Sabbir Hossain

hossain.sabbir17@gmail.com | (647) 545-8842 | linkedin.com/in/itssabbir | github.com/itsSabbir | sabbir.ca
Toronto, ON | Open to relocation (US/Can) | TN Visa Auth.

Skills

Programming Languages: Python, SQL, R, C, JavaScript, TypeScript, Bash, Flask, REST/GraphQL APIs
Data Engineering: Apache Spark, PySpark, Apache Airflow, Apache Kafka, Teradata, PostgreSQL, MongoDB, MySQL, ETL/ELT Pipelines, Data Modeling (Kimball), Data Warehousing, Data Quality, Batch and Stream Processing, SCD Types
Cloud and DevOps: AWS (S3, EC2, RDS), Docker, Linux, Git, CI/CD, GitHub Actions, Distributed Systems
ML and Analytics: PyTorch, TensorFlow, Scikit-learn, Data Structures and Algorithms, Query Optimization
Practices: Agile/Scrum, JIRA, Confluence, Technical Documentation, Microservices, TDD

Technical Highlights

- Shipped a **78-attribute MicroStrategy** analytics cube integrating 4+ 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.
- Recovered **130K+ records** and fixed 22K+ misattributions via **SCD Type 2** temporal joins, expanding historical coverage from 1 to 9+ months.
- Engineered ETL pipelines processing **750+ TB** of multi-omics data on HPC clusters at **Johns Hopkins University**, accelerating biomarker discovery by **40%**.
- Promoted to **Technical Gatekeeper** (3 months) and designated **Production Backup Owner** (6 months); **Harvard NCRC 2024 Plenary Speaker** (1 of 12 from 5,000+).

Work Experience

Bell Canada

Data Platform Engineer | Toronto, ON - Remote, Full Time | Dec 2025 - Present

- **Enterprise Analytics Platform Delivery:** Shipped a 78-attribute MicroStrategy analytics cube integrating operations API (SmartPath), asset management (Maximo), billing system (IPACT), and directory services (LDAP). Built derived metrics (Median Hours, Distinct Request Count), heatmap thresholds, and cross-filter interactivity, then deployed Dev to Pre-Prod to Prod with director sign-off.
- **Architectural Decision-Making and Scope Protection:** Reduced build cost and risk by using a SQL view (v_fact_smtpth_nts_timesheet) instead of a physical fact table for event-to-request joins and duration metrics. Blocked out-of-scope Maximo integration requests and expanded from 13 columns to a future-proof 78-attribute design to stop recurring ad hoc cycles.
- **Production System Ownership and Operational Resilience:** Provided escalation coverage as backup owner for the SmartPath production API during a company-wide code embargo supporting 12+ dependent pipelines. Resolved critical ETL failures in residual error staging tables (Jarvis residual error tables) and corrected MicroStrategy metric misconfiguration under time pressure.
- **Cross-Domain Technical Investigation:** Ramped from zero domain knowledge to full system comprehension in 2 weeks for the CS Attack pipeline spanning Salesforce, event operations stage (EOM), billing system (IPACT), downstream processing stage (TTI), and SmartPath. Built a repeatable field discovery and join-proof method using the Teradata system catalog (DBC.Columns), validating source population and join feasibility with evidence.
- **Engineering Workflow and Governance:** Co-authored the engineering workflow standard separating investigation (feasibility, validation) from implementation (code, test, deploy). Established Jira tracking, in-scope vs out-of-scope boundaries, and a mandatory front-door ticket intake for budget accountability.
- **Leadership Transition and Strategic Documentation:** Owned continuity through a leadership transition by refactoring NTS Confluence pages into Executive Summary plus Technical Appendix. Became the canonical knowledge source across NTS, SmartPath, and CS Attack while standardizing visual-first director reporting.
- **Cloud Migration and Modernization Strategy:** Supported Q1 2026 migration to Google Cloud Platform and BigQuery by pivoting Teradata wide tables to a snowflake schema for better fault isolation. Prepared a roadmap to replace SAS Data Integration Studio with Cloud Composer (Apache Airflow) DAG orchestration.

Bell Canada

Data Engineer | Toronto, ON - Remote, Full Time | Jun 2025 - Nov 2025

- **Enterprise Data Platform Architecture:** Built and productionized the NTS pipeline on Teradata using a three-tier ETL and ELT architecture: staging, warehouse, and analysis layers. Integrated REST API event streams (SmartPath), legacy ERP (Maximo), billing system (IPACT), and directory services (LDAP) via Python and SAS Data Integration, enforcing data contracts and Kimball dimensional modeling patterns.
- **Algorithmic Engine Design and Optimization:** Built a stateful sessionization algorithm in Python to fix event sequencing defects, refactoring a flawed sequential method into a robust two-pass group-by propagation model. Achieved deterministic mapping across distributed agent sessions by identifying anchor events and backfilling request identifiers to preceding and succeeding events.
- **Compute Resource and SLA Optimization:** Reduced query latency by 83 percent (12 minutes to 2 minutes) on a join over 23 million rows by replacing dynamic runtime computation with a materialized pre-aggregation layer. Eliminated production timeouts and stabilized nightly SLA compliance through query optimization and static reference architecture.
- **Data Reliability and Root Cause Analysis:** Expanded analytical coverage by 800 percent (1 month to 9+ months) by running a full root cause analysis (RCA) on a hardcoded 30-day lookback filter causing systemic data drift. Executed a historical recovery program recasting 28,000+ and 50,000+ records, raising ticket match accuracy to the highest level since system inception.
- **Data Integrity and Temporal Modeling:** Fixed historical attribution defects by implementing SCD Type 2 temporal joins on creation date to resolve employee hierarchy changes against directory services data. Replaced volatile login identifiers with a stable natural key (agent email) to preserve data lineage integrity for historical reporting.

- **Pipeline Resilience and Granularity:** Prevented duplicate data during retries by enforcing idempotency and atomic writes via composite upsert keys (request, ticket, and configuration item identifiers). Applied COALESCE, UPPER, and TRIM sanitization for all join conditions and enforced pre-aggregation patterns, preserving model granularity across all environments.
- **Modular Architecture and Migration:** Refactored monolithic SQL into modular clean and calculate transformation stages, a pattern analogous to dbt staging and marts. Created a unified view abstraction layer merging legacy and modern structures, enabling zero-downtime migration for downstream BI consumers.
- **Observability and Monitoring Infrastructure:** Built a Python and Apache Airflow observability module with configuration-driven validation checks orchestrated via DAG modules across 12+ pipelines. Reduced debugging time by 60 percent and prevented 15+ monthly data quality incidents through automated schema validation, anomaly detection, and threshold alerting.
- **Technical Governance and Documentation:** Promoted to Technical Gatekeeper within 3 months to govern the domain by enforcing defensive coding standards. Authored end-to-end validation documents including ERDs, data flow diagrams, and count-by-stage proofs, establishing the team standard for peer review and knowledge transfer.
- **Requirements and Stakeholder Management:** Led requirements gathering and technical feasibility assessments for new data pipeline initiatives. Translated business needs into technical architectures and implementation roadmaps by facilitating cross-functional alignment sessions with stakeholders.
- **Data Analysis and Executive Communication:** Built visualizations and presented pipeline performance and data quality metrics to directors, team leads, and BI analysts. Drove buy-in for platform modernization initiatives through data-driven executive reporting.

Johns Hopkins University

Bioinformatics Software Development Research Assistant | Baltimore, MD - Remote, Part Time | Sept 2022 - Present

- **Open-Source Platform Architecture:** Reduced analysis load times by 83 percent through optimized caching on a full-stack bioinformatics platform supporting 100+ global researchers. Built using Python, R, JavaScript, and C with microservices architecture, SOLID principles, and Docker containerization.
- **Scalable ETL and Big Data Processing:** Engineered scalable ETL pipelines processing over 750 terabytes of multi-omics data on HPC clusters. Accelerated biomarker discovery by 40 percent using Python, R, SQL, and machine learning models including SVM-RFE and Random Forest.
- **Automated Data Quality and ML:** Improved data integrity by 30 percent by implementing automated data quality checks and anomaly detection using unsupervised ML (K-Means, DBSCAN) with TensorFlow within CI/CD pipelines. Validated biomarker analysis software using TensorFlow, Keras, and Scikit-learn.
- **Interactive Visualization:** Built interactive data visualization dashboards for molecular modeling and educational use using Shiny, React, and D3.js. Improved usability and accessibility for researchers working with complex genomic datasets.
- **API Development and Integration:** Developed and optimized REST and GraphQL APIs to support real-time data access and model simulations across research modules.
- **Cloud Infrastructure and DevOps:** Configured AWS environments including EC2 and S3 and automated testing and deployment workflows with GitHub Actions. Improved reliability and collaboration through CI/CD automation.
- **Data Governance and Compliance:** Applied secure data management and governance practices to ensure compliance with institutional privacy and research ethics standards. Maintained data lineage documentation for audit trails.
- **Cross-Functional Collaboration:** Collaborated with cross-functional experts including oncologists and statisticians to align computational workflows with research goals. Mentored peers on HPC and reproducible software practices.
- **Research and Technical Communication:** Authored multiple 35-page research manuscripts featuring interactive dashboards and reproducible analyses. Presented award-winning research at ABRCMS and Harvard NCRC, selected as 1 of 12 plenary speakers from 5,000+ applicants.

University of Toronto

Software Development Research Assistant | Toronto, ON - Hybrid, Part Time | Sept 2019 - Apr 2024

- **Full-Stack Platform Engineering:** Reduced analysis effort by 30+ hours per week across 7 research teams by engineering full-stack bioinformatics platforms. Built automation using Python, R, C, and Java with object-oriented patterns.
- **SDLC and Requirements Translation:** Owned the full software development life cycle (SDLC) including requirements, architecture, implementation, testing, deployment, and maintenance. Translated multidisciplinary research requirements into production-grade software.
- **DevOps and Containerization:** Cut setup and configuration time by 50 percent by implementing Docker-based DevOps workflows to eliminate environment drift. Enabled reproducible and scalable computation.
- **Performance Optimization and UX:** Improved UI render times by 45 percent for large genomic datasets by optimizing data visualization in Next.js and Tailwind CSS.
- **Engineering Leadership and Agile:** Led Agile Scrum adoption and mentored a team of 5 junior developers. Increased throughput and strengthened cross-team collaboration.

Education

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

- GPA: 3.96 / 4.0 | Relevant Coursework: Data Structures and Algorithms, Software Design, Systems Programming, Algorithm Design and Analysis, Theory of Computation, Operating Systems, Database Systems, Machine Learning, Distributed Systems, Cloud Computing, Computer Networks, Applied Bioinformatics, Statistics and Probability

Projects

Image Processing Pipeline Server

- Architected a high-performance multi-threaded C server for real-time image processing using POSIX threads and sockets, handling 100+ concurrent clients with under 100ms latency.
- Implemented TDD (Python integration tests, shell scripts) and CI/CD, demonstrating 30% faster processing vs. baseline.
  Key Tech: C (pthreads), Python, Linux/Unix, Sockets, Multithreading, CI/CD, TDD

Automated Anomaly Detection System

- Engineered a full-stack anomaly detection platform using Node.js/Express backend to orchestrate Python (YOLOv5, LSTM) processing of video uploads, storing frames on AWS S3 and alerts in AWS RDS (PostgreSQL).
- Developed React/TypeScript/MUI frontend with dynamic alert filtering; designed for Docker deployment to AWS EC2 with CI/CD.
  Key Tech: Node.js, Express, Python (PyTorch, YOLOv5), PostgreSQL, React, TypeScript, AWS (EC2, S3, RDS), Docker

MicrobiomeExplorer R Package (Open Source)

- Created a modular R package for 16S rRNA/metagenomic analysis, integrating ETL pipelines, Bioconductor stats, and interactive Shiny visualizations.
  Key Tech: R, Shiny, Python, Bash, Bioconductor, ETL, Data Visualization

Bioinformatics Pipeline for Gene Expression Analysis

- Built an end-to-end containerized (Docker) bioinformatics pipeline using Nextflow for reproducible RNA-seq analysis (DESeq2, GSEA), reducing manual effort 40%.
  Key Tech: Nextflow, R (Bioconductor), Python, Docker, Bash, Workflow Automation

Red Blood Cell Counter

- Engineered a C application using image processing algorithms (segmentation, flood-fill) for automated RBC counting, reducing false positives 30%.
  Key Tech: C, Image Processing, Algorithm Design, Data Structures, Memory Management

Awards and Achievements

- **Plenary Speaker**, National Collegiate Research Conference (NCRC) Harvard 2024 - Selected as 1 of 12 from 5,000+ applicants
- **Best Detailed Oral Presentation**, ABRCMS Conference 2023 - Top presenter; selected from 80 oral presenters out of 6,500+ attendees
- **Best Poster Presentation**, ABRCMS Conference 2024 - Competed among 150+ graduate-level presenters
- **Poster Presentation**, National Collegiate Research Conference (NCRC) Harvard 2024
- **Friends Of Arts And Science Award** in Computer Sciences, University of Toronto (2022, 2023, 2024)
- **Friends Of Arts And Science Award** in Physical And Life Sciences, University of Toronto (2022, 2023, 2024)