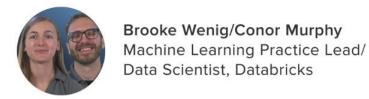
Lesson 3.1: Engineering Data Pipelines

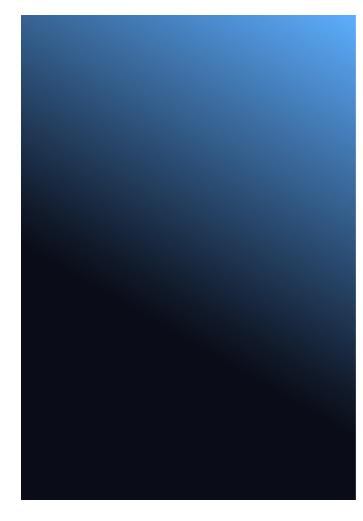
DISTRIBUTED COMPUTING WITH SPARK SQL

Engineering Data Pipelines





Slide 2: Welcome Back!



Welcome Back!

Pipelining: process of moving data through an application

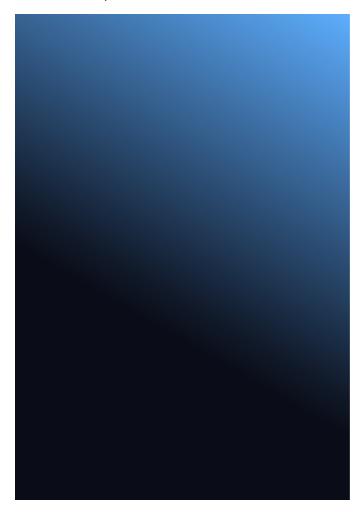
Data ecosystems:

Legacy data warehouses

Databases

Data lakes

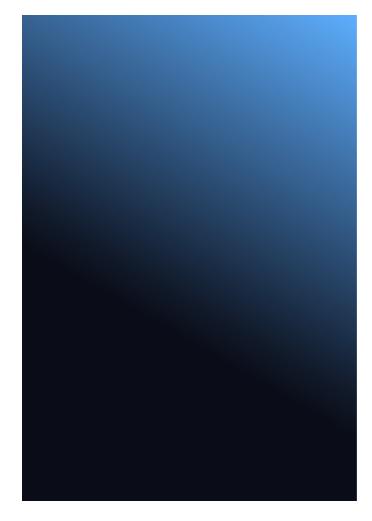
Slide 3: What Spark Has to Offer



What Spark Has to Offer

A unified way to access data where it lives

Slide 4: Module Preview



Module Preview

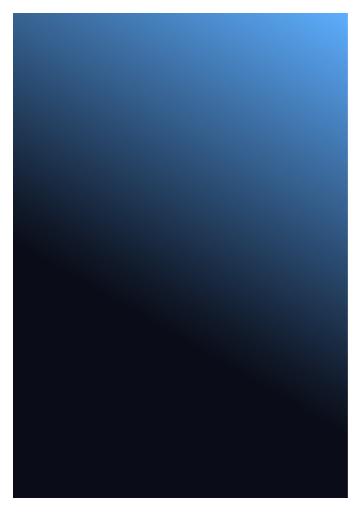
Introduce general demand to data applications

Access data in various formats

Compare and contrast data formats

Explore JSON, schemas, and parallel data writes

Slide 5: Learning Objectives



Learning Objectives

Create an end-to-end pipeline that:

Reads data

Transforms it

Saves end result