Akanksh Gatla

716-808-2702 | akankshgatla@icloud.com | Akanksh — Linkedin | Akanksh — Github | Akanksh — Portfolio

Summary

High-skilled data professional with **5+** years of cloud experience with data mining, data modeling, and data transformation (**ETL/ELT**) on multi **Terabyte** at distributed platforms. Possess expertise in programming tools like **Python, R, Spark, SQL** and complete maintenance in **Software Development Life cycle (SDLC**) like **Agile/Waterfall**. Cloud-based solutions using **Azure** services, incorporating **CI/CD** pipelines and orchestration through automation tools like **Terraform**, **Azure Data Factory (ADF)**. Working on an **OLTP / OLAP** environment that includes production and development databases in **SQL Server**, leveraging **Databricks** for implementing reporting systems and analytics for healthcare clients. Deep understanding of **Big Data** tools on **Healthcare** Domain, like **Apache Hadoop, Spark, Solr, Hive, Kafka, and Airflow**.

Experience

Data Engineer February 2023 – Present

Company: Unity Population Health , Client: StarMed *Dublin, Ohio*

- Designing a **SQL Server** database migration to **Azure SQL**, leading to a **60%** reduction in operational costs due to cloud optimization.
- Spearheading **VBC** solution that enabled data-driven decisions, increasing revenue by **18%** through targeted marketing strategies.
- Building an automation system using **Python** to generate lists of non-compliant patients based on clinical, functional and operational checks defined by **Value-Based Care** (VBC), and display them on provider's Electronic Medical Records (EMR).
- Implementing a **Unified Data Systems** (UDS) reporting system using **Databricks** (**Spark SQL**) and **Power BI** to create reports for healthcare clients focusing on clinical compliance.
- Working on performance tuning of long-running Spark Jobs using DataFrames, **Spark-SQL**, memory tuning, **Spark YARN**
- Handling sensitive **Electronic Medical Record** (EMR) data using APIs, retrieve results in JSON format, and ensure secure handling and compliance with data privacy regulations.
- Successfully led projects (**Patient Engagement, Remote Patient Monitoring, UDS Reporting**) through the entire development cycle, driving down storage costs by **25%**, increasing customer satisfaction by **20%**, streamlining integration and profiling processes by **40%**, and more.

Data Engineer January 2020 – July 2022

Company: Unity Population Health , Client: Heart of Ohio *Hyderabad, India*

- Operated complex data workflows using **Apache Airflow** to automate **ETL** processes from **SQL Server** and perform **Stored Procedure** optimization for seamless data pipeline execution to automate repetitive tasks.
- Managed data integration from multiple health kits, facilitating **Bi-directional** data flow with iOS (**Apple Health**) and Android (**Google Fit**) platforms.
- Integrated patient vitals seamlessly into the **EMR** platform**(Epic)**, utilizing Unity-verified **API** to load live results such as weight, blood glucose, blood pressure, heart rhythm, ECG and pulse for remote patient monitoring.
- Orchestrated the transition to an in-memory database system, reducing data access latency by over **70%**.
- Integrated **OAuth 2.0** for accessing data from Electronic Medical Records **(EMR)** and conducted **Risk Analysis** of monitoring vital data using clustering algorithms to predict dominantly affected patient groups.
- Created a robust **Patient Chatbot** on **Azure** cloud for scalable and efficient deployment of the chatbot, incorporating advanced features for scheduling appointments, conducting mental health screenings, and managing remote patient monitoring.
- Integrated **APIs** from Electronic Medical Record(**EPIC, Allscripts**) to manage and process patient data, ensuring seamless communication between the chatbot and healthcare systems.
- Managed project tasks and communication channels using **Jira** for task tracking and **MS Teams** for real-time collaboration, facilitating cross-functional teamwork and project delivery efficiency.

Junior Data Engineer *September 2019 - December 2019*

- Data Integration ecosystem from source to **Snowflake** schema with **Airflow** Orchestration.

- Created code snippets in **Databricks** notebooks to handle **3.8** million rows with **Spark SQL** and **Python** to execute comprehensive predictive **Risk analysis** framework on clinical data.
- Performed **Standard Scaling** , **PCA**, **Agglomerative** and **K-means** clustering to identify patterns and trends in patient data based on Diagnosis, Lab, Immunizations and Screening tests enhance the understanding of patient risk profiles.
- Employed **MLflow** for end-to-end Machine learning life-cycle management, including experiment tracking, model versioning and deployment, ensuring producibility and scalability of the project.
- **Automated** Clustering Analysis for **350+** ICD codes and draw conclusions with dashboard using **Python** script to visualize trends in vitals and lab tests, facilitating data-driven decision-making for healthcare providers.

Data Scientist Intern April 2019 – August 2019

Company: The Spark Foundations *Hyderabad, India*

- Established automated data pipelines using **Python** and **PostgreSQL** for **15** disparate data sources, and facilitated a **40%** faster data retrieval rate by re-designing database indexing strategies from scratch.
- Optimized **SQL scripts** and **Stored Procedures** which resulted in a **35%** improvement in app response.
- Performed **Feature Selection**(forward selection, backward elimination, and stepwise approach), **Dimensionality Reduction**(PCA) methods to figure out significant variables.
- Built an **Ensemble** model using **Machine Learning** algorithms to forecast student performance, achieving an accuracy rate of **84%** on **Random Forest Regressor**, contributing to a robust predictive framework.
- Enabled real-time dashboards using **Tableau** to visualize core student **KPIs**, enabling students to identify potential areas for improvement.

Technical Skills

Languages: Python 3.x, R, C/C++, SQL, T-SQL, Spark SQL, PySpark, HTML/CSS, Java
Database: PostgreSQL, SQL Server, MSSQL, MySQL, SQLite, Snowflake, MongoDB
Methodologies: Agile, Scrum, Waterfall
Operation Systems: Linux (Any Distro), Unix, Windows
Cloud Services: Azure, GCP
Apache: Hadoop, Spark, Kafka, Solr, Pig, Hive
Data Integration: ETL/ELT, Azure Data Factory, Erwin Modeling, Airflow, Databricks
Statistical Methods: Hypothetical Testing, ANOVA, Time Series, Confidence Intervals, Bayes Law, Principal Component Analysis (PCA), Dimensionality Reduction and Cross-Validation
Business Intelligence and Predictive models: Regression analysis, Decision Tree, Random Forest, Support Vector Machine, Neural Network, K-Means Clustering, KNN, Ensemble, Natural Language Processing
Data Visualization: Tableau, Microsoft Power BI, Matplotlib, Seaborn, Plotly, Microsoft Excel
Machine Learning: Regression, Clustering, SVM, Decision trees, Classification, Recommendation systems
ETL/Data Warehouse Tools: Web Intelligence, Talend, Informatica, Tableau, Data Modeling Star-Schema Modeling, Snowflake-Schema Modeling, and Fact and Dimension tables, Pivot Tables
Frameworks: Flask, Django, Streamlit (No code solution)

Certification

- **Databricks Certified Data Engineer Associate**
- **Microsoft Certified: Azure Data Engineer Associate**
- **Academy Accreditation - Generative AI Fundamentals, Databricks**
- **Academy Accreditation - Databricks Lakehouse Fundamentals**
- **Introduction to Generative AI, Google**
- **Data Analysis with Python: Zero to Pandas, Jovian**
- **Python & SQL Certificate, Hackerrank**
- **Introduction to Data Science**

Education

State University of New York

Master of Science in Computer Science *Buffalo, New York*

Lovely Professional University

Bachelor of Science in Computer Science *Punjab, India*