# Rahul Waghulde

# Contact: +1 929-564-5130

# Email: waghulde.rahul16@gmail.com

**PROFESSIONAL SUMMARY**

Experienced and adaptable Generative AI Specialist and Data Engineer proficient in Azure, Amazon, and Google Cloud ecosystems. Skilled in crafting and overseeing comprehensive data pipelines spanning various platforms. Expertise includes harnessing Azure AI and machine learning studio, with a deep understanding of both data science and engineering domains. Demonstrated success in optimizing GPU performance, implementing RAG vector search, and leveraging NLP and LLM technologies.

* Around 10+ Years of Information Technology experience with Data Engineering and report development – a good mix of Data Science, Business Intelligence, Report Development, Data Analytics, Data Warehouses – Data Marts etc.
* Experience in Agile Methodologies, Scrum stories, and sprints experience in a Python-based environment, along with data analytics, and data wrangling.
* Experience in **Data Science/Machine Learning** in different domains such as **Data Analytics, Machine Learning (ML), Predictive Modelling, Natural Language Processing (NLP), and Deep Learning algorithms.**
* Proficient at a wide variety of Data Science programming languages **Python, R, SQL, PySpark, Sci-kit Learn, NumPy, SciPy and Pandas, NLTK, TextBlob, Genism, SpaCy, Keras and TensorFlow.**
* Experienced in fine-tuned **LLMs**, including **GPT-3.5** and **GPT-4**, for various natural language processing tasks, such as text generation, summarization, and translation.
* Ability to integrate and optimize **LLMs** and deep learning models on edge devices, ensuring efficient performance in resource-constrained environments.
* Experienced in Integrating **SARIMA** models within a broader analytics framework, providing stakeholders with actionable insights on seasonal trends and cyclic patterns.
* Strong knowledge **Apache Spark**, **Scala**, **Python**, R and **MLOps**
* Experienced in facilitating the entire lifecycle of a **data science** project: Data Cleaning, Data Extraction, Data Pre-Processing, Dimensionality Reduction, Algorithm implementation and Validation.
* Expert in **Machine Learning** algorithms such as Ensemble Methods (Random forests), Linear, Polynomial, Logistic Regression, Regularized Linear Regression, Support Vector Machines (SVM), Deep Neural Networks, Extreme Gradient Boosting, Decision Trees, K-Means, K-NN, Gaussian Mixture Models, Naive Bayes.
* Proficient in Creating, Debugging, Scheduling, and Monitoring jobs using Airflow for ETL batch processing to load into Snowflake for analytical processes.
* Experience on Migrating SQL database to **Azure Data Lake, Azure Data Lake Analytics, Azure SQL Database, Data Bricks,** and **Azure SQL Data Warehouse**and Controlling and Migrating On-premises databases to Azure Data Lake store using Azure Data factory.
* Experience in using Jira/Azure Dev Ops for ticketing and tracking issues and Jenkins for continuous integration and continuous deployment.
* Experienced in Data Integration Validation and Data Quality controls for ETL process and Data Warehousing using MS Visual Studio **SSIS, SSAS, SSRS**.
* Experience with version control tool – Git and build tools like Apache Maven/Ant.
* Hands on experience with big data tools like Hadoop, Spark, Hive, Pig, PySpark, Spark SQL, PySpark
* Skilled in data management, including **Data munging, Data cleaning, Data Analytics, Data Visualization,** and Big Data ecosystems using **Hadoop, Hive, HDFS, MapReduce**, Spark, Airflow, Snowflake, Teradata, Flume, Yarn, Oozie, and Zookeeper.
* Solid understanding of **Agile Methodologies, Scrum stories** and **sprints in a SQL** and Oracle-centric environment, bolstered by robust **data analytics and data wrangling skills**.
* Capable of facilitating the entire lifecycle of a data project: Data Cleaning, Data Extraction, Data Pre-Processing, Dimensionality Reduction, Algorithm implementation, Back Testing, and Validation.

**TECHNICAL SKILLS:**

|  |
| --- |
| **Big Data Technologies:**  Azure Data Factory, Hadoop, MapReduce, HDFS, Hive, HBase, NiFi, Airflow, Apache Spark. |
| **Databases:**  Azure CosmosDB, Postgresql Flexible Server, MySQL, Azure SQLdb, MongoDB |
| **Programming:**  Python, R, Java, JavaScript, React, Shell script, SQL, markdown |
| **Machine Learning:** LLM, LSTM, RESNET-50, RNN, CNN, Regression (Linear and Logistic), Decision trees,  Random Forest, SVM, KNN, PCA. |
| **ML Frameworks:** Promptflow, Langchain, Pandas, Keras, NumPy, TensorFlow, Scikit-Learn, NLTK. |
| **Cloud Technologies:**  Azure, AWS, GCP |
| **Azure Tools:** Open AI, Machine Learning Studio, Promptflow, Azure notebooks, Azure Databricks |
| **AWS Tools:** EC2, S3, Glue, Athena, AWS Bedrock |
| **Versioning tools:**  Git, GitHub, bitbucket |
| **Operating Systems:**  Windows, Ubuntu Linux, MacOS |

**PROFESSIONAL EXPERIENCE:**

**Bank of America - San Antonio, TX Feb 2023 to Till Date**

**Role: Data Scientist / Gen-AI Data Engineer**

**Project Overview:** At Bank of America, I designed and implemented end-to-end data pipelines using AWS, GCP, and Azure platforms. I developed AI and ML models using Azure OpenAI and worked on RAG vector search implementations for text generation and NLP tasks. I was responsible for automating model deployments, improving SARIMA models for time-series forecasting, and collaborating with cross-functional teams to ensure efficient model performance in production.

**Responsibilities:**

* Designed and implemented end-to-end data engineering pipelines using Azure Data Factory, Azure SQL, Blob storage, and Data Lake Analytics, ensuring efficient and scalable data integration and processing.
* Developed time-series forecasting models using **SARIMA** and **Prophet**, improving forecasting accuracy by 20%, resulting in better inventory management and resource allocation for banking systems.
* Utilized **MLOps** tools such as **Vertex AI**, **MLflow**, and **Kubeflow** to automate the lifecycle of machine learning models, including training, evaluation, and deployment, reducing time-to-production by 30%.
* Integrated time-series models with **TensorFlow Extended (TFX)** and **Kubeflow**, facilitating seamless deployment of demand forecasting models into production environments, optimizing resource planning.
* Developed and deployed machine learning models for predictive analytics using **Python** libraries such as **TensorFlow**, **PyTorch**, and **scikit-learn**, enhancing financial trend predictions and loan risk analysis.
* Utilized **Apache Spark** in Synapse for distributed data processing, enabling large-scale analytics and fast data exploration.
* Developed and maintained scalable **Java-based** applications for various business requirements, implementing key **OOP principles** such as **inheritance**, **encapsulation**, and **polymorphism** to improve code modularity and maintainability.
* **Implemented comprehensive testing strategies** using **JUnit** and **Mockito**, including unit, integration, and functional testing, ensuring code reliability and reducing production defects.
* Monitored performance and resource usage using **Azure Synapse Studio** and integrated **Azure Monitor** to ensure continuous operational efficiency.
* Optimized machine learning algorithms for performance improvement using **Apache Spark** and **Azure Databricks**, enhancing the scalability and efficiency of banking data models.
* Implemented real-time streaming solutions using **Azure Stream Analytics** and **Kafka**, enabling the processing of high-volume, time-sensitive financial data, and ensuring compliance with latency requirements.
* Built predictive models for financial applications using advanced algorithms like **Random Forests**, **Gradient Boosting**, and **Neural Networks**, delivering high-accuracy predictions for loan defaults and customer churn.
* Developed and deployed conversational AI chatbots using frameworks like **OpenAI’s GPT** and **PyTorch**, automating customer interactions and improving response time for banking services.
* Programmed **Python** utilities with **SciPy**, **NumPy**, and **Pandas** for data processing, statistical analysis, and model optimization, enhancing the overall accuracy and performance of financial models.
* Leveraged **Azure App Services** and **Azure Kubernetes Service (AKS)** to deploy and manage scalable AI and machine learning applications, ensuring high availability and performance for critical banking applications.
* Built **ETL** workflows using **Informatica PowerCenter** to extract, transform, and load data from various financial systems into target databases, improving data availability for reporting and analysis.
* Designed **REST APIs** with **Spring Boot**, using annotations like **@RestController**, **@RequestMapping**, and **@PathVariable** to streamline development and enhance API security with **Spring Security** and **JWT tokens**.
* Automated the deployment of machine learning models into production environments using **CI/CD pipelines** integrated with **Jenkins** and **GitLab CI**, reducing manual deployment efforts by 50%.
* Developed deep learning models using transformer-based architectures like **BERT** and **GPT** for natural language processing tasks, including sentiment analysis and named entity recognition, improving customer sentiment tracking.
* Utilized **Azure’s PostgreSQL Flexible Server** and **Cosmos DB** with **pgvector** extension to enable fast vector search for **Retrieval Augmented Generation (RAG)** and enhance the accuracy of AI-powered financial document searches.
* Implemented demand forecasting models in **Python** using **SARIMA** and **Prophet**, automating the forecasting process and reducing manual effort by 50%, leading to better financial planning and decision-making.
* Performed **Exploratory Data Analysis (EDA)** using **Python** libraries such as **Pandas**, **NumPy**, and **matplotlib** to uncover critical insights from financial datasets, enhancing data-driven decision-making.
* Developed machine learning-based fraud detection systems using techniques like **XGBoost**, **Support Vector Machines (SVM)**, and **Random Forests**, improving the identification of suspicious transactions in banking systems.
* Led the migration of legacy **ETL** processes to modern data platforms like **Azure Data Factory**, ensuring smooth data migration, zero data loss, and improved operational efficiency for banking clients.
* Utilized **PyTorch** for training and deploying deep learning models focused on time-series predictions, improving the forecasting of financial metrics and customer behavior patterns by 20%.

**Environment:** Azure Open AI, Azure SQL DB, Azure Notebooks, Azure Data Factory, Azure Functions, Power BI, Promptflow, Langchain, PostgreSQL Flexible server, pgvector extension, CosmosDB, Sharepoint, R, Python, Agile.

**BNY Mellon – Pittsburgh, PA Jan 2022 to Feb 2023**

**Role: Data Scientist / Gen-AI Data Engineer**

**Project Overview:** In this role, I focused on developing custom LLMs and implementing machine-learning solutions for domain-specific applications. I worked on automating the data pipeline using PySpark, AWS Glue, and Apache Spark, and developed time-series models to forecast data trends. Additionally, I was involved in automating ML workflows, building classification models, and improving data processing using big data tools in a cloud-based environment.

**Responsibilities:**

* Designed and set up an Enterprise Data Lake to support use cases including analytics, data processing, storage, and reporting for voluminous and rapidly changing financial data.
* Developed **PySpark scripts** to encrypt raw data using hashing algorithms, enhancing data security for sensitive client information.
* Implemented **LSTM and GRU networks** to solve vanishing gradient issues, improving model stability and performance in complex time series prediction tasks for financial data forecasting.
* Fine-tuned **OpenSearch queries and index mappings**, improving query response times by 30% for search-based financial applications.
* Built **data lakes** using **Azure Data Lake Storage (ADLS)** integrated with **Azure** **Synapse** to handle structured, semi-structured, and unstructured data.
* Streamlined **data ingestion processes** by integrating **Apache Spark** and **SQL** pools for real-time and batch processing in **Azure** **Synapse**.
* Converted **Hive/SQL queries into Spark transformations** using **Spark RDDs** and **Python**, optimizing query performance for large-scale datasets.
* Automated **data cataloging** using **AWS Glue Catalog** with crawlers, enabling SQL query operations on S3-stored financial data.
* Utilized **Spring Boot's dependency injection (DI) and Aspect-Oriented Programming (AOP)** features to reduce code coupling, enhance maintainability, and simplify cross-cutting concerns like logging, security, and transaction management.
* Developed **custom machine learning pipelines** using **PySpark** and **AWS Glue** to automate demand forecasting and supply chain processes, increasing operational efficiency by 25%.
* Integrated **Bedrock with AWS services** like **S3**, **Lambda**, and **SageMaker** to enhance performance, scalability, and machine learning capabilities in data-driven banking applications.
* Managed **security groups in AWS**, ensuring high availability, fault tolerance, and auto-scaling using **Terraform templates** for reliable cloud infrastructure in banking systems.
* Automated **machine learning lifecycle** by integrating **PyTorch** with **Kubeflow** and **MLflow**, enabling continuous training, versioning, and deployment of financial models.
* Deployed **and** managed **OpenSearch clusters** in AWS, ensuring optimal configuration and disaster recovery capabilities for financial search-based applications.
* Developed **and** deployed **domain-specific LLMs**, tailoring them for financial applications to meet BNY Mellon’s unique business requirements.
* Developed **AI-powered chatbots and virtual assistants** using LLMs to enhance customer interaction through natural language understanding, integrated with APIs like **Hugging Face** and **LangChain** for seamless pipeline automation.
* Implemented **XGBoost models** using **Python** for statistical modeling, determining predictive probabilities in financial forecasting models.
* Built **classification models** like **Logistic Regression**, **SVM**, and **Random Forest** to predict customer churn rates in the banking sector, improving retention strategies.
* Developed **ETL pipelines** using **Python** and **PySpark**, automating data ingestion and processing for large-scale data warehousing in BNY Mellon’s financial systems.
* Performed **data cleaning and feature selection** using PySpark’s ML packages and deep learning frameworks like **TensorFlow** and **Keras**, optimizing model performance for financial analysis.
* Developed **sentiment analysis models** using **Hugging Face** for monitoring customer feedback and improving banking services, identifying areas for improvement through natural language processing.
* Created **Lambda functions with Boto3** to automate EC2 cost optimization by deregistering unused AMIs across application regions, reducing operational costs.
* Developed **ETL** jobs **for AWS Glue and EMR** for efficient data transformation and migration between **Amazon S3** and **Athena**, enabling real-time data querying.
* Employed **LangChain for automating customer onboarding**, chaining NLP tasks like document parsing, data validation, and regulatory checks into a single pipeline, reducing manual intervention.
* Implemented **Airflow DAGs** for scheduling and monitoring ETL batch jobs, loading financial data into **Snowflake** for analytics, ensuring efficient job management and orchestration.

**Environment:** AWS EMR, EC2, S3, RDS, Athena, Glue, Auto Scaling, Elastic Search, Lambda, Amazon SageMaker, Apache Spark, HIVE, Map Reduce, Snowflake, Python, Tableau, Agile.

**ASFS Wealth Management – Mumbai, India July 2018 to Feb 2021**

**Role: Data Scientist**

**Project Overview:** As a Data Scientist, I was responsible for developing machine learning models to improve financial decision-making processes. I worked on predictive analytics, feature engineering, and model optimization using TensorFlow, Keras, and PySpark. I also developed advanced ETL pipelines to integrate financial datasets and collaborated with teams to implement deep learning models, which were deployed on GPU-based platforms for faster processing.

**Responsibilities:**

* Developed, maintained, and optimized machine learning models using TensorFlow, Keras, NumPy, Scikit-Learn, and tf.Data API in Python, improving the predictive performance for wealth management portfolios.
* Conducted **Exploratory Data Analysis (EDA)** and cleaned fingerprint image datasets stored on a local network using **Python**, ensuring data quality through the **NFIQ algorithm**, and visualized the distributions using **histograms** for in-depth analysis.
* Implemented machine learning algorithms such as **Generalized Linear Models (GLM)**, **SVM**, **Random Forest**, **Boosting**, and **Neural Networks** using **R** and **Spark (PySpark, MLlib)**, enhancing client investment prediction accuracy.
* Performed data imputation using the **Scikit-learn** package in **Python** to handle missing data, improving the quality and reliability of financial datasets for model training.
* Designed and implemented system architecture for a cloud-hosted solution on **Amazon EC2**, ensuring scalability, security, and cost-effectiveness for data-driven wealth management systems.
* Extracted data from **AWS S3** buckets, performing large-scale data pulls to retrieve relevant wealth management data for further analysis and model building.
* Leveraged **AWS Transcribe** to process and analyze call transcripts, applying text processing techniques to derive insights from client interactions and service feedback.
* Developed **MapReduce** and **Spark Python** modules for predictive analytics and machine learning on **Hadoop** in **AWS**, enabling faster processing of large wealth management datasets.
* Transformed image datasets into **protocol buffers**, serialized them, and stored the data in **TFRecord** format, ensuring efficient handling of large datasets for deep learning models.
* Optimized data ingestion pipelines using **tf.Data API** for **TFRecord** files, enabling scalable model training on datasets larger than the available CPU memory by streaming data over the network.
* Performed **GPU-based training** using **Half Precision FP16** on **Nvidia Titan RTX** and **Titan V GPUs** for **TensorFlow**, accelerating model training for large-scale wealth management predictive models.
* Automated hyperparameter tuning and model optimization, running over 50 variations of model configurations and generating automated reports, reducing time-to-insight for model selection.
* Maintained and retrained models created by other data scientists, optimizing them for wealth management data by applying different datasets and improving overall model accuracy.
* Created custom tools for data scientists, enhancing their productivity in exploring and processing wealth management data, as well as automating repetitive data science tasks.
* Productized **TensorFlow** models by converting them into a format compatible with **C++** and **Android applications**, facilitating the integration of predictive models into existing financial systems.
* Applied **Transfer Learning** using **ResNet50** with fingerprint images by freezing lower layers and retraining upper layers, boosting the performance of financial client verification models.
* Visualized internal layers of **CNNs** by creating **Class Activation Maps (CAM)**, enabling better understanding of how models process client-related data in wealth management.
* Evaluated model performance using **Validation** and **Testing sets** to avoid overfitting, and measured accuracy with **Confusion Matrix** and **ROC Curves**, ensuring reliable predictions for wealth management insights.

**Environment:** Hadoop, Agile, MapReduce, Snowflake, Spark, Hive, Kafka, Python, R, Airflow, JSON, AWS, EC2, S3, Athena, Glue, AutoScaling, EKS, ELB, Tensorflow, Keras

**Infosys – Mumbai, India June 2015 to June 2018**

**Role: Data Scientist**

**Project Overview:** In this role, I worked on implementing machine learning models for customer satisfaction analysis and predictive modeling. My primary focus was on processing large datasets, developing algorithms using Python and R, and implementing NLP models to analyze customer feedback. I also contributed to building cloud-hosted solutions using AWS and collaborated closely with the business to design scalable data architecture.

**Responsibilities:**

* Analyzed business requirements and mapped functional specifications to design **MapReduce** programs and algorithms for large-scale data processing.
* Worked with various data formats such as **JSON** and **XML** and implemented machine learning algorithms in **Python**, enhancing data-driven decision-making processes.
* Set up storage and data analysis tools in **Amazon Web Services (AWS)**, leveraging cloud infrastructure for scalable data management and computation.
* Utilized **R packages** such as **knitr**, **dplyr**, **SparkR**, **Causal Infer**, and **Space-Time** for statistical modeling and data analysis in various business use cases.
* Built **data lakes** using **Azure Data Lake Storage (ADLS)** integrated with **Azure** **Synapse** to handle structured, semi-structured, and unstructured data.
* Streamlined **data ingestion processes** by integrating **Apache Spark** and **SQL** pools for real-time and batch processing in **Azure** **Synapse**.
* Applied **Python libraries** like **Pandas**, **NumPy**, **Seaborn**, **SciPy**, **Matplotlib**, **Scikit-learn**, and **NLTK** to develop machine learning algorithms, improving insights derived from structured and unstructured data.
* Conducted **Natural Language Processing (NLP)** to determine customer satisfaction and enhance customer experience through sentiment analysis and text classification.
* Designed and implemented **Statistical models**, **Predictive models**, and **Enterprise Data Models**, managing data life cycle in both **RDBMS** and **Big Data** environments to streamline business intelligence processes.
* Performed **Data Quality Validation** to verify **Critical Data Elements (CDE)** and identified anomalies, ensuring data integrity across systems.
* Participated in all phases of **Data Mining**, including **Data Collection**, **Data Cleaning**, **Development**, **Validation**, **Visualization**, and performed **Gap Analysis** to optimize data processes.
* Worked with **NoSQL databases** like **Cassandra** to handle large volumes of unstructured data in distributed systems.
* Programmed utilities in **Python** using **SciPy**, **NumPy**, and **Pandas** for automating data processing tasks, improving operational efficiency.
* Implemented classification models using supervised algorithms such as **Logistic Regression**, **Decision Trees**, **K-Nearest Neighbors (KNN)**, and **Naive Bayes**, improving prediction accuracy for business applications.
* Handled various file formats such as **Text files**, **Sequence Files**, **Avro**, **ORC**, and **Record Columnar CRC**, ensuring efficient data storage and retrieval.
* Used **AWS services** like **EC2** and **S3** for handling small data sets, enhancing scalability and availability of data storage and processing.
* Designed **3NF Data Models** for **ODS** and **OLTP systems**, as well as **Dimensional Data Models** using **Star** and **Snowflake Schemas**, optimizing database performance.
* Updated **Python scripts** to match training data with database records stored in **AWS Cloud Search**, automating response label assignment for document classification.
* Created **SQL tables** with referential integrity and developed complex queries using **SQL**, **SQL PLUS**, and **PL/SQL**, optimizing data retrieval and manipulation for reports.
* Designed and developed **Use Case**, **Activity Diagrams**, and **Sequence Diagrams** using **UML** and **Visio**, supporting software development and business process visualization.
* **Environment:** AWS, R, Machine learning-Algorithms, Anaconda, Predictive Analytics, Deep Learning- Algorithms, CNN, HCNN, Python, Data Mining, Data Collection, Data Cleaning, Validation, HDFS, Hive, OLAP, Metadata, MS Excel, SQL, and MongoDB.

**TCS – Mumbai, India June 2013 to June 2015**

**Role: Junior Data Scientist**

**Project Overview:** As a Junior Data Scientist, I was responsible for developing machine learning models and data pipelines for various predictive analytics projects. I focused on data cleaning, preprocessing, and building scalable ETL solutions, which enabled efficient data flow and actionable insights. I worked on multiple cross-functional projects, collaborated with business analysts and engineers, and delivered data-driven insights to improve business performance.

**Responsibilities:**

* Designed and implemented predictive models using machine learning algorithms such as Linear Regression, Decision Trees, and Random Forests to address business challenges like customer churn prediction and sales forecasting.
* Performed extensive data cleaning and preprocessing using **Pandas** and **NumPy**, handling missing values, feature scaling, and data normalization, ensuring high-quality input data for model training.
* Developed **SQL-based ETL pipelines** for automating data extraction, transformation, and loading from multiple sources, ensuring efficient data ingestion into data warehouses for further analysis.
* Evaluated models using performance metrics such as **accuracy**, **precision**, **recall**, and **F1-score**, and improved model performance through **hyperparameter tuning** techniques like grid search and cross-validation.
* Collaborated with cross-functional teams, including business analysts, software engineers, and project managers, to deliver data-driven solutions aligned with business objectives and integrated machine learning models into existing systems.
* Built interactive dashboards using tools like **Tableau** and **Matplotlib** to visualize data insights, enabling business stakeholders to make informed decisions based on real-time trends and key performance metrics.
* Engineered new features from existing data, improving model accuracy by 20%, and applied **dimensionality reduction** techniques like **Principal Component Analysis (PCA)** to streamline datasets and reduce model complexity.
* Automated data collection and reporting using **Python scripts**, reducing manual work by 40%, and developed automated weekly and monthly performance reports for senior management.
* Conducted **Exploratory Data Analysis (EDA)** to uncover hidden patterns and correlations in data, presenting actionable insights through well-structured reports to support strategic decision-making.
* Worked with **Hadoop** and **Apache Spark** for processing large datasets, optimizing processing time by 50% for high-volume tasks such as user behavior analysis and customer segmentation.

**Environment:** Python, Pandas, NumPy, Sci-kit Learn, TensorFlow, SQL, Tableau, Apache Spark, Hadoop, Jupyter Notebook, Matplotlib, Seaborn, Excel, AWS (S3, EC2), Git, Jenkins, Linux (Ubuntu), Windows