Kaushik Patil

(857) 544-8130 • patil.kau@northeastern.edu
https://www.linkedin.com/in/kaushikpatil/ • https://github.com/kaushikpatil2911/ • Austin, Texas, USA

EDUCATION

Northeastern University, Boston, MA — **Sep 2019 - May 2021**
Master of Science in Information Systems — **GPA:** 3.51 / 4.0

University of Mumbai, Mumbai, India — **Sep 2014 - Jun 2018**
Bachelor of Engineering in Electronics and Telecommunication Engineering

TECHNICAL SKILLS

Programming Languages:	Python 3.x, SQL, Java, Bash
Big Data & Distributed Systems:	Apache Spark, Hadoop, Hive, Databricks
ETL & Orchestration:	Apache Airflow, Talend, dbt, Alteryx
Databases & Warehousing:	PostgreSQL, MySQL, MongoDB, Redis, Amazon Redshift, Snowflake, Google BigQuery
Streaming & Messaging:	Apache Kafka, RabbitMQ
Cloud Platforms:	AWS (Lambda, EMR, S3, Glue), Google Cloud Platform, Microsoft Azure
Infrastructure & DevOps:	Docker, Kubernetes, CI/CD, Git
Visualization & Monitoring:	Power BI, Tableau, Kibana

EXPERIENCE

Informatica — Toronto, Ontario, Canada — **Jun 2024 - Present**
Senior Data Governance Specialist - Support — Austin, Texas, USA — **Aug 2021 - Jun 2024**

- Extracted, supported, and optimized large-scale **metadata ingestion**, **data lineage**, and **data integration processes and workflows** using Informatica's Cloud Data Governance and Catalog (CDGC) and on-prem Enterprise Data Catalog (EDC), resolving 1,000+ complex technical issues with a 98% customer satisfaction score.
- Collaborated with platform engineering and product teams to debug and replicate **data ingestion and integration** issues, and tested connector-level patches, contributing to 25+ product enhancements per year.
- Led integration and configuration of CDGC/EDC with 20+ modern data platforms (e.g., **Snowflake**, **Google BigQuery**, **Oracle**, **Power BI**, **Tableau**), leveraging **SQL, REST APIs**, and **cloud-native tools** to improve ingestion accuracy and reduce setup time by 40%.
- Tuned and scheduled **metadata ingestion jobs**, leveraging performance metrics to optimize extraction efficiency, reduce latency, and increase throughput across both **cloud (IDMC)** and hybrid environments.
- Applied **data pipeline optimization** techniques and troubleshooting methods to resolve failures related to **Spark**-based profiling, schema drift, and large-scale data source changes.

RAWsb Technologies Inc — Boston, United States — **Jun 2021 – Aug 2021**
Software Development Engineer Intern

- Designed a COVID-19 Information Search Assistant prototype, improving search relevance by **15%** by integrating data from three online sources and processing **thousands** of queries during development and developed core components of a distributed web crawler and caching system in C, handling **tens of thousands** of web pages to support initial **SparkSQL** integration

ACADEMIC PROJECTS

Medical Plan RESTful Application — **Jan 2021 – Apr 2021**

- Built a RESTful API with full CRUD operations and JWT-based auth, supporting test workloads of 500+ simulated users/day with **10% lower error rates**
- Implemented asynchronous data indexing using **RabbitMQ** and **Elasticsearch**, achieving search response times under 2 seconds
- Leveraged **Redis** for JSON document storage with **<10ms** latency and visualized search trends via **Kibana**, aiding usage analysis during testing

IMDb Data Analysis and Visualization — **Sep 2020 – Dec 2020**

- Extracted and profiled 10M+ rows from IMDb-related datasets, improving data quality by **25%** using Talend-based cleaning pipelines
- Streamlined ETL workflows using **Talend** and **Alteryx**, reducing transformation time by **20%** while supporting 4+ dimensions with SCD logic
- Created **6+ dashboards** in **Power BI & Tableau** to visualize box office trends and actor/movie insights, improving reporting turnaround by **35%**

Amazon Sports Reviews Big Data Analysis — **May 2020 – Aug 2020**

- Implemented and optimized **MapReduce** & **Hive** jobs to process **5+ million** Amazon Sports Reviews, reducing ETL pipeline runtime by **42%**
- Improved Java EE integration for MapReduce, decreasing code development and execution time by **25%** across **15+ MapReduce jobs**
- Stored over **500GB** of processed intermediate data in **MongoDB** clusters, increasing data retrieval speed by **18%** for downstream analytics
- Published **10+ Tableau dashboards** enabling stakeholders to analyze sentiment trends, resulting in a **20% increase** in marketing accuracy

Sentiment Analysis of Amazon Product Reviews — **Sep 2019 – Dec 2019**

- Processed **100,000+** Amazon reviews across 3 product categories, balancing dataset to **80% positive and 20% negative** classes using TF-IDF
- Trained and compared **5 machine learning models**, achieving highest accuracy of **87% with Naive Bayes** on test set of 20,000 reviews
- Improved model performance by **15%** through feature engineering and hyperparameter tuning across 10+ experiments
- Generated **5+ visualizations** including confusion matrices and sentiment distribution plots in Jupyter Notebooks to communicate insights