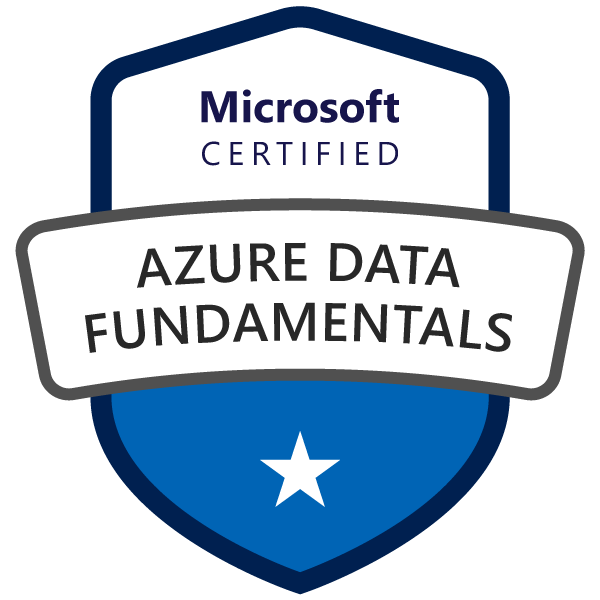
**SAI PRANEETH KOMMU**

**Data Engineer**  
+1 4694434077 | [praneeth07072000@gmail.com](mailto:praneeth07072000@gmail.com) | [Linkedin](https://www.linkedin.com/in/kommusaipraneeth/)

**PROFESSIONAL SUMMARY**

* Over **6 Years of Experience** in designing and developing robust **Data Pipelines** using **Python**, **PySpark**, and **SQL** to support large-scale data processing across cloud and on-premise environments for various industry domains.
* Proficient in big **data technologies** including **Apache Spark**, **Apache **Kafka****, and the **Hadoop Ecosystem (HDFS, Hive, MapReduce)** for efficiently processing, analyzing, and transforming high-volume, high-variety datasets.
* Expert in implementation of Azure Data solutions, **ETL/ELT data migration** solutions using **Azure Databricks**, **Azure Data Factory, Azure Data Lake**, hybrid Linux / Windows and Structured Query Language (SQL) databases.
* Skilled in deploying and managing cloud-based data platforms on **AWS**, **Azure**, and **Google Cloud Platform (GCP)**, leveraging services like **AWS Glue**, **Azure Synapse Analytics,** and Google BigQueryfor scalable storage and compute.
* Proficient in programming with Python including **Big Data technologies** like Apache Spark, Hadoop, Hive.
* Led projects involving the installation, configuration, patching, and upgrading of **Microsoft SQL Server** (2014, 2017, 2019, 2022), including **data migration** between versions.
* Extensive experience in **containerization** and **orchestration** using **Docker** and **Kubernetes** for packaging and deploying data processing applications based on microservices architecture in distributed environments.
* Experienced in designing and implementing secure **RESTful APIs** using **OAuth 2.0** and **JWT** standards.
* Deep understanding of **data warehousing** principles and expertise building optimized data warehouses using **Amazon **Redshift**, **Azure Synapse Analytics**** and **Google **BigQuery**** for business intelligence (BI) and reporting.
* Skilled in developing real-time streaming data solutions using **Azure Event Hub**, **Apache Kafka**, **Apache** **Spark Streaming** and **Google Pub/Sub** to enable low-latency data analytics, streaming analytics, and operational insights.
* Expertise in building complex **Extract, Transform, Load (ETL)** packages using tools like Azure Data Factory, SSIS, merging varied data sources into a unified data model.
* Knowledgeable in implementing **Role-Based Access Control (RBAC)**, data acquisition, **data governance**, data profiling, **data quality** and frameworks such as **Great Expectations** and **Apache Ranger** to ensure compliance and accuracy across Microsoft Fabric data pipelines.
* Experienced in setting up **Continuous Integration and Continuous Deployment** (**CI/CD) pipelines** for data projects using **Jenkins**, **Azure DevOps**, and **Git**, enabling automated testing, version control, and streamlined deployments.
* Worked with Avro, **Parquet**, and semi-structured **JSON** data; converted formats using **PySpark DataFrames**.
* Expert in **OLTP** and **OLAP** data modeling, **SQL performance tuning,** and database normalization.
* Collaborative team player, adept at working closely with data scientists, business analysts, and stakeholders to translate complex business requirements into scalable data engineering solutions.
* Strong analytical skills in optimizing **SQL queries** and data workflows to improve performance, reduce cost, and accelerate report generation.
* Hands-on experience across the complete **Software Development Life Cycle (SDLC)** using **Agile** and hybrid methodologies, with active contributions to sprint planning, retrospectives, and backlog grooming.
* Excellent stakeholder communication skills, capable of presenting technical concepts to non-technical stakeholders.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **Category** | **Skills** |
| Programming Languages | Python, SQL, PySpark, Scala, T-SQL, R, Java, Shell Scripting, Powershell |
| Big Data & ETL Tools | Apache Spark, Azure Events Hub, Apache Kafka, Hadoop (HDFS, MapReduce, Hive, HBase), Spark Streaming, Airflow, DBT, Informatica |
| Cloud & Data | Microsoft Fabric, Azure Data Factory, Azure Databricks, Azure Synapse Analytics, HDInsight, AWS, S3, Redshift, Glue, CloudWatch, Google Cloud, BigQuery, Snowflake |
| Data Warehousing & Databases | Azure Data Lake Storage, ADLS, AWS Redshift, Google BigQuery Azure SQL, Azure Cosmos DB, MS SQL Server, PostgreSQL, MySQL, MongoDB |
| Data Modelling | Star & Snowflake Schema, OLTP, OLAP, Data Management, Data Governance |
| Data Visualization & Analysis | Microsoft Excel, Microsoft Power BI, Tableau, Statistical Analysis |
| API & Microservices | REST API, JWT/OAuth2 Authentication, API Gateways, Microservices Architecture |
| Data Administration | Performance Tuning, Backup & Recovery, High Availability, Replication, Clustering, SQL |
| Monitoring & DevOps | CI/CD, Github Actions, Azure Monitor, Docker, Jenkins, Prometheus, Grafana, Agile |
| Soft Skills | Communication, Cross-Functional Collaboration, Analytical Thinking, Business Acumen |

**PROFESSIONAL EXPERIENCE**

Client: **Bank of America**

**Senior Data Engineer | May 2024 – Present**

* Designed and implemented scalable **data ingestion** pipelines using **Apache Kafka**, **Python**, and **Spark Streaming** to process millions of financial transactions, enabling real-time availability and low-latency processing for fraud detection.
* Developed complex **batch** and **streaming ETL workflows** with **PySpark,** orchestrated them using **Apache Airflow** and **dbt** for automated scheduling, error handling, and alerting, improving pipeline reliability and minimizing downtime.
* Designing and developing **Azure stream analytics** jobs to process real time data using **Azure Event Hubs**.
* Enhanced the data lake architecture by leveraging **Azure Data Lake Storage Gen2** storage with lifecycle policies and data tiering, reducing storage costs and improving data accessibility for analytics teams.
* Led migration of **100+ legacy Hadoop jobs** to Azure-native platforms (**ADF, Databricks, HDInsight**), improving scalability by **50%**, reducing operational overhead by **30%**, and cutting maintenance hours by 15/week.
* Built containerized microservices for data processing using **Docker,** and deployed them on **Kubernetes clusters to** enabling scalable, flexible data transformation aligned with cloud-native architecture principles.
* Integrated machine learning models into production machine learning pipelines using **MLflow** for model tracking and deployment, enabling predictive analytics for credit risk assessment and customer segmentation.
* Developed and maintained secure, high-performance **RESTful APIs** with **OAuth 2.0** authentication to expose processed data for real-time analytics and reporting across internal applications and authorized external clients.
* Optimized **SQL** & **PySpark** datasets, reducing report load times by **40%** and improving **Power BI dashboard** refresh rates by **35%**, supporting timely financial reporting.
* Automated **continuous integration** and delivery of data pipelines using **Jenkins** and managed source control with **Git**, facilitating faster deployments and reducing manual errors in the production environment.
* Collaborated closely with business analysts and data scientists to translate complex financial reporting requirements into optimized **SQL** queries and **PySpark** scripts for analytical dashboards.
* Implemented automated data validation and quality checks using **Great Expectations**, ensuring Data Integrity and Governance and consistency across multi-stage data pipelines that support critical banking applications.
* Delivered knowledge transfer sessions on cloud-native data engineering practices, container orchestration, and pipeline optimization, improving team skill sets and accelerating project delivery timelines.
* Mentored junior engineers in developing efficient **Spark** jobs, debugging complex data processing issues, and adopting DevOps best practices, resulting in improved overall team productivity and quality of deliverables.

Client: **Cholamandalam MS General Insurance**

**Data Engineer | Aug 2021 – Aug 2023**

* Designed and maintained scalable **ETL/ELT** pipelines using **Python**, **Apache Spark**, and **SQL** to efficiently process high-volume insurance claims data, enabling accurate and timely reporting for operational decision-making.
* Built real-time streaming solutions with **Apache Kafka** and **Spark Streaming**, reducing data latency by **60%** and improving fraud detection response times by **40%** in claims workflows.
* Architected and optimized cloud-based **data warehousing** solutions using **Azure Synapse Analytics**, enabling fast and cost-effective querying of large datasets for business intelligence and analytics teams.
* Automated infrastructure provisioning and data pipeline deployments using **Terraform** and **Azure DevOps CI/CD**, resulting in consistent, repeatable deployments and reducing manual configuration errors.
* Implemented strict data security and compliance controls using **Azure IAM** and **Azure Purview APIs** to protect sensitive insurance data and maintain regulatory compliance.
* Containerized data ingestion and transformation microservices using **Docker**, orchestrating deployments with Kubernetes clusters to ensure scalability, high availability, and ease of management.
* Building **KPI’s, measures, calculated columns** and calculated tables using **Data analysis expressions** (**DAX**). That helped users to use the **PowerBI** in a Self-Service way.
* Integrated machine learning models into production pipelines through Kubeflow, enabling predictive analytics for identifying fraudulent claims and optimizing underwriting processes.
* Designed and maintained **100+ ETL/ELT workflows** using **Azure Data Factory, Airflow,** and **dbt,** increasing pipeline transparency by **60%** and reducing job failure resolution time by **50%** through improved monitoring and alerting.
* **Optimized SQL queries** and indexing strategies, improving report **performance by 2x** and reducing cloud resource costs by **30%** on **Azure Synapse** and BigQuery warehouses.
* Collaborated with data scientists and analysts to prepare and clean large datasets using **PySpark** and **Pandas**, facilitating accurate and efficient model training.
* Established automated data quality checks using **Great Expectations**, ensuring that incoming data met business-defined standards for accuracy and completeness.
* Developed and deployed secure **REST APIs** with **JWT-based authentication**, enabling real-time access to data for **10+** internal applications and improving data access control and auditability.
* Automated routine system and data management tasks using **Unix shell scripting**, enhancing operational efficiency and reducing human errors.
* Engaged regularly with cross-functional stakeholders to gather evolving requirements and deliver technical solutions aligned with business goals.
* Documented data engineering processes, architecture designs, and workflows thoroughly, supporting knowledge sharing and onboarding of new team members.

Client: **Trinity Health**

**Data Engineer | Jun 2018 – Aug 2021**

* Developed end-to-end data pipelines using **Python, SQL,** and **Apache Spark** to process over **30 million** EHR and claims records monthly, supporting clinical analytics that informed **95%** of patient care reporting initiatives.
* Built and optimized batch processing workflows in **Apache Airflow**, **DBT** enabling automated scheduling, monitoring, and failure recovery for critical healthcare data jobs.
* Leveraged **Azure Data Factory** to orchestrate data ingestion from multiple on-premises and cloud sources, ensuring reliable and timely availability of healthcare data.
* Created scalable **data warehouse** solutions on **Microsoft Azure Synapse**, integrating diverse healthcare datasets for unified analytics and improved patient care insights.
* Implemented robust data quality frameworks using Great Expectations, ensuring integrity and accuracy of clinical and operational data consumed by analysts and clinicians.
* **Optimized SQL queries** and schema designs, reducing report query time by **70%** and enhancing performance of healthcare analytics dashboards accessed by **200+** clinicians and analysts.
* Utilized **Power BI** in collaboration with BI teams to create interactive dashboards visualizing patient outcomes, operational metrics, and financial performance.
* Developed secure **REST APIs** for exposing aggregated healthcare data to internal applications with **OAuth 2.0** based authentication, enhancing interoperability across systems.
* Configured **Azure Monitor** and **Log Analytics** to track **100+ data pipelines**, reducing downtime incidents by **60%** and enabling proactive alerts with **<5 min response time** for critical failures.
* Collaborated with clinical data scientists to prepare and clean datasets for advanced **machine learning models** predicting patient readmissions and treatment efficacy.
* Participated in **Agile** ceremonies, contributing to **sprint planning**, backlog grooming, and retrospectives to ensure timely delivery of data engineering tasks.
* Supported migration of **15+ legacy databases** to **Azure SQL** and **Synapse**, improving scalability by **40%**, reducing backup time by **75%**, and enhancing disaster recovery through automated cloud snapshots.

**EDUCATION**

**Master of Science (M.S.) in Computer Science**  
New Jersey Institute of Technology, *Newark, New Jersey*

**CERTIFICATION**

* Microsoft Certified: **DP 900** - Azure Data Fundamentals Certification
* Microsoft Certified: **DP‑700** – Azure Associate Fabric Data Engineer Certification