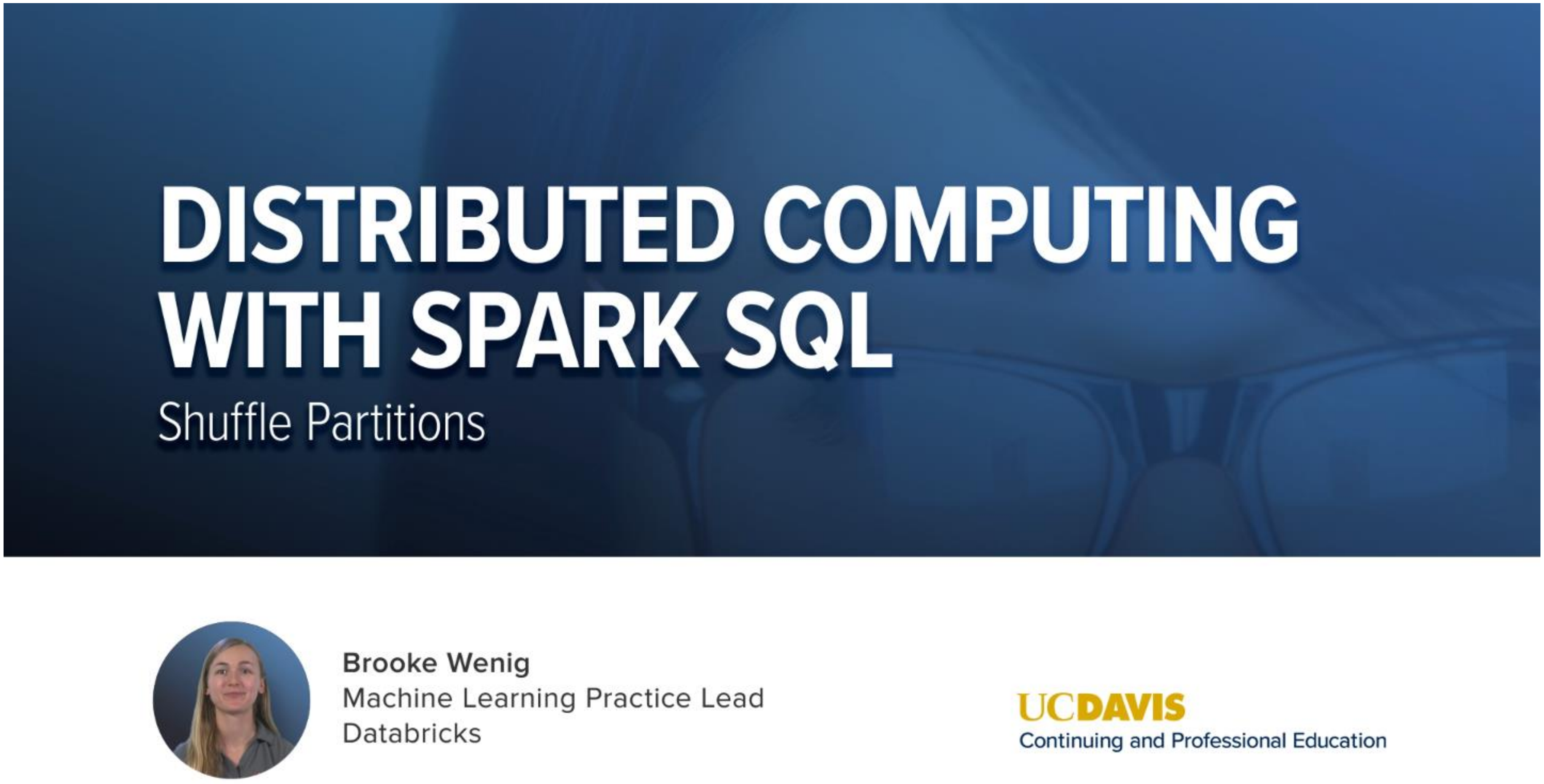
Lesson 2.4: Shuffle Partitions



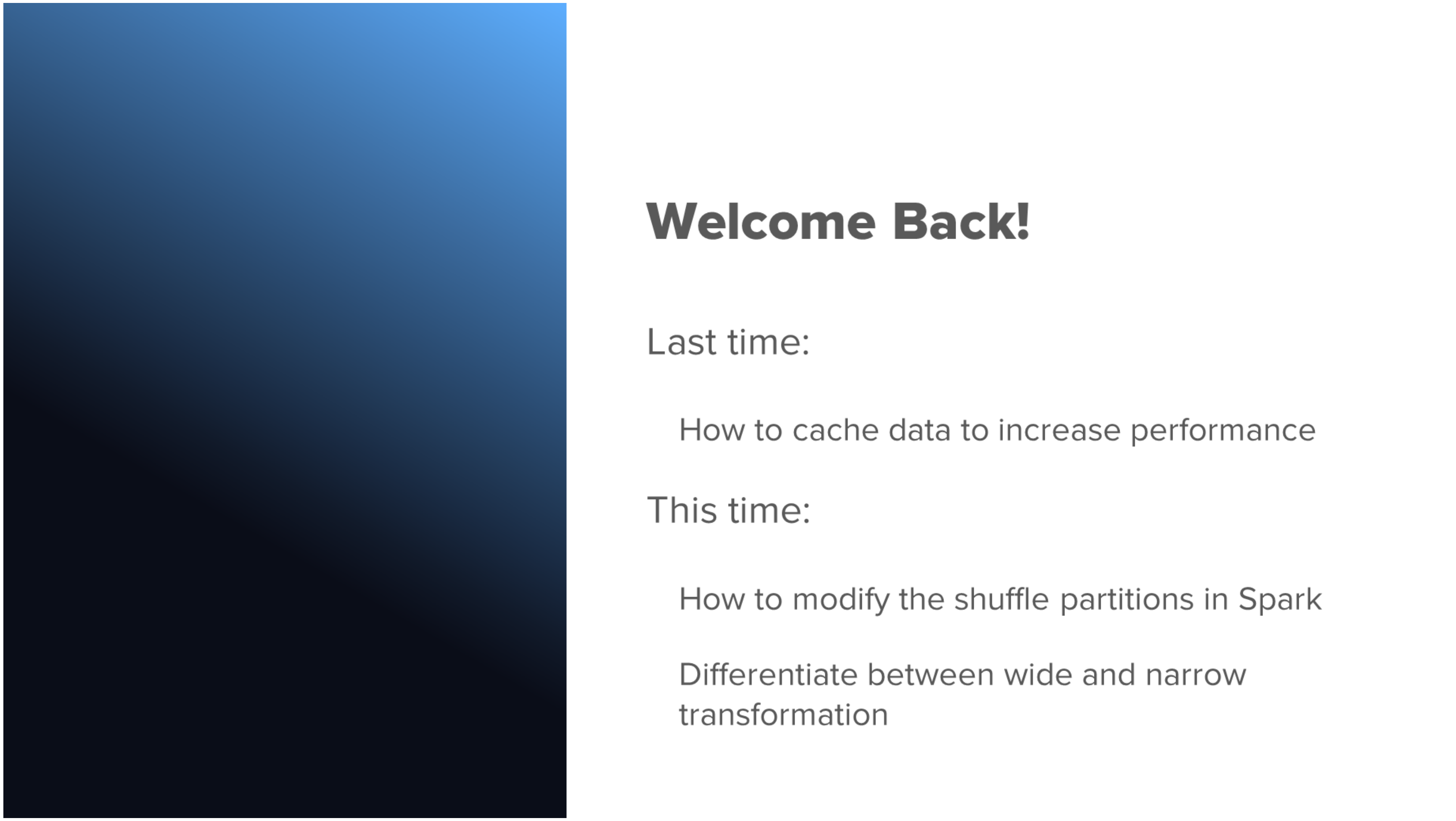
# DISTRIBUTED COMPUTING WITH SPARK SQL

## Shuffle Partitions

**Brooke Wenig**
Machine Learning Practice Lead
Databricks

UCDAVIS
Continuing and Professional Education

Slide 2: Welcome Back!



## Welcome Back!

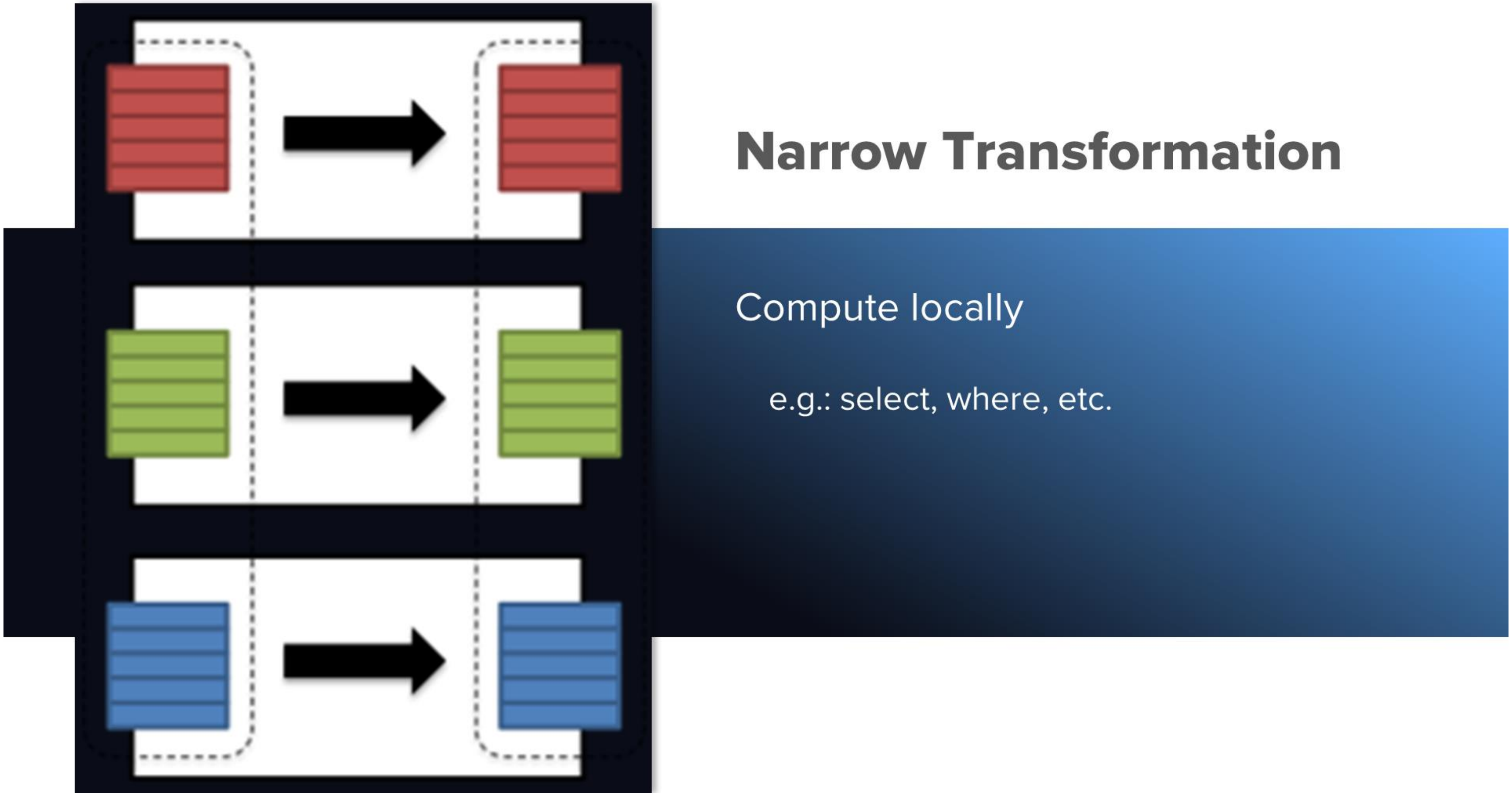
Last time:

- How to cache data to increase performance

This time:

- How to modify the shuffle partitions in Spark
- Differentiate between wide and narrow transformation

Slide 3: Narrow Transformation

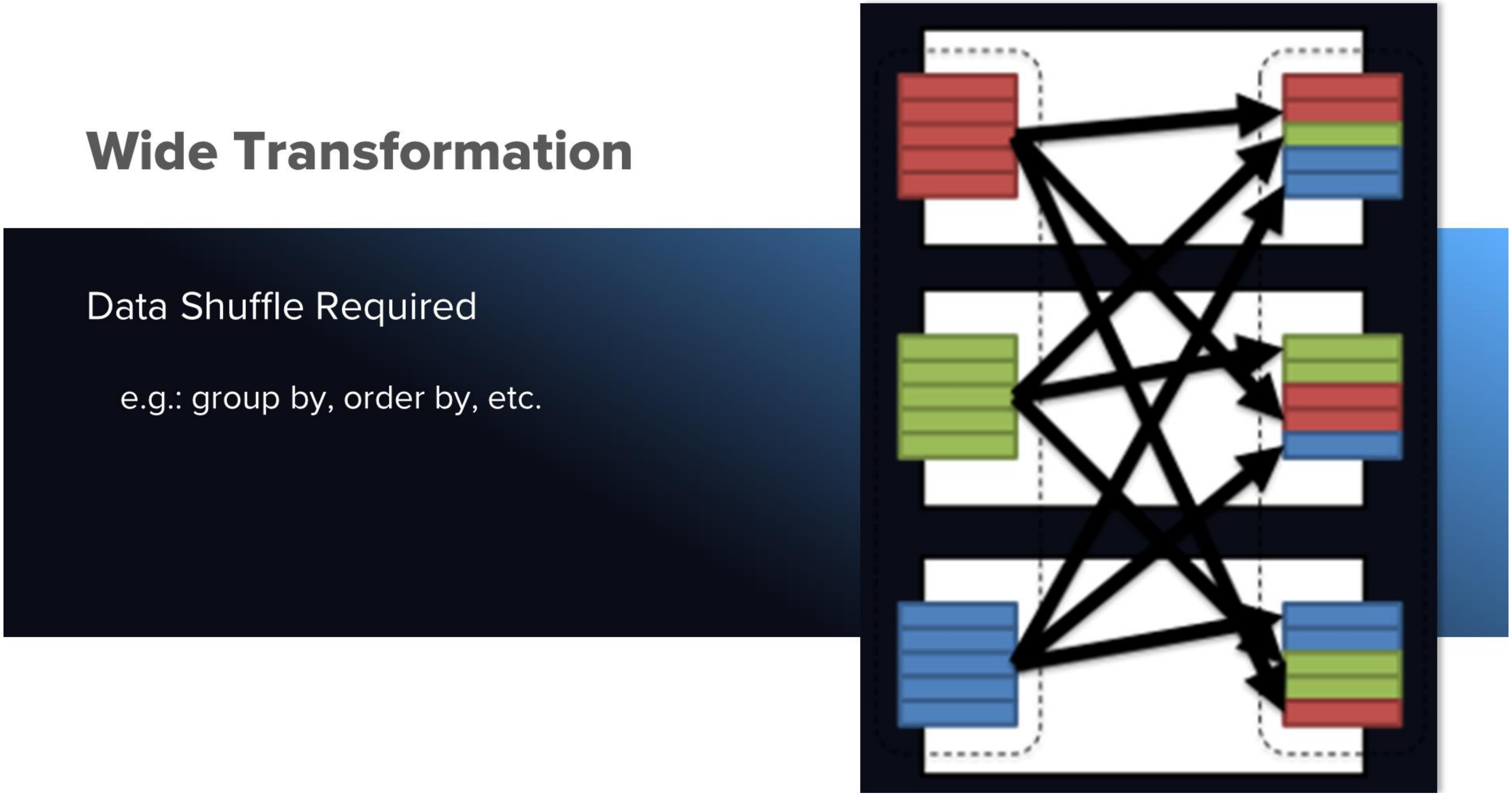


## Narrow Transformation

Compute locally

- e.g.: select, where, etc.

Slide 4: Wide Transformation

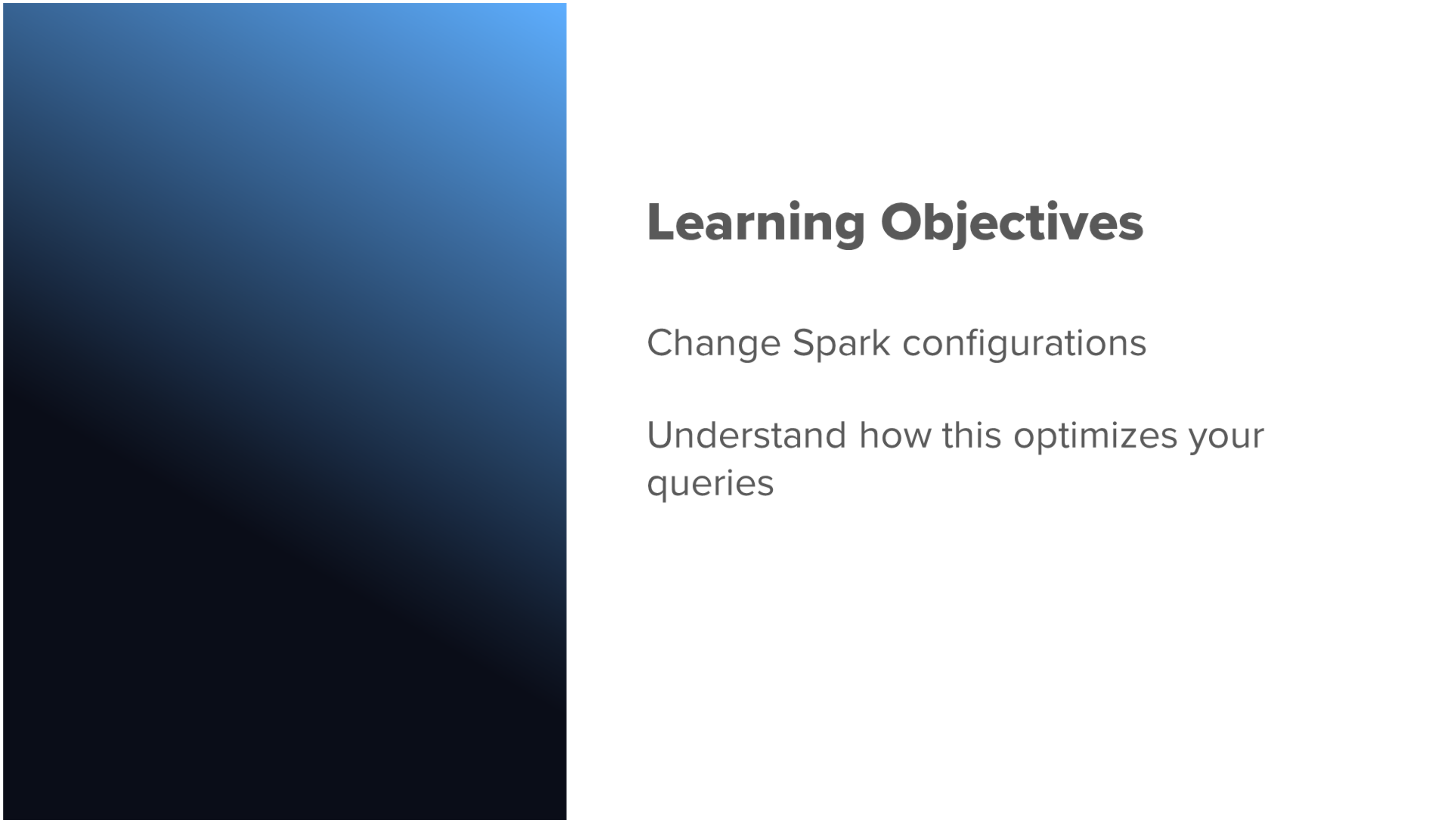


## Wide Transformation

Data Shuffle Required

e.g.: group by, order by, etc.

Slide 5: Learning Objectives



## Learning Objectives

Change Spark configurations

Understand how this optimizes your queries