# Gustavo

## SUMMARY

Gustavo is a data engineer with over 10 years of experience in data engineering and cloud computing. He designs, develops, and implements data solutions using AWS and Python to support the company's data-driven products and services.

His expertise includes data mining, integration, transformation, quality assurance, data modeling, performance testing, and pipeline orchestration. Gustavo utilizes technologies such as PySpark, Kafka, Airflow, and AWS Glue to build scalable and reliable data pipelines that handle data from various sources and formats.

He works closely with data analysts, data scientists, and product managers to align data solutions with customer and user needs. His mission is to leverage data and technology to create value and impact for both business and society.

Gustavo has a C1 level of English.

## CORE SKILLS

- ETLs
- Python (Pandas, Dask and Polars)
- PySpark (AWS EMR, AWS Glue Jobs, and Google DataProc)
- SQL (SQL Server, Postgre, MySQL)
- NoSQL (AWS DynamoDB, Google Big Table, and MongoDB)
- AWS (Glue, S3, Lambda, EMR, CloudFormation, CloudWatch, Redshift, Athena, Aurora, SNS, Kinesis and MSK)
- Google Cloud Platform (Cloud Storage, DataProc, BigQuery, PubSub, Cloud Functions, Cloud Run)
- Airflow (Google Cloud Composer and MWAA)
- Data Visualization (Power BI, Tableau, and Plotly)
- Snowflake
- dbt

## WORK HISTORY

### XXXXXX

Data Engineer

*July 2022 - Present*

- Created Data Lakehouse using S3 and Glue Data Catalog.

- Data ingestion integration with several data providers from different sources like REST API, GraphQL, CDC, DMS, PostgreSQL, and MongoDB
- Designed pipelines templates for AWS Glue using AWS CloudFormation
- Pipeline migration from SQL to PySpark
- IaC migration from CloudFormation to Terraform
- Streaming data ingestion pipelines with AWS Kinesis + Apache Flink
- Near real time pipelines with AWS MSK + Apache Flink
- Batch pipeline orchestration with Apache Airflow
- Skills: Apache Flink • Apache Kafka • dbt • Data Engineering • Kubernetes • Terraform • Apache Airflow • Snowflake • Amazon Web Services (AWS) • Python • SQL • Google Cloud Platform (GCP) • Apache Spark

**number8**

Data Engineer

*February 2022 - February 2023*

- Lakehouse architecture development with AWS Glue Data Catalog.
- Pipeline orchestration with AWS Glue Jobs, Triggers, Crawlers with Workflow
- Ingestion of REST API, GraphQL, MongoDB and AWS DMS sources
- Resource provisioning with AWS CloudFormation
- Full pipeline running in AWS Glue, with Workflow, Triggers, Crawlers, and Jobs getting data from different sources like MongoDB, MySQL, and PostgreSQL and creating a Glue Data Catalog to connect through Redshift Spectrum and Athena or directly in Redshift Storage
- Making data available in AWS Redshift
- Skills: Data Engineering • Kubernetes • Terraform • Apache Airflow • Amazon Web Services (AWS) • Python • SQL • Google Cloud Platform (GCP) • Apache Spark

**ANBIMA**

Data Engineer

*April 2020 - February 2022*

- Created Data Lakehouse using S3 and Glue Data Catalog.
- • Pipeline migration from on-premises servers and legacy and graphical interfaces tools to AWS Glue with Python and PySpark Jobs, orchestrating with Triggers and Crawlers through a Workflow.
- • 10TB of data migrated from SQL Server (2008 – 2012) to AWS S3, Redshift Spectrum, and Athena
- Data modeling to provide a self-service schema for business analysts and integrated with Power BI from Redshift.

- Improved critical daily pipeline from 8 hours to 20 minutes, handling around 60 GB/daily data.
- Conducted data journey workshops for business analysts.
- Created pipeline templates for easy maintenance.
- Designed CI/CD pipeline with AWS CodePipeline and CloudFormation
- Data ingestion integration with several data providers from different sources like REST API, GraphQL, and CDC.
- Pipeline orchestration with AWS MWAA (Amazon Managed Workflows for Apache Airflow) using EMR clusters and Lambda
- On-premise Oracle Database 12c with OLTP and OLAP models migrate to AWS. The OLTP model was migrated to DynamoDB, improving the application team time to market development, and the OLAP model was migrated to a single node Redshift cluster.
- Skills: Data Engineering • Terraform • Apache Airflow • Amazon Web Services (AWS) • Python • SQL • Google Cloud Platform (GCP) • Apache Spark

**Febrafar**
Data Engineer
*March 2018 - March 2020*

- Responsible for implementing a data-driven culture, migrating all the Excel reports to Python and PySpark
- Pentaho pipeline migration to PySpark increasing performance by 800%
- Created Data Lake using Google Cloud Storage and Google BigQuery
- Data modeling to provide a self-service schema for business analysts and integrated with Power BI.
- Created automated pipeline using AWS StepFunctions, Lambda, and EMR to deliver 10k+ personalized reports for customers
- Pipeline orchestration with Airflow (Google Cloud Composer) andDataProc that handled ingestion of 3TB data/monthly from SQL Servers and FTP Servers that created dozens of star schemas and denormalized models for customer consumption
- Skills: Data Engineering • Amazon Web Services (AWS) • Python • SQL • Google Cloud Platform (GCP) • Apache Spark

**Roche**
Business Intelligence Engineer
*March 2018 - March 2020*

- Created data warehouse in SQL Server and PySpark to obtain data from Salesforce, making it possible to develop KPI of customer's journey.
- Developed KPIs and dashboards in Power BI, allowing a full view of the forecast process in the organization and strategic focus.

- Migrated VBA and Excel reports to Python
- Skills: Data Analysis • Python • SQL

**Banco Santander**
Business Intelligence Analyst
*August 2014 - June 2017*

- Automated Excel reports, reducing errors and inconsistency.
- Improved flow of information, developing a centralized pipeline using SQL Server Triggers and Stored Procedures.
- Developed KPIs and dashboards for strategic focus.
- Skills: Data Analysis • SQL

## EDUCATION

Tech Degree: Computer Science Programming
Faculdade Impacta, São Paulo, Brazil
2020/01 – 2021/12

MBA: Data Engineering
XP Educação, São Paulo, Brazil
2022/07 – current

## LANGUAGES

Portuguese: Native

English: C1