Pyspark all in ... ata management

airflow_kaf ... dation.docx

Associate ... pdated.pdf

kafka_vali ... guide.docx

Oozie Mcqs

Pyspark ass ... milestone

Pyspark Handbook v1

Page 1 / 28

Copy of clo ... rk-2016.pdf

```
# ----------------------------------------
# 1. Data Ingestion / Data Collection
# ----------------------------------------
spark = SparkSession.builder.appName("DataEngineeringFullProject").getOrCreate()

print("Step 1: Ingesting Data...")

# Load CSV File
df_raw = spark.read.option("header", True).csv("/mnt/data/raw/source_data.csv")
df_raw.show(3)

# Load JSON File
df_json = spark.read.option("multiline", True).json("/mnt/data/raw/source_data.json")
df_json.show(3)

# Load Parquet File
df_parquet = spark.read.parquet("/mnt/data/raw/source_data.parquet")
df_parquet.show(3)

# Simulate Kafka Stream Ingestion (static DF as demo)
kafka_simulated_df = df_raw.withColumn("source", lit("kafka"))
kafka_simulated_df.show(3)

# Merge datasets (if applicable)
df_combined = df_raw.unionByName(df_parquet, allowMissingColumns=True)

# ----------------------------------------
# 2. Data Storage / Data Lake Organization
# ----------------------------------------
print("Step 2: Storing raw data in Bronze layer...")
```

Pyspark al ... anagement