KARAN PRATAP SINGH

+1(469) 332-2869 • karan.corporate25@gmail.com • LinkedIn • Github • Dallas, TX

EDUCATION

Master of Science (M.S.) - Business Analytics Aug 2022 - May 2024
The University of Texas at Dallas, Texas, TX

Bachelor of Technology (B.Tech.) - Computer Science Aug 2014 - Jun 2018
Dr. A. P. J. Abdul Kalam Technical University, Ghaziabad, U.P

SKILLS

Programming Languages:	SQL, Python, R, SAS, SSIS, Linux, SQL Server, Oracle, Mysql, Nosql.
Libraries:	Numpy, Pandas, Dplyr, Scikit-learn
Cloud Platforms:	AWS S3, EC2, Redshift, QuickSight, Lambda, AWS Athena, ADLS Azure Data Factory, Azure Databricks, Snowflake
Big Data Technologies:	Hive, Hadoop, HDFS, Splunk, Apache Spark, Apache Flume, Apache Kafka, HBase.
Data Visualization /Analysis:	Alteryx, Power BI, Visio, Cognos, Tableau, MS Excel.
Version Control:	Git, Jenkins, Docker.
Statistics & AML:	Hypothesis Testing, Logistic Regression, ANOVA, K-Means, KNN, Random Forest.
Frameworks:	Data Warehousing, Data Modeling, REST API, CI/CD, Agile Methodology.
Certifications:	AWS Solutions Architect Associate, Azure Data Engineer Associate, Google Analytics.

EXPERIENCE

Senior Software Engineer Apr 2021 - Jul 2022
Qualitest *Noida, India*

- Implemented AWS Glue PySpark pipelines for Sales and forecast data, achieving 30% faster extraction, aggregation, and consolidation and automated Lambda functions for S3 events, reducing deployment time by 50%.
- Executed hundreds of PySpark jobs in AWS Glue and EMR to execute data transformations based on STMs, orchestrated via Apache Airflow to automate workflow sequencing, adding business value and reducing data processing time by 30%.
- Developed and tested ETL data pipeline for data retrieval to extract sales data from Hive and send to target vendor for UPC forecast sales analysis, achieving a strong data architecture and a 20% increase in data processing.
- Optimized a data pipeline solution design by migrating from Hive to Apache Spark and performed debugging of migration algorithms based on key performance indicators, resulting in 25% faster processing and improving system architecture.
- Implemented ETL pipeline leveraging AWS Lambda, AWS Glue, and Amazon Athena to analyze partnered credit sales trends and customer segmentation, leading to a 30% reduction in query response times for proactive decision-making.

Software Engineer Jun 2018 - May 2021
QA Infotech *Noida, India*

- Established a batch processing pipeline using PySpark to identify the replenishment data based on the purchase order transactions across all the channels and performed root cause analysis for management reporting, resulting in a 30% improvement in real-time insights.
- Scripted Boto3 to access AWS endpoints, generating metadata tables for service usage to fuel Quicksight dashboards. Enabled insights on performance, SLA targets, infrastructure code, data lineage, and data observability.
- Designed and developed a comprehensive solution architecture utilizing Azure Databricks and matched the ingested data with the Machine learning models before pushing it to Azure data lake storage, resulting in a 25% increase in data processing efficiency.
- Performed the Testing of the ETL pipeline and the changes in data models based on the new feature guidelines on JIRA.

Data Analyst Intern Jun 2017 - Nov 2017
Nirwani Technologies Pvt. Ltd., India *Delhi, India*

- Created SQL queries and provided database solutions to the R&D team from customer usage reports, improving data quality business objectives by 15% and enhancing team collaboration.
- Automated failure handling with a scripting language like Python and created scripts for data flow and data collection decreasing manual intervention by 12%.
- Increased Enterprise Data Warehouse flexibility by 70% and reduced the cost of re-engineering by 20%

PROJECTS

- **Conagra Hackathon (FMCG Data):** Led analysis of present FMCG market trends, providing growth recommendations for Conagra Brands. Identified opportunities in this segment using XGBoost and Random Forest , revealed a previously untapped market of $67 million, enhanced the product portfolio, and recommended effective cost-effective merchandising strategies. The analysis prompted the introduction of an additional product for substantial weekly sales enhancement
- **Stock Market Analysis** Implemented a full-stack real-time stock market analytics pipeline with stream ingestion with Alpha Vantage API, leveraging Apache Flume, Python, HDFS, and Spark RDDs for 5 stocks. (
- **Analyzing Powerlifting Dataset:** Utilized data analysis to evaluate lifts' impact on powerlifting outcomes, deploying Machine learning models: linear regression, decision trees, random forests, SVMs, and neural networks, for insights. Showcased expertise in data analytics, stats, RDBMS, and R programming.
- **Cab Fares Data Analysis :** Examined the data for Uber and Lyft rides in Boston to understand what factors affect cab prices using statistical modeling, business intelligence, and data quality management.
- **Truck Driver Hazard Identification :** Conducted comprehensive analysis of truck fleet data using Impala, resulting in precise transportation forecasts and the extraction of actionable insights to optimize operations. Employed Tableau for advanced data visualization flexibility, enabling the identification of patterns and the application of statistical innovation methods to identify high-risk truck drivers, thus significantly improving road safety measures