Karan Desai

8572308071 | karan.desai909@gmail.com | linkedin.com/in/mrkarandesai | Boston MA

Experience

Cloud Software Engineer, *Trilagen* | Boston, MA *Jul 2023 - Present*

- Scripted 400TB data transfer from EBS volumes to S3-Glacier resulting in 98% storage cost savings.
- Migrated 2000-user On-Premises Active Directory with Domain Name to AWS Managed AD.
- Built AWS Glue-based ETL pipeline to transfer 5TB unstructured data from S3 to RDS by leveraging PySpark's EMR operations.
- Automated data extraction from 100k document-images using AWS Textract, stored in Azure SQL Server, saving USD 800k.
- Migrated on-premises SQL Server to RDS-Custom, boosting IO ops by 40% with minimal downtime.
- Implemented OAuth2 architecture using Cognito User Pool ensuring validation and authentication of API Gateway calls to the backend.
- Hosted Web-App on AWS Amplify utilizing AWS CodeCommit.
- Constructed cloud infrastructure (VPC, Subnets, EC2, ASG, ALB, NATs, Postgres DB) for seamless Web App migration to AWS.
- Streamlined deployment of Docker container on AWS ECS using CodeBuild and Puppet.

Data Science Co-op, *Tausight* | Boston, MA *Jan 2022 - Jul 2022*

- Crafted queries on GCP's BigQuery and created new schemas resulting in 10% data retrieval efficiency.
- Orchestrated automated batch execution of PySpark jobs on GCP's Dataproc clusters using Airflow DAGs with 90% time efficiency.
- Employed Bi-directional LSTMs for PHI Document Classification resulting in a 3% precision boost without compromising accuracy.
- Developed Statistical Model with 60% accuracy for Anomaly Detection of fraudulent IP addresses.

Data Scientist, *Quality Kiosk Technologies* | Mumbai, India *Aug 2019 - Dec 2020*

- Assembled production-grade data pipeline which consolidated 150M raw records from 30+ sources using Kafka and PySpark.
- Automated information retrieval analytics and from Log Data using NLP and unsupervised ML, achieving 80% time-saving.
- Trained a Supervised ML model on 30k data points to classify reviews into predefined classes with over 70% accuracy.
- Leveraged Deep Learning to detect and address functionality issues in banking app screenshots posted by users.
- Developed Customer Segmentation and Churn model based on spending patterns, increasing ROI and customer retention by 10%.

TA - Machine Learning, *Northeastern University* | Boston, MA *Sep 2021 - May 2023*

- Conducted Python workshops and trained over 250 students across five semesters at Northeastern University.
- Mentored over 100 groups on Machine Learning projects and conducted Python code reviews for their implementation.
- Guided students with Supervised & Unsupervised Machine Learning and Neural Networks concepts.

Certificates

Aug 2023 **AWS Certified Solutions Architect – Associate**
June 2023 **AWS Certified Cloud Practitioner**
Mar 2023 **Google TensorFlow Developer Certificate**
Aug 2019 **PGP in Data Science, Big Data, and Data Analytics in association with IBM**

Education

May 2023 **MS - Data Analytics Engineering** *Northeastern University* | Boston, MA

Skills

Programming	Python, R, PySpark, PyTorch, NumPy, Pandas, NLTK, Scikit-Learn
Data Engineering	MSSQL, MySQL, Apache Airflow, Directed Acyclic Graphs – DAG, Docker, Kubernetes
Visualization	Plotly, Matplotlib, Seaborn, GGplot, Tableau, AWS QuickSight, QlikSense
GCP	DataProc, Compute Engine, Cloud SQL, BigQuery, Cloud DataFlow
AWS	Redshift, DB Migration, Lambda, API Gateway, Fargate, Textract, IAM, RDS, Kinesis, VPC, Security Groups, Cognito

Academic Projects

Online News Popularity Prediction – PySpark *Apr 2023*

- Trained a Linear Regression model on a dataset of 400k instances with PySpark's MLlib
- Conducted parallelized EDA, Pre-Processing, Modelling, and Evaluation on Discovery Clusters with 30 nodes, resulting in a 40% reduction in execution time while maintaining an 86% accuracy rate

Criminal Activity Analysis - Tableau Dashboard link *Dec 2022*

- Developed 7 interconnected visualizations spanning demographic analysis, geographic mapping offering insights into the complex dynamics of crime and socioeconomic factors across communities

Image Analysis using Neural Nets *Dec 2021*

- Classified 60,000, 32x32 color images, divided into 100 classes, further grouped into 20 super classes
- Transfer Learning Learning proved to be the most efficient approach with 82% train and 78% test accuracy

Customer Satisfaction Based on Complaints *May 2019*

- Trained 4 Supervised Classification models with 100K data points to predict customer satisfaction for their grievances
- Recruited Logistic Regression as final model for the task with below-par accuracy but best recall rate of 90%