Delta Lake

ACID tranzakciók
Big Data rendszerekben

Tóth Zoltán
CTO Datapao

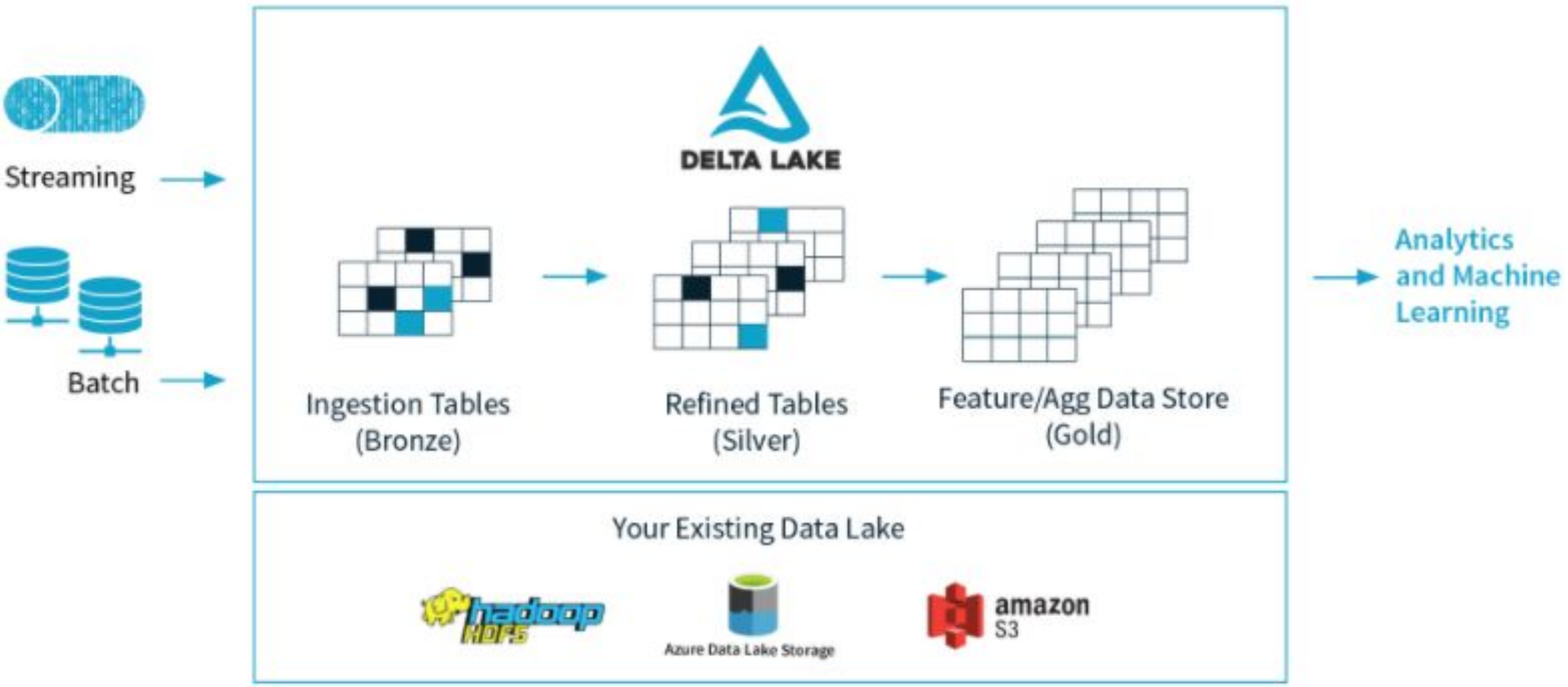
Budapest Data - 2020 szeptember 23.


I'M SORRY
>_<

Bemutatkozás - Tóth Zoltán

Delta Lake - delta.io

Delta Lake is an open-source storage layer that brings ACID transactions to Apache Spark™ and big data workloads.



ACID

Atomicity	• File alapú transaction log • Cloud storage/HDFS file írás atomikus
Consistency	• Az adatbázis tartha a definiált invariánsokat • *C in ACID was “tossed in to make the acronym work”**
Isolation	• Konkurrens írások nem interferálnak • Megoldás: File alapú, szekvenciális transaction log (atomic) olvasása/írása
Durability	• Nem veszik el az adat • A Cloud storage/HDFS garantálja

*Martin Kleppmann: Designing Data Intensive Applications, page 225

További feature-ök

Schema Evolution

- NOT NULL → NULLABLE
- Új oszlop hozzáadása
- Típusok “upcast”-olása (pl Int -> Long)

*Martin Kleppmann: Designing Data Intensive Applications, page 225

További feature-ök

Adat-változtatás

- DELETE
- UPDATE
- UPSERT/MERGE

*Martin Kleppmann: Designing Data Intensive Applications, page 225

További feature-ök

Time travel

A delta tábla korábbi snapshotjain is könnyen lehet dolgozni.

*Martin Kleppmann: Designing Data Intensive Applications, page 225

Bemutatkozás

DEMO

További feature-ök

Mikor használd?

- Spark 3-at használsz, és ...:
  - Konkurrens írásra van szükséged
  - UPDATE/DELETE/MERGE műveleteket végzel

Mikor ne használd?

- Nem Sparkot használsz (= nincs Delta támogatás)
- Nincs sok adatod és egy fájl/SQLLite épp elég
- Egy “writer” ír egyszerre appenddel/overwrite-tal mindig

  (ilyenkor használj parquet-t, ami sokkal gyorsabb)

*Martin Kleppmann: Designing Data Intensive Applications, page 225

Köszönöm!

DATA PAO

Contact

Tóth Zoltán

zoltan@datapao.com

+36 30 291 3599

LinkedIn: zoltanctoth

Demo Notebook: https://datapao.blob.core.windows.net/datapao-public/Delta Demo.dbc