**Sai Kaushik Mathe**

**DATA ENGINEER**

**(216) 423-7973 |** [**s.mathe3107@gmail.com**](mailto:s.mathe3107@gmail.com) **| OH |** [**LinkedIn**](https://www.linkedin.com/in/sai-kaushik-mathe-066a95251/)

# SUMMARY

* Data engineer with five years of experience in designing, implementing, and optimizing data solutions to support business objectives. Proficient in leveraging a diverse set of technologies and tools to build scalable and efficient data pipelines, enabling data-driven decision-making and business insights. Skilled in managing and processing large volumes of structured and unstructured data, with a strong focus on data quality, reliability, and performance. Demonstrated expertise in cloud platforms such as Azure, AWS, and Snowflake, as well as proficiency in technologies like Python (NumPy, Pandas, Matplotlib, SciPy, Scikit-learn, Seaborn), SQL, Apache Airflow, Azure Synapse and Scala. Proven ability to collaborate effectively with cross-functional teams to deliver impactful data solutions and drive organizational success.

# SKILLS

* + **Programming Languages**: SQL, Python, R, Java, Unix, Pyspark, Scala
  + **Big Data Ecosystems**: Hadoop, MapReduce, Hive, Spark
  + **Cloud Platforms**: AWS (EMR, EC2, S3, Glue, Redshift), Azure (ADF, Databricks, Synapse), Google Cloud
  + **ETL/DWH Tools**: Tableau Prep, Alteryx, Informatica Power Center, Data Stage, SSIS, Ab Initio
  + **Visualization Tools**: Tableau, Power BI, Tibco Spotfire
  + **Database Systems**: MS SQL Server, Teradata, Oracle, PostgreSQL, MongoDB
  + **Methodologies**: Agile, Scrum, Waterfall, SAFe
  + **Machine Learning**: Regression, Decision Trees, Random Forests, SVM, Neural Networks, NLP
  + **Version Control**: Git, GitHub, SVN, Jira

# EDUCATION

* **Masters in computer and information science | Cleveland State University *Dec 2023***
* **Bachelor of Engineering in Computer Science | Jawaharlal Nehru Technological University *May 2018***

# EXPERIENCE

**Carelon, Indianapolis, CT | Azure Data Engineer Jan *2023 – Mar 2024***

* Developed smart pipelines using Pyspark and Azure Databricks, reducing data processing time by 30%.
* Created real-time data streaming architectures with Azure Event Hubs and Apache Kafka, improving data freshness by 25%.
* Developed and deployed high-performance ETL pipelines using PySpark and Azure data factory, handling complex data transformations and load operations.
* Utilized Azure Events Hubs and Apache Kafka for real time data ingestion and streaming analytics, resulting in 30% enhancement in data freshness and availability.
* Implemented advanced performance tuning techniques in ETL pipelines, reducing data processing time by over 25% through optimization of Spark jobs and resource allocation.
* Designed a scalable data lake architecture with Cosmos DB, facilitating efficient storage and high-performance querying of large health datasets, reducing database management time by 35%.
* Developed and integrated resume points in Stream Sets, minimizing data processing downtime by X% and increasing pipeline reliability and fault tolerance in handling health data.
* Engaged with data analysts and health department stakeholders to gather requirements and design tailored data solutions within Palantir Foundry, enabling advanced health analytics and reporting.
* Collaborated with Quality Engineering to define and execute rigorous testing protocols, utilizing unit tests, integration tests, and data validation checks to ensure accuracy and reliability of health data pipelines.
* Ensured compliance with HIPAA regulations by implementing robust data privacy and security measures, including data anonymization, encryption, and access control specific to health data.
* Automated ETL processes using Airflow, enhancing data pipeline reliability and reducing downtime by 20%.
* Designed and deployed interactive dashboards in Power BI, improving stakeholder decision-making efficiency.

**Humana, Louisville, KY | Data Engineer Feb *2022 – Dec 2022***

* Actively participated in the analysis, design, and development phases of the SDLC, employing agile methodologies. Utilized tools such as JIRA for task management, sprint planning, and issue tracking, and GitHub for version control to ensure efficient collaboration and code management.
* Developed Spark applications utilizing AWS EMR for distributed data processing, enabling scalable and high-performance data transformations and analytics.
* Designed, implemented, and optimized cloud-based data pipelines using AWS services, including Amazon S3 for data staging and AWS Redshift for large-scale data migration and analytics.
* Implemented custom data validation scripts and automated tests to ensure data integrity and consistency throughout the data processing pipeline.
* Implemented infrastructure as code (IaC) principles using AWS CloudFormation or Terraform for automated provisioning of resources, reducing deployment time and ensuring consistency across environments.
* Utilized Stream Sets' advanced features such as schema evolution and error handling to build resilient and adaptable data pipelines.
* Designed and implemented end-to-end data pipelines using Stream Sets Data Collector, leveraging its intuitive UI and extensive library of connectors for efficient data ingestion, transformation, and routing.
* Established and optimized CI/CD pipelines for multiple data projects using AWS DevOps services like AWS Code Pipeline, AWS Code Build, and AWS Code Deploy.
* Conducted technical workshops and knowledge-sharing sessions to disseminate best practices and promote a culture of technical excellence, resulting in a 95% satisfaction rate among stakeholders.
* Conducted comprehensive code reviews for data engineering codebases hosted on AWS, ensuring adherence to best practices, performance optimization, and maintainability.
* Designed, implemented, and maintained Apache Airflow Directed Acyclic Graphs (DAGs) to orchestrate complex ETL (Extract, Transform, Load) workflows and data pipelines.

**Sanus Software Solutions Pvt. Ltd, India | Data Engineer *Jul 2017 - Sept 2021***

* Migrated data to cloud using AWS and Azure, ensuring 99.9% data accessibility and reliability.
* Created Spark applications for data extraction and transformation, improving data processing speed by 25%.
* Developed Power BI dashboards, providing actionable insights to stakeholders.
* Managed cloud-based data pipelines using Azure Data Factory, enhancing data integration processes.
* Developed ETL mappings in Informatica Power Center, optimizing data transformation processes.
* Created SAP Business Objects dashboards, enhancing data visualization and reporting capabilities.

# PROJECTS

1. **Sentinel Health: Disease Surveillance System**

* Implementation of distributed data processing frameworks like Apache Spark for scalable analysis of large volumes of health data in real-time.
* Integration of machine learning models for anomaly detection and pattern recognition within the data pipeline to automate outbreak identification.
* Utilization of cloud-based infrastructure for elasticity and cost-effectiveness in managing fluctuating data loads and computational demands.

1. **Cloud Data Pipeline Optimization**

* Implementation of distributed data processing frameworks such as Apache Spark on AWS EMR to enable scalable analysis of large volumes of data, ensuring efficient processing and real-time insights.
* Utilized Infrastructure as Code (IaC) principles with AWS CloudFormation and Terraform to automate resource provisioning, ensuring consistent deployment across environments and reducing manual configuration overhead.
* Designed and managed Apache Airflow Directed Acyclic Graphs (DAGs) to orchestrate complex ETL workflows, providing robust pipeline orchestration capabilities for seamless data flow management and task scheduling.

# CERTIFICATIONS

* Azure Fundamentals (AZ-900)
* Microsoft and LinkedIn Career Essentials in Software Development
* Career Essentials in Data Analysis by Microsoft and LinkedIn
* Oracle Certified Junior Associate (OCAJ)
* CCNA Routing and Switching Course by CISCO