# Summary

1	Version	3
	Introduction	
	Data Sourcing/Extraction	
	Data Transformation and Data Analysis	
5	Data Loading	14
6	Conclusion	16
7	References	17

# 1 Version

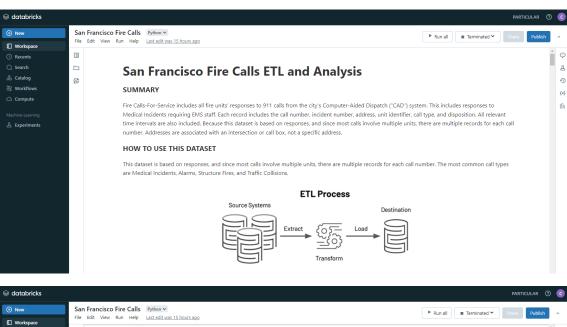
This document was created by Cleber Zumba de Souza and can be distributed freely, as long as the source is mentioned.

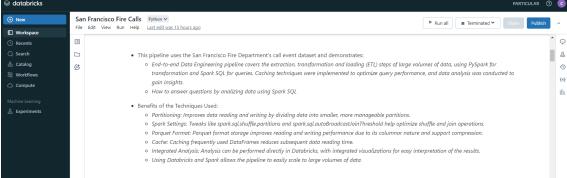
Versão	Ação	Data
1.0	Document creation	21/06/2024

## 2 Introduction

### San Francisco Fire Calls

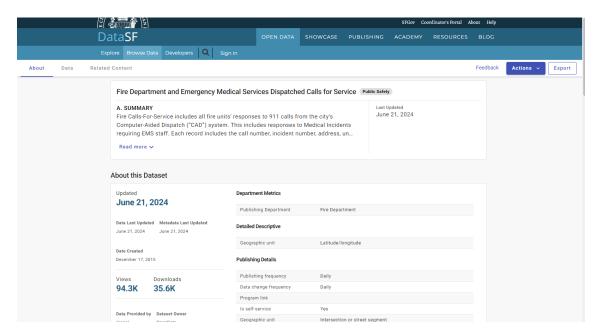
This Databricks notebook is the end-to-end, showing how to use DataFrame and Spark SQL for common data analytics patterns and operations on a San Francisco Fire Department Calls dataset.



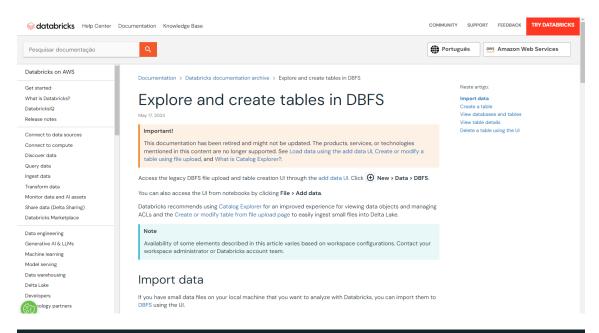


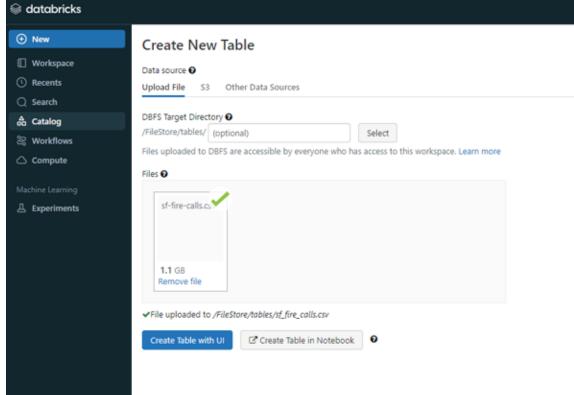
# 3 Data Sourcing/Extraction

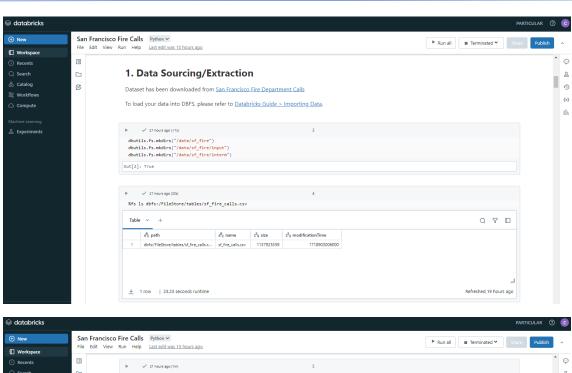
# Dataset has been downloaded from <u>San Francisco Fire</u> <u>Department Calls</u>

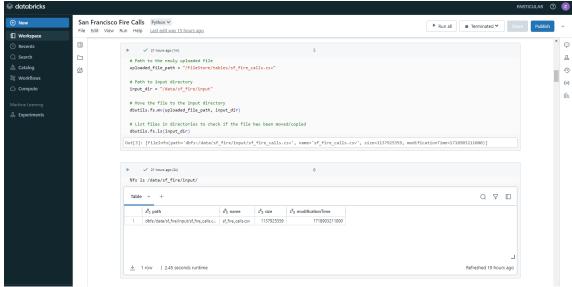


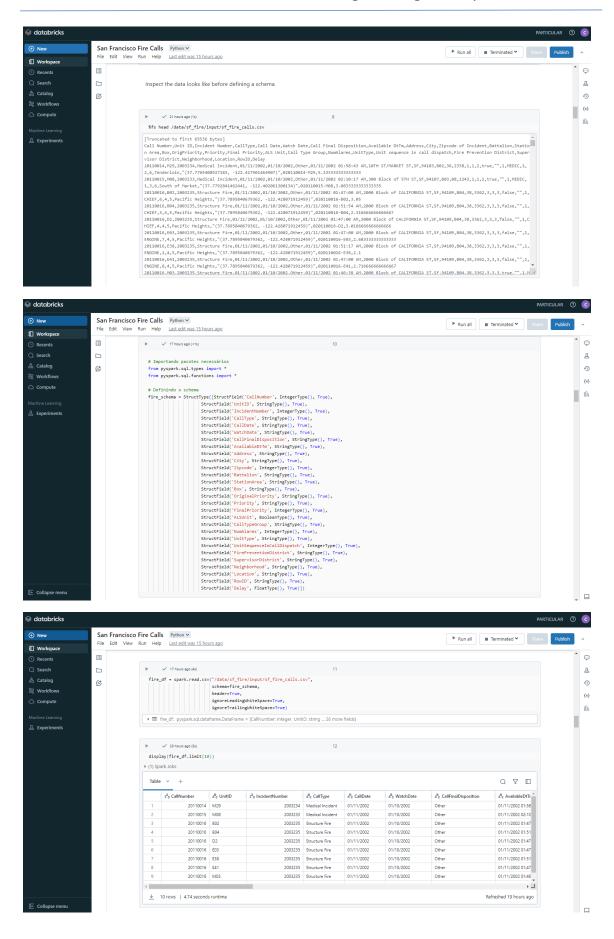
# To load your data into DBFS, please refer to <u>Databricks Guide ></u> <u>Importing Data</u>.



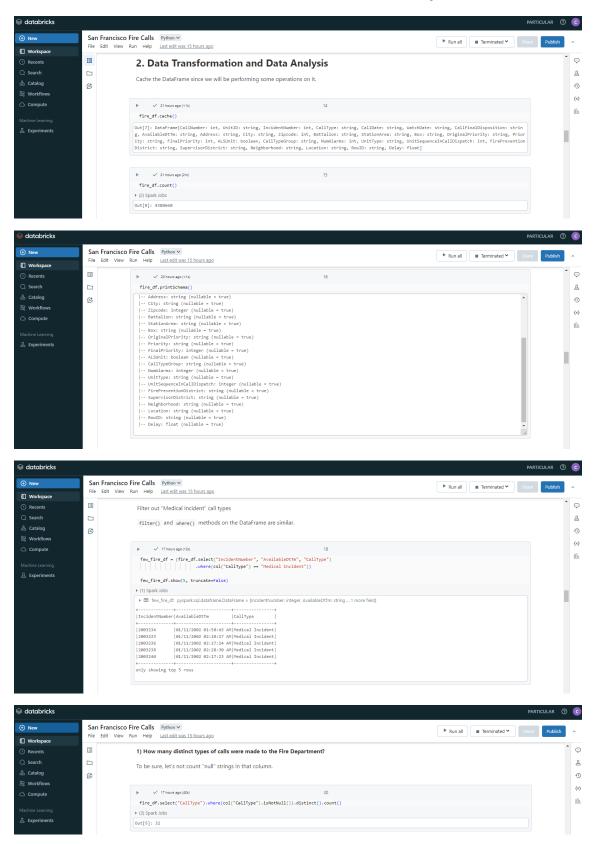


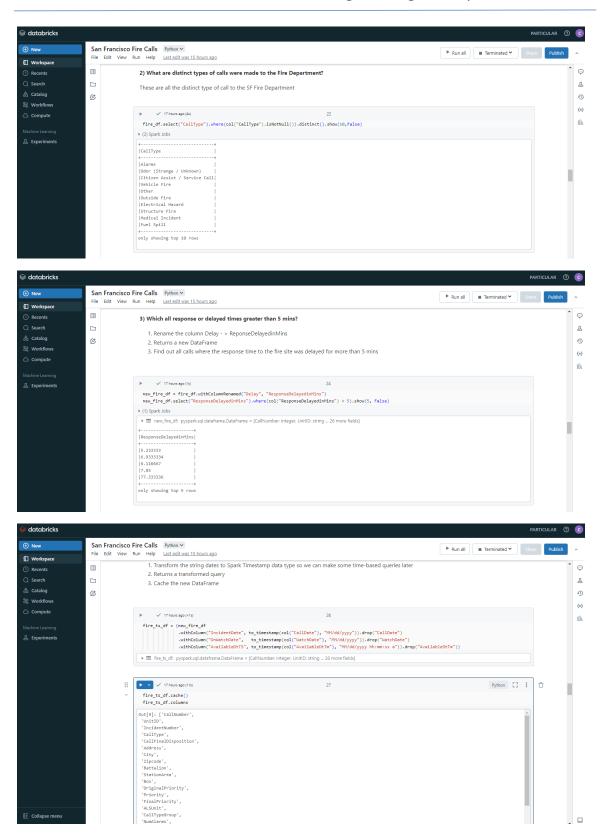


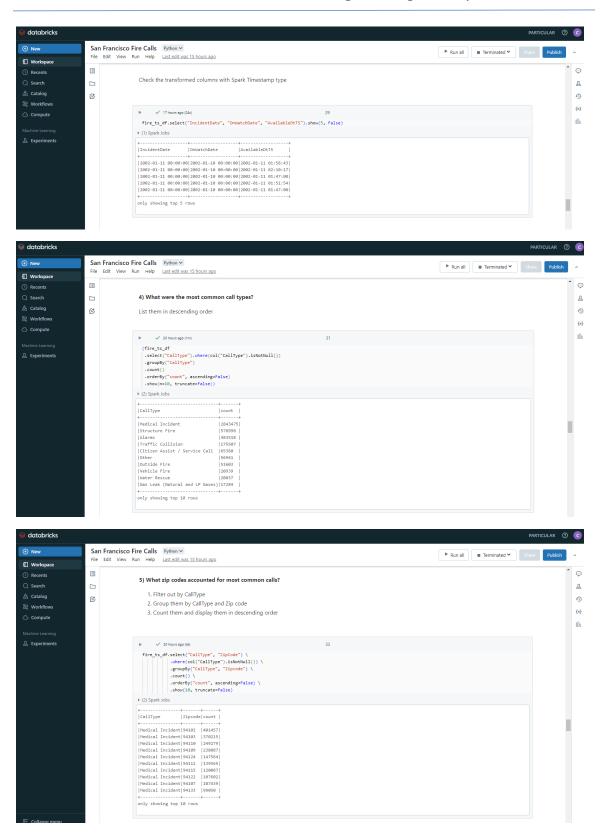


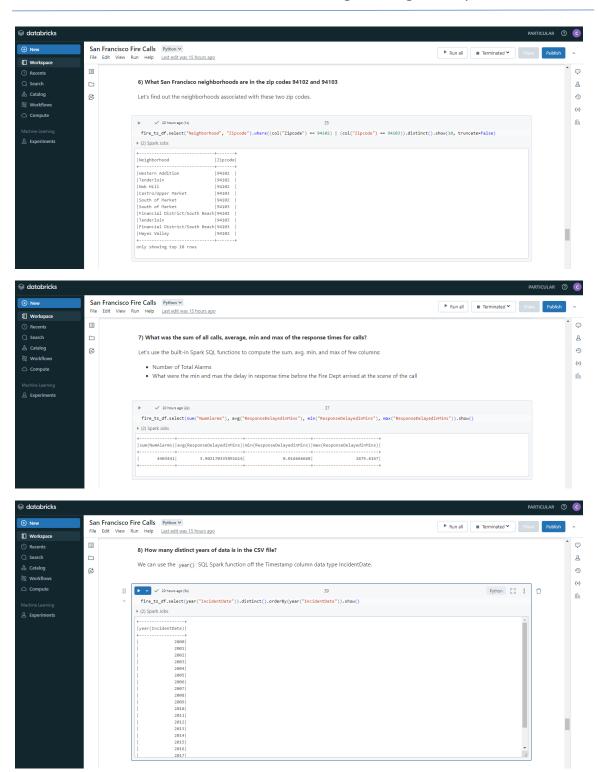


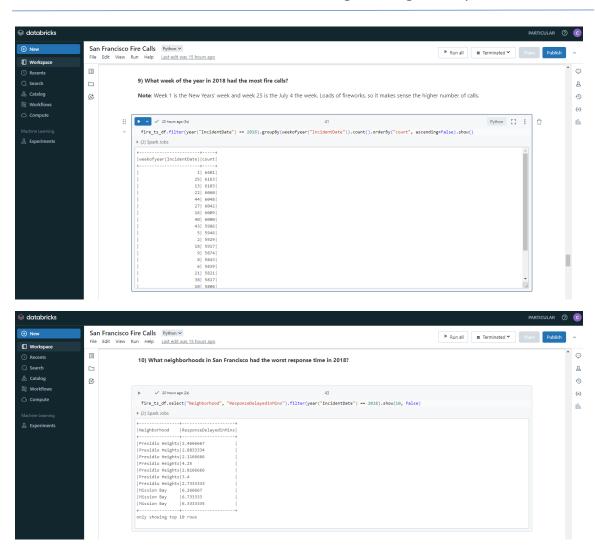
# 4 Data Transformation and Data Analysis



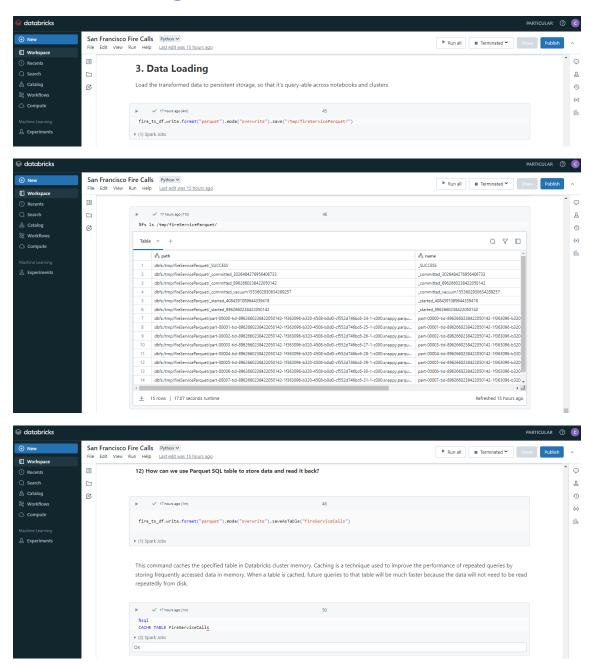


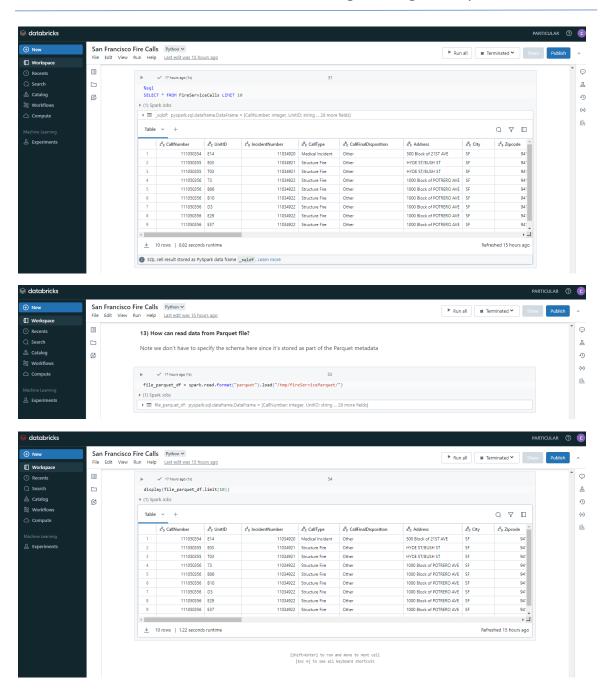






# 5 Data Loading





# 6 Conclusion

In this article, we explored the powerful capabilities of PySpark and SparkSQL within the Databricks platform, focusing on data analysis and transformation. Through our hands-on examples and code implementations, we demonstrated how to effectively leverage these tools to process and analyze large datasets efficiently.

We perform data extraction, transformation, and loading (ETL) processes, highlighting the seamless integration of PySpark for scalable data processing. Additionally, we utilized SparkSQL to perform queries, showcasing its ability to handle large-scale data analytics with ease.

We use techniques the performance optimization techniques, such as caching, which are essential for achieving high-performance data processing. By following best practices and utilizing the features of Databricks, we were able to enhance the efficiency and speed of our data analysis workflows.

I hope this article serves as a helpful guide for those looking to harness the power of PySpark and SparkSQL in their data engineering and analytics projects. The journey of data analysis is filled with continuous learning and innovation.

Thank you for taking the time to read this article. Your feedback and insights are always welcome as we continue to learn and grow in the ever-evolving field of data science.

Happy coding!

Cleber Zumba de Souza

# 7 References

PARSIAN, Mahmoud. **Data Algorithms with Spark**. Sebastopol, California, United States: O'Reilly Media, 2022.