+919913978738
Bengaluru, Karnataka
parth.soni78738@gmail.com

Parth Soni

Senior Data & AI Engineer
AWS Community Builder | Content Creator

TheParthSoni.github.io
github.com/theparthsoni
linkedin.com/in/parthasoni
youtube.com/@soniparth

Technology Development Professional with **8+ years** of experience in **cloud-agnostic** and **platform-agnostic Data and AI** solution design across **AWS, Azure, and hybrid** cloud environments. Proven expertise in **Generative AI, Data Engineering, Streaming Systems, Machine Learning, and Modern Data Architectures**. Strong hands-on skills in **SQL, Python, Java, C#, and core Data Structures and Algorithms**. Adept at developing **code-agnostic**, scalable systems for **Data Warehousing, Analytics, and AI/ML workloads**, with a strong focus on adaptability, optimization, and delivering enterprise-grade, **AI-first solutions**.

Skills

Tools and Languages	Python, PySpark, Scala, SQL, C#, .NET, Bash/Shell, Terraform, Docker, Git, Kubernetes
Big Data & Streaming	Kafka, Flink, PyFlink, Debezium, CDC, Spark, Hadoop, Delta Lake, Iceberg, Apache Doris
Data Engineering & ETL/ELT	ETL/ELT Pipelines, Real-time Streaming, Data Warehousing, Data Modeling, dbt, Query Optimization, NoSQL (DynamoDB)
ML & GenAI	SageMaker, Azure ML, MLflow, scikit-learn, LLMs, RAG, LangChain, Pinecone/PGVector
Cloud & DevOps	AWS (S3, EMR, Glue, Lambda, DynamoDB, SageMaker, Redshift), Azure (Synapse, ADF, Azure ML), CI/CD (GitHub Actions), IAM, VPC
Business Intelligence	SSIS, SSRS, SSAS, SQL Server, Performance Tuning
Communication & Leadership	Public Speaking, Technical Writing, Mentoring, Client Communication

Technical Experience

Senior Data Engineer — **June 2022 — Present**
Planet Payment — *Bengaluru, Karnataka*

- Designed and implemented real-time data pipelines using **CDC, Kafka, FlinkPyFlink, Debezium, AvroProtobuf, and Kraft** clusters, with **Amazon DocumentDB** as the event store.
- Built large-scale batch and streaming pipelines with **Python, PySpark, Pandas, and dbt**, processing multi-terabyte datasets and reducing runtime by **30–50%**.
- Developed **modular** ingestion, validation, and **schema-enforcement frameworks in Python** to standardize enterprise data flows.
- Designed and deployed ETL/ELT workflows on A**WS EMR, SFN, Glue, Lambda, and S3**, and implemented metadata-driven pipelines using **Glue (PySpark/Scala)**.
- Built cost-optimized **Spark** environments on **EMR** using autoscaling, spot instances, and cluster reuse.
- Implemented scalable **NoSQL** ingestion and querying using **DynamoDB** with optimized PK/GSI designs.
- Architected secure production workloads using **IAM, VPC endpoints, encryption, and fine-grained access policies**.
- Designed **ML pipelines** using **AWS SageMaker** and **Azure ML** including **feature processing, training, evaluation, and CI/CD integration**.
- Built **Databricks Medallion Architecture** and optimized cost/performance across **S3, Glue, Redshift, and Databricks**.
- Developed **real-time GenAI and LLM-enabled workflows**, including **RAG pipelines** using **LangChain, PGVector/Pinecone, and OpenAI embeddings**.
- Implemented **ML governance**, monitoring, and **drift detection** with **MLflow** and **Azure ML**.
- Modernized BI systems (**SSIS, SSRS, SSAS, SQL Server 2019**) and integrated **Datadog APM** for on-prem applications.
- Developed **Terraform-based IaC** and automated deployment pipelines with **GitHub Actions**; containerized workloads using **Docker and shell scripts**.
- Collaborated with Data Science teams on ML and data products supporting **deep learning models** for **fraud detection and revenue optimization** use cases.
- Mentored engineers on **scalable data architecture, distributed systems**, and best coding practices.
- Conducted **code and design reviews** to ensure high-quality, maintainable data products.
- Developed operational and data quality metrics to ensure **SLA adherence** and pipeline reliability.
- Advised product engineering teams on **data modeling, ingestion strategies, and architectural decisions**.
- Used **Apache Hive, ORC, Parquet, and Delta for large-scale analytical workloads.**
- Developed **business and operational metrics** supporting analytical and **ML-driven decision-making**.

Senior Application Engineer — **April 2020 — June 2022**
Thomson Reuters — *Hyderabad, Telangana*

- Designed and optimized scalable ETL workflows and data processing interfaces using **Spark and Hadoop** to support ONESOURCE applications.
- Built a centralized **data warehouse (star/snowflake)**, automating manual workflows and reducing user effort by **75%**; engineered high-scalability streaming systems.
- Improved transactional and analytical query performance through **index optimization and SQL tuning**.

- Developed and maintained **Stored Procedures, Triggers, Views, SSIS packages, and complex SQL scripts** for production operations.
- Reduced data load times by **80%** through innovative enhancements to ONESOURCE GTM's data ingestion requirements.
- Led migration of GTM products from on-prem to cloud platforms including A**zure Synapse, Azure SQL, ADF, ADLS, and AWS S3, Lambda, Glue, Kinesis, and EMR.**
- **Tuned Spark workloads (executors, memory, partitions)** reducing compute cost by **25–40%**; optimized Glue/EMR jobs with **partition pruning, broadcast joins, and efficient file fo**rmats (Delta, Parquet, ORC).
- Collaborated with architects and BI teams to define **data contracts, SLAs**, and **integration patterns**; delivered curated datasets enabling **real-time analytics and ML workloads**.

Software Development Engineer **April 2018 — March 2020**

Integration Point - Thomson Reuters *Vadodara, Gujarat*

- Developed **C#/.NET components for next-gen GTM products**, covering end-to-end **SDLC including design, development, testing, and maintenance.**
- Built **Salesforce → GTM** data synchronization workflows by developing **C#/.NET APIs** integrated with **Azure Data Factory and NoSQL storage.**
- Created a **SQL Monitor tool**to automate production alerts and improve incident visibility, collaborating closely with DBA teams.
- Monitored application and database performance using **Datadog, JSON logs, and SQL diagnostics.**
- Led the ConfigRequests team, implementing workflow configuration changes and environment validations on **Azure DevOps.**
- Served as **Information Security Ambassador**, performing architecture reviews, security assessments, and contributing to cloud security initiatives.

Software Development Intern **June 2017 — July 2017**

L&T Hydrocarbon Engineering Ltd *Vadodara, Gujarat*

- Full stack development over HR Junction and Resource booking project by developing modules with the help **ASP.Net, JavaScripts, T-SQL and Azure Cloud.**

Education

Bachelor of Engineering – Information Technology 2014 — 2018

Gujarat Technological University CGPA — 8.73

Senior Secondary Education - Science 2012 — 2014

Kendriya Vidyalaya, CBSE CGPA — 8.20

Certifications

AWS Certified **Solutions Architect — Professional**	2025 — 2028
AWS Certified **Machine Learning — Specialty**	2025 — 2028
AWS Certified **Cloud Practitioner**	Issued 2022
Microsoft Certified: Azure **Data Engineer** Associate	2021 — 2025
Microsoft Certified: Azure **AI** Fundamentals & **Data** Fundamentals	Issued 2021
Scrum Fundamentals Certified (**SFC**)	Issued 2020

Public Speaking

Practical Insights on Building Large-Scale Analytics Architectures

Apache Doris Summit 2025 *https://tinyurl.com/ApacheDorisSummitYT*

Showcasing high-performing data platform architecture(s) with Apache Doris and AWS

- AI-native real-time feature store: 70% cheaper than DynamoDB/Redis combo
- AI-powered intelligent log analytics
- Lakehouse intelligence layer on AWS
- Data platform with intelligent archival

Migrating from Snowflake to Apache Doris: Real-World Case Study

Apache Doris Webinar *https://tinyurl.com/SFtoAWSDoris*

- Led the migration from Snowflake to Apache Doris, gaining deep hands-on experience and sharing insights to enable and upskill big data teams across the world.

Planet | TechTalks - Azure & AWS Data Engineering Services

Planet Tech Summit *http://tinyurl.com/ParthAWS*

- Mentorship and knowledge sharing to built and grow cloud community. *http://tinyurl.com/ParthAzure*

AWS Community Builder

AWS Events & Meetups

- AWS Community Builder recognized for delivering impactful talks on cloud-native AI, data platforms, and GenAI applications.