

# Week 4

Assignment Project Exam Help

FIT5202 Big Data Processing https://powcoder.com



# Week 4 Agenda

- Week 3 Review Assignment Project Exam Help
- Review of Parallel Joins
- Dataframe operations with pyspark https://powcoder.com
  - Sort
  - Distinct
  - Groupby

- Dataframe UDFs(User Defined Functions)
- TODO: Combining various operations to write dataframe gueries.
- Working on Assignment 1



#### Week 3 Review

- Spark Join Strategies Signment Project Exam Help

   Broadcast Hash Join

  - Sort Merge Join
  - Shuffled Hash Join

https://powcoder.com

- **Parallel Joins** 
  - Inner, Outer, Left, Right, Left Anti, Left Semi powcoder
- 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. Project Exam Help

https://powcoder.com

Add WeChat powcoder

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



# **PySpark Cheatsheet**

https://powcoder.com



#### **Spark Execution Plan**

Logical Plan Optimized Logical Plan
Physical Plans

Cost Based Optimization

Cost Based Optimization

Heilip and resource taken by each strategy) Cost Model https://powcoder.com Selected Physical Plan Add WeChat powcoder Code Planning Generation SQL Query Selected Optimized Unresolved Physical Logical Plan Physical **RDDs** Logical Plan Logical Plan Plan DataFrame Catalog

## **Lab Instructions and Demo**

Assignment Project Exam Help

https://powcoder.com



#### **Thank You!**

See you next week.

Assignment Project Exam Help

https://powcoder.com