Apache Spark Application Performance Tuning

Spark Architecture

- RDDs
- DataFrames and Datasets
- Lazy Evaluation
- Pipelining

Data Sources and Formats

- Available Formats Overview
- Impact on Performance
- The Small Files Problem

Inferring Schemas

- The Cost of Inference
- Mitigating Tactics

Dealing With Skewed Data

- Recognizing Skew
- Mitigating Tactics

Catalyst and Tungsten Overview

- Catalyst Overview
- Tungsten Overview

Mitigating Spark Shuffles

- Denormalization
- Broadcast Joins
- Map-Side Operations
- Sort Merge Joins

Partitioned and Bucketed Tables

- Partitioned Tables
- Bucketed Tables
- Impact on Performance

Improving Join Performance

- Skewed Joins
- Bucketed Joins
- Incremental Joins

Pyspark Overhead and UDFs

- Pyspark Overhead
- Scalar UDFs
- Vector UDFs using Apache Arrow
- Scala UDFs

Caching Data for Reuse

- Caching Options
- Impact on Performance
- Caching Pitfalls

Workload XM (WXM) Introduction

- WXM Overview
- WXM for Spark Developers

What's New in Spark 3.0?

- Adaptive Number of Shuffle Partitions
- Skew Joins
- Convert Sort Merge Joins to Broadcast Joins
- Dynamic Partition Pruning
- Dynamic Coalesce Shuffle Partitions