RAJAT KAPGATE

Open to Relocation | +1 (812) 553-2822 | rkapgate@iu.edu | linkedin.com/in/rajat-kapgate | github.com/rajatk9962

EDUCATION

Indiana University Bloomington, USA **Aug 2023 – May 2025**
Master of Science in Data Science **GPA: 3.94/4.00**
Coursework: Data Mining, Machine Learning, Advanced Database Concepts, Statistics, Data Visualization, Algorithms

University of Mumbai, India **Aug 2017 – Jun 2021**
Bachelor of Engineering in Computer Engineering **GPA: 3.50/4.00**
Coursework: Big Data Analytics, Elements of Artificial Intelligence, Advanced DB, Data Structures, Exploratory Data Analysis

PROFESSIONAL EXPERIENCE

Research Data Analyst | Indiana University School of Optometry, Bloomington, USA **May 2024 – Present**

- Analyzed 500+ GB of infant data, performing **statistical analysis** to correlate **head and eye movement** for identifying potential early markers of **infant eye disorders**, as part of an NIH-funded study (**Grant EY032897**).
- Preprocessed **time series data** using techniques like IQR filtering, rolling sum, low pass filtering and non-max suppression to isolate meaningful head motion segments, enhancing data quality by **40%**.
- Performed **ANOVA** across five age groups, identifying statistically significant motor control variations with a p-value less than 0.05.
- Engineered **advanced data visualizations**, including head movement reconstruction with Unity and Open3D, KDE, polar plots, and correlation maps to analyze infant head dynamics.

Data Science Co-op | Boehringer Ingelheim Pharmaceuticals, Ridgefield, USA **May 2024 – Nov 2024**

- Orchestrated an ETL pipeline to process **2M+ drug price data points** from the Nuro API, optimizing SQL workflows on **AWS RedShift**, automating with cron jobs, and storing results in **AWS S3**.
- Implemented a **FinOps** cost analytics dashboard with **Streamlit**, integrating cloud cost data for better visibility. Monitored Jenkins **CI/CD pipelines** and performed root cause analysis, uncovering inefficiencies and cutting compute costs by **$200K+.**
- Accomplished a **70%** reduction in **reporting** turnaround for the drug Jardiance by leveraging **Large Language Models, LangChain** CSV **agents**, **Azure** Chat API, and Python-pptx**,** leading to an increase in decision making efficiency.
- Devised a **Retrieval-Augmented Generation (RAG)** system on proprietary organizational data using **Azure GPT-4o** and **FAISS** for vector-based similarity search, cutting research effort by 40%.

Data Analyst | TCS Research, Mumbai, India **Jun 2021 – Present (On sabbatical leave)**

- Utilized **Google BigQuery** for crafting intricate database queries and harnessing BI Tools such as **Tableau** and **Power BI** to craft impactful dashboards. Managed a high-performing team of IT professionals, resulting in a **40%** increase in project efficiency.
- Improved employee retention by **5%** by leveraging **SAP** data to integrate **KPIs** (attrition rate, turnover, tenure) with **multivariate forecasting**, identifying and remediating 3 critical attrition drivers and informing targeted retention strategies.
- Developed **Excel**-based financial models incorporating **macros** and advanced functions (VLOOKUP, INDEX-MATCH) to automate team-level expense reporting, increasing accuracy and reducing manual effort by **40%.**
- Served as a key resource for data science, **mentoring** six associates and simplifying complex analytics for managers, leading to a **20%** improvement in decision-making and team efficiency in visualization and reporting.
- Led an **innovative** deep learning project, building a MultiStream CNN-LSTM model for Indian Sign Language recognition. Reduced parameters from **300K+ to 2.5K**, improving convergence by 70% and achieving **24.4% WER** on the RWTH Phoenix benchmark.

ACADEMIC PROJECTS

Serverless ETL Pipeline on AWS for Sales Analytics (AWS Data Engineering)

- Built an end-to-end serverless ETL pipeline using **AWS CloudFormation**, **Glue (PySpark), S3**, and **Redshift** to ingest, transform, and aggregate sales data, automating job orchestration and schema management via Glue Crawlers and interactive notebooks.

Big Data & Climate Change Prediction Pipeline (PySpark)

- Engineered a scalable data pipeline using MongoDB and PySpark to ingest and process global temperature, performing distributed ETL and forecasting trends via linear regression; deployed on JetStream2 with cloud-based visualization for high-volume datasets.

Patient Outcomes (Tableau)

- Unearthed **5 key trends** in MIMIC-III dataset through **data visualization**, providing **actionable insights** into patient outcomes.

A/B Testing & Marketing Campaign Optimization

- Conducted A/B testing and regression analysis on 365-day Facebook and Google AdWords data to assess conversions and cost efficiency. Used hypothesis testing and cointegration analysis to optimize ad spend allocation and improve ROI.

TECHNICAL SKILLS

- **Programming Languages:** Python, R, SQL, C++, Java, JavaScript, ReactJS, HTML, CSS, XML
- **Big Data & Processing:** Apache Spark, PySpark, Hadoop, Kafka, Snowflake
- **Workflow & ETL:** Airflow, AWS Step Functions, Glue, dbt, Spark SQL, Delta Lake, Jenkins
- **Databases:** PostgreSQL, MySQL, MongoDB, DynamoDB, Redshift, BigQuery, Parquet
- **Business Intelligence:** Tableau, Power BI, Looker Studio, ggplot2, GeoPandas, Seaborn, Excel, NumPy, Pandas
- **Cloud Tools:** AWS (S3, Glue, Lambda, Redshift, CloudFormation), Azure (Data Factory, Databricks), GCP (BigQuery, Dataflow)