Harsh Ekambe

[**hekambe400@gmail.com**](mailto:hekambe400@gmail.com) **| 313-638-4053 |Lake Hiawatha, NJ**

**LinkedIn:** <https://www.linkedin.com/in/harsh-ekambe-09b076210/>

# Career Summary

* Experienced Data Engineer with 8 years of expertise in designing, implementing, and optimizing scalable cloud-based data solutions across healthcare and various industries.
* Proficient in leveraging AWS services including Glue, Redshift, Athena, and Lambda to streamline ETL pipelines, improve data accessibility, and reduce processing times by up to 50%.
* Skilled in integrating large-scale data processing frameworks like Apache Airflow and Kafka to optimize workflows, ensuring performance, cost-efficiency, and scalability.
* Strong experience in machine learning model development, utilizing AWS SageMaker to deploy models that enhance predictive analytics and improve decision-making in compliance with industry standards.
* Proficient in data visualization using Tableau, enabling stakeholders to make informed decisions by creating interactive dashboards and reports that track data trends.
* Expertise in data wrangling techniques and building pipelines to prepare high-quality inputs for machine learning models, ensuring accuracy and reliability.
* Adept at collaborating with cross-functional teams to align data solutions with business goals, driving data-driven insights and enhancing organizational performance.

**Skills**

* **Languages & Tools:** Python, SQL, AWS Glue, Redshift, Lambda, S3, CloudFormation, IAM, VPC, Tableau, Git, ThunderClient, Git, R, C++
* **AWS Services:** Glue, Redshift, Athena, Lambda, S3, CloudFormation, IAM, VPC
* **Data Platforms & Tools:** Apache Airflow, Hadoop, Spark, AWS SageMaker, dbt, kafka, EPIC
* **Data Analysis and Engineering**: Pandas, Numpy, Scipy, Matplotlib, Keras, Tensorflow, PyTorch, NLTK, SciKit Learn, Apache Spark, Hadoop, Airflow, Kafka, dbt, Azure Databricks
* **Machine Learning**: Supervised and Unsupervised Algorithms, Deep Learning, GenAI, GAN, VAE, LLM, XGBoost, Computer Vision, NLP, autoencoders

**Education:**

* **Bachelor of Science in Mechanical Engineering,** Mumbai University, India.

**Professional Experience:**

**Michigan Health Information Network (MiHIN), East Lansing, MI || Jan,2024 – Present**

**Data Engineer**

# Responsibilities:

* **Designed and optimized ETL pipelines:** Leveraged AWS Glue and Redshift to process and transform large datasets, achieving a 50% improvement in processing time. Spearheaded the development of scalable ETL processes to handle complex data workflows.
* **Automated data workflows:** Implemented AWS Lambda and Step Functions to streamline data operations, reducing manual intervention by 40% and improving overall efficiency.
* **Managed data storage and retrieval:** Ensured secure and efficient data handling in S3, optimizing storage policies and retrieval processes for faster data access.
* **Data Integration with Informatica:** Employed Informatica PowerCenter to design and manage data integration workflows, ensuring seamless data transformation, loading, and validation processes across multiple systems, enhancing data accuracy and quality by 30%.
* **ETL pipeline**: Engineered and deployed ETL pipelines using Apache Airflow to extract complex HL7 FHIR data from over 15 FHIR resources of health data standards like CPT, ICD-10, OID and transforming it for comprehensive exploratory data analysis, elevating ETL process by 40%
* **ML Models**: Utilized AWS SageMaker for constructing and executing machine learning models, improving predictive analytics while adhering to HIPAA guidelines.
* **Data Integration with Healthcare Systems:** Integrated data from healthcare systems such as Epic, automating data extraction, transformation, and loading processes to support real-time analytics and improve decision-making across healthcare teams**.**
* **Data visualization:** Created data visualizations to track and analyze EPIC EMR data trends over time, using time series forecasting models such as ARIMA, ANOVA and exponential smoothing to enhance interpretation of healthcare data patterns and improve statistical analysis.

**Environment:** Python, SQL, AWS Glue, AWS SageMaker, Redshift, Lambda, S3, CloudFormation, Informatica, Step Functions

# Interoperability Institute, Ann Arbor, MI || Feb 2022 - Dec 2023

# Data Scientist

**Responsibilities:**

* **Synthetic FHIR Data Processing:** Collaborated on processing and cleaning synthetic FHIR data using Python and Jupyter Notebook, improving data quality by 20%. Developed robust data pipelines to manage and preprocess healthcare data.
* **Machine Learning:** Applied supervised and unsupervised machine learning algorithms like Linear Regression, KNN, Logistic Regression, Decision Trees, Random Forest, SVM, Clustering, PCA, Anomaly Detection to Synthetic FHIR data, achieving improvement in predictive analytics accuracy.
* **Data Visualization:** Created detailed visualizations of patient condition trends, aiding healthcare professionals in data-driven decision-making. Communicated complex analytical findings to both technical and non-technical audiences.
* **Time Series Forecasting:** Visualized patient, condition, observation FHIR resources data trends over time to utilize ARIMA model forecasting, enhancing data comprehension.
* **Collaborative Data Projects:** Worked closely with data scientists and domain experts to translate business requirements into technical solutions, ensuring alignment with organizational goals.
* **Data wrangling:** Employed data wrangling techniques to handle missing data, outliers, and data normalization, securing high-quality inputs for machine learning models.
* **Model Deployment and Monitoring**: Deployed machine learning models on AWS SageMaker, incorporating real-time monitoring and automated retraining pipelines to maintain model accuracy and adapt to evolving data patterns.
* **Data Integration**: Utilized Matillion for seamless data integration, enabling efficient ETL processes to extract, transform, and load data from various sources into cloud data warehouses, enhancing the scalability and flexibility of data-driven applications in healthcare.
* **Epic Data Extraction and Transformation:** Designed workflows to extract clinical and administrative data from Epic EMR, integrating it with AWS Redshift for advanced analytics, improving data access and reducing report generation times by 30%.
* **CQL**: Enforced Clinical Query Language (CQL) to standardize the expression of clinical logic following protocols like SFTP, PHINMS and S3.

**Environment:** Python, Jupyter Notebook, ARIMA, Machine Learning, FHIR, REST API, Supervised and Unsupervised ML algorithms, Matillion, Mirt, EPIC, CQL

# Exadatum, Pune, India || Jan 2020 – Dec 2021

# Senior Data Engineer

**Responsibilities:**

* **API and ETL Development:** Developed and maintained Python-based APIs and ETL jobs using PySpark on on-premises clusters, significantly enhancing data processing efficiency and reliability.
* **Event-Driven Data Processing:** Integrated AWS Glue jobs with Lambda to enable event-driven data workflows, reducing latency and ensuring timely data availability.
* **CI/CD Implementation:** Led the implementation of CI/CD pipelines using Jenkins and Airflow, which improved deployment speed by 30% and reduced system downtime during updates.
* **Informatica for Data Processing:** Utilized Informatica PowerCenter to fine-tune and optimize large-scale data processing workflows, improving data pipeline performance by 25% and reducing data processing times, ensuring faster delivery of actionable insights.
* **Data Security and Compliance**: Implemented data encryption and access control policies within AWS environments, ensuring compliance with industry standards and safeguarding sensitive information.
* **Performance Tuning**: Optimized ETL jobs and query performance on AWS Redshift, resulting in a significant reduction in query execution time, which enhanced the overall efficiency of data processing workflows.

**Environment:** Python, PySpark, AWS Glue, Lambda, Jenkins, Airflow, CI/CD, Docker, PostgreSQL, HDFS, Informatica

# Application Software Technology, Pune, India || Sep 2017 - Dec 2019

# Data Engineer

**Responsibilities:**

* **Backend Data Engineering:** Designed and implemented backend data access modules using SQL and PL/SQL and Oracle databases, improving data access efficiency and system performance.
* **Large-Scale Data Processing:** Utilized AWS services like EMR and Redshift to process and analyze large-scale datasets, optimizing workflows for both performance and cost-efficiency.
* **Automated Data Pipelines:** Developed automated data pipelines, reducing manual data processing tasks by 50% and ensuring consistent data availability across systems.
* **Data Warehouse Optimization:** Managed data warehouses, optimizing schemas and tuning performance to support large-scale data analytics and reporting.

**Environment:** Python, R, Hadoop, Spark, AWS EMR, Redshift, SQL, Oracle, PL/SQL, Tableau

# Rudder Analytics, Pune, India || Jun 2016 - Aug 2017

# Data Analyst

**Responsibilities:**

* **Data Modeling & Database Design:** Designed and implemented relational data models using MySQL Workbench. Developed complex stored procedures and triggers to maintain data integrity, enhancing database performance for key business operations.
* **Advanced Data Analysis**: Conducted in-depth analysis of large datasets using Python and Hadoop, delivering actionable insights that guided strategic business decisions and improved operational efficiency.
* **ETL & Data Warehousing:** Built efficient ETL pipelines and executed Hive queries to load and transform data into the warehouse, streamlining data storage and retrieval processes for large-scale analytics.
* **Data Integration:** Integrated data from various sources into a unified data warehouse, improving the accuracy, consistency, and accessibility of business reports across departments.
* **Performance Optimization:** Optimized Hive queries and refined data storage strategies, resulting in a 25% improvement in ETL process efficiency and significantly faster data retrieval for business users.
* **Data Governance:** Implemented data governance policies using Informatica Axon, ensuring compliance with regulatory standards and improving the traceability of data usage across the organization.
* **Data Quality Assurance**: Utilized Informatica Data Quality tools to automate data validation and cleansing processes, boosting data accuracy and reliability by 20%, which led to better business reporting and decision-making.

**Environment:** Python, PL/SQL, Oracle, SQL, Hive, Hadoop, ETL, Informatica

**Projects**: https://github.com/theharshbot

* **Emotion-Based Music Player:** Developed a deep learning-based music player that creates dynamic playlists based on emotional states, achieving 89% accuracy. The system utilizes facial recognition technology to detect emotions such as happiness, sadness, anger, or surprise from real-time video feeds. Once the emotion is identified, the player selects and curates a playlist that matches the user's current mood. The project achieved an 89% accuracy rate in emotion detection and playlist matching, significantly enhancing the user experience by providing emotionally resonant music recommendations.
* **GenAI for Business Analysis:** Fine-tuned GPT-3.5-turbo using OpenAI API to enhance its capabilities for analyzing customer complaints, improving response accuracy. The project involved training the model on a large dataset of customer interactions, enabling it to better understand and address specific business issues. As a result, the model demonstrated improved accuracy in identifying the root causes of customer complaints and providing actionable insights. This enhancement led to a notable increase in customer satisfaction by enabling more effective and empathetic communication with users.
* **HAPI-FHIR ETL Pipeline with Apache Airflow**: Implemented a robust ETL pipeline using Apache Airflow to automate the extraction, transformation, and loading of patient data from the HAPI FHIR server into a PostgreSQL database, enabling efficient processing of FHIR patient records. Developed a structured transformation process that converts raw JSON data into a normalized format, extracting critical patient attributes such as ID, name, gender, birth date, and address for enhanced analytics and reporting capabilities. Designed the pipeline to handle large datasets through pagination, ensuring efficient retrieval and processing of multiple FHIR bundles while dynamically creating the target PostgreSQL table when necessary for data integrity and organization.
* **Automated ETL Pipeline for EPL Football clubs** Developed an ETL pipeline using Apache Airflow to automate the extraction, transformation, and loading of Premier League football club data from Wikipedia into a PostgreSQL database, enhancing scalability and maintainability. Scraped and cleaned web data using requests, BeautifulSoup, and Pandas, ensuring accuracy and consistency of the data for further analysis and reporting. Integrated PostgreSQL for storing processed data, while leveraging Docker for containerizing the services, enabling seamless deployment and efficient orchestration of the pipeline.