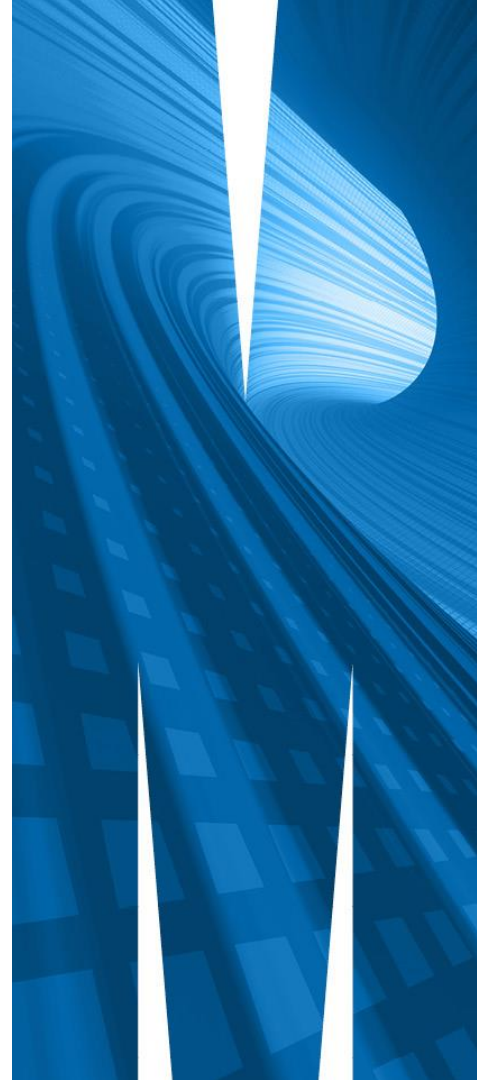


Week 3

FIT5202 Big Data Processing



Week 3 Agenda

- Week 2 Review
- Assignment 1 Briefing
- Join Strategies in Spark
- Joins used in Spark
- Understanding Query Execution Plan and DAG in Spark UI (Demo)

Week 2 Review

- Understanding SparkSession and SparkContext
- Data Partitioning in RDDs
- Parallel Search and RDDs with multiple conditions
- RDDs vs DataFrames
- Partitioning in DataFrames
- Searching and Filtering in DataFrames
- Using SparkSQL With DataFrames.

Solution – Week 2

```
from pyspark.sql.functions import col

bank_df = df_round

## The functions to filter in dataframes are similar to the functions in RDD. Thus, the steps are:
# 1. Search the records with balance between 1000 and 2000
bank_df = bank_df.filter(col("balance")>1000)\
    .filter(col("balance")<2000)

# TODO:
# 2. Also in the same dataframe, search the records with primary or secondary education and age less than 30
# bank_df =
#### SOLUTION:
bank_df = bank_df.where(col("education").isin(["primary", "secondary"])).filter(col("age")<30)

# TODO:
# 3. Also filter with those who are married
# bank_df =
#### SOLUTION:
bank_df = bank_df.filter(bank_df["marital"]=="married")

# TODO:
# 4. Display the previous attributes plus the information of day, month and deposit
# bank_df =
#### SOLUTION:
bank_df = bank_df.select("age", "education", "balance", "day", "month", "deposit")

# 5. Display the records
bank_df.show()
```

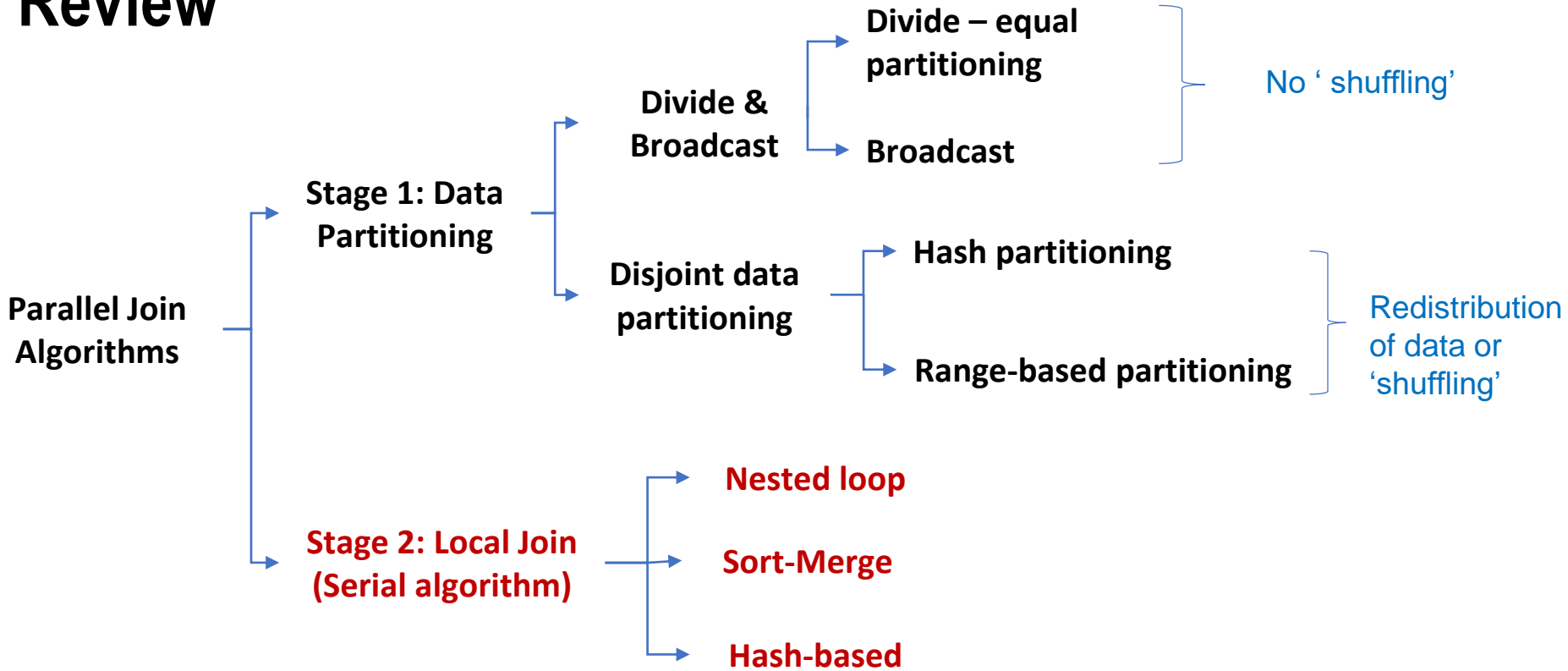
Joining Strategies By Spark

Why do we need different join strategy?

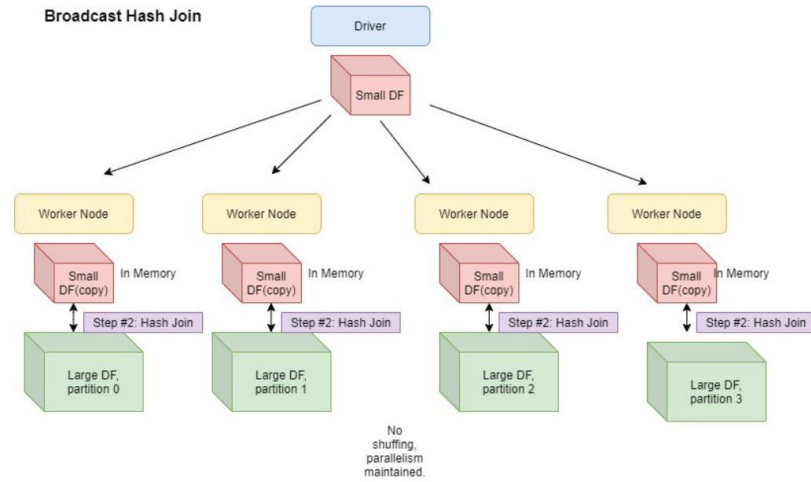
1. Shuffle Sort Merge Join (default)
2. Broadcast Hash Join
3. Shuffle Hash Join

<https://towardsdatascience.com/strategies-of-spark-join-c0e7b4572bcf>

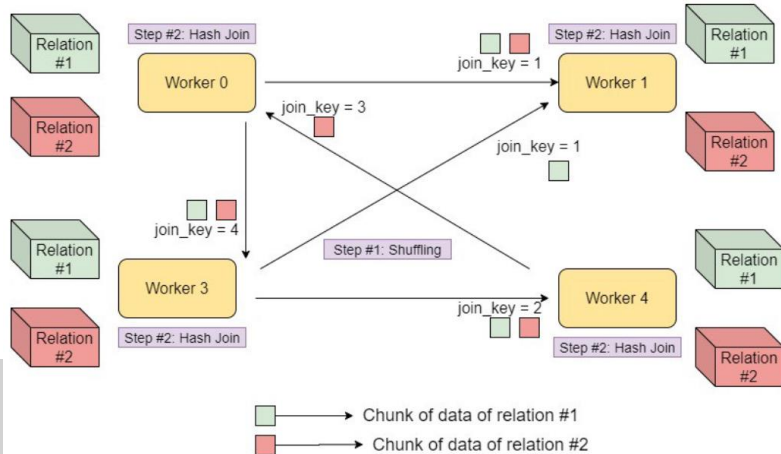
Review



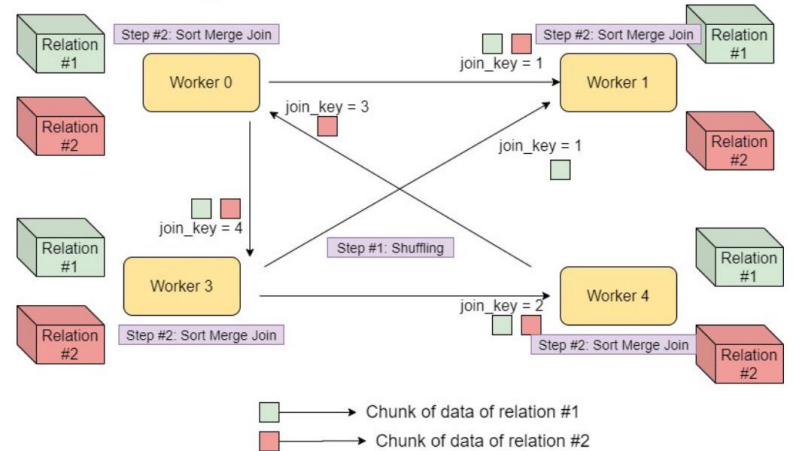
Broadcast Hash Join



Shuffle Hash Join



Shuffle Sort-Merge Join

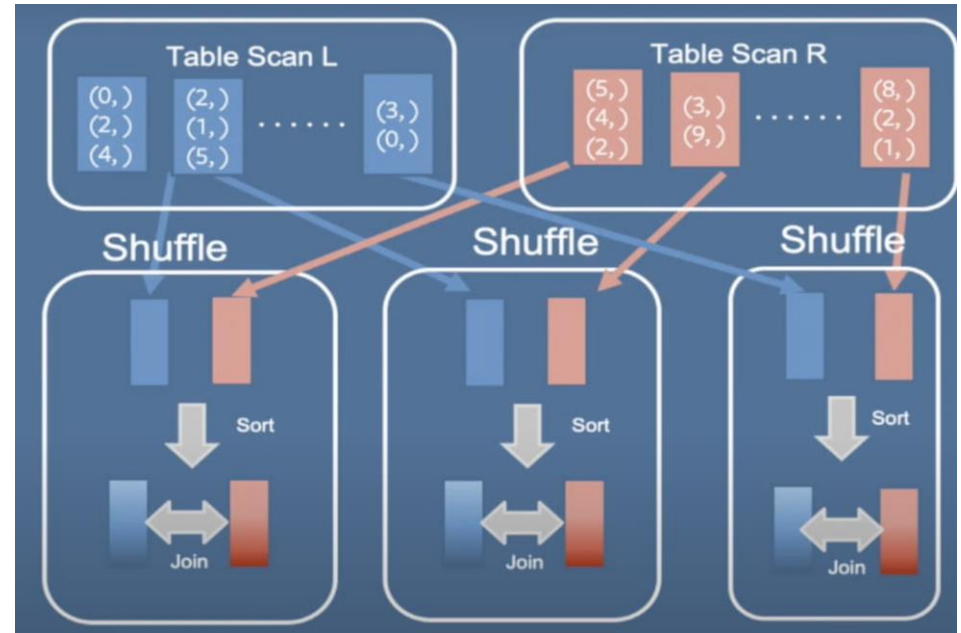


Join Strategies in Spark

Sort-Merge Join

1. Data from both tables are shuffled (when no bucketing was done)
2. Data are then sorted based on the key and perform sort-merge join on each executor

Most Scalable joining strategy in Spark

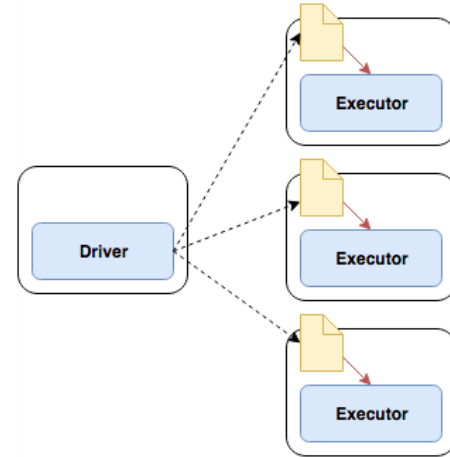


https://databricks.com/session_eu19/spark-sql-bucketing-at-facebook

Join Strategies in Spark

Broadcast Hash

1. Smallest table is broadcasted to each node (using broadcast variable)
2. Joining then takes place
3. Eliminates the shuffling process for bigger dataset.
4. Threshold can be Configured using “spark.sql.autoBroadcastJoinThreshold”
5. Default is 10 MB



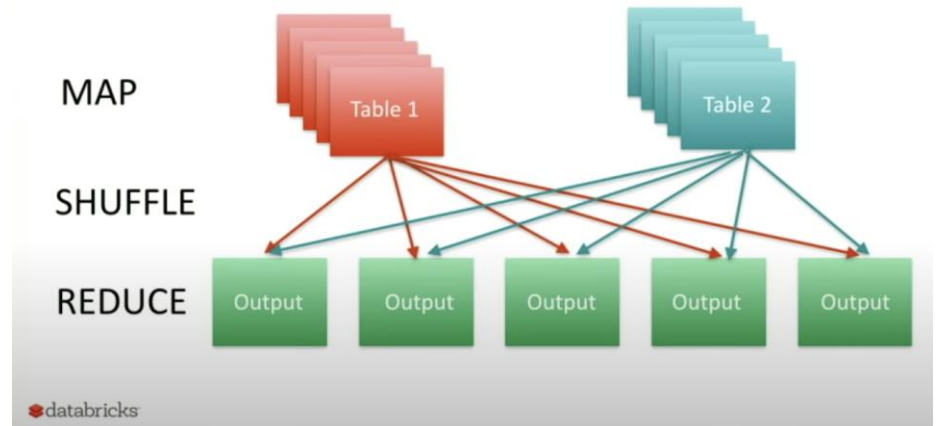
<https://henning.kropponline.de/2016/12/11/broadcast-join-with-spark/>

Join Strategies in Spark

Shuffle Hash

1. When both data frames are partitioned using same partitioner. So join keys will fall into the same partition.

Shuffle Hash Join



<https://databricks.com/session/optimizing-apache-spark-sql-joins>

Join Strategies in Spark

When to use which?

1. When both datasets are too big : Sort-merge join/ Shuffle-hash join
2. When one table is big and another is small : Broadcast Hash Join

Other Core Joins

- INNER
- OUTER
- LEFT
- RIGHT
- LEFT SEMI
- LEFT ANTI

Query Execution Plan and DAG Demo

Thank You!

See you next week.