John García

jhgarciam@gmail.com   +1 (917) 710-2437   linkedin.com/in/jhgarciam

Summary

Senior Data Engineer & Technical Lead with deep experience building scalable data solutions across fintech, healthcare, retail, and research. I specialize in modernizing data infrastructure, leading teams, and delivering high-impact pipelines in cloud environments. Proven ability to build from scratch, migrate legacy systems, and optimize for cost and performance.

- Data Engineering at Scale: Designing and optimizing pipelines for large datasets (200M+ records), complex signals, and time series using AWS, EMR, Airflow, Spark, and dbt
- Technical Leadership: Led teams of 4-6 engineers across cloud migrations, platform builds, and analytics initiatives. Comfortable running standups, architecture reviews, and stakeholder presentations
- Pipeline Optimization: Re-architected critical data pipelines reducing processing time from 24 hours to 45 minutes (95% reduction) and from days to hours on big data workloads
- Cloud Migration: Led multi-phase migrations from on-prem (Oracle, SQL Server, Informatica) to GCP and from Snowflake to Postgres, balancing compliance, cost, and performance
- Cost Optimization: Achieved ~40% reduction in processing costs through architecture redesign and ~35% annual savings via on-demand EMR infrastructure
- Data Quality & Governance: Built custom validation frameworks with human-in-the-loop approval, dbt testing, and column-level encryption for PII in banking environments
- Real-Time Systems: Architected streaming platforms processing 50K events/second with sub-5-second latency using Kafka and Snowflake
- ML & Analytics Enablement: Served data for segmentation, churn, attrition, and demand forecasting models. Built recommendation systems and predictive models for retail and entertainment
- Platform Builder: Built data platforms from scratch when none existed, including real-time data collection systems replacing vendor software
- Cross-Functional Partnership: Partner with finance, marketing, operations, and data science teams to translate business needs into data solutions
- Long-Term Client Trust: Maintained 5+ year consulting engagement alongside full-time roles; 10+ year research partnership across two stints
- Full-Stack Capability: Deep expertise in Python, SQL, and R with foundations in algorithm implementation, ML deployment, and statistical analysis
- Cloud Native: Production systems on AWS (EMR, S3, RDS, Lambda, Bedrock) and GCP (BigQuery, Dataflow, Cloud Storage)
- Research Background: 6 peer-reviewed publications in neuroimaging; experience with HPC, signal processing, and scientific computing
- Fluent in English and Spanish

Experience

Waypoint Building Group — Remote

Staff Data Engineer (previously Senior Data Engineer via Terminal Labs) — Mar 2021 - Current

- **Multi-Tenant Data Platform Migration**: Led the architectural design and execution of a multi-tenant data platform migration from a centralized Snowflake warehouse to isolated Postgres databases per client (12+ databases, ~1TB total). Driven by client compliance mandates and cost optimization; achieved ~40% reduction in processing costs. Evaluated and selected Postgres over alternative data stores to meet performance and financial regulatory requirements; implemented infrastructure as code with Terraform and executed a zero-downtime cutover.
- **Data Pipeline Optimization**: Drove architecture and optimization of automated ETL pipelines using Airflow, dbt, Python, and AWS (S3, RDS, Lambda), reducing end-to-end data processing time from 24 hours to 45 minutes (>95%

reduction). Replaced manual CSV processing and ad-hoc runs with fully automated, scheduled workflows. Improved system reliability, data freshness, and downstream analytics performance.

- **Data Platform Ownership & Operations**: Owned the data platform end-to-end, setting operational standards for monitoring (New Relic), custom alerting, data quality enforcement, compliance adherence, and incident response. Built custom data quality checks comparing daily data against historical baselines with human-in-the-loop approval for anomalies before production promotion. Served as the staff-level technical owner and sole data engineer enabling analytics and reporting across the organization.
- **Data Warehouse & Analytics**: Designed and maintained the enterprise data warehouse as the system of record for financial and leasing analytics, supporting executive reporting and business intelligence for real estate clients. Led experimental AI/LLM agent integration using AWS Bedrock, enabling natural language access to data through tool-augmented responses.
- **Real Estate Platform Integrations**: Built and maintained data ingestion pipelines integrating with major property management platforms (Yardi, MRI, VTS, RealPage, ProForma) via APIs. Standardized disparate data formats into unified schemas for financial and leasing analytics.
- **Stakeholder Partnership & Documentation**: Partnered with finance and operations teams to translate business requirements into data solutions. Maintained comprehensive onboarding documentation and kept repository standards current. Managed AWS cost monitoring with spike alerting to control infrastructure spend.
- **Data Modeling & Testing**: Implemented snapshot-based data modeling for historical tracking. Built dbt test suite for data integrity validation (uniqueness, referential integrity) triggered on each data load. Managed per-client database credentials via AWS Secrets Manager.

Quantum Retail Remote
Data Engineering Consultant Sept 2019 - Current

- **Big Data Pipeline Modernization**: Maintain and upgrade legacy big data pipelines on AWS (S3, EMR, Hive, Spark) processing 200+ million records across 1,500+ stores and hundreds of thousands of SKUs for retail apparel clients. Modernized legacy shell script and Java Spring infrastructure by automating cluster provisioning with Python and upgrading hardcoded EMR versions. Reduced processing time from days to hours.
- **On-Demand EMR Infrastructure**: Architected on-demand EMR cluster provisioning that spins up, executes jobs, and terminates automatically based on workload. Optimized cluster configurations per job type, achieving ~35% annual cost savings.
- **Store Clustering Analytics**: Support K-means clustering pipeline that segments 1,500+ stores into 18 clusters based on demand gross margin and gross margin per style-color. Process runs 320 random starts with 200 iterations each to avoid local minima. Output drives assortment and allocation decisions.
- **Demand Forecasting & Inventory Optimization**: Support seasonal sales forecasting and deep-dive analytics for pack recommendations and size optimization using R and Spark. Deliver results via CSV data feeds and Excel reports to client merchandising teams.
- **Data Quality & Client Delivery**: Perform outlier detection and validation checks on retail datasets. Support 2 direct clients and 3 additional clients through platform maintenance. Engage seasonally (holiday, spring, summer, fall) for forecasting cycles.
- **Long-Term Client Partnership**: Maintained 5+ year consulting engagement, serving as trusted technical partner for retail analytics infrastructure. Provided continuity across seasonal planning cycles.

BairesDev (Outcome Health) Remote
Senior Data Engineer Sept 2020 - 2021

- **Real-Time Ad Analytics Platform**: Architected and built a real-time streaming platform processing 50K events/second from 300+ healthcare waiting room screens, measuring ad impressions and dwell time with sub-5-second latency. Tech stack: AWS, Kafka (Confluent), Snowflake.
- **Legacy Platform Migration**: Led architectural design and MVP delivery for migrating legacy Treasure Data batch pipelines to Kafka and Snowflake, improving data freshness and enabling real-time ad effectiveness analytics. Validated new pipeline output against legacy system to ensure data consistency.
- **Event Schema Design**: Collaborated with device software team on event logging specifications. Provided recommenda-

tions on JSON schema design and field structure to support downstream analytics requirements.

- **Data Quality & Deduplication**: Implemented deduplication logic to ensure accurate impression and dwell time metrics across high-volume streaming data.
- **Cross-Functional Delivery**: Worked within a team of 6 data engineers delivering data infrastructure supporting ad sales, client reporting, and internal analytics. Visualization layer powered by Power BI.

Scotiabank Colpatria Bogotá, Colombia
Lead Data Engineer, Retail Analytics Dec 2018 - Sept 2020

- **Cloud Migration Leadership**: Led a team of 5 contract engineers migrating the on-prem analytics platform (Oracle, SQL Server, Informatica) to Scotiabank's global GCP infrastructure. Migration executed in 5 phases covering hundreds of tables and ~90 pipelines. Drove initiative to align with Canada HQ's standardization mandate while achieving cost savings and performance improvements.
- **Target Architecture Design**: Designed target architecture using Cloud Storage, Dataflow, and BigQuery with Avro-based ingestion pipelines. Presented architecture to stakeholders across business units and regional leadership. Avro format enabled schema evolution and supported column-level encryption for PII compliance.
- **Retail Analytics Platform**: Owned end-to-end data engineering for credit card spend analytics, customer segmentation, and attrition modeling. Supervised ingestion, ETL, data modeling, BI, and ML serving layers, reducing time-to-insight for marketing teams.
- **Data Science Enablement**: Served data to data science teams for segmentation and attrition models. Ensured data availability and freshness for model training and scoring pipelines. Supported analytics workflows via Jupyter notebooks.
- **BI & Stakeholder Delivery**: Partnered with marketing team to deliver faster access to analytics. Visualization layer powered by Power BI and internal tools.
- **Team & Process Management**: Ran standups, grooming sessions, and sprint planning for contractor team. Coordinated across business units, compliance, and Canada HQ throughout multi-phase migration.
- **Security & Compliance**: Implemented column-level encryption for PII handling in banking datasets. Managed access controls and audit requirements for sensitive financial data.

Yogome, Inc. Mexico City / San Francisco, CA
Senior Data Scientist May 2017 - Aug 2018

- **Data Platform Build-Out**: Hired as Data Scientist but pivoted to building the company's first data platform from scratch — no infrastructure existed. Designed and implemented end-to-end pipeline architecture to enable analytics and ML capabilities.
- **Gaming Analytics Pipeline**: Led a team of 4 (2 data scientists, 2 data engineers) building data pipelines using Luigi for orchestration, Mixpanel for event ingestion, Vertica for warehousing, and Tableau for visualization. Processed gaming and marketing events across 100+ educational mobile games.
- **Event Data Transformation**: Transformed complex nested JSON event data from Mixpanel into structured tabular format in Vertica. Chose Vertica for fast time-to-value given the semi-structured source data.
- **ML Models for User Behavior**: Built churn prediction and user movement models to support acquisition reporting and retention analysis. Models fed into marketing and product decision-making.
- **Ad Spend & Growth Analytics**: Developed ad spend optimization analysis for Facebook and Apple campaigns. Tracked MAU and acquisition metrics to inform marketing spend allocation.
- **Stakeholder Reporting**: Built Tableau dashboards for marketing, product, and executive teams. Delivered insights on user engagement, retention, and campaign performance across 100+ games.

Analizan Mexico City
Data Scientist Mar 2015 - Nov 2016

- **Recommendation Systems**: Built collaborative filtering and content-based recommendation systems for movies and products for a major Mexican theater chain. Leveraged transaction history, viewing behavior, and demographic data to personalize recommendations.
- **Customer Transaction Modeling**: Developed predictive models for next-purchase behavior using Python, R, and SQL Server. Models delivered via batch scoring and reports to support marketing campaigns.

- **New Location Profitability Analysis**: Built predictive models for profitability of new theater locations. Features included local demographics, foot traffic, real estate costs, and competitive landscape analysis.
- **Team & Delivery**: Worked within a 5-person analytics team at a consulting firm serving retail and entertainment clients. Delivered models that contributed to revenue optimization for client merchandising and expansion decisions.

NYU School of Medicine - Center for Neuromagnetism New York, NY
Research Engineer Jul 2005 - Dec 2014; May 2016 - Nov 2017

- **Brain Signal Analysis Pipelines**: Developed analysis pipelines for MEG/EEG datasets (250 channels, 30+ subjects per study, ~450GB per study) using MATLAB and Python. Built source localization models using minimum norm estimation for neurological conditions (epilepsy, Parkinson's, autism) and basic science research including visual processing and auditory response studies.
- **High-Performance Computing**: Leveraged NYU HPC cluster for computationally intensive analysis including frequency analysis and FFT (Fast Fourier Transform) processing across large neuroimaging datasets.
- **Data Collection Platform**: Designed and built real-time data collection platform replacing limited vendor software (CTF). Integrated hardware acquisition with custom analysis workflows, monitoring CTF output and triggering downstream processing automatically.
- **Research Enablement & Training**: Trained PhDs and lab technicians on data analysis, MATLAB, and research protocols. Enabled non-technical researchers to run standardized analysis pipelines independently.
- **Publications & Scientific Output**: Co-authored 6 peer-reviewed publications. Contributed signal processing and analysis methodology to studies across multiple neurological and cognitive research domains.
- **Long-Term Research Partnership**: Maintained 10+ year engagement across two stints, returning in 2016 to complete grant-funded projects. Provided continuity and institutional knowledge across multiple research initiatives.

Skills

- **Languages**: Python, SQL, R, Unix scripting, Java
- **Data & Orchestration**: Airflow, dbt, Luigi, Kafka, Spark, Hive, Hadoop
- **Databases & Warehouses**: Postgres, Snowflake, BigQuery, Vertica, Redshift, SQL Server
- **Cloud & Infrastructure**: AWS (EMR, S3, RDS, Lambda, Bedrock), GCP (Dataflow, BigQuery, Cloud Storage), Terraform, Docker
- **Analytics & ML**: pandas, Scikit-learn, Hugging Face, MATLAB
- **Visualization & Monitoring**: Tableau, Power BI, New Relic

Education

New York University New York, NY
Master of Science, Computer Science 2012

Relevant Coursework: Operating Systems, Analysis Of Algorithms, Machine Learning (Y. LeCun), Math for Computer Science, Data Mining

National University of Colombia Bogotá, Colombia
Bachelor of Engineering, Systems Engineer 2004

Publications

- **On the Feasibility of Real-Time Prediction of Driver's Traffic Accident Risk with Auto ML using Demographics and Brain Concentration Levels Information** (2022)
  Garcia J, *Arias-Rojas W, Hernandez G* — IEEE International Conference on Vehicular Electronics and Safety (ICVES)
- **Noninvasive muscle activity imaging using magnetography** (2020)
  Llinás R, Ustinin M, Rykunov S, Walton K, Rabello G, **Garcia J**, *Boyko A, Sychev V* — Proceedings of the National Academy of Sciences

- **Differential modulation of rhythmic brain activity in healthy adults by a T-type calcium channel blocker: an MEG study** (2017)
  Walton K, Maillet E, **Garcia J**, *Cardozo T, Galatzer-Levy I, Llinás R* — Frontiers in Human Neuroscience
- **Reconstruction of human brain spontaneous activity based on frequency-pattern analysis of magnetoencephalography data** (2015)
  Llinás R, Ustinin M, Rykunov S, Boyko A, Sychev V, Walton K, Rabello G, **Garcia J** — Frontiers in Neuroscience
- **Theta and gamma oscillations typical of thalamocortical dysrhythmia in patients with trigeminal neuralgia pain** (2012)
  Walton K, **Garcia J**, *Rabello G, Delfino J, Llinás R* — Research Publication
- **Magnetic sources of the M50 response are localized to frontal cortex** (2008)
  Garcia-Rill E, Moran K, **Garcia J**, *Findley WM, Walton K, Strotman B, Llinás R* — Clinical Neurophysiology
- **Abnormal brain activity in patients with complex regional pain syndrome (CRPS) type I** (2008)
  Dubois M, Rojas-Soto D, **Garcia J**, *Levacic D, Walton K, Llinás R* — Research Publication
- **Minería de datos difusa: descubrimiento de reglas asociativas difusas** (2004)
  Garcia J, *Ortiz J* — Revista Vínculos