Patrick Siu

(917) 705-9998 | psiu003@gmail.com | Millburn, NJ

Linkedin | Portfolio | Github

Executive Summary

Lead Data Engineer architecting petabyte-scale AI platforms (4PB+, 30k+ cores) that operationalize GenAI, RAG, and Agentic workflows. Currently leading a 12-person team to deliver 270x performance optimizations and detection pipelines catching 14M+ daily threats. Expert in Python, Spark, and Kubernetes, driving secure, scalable infrastructure for high-impact security operations.

Technical Skills

Languages & Compute: Python (Expert), PySpark, SQL, R, Java, C++, Bash/Shell.
AI & Machine Learning: GenAI Ecosystems (Agentic Workflows, RAG, MCP), NLP (Sentiment Analysis, Topic Modeling), Computer Vision, Graph Algorithms, Neo4j.
Big Data & Infrastructure: AWS, GCP, Docker (Containerization), Kubernetes, Hive, HDFS, Hadoop Ecosystem.
Data Warehousing & ETL: Redshift, PostgreSQL, Airflow, Automated Data Pipelines, CI/CD.
Visualization & Analytics: Tableau (Advanced/Certified), R Shiny, Adobe Analytics, Business Intelligence Dashboards.

Experience

Lead Data Engineer (Financial Crimes) | **TikTok U.S. Data Security** | **5/2024 – Present**

- **Team Leadership & Strategy:** Built and led a 12-person engineering team from the ground up, establishing core SDLC standards, code review protocols, and modular design patterns that boosted output and auditability.
- **Petabyte-Scale Architecture:** Architected a multi-region data platform on ByteCloud (abstracted layer over Oracle Cloud and AWS) managing 4PB+ of data. utilized Dorado (Databricks / Spark) and HDFS to optimize storage costs and query performance.
- **GenAI & LLM Implementation:** Delivered production-grade LLM ecosystems by integrating Agentic AI workflows, Model Context Protocol (MCP), and RAG pipelines. Built using LangChain, VikingDB (in-house Milvus equivalent), and Gemini 3 to ensure high-fidelity, context-aware interoperability.
- **Orchestration & Automation:** Automating critical investigative operations and reducing manual effort by implementing Dorado modular multi-stage pipelines for complex DAG dependency management and leveraging TCE (in-house Kubernetes) for scalable execution.
- **Data Governance & Quality:** Drove a data quality transformation by instituting a custom data governance ruleset, a centralized data library, and proactive monitoring to ensure department-wide trustworthiness and secure access control.
- **Analytics & Observability:** Delivered 20+ dashboards for risk and operations metrics using Aeolus (in-house Tableau), supported by a custom contextual alerting system with direct integration with Lark (in-house Slack messaging)

Specialist (Data Engineer: Data Scanning) | **TikTok U.S. Data Security** | **5/2023 – 5/2024**

- **Containerized Scale Architecture:** Designed and implemented a massive-scale compliance scanning framework using a custom distributed orchestration framework leveraging the Dorado API to manage parallel processing across 30,000 compute cores and 60 TB memory.
- **Performance Optimization:** Achieved a **270x increase in scanning speed** by refactoring legacy scripts and automation structure for HDFS, MySQL, MongoDB, Redis, and Elastic Search. Optimized resource contention using throttling and load detection to prevent overwhelming sensitive data sources such as production databases.
- **Big Data at Scale:** Extended privacy scanning coverage to **6 exabytes** of data by engineering a custom bridge through Linux to integrate Nuix scanning capabilities directly with **HDFS** storage, before we created our own in-house capability
- **Regulatory Impact:** Generated compliance reports from our scanning system, which were publicly presented as primary evidence during Congressional hearings on data security.

Specialist (Data Engineer: Bot Detection) **Bytedance (TikTok)** **10/2021 – 5/2023**

- **Bot Detection Architecture:** Led a team of 9 to engineer a daily automated pipeline processing 1 billion user accounts to detect **14M inauthentic accounts** daily.
- **Defense Efficacy:** Achieved a **94% incremental detection rate** (catching actors missed by standard defenses) by engineering features based on behavioral velocity such as coordinated spike engagement, activity interarrival times, no sleep patterns, etc.
- **App Intel Monitoring:** Built a scraping and monitoring system using Python+Beautiful Soup and DAG Hive orchestration to inventory the global Bytedance app footprint. Automated real-time alerts via **Lark Webhooks** to flag privacy vulnerabilities and popularity shifts.
- **Forensic Investigation:** Served as the dedicated technical investigator for 3 P&C litigation cases, executing complex forensic data retrieval queries across multiple departments to prepare corporate depositions under tight legal deadlines.
- **Thought Leadership:** Selected by professor Dr. Holly Russo and Dr. Randy Garrett to lecture on Data Visualization architectures at **George Mason University**, representing Bytedance's technical expertise to graduate-level Data Science cohorts.

Principal Data Scientist (DataCloud) **ADP** **11/2020 – 10/2021**

- **AWS Pipeline Automation:** Engineered automated data delivery pipelines utilizing AWS Step Functions for the Data Licensing business, replacing manual workflows to improve reliability and speed.
- **Monetization Analytics:** Architected a ”live” revenue and KPI dashboard suite for API monetization using Tableau, enabling real-time decision-making on high-volume transaction data.

Lead Data Scientist (Infrastructure & Strategy) **ADP** **11/2018 – 11/2020**

- **Greenfield Infrastructure:** Built the BI and analytics infrastructure from the ground up for the ”Wisely Direct” startup unit, utilizing **Docker** containers for portability and Airflow for orchestration.
- **Data Silo Unification:** Consolidated 13 disparate data silos into a unified **Redshift** and **Postgres** data warehouse, enabling the first comprehensive view of the business.
- **Strategic ETL:** Developed end-to-end **Python** ETL pipelines to support a $700M digital transformation portfolio, strategically guiding funding allocation through data-driven retention modeling.

Lead Data Scientist (NLP & Client Experience) **ADP** **5/2014 – 11/2018**

- **NLP & Text Mining:** Engineered an enterprise-grade NLP engine using R openNLP to classify **9M+ unstructured service logs**, achieving **88% precision** and expanding topic coverage from 26 to 442 categories.
- **Speech Analytics:** Delivered a Proof-of-Concept (POC) speech-to-text pipeline using Nice CX combined with sentiment analysis to identify key call volume drivers.
- **Enterprise Visualization:** Stood up and automated the first enterprise-wide **Tableau** dashboard ecosystem to standardize performance measurement across all business units.

Prior Experience

- **Product Manager** | Thomson Reuters (2012 – 2014)
- **Product Manager** | Dun & Bradstreet, Inc. (2012 – 2012)
- **Product Manager** | Canon U.S.A., Inc. (2007 – 2012)
- **Product Manager** | New Era Pump Systems, Inc. (2005 – 2007)
- **Product Manager** | PCW Microsystems Inc. (2003 – 2005)

Education

Stony Brook University B.S. in Computer Science **Class of 2003**

Icebreaker Topics and Personal Interests

Deep learning image recognition · Long-distance motorcycle touring & engine repair · Scuba diving & snowboarding · Volunteer coach on Devils Youth Mites Hockey team (USA Hockey Certified CEP level 2)