Haritha gurram
Senior Data Scientist
Phno: +12148430748| harithagh.03@gmail.com

PROFESSIONAL SUMMARY

- Senior Data Scientist with 10+ years designing and deploying advanced analytics, machine learning models, and data-driven solutions, enabling predictive insights, real-time decision support, and business optimization across M&A, healthcare, finance, and e-commerce domains.
- Designed and implemented **ETL workflows** using **GCP Dataflow**, **Apache Spark**, and **Talend**, integrating data from **GCP Pub/Sub**, **Cloud SQL**, **Hadoop**, and **Azure Data Lake**, ensuring structured data availability in **BigQuery** and **Snowflake** for analytics, reporting, and downstream processing pipelines.
- Developed batch and streaming pipelines using **Apache Spark**, **PySpark**, **Spark Structured Streaming**, and **AWS S3**, transforming raw datasets into enriched formats while maintaining data lineage with **Apache Atlas**, **Monte Carlo**, and **Bigeye** for operational reporting and analytics compliance.
- Processed and modeled large-scale datasets using **Python**, **pandas**, **NumPy**, **SciPy**, and **SQL**, combining structured and unstructured data from **Azure SQL Database**, **Azure Blob Storage**, **Amazon RDS**, and **Redshift** to support financial, healthcare, and manufacturing analytics workloads.
- Managed predictive and machine learning models using **PyTorch**, **Keras**, **LightGBM**, **SARIMA**, **Scikit-learn**, **XGBoost**, **Prophet**, and **ARIMA**, evaluating performance with **SHAP**, **MLflow**, and **Optuna**, preparing datasets from **Azure Synapse**, **Databricks**, and **Snowflake** for training and validation.
- Processed text and NLP datasets using **spaCy**, **NLTK**, **regex**, **PyPDF2**, **pdfplumber**, **textract**, and **FHIR**, integrating healthcare and structured datasets to create searchable, standardized pipelines supporting **HL7** compliance across multiple environments.
- Built interactive dashboards and reports using **Power BI**, **Tableau**, and **Looker**, sourcing data from **BigQuery**, **Snowflake**, **Azure Synapse**, and **Redshift**, visualizing KPIs, operational metrics, and predictive insights for finance, healthcare, and manufacturing stakeholders.
- Monitored pipelines and infrastructure using **GCP Stackdriver**, **AWS CloudWatch**, **Azure Monitor**, **Prometheus**, and **Datadog**, implementing alerting, logging, and dashboards to ensure reliability, high availability, and rapid failure detection across **AWS**, **GCP**, and **Azure** environments.
- Managed cloud compute and container environments with **Kubernetes**, **GCP GKE**, **AWS EKS**, **Azure Virtual Machines**, and **EC2**, deploying containerized workflows using **Docker** and orchestration frameworks such as **Kubeflow** for scalable and reproducible analytics processing.
- Orchestrated CI/CD pipelines using **Git**, **GitLab CI/CD**, **GitHub Actions**, **Jenkins**, and **SageMaker Pipelines**, enabling version control, automated testing, and reproducible deployments for **Python**, **SQL**, and **PySpark** workflows across multi-cloud ecosystems.
- Managed large-scale cloud data warehouses including **Snowflake**, **BigQuery**, **Amazon Redshift**, and **Azure Synapse Analytics**, integrating ETL data from **AWS S3**, **Azure Data Lake**, **Azure Blob Storage**, and **GCP Cloud Run** for analytics-ready reporting datasets.
- Performed data quality validation and governance using **Collibra**, **Apache Atlas**, **Monte Carlo**, **Bigeye**, **Alation**, **Great Expectations**, and **Deequ**, ensuring lineage-tracked, compliant datasets aligned with **HIPAA**, **HL7**, and **FHIR** regulatory standards.
- Built and deployed ML inference workflows using **SageMaker**, **Triton Inference Server**, **FastAPI**, **TensorFlow**, and **PyTorch**, integrating **LoRA**, **PEFT**, **T5**, **BERT**, **LLaMA**, **GPT-3.5**, and **GPT-4 embeddings** with vector stores including **FAISS**, **Weaviate**, **Milvus**, and **Chroma**.

- Conducted advanced analytics using **Python**, **pandas**, **NumPy**, **SciPy**, **Seaborn**, **Matplotlib**, **Darts**, **LIME**, and **Hyperopt**, performing time-series forecasting, experimentation, and optimization to improve predictive accuracy across operational and healthcare datasets.
- Developed data pipelines and orchestration workflows with **Apache Beam**, **Databricks**, **Azure Databricks**, **dbt**, **Azure Data Factory**, and **Azure Synapse**, transforming structured and unstructured datasets into **Snowflake**, **BigQuery**, and **Redshift** for reporting and analytics.
- Managed cloud security and compliance using **AWS IAM**, **KMS**, **AWS VPC**, ensuring encryption, access control, and governance across **AWS**, **Azure**, and **GCP** pipelines aligned with **HIPAA** and **FHIR** standards.
- Conducted model monitoring and observability using **EvidentlyAI**, **Weights & Biases**, **WhyLabs**, **Grafana**, and **Prometheus**, tracking performance, drift, and operational anomalies for ML workflows deployed on **SageMaker**, **Kubeflow**, and **Azure ML**.
- Processed and prepared analytics datasets using **Azure ML**, **Databricks**, **Spark MLlib**, **Python**, **pandas**, **NumPy**, and **SQL**, integrating data from **Azure Blob Storage**, **Azure SQL Database**, **AWS S3**, **Amazon RDS**, and **Redshift**.
- Implemented workflow automation and orchestration using **Apache Airflow**, **Prefect**, **AWS Glue**, **Step Functions**, and **Lambda**, managing ETL scheduling, notifications through **SQS** and **SNS**, ensuring reliable and timely multi-cloud data delivery.

TECHNICAL SKILLS

- **Cloud Platforms:**
  - AWS (Lambda, S3, RDS, Dynamo DB, Neptune, Glue, Redshift, EC2, SNS/SQS, Cloud Watch, Athena, VPC, KMS, Route 53, CodePipeline, Cloud Formation, CLI, ALB/ELB)
  - Azure (Data Factory, Data Lake, Databricks, Synapse, SQL DB, HDInsight, DevOps, Functions, Logic Apps, Monitor, Cosmos DB, ML Studio)
  - GCP (BigQuery, Storage, Dataflow, Dataproc, Composer, Cloud Functions, IAM, Monitoring, Security Centre)
- **Programming & Scripting:** Node.js, Python, Scala, Java, Shell, JavaScript, FastAPI, Flask
- **Big Data & Data Engineering:** Apache Spark, Hadoop (HDFS, MapReduce, YARN, Hive, Pig, HBase), Flink, Kafka, Sqoop, Oozie, Nifi, Delta Lake, Snowflake
- **ETL & Workflow Orchestration:** Airflow, AWS Glue, DBT, Informatica, Talend, DataStage, Azure Data Factory, Cloud Composer, Fivetran
- **DevOps & CI/CD:** Jenkins, Git, Bit bucket, Ansible, Terraform, Docker, Kubernetes, MLflow, Cloud Formation, CI/CD Pipelines, Github, Chef, Puppet
- **Databases:** RDBMS, Oracle, SQL Server, PostgreSQL, MySQL, Teradata, DB2, MongoDB, Cassandra, Cosmos DB, Dynamo DB, Neptune, HBase
- **Visualization & Reporting:** Power BI, Tableau, SSRS, Excel, Cognos, SPSS
- **Testing, Monitoring & Automation:** Unit Testing, Cloud Watch, Azure Monitor, GCP Logging, Automated Testing, Alerting, Custom Dashboards, Atlan, SODA
- **Methodologies & Concepts:** OOP, Design Patterns, UML, Micro-services, Distributed Architecture, Erwin, Agile, Scrum, SDLC, Infrastructure as Code
- **ML & Data Science Libraries:** Pandas, NumPy, Scikit-learn, TensorFlow, Matplotlib, Seaborn, Scipy
- **Algorithms:** Decision Tree, Random Forest, Logistic Regression, Gradient Boosting, Support Vector Machine (SVM), K-Nearest Neighbor (KNN).
- **AI/ML:** SageMaker, Comprehend, PyTorch, TensorFlow, Huggingface, MLOps tools (Kubeflow, MLflow, Vertex AI Pipelines), Gen AI
- Deployment, Regulatory Data Handling (HIPAA, SOX)

PROFESSIONAL EXPERIENCE

Client: Walgreens **Deerfield, IL**
Role: Senior Data Scientist **January 2024 – Current**

- Designed end-to-end data pipelines using **Python**, **pandas**, **NumPy**, and **PySpark** integrated with **Spark MLlib**, ingesting clinical and pharmacy datasets from **AWS S3**, **Redshift**, and **RDS**, performing data cleaning, feature engineering, and transformations to feed machine learning models for healthcare predictions.
- Developed NLP pipelines for clinical documents using **spaCy**, **NLTK**, **regex**, **PyPDF2**, **pdfplumber**, and **textract**, extracting key entities and relationships from **FHIR** and **HL7** records, transforming unstructured text into structured datasets for predictive analytics and semantic search workflows.
- Built transformer-based models using **Hugging Face Transformers**, **T5**, **BERT**, **LLaMA**, **GPT-3.5**, and **GPT-4**, fine-tuning with **LoRA** and **PEFT**, to classify clinical text, predict patient outcomes, and automate knowledge extraction in healthcare documentation.
- Implemented semantic search and vector retrieval workflows using **OpenAI embeddings API**, **sentence-transformers**, **FAISS**, **Weaviate**, **Milvus**, and **Chroma**, integrating embeddings into **LangChain** pipelines for efficient clinical question-answering and decision support systems.
- Deployed machine learning models into production using **FastAPI**, **Triton Inference Server**, **Kubernetes**, **Docker**, and **SageMaker**, orchestrating CI/CD pipelines with **Jenkins** and **Git**, ensuring scalable inference, monitoring, and real-time patient analytics.
- Managed end-to-end workflow orchestration using **Apache Airflow**, **Prefect**, **Kubeflow**, and **SageMaker Pipelines**, scheduling data ingestion, preprocessing, model training, evaluation, and deployment while maintaining lineage, logging, and error handling for clinical ML pipelines.
- Ensured data quality and compliance using **Great Expectations**, **Deequ**, and custom validation scripts, monitoring healthcare data in **AWS Glue Catalog** and **Redshift**, detecting anomalies, validating transformations, and maintaining **HIPAA** standards across all datasets.
- Conducted exploratory data analysis and visualization using **Python**, **NumPy**, **SciPy**, **Matplotlib**, **Seaborn**, **R**, and **Tableau**, analyzing prescription trends, patient outcomes, and healthcare patterns to provide actionable insights for operational and clinical decision-making.
- Implemented scalable ETL pipelines using **AWS Glue**, **Lambda**, **Step Functions**, **SQS**, and **SNS**, extracting, transforming, and loading structured and unstructured clinical data into **S3** and **Redshift** for downstream analytics and predictive modeling.
- Built predictive models for readmission, demand forecasting, and patient outcomes using **Scikit-learn**, **XGBoost**, **Prophet**, **ARIMA**, and **Keras**, tuning hyperparameters with **Optuna**, and explaining model predictions using **SHAP** to improve clinical decision support.
- Conducted model tracking and experiment management using **MLflow**, **Weights & Biases**, and **EvidentlyAI**, capturing metrics, model versions, and performance drift for machine learning workflows deployed across healthcare datasets.
- Developed cloud pipelines for NLP extraction and embedding generation using **AWS S3**, **EC2**, **EKS**, **Lambda**, **SageMaker**, and **Bedrock**, ensuring secure, scalable, and reproducible workflows for large-scale healthcare data processing.
- Consolidated structured, semi-structured, and unstructured data using **Snowflake**, **BigQuery**, **AWS Redshift**, and **Apache Spark**, integrating multiple clinical and pharmacy sources into unified datasets for analytics and machine learning applications.
- Implemented real-time streaming pipelines using **Apache Kafka**, **Spark Structured Streaming**, **AWS CloudWatch**, and **Datadog**, enabling immediate processing of patient events, prescription transactions, and operational alerts for healthcare monitoring.

- Designed recommendation and personalization models for healthcare services using **TensorFlow**, **PyTorch**, **Keras**, and **BERT**, deploying predictive outputs through **FastAPI** endpoints to automate patient-specific clinical suggestions and interventions.
- Automated data validation, anomaly detection, and drift monitoring using **Great Expectations**, **EvidentlyAI**, and **custom Python scripts**, ensuring continuous data quality and integrity across ingestion, transformation, and model training pipelines.
- Built semantic search and retrieval pipelines for clinical documentation using **sentence-transformers**, **FAISS**, **Weaviate**, and **Milvus**, enabling physicians and pharmacists to quickly access relevant records and accelerate patient care decisions.
- Collaborated with teams using **Jira**, **Confluence**, and **Miro**, capturing requirements for predictive analytics, NLP, and reporting dashboards, documenting workflows, and ensuring clear communication for healthcare machine learning projects.
- Implemented CI/CD pipelines for healthcare ML applications using **Git**, **Jenkins**, **Docker**, and **Kubernetes**, enabling reproducible training, automated testing, and smooth deployment of predictive models into production environments.
- Designed analytics dashboards with **Tableau**, **Power BI**, and **AWS QuickSight**, integrating predictive insights and patient health trends to support executive-level decision-making and operational improvements in pharmacy services.
- Orchestrated LLM workflows using **LangChain** and **Haystack**, integrating multiple models (**GPT-4**, **Claude**, **Mistral**) to summarize medical literature, extract knowledge, and generate actionable insights for clinical teams.
- Managed secure cloud infrastructure using **AWS VPC**, **IAM**, **KMS**, **EKS**, and **Lambda**, ensuring protected storage of sensitive healthcare data and compliant deployment of machine learning pipelines.
- Conducted advanced feature engineering using **Python**, **pandas**, **NumPy**, **regex**, and **Spark MLlib**, preparing structured and unstructured healthcare data for predictive modeling and machine learning workflows.
- Performed deep learning experiments with **Transformers**, **T5**, **BERT**, **LLaMA**, and **GPT-3.5**, leveraging **PEFT** and **LoRA** for fine-tuning models tailored to clinical classification, NLP, and knowledge extraction tasks.
- Developed knowledge validation frameworks using **Collibra**, **EvidentlyAI**, and **custom scripts**, ensuring data governance, lineage tracking, and accurate, compliant predictions for healthcare ML applications.
- Designed document extraction pipelines using **PyPDF2**, **pdfplumber**, **textract**, and **Comprehend**, converting insurance claims, forms, and patient records into structured datasets ready for analysis and modeling.
- Optimized model training and inference using **Triton Inference Server**, **Kubeflow**, **SageMaker**, and **FastAPI**, reducing latency, scaling predictions, and improving clinical decision support responsiveness.
- Built analytics workflows using **Hadoop**, **Apache Hive**, **Spark MLlib**, and **AWS Glue**, processing massive healthcare claims and pharmacy datasets to generate predictive insights for operations and patient outcomes.
- Monitored production ML models using **MLflow**, **Weights & Biases**, and **EvidentlyAI**, detecting performance degradation, drift, and anomalies, triggering retraining pipelines, and ensuring continuous high-quality predictions for healthcare applications.

Environment: Python, pandas, NumPy, PySpark, Spark MLlib, AWS S3, Redshift, RDS, spaCy, NLTK, regex, PyPDF2, pdfplumber, textract, FHIR, HL7, Hugging Face Transformers, T5, BERT, LLaMA, GPT-3.5, GPT-4, LoRA, PEFT, OpenAI embeddings API, sentence-transformers, FAISS, Weaviate, Milvus, Chroma, LangChain, FastAPI, Triton Inference Server, Kubernetes, Docker, SageMaker, Jenkins, Git, Apache Airflow, Prefect, Kubeflow, SageMaker Pipelines, Great Expectations, Deequ, AWS Glue Catalog, SciPy, Matplotlib, Seaborn, R, Tableau, AWS Glue, Lambda, Step Functions, SQS, SNS, Scikit-learn, XGBoost, Prophet, ARIMA, Keras, Optuna, SHAP, MLflow, Weights & Biases, EvidentlyAI, EC2, EKS, Bedrock, Snowflake, BigQuery, Apache Spark, Apache Kafka, Spark Structured Streaming, AWS CloudWatch, Datadog, TensorFlow, Power BI, Haystack, Claude, Mistral, AWS VPC, IAM, KMS, Collibra, Comprehend, Hadoop, Apache Hive, HIPAA

Client: Costco Seattle, WA
Role: Sr. Senior Data Scientist March 2021 – December 2023

- Designed and implemented end-to-end ETL pipelines using **Python**, **PySpark**, **Apache Spark**, and **Azure Data Factory**, ingesting structured and semi-structured retail sales data from **Azure Data Lake** and **Snowflake**, performing data cleaning, transformation, feature engineering, and aggregations to enable downstream predictive analytics for inventory and demand forecasting.
- Built and deployed machine learning models using **PyTorch**, **Keras**, **Azure ML**, **LightGBM**, and **SARIMA**, training on historical sales, seasonal trends, and promotional data, validating performance, and deploying models into production to optimize pricing, inventory management, and sales forecasting workflows.
- Conducted exploratory data analysis and visualization using **Python**, **R**, **Seaborn**, **Matplotlib**, **Tableau**, and **Power BI**, analyzing retail KPIs, sales trends, and customer purchase behavior, generating actionable insights for merchandising, category management, and executive dashboards.
- Managed cloud-based infrastructure and orchestration using **Azure**, **Azure Synapse**, **Azure Synapse Analytics**, **Azure Databricks**, **Databricks**, and **Kubeflow**, automating end-to-end ML workflows, including ingestion, feature engineering, model training, and batch inference pipelines.
- Implemented scalable data pipelines using **Azure Data Factory**, **Azure Functions**, **Hadoop**, **HDFS**, **dbt**, and **Apache Beam**, ingesting multi-source retail datasets, applying transformations, validating data integrity, and loading outputs to **Snowflake** and **Azure Data Lake** for analytics.
- Optimized predictive model performance using **Hyperopt**, **LightGBM**, **PyTorch**, and **Keras**, performing hyperparameter tuning, cross-validation, and ensemble modeling to improve demand forecasting accuracy across seasonal and promotional campaigns.
- Monitored and tracked model performance using **Prometheus**, **WhyLabs**, and **Azure Monitor**, implementing alerts for drift, anomalies, and degradation, triggering retraining and pipeline adjustments to maintain reliable predictions for retail operations.
- Built advanced feature engineering workflows using **Python**, **Pandas**, **NumPy**, **dbt**, and **PySpark**, transforming transactional, inventory, and customer datasets into high-quality features suitable for regression, classification, and time-series modeling.
- Developed customer segmentation and recommendation systems using **PyTorch**, **LightGBM**, and **Keras**, integrating demographic, transactional, and loyalty program data to optimize targeted promotions, upselling strategies, and marketing campaigns.
- Implemented batch and near real-time data processing pipelines using **Azure Service Bus**, **Azure Functions**, **Apache Spark**, **PySpark**, and **Databricks**, ensuring timely ingestion, processing, and aggregation of point-of-sale, inventory, and supply chain data.
- Created reproducible ML workflows and CI/CD pipelines using **Git**, **GitHub Actions**, **Azure ML**, and **Kubeflow**, automating model training, testing, deployment, and monitoring for predictive analytics projects across retail datasets.
- Conducted statistical analysis and hypothesis testing using **Python**, **R**, **SciPy**, and **SQL**, validating assumptions on promotions, seasonal trends, and inventory movements, providing data-driven recommendations for business strategy optimization.
- Developed dashboards and reporting pipelines using **Power BI**, **Tableau**, **Python**, and **R**, integrating predictive insights and historical trends to provide retail leadership with actionable visualizations for decision-making.
- Consolidated structured and unstructured retail datasets from multiple sources using **Azure Data Lake**, **Snowflake**, **Azure Blob Storage**, and **Databricks**, ensuring clean, normalized, and consistent data for analytics and ML modeling pipelines.
- Automated anomaly detection and data validation workflows using **Python**, **dbt**, and **LIME**, identifying outliers in sales, inventory, and customer datasets, flagging irregularities, and notifying relevant teams for corrective action.

- Designed seasonal and promotional forecasting models using **SARIMA**, **LightGBM**, **PyTorch**, and **Keras**, incorporating historical sales, weather patterns, and holiday effects, enabling precise demand planning and inventory optimization.
- Performed time-series analysis and trend decomposition using **Darts**, **SARIMA**, **Python**, and **R**, integrating outputs into ML workflows to improve long-term forecasting and optimize supply chain planning across retail stores.
- Conducted data governance and lineage tracking using **Alation**, **dbt**, **Azure Synapse Analytics**, and **Azure**, maintaining consistent metadata, tracking transformations, and ensuring compliance with internal retail data standards.
- Monitored retail ML pipelines using **Prometheus**, **WhyLabs**, and **Grafana**, tracking performance metrics, pipeline throughput, and data quality, enabling proactive issue resolution and continuous improvement in data workflows.
- Built recommendation and personalization models using **PyTorch**, **LightGBM**, **Keras**, and **Python**, analyzing customer transactions and loyalty data to suggest targeted products and promotions, enhancing customer engagement and increasing basket size.
- Implemented version control and collaboration workflows using **Git**, **GitHub Actions**, **Jira**, and **Confluence**, ensuring reproducibility, proper documentation, and coordination across data science and engineering teams for all retail analytics projects.
- Designed data transformation pipelines using **dbt**, **Apache Beam**, **PySpark**, **Azure Data Factory**, and **Azure Functions**, processing large-scale transactional data for operational analytics, reporting, and predictive model inputs.
- Developed end-to-end predictive analytics pipelines integrating **Python**, **R**, **SciPy**, **Seaborn**, **Tableau**, and **Power BI**, generating insights for marketing campaigns, inventory optimization, and customer behavior analysis.
- Performed advanced statistical modeling using **LightGBM**, **SARIMA**, **PyTorch**, **Darts**, and **Keras**, identifying key sales drivers, forecasting demand, and recommending inventory allocations for retail store networks.
- Implemented cross-functional dashboards using **Tableau**, **Power BI**, and **Grafana**, integrating ML outputs, KPI tracking, and operational metrics, enabling executives and operations teams to make data-driven decisions across merchandising and supply chain.

Environment: Python, PySpark, Apache Spark, Azure Data Factory, Azure Data Lake, Snowflake, PyTorch, Keras, Azure ML, LightGBM, SARIMA, R, Seaborn, Matplotlib, Tableau, Power BI, Azure, Azure Synapse, Azure Synapse Analytics, Azure Databricks, Databricks, Kubeflow, Hadoop, HDFS, dbt, Apache Beam, Hyperopt, Prometheus, WhyLabs, Azure Monitor, Pandas, NumPy, Azure Service Bus, Git, GitHub Actions, SciPy, SQL, LIME, Darts, Alation, Grafana, Jira, Confluence, Azure Blob Storage

Client: Transunion **Chicago, IL**
Role: Data Scientist **May 2018 – February 2021**

- Designed and executed ETL pipelines using **GCP Dataflow** and **Apache Spark**, integrating diverse datasets from **GCP Pub/Sub** and **Cloud SQL**, ensuring high-quality, structured data delivery to **BigQuery** for downstream analytics and reporting.
- Developed advanced analytics models in **Python** and **PySpark**, applying machine learning techniques to detect anomalies in financial transactions, and operationalized models through **GCP Cloud Run** for automated real-time scoring.
- Implemented data orchestration workflows combining **GCP Dataproc**, **Hadoop**, and **Talend**, automating batch and streaming processes while ensuring compliance with banking data standards and efficient processing of large-scale datasets.
- Monitored and optimized cloud infrastructure using **gcp Stackdriver**, **Kubernetes**, and **GCP GKE**, maintaining system reliability, autoscaling configurations, and alerting mechanisms for real-time transaction and analytics pipelines.
- Conducted data quality checks and lineage tracking with **Apache Atlas**, **Monte Carlo**, and **Bigeye**, ensuring end-to-end visibility into critical banking datasets, while validating accuracy and integrity for regulatory reporting.

- Built robust CI/CD pipelines in **GitLab CI/CD** and **Docker**, integrating automated testing and deployment for **Python** and **PySpark** applications, enabling consistent updates to production banking analytics systems.
- Developed SQL-based data models in **BigQuery** and **Cloud SQL**, transforming raw transactional data into actionable insights and supporting dashboard creation for banking risk assessment and reporting.
- Designed metadata and data governance frameworks using **Power Designer** and **Apache Atlas**, establishing structured documentation of data assets, lineage, and compliance workflows for internal banking audits.
- Collaborated with cross-functional teams using **Confluence**, **Jira**, and **Trello**, documenting requirements, tracking progress, and ensuring seamless coordination of data pipeline development and model deployment projects.
- Leveraged **Looker** to develop interactive dashboards and visual analytics, integrating data from **Snowflake** and **BigQuery**, providing business stakeholders with real-time insights into credit scoring and risk metrics.
- Automated anomaly detection workflows using **Python** and **Apache Spark**, processing large-scale banking transaction logs ingested through **GCP Pub/Sub**, performing real-time analysis, and triggering alerts for suspicious, fraudulent, or high-risk activities to support proactive risk management in banking operations.
- Designed and deployed containerized data solutions using **Docker** and **GCP Cloud Run**, ensuring reproducible, scalable, and maintainable management of machine learning models, analytics applications, and real-time processing pipelines across banking data environments, simplifying deployment and reducing operational overhead.
- Maintained complete version control and collaboration of analytics code and data pipelines using **git** and **gitlab**, enabling seamless team development, rollback capabilities, and consistent delivery of data workflows, machine learning models, and banking analytics solutions in complex, multi-developer environments.
- Conducted performance tuning and query optimization in **BigQuery** and **Snowflake**, improving processing efficiency for massive banking datasets, reducing latency, accelerating reporting, enhancing data-driven decision-making, and ensuring rapid availability of insights for transactional, risk, and compliance analytics.
- Integrated batch and streaming data pipelines using **Talend**, **GCP Dataflow**, and **GCP Pub/Sub**, streamlining ingestion from multiple banking sources, ensuring reliable, near real-time data availability, and maintaining consistency, accuracy, and operational efficiency across critical financial data processing workflows.
- Monitored and maintained production data pipelines with **gcp Stackdriver**, **Kubernetes**, and **GCP GKE**, implementing automated alerting, proactive failure detection, and scaling solutions to minimize downtime, ensuring continuous availability and stability for mission-critical banking analytics and operational systems.
- Implemented predictive analytics and credit risk modeling using **Python**, **PySpark**, and **Apache Spark**, leveraging structured and unstructured banking datasets to generate accurate risk scores, forecast potential defaults, and provide actionable insights for portfolio management and strategic decision-making.
- Developed end-to-end workflow documentation and process maps using **Confluence**, **Jira**, and **Power Designer**, ensuring traceability, knowledge sharing, and regulatory compliance, while supporting cross-functional teams in understanding, auditing, and maintaining banking analytics pipelines and data governance practices.

Environment: GCP Dataflow, Apache Spark, GCP Pub/Sub, Cloud SQL, BigQuery, Python, PySpark, GCP Cloud Run, GCP Dataproc, Hadoop, Talend, gcp Stackdriver, Kubernetes, GCP GKE, Apache Atlas, Monte Carlo, Bigeye, GitLab CI/CD, Docker, SQL, Power Designer, Confluence, Jira, Trello, Looker, Snowflake, Git

Client: Honeywell **Charlotte, NC**

Role: Big Data Engineer **December 2016 – April 2018**

- Designed and implemented ETL pipelines using **Apache Spark** and **PySpark**, processing large-scale manufacturing sensor data stored in **Azure Blob Storage** and **Azure SQL Database**, transforming it for analytics, and loading the curated datasets into **Azure Synapse** for downstream reporting and operational insights.
- Developed and deployed predictive maintenance models using **Python** and **Apache Spark**, integrating historical machine performance data from **Azure SQL Database**, running computations on **Azure Virtual Machines**, and delivering real-time risk scores to optimize equipment uptime and reduce unplanned manufacturing downtime.
- Built automated dashboards and interactive visualizations in **Power BI**, connecting data from **Azure Synapse** and **Azure SQL Database**, enabling production managers and engineers to monitor equipment efficiency, track key metrics, and make informed decisions in near real-time.
- Orchestrated batch and streaming data workflows using **Apache Spark** and **PySpark** on **Azure Virtual Machines**, ingesting manufacturing IoT streams from **Azure Blob Storage**, ensuring reliable, scalable, and efficient processing for operational intelligence and predictive analytics.
- Implemented monitoring, logging, and alerting frameworks with **azure monitor** and **Azure SQL Database**, proactively tracking pipeline failures, optimizing performance, and maintaining system reliability across distributed manufacturing data pipelines and real-time analytics environments.
- Maintained CI/CD pipelines for data applications using **git** and **Jenkins**, managing version control, automated testing, and deployment of **Python** scripts and **Apache Spark** jobs to **Azure Virtual Machines**, ensuring consistent, reproducible releases of production-grade analytics workflows.
- Conducted data quality validation and performance optimization using **SQL** and **Azure SQL Database**, analyzing large manufacturing datasets, tuning queries, and ensuring accurate, efficient retrieval of operational and transactional data for reporting and predictive analytics.
- Collaborated with cross-functional teams using **Jira** and documentation standards, gathering requirements, tracking tasks, and coordinating deployment of **Python**, **Apache Spark**, and **Power BI** solutions, ensuring timely delivery of manufacturing insights and actionable analytics across business units.

Environment: Apache Spark, PySpark, Azure Blob Storage, Azure SQL Database, Azure Synapse, Python, Azure Virtual Machines, Power BI, azure monitor, git, Jenkins, SQL, Jira.

Client: Cipla **Hyderabad, India**

Role: ETL/Datawarehouse Developer **July 2015 – October 2016**

- Designed and implemented ETL pipelines using **Apache Spark** and **PySpark**, extracting healthcare datasets from **Amazon S3** and **Amazon RDS**, transforming data according to **HL7** and **FHIR** standards, and loading it into **Amazon Redshift** for downstream analytics and reporting.
- Developed complex SQL queries and data transformation scripts using **SQL** and **Python**, ensuring compliance with **HIPAA** regulations, processing large-scale healthcare datasets, and enabling accurate reporting and analytics for patient care and operational decision-making.
- Orchestrated batch and streaming data workflows using **Apache Spark**, **PySpark**, and **AWS EC2**, ingesting real-time and historical healthcare data from **Amazon S3**, ensuring scalable, reliable, and efficient processing for data warehouse operations.

- Built interactive dashboards and visualizations in **Tableau**, sourcing data from **Amazon Redshift** and **Amazon RDS**, providing real-time insights into patient outcomes, claims processing, and operational efficiency to healthcare stakeholders.
- Implemented monitoring, logging, and alerting frameworks using **aws cloudwatch** and **AWS EC2**, tracking ETL job failures, system performance, and ensuring high availability and reliability of healthcare data pipelines.
- Managed CI/CD deployments of ETL scripts and workflows using **Python** and **SQL**, integrating version control and automated testing, ensuring reproducible and consistent deployment of data warehouse jobs across **AWS** environments.
- Conducted data quality validation and performance optimization in **Amazon Redshift** and **Amazon RDS**, tuning SQL queries, validating transformed data, and ensuring timely, accurate availability of healthcare information for analytics and reporting.
- Collaborated with cross-functional teams to document ETL workflows and maintain compliance with **HIPAA**, **HL7**, and **FHIR** standards, ensuring proper handling, lineage tracking, and governance of sensitive healthcare datasets in the **AWS** ecosystem.

Environment: Apache Spark, PySpark, Amazon S3, Amazon RDS, HL7, FHIR, Amazon Redshift, SQL, Python, HIPAA, AWS EC2, Tableau, aws cloudwatch, AWS

EDUCATION DETAILS

- Rajiv Gandhi University of Knowledge Technologies, Rk valley — August 2011- May2015
  Bachelor of Technology in Civil Engineering — GPA: 8.58