Omar Leonardo Sánchez Granados

omarleonardosanchez@hotmail.com | LinkedIn: omar-sanchez-tb1 | +1 (412)-909-7301 | GitHub: Omarlsg98

With three years of data engineering expertise and a solid foundation in machine learning, I have cultivated a comprehensive skill set that spans from data source to consumption. At Minka, I demonstrated a proactive approach by building a Data Warehouse from scratch, while at Factored, I honed my skills in designing and deploying a variety of data pipelines. My academic achievements at Carnegie Mellon University, including a high CQPA, underscore my commitment to impactful technological advancement through continuous learning and improvement. I bring to the table a robust track record of designing and executing large-scale, data-centric solutions, coupled with an avid interest and growing proficiency in AI.

SKILLS

Programming languages: Python, SQL, Java, Bash, JavaScript (Node.js), Cypher (Neo4J)
Databases: MySQL, Google BigQuery, PostgreSQL, Neo4J, AWS Redshift, AWS S3, AWS DynamoDB
Tools: PyTorch, Scikit-Learn, Pandas, NumPy, DBT, Apache Airflow, Looker (Data Studio), PySpark, Google Pub/Sub, Google Dataflow, AWS Lambda, AWS Step functions
Others: AWS, GCP, PyTest, WanDB, Selenium, REST API, Jupyter, RegEx, Git, Docker, Unix, IaC/Terraform, Jenkins

EXPERIENCE

Data Engineer Intern, Amazon — May 2023 - Aug 2023

- Designed and developed an AWS-based batch NLP service for cleaning, storing, processing, and aggregating open-ended questions using AWS S3, AWS Athena, AWS Step Functions, AWS Lambda, and AWS Comprehend while employing only one-third of the time allotted for this.
- Established a full CI/CD pipeline with unit and integration testing, for the deployment of the solution using AWS CloudFormation, PyTest, and internal DevOps tool.
- Integrated the NLP service with a third-party API, serving as the user interface for enhanced UX.

Data Engineer Trainee, Factored — Apr 2022 - Jul 2022

- Retrieved data from REST APIs to generate automated reports using Python, AWS Lambda, Docker, AWS S3, AWS DynamoDB, AWS SNS, and AWS Step Functions. Deployed using Serverless Framework.
- Designed and implemented a layered data lake using AWS S3, the AWS Glue data catalog, and PySpark.
- Developed a pipeline to extract data from both PostgreSQL and S3 to create a data warehouse following Kimball's modeling technique in AWS Redshift using Python, Airflow on AWS EC2, AWS Glue, Apache Spark, DBT, and SQL.
- Implemented CI/CD pipelines to automate infrastructure deployment using GitHub Actions and Terraform.
- Consumed and transformed streaming data to produce aggregated statistics, store data, and create real-time dashboards using Python, AWS Kinesis, AWS S3, and Splunk.

Data Engineer, Bluetab (an IBM Company) — Dec 2021 - Apr 2022

- Built, tested, and documented several data pipelines using Apache Spark, Control-M, and other proprietary big data tools including Datio for an international bank.
- Deployed hundreds of Control-M data workflows and automated the creation of workflow-definition XML files using Python, RegEx, and Pandas, decreasing by a factor of 4 the time-to-production of new pipelines.
- Ensured data quality and enforced rules of completeness, consistency, integrity, etc. according to data governance expectations.

Software Engineer Analyst, Scotiabank — May 2021 - Dec 2021

- Built a tool using Selenium Web Driver, Python, Bash, and Docker to automatically execute tests to evaluate the impact of change on risk metrics, making the whole process four times faster.
- Led and orchestrated deployments from development to production using Jenkins and Bitbucket.

Support Data Analyst, Minka, Inc. — Feb 2020 - Apr 2021

- Designed, built, and maintained single-handedly and proactively a GCP-hosted Bigquery data warehouse to support complex operation reports, accounting reports, and ad hoc analysis for an online pay tech company in Latin America.

The data warehouse enabled a holistic view of a range of data sources (i.e. MySQL, Google Datastore, Google Cloud Logging, CSV files, spreadsheets, Neo4J/API). Leveraged the power of GCP (Cloud functions, Dataflow/Apache Beam, Dataprep, Bigquery, Google Cloud Storage, Pub/Sub), Data Build Tool (DBT), Javascript, and Python. This helped the business fulfill several data needs and surface costly mistakes that required attention by the C-suite.

- Designed, built, and maintained various analytical reports using SQL and Data Studio (now Looker) oriented to C-level executives.
- Implemented integrations between Google Bigquery and Spreadsheets to serve different reports and tables to provide a familiar and friendly interface to the warehouse to non-technical users.
- Developed several tools using Javascript and Node.js to automate and optimize operative processes decreasing by a factor of 10 the errors to be manually checked and fixed.
- Identified, analyzed, and resolved complex errors within the service using SQL and Bigquery, leveraging a deep understanding of the REST API services and synchronous/asynchronous messaging styles.
- Automated several REST API tests using Postman and Javascript, decreasing the testing time to just one-fifth.

EDUCATION

MS, Information Systems Management - Business Intelligence & Data Analytics, **Carnegie Mellon University** (2023)

- CQPA: 4.12/4.0
- Coursework: 11-785 Introduction to Deep Learning, 10-601 Introduction to Machine Learning, 11-667 Large Language Models, 95-828 Machine Learning for Problem Solving, 95-865 Unstructured Data Analytics.
- Worked as a Teaching Assistant (TA) for the courses Unstructured Data Analytics for two semesters and Object Oriented Programming with Java for 1 semester. Served as an officer at the Colombian Student Association.

BS, Economics and International Finance, Universidad de La Sabana (2021)

BS, Informatics Engineering, Universidad de La Sabana (2020)

- CQPA: 4.5/5.0
- Served on the team for the robotics football world championship “Robocup” in 2019 that reached semi-finalist status in the “Shield Category”. Developed a path planning algorithm and main behaviors for the striker role.

COURSES: ***Apache Airflow Fundamentals Certified*** and ***Apache Airflow DAG Authoring Certified*** (Astronomer Certifications), ***Data Engineering with Google Cloud*** (Coursera), ***Data Engineering Nanodegree*** (Udacity), ***AI Product Manager*** (Udacity), ***Deep Learning Specialization*** (Coursera).

PROJECTS

- Supply Chain Warning System - Designed a system to proactively identify weak points in the supply chain to prevent and mitigate multi-million dollar disruption scenarios. Trained several Neural Networks (including Graph Neural Networks GNN) using TensorFlow to learn topological patterns of the supply chain extracted with Neo4J, predict the impact, and suggest vulnerable nodes via backpropagation on the input topologies.
- ClaudioLM - Designed and trained a hierarchical token-based variant of MusicLM transformer architecture for text-conditioned generation in the raw audio domain using PyTorch. Implemented a pipeline to download, clean, and trim songs, and soft-label them automatically employing LLMs using AWS ECS and AWS S3.
- OpenSea-Autobot - Developed scripts using Selenium and Python to extract social media information and automate the process of publishing and listing NFTs on the OpenSea NFT marketplace.