PHANI RAJESH KULKARNI

📞 (913)-413-7115, ✉️ kulkarniphani@gmail.com, 🔗 LinkedIn, ⬡ Github, Leetcode

Results-driven Data Engineer with 3+ years of experience in Big Data, ETL, SQL, and Python. Proven expertise in Azure and AWS cloud platforms, along with a strong background in developing and optimizing ETL pipelines, data migration, and implementing CI/CD processes. Skilled in Spark, Python, Docker and various data management tools. Experienced in Agile methodologies and am eager to leverage my technical skills and professional experience to drive innovative solutions and contribute to the success of a forward-thinking team that values innovation and excels in niche technologies.

EDUCATION

University of Missouri Kansas City, Missouri   **January 2023 – May 2024**
Master of Science in Computer Science, specialization in Data Science   *GPA: 3.73 / 4.00*
Course Highlights: Big Data, Data Science, Database, Design and Analysis of Algorithms, Cloud Computing, NLP

Savitribai Phule Pune University, India   **August 2016 – May 2020**
Bachelor of Technology-BTech, in Computer Engineering   *GPA: 8.86 / 10.00*
Course Highlights: Data Structures, Object-Oriented Programming, BigData, Machine Learning

TECHNICAL SKILLS

Languages/Library/Frameworks: Python, JavaScript, R Programming, React

Databases/Datawarehouse: PostgreSQL, MySQL, Snowflake

DevOps/Cloud/ETL/Big Data Stack: Hadoop, Spark, Kafka, Azure Data Factory, Data Lake, Databricks, ETL, AWS-SQS, AWS Elastic Beanstalk, Git, Docker, Apache Nifi, Apache Airflow

PROFESSIONAL EXPERIENCE

BigData Developer - Tata Consultancy Services, Hyderabad, India   *June 2022 – December 2022*
- Ingested health insurance data from a traditional Teradata system into Hive tables using Sqoop.
- Employed various data loading techniques, primarily using incremental loading to ensure data consistency and efficiency.
- Designed and created multiple databases and tables in Hive to structure and organize ingested data effectively.
- Implemented transformations based on inbound case logic to create verification columns, determining patient insurance eligibility using PySpark
- Ensured data transformations adhered to business rules and improved data quality for downstream analysis.
- Coordinated with cross-functional teams to ensure data accuracy and readiness for analysis.
- Technologies/Languages/Tools: BigData, Sqoop, Hive, PySpark, Databricks

Data Engineer - Tata Consultancy Services, Hyderabad, India   *August 2020 – May 2022*
- Extracted, transformed, and loaded data from multiple sources into the backend SQL server database using Python Scripts, streamlining the data integration process and reducing manual data entry by 80%.
- Led the migration of on-premise SQL servers to Azure SQL, enhancing scalability and reducing maintenance overhead.
- Collaborated effectively with downstream teams to procure data from APIs, gaining a solid understanding of API Gateway functionalities
- Led the initiative to develop and deploy an Azure DevOps pipeline, facilitating continuous integration and deployment for the ETL process, resulting in a 30% increase in overall operational efficiency.
- Technologies/Languages/Tools: SQL, Python, React, Eclipse IDE, SQL, Apache-Spark, Azure Essentials, Azure Services(ADF, Databricks, ADLS, Synapse)

Data Engineer Intern – Speed Techserve Pvt. Ltd, Pune, India   *May 2019 – June 2020*
- Developed and managed Python-based pipelines for data loads from diverse sources including S3, MySQL, and Oracle. Extracted data using APIs and efficiently loaded it into S3, incorporating triggers for automation.
- Optimized data flow from various sources, ensuring seamless integration and data accuracy.
- Built insightful dashboards using Power BI to provide actionable business intelligence and support data-driven decision-making.
- Technologies/Languages/Tools: Amazon S3, SQL, Python, PowerBI

PROJECTS

Hire Human | Hackathon, Microsoft
June 2024
- Developed an application for real-time interview capture and analysis using advanced NLP algorithms, including the WaterMark method, achieving industry-level accuracy in response classification.
- Implemented frontend components using React for audio capture and encoding, and integrated AssemblyAI for precise voice diarization
- Designed a Flask backend integrating OpenAI for interviewee identification among multiple speakers and classification of responses as LLM-generated, human, or mixed using Azure OpenAI services.
- Technologies/Languages/Tools: Azure OpenAI, Cosmos DB, Flask, LLM, React

Dockerized Data Processing Pipeline with AWS SQS and PostgreSQL | Fetch
- Designed and implemented a Dockerized solution to read and process JSON data from AWS SQS, ensuring secure data handling by masking PII with SHA-256 hashing before storing it in PostgreSQL.
- Created a robust ETL pipeline in Python to automate the data flow from AWS SQS to PostgreSQL, including table creation, data parsing, and secure storage.
- Developed comprehensive Docker and Docker Compose configurations to streamline the deployment and orchestration of multiple services, including LocalStack for SQS emulation and PostgreSQL
- Technologies/Languages/Tools: Docker, Python, Amazon Simple Queue Service

Spotify Data Insights Pipeline
May 2024
- Orchestrated the ingestion of Spotify data in JSON format by uploading it to the S3 raw data repository.
- Configured automated daily triggers using Amazon CloudWatch to streamline the data extraction process.
- Processed raw JSON data to extract key information. Automated the transformation process with triggers to ensure real-time data processing upon availability in the raw folder.
- Created Snowflake tables to store the transformed data. Utilized Snowpipe for continuous data loading from S3 to Snowflake tables.
- Conducted data analysis on the transformed data to derive insights.
- Technologies/Languages/Tools: Python, PostgreSQL, AWS(S3, CloudWatch, Lambda), Snow-pipe, Snowflake

Olympics Data Analysis | University of Missouri Kansas City, Missouri
November 2023 – December 2023
- Effectively designed and executed an ETL (Extract, Transform, Load) process through Azure Data Factory.
- Designed container structure within the storage account for organizing raw and transformed data files.
- Established Databricks Workspace and utilized Apache Spark for data transformations, including null value checks and schema inference. Loaded transformed data back into Azure Data Lake Storage Gen2 for downstream analysis.
- Implemented data transformations with Azure Databricks to optimize data accuracy and prepare it for advanced analytics.
- Leveraged Azure Synapse Analytics to adeptly perform complex SQL queries, uncovering noteworthy trends and patterns.
- Technologies/Languages/Tools: Python, SQL, Azure Data Factory, Azure Synapse, and Databricks

Student Chatbot | University of Missouri Kansas City, Missouri
October 2023 – November 2023
- Conducted data preprocessing using NumPy, and NLTK on a CSV dataset with user queries and responses. Designed and implemented a user-friendly chat interface in HTML.
- Applied TFIDF and cosine similarity techniques for effective response matching.
- Integrated the chatbot into a web application using Python Flask and deployed on the Cloud using AWS.
- Technologies/Languages/Tools: Python, Natural Language Processing, Machine Learning

Data Query Optimization | University of Missouri Kansas City, Missouri
June 2023 – July 2023
- Conducted a comprehensive comparative analysis of query time using Spark SQL engine on CSV and Parquet file formats; identified Parquet as the optimal format, reducing query response time by 50%.
- Technologies/Languages/Tools: Hadoop, Spark

PUBLICATION

- Published a paper "Divyahast Gesture Vocalization: A Case Study" in The International journal of analytical and experimental modal analysis, ISSN NO:0886-9367,Volume XII, Issue VI, Page No:896 - June 2020.

- Published a paper "Turbulent Flow of Water-Based Optimization for WMSNs-Based Image Transmission ModelMultimedia", Resbee Publishers, Multimedia Research (MR), https://doi.org/10.46253/j.mr.v4i2.a3, Vol.4 No.2 2021