Rohan Katari

Mobile: 9803710137 | email: katarirohan99@gmail.com | LinkedIn: linkedin.com/in/rohankatari | GitHub: rohan818

PROFESSIONAL SUMMARY

- Accomplished **Data Engineer** with **over 4 years** of experience specializing in the design, implementation, and maintenance of scalable **data pipelines** and **architectures**.
- Demonstrated proficiency in utilizing a wide array of **programming** languages including **Python, SQL, PL/SQL, Scala, Java,** and **Bash** delivering robust data solutions.
- Extensive knowledge of database technologies, with expertise in managing and optimizing databases such **as Microsoft SQL Server, MySQL, PostgreSQL, Oracle DB, Elasticsearch, MongoDB** and various other **NoSQL** solutions. Adept in ensuring data integrity, availability, and performance.
- Skilled in **ETL, Data Pipeline and Data Warehousing** tools to streamline data extraction, transformation, and loading processes. Excel at automating and optimizing data flows to support **business intelligence** and **analytics** initiatives.
- Proficient in **Big Data Frameworks**, with a strong command over **Apache Spark** and **Hadoop** ecosystem to process large datasets efficiently and developing scalable big data processing applications to meet evolving data analysis needs.
- Expertise in **Data Analysis and Visualization**, utilizing tools including **Kibana, Power BI,** and **Tableau** to derive actionable insights and create interactive dashboards and reports from complex datasets.
- **Certified Associate in AWS and Microsoft Azure**, proficient in deploying, managing, and optimizing cloud-based data solutions using AWS and Azure services, ensuring scalable, secure, and cost-effective data infrastructure.
- Versatile skill set including **Web Development** and **DevOps** with experience in application development, containerization and testing. Proficient in deploying and managing applications in **Linux** and **Windows** environments.

TECHNICAL SKILLS

Programming	Python, SQL, PL/SQL, T-SQL, Scala, Java, Bash
Database Technologies	Microsoft SQL Server, MySQL, PostgreSQL, Oracle DB, Elasticsearch, MongoDB, Teradata, NoSQL
ETL and Data Pipeline Tools	Apache Airflow, Apache NiFi, Apache Kafka, Informatica ETL Tools, IBM DataStage, Oracle Data Integrator
Big Data Frameworks	Apache Hadoop (HDFS, MapReduce, YARN, Hive), Apache Spark (Spark SQL, PySpark, Spark Streaming)
Data Analysis and Visualization	Pandas, NumPy, Kibana, Power BI, Tableau
Cloud Platforms and Services	**AWS (Amazon Web Services):** IAM, EC2, S3, Lambda, VPC, RDS, DynamoDB, Redshift, Elastic MapReduce, Kinesis, Glue, Athena, CloudFormation **Microsoft Azure:** Data Factory, Synapse Analytics, SQL Database, Data Lake Storage, Databricks, Stream Analytics, Blob Storage
Web Development	Django Application Development, JavaScript, HTML5, CSS, REST API, OpenAPI Specification, Swagger, D3.js, Python Dash
DevOps and Other Skills	Docker, Kubernetes, Git, Terraform, Snowflake, Data Warehousing, Data Lakes, Linux, Microsoft Windows

CERTIFICATIONS

- **Microsoft Certified Azure Data Engineer Associate, Jan 2024-Jan 2025**, Credential ID- D73E54002CB2CD0E (link)
- **AWS Certified Developer – Associate, Dec 2023 – Dec 2026**, Validation number- QDWBT0RDYMQQ1ZGQ (link)

WORK EXPERIENCE

Data Engineer   **Sep 2022 – Dec 2023**
University of North Carolina at Charlotte   Charlotte, NC, USA

- Worked on creating a **Django** web application for research administration staff and faculty to upload and view grant-related documents, enhancing restricted user accessibility and document management with **AWS S3** storage.
- Developed a data pipeline using **Apache Airflow** orchestrating PDF parsing, streaming data and loading into AWS RedShift for data analysis, reducing manual data extraction efforts by 45%.
- Utilized **Apache Kafka** in the data pipeline to efficiently handle streams of data extracted from PDFs, enabling scalable and fault-tolerant data transfer between services.

- Optimized **AWS RedShift OLAP** performance and table storage by tuning queries, vacuuming tables and adjusting data distribution styles for faster load times and query execution for large datasets.

Software Engineer   **Jun 2021 – Jun 2022**
Tata Consultancy Services (TCS)   Hyderabad, India

- Implemented a data pipeline for **massive** Call Detail Records (CDRs) **data extraction** and **ingestion**, enhancing reliability with **Slack** notifications for immediate pipeline failure alerts.
- Built **Apache Airflow DAGs** for CDR data preprocessing and billing activities, reducing processing time by 30% and ensuring efficient monthly billing.
- Developed **Spark SQL** scripts to perform complex data aggregations, joins, and analytics on consumer subscription **big datasets**, preparing data for insightful reporting and ad-hoc analysis in **Azure Synapse Analytics**.
- Utilized **Oracle Data Integrator (ODI)** to design and populate databases and data warehouses and that are optimized for analyzing telecommunications data, supporting BI and reporting tools.
- Configured **Kubernetes** for orchestrating containerized application deployments, ensuring high availability, load balancing, and automated scaling for data processing tasks.

Data Engineer   **Jan 2020 – May 2021**
Dr. BR Ambedkar Open University   Hyderabad, India
- Implemented an **ETL** data pipeline solution leveraging **Apache NiFi** and **Python** to automate the extraction of new student enrollment data from multiple affiliated institutes via **REST APIs**, ensuring efficient data ingestion into **Elasticsearch**.
- Created comprehensive **Kibana** dashboards to visualize enrollment and student course selection trends, providing key insights aiding strategic decision-making.
- Enhanced query performance on historical data by implementing data partitioning and bucketing strategies on **Hive** tables stored in **HDFS**, facilitating faster insights for analysis and reporting.
- Developed and maintained **Power BI** dashboards and reports to track and analyze **KPIs** across departments, utilizing **DAX** queries for complex **data modeling**.

EDUCATION

Master of Science, Computer Science   **Aug 2022- Dec 2023**
University of North Carolina at Charlotte, NC, USA
Coursework: Cloud Computing for Data Analysis, Big Data Analytics for Competitive Advantage, Software Systems Design and Implementation, Network-Based Application Development

Bachelor of Technology, Computer Science and Engineering   **Aug 2017- Jun 2021**
JNTU, Hyderabad, India
Coursework: Data Structures and Algorithms, Database Systems, Data Warehousing and Data Mining, Cloud Application Development, Information Security

PROJECTS

SeeClickFix API Data Pipeline | GitHub
- Implemented a robust **ETL data pipeline** leveraging **Apache NiFi** and **Python**, to ingest current and historical records 311 service request data from the SeeClickFix API, for a specified county. Utilized Elasticsearch for efficient data storage and indexing, enabling advanced **data visualization** through custom Kibana dashboards.

e-Learning Platform using Django | GitHub
- Developed a **Django**-based e-learning platform featuring a dynamic Course Management System (CMS), flexible content models, and a robust student registration system. Incorporated Django Channels for real-time chat, a **RESTful API**, and utilized **Docker** for deployment, enhancing user engagement and streamlining course creation and enrollment.

Newsfeed Data Pipeline | GitHub
- Created an application to aggregate the latest news from various RSS Feeds and make them accessible to users via a convenient API. Leveraging **Apache Airflow, Kafka**, **Kafka-Connect** Connectors, **MongoDB**, **MinIO**, **Elasticsearch** and **Kibana**, the data pipeline orchestrates periodic **web scraping** tasks to gather, process, store and visualize news data.