Spark & Clusters

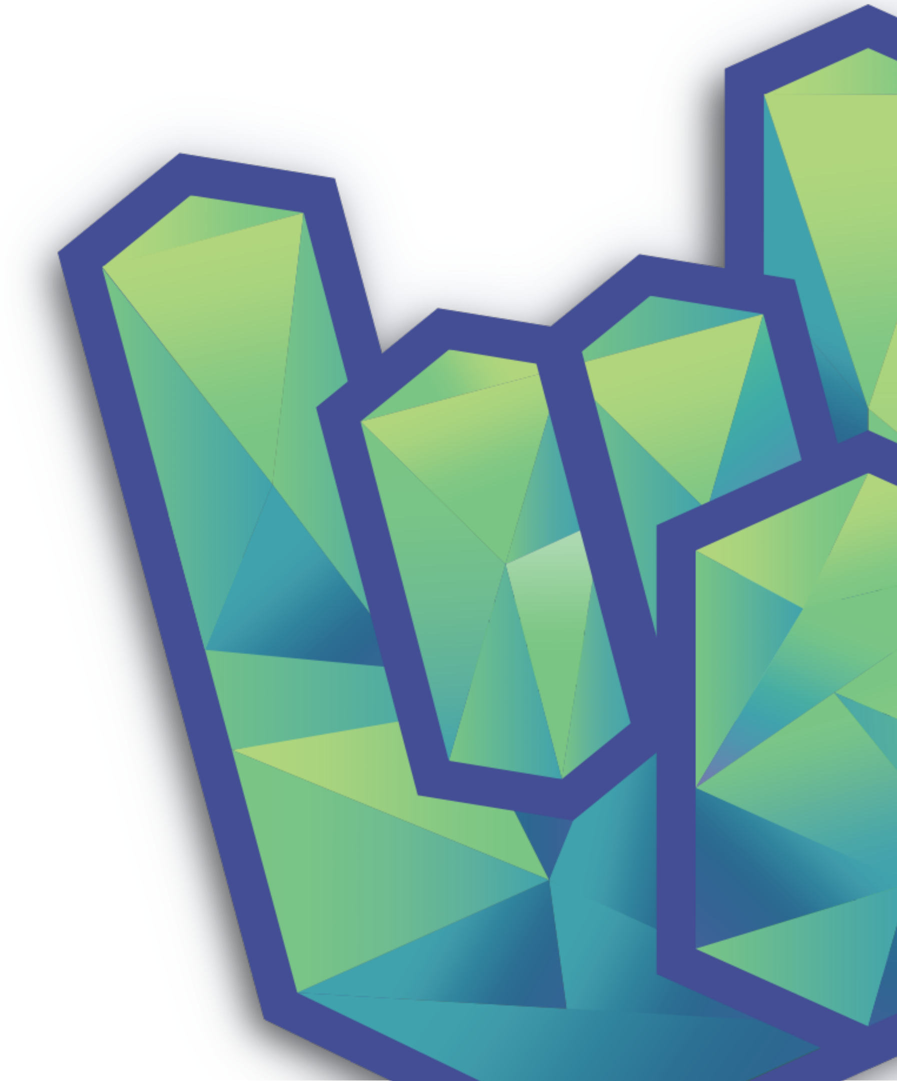
Objective

Show how Spark runs on a cluster

Anatomy of a Spark job

- stages
- tasks
- shuffles
- the DAG

The Spark UI



Execution Terminology

A job has stages, a stage has tasks.

stage = a set of computations between *shuffles*

task = a unit of computation, per partition

the DAG = graph of RDD dependencies

Takeaways

spark-shell for interactive Spark REPL

Spark UI to investigate, debug and track jobs

The anatomy of Spark jobs

- stage = all computations in between two shuffles
- task = unit of computation per partition
- the DAG = the graph of RDD dependencies

Spark optimizations and step generation

Physical plans

Spark rocks