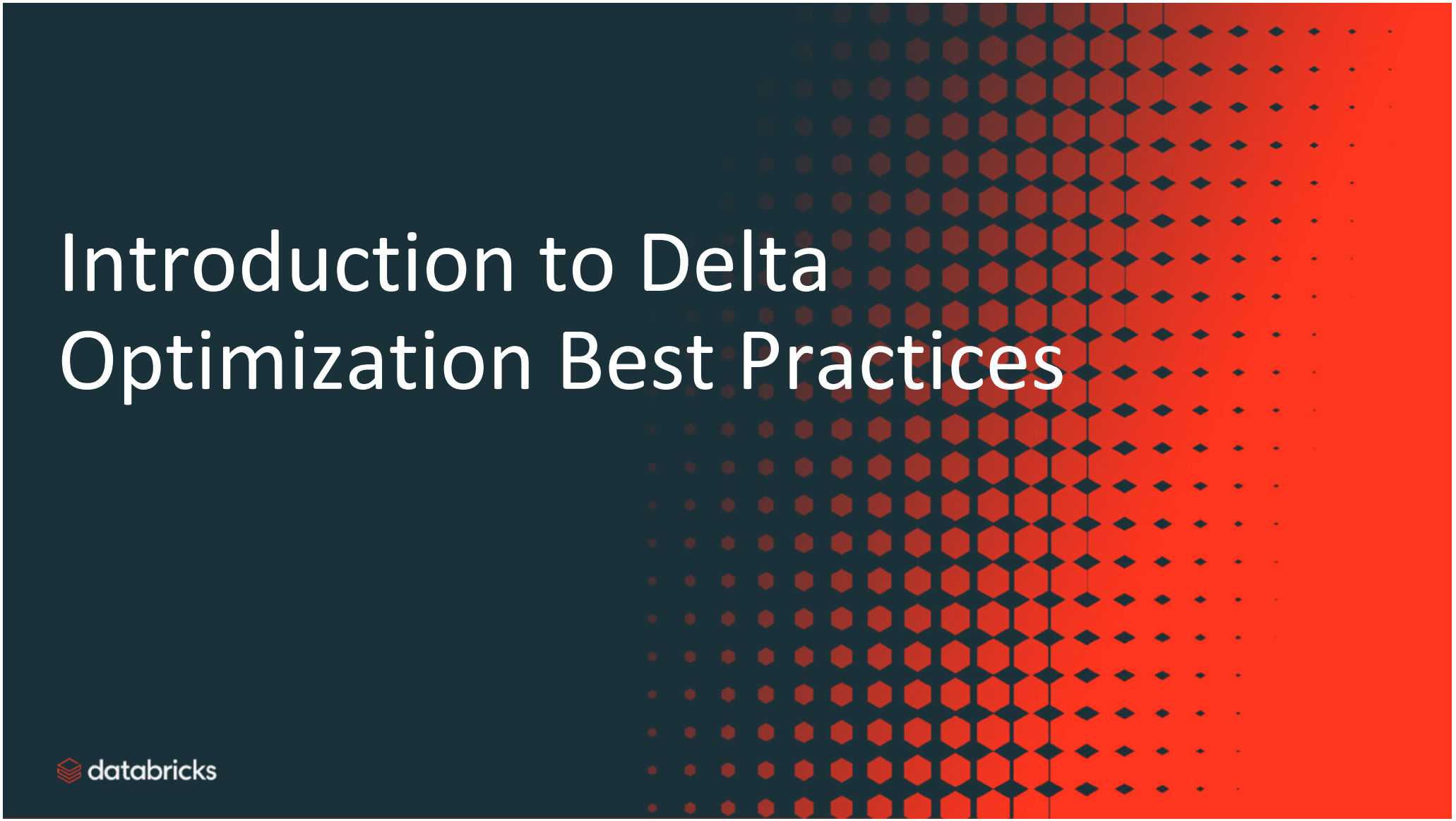

databricks

Introduction to Delta Optimization Best Practices



The Emergence of Data Lakes

Really cheap durable storage.
High durability. Cheap. Infinite scale.

Store all types of raw data.
Video, audio, text, structured, unstructured

Open standardized formats.
Parquet format, big ecosystem of tools operate directly on these file formats.

Challenges with data lakes

Hard to append data. Adding newly arrived data leads to incorrect reads**.**

Modification of existing data difficult. GDPR/CCPA requires making fine grained changes to existing data lake.

Jobs failing mid way. Half of the data appears in the data lake, the rest is missing.

Challenges with data lakes

Real-time operations hard – mixing streaming and batch leads to inconsistency.

Costly to keep historical versions of the data – regulated environments require reproducibility, auditing, and governance.

Difficult to handle large metadata – for large data lakes the metadata itself becomes difficult to manage.

Challenges with data lakes

"Too many files" problems. Data lakes not great at handling millions of small files.

Fine grained access control difficult. Enforcing enterprise-wide role-based access control on data difficult.

Challenges with data lakes

Hard to get great performance – partitioning the data for performance error-prone and difficult to change.

Data quality issues. Hard to ensure that all the data is correct and has the right quality.