Ankita Sanas

1 Years as Data Engineer at Tech Mahindra

CONTACT

9022865174

sanasankita04@gmail.com

India, Pune, 900 Sadashiv Peth

EDUCATION

2021 - 2025

Bachelor of Science (BS)

Yashoda Techical Campus,Satara, India, Satara

HOBBIES

- Dancing
- Traveling

LANGUAGES

- Marathi,Hindi,English

PERSONAL INFO

- Date of birth: 4 December 2003
- Place of birth: Koregaon
- Nationality: Indian

PROFESSIONAL SUMMARY

1 Years as Data Engineer at Tech Mahindra From Dec 2024 to Till Date

Data Engineer with 1+ years of experience in Data Engineering, Big Data technologies and Data Analysis, . Proficient in Python, PySpark, Databricks SQL, Hadoop, Hive , Azure Cloud with expertise in building and managing scalable data pipelines using Azure Data Factory. Skilled in Data Mining, Data Preparation, Data modeling and ETL processes, ensuring high-quality data flow for analytical and business needs. Experienced in handling large datasets, implementing Machine Learning algorithms, and working with cloud platforms for data storage and processing. Strong knowledge of Proof of Concepts (PoC) and gap analysis to drive data-driven solutions for enterprise applications.

EXPERIENCE

Data Engineer 2024 - Now

Kizer, United States

Kizer Project – Automated Medical Transcription Pipeline

Domain: Healthcare

Tech Stack: Python 3.11, Whisper API, AWS Lambda, Amazon S3, AWS Glue, Amazon Athena, Terraform, AWS CloudWatch, Docker, Nginx

Problem Statement:

Healthcare providers manage a massive volume of **patient–doctor audio conversations** containing critical diagnostic and treatment information. These recordings are usually **unstructured and unsearchable**, making them unsuitable for analytics. Manual transcription was **slow, error-prone, costly, and non-compliant**. The goal was to design a **secure, automated, and scalable transcription pipeline** to make medical conversations **searchable, queryable, and AI-ready**, while ensuring compliance, cost efficiency, and scalability for real-time healthcare analytics.

Goals:

- Automate transcription of medical audio using Whisper API.
- Store transcriptions securely in **AWS S3** with partitioning and metadata tagging.
- Enable structured querying and reporting via **AWS Athena**.
- Ensure scalability and cost efficiency using **serverless AWS Lambda**.
- Lay foundation for **future NLP/AI use cases** (summarization, symptom/diagnosis extraction).
- Implement **infrastructure automation with Terraform** and **monitoring with CloudWatch**.
- Enforce data security and compliance with **IAM roles and encryption**.

Key Achievements:

- Designed and deployed an **end-to-end audio-to-text transcription pipeline** using Whisper API and AWS services.
- Reduced **manual transcription effort by 90%**, accelerating healthcare documentation.
- Built a **serverless architecture** with AWS Lambda for scalability and cost-effectiveness.
- Delivered **structured, queryable data** via AWS Glue & Athena, improving diagnosis analytics accuracy.
- Established a **foundation for advanced NLP/AI models** (summarization, medical insights, sentiment analysis).
- Achieved **cost optimization** through S3 lifecycle policies and partitioned storage strategies.
- Enhanced **system reliability** with infrastructure-as-code (Terraform) and monitoring (CloudWatch).
- Strengthened **data security & complian**

SKILLS

Python
SQL
Big Data
Data Warehousing
Data Mining
ETL
Apache Hadoop
Apache Spark
NoSQL
Data Modeling
Cloud Computing
Data Visualization
Machine Learning
Data Analysis
Business Intelligence
AWS
Azure
Data Pipeline

Azure Data Factory