Big Data

Spark SQL (Jobsheet 9)

Dosen:

M. Hasyim Ratsanjani, S.Kom., M.Kom.



Dibuat Oleh:

Nurhaliza Anindya Putri 2241720016

Kelas: TI - 3D

POLITEKNIK NEGERI MALANG

2025

1. Implementasi Python Spark (di Dockers)

No	Gambar	Deskripsi
	from pyspark.sql import SparkSession # Buat SparkSession spark = SparkSession.builder.appName("ExtractCSV").getOrCreate() # Extract - Membaca file CSV df = spark.read.csv("sales_data.csv", header=True, inferSchema=True) # Tampilkan beberapa baris pertama df.show(5) # Cek struktur data df.printSchema()	Extract: Baca data dari file CSV (sales_data.csv).
	Section Sect	Hasil ss dari ekstract sales_data.csv
	from pyspark.sql.functions import col # Filter transaksi dengan Revenue > 100 df_filtered = df.filter(col("Revenue") > 100) # Tampilkan hasil filter df_filtered.show(5)	Filter transaksi dengan Revenue > \$100.

Bardus	Hasil ss filter transaksi dengan revenue >100
from pyspark.sql.functions import sum	Hitung total penjualan per kategori.
# Hitung total penjualan per Product_Category df_result = df_filtered.groupBy("Product_Cate gory") \	
.agg(sum("Revenue").alias("total_s ales"))	
# Tampilkan hasilnya df_result.show()	
Product_Category total_sales Clothing 8198902 Accessories 13559164 Bikes 61782134	Hasil ss hitung total penjualan per kategori
# Load - Simpan hasil ke Parquet df_result.write.mode("overwrite").p arquet("output_sales.parquet")	Load: Simpan hasil ke Parquet.
from pyspark.sql.functions import month	Hitung total pendapatan per bulan.
# Pastikan kolom Date bertipe DateType from pyspark.sql.functions import to_date	
# Convert string ke DateType kalau perlu df = df.withColumn("Date", to_date(col("Date"), "yyyy-MM-dd"))	
# Extract bulan dari kolom Date df_with_month = df.withColumn("Month",	

```
month(col("Date")))
# Hitung total pendapatan per bulan
df monthly revenue =
df with month.groupBy("Month")
.agg(sum("Revenue").alias("total\_r
evenue")) \
  .orderBy("Month")
# Tampilkan hasil
df monthly revenue.show()
                                       Hasil ss Hitung total pendapatan
 |Month|total_revenue|
                                       per bulan.
         7005895
          6834583
          7347164
          7602750
    41
          8836763
          9043008
    6
          5721459
          5711193
          5841885
          59950791
   10
          6244298
   11
          9086931
                                      Identifikasi 5 produk terlaris.
from pyspark.sql.functions import
desc, sum
# Kelompokkan berdasarkan kolom
Product dan hitung total Revenue
df top products =
df.groupBy("Product") \
.agg(sum("Revenue").alias("total r
evenue")) \
  .orderBy(desc("total revenue")) \
  .limit(5)
# Tampilkan hasil
df top products.show()
```

Product total_revenue Road-150 Red, 62 3829416 Mountain-200 Blac 3366248 Road-150 Red, 52 3180840 Road-150 Red, 56 3158805 Mountain-200 Silv 3081078	Hasil ss Identifikasi 5 produk terlaris.
# Simpan hasil ke Parquet df_top_products.write.mode("over write").parquet("top5_products.par quet")	Simpan hasil dalam format Parquet.
from pyspark.sql.functions import month, sum # Ambil bulan dari kolom Date df_with_month = df.withColumn("Month", month("Date")) # Kelompokkan berdasarkan bulan dan hitung total pendapatan df_monthly_revenue = df_with_month.groupBy("Month") .agg(sum("Revenue").alias("total_r evenue")) # Tampilkan hasil df_monthly_revenue.show() # Simpan hasil ke Parquet df_monthly_revenue.write.mode("o verwrite").parquet("monthly_revenue.ue.parquet")	Pendapatan perbulan

Month total_revenue 12 9086931 1 7005895 6 9043008 3 7347164 5 8836763 9 5841885 4 7602750 8 5711193 7 5721459 10 5995079 11 6244298 2 6834583	Hasil ss Pendapatan perbulan
from pyspark.sql.functions import sum, desc # Kelompokkan berdasarkan produk dan hitung total pendapatan df_top_products = df.groupBy("Product") \ .agg(sum("Revenue").alias("total_r evenue")) \ .orderBy(desc("total_revenue")) \ .limit(5) # Tampilkan hasil df_top_products.show()	Identifikasi 5 Produk terlaris
Product total_revenue Road-150 Red, 62 3829416 Mountain-200 Blac 3366248 Road-150 Red, 52 3180840 Road-150 Red, 56 3158805 Mountain-200 Silv 3081078	Hasil ss Identifikasi 5 Produk terlaris
# Simpan hasil ke Parquet df_top_products.write.mode("over write").parquet("top5_products.par quet")	simpan dalam format parquet