```
from pyspark.sql import SparkSession

# Initialize a Spark session
spark_session = SparkSession.builder \
    .appName("Parquet Load and Save Operation") \
    .getOrCreate()

# Read data from CSV file
data_frame = spark_session.read.format("csv")\
.option("header", "true")\
.option("inferSchema", "true")\
.load("s3://labs-jhabyar/nasa/data.csv")

# Store data as Parquet in HDFS
data_frame.write.parquet("/user/hadoop/parquet_data")

# Print the schema of DataFrame
data_frame.printSchema()

# Create a temporary SQL view
data_frame.createOrReplaceTempView("accessLogsView")

# SQL to find top 5 hosts by file access count
query_most_active_hosts = """
    SELECT host, COUNT(DISTINCT url) AS unique_file_count
    FROM accessLogsView
    GROUP BY host
    ORDER BY unique_file_count DESC
    LIMIT 5
```

```
"""
active_hosts_df = spark_session.sql(query_most_active_hosts)
active_hosts_df.show()

# SQL to find top 5 frequently accessed files
query_top_accessed_files = """
   SELECT url, COUNT(*) AS hits
   FROM accessLogsView
   GROUP BY url
   ORDER BY hits DESC
   LIMIT 5
"""
frequently_accessed_files_df = spark_session.sql(query_top_accessed_files)
frequently_accessed_files_df.show()

# SQL to find top 5 files by total traffic
query_high_traffic_files = """
   SELECT url, SUM(bytes) AS total_bytes
   FROM accessLogsView
   GROUP BY url
   ORDER BY total_bytes DESC
   LIMIT 5
"""
high_traffic_files_df = spark_session.sql(query_high_traffic_files)
high_traffic_files_df.show()

# Working with Parquet data
parquet_data_path = "hdfs:///user/hadoop/parquet_data/"
df_from_parquet = spark_session.read.parquet(parquet_data_path)
```

```
df_from_parquet.printSchema()

# Create a view from Parquet data
df_from_parquet.createOrReplaceTempView("parquetView")

# SQL query for most accessed files from Parquet data
query_parquet_top_files = """
   SELECT url, COUNT(*) AS hit_count
   FROM parquetView
   GROUP BY url
   ORDER BY hit_count DESC
   LIMIT 5
"""
top_files_from_parquet_df = spark_session.sql(query_parquet_top_files)
print("Most Accessed Files from Parquet Data:")
top_files_from_parquet_df.show()

# SQL query for top traffic files from Parquet data
query_parquet_traffic = """
   SELECT url, SUM(bytes) AS total_traffic
   FROM parquetView
   GROUP BY url
   ORDER BY total_traffic DESC
   LIMIT 5
"""
top_traffic_from_parquet_df = spark_session.sql(query_parquet_traffic)
print("Files with Highest Traffic from Parquet Data:")
top_traffic_from_parquet_df.show()
```