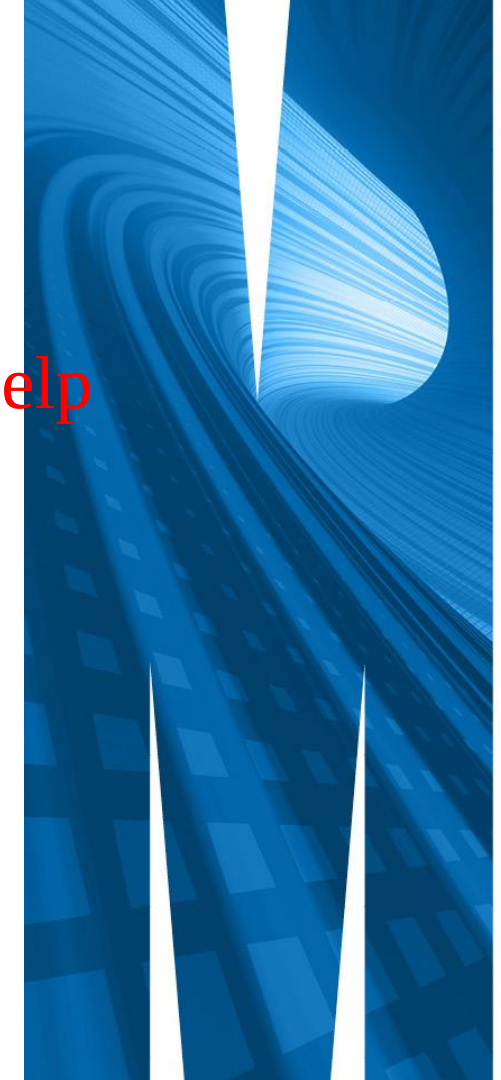
MONASH University

Week 4

FIT5202 Big Data Processing



Week 4 Agenda

- Week 3 Review
- Review of Parallel Joins
- Dataframe operations with pyspark
  - Sort
  - Distinct
  - Groupby
- Dataframe UDFs(User Defined Functions)
- TODO : Combining various operations to write dataframe queries.
- Working on Assignment 1

Week 3 Review

- Spark Join Strategies
  - Broadcast Hash Join
  - Sort Merge Join
  - Shuffled Hash Join
- Parallel Joins
  - Inner, Outer, Left, Right, Left Anti, Left Semi
- Understanding Query Execution plans from Spark UI DAG

Week 3 Review

Why Full outer join uses “sort-merge”?

BHJ is not supported for full outer join.

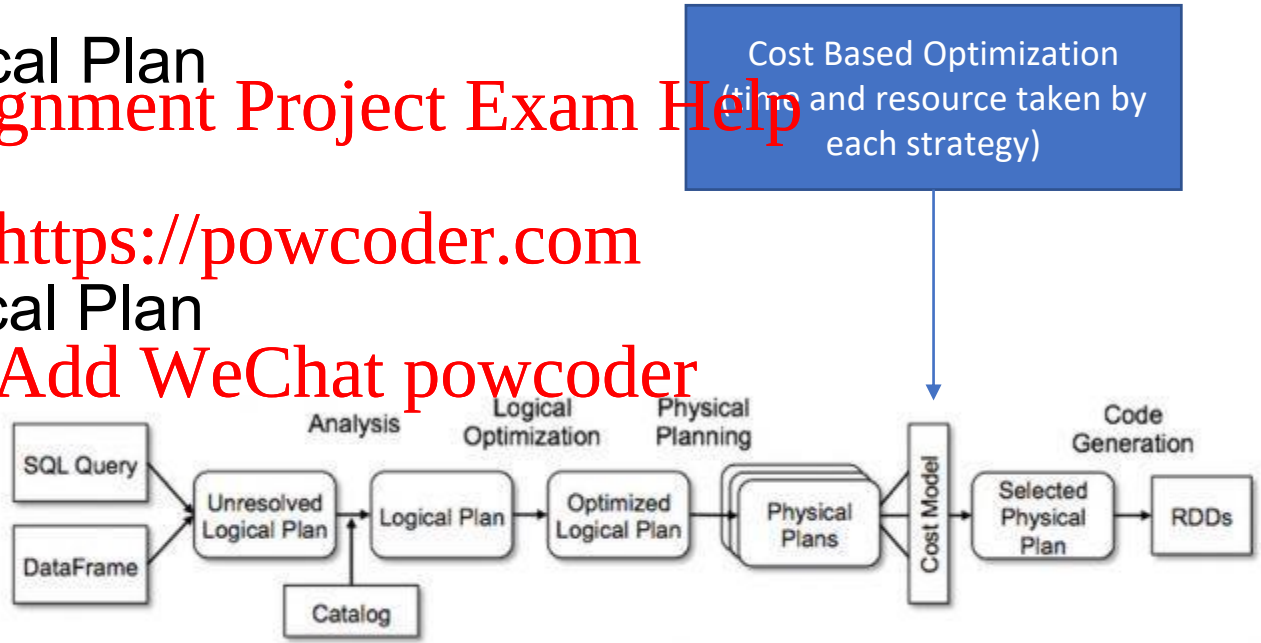
Ref : https://sujithjay.com/spark/broadcast-joins

PySpark Cheatsheet

https://s3.amazonaws.com/assets.datacamp.com/blog_assets/PySpark_SQL_Cheat_Sheet_Python.pdf

Spark Execution Plan

- Logical Plan
- Optimized Logical Plan
- Physical Plans
- Cost Model
- Selected Physical Plan



Lab Instructions and Demo

Thank You!

See you next week.