(973) 519-5397
Cleveland, Ohio
naveen.kannan12@gmail.com

Naveen Kannan

Data Engineer

Portfolio: Data Science Blog
github.com/naveenk2022
linkedin.com/in/naveenkannan1

Three (3) years expert skills in database development (PostgreSQL, MySQL, Hadoop) using ETL/ELT tools (SQLAlchemy, Alembic, Spark), developing production level REST APIs (FastAPI, Flask) and deploying scalable, containerized (Docker) workflows in HPC (High Performance Computing), AWS (EC2, RDS) and Azure (VM) cloud platforms.

Skills

Programming Languages	Python, R, Bash, SQL, LaTeX, Markdown
Scripting/Automation	Unix Shell Scripting, Ansible
Data Engineering	Apache Spark, PySpark, Hadoop (HDFS, YARN, MapReduce), Hive
ETL and Data Pipelines	Airflow, SQLAlchemy, Alembic, Pydantic, Pandas, Dask
Relational Databases	PostgreSQL, MySQL, HiveQL
Big Data Storage	Hive, Parquet, ORC, Data Lake Storage (HDFS, Azure Blob)
Machine Learning	Tensorflow, YOLOv4, Scikit-learn, PyTorch
Distributed Computing	SLURM, YARN, HPC Job Scheduling
DevOps	Ansible, CI/CD Pipelines, Infrastructure as Code
Database Backup	pgBackRest, Unix Shell Scripts
Containerization	Docker, Docker Compose, Docker Swarm
Version Control	Git, GitHub, GitLab
Data Analysis	Geospatial Analytics, Statistical Analysis
Web Frameworks	FastAPI, Django, Flask

Technical Experience

Data Engineer **July 2024 — Present**
Trailhead Biosystems *Cleveland, Ohio*

- Deployed and configured PostgreSQL 17 on Azure VMs, implementing pgBackRest for automated daily backups and disaster recovery, achieving 99.99% data availability.
- Designed a fully normalized relational data model using SQLAlchemy and versioned schema migrations with Alembic, integrating 5+ previously siloed data sources to enable unified, cross-functional analytics.
- Optimized high-complexity SQL queries, reducing average query response times by 30%, significantly improving performance for dashboards and internal reporting tools.
- Hardened Azure infrastructure security by configuring a site-to-site VPN for on-premises access and deploying Sophos endpoint VPN for remote access, eliminating insecure public exposure and aligning with enterprise security policies.
- Developed and deployed RESTful APIs using FastAPI, served via Gunicorn behind Nginx, and secured access through Microsoft Entra ID (Azure AD) authentication, enabling scalable, enterprise-grade data services.
- Built a Django-based frontend for data entry and visualization, deployed on Azure with secure intranet access via Microsoft Entra ID (Azure AD), streamlining workflows and reducing manual data entry errors by over 70%.
- Migrated legacy Excel-based tracking systems to a centralized PostgreSQL database, improving data traceability, reducing duplication, and enhancing cross-team accessibility and reporting efficiency.
- Engineered Python ETL pipeline to extract proprietary model coefficient data from 5+ years of HTML reports, transforming 250+ unstructured documents into a normalized SQL database with 99.7% accuracy.
- Designed optimized data model that reduced query execution time by 85% while enabling cross-model comparisons previously impossible with legacy systems.
- Developed interactive analytics dashboard that decreased data access time from days to minutes, driving 40% faster decision-making for 30+ stakeholders and supporting $2M in annual investment decisions.

Research Associate **May 2023 — July 2024**
Department of Population and Quantitative Health Sciences, Case Western Reserve University *Cleveland, Ohio*

- Configured a PXE boot server using dnsmasq to automate remote deployment of RHEL and Ubuntu images across an 18-node bare-metal compute cluster, updating the outdated images,and reduced installation time by over 200%.
- Developed custom Ansible roles and shell scripts to fully automate the deployment and configuration of Apache Hadoop (HDFS, YARN, MapReduce), Hive, and Apache Spark across an 18-node bare-metal cluster, reducing setup time by 80% and ensuring consistent, repeatable environments for large-scale data processing.

(973) 519-5397
Cleveland, Ohio
naveen.kannan12@gmail.com

Naveen Kannan

Data Engineer

Portfolio: Data Science Blog
github.com/naveenk2022
linkedin.com/in/naveenkannan1

- Configured and optimized YARN, Spark, and MapReduce resource allocation parameters—such as executor memory, cores, and container sizing—to improve cluster utilization by 40% and reduce job runtime latency by up to 60% for large-scale genomic workflows.
- Designed and implemented Spark and PySpark pipelines to annotate and analyze 362 million structural genetic variants associated with Alzheimer's disease, leveraging distributed computing in a high-performance cluster to accelerate data processing by 10x and enable scalable downstream analysis for ADSP (Alzheimer's Disease Sequencing Project) collaborators.
- Integrated PostgreSQL with Hadoop/Spark using JDBC to automate schema generation and perform scalable ETL/ELT of over 4 TB of genomic flatfiles (VCF, BAM, CRAM) into a PostgreSQL database, reducing manual data onboarding time by 85% and enabling efficient downstream querying and analytics.
- Collaborated with cross-functional research teams including scientists, bioinformaticians, and data analysts to define data requirements and develop Spark-based pipelines that transformed raw genomic research outputs into curated datasets, accelerating regulatory reporting and downstream analytics by 70%.
- Conducted root cause analysis and performance tuning for Spark and Hive jobs using execution plans and job history logs, reducing job failures and increasing throughput for critical analytics workloads.
- Deployed bioinformatics pipelines installed with Conda, PIP and Mamba via Docker containers within an AWS EC2 instance, connecting to an S3 bucket containing several hundred gigabytes of genomic data for data annotation.
- Developed and trained a Generative Adversarial Network (GAN) using PyTorch and TensorFlow on the 1000 Genomes dataset to generate synthetic haplotype data that passed Hardy-Weinberg equilibrium testing. Validated the synthetic data through PCA analysis, demonstrating comparability to real data in aggregate metrics for genetic variation.

Graduate Research Assistant **January 2023 — May 2023**
Department of Population and Quantitative Health Sciences, Case Western Reserve University *Cleveland, Ohio*

- Engineered Docker containers for streamlined deployment of specialized R packages and Python modules managed by Conda and Mamba, optimizing analyses for single-cell RNA datasets.

Graduate Research Assistant **May 2022 — December 2022**
GIS Health and Hazards Lab, School of Medicine, Case Western Reserve University *Cleveland, Ohio*

- Developed Python scripts to extract and process video frames from geospatial video footage of refugee camps in the Democratic Republic of Congo following the Mt. Nyiragongo eruption, enabling frame-by-frame analysis at scale.
- Implemented and fine-tuned a YOLOv4 object detection model to identify refugee tents within individual frames with >92% detection accuracy, using a custom-trained dataset for high performance in low-contrast, real-world conditions.
- Automated geotagging of detected objects by associating each video frame with its embedded GPS metadata, enabling precise spatial mapping of tent locations across multiple camp zones.
- Generated interactive geospatial heat maps of refugee tent distribution using GeoPands, facilitating rapid situational awareness.
- Reduced manual image annotation and geospatial analysis time by over 90%, significantly accelerating data delivery for emergency response coordination.
- Contributed to peer-reviewed research publication, showcasing the use of computer vision and geospatial analytics for post-disaster humanitarian applications.

Junior Resident Doctor **July 2020 — September 2020**
Department of Psychiatry, Saveetha Medical College *Chennai, India*

- Evaluated patients and diagnosed psychiatric illnesses in an outpatient setting.

Junior Resident Doctor **March 2019 — March 2020**
Madras Medical College *Chennai, India*

- Rotated through the following departments: Internal Medicine, Pediatrics, General Surgery, Obstetrics and Gynecology, Community Medicine, Psychiatry, Emergency Trauma Ward, Labor Ward.

Education

MSc in Biomedical and Health Informatics , *Case Western Reserve University* January 2022 — August 2023
Bachelor of Medicine and Bachelor of Surgery, *Madras Medical College* September 2016 — March 2020