[LinkedIn](https://www.linkedin.com/in/akshay--saxena/) [Github](https://github.com/akshaysaxena323)

**Professional Summary:**

• Senior Data Engineer with 10+ years of experience in designing, building, orchestrating and maintaining scalable data pipelines and ETL processes. Proficient in leveraging technologies such as SQL, Python, and Apache Spark to transform raw data into actionable insights. Skilled in cloud platforms (e.g., AWS, GCP) and data warehousing solutions (e.g., Redshift, BigQuery). Experienced in automating workflows using CI/CD, ensuring high performance, security, and maintainability of data systems.Strong background in data modeling, database design, and optimization techniques. Adept at collaborating with cross-functional teams to drive data-driven decision-making and enhance data accessibility. Committed to implementing best practices in data governance and data quality assurance. Experience in Generative AI, Text Analytics, developing different Statistical Machine Learning, Data Mining solutions to various business problems and generating Data Visualizations using Python natural language processing.

**Technologies Used:**

|  |  |
| --- | --- |
| **Languages:** | Python (NumPy, pandas, scikit-learn, Tensorflow, PyTorch), R, SQL, Scala, Unix/Shell scripting, Selenium |
| **Data Analysis & Visualization:** | Pandas, scipy, statsmodels, dash, pyspark, plotly, matplotlib, seaborn, bokeh, tableau, streamlit |
| **BigData Technologies/ Databases:** | Apache Spark (PySpark), Hadoop, Apache Kafka,Snowflake, MySQL, PostgreSQL ,MongoDB, Hive, GCP Techstack - Cloud Composer, Cloud Dataproc, Cloud Dataflow, Cloud IAM, Cloud Logging, Cloud Vision, Document AI |
| **Development Tools & Environments:** | Jupyter Notebooks, JupyterLab, VS Code, GCP, Azure, Apache Airflow, Kubernetes, Docker, Gitlab,Cloud Shell, Virtual Machines |

**Professional Experience:**

**Client:Tenet HealthCare May 2022 – Present**

**Project #1: Cloud Data Migration and ETL Pipeline Automation – Supply Chain**

**Role**: Senior Data Engineer – Clinical
**Tools/Technologies**: Google Cloud Platform (GCP), BigQuery, Apache Airflow, SQL Server, ETL, Docker, Kubernetes, PySpark, Apache Hive, GCS, Dataflow

* **Project Overview**: Led the migration of clinical data systems from SQL Server to Google Cloud Platform (GCP), ensuring seamless integration of new systems with legacy applications.
* **Key Achievements**:
	+ Migrated multiple SQL Server tables, views, and stored procedures to GCP, optimizing performance through the use of BigQuery.
	+ Designed and implemented an ETL pipeline in Apache Airflow to automate data extraction, transformation, and loading (ETL), ensuring timely and accurate processing of large-scale clinical data.
	+ Snowflake/Snowpark framework built to automate the process from staging to Dimension and Fact table loads for each functional area.
	+ Automated data pipelines in Azure, Python and Talend to process data into snowflake.
	+ Leveraged Google Cloud Functions to develop an event-driven architecture for monitoring Google Cloud Storage (GCS), triggering an automated ETL pipeline in Google Dataflow to ingest data from multiple source systems into BigQuery.
	+ Optimized SQL Server stored procedures to work seamlessly in BigQuery, significantly improving performance and ensuring compatibility with GCP's cloud-native capabilities.
	+ Integrated Dataflow pipelines with multiple external systems like Oracle, SQL Server, and AS400, ensuring smooth data extraction and transformation.
* **Outcome**: The migration resulted in a significant increase in performance, scalability, and maintainability, improving data accessibility across cloud platforms.

**Project #2: NCDR Optimization and Containerization**

**Role**: Senior Data Engineer – Clinical
**Tools/Technologies**: Google Cloud Platform (GCP), Kubernetes, Docker, Apache Airflow, BigQuery, SQL Server, NCDR, Cloud Functions

* **Project Overview**: Led the optimization of National Cardiovascular Data Repository (NCDR) by transitioning from a legacy sequential process to a containerized solution, significantly improving system efficiency and scalability.
* **Key Achievements**:
* Replaced traditional login mechanism with Two-Factor Authentication (2FA), requiring users to verify identity using One-Time Passwords (OTP) sent via SMS.
	+ Designed and implemented a containerized solution using Docker and Kubernetes, achieving an 80% reduction in runtime for concurrent user requests compared to legacy systems.
	+ Performed Snowflake optimization by cleaning unused tables, WH resizing, Auto suspension consistency and tagging framework.
	+ Implemented data ingestion, streaming and transformations process by Snowflake. ELT automation framework with Snow SQL, External functions, tasks, Streams, Snow pipe, stored procedures, and custom functions using JavaScript, Python, AWS and ECS.
	+ Freed up CPU and memory resources, improving system performance and resource allocation for other critical applications.
	+ Designed and implemented an ETL orchestration framework using Apache Airflow, reducing manual intervention and enhancing workflow reliability and monitoring.
* **Outcome:** The system performance was greatly improved, with a significant reduction in processing time and enhanced scalability for future growth.

**Project #3: Generative AI for Healthcare**

**Role**: Senior Data Engineer – Clinical
**Tools/Technologies**: Google Cloud Platform (GCP), Gemini LLM, BigQuery, Document AI, Natural Language Processing (NLP), AI/ML, Cloud Functions

* **Project Overview**: Developed a Generative AI (GenAI) solution to automate the generation of case summaries and diagnosis reports using Gemini LLMs. This helped to streamline the documentation process in clinical workflows and improve operational efficiency.
* **Key Achievements**:
* Developed a Retrieval-Augmented Generation (RAG) application leveraging Gemini LLMs to summarize case notes and generate comprehensive diagnosis reports.
* Worked on CI/CD tool Jenkins to automate the build process from version control tool into testing and production environment.
* Implemented a solution for analyzing and extracting key data points from contractual agreements using Google Cloud Document AI, enabling efficient searching, summarization, and classification of complex legal documents.
* Leveraged Natural Language Processing (NLP) to improve the accuracy of document extraction, reducing the time spent on manual document review.
* **Outcome**: This project improved the accuracy and speed of data extraction and case documentation, reducing manual workloads and enabling more timely decision-making.

**Project #4: Cloud Data Ingestion and Transformation Framework**

**Role**: Senior Data Engineer – Clinical
**Tools/Technologies**: Google Cloud Platform (GCP), Apache Airflow, Apache Spark, BigQuery, Python, Cloud Functions, Dataflow, SQL Server

* **Project Overview**: Developed and implemented a robust framework for ingesting, transforming, and processing data in Google Cloud Platform using cloud-native tools such as Cloud Functions and Dataflow. The solution supported both batch and real-time data pipelines.
* **Key Achievements**:
* Developed an event-driven File Watcher using Google Cloud Functions to trigger Dataflow pipelines for automatic data processing upon new file arrivals in Google Cloud Storage (GCS).
* Engineered a scalable data pipeline in Apache Spark (PySpark) for transforming large clinical datasets, processing millions of records with minimal latency.
* Postgres SQL functions implemented to transit and persist the data into multiple layers process and code review for evaluating the data quality.
* Designed a highly configurable data transformation framework in Apache Airflow to orchestrate complex data pipelines, dynamically generating Airflow DAGs based on metadata stored in BigQuery.
* Automated the execution of data transformations with Apache Airflow, ensuring data quality, accurate job execution monitoring, and failure logging.
* **Outcome**: The solution streamlined data ingestion and transformation tasks, significantly reducing processing time and improving operational efficiency.

**Highmark Inc. – BlueCross. BlueShield. October 2018 – May 2022**

**Role**: Data Analytics Engineer – Specialty Pharmacy

**Project #5:** Specialty Pharmacy Analytics and Cost Optimization
**Environment**: Hadoop, Apache Hive, Teradata, Python, R, Tableau, SQL Server, Python

**Description**: Highmark is a health insurance company focused on providing pharmacy services and managing clinical data for optimal healthcare delivery. This project aimed at building an analytic layer on top of Highmark's existing Enterprise Data Warehouse (EDW) to analyze specialty drug costs and improve profitability through data-driven insights. The work involved migrating data from Teradata to the Hadoop ecosystem, leveraging data science models, and building analytical dashboards to enhance clinical decision-making and client reimbursement strategies.

**Roles and Responsibilities**:

* Data Migration & Infrastructure Optimization:
Led the migration of Teradata tables to Apache Hive tables, enabling seamless integration with the Hadoop ecosystem and improving the scalability and performance of data pipelines. This migration provided Highmark with the ability to process large volumes of specialty drug data more efficiently.
* Exploratory Data Analysis (EDA) on Specialty Drug Usage:
Conducted exploratory data analysis (EDA) on specialty drug usage, leveraging members' medical histories and personal information. Identified patterns and trends in drug consumption that helped shape strategic decisions, contributing to cost optimization and improved profit margins.
* Real-Time Dashboard Development:
Designed and implemented a real-time dashboard using Tableau to visualize key drug consumption metrics. The dashboard provided cross-functional teams with easy access to drug utilization insights, fostering collaboration and enabling faster, data-driven decision-making.
* Automated Data Extraction with Python:
Developed Python scripts to automate the data extraction process from the Chronic Medication Service (CMS) directly into the Hadoop ecosystem, bypassing intermediate Teradata storage. This automation streamlined the data pipeline, reducing processing time and enhancing the overall data extraction workflow.
* Site of Care Development Project:
Led the Site of Care Development project, analyzing claims rates from Walgreens and comparing them against optimal payment models. Utilized R, Python, Tableau, and Apache Hive to provide actionable insights and recommendations to optimize client reimbursement strategies.
* Advanced Drug Analysis with Data Science Models:
Performed advanced drug analyses using data science models and pharmacy analytics to provide tailored insights to various internal teams. These insights helped fulfill service requests and contributed to better operational effectiveness, enhancing Highmark's ability to serve clients effectively.
* Retrospective Claims Analysis for Hemophilia Products:
Executed a retrospective claims analysis of bleeding rates among hemophilia factor products. Identified trends and patterns to inform clinical decision-making, ultimately leading to improved patient outcomes and optimized care strategies for high-cost specialty medications.

**Outcome**:

The project significantly enhanced Highmark’s ability to analyze specialty drug costs and consumption, leading to better reimbursement strategies, cost optimization, and improved operational efficiency. The migration to Hadoop/Hive and automation of data pipelines greatly increased processing efficiency, while the development of real-time dashboards and advanced analytics drove improved decision-making across the organization.

**UnitedHealth Group May 2013 - July 2016**

*Software Engineer – Data Engineering*

**Project #6 *:*****Healthcare Analytics and Predictive Modeling****Environment*:***Python, Big Data (Hive, Pig, HBase, Spark), R, Weka, Octave, Keras, TensorFlow, Amazon Redshift, Apache Cassandra, Kafka

**Description**: UnitedHealth Company is a leading healthcare product development firm committed to transforming healthcare and enhancing consumer engagement through technology. The project focused on utilizing Python and machine learning algorithms to analyze healthcare data and provide insights for better decision-making. The work involved data wrangling, predictive modeling, and working with various big data technologies to process and analyze large healthcare datasets, with a strong focus on improving operational efficiencies and healthcare outcomes across the United States.

**Roles and Responsibilities:**

* Data Wrangling & Predictive Modeling:
Extensively used Python for data wrangling, preprocessing, and building predictive models to gain insights from healthcare datasets. Leveraged machine learning algorithms to predict healthcare trends and outcomes, providing actionable insights to improve consumer engagement and healthcare delivery.
* Big Data Technologies:
Computed and analyzed large datasets using MapReduce, Hive, Pig, HBase, and Spark. Processed and transformed data to enable scalable analysis and support real-time data processing for healthcare analytics applications.
* Data Science Toolkits:
Utilized common data science toolkits such as R, Weka, NumPy, Rattle, and Octave to build models and perform complex data analysis, helping to uncover patterns and trends in large healthcare datasets.
* Collaborated with data engineers, web developers, and product managers to deploy machine learning models in production. Contributed to the integration of models into existing healthcare systems, ensuring they were operationalized for end-user use.
* Translated business requirements into technical specifications, ensuring alignment between business objectives and technical deliverables. Worked on the documentation part of the project to ensure smooth knowledge transfer and project transparency.
* Knowledge Transfer & Training:
Provided knowledge transfer sessions to new joiners, enabling them to understand the functional and technical aspects of the project. Ensured the team was well-versed in using the tools and technologies employed for predictive modeling and big data processing.
* Built and maintained the ETL infrastructure for diverse data sources, ensuring the smooth extraction, transformation, and loading of healthcare data into appropriate systems for analysis.
* Utilized NoSQL databases like Amazon Redshift and Apache Cassandra to enable scalable, efficient data storage and fast data retrieval. These databases were crucial in processing large volumes of healthcare data, ensuring high performance for real-time analytics and reducing the data storage complexity.
* Built and maintained the ETL infrastructure for diverse data sources, ensuring the smooth extraction, transformation, and loading of healthcare data into appropriate systems for analysis.

**EDUCATION**

* Master of Science in Computer Information Systems (Concentration in Data Science)
University of Houston-Clear Lake, Houston, TX - 2018
* Bachelor of Science in Information Technology
Rajasthan Technical University, Kota, India - 2012Prictive Analytics, Predictive Analytics, Diagnostic Analysis, Diagnostic Analysis, Diagnostic Analysis,Predictive Analytics, Feature Engineering, Data Processing, Feature Engineering, Data Processing Feature Engineering, Data Processing, ETL, ETL, ETL,Business Analytics, Business Analytics, Business Analytics,Kafka, Kafka, Kafka, Docker, Docker, Docker, Kubernetes, Kubernetes, NoSQL, NoSQL, NoSQL,Data Integration, Data Integration, Data Migration, Data Migration, Data Quality, Data Quality, Data Governance, Data Governance,Batch Processing, Batch Processing, Batch Processing,Data Visualization, Data Visualization,Data Pipeline, Data Pipeline, Data Pipeline, Data Modeling, Data Modeling, Data Modeling,Stream Processing, Stream Processing, Stream Processing