DATAHACK

Proyecto Apache Spark - Jorge de Dios

EJERCICIO 4

Descripción de cada 'caja' en el plan de ejecución del ejercicio 3 en donde se realizó un join entre dos dataframes (df_food_inspections_cleaned y df_risk_description_cleaned)

Stage 211 (skipped)

- Scan csv: Spark escanea y lee los datos desde un archivo CSV.

- WholeStageCodegen: Optimizador que ayuda a reducir la sobrecarga y mejorar la eficiencia en el procesamiento de los datos leídos del CSV.

- Exchange: Redistribuye los datos ente los nodos para preparar la operación de join, generando una etapa de intercambio.

Stage 212

- Scan csv: Similar a la caja en el Stage 211, esta operación escanea los datos desde la fuente CSV.

- ShuffleQueryStage: Recibe datos redistribuidos (shuffle) del Exchange en el Stage 211 para la operación de join.

- WholeStageCodegen: Al igual que en el Stage 211, esta caja optimiza la ejecución de los datos combinados y finaliza la ejecución del join.