ECBS 5334 - Data Engineering 4
Big Data Computing with Apache Spark

Zoltan C. Toth
tothz@ceu.edu

https://www.linkedin.com/in/zoltanctoth/

CEU CENTRAL EUROPEAN UNIVERSITY

Schedule

- We are working from 13:30 - 19:20 CET

Agenda

This week

1. Big Data Computing - History overview
2. Apache Spark Intro + Databricks Workspace Setup
3. Apache Spark - Data Analytics Basics

Next week

4. More Apache Spark Data Analytics
5. A bit of Spark internals (optional)

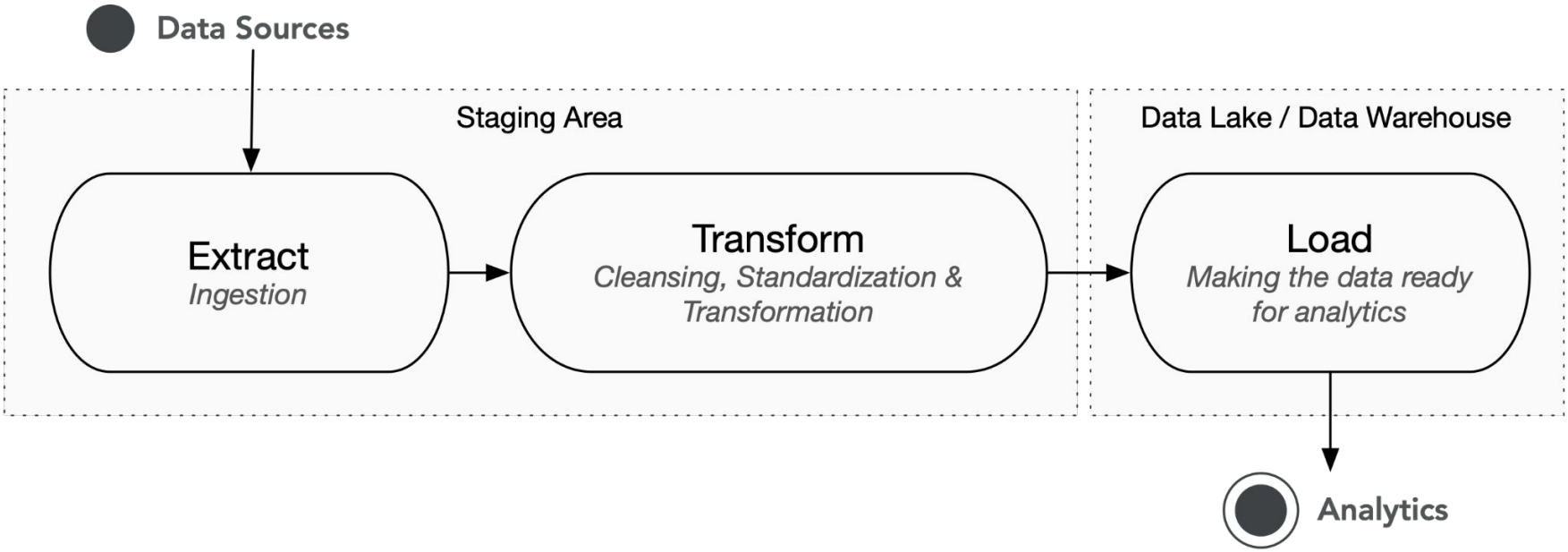
To pass this course

1. Don’t miss more than 25% of the sessions
2. There will be a single deliverable (A Databricks Notebook), deadline **2 Jun 2023 23:59.**
3. Grading will be based on instructions provided for you later.

Traditional Data Warehousing problems

- Data in a Data warehouse is expensive:
  - Licensing
  - Hardware / Operations
- Alternative: Working off plain files (CSV, JSON, …):
  - Slow
  - ACID problems: Concurrent read and write, consistency, …
- OLAP design doesn't scale well

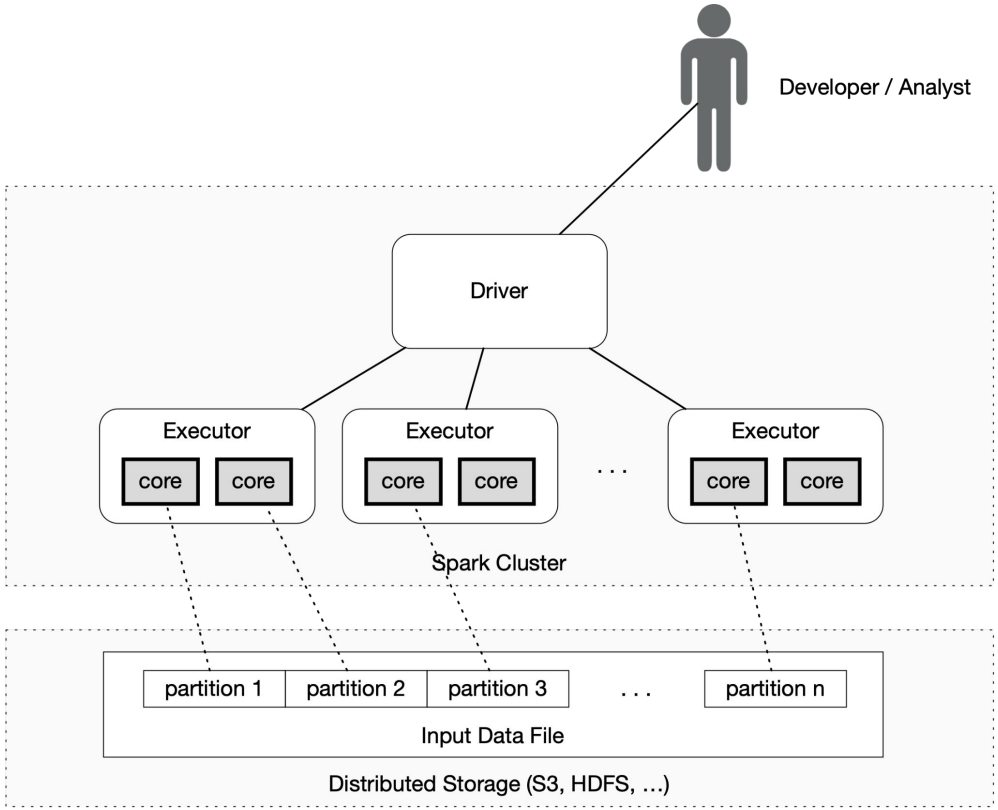
Solution - Classic ETL



Entering the era of Big Data

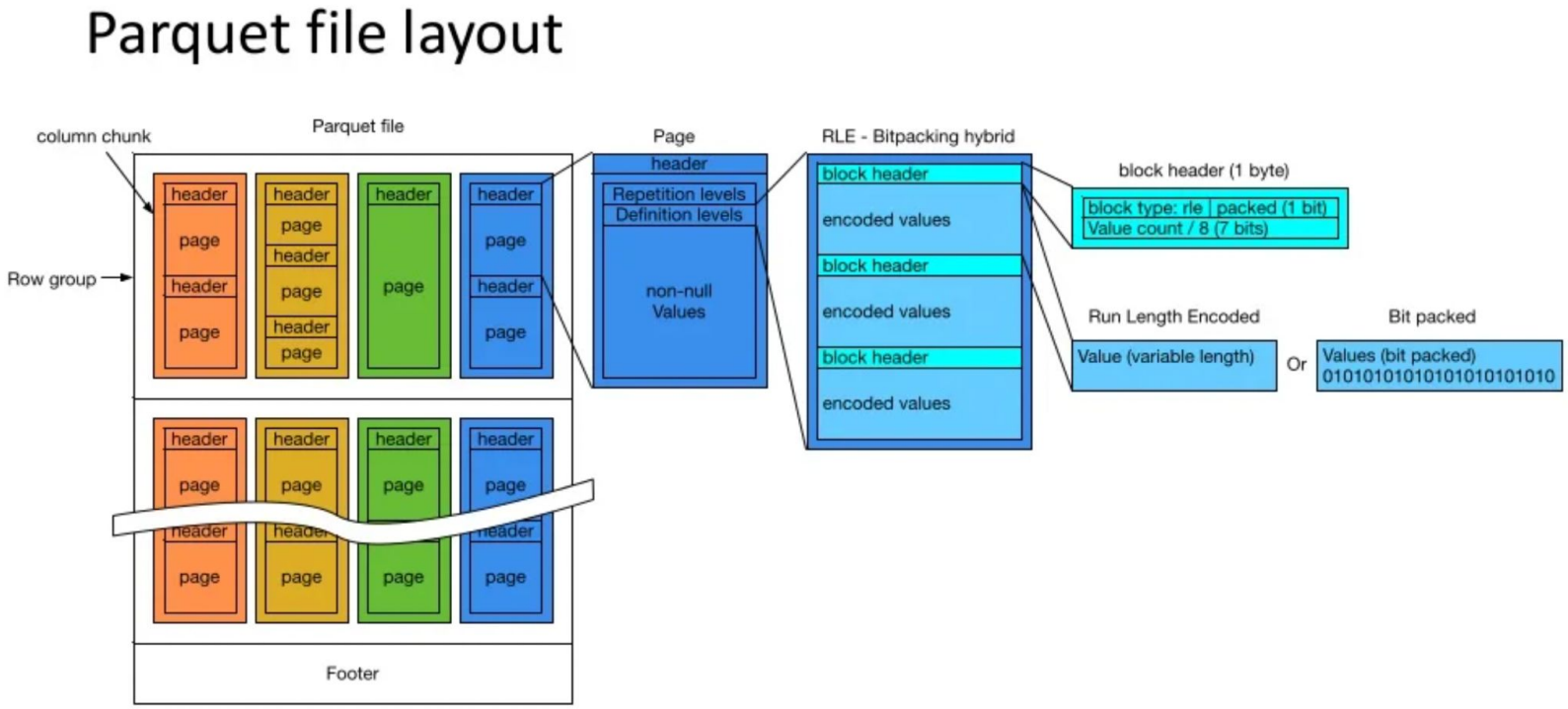

APACHE
Spark
TM

Spark Architecture

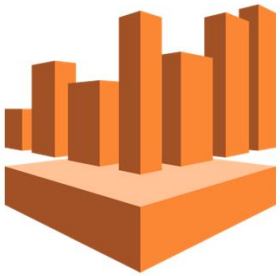


Scalable storage in the Cloud

High-performance file-formats



Data Warehousing meets Big Data (ACID guarantees)



Amazon Athena

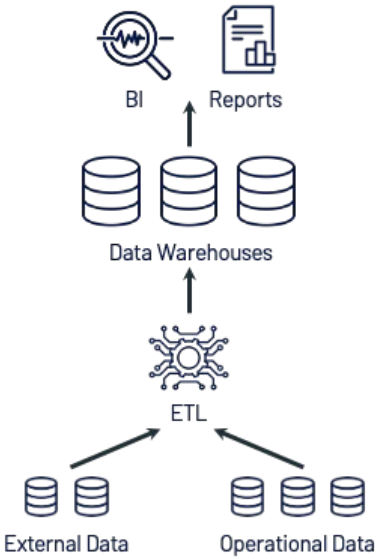
DELTA LAKE

The Data Warehouse meets the Data Lake

databricks

snowflake

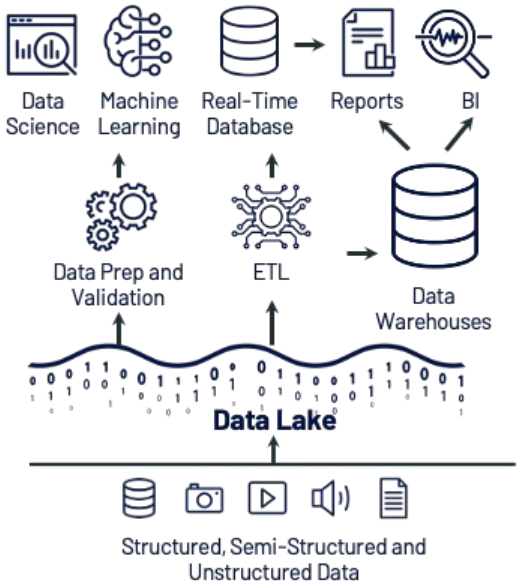
Data Warehouse



credits: Databricks

Data Warehouse

Data Lakes

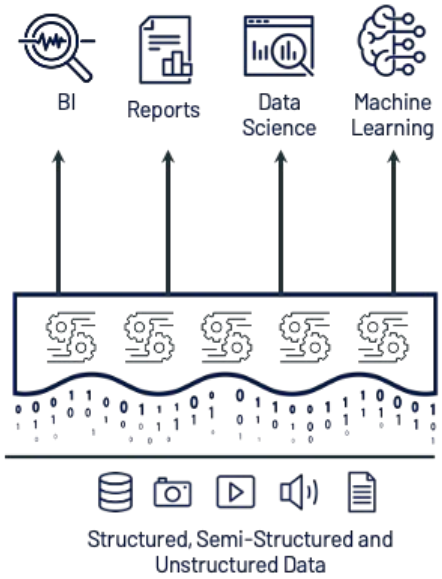


credits: Databricks

Data Warehouse

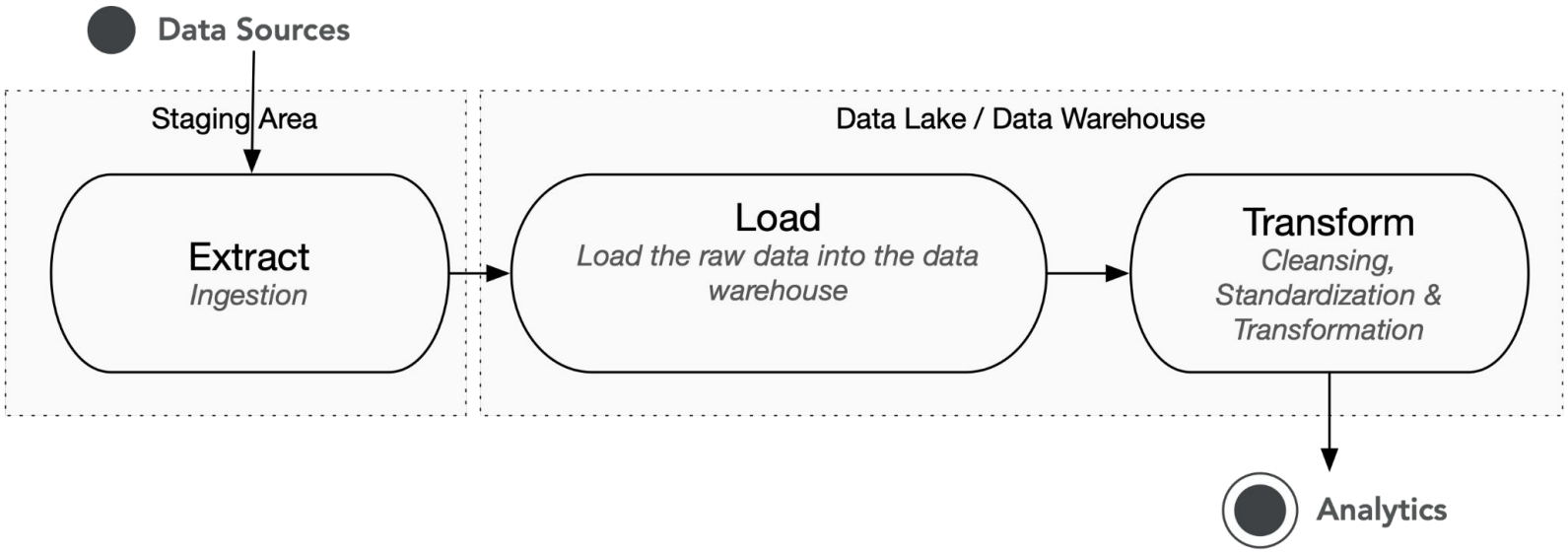
Data Lakes

Lakehouse

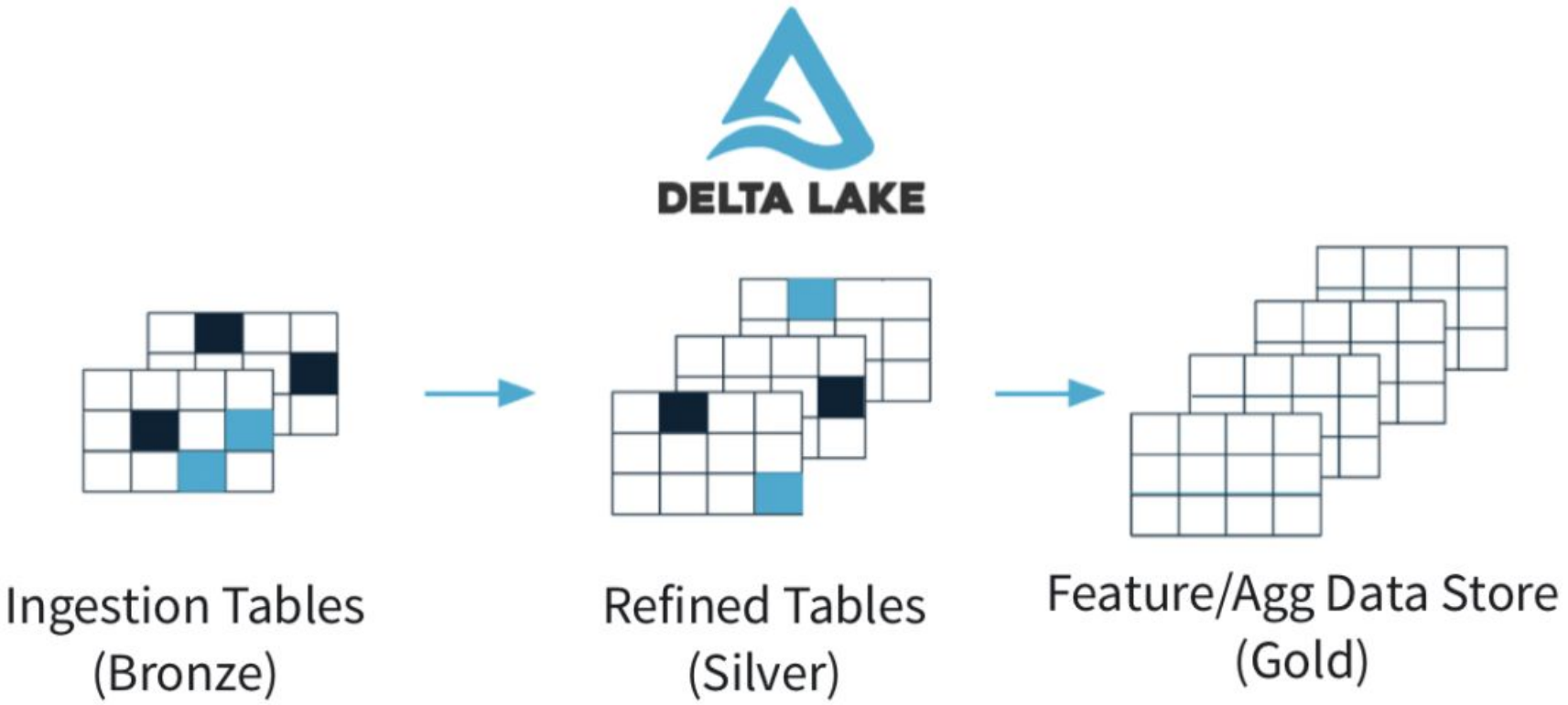


credits: Databricks

Cheap DWH Storage: The advent of ELT



The Medallion Architecture (see https://delta.io)



The “Modern Data Stack”