Priyam Deepak Choksi

choksi.pr@northeastern.edu | +1 (857) 763-8346 | linkedin/choksipriyam | github.com/priyam-choksi

EDUCATION

Master of Science in Information Systems – Northeastern University, Boston September 2023 - May 2025
Courses : Adv Data Science, Big-Data Systems, Data Science, Database Design and Management
Master of Science in Information Technology – University of Mumbai, India May 2021- June 2022
Bachelor of Science in Information Technology – University of Mumbai, India June 2018 - May 2021
Courses : Data Structures & Algorithms, Business Intelligence, AI, Statistics, Cloud Computing, Computer Networks

TECHNICAL SKILLS

Programming	: Python, SQL, R, Scala, Java, C++, Typescript
Databases	: MySQL, MongoDB, PostgreSQL, Oracle 11g, Snowflake, Apache Hive
Big Data and Cloud	: AWS (S3, EC2, Lambda), Azure, Spark, Kafka, Athena, Databricks, Docker, Kubernetes
Data Science	: TensorFlow, PyTorch, Keras, Pandas, NumPy, spaCy, LLM, BERT, Word2Vec, NLTK, Deep Learning
BI/ETL	: Power BI, Tableau, Excel, Looker, Alteryx, Talend, Cloudera, Informatica, Domo, dbt

WORK EXPERIENCE

Data Engineer – Heeva Infra, India January 2023 - April 2023

- Engineered scalable data pipelines using **Apache Kafka, AWS Lambda and PostgreSQL** to automate the flow of customer and infrastructure data, reducing manual errors by **28%** and ensuring efficient real-time data processing
- Architected a data ingestion pipeline with **dbt**, integrating sales, inventory, and customer data from **8+** sources, reducing data latency by **25%** (from 100ms to 75ms), and enabling real-time analytics
- Built and maintained a multi-dimensional data warehouse using **Snowflake**, optimizing queries, partitioning, and indexing to cut storage costs by **9%** and enhance query performance by **15%**
- Improved ETL jobs by implementing **SCD framework**, automated data validation scripts, and error handling in **Python** and **SQL**, increasing data reliability by **30%** and reducing ETL processing time by **20%**
- Collaborated with senior engineers in developing a **continuous integration/continuous deployment (CI/CD)** pipeline, reducing deployment time by **24%** and increasing system reliability

Machine Learning and Data Science Intern – Fasttrack Software, India June 2022 - December 2022

- Developed and fine-tuned **CNN** using **TensorFlow** improving image classification accuracy by **12%** for product identification which enhanced the accuracy and efficiency of the company's inventory management system
- Assisted in deploying data-driven ML models using **AWS SageMaker**, enhancing real time prediction capabilities and reducing system response time by **20%** improving user experience
- Conducted statistical analysis using **R** and **NLP (Natural Language Processing)** to analyse customer behaviour data, enhancing marketing campaign effectiveness by **15%** and indirectly boosting customer service and engagement
- Worked with cross-functional teams to integrate **machine learning** solutions into existing business processes and monitoring dashboards, increasing operational efficiency by **20%**

PROJECTS

Real-Time Stock Market Data Processing (Python, Apache Kafka, AWS EC2, S3, Glue, Athena)

- Developed a real-time data streaming architecture with **Apache Kafka** and **AWS EC2** for instant market analytics
- Configured **AWS S3** & **Glue** to store and catalog data, reducing retrieval times by **20%** for improved strategic analysis
- Optimized a **Python**-based pipeline to process up to **5GB** daily, enhancing reliability and analysis capabilities

Sales & Purchasing Data Warehouse Integration (AWS, Talend, SQL, ETL, dbt, Tableau, PowerBi, Snowflake)

- Enhanced query efficiency by **25%** with a **Talend**-built data warehouse using **SQL, PostgreSQL, MySQL, and Oracle**
- Developed interactive dashboards in **Looker, Tableau, & Power BI,** increasing data analysis & visualization by **15%**
- Improved data accuracy by **20%** & enhanced decision-making with SCD Type 1, associative tables, and outriggers

Scalable Data Pipeline with Kafka and Cassandra (Apache Airflow, Apache Kafka, Apache Spark, Cassandra, Docker)

- Built a data pipeline with **Airflow**, **Kafka**, and **Spark**, deployed using **Docker**, and stored processed data in **Cassandra**
- Automated data ingestion and task scheduling using **Airflow**, ensuring real-time processing and reliable data flow
- Enhanced data retrieval speed and fault tolerance with **Cassandra**, improving system performance and availability