MRUNMAYEE DHAPRE

San Jose | +1(408)422-7225 | mrunmayee.dhapre.98@gmail.com | LinkedIn | GitHub | Medium

EDUCATION

M.S., Computer Science (Big Data, Machine Learning, Computer Networks, Graph Neural Networks) **Aug 2023 - May 2025**
San Jose State University, San Jose
Course Work:

B.Tech, Computer Engineering **Sep 2016 - Sep 2020**
VJTI (Mumbai University), Mumbai

TECHNICAL SKILLS

- **Languages and Web Dev:** Python, Java, Javascript, SQL, PySpark, Flask, Django, FastAPI, Angular, React, CSS, Streamlit.
- **AI-ML:** Keras, Spark MLlib, ARIMA, RNN, LSTM, GRU, CNN, LangChain, OpenAI, HuggingFace, LlamaIndex, AstraDB.
- **Big Data and ETL Tools:** Hadoop, Hive, Spark, Pandas, Snowflake, Delta Lake, BigQuery.
- **Database Management:** MSSQL, MySQL, CosmosDB, Cassandra, MongoDB, HBase, Oracle.
- **Cloud Platforms:** Microsoft Azure **(**Kubernetes, Stream Analytics, ML Studio), Amazon Web Services, Google Cloud Platform.
- **Miscellaneous:** Agile development, NVIDIA DGX, Grafana, Root Cause Analysis, Unit Testing (pytest, unittest).

WORK EXPERIENCE

Data Engineer, Fractal Analytics, Mumbai (Python, SQL, Azure) **Sep 2020 - Jul 2023**

- Designed a warehousing solution to unify all on-premise deployments, resulting in a 66% increase in operational efficiency.
  - Devised a scalable **event-driven microservices architecture** reducing the onboarding time for a new warehouse by 50%.
  - Led a backend team of 4-6 developers as a senior member to build Extract-Transform-Load (**ETL**) pipelines and **REST** APIs.
  - Streamlined the database design to accommodate real-time and historical data across 3 microservices, equipped for scaling.
  - Delineated dataflows to calculate approximately 150 warehouse metrics for dashboard integration.
  - Managed the deployment of the **Flask** (python) application on **Kubernetes** cluster in an agile development cycle.
- Optimized Root Cause Analysis by 60% developing a **Power BI dashboard** to visualize errors within application logs.
- Spearheaded POCs involving **Azure Data Factory, Stream Analytics, graphQL** for the application with an SLA of 1 min.

INTERNSHIPS

Research Assistant, San Jose State University, San Jose (Airflow, Keras, ARIMA) **Sep 2023 - May 2024**

- Modeled the groundwater data (8.5m records) from SRNL for a research project under the Department of Energy.
- Employed ARIMA, Prophet, RNN models for Time Series Forecasting on HPC devices minimizing training time by 67%.
- Developed Machine Learning pipeline to retrain the model periodically with Airflow automating workflows by 50%.

Software Intern, Vorbitech Solutions, Pune (Angular, PHP, Flask) **Jun 2019 - Jul 2019**

- Executed features for a compliance management solution in Angular, achieving production deployment within 2 weeks.
- Orchestrated a robust web scraping application in Flask to detect dead links within web pages housing approximately 1000 links.

ACADEMIC PROJECTS

FinanceAgent, Rag-a-thon (GenAI, LlamaIndex, OpenAPI) **Feb 2024**

- Developed a chatbot to provide information on quarterly earnings and share prices of major corporations, as a financial assistant.
- Configured RAG Agent for data from 2 types of sources: PDFs (stored in vector database) and CSV files
- Established streaming communication between backend (FastAPI) and frontend (React) for better user experience.

Predictive Modeling for Operational Technology (OT) Application (Keras, LSTM) **Sep 2019 - May 2020**

- Achieved a remarkable accuracy of 96.8% by performing Time Series Forecasting on Industrial Control System logs.
- Innovated hybrid LSTM and GRU models for anomaly detection, leveraging the NVIDIA DGX server equipped with 4 GPUs.
- Published the findings in a technical paper featured in the International Journal for Artificial Intelligence, 2021.

ACHIEVEMENTS

- Achieved 3rd place in the Fractal Hackathon 2023 by developing a robust machine learning model for predicting credit defaults.
- Earned a double promotion in recognition of outstanding performance at Fractal Analytics in 2023.

- **Tell me about yourself**

- Currently grad student majoring in CS at SJSU
- Bachelors in CE in 2020.
- Data Engineer at Fractal for 3 years, designing ETL pipelines to optimize warehouse operations
- developed expertise in Python and SQL (and gained hands-on experience in other Azure cloud offerings like Functions, ServiceBus, Kubernetes, PowerBI, Data Factory, Monitor, etc).
- Currently also a research assistant working on machine learning pipelines for environmental monitoring. Techstack: Airflow with Keras.
- As my semester ends, I am looking for a summer internship to get some more industry experience in the data engineering domain.

- **Why Lending Club?**

After participating in a hackathon exploring models for credit default prediction, I developed an interest in this domain. Lending Club has been in this industry for more than a decade now. I think it has grown quite stably, establishing itself as an industry leader. It is one of the best companies for my journey in this domain.

- **Why this role?**

I attended the data tour by Snowflake last year in Santa Clara and have been fascinated by the product ever since. I have had some hands-on with the product through workshops, etc but I want to use it in a professional setting and see it in action. I am looking forward to working on it.

- **Tell me about your research as RA**

We are given a groundwater dataset monitored by SRNL and are tasked with drawing insights from it. Currently I am building a time series forecasting model to predict the concentrations of different contaminants. These models can be incorporated into early warning systems. I am also working on building automated pipelines to update these models on new data from the sensors in near real time.

- **Tell me about a time when you had to deal with a challenging project in work experience:**

When I first joined Fractal, I was assigned to a team designing the architecture for a warehousing and dashboarding project. The client had about 30 warehouses across Europe with different systems. Our application would read from these, store in a common datastore and display across a consistent dashboard. I was part of the ETL team which fetched data from source, performed transformations and calculations and fed it to the dashboard.

The goal was to design a system that would minimize onboarding time if a new site was to be added. This was the first problem as different sites had different data sources as well as different numbers of data sources. These data sources were generally SQL and Oracle databases. The queries for data extractions, meaning of columns as well as the granularity of data also varied based on the source system. A lot of transformation was needed before we could use this data.

The next complication was Change Data Capture. We needed to trigger our pipelines when a change was detected in the source system but there was no CDC mechanism on the client systems. We

tried different tools, like Azure Data Factory and Stream Analytics but without CDC it was difficult. This meant that we had to create such a mechanism ourselves.

Another requirement was that we had to keep things as configurable as possible, from data sources to UI. So the data sources could be different and multiple for each site. The dashboard could contain a varying number of KPIs which could be altered on an adhoc basis. This becomes more convoluted when you add the fact that we needed these KPIs calculations to be configurable. As in, a site could ask for a particular KPI with certain calculations and another site could ask for the same KPI with different calculations.

To fix these three issues our lead engineer proposed the micro services architecture a service for each complication. We separated our CDC in a service that would monitor the source system, capture data change and push it to a message queue. We designed the KPI calculation service which read these changes, performed the KPI calculations and stored some of it back to the database. This service also pushed the updates to another queue which would be read by the UI. The configurations were managed by a service that contained the mapping of each Site and KPI to the version it needed. The KPI service received these details from this service and performed the calculations accordingly.

As I had gained a very good business understanding, I designed the schema for the datastore and the data flows. So while we were designing these services and their tasks and in these data flows, I had a call with the team lead almost everyday to explain to him my approach and get everything validated. And he would help me if I was stuck somewhere. But as someone fresh out of college this was an extremely challenging task.

- **Tell me about a time you used data to improve business decisions / Can you describe a situation where you had to work on a complex coding problem as part of a team? How did you collaborate with your colleagues to find a solution?**

When I was working at Fractal, we were designing a warehouse management system for a client with about 30 warehouses across Europe. An important KPI which we had to recalculate periodically and dashboard was product availability. This basically checks the availability of all products in a particular order to complete it. We had to take into account different parts of inventory in the warehouse as well as incoming and production inventory. The outgoing orders also provided the quality requirements for the product which could be very rigid or relatively lenient. The allocation strategy used would then allocate the goods based on these, trying to optimize the use. The operator was able to see the locations of these products to allocate them and perform other operations like cut products from an order, etc. The allocation strategy and inventory specifications were very complicated and needed a thorough understanding of the business use case. As I already had a decent business understanding, this task was given to me. To understand the requirements, I conducted elaborate discussions with the consultants, going back to the client whenever required. Finally after about 1-2 months I was able to deliver the module.

- **Talk about your experience in creating dashboards and visualizations:**

There was a project in our company where the client wanted to see the errors happening in the system at any time for a set duration. I was assigned to create a dashboard for the same in PowerBI. I created a query to fetch error logs from Azure Monitor for the timeframe set dynamically by the user. This data was also categorized by the errors and presented as a hierarchical bar chart. The dashboard was frequently used by consultants to present to our clients and by developers to identify critical errors. Hence it significantly optimized the root cause analysis.

- **Explain your most exciting project (Spark)**

Recently I took a course in Machine Learning for Big Data. Here, we were introduced to various big data tools like Hadoop, MapReduce, Spark, Hive, etc. We also had to create a project in one of these towards the end of this course in groups of two. My partner and I decided to work with fake movie review detection with sentiment analysis as we believed parallel computing would be extremely helpful in this case. To get authentic data, we had webscraped this dataset from the rotten tomato website. We then imported this data into the databricks community cluster on Spark and performed the NLP preprocessing on it - cleaning, tokenization, stemming, lemmatization and embedding. Different models were tried with word2vec embeddings like random forest, SVM, neural networks, etc.

The trickiest part for this was the labeling. As this was a webscraped dataset, it did not have any labels. We debated on various approaches for labeling - the most naive one being marking the polar reviews as fake or biased. Finally we settled on the approach of attaching a sentiment score to the review and assigning fake of the sentiment score does not match the rating provided. I have written a blog about these on my page in medium.

- **Tell me about a project you are most proud of:**

During my undergraduate studies, I undertook a significant project focusing on time series analysis, where I applied recurrent neural networks (RNNs) to forecast system logs within an industrial control system. This involved developing and training RNN models to predict subsequent events, ultimately aiding in anomaly detection. I took a novel approach by crafting an architecture that integrated various embedded RNN outputs. We detailed it in a published paper for the International Journal of Artificial Intelligence. I am very proud of my work in this project as it was something I learnt from scratch and I loved to come up with innovative solutions for the problem.

- **Talk about a project not in your resume**

FinanceAgent

Recently I participated in a GenAI hackathon on Retrieval Augmented Generation, hosted by LlamaIndex. Here we formed a group of 4 and worked on creating a chat agent that could answer financial queries for novice traders. This agent was given access to the quarterly earnings and share prices of about 10 MNCs as context and asked questions about the same. The quarterly earnings were in PDF format and share prices were a CSV file. Based on the details mentioned in the question, the agent would identify the appropriate source to query and provide answers. We used the OpenAI LLMs and the Llamma Index libraries for this, with FastAPIs at backend and React at the front end. I specifically worked on the backend as well as setting up the agent. The integration and streaming communication were very complicated to accomplish and took a significant amount of time, second only to prompt engineering for the LLMs.

Book Recommendation:

After attending the Google DevFest, I got interested in GenAI, especially Retrieval Augmented Generation. To try this out, I worked on a book recommendation system with a goodreads dataset. The dataset contained books with descriptions. This was chunked and embedded and added to the vector

database. An LLM then ran on top of this database to perform question-answering. I asked the chatbot to provide recommendations based on the book I like. Then, I compared the genres of the books recommended.

- **Tell me a strength of yours:**

I enjoy trying out new approaches or innovative ideas. Whenever given a task, I try to think of an approach I might not have used before. And I believe I have a faster learning curve on new technologies/languages than others. In my previous organization, we were designing the architecture for a warehouse management system. During this time, I worked a lot on the POCs for the architecture, which involved switching technologies quite a bit. I really enjoyed it. During my undergrad, I was part of a project involving predictive analysis with RNN. I came up with a novel architecture for this model combining different embedded RNN outputs. Currently I am a research assistant working on machine learning for groundwater monitoring. In this, I try implementing different models with the dataset provided by SRNL.

- **Tell me a weakness of yours:**

Previously, I struggled to articulate my thoughts systematically. This became evident during work when explaining complex concepts to colleagues. To address this, I began preparing detailed, organized notes before meetings. This practice not only improved my communication skills but also extended to project management, where I now systematically document and share my experiences through blogging. This approach has significantly enhanced my ability to convey ideas clearly and coherently.

- **Explain a time you rigged the system**

While applying to companies, there is a company that allows only a couple of applications my a candidate. But I believed I was suitable for more than 2 roles at that company. So I applied with a different email ID. However, as I had the same mobile number that did not work. So I added my international mobile number(india). It accepted the number and directed me to verification. However, I did receive any message from the company for OTP. Dejected, I was about to give up on the idea when I received an OTP for some random activity. I was confused but cautious as I did not know its origin. Just to try something out, I attempted adding this OTP to the verification portal and it actually worked. I was able to send another application to the company.

- **Tell me a time you failed on a project:**

While working on the warehousing project, we encountered a challenge with the pop-out component, which displayed detailed item information. The data involved was complex, and we needed to decide how to handle it within our microservices architecture. One suggestion was to fetch the data directly from the source when the pop-out was clicked to save storage space. However, I expressed concern about potential data redundancy and the loss of valuable information for future analytics. Despite raising these concerns, we proceeded with the suggested approach based on senior engineer advice. During the client presentation, the flaw in the architecture became apparent when scalability issues were pointed out. This led to a complete overhaul of the architecture, requiring significant rework and a month of effort. Looking back, I regret not advocating more strongly for a different approach.

- **Tell me about a time you saved the day.**

One evening, around 10 pm, my team lead urgently summoned me to the war room to address a critical issue: the dashboard wasn't updating, and the client had raised concerns. In the war room call were our project manager, tech lead, DevOps engineer, front-end engineer, and myself, all collaborating to diagnose the problem. It became evident from the logs that the ETL pipelines weren't triggering, leading us to suspect an Azure-related issue. We reached out to Azure support while continuing to analyze the logs. After about two hours of intensive debugging, we identified the root cause: a network patch implemented by the client. Unfortunately, reverting this patch would take another day, and the client couldn't afford the downtime. Recognizing the urgency of the situation, I proposed leveraging our development environment triggers to initiate our production pipelines as a temporary solution. This quick fix allowed the system to remain operational until the patch was reverted, earning praise from my tech lead for my resourcefulness. The patch was successfully reverted the following day, resolving the issue.

- **Tell me about a time you had to deal with a difficult employee/confrontation**

Our system seemed to have some intermittent performance issues with multiprocessing in Flask. A newly appointed tech lead proposed switching to the Gunicorn server to address the problem. We implemented and deployed the change, only to face failures in production. The service was functional, but several ETL pipelines were underperforming and experiencing delays. After spending some time on RCA, we realized that the change might negatively impact the client if not resolved promptly. Despite facing resistance from the tech lead, I advocated for reverting the change to mitigate the issue. After persistent persuasion, the tech lead finally agreed to revert the modification. Subsequent investigation revealed compatibility issues between Kubernetes and Gunicorn, as evidenced by the restoration of normal pipeline functionality post-reversion.

- **Share an example of a project where you had to meet a tight deadline.**

There was a component which was to be delivered within a few weeks, we had actually allotted enough time to it and almost wrapped up development for it. But then there was a structural change that came in. Our consultant informed us that the logic for calculation was changed. As we had already committed to the previous deadline we could not negotiate as they argued this was the logic all along. So me and my colleague had worked through the weekend to deliver it on Monday. We were able to reuse some components but had to revise a lot and thorough testing had to be done before the release.

- **Why do you want to work in Data Engineering?**

Ever since I heard about big data in my undergrad I was fascinated by it. The applications it has are enormous. After experimenting with machine learning, I have come to understand the importance of having clean and meaningful data. I want to work in this domain to impact change.

- **How do you inspire your team?**

I believe one should always lead by example. I have seen most of my managers do that. If they work for 12 hrs only then they ask me to work for 10hrs. Another key aspect is communication. I will try my best to be transparent and honest with them. Positive and negative reaffirmations work on people depending on their personality. Hence knowing and understanding people and then applying the appropriate approach is important.

- **Talk about your work experience**

Onboarding stage in my previous company involved teaching us everything about cloud and data. We covered 3 cloud platforms: GCP, Azure and AWS. I was trained further in Azure covering mainly Synapse, ML Studio, Stream, ServiceBus,Event Grid, Data Lakes. I was also trained in ETL Tools like Databricks, Data Factory, Knime and dashboarding tools like Power BI, PowerApps, Tableau. We were also given lectures on business understanding of the CPG domain ranging from Supply chain to marketing, sales, etc. Data vendors like neilson and terminologies in point of sale data were also explained. During this I also completed an Azure data engineer exam.

My first project was Warehouse management and dashboarding. We were to automate dashboards in 60 warehouses across europe. They needed a unified interface for different data sources, minimum onboarding time, highly configurable UI, automations to reduce manpower, and high scalability. As a member of the ETL team, I conducted multiple POCs gaining invaluable business and data flows understanding. I attempted tools like ADF, Stream Analytics but they did not work without CDC. We finally settled on event driven microservices architecture. This was a flask application, entirely in python and SQL for query fetching. It was deployed on Kubernetes. I also worked in designing dataflows and schema for relational databases.

- **What's your ideal next role? What's your plan 5 years down the line?**

I want to continue working in the data domain, potentially working as a data architect or a machine learning engineer. I enjoy coming up with new and innovative designs and building end to end data solutions - from data collection to ML modeling.

- **What is one of the biggest decisions you've had to make and why did you do so?**

I was doing good work in my previous organization. Coming here for my Masters was probably one of the biggest decisions, definitely the hardest one. I wanted to advance my career and better myself by stepping out of my comfort zone. So I came here.

- **Describe a good working environment?**

Transparency

- Approachable open communication vertical
- Collaboration horizontal
- mutual respect among team members

Importance to Learning

- Technical learning
- Innovative ideas

- **Why do I match the role?**

In a hackathon at Fractal analytics, we worked on a case study trying to predict whether a debtor is likely to default on his/her loan and won the 3rd prize. I have also completed a project

performing ML in Spark (where I used Pyspark). A blog on it is also published.I published a technical paper on time series forecasting for industrial control systems.