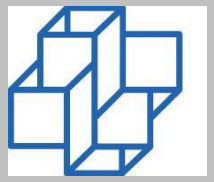
Escola de Verão LNCC 2022
Jornada em Ciência de Dados

Delta Lake - Camada de Armazenamento

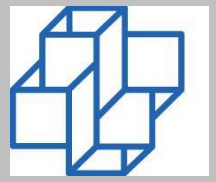
Fabio Porto (fporto@lncc.br)
Gustavo Decarlo (gdecarlo@posgrad.lncc.br)
LNCC – CCC - DEXL Lab
http://dexl.lncc.br

Apresentação



- Sou aluno da pós graduação do LNCC no programa de Mestrado
- Tenho mais de 6 anos trabalhando no mercado com Engenharia de dados, passando pela Globo, Stone e atualmente estou na Thoughtworks

Sumário

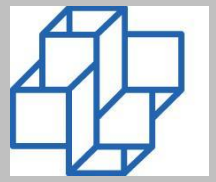


- Cenário de Big Data
- Sobre o Delta Lake
- Principais contribuições na indústria e no cenário de Big Data
- O Delta Lake está sozinho?
- Vamos as mãos na massa

Cenário de Big Data

Como era antes?

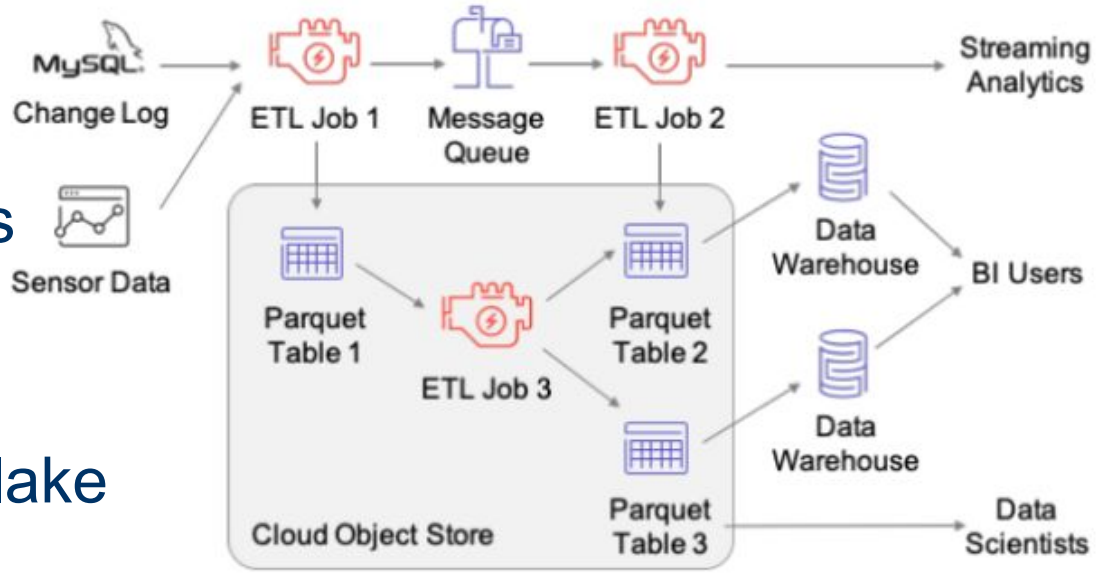
Cenário de Big Data - Sobre Datalake



- Proposta com armazenamento distribuído de dados
- Possibilitar ter vários tipos de formatos de dados, como estruturados, semi-estruturados e não estruturados
- Atomização dos dados, não possibilitando de forma fácil updates, upserts e merge dos dados
- Tradicionais data lakes usam o Hadoop com o HDFS como seu repositório
- Outro ponto importante permitir ter o data lake com diversos tipos de arquivos, como csv, json, parquets, orc e etc.

Cenário de Big Data

- Atualização contínua dos dados
- Problema com esquemas
- Particionamento dos dados
- Gestão dos dados
- Separação do data lake do warehouse
- Leis de proteção aos dados.

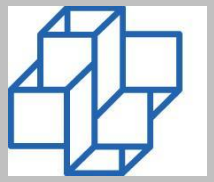


(a) Pipeline using separate storage systems.

DEXL
DATA EXTREME LAB

Sobre o Delta Lake

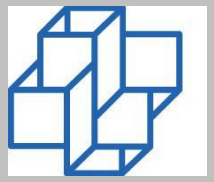
O Foco e um pouco de história



- Objetivos:
  - Ser uma camada de tabelas em cima de storages da nuvem* com operações ACID (Atomicity, Consistency, Reliability e Durability). Como foco inicial.
  - Dotar os usuários de poderem atualizar múltiplos objetos de uma só vez, e permitir alta escalabilidade de leitura e escrita desses objetos.
  - Ter um fácil acesso e gestão dos metadados das tabelas.
- Começou a ser oferecido em 2017 de forma comercial pela Databricks e 2019 como um projeto Open Source.

* Não somente nuvem, temos para o HDFS também.

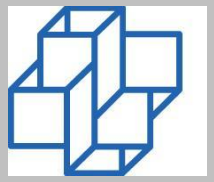
Principais Características



- Time travel
- SQL DDL, como: CREATE, ALTER
- SQL DML, como: UPSERT, DELETE e MERGE.
- Eficiência no Streaming I/O
- Caching
- Data layout optimization
- Evolução do esquema
- Log de auditoria

Principais contribuições na indústria e cenário de Big Data

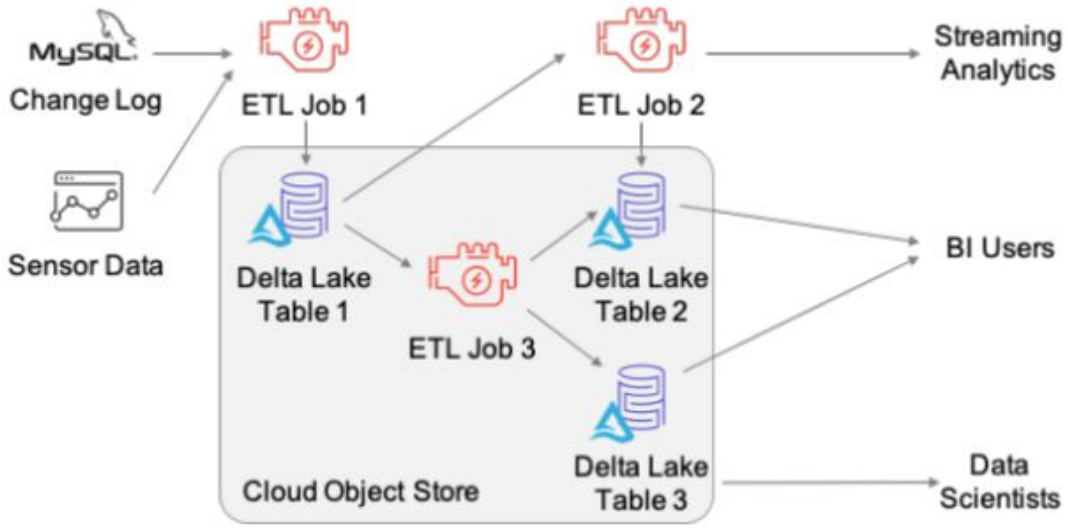
Contribuições



- Umas das principais é o custo da redundância dos dados em diferentes data warehouses e sistemas de storage.
- Outro ponto traz uma abstração importante na gestão dos dados e transparência das operações nos dados com uma camada de metadados. Esse ponto de governança no mercado é complexo, por conta das leis de proteção aos dados.

Contribuições

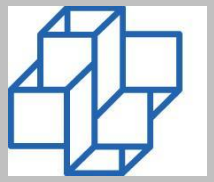
- Reduz a complexidade da arquitetura da solução das plataformas de dados. Isso para a indústria é crucial, pois é tempo e investimento.



(b) Using Delta Lake for both stream and table storage.

O Delta Lake está sozinho?

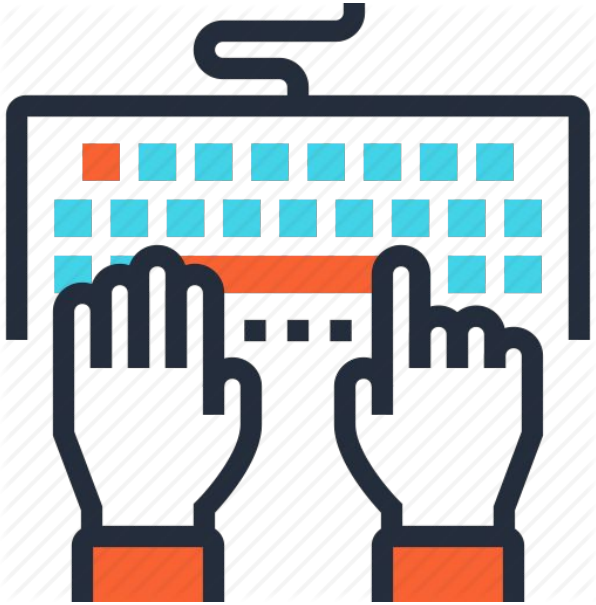
O Delta Lake não está sozinho



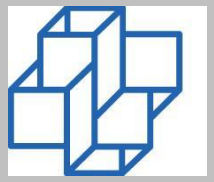
- Apache Iceberg - Projeto nasceu no Netflix e similar ao Delta Lake em ser um storage layer.
- Apache Hudi - projeto também similar ao Delta Lake e ao Apache Iceberg

* https://iceberg.apache.org/
* https://hudi.apache.org/

Vamos as mãos na massa



Link: https://colab.research.google.com/drive/1QLIu4sDNQdXfji9vDC4t70qPeod2ehb3



Obrigado!
Perguntas?