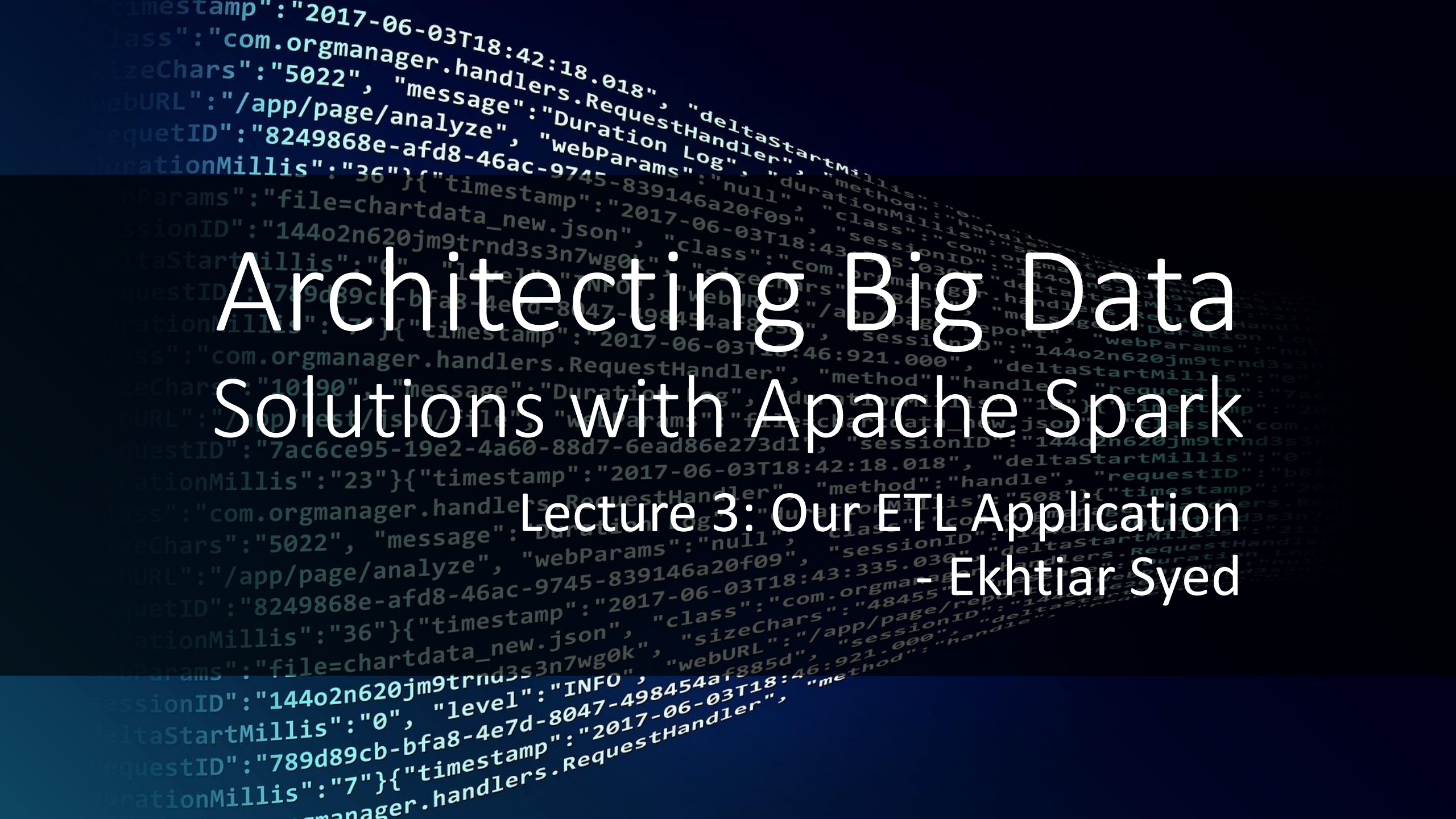
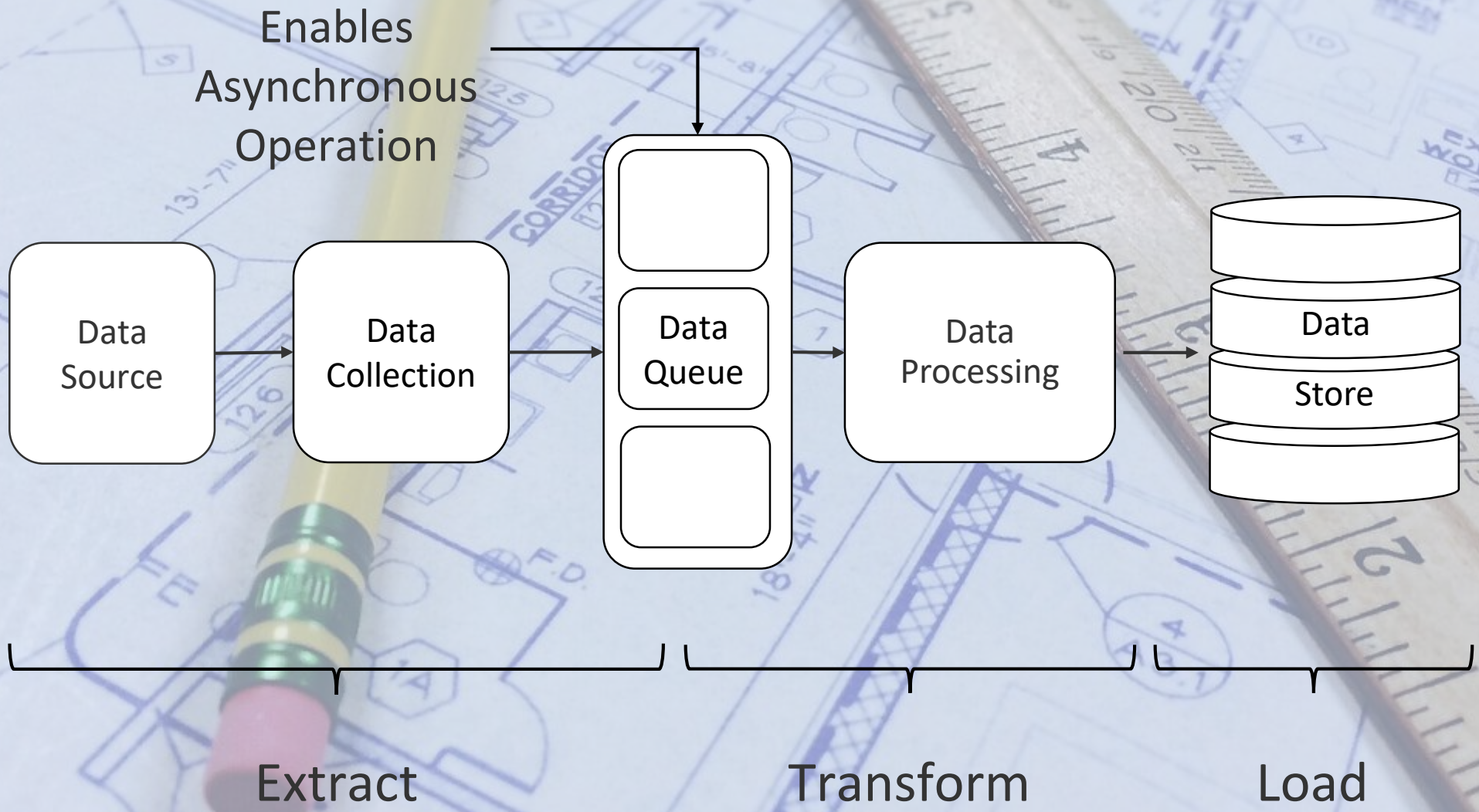


Architecting Big Data Solutions with Apache Spark

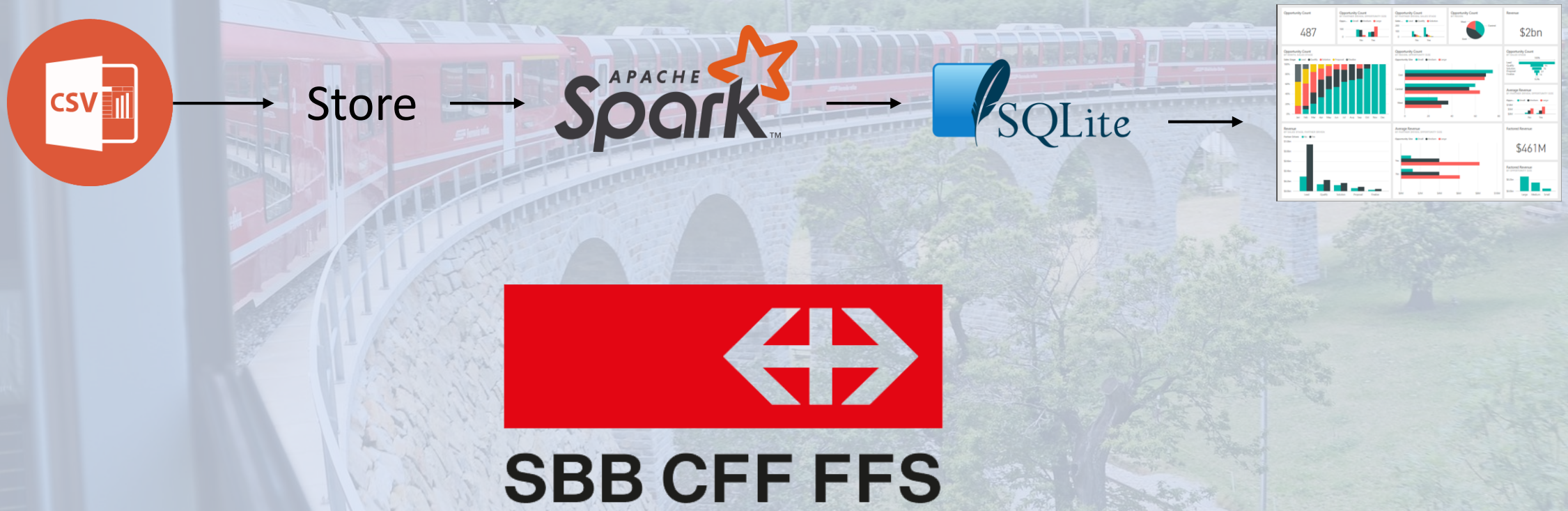
Lecture 3: Our ETL Application

- Ekhtiar Syed

Reference Architecture: Batch Oriented Data Pipelines



Our Project



SBB CFF FFS

Technology Mapping for Our Course

Engagement Layer	Superset, matplotlib, Seaborn
Analytics Layer	Apache Spark
Integration Layer	Scrapy
Persistence Layer	Local File System
Infrastructure Layer	databricks, amazon EMR

We will use visualization tools like Superset to visualize The result of our analytics layer.

We will use spark to solve complex analytics, machine learning and streaming problems. This is another focus area of our course!

We will use Spark to collect and consume data from desperate sources. This is one of the focus areas!

Persistence layer, and Relational and Non-relational databases are a topic of it's own. I will touch upon only the Parquet file format as our persistence layer.

We will use Databrick's community edition as our learning platform. If there is time, I would like to touch upon Elastic MapReduce on Amazon Web Services.

Source: Spark for Python Developers – Amit Nandi

Let’s Start The Practical Part!