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, Remote, Full-Time

- **Architected and productionized** the mission-critical **Network Ticket Service (NTS)** data pipeline using **Python**, **SAS DI**, and **advanced SQL**, integrating **disparate operational systems** (**Maximo**, **SmartPath API** event streams) into a **Teradata analytical warehouse** through **multi-stage Extract Transform Load (ETL workflows**, **schema-versioned loads**, and **enforceable data contracts** across **DEV/QA/PROD** environments.
- **Diagnosed and resolved** systemic data integrity drift in the NTS pipeline by conducting a full-stack **Root Cause Analysis (RCA)** to correct a misaligned filtration invariant and establishing an **ongoing historical data recovery program**; **executed** three staged recasts correcting **28k+**, **50k+**, and several hundred high-complexity edge-case records, **expanding** analytical coverage from 1 to 9+ months (**+800%**) and **raising** ticket match accuracy to the highest level since system inception.
- **Re-architected** a critical performance anti-pattern by replacing a direct join to a **23-million-row live table** with a **pre-aggregated static reference** design, cutting query runtime from 12 mins to 2 mins and stabilizing nightly **SLA compliance**.
- **Engineered** a stateful **Python** algorithm for duration capture and sessionization by refactoring a flawed sequential method into a robust **two-pass group-by propagation model**; correctly propagated `request_id` across event sequences, achieving **zero calculation defects** in QA validation.
- **Instituted** rigorous data-warehouse design standards: implemented **composite UPSERT keys** (`request_id`, `ticket_number`, `CI_number`), applied `COALESCE/UPPER/TRIM` hygiene for joins, and enforced pre-aggregation patterns for idempotent loads—**eliminating data inflation** and preserving model granularity.
- **Standardized** KPI logic by isolating all **CASE-based derivations** in a dedicated **calculations** extraction layer; improved maintainability, ensured business-ready metrics, and accelerated peer reviews through clear separation of transformation vs. analytics code.
- **Promoted** to **technical gatekeeper** for the NTS data domain within **3 months**, leading stakeholder meetings, peer code reviews, and architectural audits (defensive coding, modularization). Began structured **cross-training in Apache Airflow and GCP orchestration concepts** to support future pipeline automation initiatives.
- **Established a rigorous proof-based problem-solving methodology**, authoring end-to-end **validation and architecture documentation** (**ERD, DFD, count-by-stage proofs**) in **Confluence** and **Jira**; this approach **de-risked complex implementations**, ensured **reproducible, logically verified data flows**, and became the team standard for **peer review** and **knowledge transfer**.

Bioinformatics Software Development Research Assistant — Sept 2022 – Present

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

- **Architected** and maintained an open-source, full-stack bioinformatics platform (**Python, R, JS, C**) using microservices and SOLID principles with **Docker**; **reduced** analysis load times 83% via optimized caching, supporting 100+ global researchers.
- **Engineered** scalable ETL pipelines processing 750+ TB multi-omics data (TCGA, etc.) on HPC clusters using **Python, R, SQL**, and ML models (**SVM-RFE, Random Forest**); **accelerated** biomarker discovery 40% and cut analysis time 40%.
- **Implemented** automated data quality/anomaly detection using unsupervised ML (**K-Means, DBSCAN** via **TensorFlow**) within CI/CD pipelines; **improved** data integrity 30% and validated biomarker analysis software (**TF, Keras, Scikit-learn**).
- **Built** interactive data visualization dashboards (**Shiny, React, D3.js**) for molecular modeling and educational use, improving usability and accessibility for researchers.
- **Developed and optimized REST/GraphQL APIs** to support real-time data access and model simulations across research modules.
- **Configured** lightweight **AWS (EC2/S3)** environments and automated testing/deployment workflows with **GitHub Actions**, improving reliability and collaboration across development teams.
- **Applied** secure data management and governance practices ensuring compliance with institutional privacy and research ethics standards.
- **Collaborated** with cross-functional experts (oncologists, statisticians) to align computational workflows with research goals and **mentored** peers on HPC and reproducible software practices.
- **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

- **Engineered** full-stack bioinformatics platforms (**Python, R, C++, Java**) using OOP patterns to automate lab workflows, **reducing analysis effort by 30+ hours/week** across **7 research teams**.
- **Translated** multidisciplinary research requirements into production-grade software solutions, **owning the full SDLC** (requirements, architecture, implementation, testing, deployment, maintenance).

- **Implemented Docker**-based DevOps workflows to eliminate environment drift and **cut setup/configuration time by 50%**, enabling reproducible and scalable computation.
- **Optimized** data visualization performance in **Next.js + Tailwind CSS**, **improving UI render times by 45%** for large genomic datasets and enhancing research usability.
- **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

Stock Market Prediction Pipeline

- Architected a real-time stock prediction system integrating Kafka streaming data, feature engineering, and ML models (RandomForest, XGBoost), achieving +/-5% prediction error. Deployed via Dockerized Flask Web Service.
  Key Tech: Python (Pandas, Scikit-learn, Flask), Kafka, Docker, ML, REST API, Streaming Data, Backend Development, Data Engineering

Random Fact Generator

- Developed a full-stack MERN application with web scraping, REST APIs (Node.js/Express), AI image generation (DALL-E), MongoDB, and Redis caching, focusing on modular architecture and QA.
  Key Tech: Node.js, Express, React, MongoDB, Redis, REST API, Web Scraping, JavaScript, MERN, Full Stack Development, Web Development

Job Application Tracker

- Architected secure ASP.NET Core Web API (C#) with RESTful CRUD endpoints (EF Core, SQL/PostgreSQL), JWT auth, and responsive Blazor/React frontend; deployed via Docker to Azure using CI/CD.
  Key Tech: .NET 8, C#, ASP.NET Core, EF Core, Blazor Server/React, SQL Server/PostgreSQL, Docker, Azure App Service, CI/CD, REST API, JWT Auth.

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)