**Ramakrishna Ainala**

**+18045054139**

**ramakrishna08us@gmail.com**

**Senior Data Engineer**

**LinkedIn URL**[**:** https://www.linkedin.com/in/ramakrishna-ainala-0a3403242](https://www.linkedin.com/in/ramakrishna-ainala-0a3403242)

**PROFESSIONAL SUMMARY:**

* With 10 years of robust experience in the data engineering field, I have developed a deep expertise in designing, implementing, and optimizing large-scale data systems and pipelines.
* Mastered Apache **Hadoop** and **AWS** **EMR**, leading to the development and management of scalable data processing infrastructures.
* Utilized Apache **Spark** and **Python** to perform advanced data transformations and analytics, enhancing data utilization.
* Implemented **MapReduce** jobs to handle large-scale data operations, improving data processing capabilities and efficiency.
* Managed Apache **Hive** databases, which involved schema design, data partitioning, and optimization tasks.
* Orchestrated complex data workflows using Apache NiFi, enabling more streamlined and efficient data movement and transformation.
* Configured and maintained Informatica workflows to automate and optimize **ETL** processes across various data sources.
* Deployed TensorFlow in cloud environments to build and train predictive models, improving the accuracy and speed of insights.
* Used **Python** to script automation tools and utilities, significantly reducing manual efforts and errors in data handling.
* Engineered real-time data streaming solutions with **Apache Kafka**, enhancing the timeliness and reliability of data availability.
* Managed **AWS** S3 buckets for secure and efficient data storage and retrieval in multiple high-availability environments.
* Designed and executed data pipelines using Terraform and **AWS** Glue, standardizing data processing patterns across projects.
* Utilized **SQL** extensively to perform complex data queries, manipulations, and aggregations, supporting critical decision-making processes.
* Applied **PyTorch** in machine learning projects to develop and deploy models that identified patterns and predicted outcomes.
* Automated data ingestion and processing tasks using Apache Airflow, which streamlined the setup of recurring data jobs.
* Integrated Databricks MLflow to manage machine learning lifecycle, including experimentation, reproducibility, and deployment.
* Leveraged **GIT** for version control, effectively managing code repositories for multiple projects in collaborative environments.
* Adopted agile methodologies to manage and deliver projects, enhancing team productivity and meeting tight deadlines.
* Created dataflow diagrams to visualize interactions between multiple internal divisions and ensure accurate data integration.
* Designed relational database schema and entity relationship diagrams (ERDs) to support the new CASE system using Microsoft Dataverse.
* Developed a Common Data Model (CDM) aligned with MS Dynamics 365 standards to support scalable, compliant, and cross-divisional data management.
* Produced interactive data visualizations and reports using Tableau, facilitating better insights into business metrics and trends.
* Implemented **AWS** DynamoDB for high-performance No**SQL** database solutions, supporting web-scale applications.
* Enhanced data quality and integration using **AWS** QuickSight for real-time analytics and business intelligence capabilities.
* Developed on-prem cloud solutions to maintain data security and compliance within regulated industries.
* Scripted complex data transformations and analytics tasks using R, enriching data sets for predictive and statistical analysis.
* Configured and managed orchestration layers using **GIT**Lab, improving the automation of CI/CD pipelines and workflows.
* Experienced in designing and deploying ETL solutions using Informatica IICS, enabling seamless cloud-based data integration and transformation.
* Delivered comprehensive documentation and training to support data governance, usability, and team onboarding processes.

**TECHNICAL SKILLS:**

* **Big Data Processing : Hadoop**, **Spark**, **MapReduce**, **Hive**, **Kafka**, **EMR**, **AWS** Glue
* **Operating systems: Linux**,Unix,VMware
* **Data Integratio :** Apache NiFi, Informatica **IICS**, Data Pipeline, Dagster, Terraform
* **Machine Learning Ops :** TensorFlow, **PyTorch**, Databricks MLflow
* **Web Technologies:** JSON, AJAX, Node.js, Angular, React, Deck.gl, Bootstrap
* **Cloud Services : AWS** S3, **AWS** DynamoDB, **AWS** **EMR**, On-prem Cloud Premises
* **Analytic Tools :** Tableau, Power BI, Google Analytics Fiddler, tJMP PRO, SAS, **Azure**,SSIS
* **Programming : Python** 2.5,2.7,3.4,3.6, **SQL**/PL,C,C++,OpenGL,webGL
* **Python Libraries/Packages** : NumPy, SciPy, Pickle, PyQt, PySide, PyTables, Data Frames, Pandas, Matplotlib, **SQL**Alchemy, HTTPLib2, Urllib2, Beautiful Soup, PyQuery
* **IDE**: PyCharm, Spyder, PyDev
* **Databases/Servers** :My**SQL**, MS Access, Powerpoint **SQL** Server, **Oracle**, **SQL**ite3,Snowflake, Cassandra, Redis, Postgre**SQL**, CouchDB, MongoDB, Apache Web Server 2.0, NginX, Tomcat, JBoss,WebLogic
* **Version Control & CI : GIT**, Jira
* **Project Management :** Agile Methodologies

**PROFESSIONAL EXPERIENCE:**

**Client: USAA, San Antonio, TX Dec 2023 to till date**

**Role: Senior Data Engineer Roles & Responsibilities:**

* Developed end-to-end data pipeline solutions integrating Apache **Hadoop** and **AWS** **EMR**, enhancing data processing capabilities.
* Configured **AWS** S3 for efficient data storage and management, improving access and retrieval times.
* Utilized TensorFlow within **AWS** environments to build and train machine learning models, boosting predictive analytics performance.
* Managed the automation of data flows using Apache NiFi, streamlining data ingestion and integration processes.
* Implemented **AWS** Glue for **ETL** processes, which improved data transformation and loading efficiency.
* Engineered data orchestration workflows with Databricks MLflow, facilitating machine learning lifecycle management.
* Designed and maintained cloud-native ETL pipelines using **Informatica IICS**, improving scalability and reducing latency in data workflows.
* Applied **Python** extensively to automate data processing tasks and develop backend utilities, increasing operational efficiency.
* Conducted **SQL** queries for data manipulation and reporting, supporting critical analytics and decision-making processes.
* Led a team using agile methodologies, ensuring timely delivery of data engineering projects and milestones.
* Integrated real-time data streaming capabilities with Apache **Kafka**, improving data availability and system responsiveness.
* Orchestrated comprehensive data source integration projects, ensuring robust data collection and aggregation.
* Enhanced business intelligence initiatives by implementing end-to-end solutions from data collection to visualization.
* Responsible for Architect,build and optimization of data pipelines and workflows to support real-time and batch processing in a cloud-based SaaS environment.
* Designed and scheduled **Dagster** jobs to manage dependencies across Snowflake, S3, and Cassandra-based ingestion pipelines.
* Integrated **Dagster** as the orchestration layer for critical batch data pipelines, enabling modular, testable workflows and improving overall pipeline observability.
* Integrated S3 with **Dagster** pipelines for asset tracking, metadata logging, and resilient data handoffs across stages.
* Designed cross-system data synchronization strategies to ensure consistency between Cassandra (for real-time applications) and Snowflake (for analytical workloads).
* Built metadata-driven pipeline components to dynamically handle schema changes between semi-structured data in Data Lake and structured Snowflake tables.
* Implemented failover and recovery mechanisms within data pipelines to handle latency-sensitive writes in Cassandra without disrupting downstream Snowflake ingestion.
* Implemented data governance practices, including access controls, data lineage and compliance with regulatory requirements like GDPR and HIPAA.
* Responsible for designing and implementing data pipelines leveraging **Azure** Storage Accounts, including **Azure** Data Lake Storage (ADLS), Blob Storage, and Table Storage.
* Deployed and managed **Azure** Static Web Applications for hosting and integrating data-driven front-end services and also implemented secure and scalable authentication workflows using **Azure** Service Principals and OAuth via Microsoft Entra ID.
* Collaborated with business teams to identify data needs and translated these requirements into technical solutions.
* Designed and implemented distributed data processing pipelines using Py**Spark** to handle large-scale, structured, and unstructured datasets.
* Monitored and maintained data systems to ensure high performance and availability under high load conditions.
* Documented all data engineering processes and solutions, providing clear guidelines and support for system users.
* Used tools like Collibra and Informatica Axon to support data governance initiatives, while maintaining a focus on integrating privacy and security considerations.
* Conducted data quality checks and validation, ensuring the accuracy and reliability of reports and analytics.
* Developed security protocols for data handling and storage, adhering to industry standards and compliance requirements.
* Utilized JIRA for project management and tracking, keeping all team members aligned and focused on deliverables.
* Trained junior engineers and team members on new technologies and best practices, fostering a culture of continuous learning.
* Optimized data retrieval and processing speeds by fine-tuning existing database and pipeline configurations.
* Participated in peer reviews and code audits to maintain high standards of code quality and system functionality.
* Delivered presentations and reports to stakeholders, effectively communicating the impact and value of datadriven projects.

**Environment:**  Apache **Hadoop**, **AWS** **EMR**, **AWS** S3, Dagster , TensorFlow, Apache NiFi, **AWS** Glue, Databricks MLflow, **Python**, **SQL**, Docker, Apache **Kafka**, JIRA.

**Client: Duke health,** **Cary, NC Aug 2021 to Dec 2023**

**Role: Data Engineer Roles & Responsibilities:**

* Engineered and maintained complex **ETL** pipelines using **AWS** S3 and **AWS** Glue, significantly improving data processing and management.
* Implemented real-time data processing frameworks with Apache NiFi, which enhanced the timeliness and accuracy of clinical data.
* Utilized Terraform to automate the provisioning of cloud infrastructure, streamlining the setup and maintenance of data pipelines.
* Utilized **Deck.gl** to develop interactive, geospatial visualizations of patient admission patterns, enabling healthcare analysts to identify regional disparities in access to care and emergency response trends.
* Integrated **Deck.gl** visual layers with **React-based dashboards** to support real-time monitoring of disease spread and hospital resource allocation across multiple geographic zones.
* Deployed Docker containers to ensure consistency across development, testing, and production environments, reducing deployment issues.
* Developed machine learning models using **PyTorch**, which facilitated advanced health data analytics and prediction tasks.
* Integrated Informatica workflows to streamline data transformations and enrichments, improving data quality and utility.
* Managed data storage solutions using **AWS** DynamoDB, providing scalable and high-performance data access for critical applications.
* Created interactive dashboards and reports with Tableau, offering actionable insights to healthcare professionals and decision-makers.
* Automated various data operations using **Python** scripts, which increased efficiency and reduced manual processing time.
* Responsible for designing,developing and maintain scalable **ETL** pipelines to process and prepare large datasets for training and inference of AI(Artificial intelligence) models.
* Developed scalable ETL workflows using **Informatica IICS** to integrate healthcare datasets across AWS and on-prem environments.
* Responsible for monitoring and maintaining AI model performance, implementing feedback loops to improve prediction accuracy and reliability.
* Configured **Azure** Key Vaults to manage secrets, keys, and certificates, ensuring secure access to data and applications.
* Configured **AWS** QuickSight for real-time analytics and visualization, enhancing data-driven decision-making within the health domain.
* Monitored and optimized data pipelines to ensure efficient data flows and system performance under peak loads.
* Conducted comprehensive data quality assessments using **SQL**, ensuring the integrity and accuracy of health records and reports.
* Implemented error handling, logging, and performance tuning in SSIS packages to ensure data integrity and processing efficiency.
* Designed and deployed **ETL** packages using **SQL** Server Integration Services (SSIS) to automate data extraction, transformation, and loading from diverse sources.
* Orchestrated data migration projects from legacy systems to cloud-based platforms, minimizing downtime and data loss.
* Applied **GIT** for version control, effectively managing project codebases and collaboration among the development team.
* Employed agile project management techniques using JIRA, ensuring timely delivery of project milestones and adherence to schedules.
* Developed data integration strategies to incorporate disparate data sources into a unified analytics framework.
* Led training sessions for team members on new technologies and data handling techniques, enhancing team skills and capabilities.
* Documented all technical processes and configurations, creating comprehensive manuals and guides for system users.
* Coordinated with cross-functional teams to align data engineering activities with broader organizational goals and patient care objectives.
* Established data governance and compliance protocols, adhering to industry regulations and best practices in data security.
* Participated in technical reviews and audits, ensuring adherence to architectural and design standards.
* Delivered end-to-end solutions that addressed the specific needs of healthcare data management and analysis, driving better health outcomes.

**Environment:** **AWS** S3, **AWS** Glue, Apache NiFi, Terraform, Docker, **PyTorch**, Informatica, **AWS** DynamoDB, Tableau, **AWS** QuickSight, **SQL**, **GIT**, JIRA.

**Client: PNC Bank, Pittsburgh, PA Dec 2019 to Aug 2021**

**Role: Big Data Engineer Roles & Responsibilities:**

* Developed data integration pipelines using Apache Flink to optimize real-time data processing, enhancing system responsiveness and scalability.
* Engineered complex **ETL** workflows in Informatica, significantly improving data aggregation accuracy and efficiency for critical business reports.
* Migrated legacy on-prem ETL processes to **Informatica IICS**, enhancing pipeline scalability, performance, and reducing deployment time in cloud environments.
* Utilized **Azure** Synapse Analytics to expand and optimize data warehousing capabilities, facilitating large-scale analytics across diverse datasets.
* Enhanced data operations with DBT, enabling scalable transformations and batch processing, which improved data handling for large datasets.
* Implemented Kubernetes to deploy and manage containerized applications, which increased system reliability and streamlined deployment processes.
* Orchestrated efficient data flows into Snowflake, improving data querying and analytics performance, which supported advanced data science initiatives.
* Used Snowflake external tables to minimize data movement from Data Lake while enabling immediate querying of raw datasets.
* Created a Dataverse-compliant Common Data Model to align with Microsoft Dynamics 365 platform.
* Developed security layers to segregate sensitive data when transferring between Cassandra and Snowflake via intermediate lake zones.
* Leveraged collaborative analytics techniques with Power BI to provide actionable insights to business stakeholders, enhancing decision-making processes.
* Architected robust data solutions using **Python** and Scala, supporting both batch and streaming data, which facilitated real-time analytics platforms.
* Configured and maintained **Azure** DevOps for continuous integration and continuous delivery pipelines, which enhanced development practices.
* Assisted in monitoring and mitigating data risks through the development of processes that ensure data quality, integrity, and security.
* Responsible for comprehensive data governance and compliance measures with Informatica, ensuring data integrity across all platforms.
* Collaborated with stakeholders to create and refine data governance policies, ensuring alignment with privacy, cybersecurity, and compliance requirements.
* Optimized data models and architectures to support evolving business needs with Scala and **Python**, improving response to market changes.
* Developed, tested, and deployed data pipelines that adhered to best practices in code quality and security with **GIT**, enhancing system robustness.
* Integrated data from various sources into **GCP** using Cloud Pub/Sub, Cloud Storage.
* Migrated on-premises databases and data warehouses to **GCP** using Dataproc, Data Fusion, and Database Migration Service.
* Analyzed and processed structured and unstructured data sets using **SQL** and **Python**, supporting data-driven decisions that improved business strategies.
* Conducted rigorous testing and quality assurance with **Python** and Scala to ensure the robustness of data products, reducing operational risks.
* Implemented disaster recovery and data backup procedures with **AWS** S3, ensuring data availability and integrity across business operations..
* Researched emerging technologies to keep data solutions at the forefront of technology with ongoing evaluations and updates.
* Established best practices for data security and compliance across all platforms with **Azure** and Informatica, ensuring robust data protection.
* Spearheaded initiatives to streamline data processing and reduce latency with Apache Flink, directly impacting system efficiency.
* Championed the adoption of cloud technologies with **Azure**, driving cost efficiencies and scalability in data operations.
* Engaged in continuous learning and development programs with Power BI and Informatica, keeping skills and knowledge up to date.

**Environment**: Apache Flink, Informatica, **Azure** Synapse Analytics, DBT, Kubernetes, Snowflake, Power BI, **Python**, Scala, **SQL**, **Azure** DevOps, **GIT**, **AWS** S3, Apache **Spark.**

**Client: Juspay Technologies, Bangalore, India May 2017 to Sep 2019 Role: Data Visualization Analyst Roles& Responsibilities:**

* Managed Microsoft **SQL** Server databases, enhancing data storage and retrieval operations for real-time transaction processing.
* Utilized **Python** to perform advanced data analytics, optimizing financial transaction monitoring and fraud detection.
* Employed R for statistical analysis and modeling, providing deeper insights into customer behavior patterns.
* Developed cloud-based architectures on on-prem cloud premises, ensuring robust and scalable data processing capabilities.
* Created and maintained data documentation, standardizing data handling procedures across the organization.
* Implemented Power BI for dynamic data visualization, offering actionable insights to business users.
* Automated data sourcing and integration using **SQL** scripts, streamlining data flows for analytics applications.
* Conducted regular data audits using **SQL**, maintaining high standards of data accuracy and compliance.
* Optimized data queries and models for improved performance and faster access to business intelligence.
* Collaborated with IT and business teams to align data strategies with organizational objectives.
* Facilitated data-driven decision-making by providing timely and relevant data reports to stakeholders.
* Enhanced data security and compliance by implementing rigorous data protection measures.
* Trained junior analysts and developers on data management best practices and technologies.
* Participated in technology upgrade projects, ensuring modernization of legacy data systems.
* Provided technical support for data-related issues, improving system reliability and user satisfaction.
* Advocated for the use of advanced analytics and data science techniques within the company.
* Developed disaster recovery plans for data systems, ensuring business continuity in emergencies.
* Led workshops and sessions on the use of data visualization tools, increasing organizational data literacy. **Environment:** Microsoft **SQL** Server, **Python**, R, Power BI, **SQL**, and data documentation tools.

**Client: Smiths Group, Bengaluru, India Nov 2015 to Apr 2017**

**Role: Spark Engineer**

**Roles & Responsibilities:**

* Designed and implemented large-scale data processing jobs using HDFS and **MapReduce**, improving data throughput and scalability.
* Orchestrated data workflows with Apache **Hive** and Oozie, enhancing automation and efficiency of data operations.
* Integrated real-time data streaming capabilities with **Kafka**, supporting immediate data availability and processing.
* Utilized Apache **Spark** for complex data analysis tasks, driving faster insights into manufacturing processes.
* Scripted data transformation and aggregation tasks using **Python**, simplifying data preparation for analytics.
* Managed database operations and queries using **SQL**, ensuring accurate and timely data for reports.
* Conducted performance tuning on **Spark** and **Hadoop** systems, maximizing resource utilization and processing speed.
* Developed data ingestion mechanisms with Apache NiFi, streamlining data acquisition from various sources.
* Implemented data security protocols, ensuring compliance with industry standards and protection of sensitive information.
* Collaborated with engineering teams to design data models that support product development and quality assurance.
* Trained new team members on big data technologies and best practices, enhancing team capabilities and knowledge.
* Monitored data systems for performance issues and anomalies, maintaining high availability and reliability.
* Documented system configurations and procedures, providing a reference for maintenance and future upgrades.
* Evaluated new technologies and tools for potential adoption to enhance the data architecture.
* Developed **ETL** pipelines to facilitate data collection, transformation, and storage processes.

**Environment:** HDFS, **MapReduce**, Apache **Hive**, Oozie, **Kafka**, Apache **Spark**, **Python**, **SQL**, Apache NiFi.