Labs 4–5 Summary: Apache Spark 3.5 DataFrames, SQL, and Query Tuning

Z5008 Big Data Lab · Even Semester 2026 · Dr. Innocent Nyalala

1. Lazy Transformations vs. Actions

Spark builds a logical plan across narrow transformations (`filter`, `select`, `withColumn`) and wide shuffle boundaries (`groupBy`, `join`, `repartition`), executing stages only when an action (`count`, `collect`, `write`) triggers the Catalyst optimizer and Tungsten physical engine.

2. Performance Tuning Principles

Broadcast small dimension tables (`F.broadcast(dim)`) to avoid full-table shuffle joins, enable Adaptive Query Execution (`spark.sql.adaptive.enabled`) to coalesce post-shuffle partitions dynamically, and compute partitioned window rankings (`RANK() OVER`) cleanly without extra self-joins.