

KARAN PRATAP SINGH

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

EDUCATION

Master of Science (M.S.) - Business Analytics

The University of Texas at Dallas, Texas, TX

Aug 2022 - May 2024

Bachelor of Technology (B.Tech.) - Computer Science

Dr. A. P. J. Abdul Kalam Technical University, Ghaziabad, U.P

Aug 2014 - Jun 2018

SKILLS

Programming Languages:

SQL, Python, R, SAS, SSIS, Linux Shell Scripting, Postgres, Mysql, Nosql.

Libraries:

Numpy, Pandas, Dplyr, Scikit-learn, Matplotlib, Ggplot2, Beautiful Soup.

Cloud Platforms:

AWS S3, EC2, ELB, EBS, RDS, VPC, AWS Redshift, QuickSight, Lambda, AWS Athena, ADLS, GCP, Azure Data Factory, Azure Databricks, Snowflake, Informatica

Big Data Technologies:

Hive, Hadoop, HDFS, Splunk, Apache Spark, Apache Flume, Apache Kafka, HBase.

Data Visualization / Analysis:

Alteryx, Power BI, Visio, 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, CICD, Agile Methodology.

Certifications:

AWS Solutions Architect Associate, Azure Data Engineer Associate, Google Analytics.

EXPERIENCE

Senior Software Engineer

Apr 2021 - Jul 2022

Qualitest

Noida, India

- Executed an end-to-end data pipeline, employing PySpark various workflows, data structures and used Airflow to streamline data, 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 credit sales analysis, achieving a 20% increase in resiliency, security and data governance.
- Optimized a data pipeline solution design by migrating from Hive QL to Apache Spark and performed debugging of migration algorithms resulting in 25% faster processing and improving system architecture.
- Implemented a real-time data analysis pipeline leveraging AWS Lambda for event-triggered processing of customer purchase data in Amazon S3, AWS Glue for data cataloging, preparation, and transformation, and Amazon Athena for interactive querying, enabling timely insights into sales trends, customer behavior, and product innovation performance.
- Implemented and automated data pipelines and data stores in Snowflake on prototype emerging business use cases, created test plans and performed quality assurance testing, increasing performance by 15%, and reducing ETL maintenance efforts by 20%.

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.
- Designed a self-serve platform using Python to perform data ingestion of 8 TB into Snowflake database. Utilized Snow SQL to visualize comprehensive data lineage across all platforms, improving data transparency, application design and reducing troubleshooting time by 8%.
- Designed and developed a comprehensive solution architecture utilizing Azure Databricks, Azure Data Lake and Azure Data Factory for distributed computing and application development, resulting in a 25% increase in data processing efficiency for proof of concept project.
- Leveraged data analysis techniques to uncover valuable business insights and informed strategic decisions. Provided production support through meticulous data maintenance, data management and effective issue management.
- Monitored and optimized the performance and scalability of data warehouse queries in GCP, identifying and addressing bottlenecks or inefficiencies in collaboration with the performance team mentorship and investigations, resulting in 15% improvement in query performance, coding and scalability.

Data Analyst Intern

Jun 2017 - Nov 2017

Nirwani Technologies Pvt. Ltd., India

Delhi, India

- Created SQL queries and provided data solutions to R&D team from customer usage reports, improving data quality and reducing final report generation time by 5%.
- Automated job failure handling with scripting language like Python and created scripts for data flow, internal systems design, analysis and data collection, decreasing manual intervention by 12%.
- Built tableau dashboard from business requirements for least sales per use cases, providing data solutions, enhancing decision-making and reducing analysis time 8%.
- Performed data wrangling and ensured data quality controls based on mentioned use case by validating and analyzing data, leading to 15% reduction in errors, and did multi-tasking by collaborating with R&D team leadership for data reconciliation, improving accuracy by 10%.
- Curated insightful data driven tableau reports derived from thorough data analysis, business intelligence, data management and decision making. Ensured data integrity and data management through timely maintenance and proficient issue management in production environments.

PROJECTS

- **Conagra Hackathon (FMCG Data):** Led analysis of present FMCG market trends, providing growth recommendations for Conagra Brands. Identified opportunities in this segment using emerging technologies like Alteryx and revealed a previously untapped market of \$67 million, enhanced the product portfolio, and recommended effective cost-effective merchandising strategies. The Alteryx analysis prompted the introduction of an additional product for substantial weekly sales enhancement
- **YouTube Analysis Project:** Extracted JSON and CSV files using AWS CLI, transformed them into tables via AWS Glue, and conducted table joins in Athena with Hive queries. Preprocessed the data innovation using AWS Lambda for data analytics using problem-solving, technical skills and analytical skills.
- **Analyzing Powerlifting Dataset:** Used data analysis techniques and assessed the impact of different lifts on powerlifting competition outcomes. The project included machine learning predictive models committed to synthesize data and highlighted the expertise in data analytics, statistics, rdbms and R.
- **Cab Fares Data Analysis :** Examined the data for Uber and Lyft rides in Boston to understand what factors affect cab prices using statistical modelling, 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, enabling the identification of patterns and the application of statistical innovation methods to identify high-risk truck drivers, thus significantly improving road safety measures