

# Architecting Big Data Solutions with Apache Spark

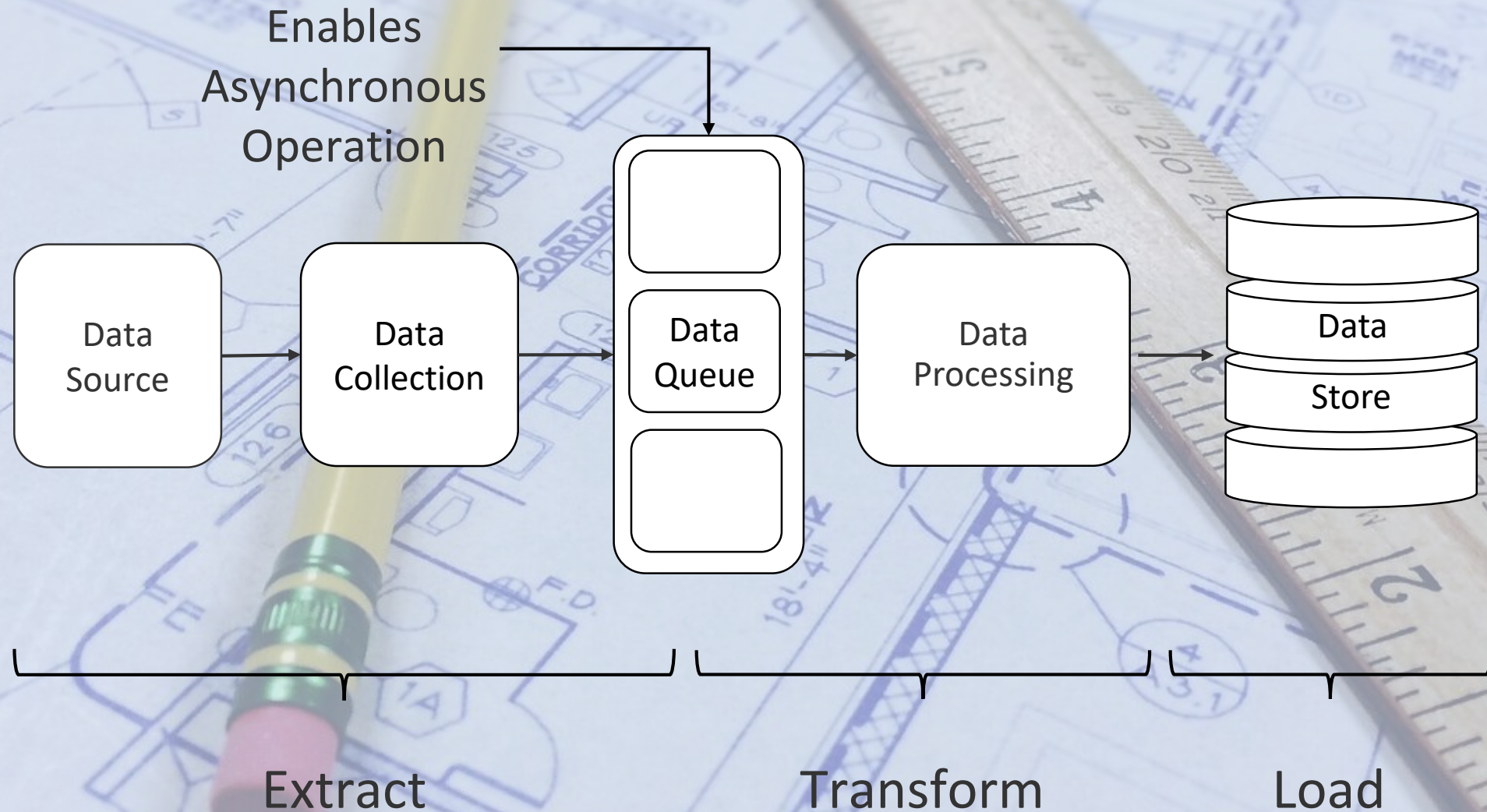
## Lecture 3: Our ETL Application

**Ekhtiar Syed**

Data Engineer/Scientist

Signify Research (formerly known as Philips Lighting)

# Reference Architecture: Batch Oriented Data Pipelines





# Our Project

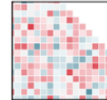


**SBB CFF FFS**



# Technology Mapping for Our Course

## Engagement Layer



Seaborn

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

## Analytics Layer



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

## Integration Layer



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

## Persistence Layer

Local File System

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.

## Infrastructure 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.





# Let's Start The Practical Part!