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.

Summary

Software and Data Engineer specializing in building and scaling resilient, data-intensive applications and pipelines. Proven track record of architecting full-stack distributed systems, remediating critical data integrity issues in high-volume production environments, and implementing robust DevOps/SRE practices. Expertise in Python, SQL, Java, and cloud-native tools (AWS, Docker) to deliver performant and reliable solutions.

Skills

Programming Languages: Python, R, SQL, C, JavaScript, TypeScript, Bash/Shell, HTML, CSS
Data Engineering & Databases: Teradata, PostgreSQL, MySQL, MongoDB, Apache Kafka, Apache Airflow, Data Modeling, SQL Optimization, Extract Transform Load (ETL), Data Pipelines, CRUD Operations, Batch Processing, Stream Processing, Data Quality, Data Validation, Data Warehousing, PySpark, Apache Spark
Cloud & Tools: AWS (EC2, S3, RDS), Docker, Linux, Git, JSON, GitHub Actions, HPC
Frameworks & Libraries: Flask, Django, Node.js, .NET, React, MERN Stack, Shiny, PyTorch, TensorFlow, Keras, Scikit-learn, D3.js, REST APIs
Software Engineering Practices: Algorithms, Data Structures, OOP (SOLID), Microservices, SDLC, CI/CD, TDD, Automated Testing, Agile, Cloud Computing, DevOps Concepts, Distributed Systems
Professional Skills: Technical Documentation, Technical Writing, Data Visualization, Cross-functional Collaboration, Problem-Solving, Scientific Computing, Leadership, JIRA, Confluence, LaTeX, Microsoft Office Suite (Excel, Powerpoint, Word), Stakeholder Management, Code Review, Requirements Gathering, System Design, Performance Optimization, Troubleshooting, Quality Assurance, Version Control, Mentoring, Presentation Skills, Root Cause Analysis, Process Improvement

Experience

Data Engineer — Jun 2025 – Present

Bell Canada, Toronto, ON – Remote, Full Time

- **Enterprise Data Platform Architecture:** Engineered and productionized the mission-critical **Network Ticket Service (NTS)** pipeline, a resilient 3-tier **ETL/ELT** architecture (Staging, Warehouse, Analysis) on **Teradata**. Integrated 4+ disparate operational systems including REST API **SmartPath** event streams, legacy ERP **Maximo**, billing system **IPACT**, and directory services **LDAP** using **Python** and **SAS Data Integration**. Enforced multi-stage workflows, schema-versioned loads, and enforceable Data Contracts across DEV, QA, and PROD using **Kimball**-style dimensional modeling patterns transferable to **Snowflake**, **BigQuery**, and **Redshift**.
- **Algorithmic Engine Design & Optimization:** Engineered a custom stateful sessionization algorithm in **Python** to resolve event sequencing defects, refactoring a flawed sequential method into a robust two-pass group-by propagation model. Developed complex state management logic to identify anchor events (T2) and backfill Request IDs to preceding (T1) and succeeding (T3) events, achieving deterministic mapping across distributed agent sessions.
- **Compute Resource & SLA Optimization:** Eliminated a critical performance anti-pattern by redesigning a high-cardinality join on a 23 million+ row live table. Replaced dynamic runtime computation with a materialized pre-aggregation layer and static reference architecture, reducing query latency by **83%** (12 mins $\rightarrow$ 2 mins) and eliminating production timeouts to stabilize nightly **SLA** compliance.
- **Data Reliability & Root Cause Analysis:** Diagnosed systemic data integrity drift by conducting a full-stack Root Cause Analysis (**RCA**) on a misconfigured temporal filter invariant with a hardcoded 30-day lookback. Executed a massive historical recovery program (recasting 28,000+, 50,000+, and edge case records), expanding analytical coverage by **800%** (1 $\rightarrow$ 9+ months) and raising ticket match accuracy to the highest level since system inception.
- **Data Integrity & Temporal Modeling:** Fixed historical attribution defects by implementing **SCD Type 2** temporal joins on creation date (between start/end dates) to accurately resolve employee hierarchy changes against **LDAP** data. Refactored legacy logic to use stable keys (Agent Email) rather than volatile Login IDs, ensuring robust historical reporting and data lineage integrity.
- **Pipeline Resilience & Granularity:** Instituted rigorous warehouse design standards by enforcing idempotency and atomic writes via composite **UPSERT** keys (Request/Ticket/CI IDs). Applied **COALESCE**, **UPPER**, and **TRIM** sanitization for all join conditions and enforced pre-aggregation patterns, eliminating data inflation during retry scenarios and preserving model granularity across all environments.
- **Modular Architecture & Migration:** Decoupled transformation logic from ingestion by refactoring monolithic **SQL** into modular Clean and Calculate stages, a pattern analogous to **dbt** staging and marts. Standardized KPI logic by isolating **CASE**-based derivations in a dedicated calculation layer, accelerating peer reviews through clear separation of transformation and analytics code and ensuring business-ready metrics. Created a unified **VIEW** abstraction layer merging legacy and modern structures, enabling zero-downtime migration for downstream **BI** consumers.
- **Observability & Monitoring Infrastructure:** Selected to architect a **Python**-based data observability framework integrated with **Apache Airflow** using DAG modules and configuration-driven checks files. Designed automated schema validation, anomaly detection, and threshold alerting for critical data quality SLAs, shifting the team from reactive debugging to proactive monitoring.
- **Technical Governance & Documentation:** Promoted to Technical Gatekeeper within 3 months to govern the NTS domain by enforcing defensive coding standards. Established a rigorous proof-based methodology by authoring end-to-end validation docs (ERDs, Data Flow Diagrams, count-by-stage proofs) in **Confluence** and **Jira**. This approach de-risked complex implementations, ensured logically verified flows, and became the team standard for peer review and knowledge transfer.
- **Engineering Leadership:** Managed development through **GitLab** feature branches and code review workflows. Mentored peers on **SQL** query optimization (CTEs, Window Functions, Execution Plans) and led cross-functional stakeholder alignment sessions with directors and business analysts.

Bioinformatics Software Development Research Assistant — Sept 2022 – Present

Johns Hopkins University, Baltimore, MD – Remote, Part Time

- **Open-Source Platform Architecture:** Architected and maintained an open source, full stack bioinformatics platform using **Python**, **R**, **JavaScript**, and **C** with microservices and SOLID principles and **Docker**. Reduced analysis load times by **83%** through optimized caching and supported more than 100 global researchers.
- **Scalable ETL & Big Data Processing:** Engineered scalable **ETL** pipelines processing over 750 terabytes of multi omics data such as TCGA on high performance computing clusters using **Python**, **R**, **SQL**, and machine learning models including **SVM RFE** and **Random Forest**. Accelerated biomarker discovery by **40%** and reduced analysis time by **40%**.
- **Automated Data Quality & ML:** Implemented automated data quality and anomaly detection using unsupervised machine learning including **K-Means** and **DBSCAN** with **TensorFlow** within CI/CD pipelines. Improved data integrity by **30%** and 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**, improving usability and accessibility for researchers.
- **API Development & Integration:** Developed and optimized **REST** and **GraphQL** APIs to support real time data access and model simulations across research modules.
- **Cloud Infrastructure & DevOps:** Configured lightweight **AWS** environments including **EC2** and **S3** and automated testing and deployment workflows with **GitHub Actions**, improving reliability and collaboration across development teams.
- **Data Governance & Compliance:** Applied secure data management and governance practices to ensure compliance with institutional privacy and research ethics standards.
- **Cross-Functional Collaboration:** Collaborated with cross functional experts including oncologists and statisticians to align computational workflows with research goals and mentored peers on high performance computing and reproducible software practices.
- **Research & Technical Communication:** Authored multiple 35 page research manuscripts featuring interactive visual dashboards and reproducible analyses and presented award winning research at **ABRCMS** and **Harvard NCRC** conferences.

Software Development Research Assistant **Sept 2019 – Apr 2024**

University of Toronto, Toronto, ON – Hybrid, Part Time

- **Full-Stack Platform Engineering:** Engineered full stack bioinformatics platforms using **Python**, **R**, **C**, and **Java** with object oriented programming patterns to automate lab workflows, reducing analysis effort by more than 30 hours per week across 7 research teams.
- **SDLC & Requirements Translation:** Translated multidisciplinary research requirements into production grade software solutions, owning the full **software development life cycle (SDLC)** including requirements, architecture, implementation, testing, deployment, and maintenance.
- **DevOps & Containerization:** Implemented **Docker** based **DevOps** workflows to eliminate environment drift and cut setup and configuration time by **50%**, enabling reproducible and scalable computation.
- **Performance Optimization & UX:** Optimized data visualization performance in **Next.js** and **Tailwind CSS**, improving user interface render times by **45%** for large genomic datasets and enhancing research usability.
- **Engineering Leadership & Agile:** Led **Agile Scrum** adoption and mentored a team of 5 junior developers, increasing throughput and strengthening cross team collaboration.

Education

B.Sc. (Hons) Computer Science, Bioinformatics & Computational Biology — University of Toronto **June 2024**

- GPA: 3.96 / 4.0 | Relevant Coursework: Data Structures & Algorithms, Software Design & Engineering Principles, Systems Programming, Algorithm Design & Analysis, Theory of Computation, Operating Systems, Database Systems, Machine Learning, Distributed Systems, Cloud Computing, Computer Networks, Applied Bioinformatics, Systems Biology, Statistics & Probability, Calculus, Programming Languages (Python, C, R, Java), Web Technologies (HTML/CSS), Microsoft Office Suite (Excel, Powerpoint, Word)

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 <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 Systems Programming, Sockets, Multithreading, Backend Development, CI/CD, TDD

Automated Anomaly Detection System

- Engineered a full-stack anomaly detection platform using Node.js/Express backend to orchestrate Python (YOLOv5 object detection, LSTM behavior analysis) processing of video uploads, storing frames on AWS S3 and alert data in AWS RDS (PostgreSQL) via CRUD APIs.
- Developed an intuitive React/TypeScript/MUI frontend with dynamic alert filtering and visual evidence display; designed for containerized (Docker) deployment to AWS EC2 with CI/CD.

Key Tech: Node.js, Express, Python (PyTorch, YOLOv5, OpenCV), PostgreSQL (AWS RDS), React, TypeScript, MUI, Axios, AWS EC2, PM2, AWS S3, Docker, REST API, CRUD Operations, ML Pipeline Engineering, Cloud Architecture, CI/CD (conceptual), Full Stack Development, Deep Learning

MicrobiomeExplorer R Package (Open Source)

- Created and developed a modular R package for 16S rRNA/metagenomic analysis, integrating ETL pipelines, Bioconductor stats, and interactive Shiny visualizations with IoT compatibility.

Key Tech: R, Shiny (Frontend/Data Viz), Python, Bash, Bioconductor, ETL, Data Visualization, Backend Logic

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, Data Pipeline, DevOps

Red Blood Cell Counter

- Engineered a C application using image processing algorithms (segmentation, flood-fill) for automated RBC counting, reducing false positives 30% with attention to detail.

Key Tech: C, Image Processing, Algorithm Design, Data Structures, Memory Management, Scientific Computing

Awards & Achievements

- Plenary Speaker, National Collegiate Research Conference (NCRC) Harvard 2024
  (Selected as 1 of 12 plenary speakers from over 5,000 applicants)
- Best Detailed Oral Presentation, ABRCMS Conference 2023
  (Top presenter in division; 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 (Awarded 2022, 2023, 2024)
- Friends Of Arts And Science Award In Physical And Life Sciences, University of Toronto (Awarded 2022, 2023, 2024)