YASH RAYTHATHA

yashbraythatha@gmail.com | (973)380-630 | LinkedIn | New York, NY

PROFESSIONAL SUMMARY

Experienced Data Engineer with expertise in designing efficient ETL pipelines, automating data workflows, and optimizing query performance to drive business insights. Proficient in Python, SQL, and AWS technologies, with a strong focus on building scalable, robust data solutions. Skilled in documenting data infrastructures to enhance collaboration and knowledge sharing. Driven to apply technical expertise in innovative projects, delivering actionable, data-driven solutions that empower forward-thinking organizations to achieve their goals.

SKILLS

Languages: Python, C/C++, Unix, HTML,CSS.
Data Pipelines and big data engineering: Pyspark, Spark SQL
Data Visualization: Tableau, Matplotlib, Seaborn
Database: SQL, Snowflake, MongoDB, Oracle
Certifications: Azure Administrator (AZ-104), Power BI Analyst Associate (PL-300), Oracle SQL Database Associate (1ZO-071)

EXPERIENCE

Data Engineer Aug 2019 - Mar 2023
HCL Technologies, Nagpur, India

- Collaborated with cross-functional teams to collect, clean, and preprocess diverse datasets, ensuring data quality and usability using Spark SQL for efficient queries.
- Led the development of efficient ETL pipelines, contributing to a 20% improvement in data processing speed.
- Automated routine tasks using Python scripts, increasing team productivity, and reducing manual errors using AWS service for improved workflow.
- Optimized spark jobs to reduce the query run time from 2 hours to 10 Minutes.
- Contributed to the documentation of data sources, methodologies, and code, ensuring comprehensive knowledge sharing within the team.

Application Support Engineer Nov 2018 - Aug 2019
Savantis Solutions, Hyderabad, India

- Improved application performance by identifying and resolving bottlenecks.
- Streamlined incident management processes, reducing downtime, and improving customer satisfaction.
- Collaborated cross-functionally to develop, test, and deploy new features and improvements.
- Conducted root cause analysis for application failures, implementing preventive measures.

PROJECTS

Twitter Data Analysis Using Apache Airflow

- Developed and managed a robust Apache Airflow ETL pipeline on AWS EC2 to automate the extraction, transformation, and loading of Twitter data from S3.
- Enhanced data integrity and streamlined processing by optimizing transformation and storage operations, resulting in improved efficiency.
- Leveraged Amazon QuickSight to create dynamic visualizations of the transformed data, facilitating advanced analytics and informed decision-making.

Biomedical Named Entity Recognition (NER) using Transformer Models

- Conducted a comparative analysis of transformer-based models (BioBERT, BERT) combined with BiLSTM and CRF for Biomedical NER on the BC5CDR dataset.
- Implemented advanced pre-processing and fine-tuning strategies to achieve the highest F1-score of 90% using BioBERT with CRF.
- Evaluated models on key metrics such as precision, recall, and F1-score, demonstrating robust entity recognition for diseases and chemicals.
- Highlighted the importance of domain-specific pretraining and structured sequence modeling for handling complex biomedical texts.

Campus Image Semantic Segmentation Using DeepLabV3

- Implemented a semantic segmentation model using DeepLabV3 with ResNet-50 backbone to classify campus images into doors, stairs, and background.
- Created and annotated a custom dataset of 200 images with pixel-wise precision using Roboflow.
- Achieved 63% mIoU and 83% pixel-wise accuracy through transfer learning, data augmentation, and hyperparameter tuning.
- Visualized segmentation results to validate model performance.

EDUCATION

Bachelor of Engineering: Computer Technology May 2017
Rajiv Gandhi College of Engineering Research & Technology - Chandrapur, Maharashtra, India

Master of Science: Data Science May 2025
University of New Haven, West Haven, CT