

HARSHAD HATTE

+1 862-405-1744 | harshadhatterusa23@gmail.com | Harrison, NJ | [linkedin.com/in/harshad-hatte/](https://www.linkedin.com/in/harshad-hatte/) | github.com/Harshad30

PROFESSIONAL SUMMARY

Data Analyst with an M.S. in Data Science (NJIT) and hands-on experience designing automated ETL/ELT pipelines, large-scale distributed data processing, and cloud data workflows using Python, PySpark, and SQL. Experienced in building high-performance data pipelines processing 2B+ telemetry records at enterprise scale and delivering analytical data stores supporting business intelligence and machine learning workloads on cloud platforms including AWS, GCP, and Azure.

EXPERIENCE

JerseyStem — Data Analyst

Aug 2025 – Present

- Developed **MYSQL** based data pipelines replacing fragmented spreadsheets across 3 departments with a single verified data source, reducing manual cross-checking and improving reporting reliability for downstream teams.
- Streamlined **Salesforce** and **GAT+** activity data into centralized reports, enabling program contributors to independently prioritize outreach tasks in real time and eliminating recurring manual data requests from operations and marketing teams.
- Partnered with Marketing, Operations, and leadership to design and deploy 4 **Looker Studio dashboards** covering pipeline activity, team performance, and school outreach metrics, giving leadership real-time visibility into **program KPIs**.

Verizon — Data Analyst

Feb 2025 – May 2025

- Engineered high-performance **ETL/ELT** pipelines in **Python** and **PySpark** processing **2B+ telemetry records**, optimized ingestion, transformation, and compression workflows while maintaining data integrity across large-scale network datasets.
- Designed automated **data ingestion** pipelines for multi-source network data, with **validation frameworks** and quality checks to catch corruption before nightly processing; pipeline fetched, cleaned, and stored data in Parquet format on university **HPC** for next-day analysis.
- Conducted large-scale **EDA** across high-volume datasets to identify usage trends and performance anomalies; findings directly informed **infrastructure planning** decisions presented to cross-functional engineering teams.

PROJECTS

Minecraft Multi-Agent LLM Research System | *Python · Node.js · FastAPI · Redis · HuggingFace · Weights & Biases · Azure ADLS Gen2 · Databricks*

- Designed **PIANO-inspired** dual-loop agent architecture comparing Claude Haiku, GPT-4o-mini, and Mistral across 6+ hours of autonomous gameplay; discovered **emergent role specialization**, models developed distinct explorer vs homekeeper strategies without explicit assignment across **multi-agent** sessions.
- Implemented **RAG-based** episodic memory (HuggingFace Sentence Transformers), **Redis pub/sub** inter-agent coordination, **goal-directed planning** module, and async FastAPI ML inference server; agent performance improved 78% session-over-session (9→21 unique items) demonstrating effective cross-session learning.
- Architected automated telemetry and event-processing pipelines capturing 5,000+ action-level records, implementing Bronze/Silver/Gold data architecture in Azure Data Lake Storage (ADLS Gen2) for scalable ingestion, validation, monitoring, and analytics workflows.
- Built **PySpark** notebooks in **Azure Databricks** to perform cross-model behavioral analysis across GPT-4o-mini, Claude Haiku, and Mistral, comparing survival rates, exploration entropy, and resource efficiency to surface model-specific decision patterns.

ReadmitIQ — Hospital Readmission Analytics Platform | *Python · XGBoost · DuckDB · ChromaDB · LangChain · Streamlit · Anthropic Claude*

- Developed end-to-end **ML pipeline** on CMS synthetic Medicare data (20,867 claims) predicting 30-day readmission risk; **XGBoost model** achieved 0.908 AUC and 86% recall, a 13.6-point improvement over logistic regression baseline with 0.907 cross-validation stability.
- Engineered 15+ clinical features including LOS, admission type, diagnosis category, and demographic risk factors to identify high-risk patterns; flagged 85+ age group (57.9% readmission rate) and mental health diagnoses (57.6%) as highest-risk segments for early intervention.
- Built **RAG-based** natural language chat interface using **ChromaDB** vector store and Claude API, enabling plain-English querying of analytics without SQL, eliminating LLM hallucination via pre-computed statistical embeddings.

VCT Pacific 2026 Data Pipeline & Competitive Analytics | *Python · MySQL · SQL · Looker Studio · Data Modeling*

- Deployed an end-to-end automated ETL pipeline** scraping 30+ VCT Pacific matches, loading data into a normalized 4-table MySQL schema with deduplication, error handling, and incremental refresh.
- Engineered SQL views computing 5+ advanced KPIs** (impact score, K/D ratio, kill contribution, rolling averages) across 14+ performance metrics including ACS, KAST, ADR, and first-blood efficiency.
- Developed a **Looker Studio** coaching dashboard analyzing ATK/DEF win rates, player heatmaps, map pool depth, and opponent agent compositions for pre-match scouting; validated by a professional VCT analyst.

EDUCATION

New Jersey Institute of Technology Newark, New Jersey, USA

Sep 2023 – May 2025

Master of Science in Data Science

TECHNICAL SKILLS

Programming: Python (Pandas, NumPy, Matplotlib, Plotly, FastAPI), R, SQL, Excel, Git

Databases: MySQL, Oracle SQL, MongoDB, Neo4j, Redis, BigQuery, AWS S3

Visualization & BI: Tableau, Power BI, Looker Studio, Weights & Biases

ML & AI: Scikit-learn, TensorFlow, PyTorch, Keras, HuggingFace Transformers, LLM APIs (Claude, GPT), RAG, Multi-Agent Systems

Cloud Platforms: Google Cloud Platform (GCP), Amazon Web Services, BigQuery, Azure (ADLS Gen2, Blob Storage, Databricks)

Data Engineering & Orchestration : Apache PySpark, Spark SQL, Azure Data Factory (ADF), DuckDB, LangChain, n8n