

Spark Azure Databricks

Databricks Spark Certification and beyond

Instructor: Prashant Kumar Pandey





Absolute Beginner to Specialization in Apache Spark and Azure Databricks





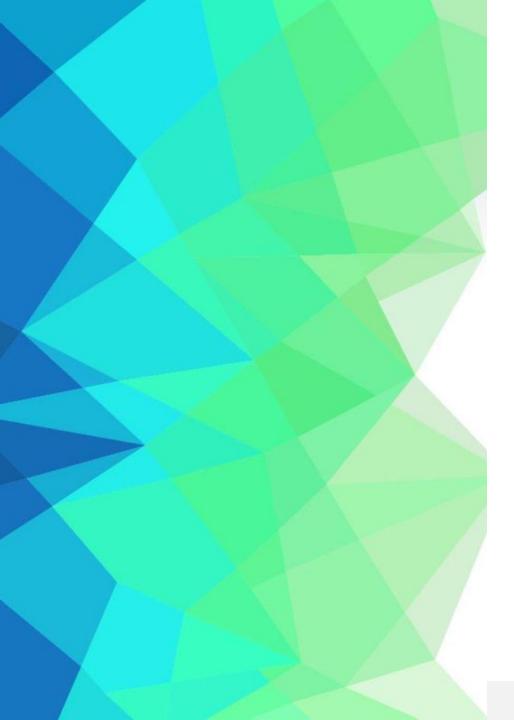
You are given a dataset at the following location.

/databricks-datasets/learning-spark-v2/sf-fire/sf-fire-calls.csv

- You are asked to do the following
 - Create a Spark data frame using the above file.
 - Create a Global Temporary View using the above data frame. Assume the view name is fire_service_calls_view.



- Create a Spark database demo_db
- Create a spark table fire_service_calls_tbl in the demo_db
 - The schema for the table must match the view schema created in the previous assignment
 - The table data file must be a parquet file
 - Make sure your notebook is re-executable
- Load data into fire_service_calls from the fire_service_calls_view



- Write Spark SQL queries to answer the following questions
 - 1. How many distinct types of calls were made to the fire department?
 - 2. What are distinct types of calls made to the fire department?
 - 3. Find out all responses or delayed times greater than 5 mins?
 - 4. What were the most common call types?
 - 5. What zip codes accounted for the most common calls?
 - 6. What San Francisco neighborhoods are in the zip codes 94102 and 94103
 - 7. What was the sum of all calls, average, min, and max of the call response times?
 - 8. How many distinct years of data are in the CSV file?
 - 9. What week of the year in 2018 had the most fire calls?
 - 10. What neighborhoods in San Francisco had the worst response time in 2018?



You are given a dataset at the following location.

/databricks-datasets/learning-spark-v2/sf-fire/sf-fire-calls.csv

- You are asked to do the following
 - Create a data frame using the above data
 - Transform the data frame to rename columns removing the space in the column names
 - Transform the data frame to fix the date and timestamp column types



- Write Spark data frame queries to answer the following questions using the data frame you created in the previous assignment
 - 1. How many distinct types of calls were made to the fire department?
 - 2. What are distinct types of calls made to the fire department?
 - 3. Find out all responses or delayed times greater than 5 mins?
 - 4. What were the most common call types?
 - 5. What zip codes accounted for the most common calls?
 - 6. What San Francisco neighborhoods are in the zip codes 94102 and 94103
 - 7. What was the sum of all calls, average, min, and max of the call response times?
 - 8. How many distinct years of data are in the CSV file?
 - 9. What week of the year in 2018 had the most fire calls?
 - 10. What neighborhoods in San Francisco had the worst response time in 2018?

