ARUNIT BAIDYA

baidyaarunit@gmail.com | linkedin.com/in/arunit-baidya | github.com/AR-UNIT | 617-935-5115

EDUCATION

Northeastern University, Khoury College of Computer Sciences — Aug 2024 – April 2026
Master of Science, Computer Science — Boston, MA, US

Anna University, PSG College Of Technology — Aug 2018 – July 2022
Bachelor of Engineering, Computer Science and Engineering — Coimbatore, India

WORK EXPERIENCE

Arcesium | Data, Backend, Cloud-Native

Software Engineer — Jul 2022 - Jul 2024, Bangalore

- Profiled and optimized **Python & Java code** performing **Spark** operations in end-to-end big data system to meet client SLOs (≤10 min time to output with 1,000–100,000 records per input file), ***improving process times by 40%.***
- Built JSON config-driven data-filtering module for Spark ETL in ***reducing compute usage by 30%*** via partition pruning.
- Optimized resource provisioning sizes (T-shirt model) in **ArgoWorkflow** and **Kubernetes, *adopted in 10+ workflows.***
- Automated CI/CD pipelines with **Jenkins, Docker,** & **Kubernetes**, integrating JUnit for *testing ensuring 85% coverage.*
- Implemented **SQL** for SCD type 2 merge, enabling bi-temporality & historical querying, state of data at point in time.
- Created LLD documentation with business logic translation, ***reducing*** cross-functional team ***inquiries by 50%.***

Software Engineering Intern — Feb 2022 - May 2022, Remote

- *Increased **ETL** throughput by 70%* via preprocessing stage to split heterogeneous data for parallel processing.
- Built data ingestion module supporting multiple input types with schema validation for data consistency.
- Boosted ETL reliability by engineering module in **Python** to extract malformed data to global kickouts Delta-table.
- Enhanced ETL observability by efficiently logging run history and storing execution stats in Delta-Tables.

Software Engineering Intern — May 2021 - Jul 2021, Remote

- Automated PDF Parsing in **Python** to identify and parse tabular information using OpenCV and Camelot.
- Implemented **CRON processing** to parse PDFs mapped against filters & annotations, ***saving 4 hours of work/week***.

PROJECTS

Distributed Scalable TODO-List API (GitHub) | Golang, Kafka, Redis, Docker, Ansible

- Designed a highly scalable REST API for TODO-List application using **PostgreSQL**, with design extensible to new data storage methods & query strategies, and **Ansible** setup for automated deployments via **GitHub CI/CD** pipeline runner.
- Deployed application using **Docker** & **Kubernetes**, with JWT authentication, rate limiting, **Kafka** for event-driven architecture, **Redis** for caching and **batch database updates,** and **Prometheus** for metrics collection.

Image Processing Application (GitHub) | Java, SOLID, MVC

- Developed an image processing application in **Java** supporting CLI, script, and GUI inputs, demonstrating **SOLID principles**, higher order functions, **object-oriented design**, **MVC**, design patterns, and **test-driven development**.

Qthreads (GitHub) | C, Multithreaded Programming

- Engineered A **user-space thread library** in **C** implementing thread management operations (create, yield, join, sleep) and concurrency management constructs (mutex, condition variable) in a **cooperative threading** model.

Simple Linux Shell (GitHub) | C, Linux

- Built Linux shell in **C** with command execution for internal & external commands, piping, & I/O redirection using **fork**.

SKILLS

Programming Languages:	Python, Java, Go, C, C++, JavaScript, HTML, CSS
Frameworks and Libraries:	REST, Redis, Swing, Pytest, JUnit, Argo Workflow, Ansible, Prometheus
Data Engineering:	Spark, Kafka, ETL, DeltaLake, Snowflake, PostgresDB, SQL, NoSQL
Cloud and DevOps:	API development, AWS, Elasticsearch, Docker, Kubernetes, Terraform
Software Development:	CI/CD, Agile Development, Test-Driven Development, Code Reviews, Git