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, Linux Shell Scripting, NoSQL, Oracle PL/SQL.
Libraries:	Numpy, Pandas, Dplyr, Scikit-learn, Matplotlib, Ggplot2, Beautiful Soup.
Cloud Platforms:	AWS S3, AWS EC2, AWS Redshift, AWS QuickSight, AWS Lambda, AWS Athena. Azure Data Lake Storage, Azure Data Factory, Azure Databricks, Snowflake
Big Data Technologies:	Hive, Hadoop, HDFS, Apache Spark, Apache Flume,Apache Kafka.
Data Analysis:	Power BI, Tableau, MS Excel.
Version Control:	Git, Jenkins, Docker.
Statistics & ML:	Hypothesis Testing, Logistic Regression, ANOVA, K-Means, KNN, Random Forest.
Frameworks:	Data Warehousing, Data Modeling, REST API, CI/CD.
Certifications:	AWS Solutions Architect Associate, Azure Data Engineer Associate, Google Analytics.

EXPERIENCE

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

- Executed an end-to-end data pipeline, employing PySpark and Airflow to streamline data and perform ETL, 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 efficiency.
- Optimized a data pipeline solution by migrating from Hive QL to Apache Spark resulting in 25% faster processing
- Implemented and automated data pipelines and data stores in Snowflake, increased performance by 15%, and reduced ETL maintenance efforts by 20%.

Data 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 to provide an omnichannel 360 view for a customer using Janus Graph, 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 and reducing troubleshooting time by 8%.
- Designed and developed a comprehensive solution architecture utilizing Azure Databricks, Azure Data Lake and Azure Data Factory, resulting in a 25% increase in data processing efficiency.
- Monitored and optimized the performance and scalability of data warehouse queries, identifying and addressing bottlenecks or inefficiencies in collaboration with the performance tuning team, resulting in 15% improvement in query performance and scalability

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

- Created SQL queries for customer usage reports, reducing report generation time by 5%.
- Automated job failure handling with Python scripts, decreasing manual intervention by 12%.
- Built power bi dashboard for least replenished goods, enhancing decision-making and reducing analysis time by 8%.
- Ensured data quality by validating inconsistent data, leading to 15% reduction in errors, and collaborated with engineering for data reconciliation, improving accuracy by 10%.

PROJECTS

- **Conagra Hackathon (FMCG Data):** Led analysis of FMCG market trends, providing growth recommendations for Conagra Brands. Identified opportunities in this segment and revealed a previously untapped market of $67 million, enhanced the product portfolio, and recommended effective merchandising strategies. The 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 using AWS Lambda.

- **Analyzing Powerlifting Dataset:** Used data analysis techniques and assessed the impact of different lifts on powerlifting competition outcomes. The project included predictive models and highlighted the expertise in data analytics, statistics, and R programming.

- **Cab Fares Data Analysis :** Examined the data for Uber and Lyft rides in Boston to understand what factors affect cab prices.