Dividing Sales Data by Month in Spark

This document explains how to divide sales data into separate tables or files based on the month using PySpark. The process involves extracting the year and month from the 'ModifiedDate' column and dynamically filtering the data for each month.

# PySpark Code for Dividing Sales Data by Month

from pyspark.sql import SparkSession  
from pyspark.sql.functions import col, to\_utc\_timestamp, from\_utc\_timestamp, lit, dayofyear, when, month, year  
  
# Initialize the Spark session  
spark = SparkSession.builder.appName("Sales Data by Month").getOrCreate()  
  
# Schema  
schema = """   
 SalesOrderID INT,  
 SalesOrderDetailID INT,  
 CarrierTrackingNumber STRING,  
 OrderQty INT,  
 ProductID INT,  
 SpecialOfferID INT,  
 UnitPrice DOUBLE,  
 UnitPriceDiscount DOUBLE,  
 LineTotal Double,  
 rowguid STRING,  
 ModifiedDate STRING  
"""  
  
# Load the CSV data  
df = spark.read.format("csv").option("header", "true").schema(schema).load("/FileStore/tables/Sales\_SalesOrderDetail.csv")  
  
# Convert 'ModifiedDate' to timestamp  
df = df.withColumn("ModifiedDate", col("ModifiedDate").cast("timestamp"))  
  
# Extract year and month  
df = df.withColumn("Year", year(col("ModifiedDate"))).withColumn("Month", month(col("ModifiedDate")))  
  
# Show the dataframe  
df.show()  
  
# Loop through the distinct years and months in the data  
distinct\_years\_months = df.select("Year", "Month").distinct().collect()  
  
for row in distinct\_years\_months:  
 year\_val = row["Year"]  
 month\_val = row["Month"]  
   
 # Filter data for the specific year and month  
 df\_filtered = df.filter((col("Year") == year\_val) & (col("Month") == month\_val))  
   
 # Show the filtered data for each month (this can be saved as a table or written to a file)  
 print(f"Data for {year\_val}-{month\_val}")  
 df\_filtered.show()  
   
 # Save each month's data to a new CSV or table  
 output\_path = f"/mnt/data/sales\_data\_{year\_val}\_{month\_val}.csv"  
 df\_filtered.write.csv(output\_path, header=True)