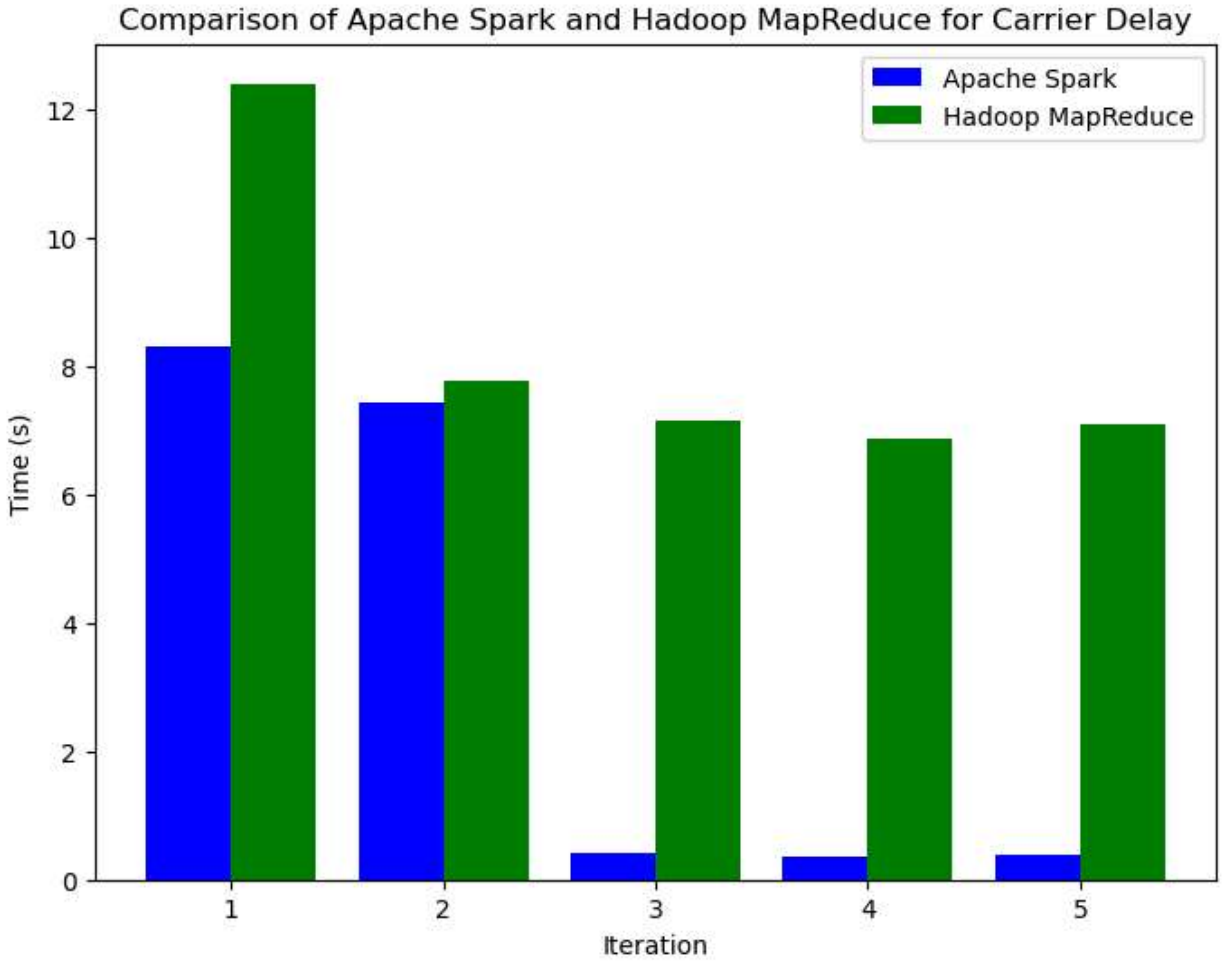
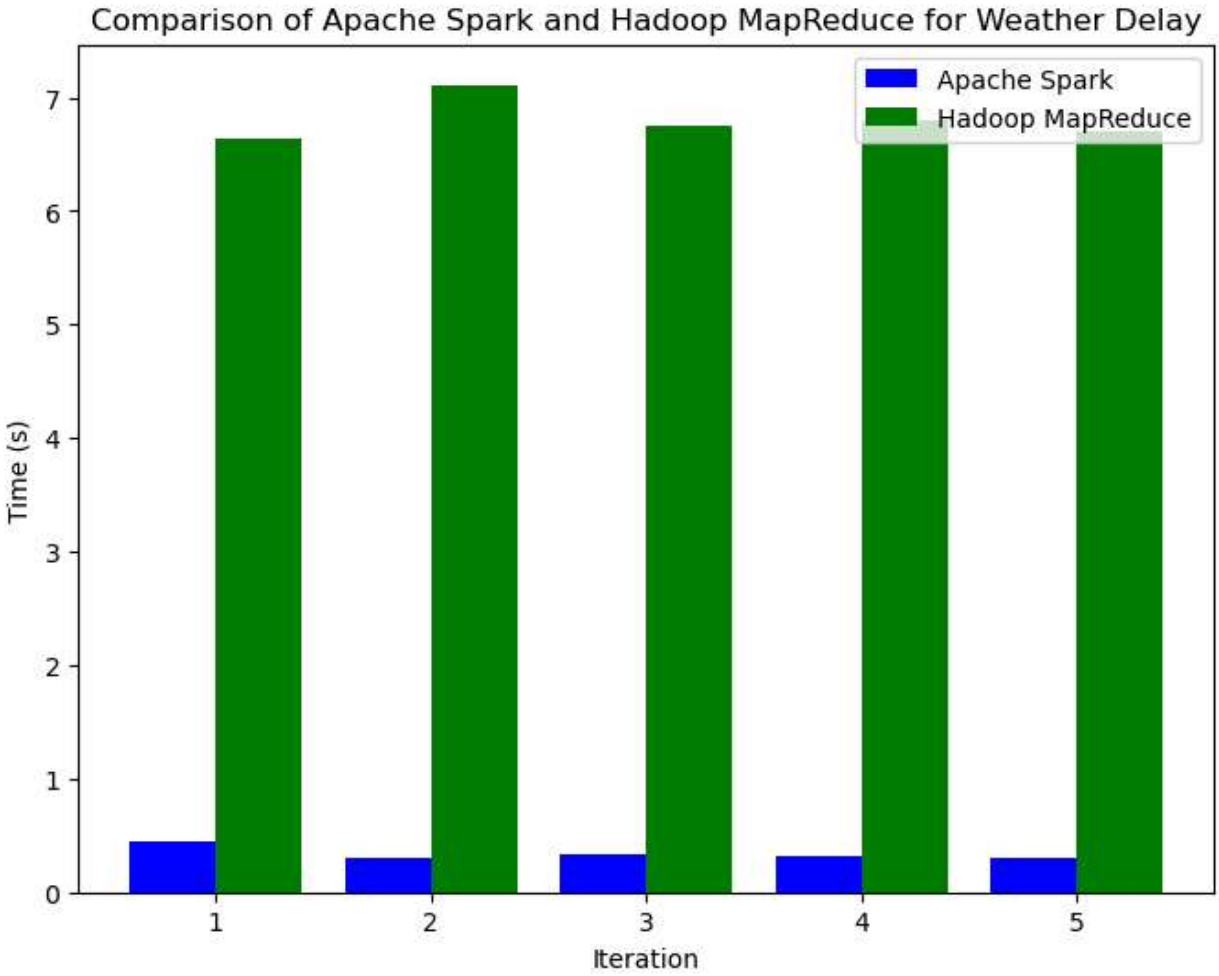
In [2]:
```
import pandas as pd

# Read the CSV file into a DataFrame
df = pd.read_csv('Apache Spark and MapReduce.csv', header=0)

# Print the column names
print(df.columns)
```

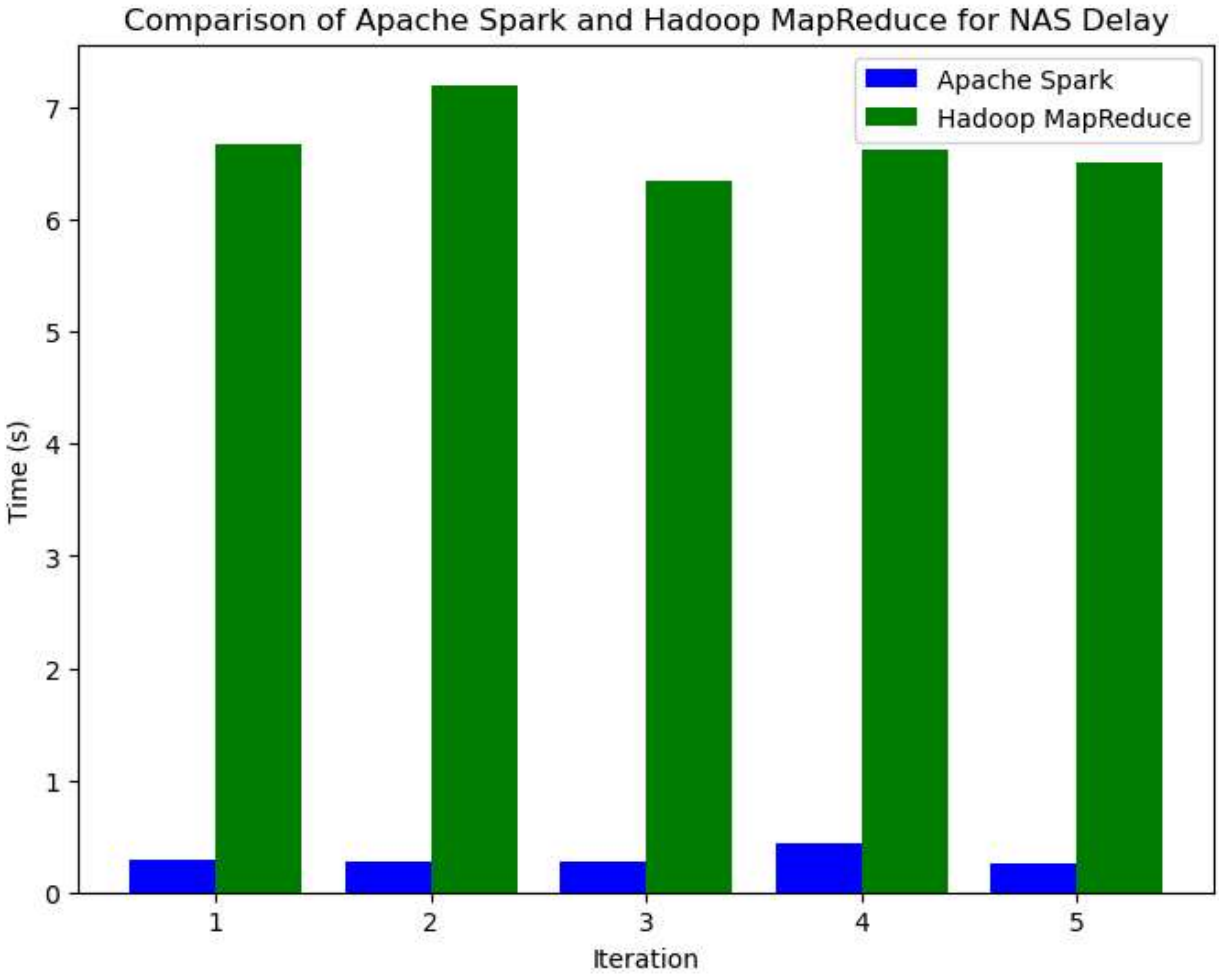
```
Index(['Query', 'Apache Spark Time (s)', 'Hadoop MapReduce Time (s)'], dtype='objec
t')
```

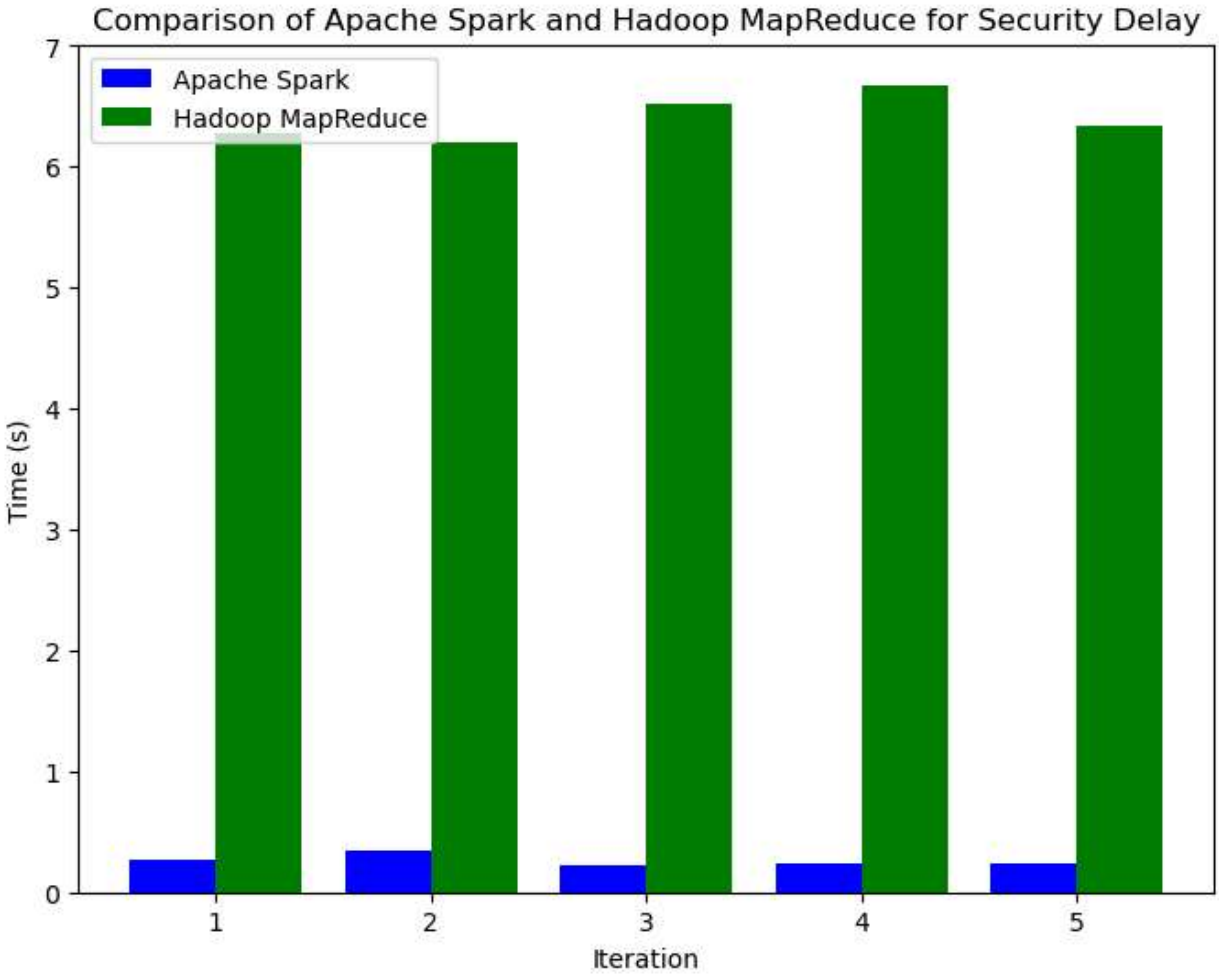
In [ ]:
```
```

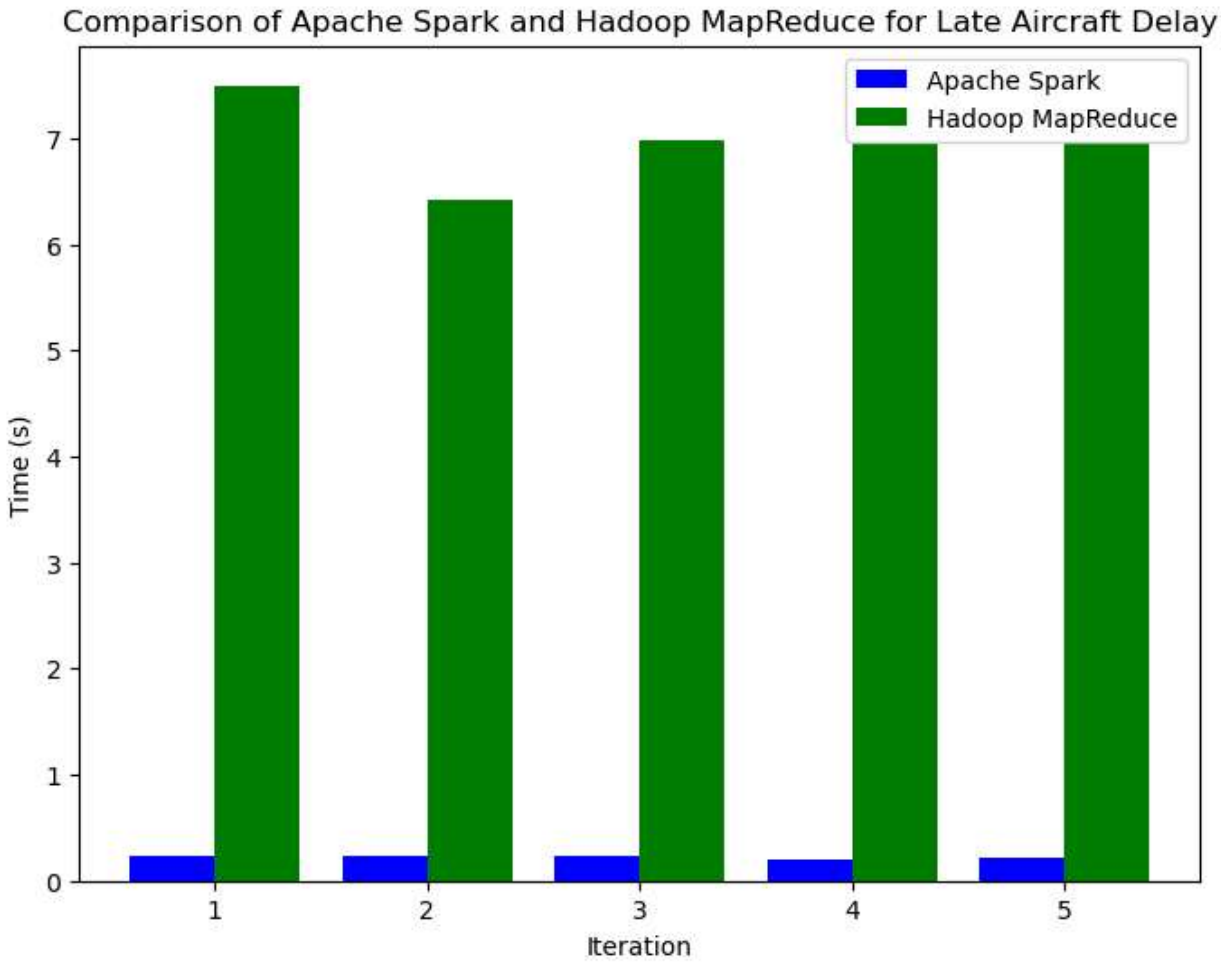
In [8]:
```
import matplotlib.pyplot as plt
import numpy as np

# Sample data for Apache Spark and Hadoop MapReduce running times for each query
all_spark_times = [
    [8.291, 7.441, 0.422, 0.35, 0.382],  # Carrier Delay query
    [0.452, 0.299, 0.33, 0.325, 0.299],  # Weather Delay query
    [0.288, 0.275, 0.276, 0.443, 0.26],  # NAS Delay query
    [0.262, 0.341, 0.221, 0.241, 0.232],  # Security Delay query
    [0.231, 0.236, 0.23, 0.201, 0.215]   # Late Aircraft Delay query
]

all_hadoop_times = [
    [12.398, 7.764, 7.138, 6.864, 7.084],  # Carrier Delay query
    [6.635, 7.111, 6.754, 6.792, 6.707],   # Weather Delay query
    [6.672, 7.202, 6.345, 6.621, 6.507],   # NAS Delay query
    [6.283, 6.201, 6.519, 6.677, 6.345],   # Security Delay query
    [7.498, 6.425, 6.982, 6.948, 6.972]    # Late Aircraft Delay query
]

# Queries
queries = ['Carrier Delay', 'Weather Delay', 'NAS Delay', 'Security Delay', 'Late Airc

# Function to plot bar chart for a specific query
def plot_graph(query):
    index = queries.index(query)
    spark_times, hadoop_times = all_spark_times[index], all_hadoop_times[index]
    iterations = np.arange(1, len(spark_times) + 1)

    plt.figure(figsize=(8, 6))
    plt.bar(iterations - 0.2, spark_times, width=0.4, label='Apache Spark', color='blu
    plt.bar(iterations + 0.2, hadoop_times, width=0.4, label='Hadoop MapReduce', color

    plt.xlabel('Iteration')
    plt.ylabel('Time (s)')
    plt.title(f'Comparison of Apache Spark and Hadoop MapReduce for {query}')

    plt.legend()
    plt.show()

# Generate bar charts for all queries
for query in queries:
    plot_graph(query)
```


Comparison of Apache Spark and Hadoop MapReduce for Carrier Delay
Apache Spark
Hadoop MapReduce
Time (s)
0
2
4
6
8
10
12
1
2
3
4
5
Iteration


Comparison of Apache Spark and Hadoop MapReduce for Weather Delay
Apache Spark
Hadoop MapReduce
Time (s)
0
1
2
3
4
5
6
7
1
2
3
4
5
Iteration


Comparison of Apache Spark and Hadoop MapReduce for NAS Delay
Apache Spark
Hadoop MapReduce
Time (s)
0
1
2
3
4
5
6
7
1
2
3
4
5
Iteration


Comparison of Apache Spark and Hadoop MapReduce for Security Delay
Apache Spark
Hadoop MapReduce
Time (s)
Iteration
0
1
2
3
4
5
6
7
1
2
3
4
5



In [17]:

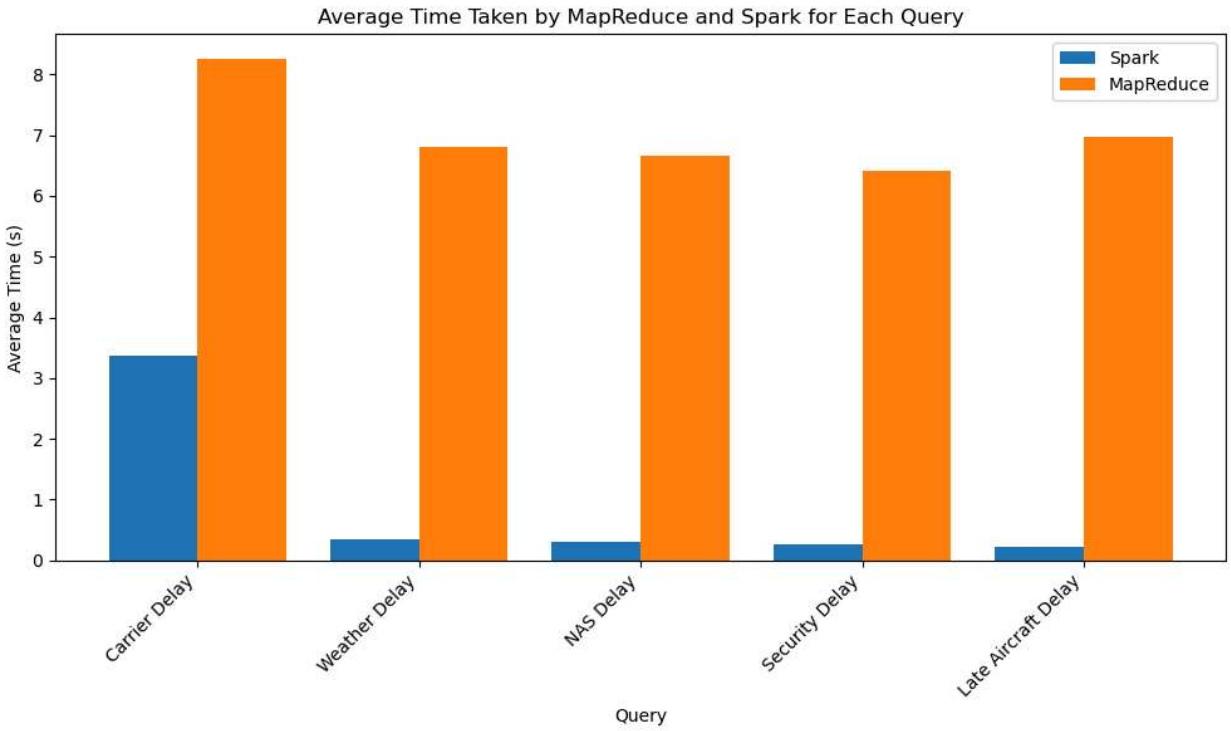
```python
from tabulate import tabulate

# Sample data for Apache Spark and Hadoop MapReduce running times for each query
all_spark_times = [
    [8.291, 7.441, 0.422, 0.35, 0.382],  # Carrier Delay query
    [0.452, 0.299, 0.33, 0.325, 0.299],  # Weather Delay query
    [0.288, 0.275, 0.276, 0.443, 0.26],  # NAS Delay query
    [0.262, 0.341, 0.221, 0.241, 0.232],  # Security Delay query
    [0.231, 0.236, 0.23, 0.201, 0.215]   # Late Aircraft Delay query
]

all_hadoop_times = [
    [12.398, 7.764, 7.138, 6.864, 7.084],  # Carrier Delay query
    [6.635, 7.111, 6.754, 6.792, 6.707],   # Weather Delay query
    [6.672, 7.202, 6.345, 6.621, 6.507],   # NAS Delay query
    [6.283, 6.201, 6.519, 6.677, 6.345],   # Security Delay query
    [7.498, 6.425, 6.982, 6.948, 6.972]    # Late Aircraft Delay query
]

# Queries
queries = ['Carrier Delay', 'Weather Delay', 'NAS Delay', 'Security Delay', 'Late Airc

# Calculate the average times for each query
spark_avg_times = [sum(times) / len(times) for times in all_spark_times]
hadoop_avg_times = [sum(times) / len(times) for times in all_hadoop_times]

# Create table data
table_data = []
```

```
for query, spark_time, hadoop_time in zip(queries, spark_avg_times, hadoop_avg_times):
    table_data.append([query, spark_time, hadoop_time])

# Display table
print(tabulate(table_data, headers=['Query', 'Spark Avg Time (s)', 'Hadoop Avg Time (s
```

```
+--------------------+----------------------+-----------------------+
| Query              |   Spark Avg Time (s) |   Hadoop Avg Time (s) |
+====================+======================+=======================+
| Carrier Delay      |               3.3772 |                8.2496 |
+--------------------+----------------------+-----------------------+
| Weather Delay      |               0.341  |                6.7998 |
+--------------------+----------------------+-----------------------+
| NAS Delay          |               0.3084 |                6.6694 |
+--------------------+----------------------+-----------------------+
| Security Delay     |               0.2594 |                6.405  |
+--------------------+----------------------+-----------------------+
| Late Aircraft Delay |              0.2226 |                6.965  |
+--------------------+----------------------+-----------------------+
```

In [18]:

```
# Calculate the average times for each query
spark_avg_times = [np.mean(times) for times in all_spark_times]
hadoop_avg_times = [np.mean(times) for times in all_hadoop_times]

# Plotting the bar graph
plt.figure(figsize=(10, 6))
plt.bar(np.arange(len(queries)) - 0.2, spark_avg_times, width=0.4, label='Spark', alig
plt.bar(np.arange(len(queries)) + 0.2, hadoop_avg_times, width=0.4, label='MapReduce',

# Adding labels and title
plt.xlabel('Query')
plt.ylabel('Average Time (s)')
plt.title('Average Time Taken by MapReduce and Spark for Each Query')
plt.xticks(np.arange(len(queries)), queries, rotation=45, ha='right')
plt.legend()

# Show plot
plt.tight_layout()
plt.show()
```


Average Time Taken by MapReduce and Spark for Each Query
Spark
MapReduce
Average Time (s)
0
1
2
3
4
5
6
7
8
Carrier Delay
Weather Delay
NAS Delay
Security Delay
Late Aircraft Delay
Query

In [ ]: