Scenario: Given a customers table generated using fake data and zipcodes csv we had to join the 2 datasets to get aggregated metrics by zipcode.

Approach:

1. Generate fake data for customers using the notebook
2. Read customers data and zipcodes csv as dataframes
3. Customers had approx. 400K records and zipcode csv was 984 MB
4. As this is the case where one of the dataframes is less than 2 GB threshold for broadcast, we chose broadcast join to join the datasets
5. First, aggregate the zipcodes dataset to get unique sum of population by zipcodes
6. Join the customer dataframe and use broadcast pyspark method to explicitly hint spark to use broadcast join
7. The below DAG shows that due to broadcast join, the zipcode dataframe was broadcasted to each executor. This avoided the shuffling of both dataframes
8. The exchange stage that we are seeing in the DAG is for groupby after joining both dataframes for combining total customers by zipcode

DAG visualization showing broadcast hash join and 2 exchange steps for groupby

Jobs Stages Storage Environment Executors SQL / DataFrame JDBC/ODBC Server Structured Streaming Connect

Details for Query 75

Download screen as png

Submitted Time: 2025/04/24 23:35:15
Duration: 11 s
Succeeded Jobs: 51 52

Expand all the details in the query plan visualization Show experimental metrics

Plan Visualization