Marcelo Peres

Big Data Engineer

Email: brmarcelo.peres@gmail.com

LinkedIn: [My Profile](#)

Github: [My Repo!](#)

I am a Big Data Engineer with 8 years of experience specializing in Python, PySpark, and Spark SQL.
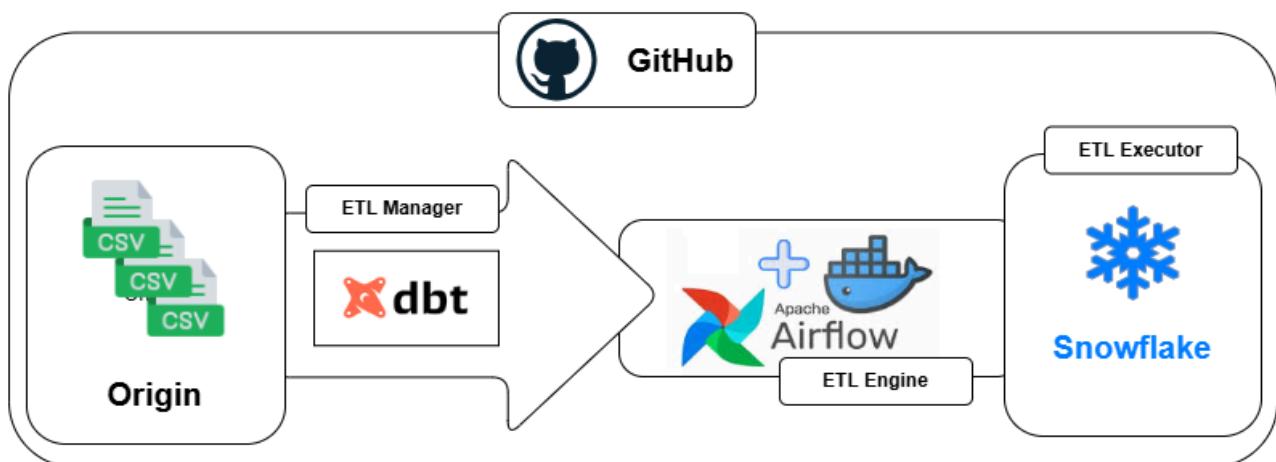
For the past 2 years, I have been working extensively in the Databricks ecosystem within the AWS platform, leading efforts to migrate complex ETL/ELT processes from EMR clusters.

I have hands-on experience with data orchestration tools, particularly Apache Airflow and Control-M, where I designed and optimized workflows for scalable data pipelines. I have been making a lot efforts to work with Airflow and enable me to do efficiently schedule, monitor, and maintain reliable data operations across distributed systems.

With a passion for collaboration, I excel in fostering teamwork by sharing knowledge and best practices to help teams grow and achieve business goals. I firmly believe that the greatest asset a professional can have is the ability to share skills and empower others without fear of competition.

My commitment to technical excellence and continuous learning drives me to stay at the forefront of emerging technologies, always delivering value and maintaining high development standards.

[Github Project Example](#)

Work Experience

Senior Big Data Engineer | Apllos Solutions | 02 years and 05 month(s)

From Jul 2023 to nowadays

Project Audlab | 11 month(s)

From Jan 2025 to nowadays

Now developing into data lake Delta Lake tables and Postgre tables for Gerdau Commercial Assets. Some tables in Postgres for apps interaction and decision making through PBI Reports.

Project GAB - Industrial Maintenance | 03 month(s)

From Sep 2024 to Dec 2024

In charge of migrate and rebuild ETL process in Pyspark. The original scope was only migrate the scripts and put in a CICD environment, and configure jobs schedule in Control-M plataform.

However, due to a delay faced through PBI workflows, the project suffered a change bringing a big part of process to databricks. New tables were created, steps were adpted and the project ajusted.

Final charge time has been reduced and the PBI charges becmes faster.

For three large tables the gains were awesome as below mentioned:

- **The first table took 01h20 before.**
- **The second one took 03h13 before.**
- **The third one took 01h30 before.**

Now each one of them is taking 10 minutes or less for loading PBI Insights.

Project Databricks Migration Audlab | 01 years and 02 month(s)

From Jul 2023 to Sep 2024

In charge of data ingestions.

Audlab is an Audit department in Gerdau Company. Gerdau is an important enterprise in Brazil that works with iron and steel, producing several tools for different segments in the world.

In this department I provide a full control of gitlab and deploying scripts.

Control - M is responsible to execute EMR clusters and scripts in python and pyspark through different assets. Producing ETL and Data Science Scripts for machine learning and data transformation.

I was one of core people helping data migration steps from AWS tools to Databricks. EMR clusters now replicated accordingly in Databricks.

The migration was well successed. Finnaly working in Databricks.

Big Data Engineer | A3 Data | 03 month(s)

Dec 2022 to Mar 2023

Datalake project - Stellantis Company | 03 month(s)

From Dec 2022 to Mar 2023

<< GCP - Google Cloud Platform >>

Responsible for creating tools to interect data at plataform.

Tools used:

- Terraform;
- Google Storage;
- BigQuery;
- Google Functions;
- Airflow;
- Python;
- SQL.

Data Engineer Consultant | Via Consulting | 05 month(s)

Apr 2022 to Sep 2022

Project - Gol Spend & Get| 01 month(s)

Ago 2022 to Ago 2022

Smiles S.A.

AWS Python lambda function that validates files to be called by an API.

Resources:

- Cloud Formation | Python with unit tests | Github release info | Jenkins to observe github uploaded code | Sonar for code quality | End of Devops stack with deploy.

Project - Zendesk Replication | 01 month(s)

Jul 2022 to Jul 2022

Continuous pushing of data in a AWS environment using tools like:

- Pyspark Spark Glue jobs | Cloud Formation | AWS Athena | Apache Hudi metadata for data governance.

Project Unimed - Stuffed Wallet | 03 month(s)

Apr 2022 to Jun 2022

Unimed Insurance

Colaborating with the team in a ETL process using tools like:

- Informatica Powercenter | PLSQL - Oracle.

The idea of the project is a campaign that supports the score of the company's brokers.

Data Engineer Consultant | BI4ALL | 10 month(s)

Jul 2021 to May 2022

Project - BI in AWS | 10 month(s)

Jul 2021 to May 2022

Manserv Company

Responsible for creating data pipelines to push data in S3 using python AWS lambda funtions. Using packages such as awswrangler, xmltodict, json and much more. Being involved in great projects with the relevant skills, accessing different APIs from several providers.

Some of these ones are:

- Volvo
- Nuntec
- Komatsu
- Komtrax
- Caterpillar and many others.

Also Transforming XML and JSON API extrations into tabular data to be recorded in parquet table format, grating a better use of S3 bucket as much as gaing performance in a compressed file format.

IT Analyst | CLIF | 02 years and 02 month(s)

Apr 2019 to Jun 2021

Systems and BI Management | CLIF | 02 years and 02 month(s)

Apr 2019 to Jun 2021

CLIF - Centro Logístico Integrado FastCargo S.A

Systems Management

- Responsible for managing the SARA system, Logix, My Checklist & Customer Service Portal;
- Application of SARA customs system update packages;
- Updating of stored procedures in SQL;
- SQL Trigger Updates;
- Module Reviews among others;
- Creating a Python script to insert and update ships and related data on the Customer Service Portal;

BI Management

Creation of reports in Power BI as:

- Registration management and control;
- Unit management and control;
- Transport management and control;
- Ship management and control;

The above reports are synchronized and loaded into Power BI containing data from 02 different databases and a loading of data from a Python Script connected to an API available by Porto Itapoá containing the dates of ship movements, thus making it possible to cross all information among those different data to better management of customer assets.