Sabbir Hossain

hossain.sabbir17@gmail.com | (647) 545-8842 | linkedin.com/in/itssabbir | github.com/itsSabbir | sabbir.ca
Atlanta Metropolitan Area, GA | Open to Relocation (US/Canada) | TN Visa Authorized

SUMMARY

Data Engineer with production ownership of mission-critical ETL/ELT pipelines serving enterprise analytics across 4+ operational systems. Proven track record of 83% query performance optimization on 23M+ row datasets, 130K+ record recovery programs, and building observability infrastructure across 12+ pipelines. Combines deep data platform engineering with full-stack development and machine learning experience to deliver scalable, reliable data solutions.

SKILLS

Data Engineering: Apache Spark, PySpark, Apache Airflow, Apache Kafka, Teradata, BigQuery, Snowflake, Amazon Redshift, PostgreSQL, MongoDB, MySQL, ETL/ELT Pipelines, Data Modeling (Kimball/Snowflake Schema), Data Warehousing, Data Quality, Data Lineage, Batch and Stream Processing, SCD Types, MicroStrategy, dbt Concepts, Data Architecture

Programming: Python, SQL, R, C, Bash, Java, JavaScript, Node.js, React, Flask, REST/GraphQL APIs

Cloud and DevOps: GCP (BigQuery, Cloud Composer, Looker), AWS (S3, EC2, RDS, Lambda), Docker, Kubernetes, Linux, Git, CI/CD, GitHub Actions, Terraform Concepts, Distributed Systems

ML and Analytics: PyTorch, TensorFlow, Scikit-learn, Data Structures and Algorithms, Query Optimization, Anomaly Detection

Practices: Agile/Scrum, JIRA, Confluence, Technical Documentation, Data Governance, Data Security, System Integration, Microservices, Test-Driven Development (TDD)

TECHNICAL HIGHLIGHTS

- Delivered 78-attribute MicroStrategy analytics cube integrating 4+ enterprise systems; secured director sign-off and production deployment within 6 weeks
- Reduced query latency 83% (12 min to 2 min) on 23M+ row joins; prevented 15+ monthly data quality incidents via automated observability and monitoring
- Recovered 130K+ records and fixed 22K+ misattributions via SCD Type 2 temporal joins, expanding historical analytical coverage from 1 to 9+ months
- Built Python/Airflow observability framework across 12+ pipelines, reducing debugging time 60% via automated validation and data lineage tracking
- Promoted to Technical Gatekeeper (3 months) and Production Backup Owner (6 months); Harvard NCRC 2024 Plenary Speaker (1 of 12 from 5,000+ applicants)

EXPERIENCE

Bell Canada | Toronto, ON – Remote　Jun 2025 – Present
Data Engineer

- Engineered and own the mission-critical NTS pipeline, a resilient 3-tier ETL/ELT architecture (Staging, Warehouse, Analysis) on Teradata integrating 4+ operational systems (SmartPath REST API, Maximo ERP, IPACT billing, LDAP directory) using Python and SAS Data Integration. Enforce data contracts, Kimball dimensional modeling, and schema-versioned deployments across DEV/QA/PROD environments.
- Designed and delivered a 78-attribute enterprise analytics platform (MicroStrategy Drill Cube) with derived metrics, heatmap conditional formatting, and cross-filter interactivity. Led architectural decisions including rejecting a proposed physical Fact Table in favor of a SQL View, and expanded scope from 13 to 78 attributes to eliminate recurring ad-hoc request cycles. Secured director-level sign-off within 6 weeks.
- Diagnosed systemic data integrity drift via full-stack Root Cause Analysis on a misconfigured temporal filter. Executed staged historical recovery (28K+, 50K+, and edge-case records totaling 130K+), expanding analytical coverage 800% (1 to 9+ months). Fixed 22K+ employee misattributions via SCD Type 2 temporal joins on LDAP hierarchy data, replacing volatile Login IDs with stable natural keys.
- Eliminated a critical performance anti-pattern by replacing a direct join on a 23M+ row live table with a materialized pre-aggregation layer, cutting query runtime 83% (12 min to 2 min) and stabilizing nightly SLA compliance. Engineered a custom stateful sessionization algorithm using two-pass group-by propagation to resolve event sequencing defects with zero calculation defects in QA.
- Built Python/Airflow observability framework with configuration-driven validation checks across 12+ pipelines. Implemented automated schema validation, anomaly detection, and threshold alerting for data quality SLAs, reducing debugging time 60% and preventing 15+ monthly incidents. Managed as production backup owner during company-wide code embargo for 12+ dependent pipelines.

- Promoted to Technical Gatekeeper within 3 months, governing the NTS domain with defensive coding standards, peer code reviews, and architectural audits. Authored validation documentation (ERDs, Data Flow Diagrams, count-by-stage proofs) that became the team standard. Leading knowledge transfer across NTS, SmartPath, and CS Attack domains; contributing to Q1 2026 GCP/BigQuery migration from legacy Teradata.

Johns Hopkins University | Baltimore, MD – Remote Sep 2022 – Present

Bioinformatics Software Development Research Assistant

- Architected an open-source full-stack bioinformatics platform (Python, R, JavaScript, C) with microservices architecture, SOLID principles, and Docker containerization. Reduced analysis load times 83% via optimized caching, supporting 100+ global researchers.
- Engineered scalable ETL pipelines processing 750+ TB multi-omics data on HPC clusters using Python, R, SQL, and ML models (SVM-RFE, Random Forest). Accelerated biomarker discovery 40% and reduced analysis time 40%. Implemented automated data quality and anomaly detection (K-Means, DBSCAN via TensorFlow) within CI/CD pipelines.
- Built interactive visualization dashboards (Shiny, React, D3.js), developed REST/GraphQL APIs for real-time data access, and configured AWS (EC2/S3) environments with GitHub Actions CI/CD. Applied data governance practices for institutional compliance and audit trails.
- Authored 35-page research manuscripts with reproducible analyses. Presented award-winning research at ABRCMS and Harvard NCRC conferences; selected as 1 of 12 plenary speakers from 5,000+ applicants.

University of Toronto | Toronto, ON – Hybrid Sep 2019 – Apr 2024

Software Development Research Assistant

- Reduced analysis effort by 30+ hours/week across 7 research teams by engineering full-stack platforms (Python, R, C, Java) with OOP patterns. Owned full SDLC from requirements to production deployment.
- Cut environment setup time 50% via Docker-based DevOps workflows. Improved UI render times 45% for large genomic datasets in Next.js/Tailwind CSS. Led Agile Scrum adoption and mentored 5 junior developers.

EDUCATION

B.Sc. (Hons) Computer Science, Bioinformatics and Computational Biology | University of Toronto June 2024

GPA: 3.96/4.0 | Coursework: Data Structures and Algorithms, Database Systems, Distributed Systems, Machine Learning, Operating Systems, Software Design, Cloud Computing, Computer Networks

PROJECTS

- **Image Processing Pipeline Server:** High-performance multi-threaded C server using POSIX threads and sockets, handling 100+ concurrent clients with <100ms latency. TDD and CI/CD with 30% faster processing vs. baseline.
- **Automated Anomaly Detection System:** Full-stack platform (Node.js/Express, React/TypeScript) orchestrating Python ML (YOLOv5, LSTM) for video processing. AWS S3/RDS storage with containerized Docker deployment.
- **Stock Market Prediction Pipeline:** Real-time prediction system integrating Kafka streaming, feature engineering, and ML models (RandomForest, XGBoost) achieving +/-5% prediction error. Dockerized Flask deployment.

AWARDS

- **Harvard NCRC 2024 Plenary Speaker** – Selected as 1 of 12 from 5,000+ applicants
- **Best Oral Presentation, ABRCMS 2023** – Top presenter from 80 oral presenters out of 6,500+ attendees
- **Best Poster Presentation, ABRCMS 2024** – Competed among 150+ graduate-level presenters
- **Friends of Arts and Science Awards, University of Toronto** – Computer Sciences and Physical/Life Sciences (2022, 2023, 2024)