SARTHAK VILAS PATEL
Data Tech Lead

Email: sarthak1287@gmail.com
Website: www.sarthakvilaspatel.com
LinkedIn: www.linkedin.com/in/sarthakvilaspatel

Summary

Accomplished Technology Leader with extensive experience driving data engineering, cloud transformation, and architecture modernization across fintech, biotech, banking, insurance, and academia. Proven ability to design and optimize large-scale data platforms, ETL pipelines, and AI-driven automation. Skilled in building cost-efficient cloud solutions, maintaining scalable big data infrastructures, and supporting seamless CI/CD pipeline integration.Experienced in leading high-impact teams, fostering innovation, and aligning data strategy with business goals to drive scalability and operational efficiency. Strong collaborator in solving complex, high-stakes problems and committed to continuous improvement and team success.

Achievements and Certifications

- AWS Blog Post on GATK-SV pipeline - Link and presented it at Advanced Scientific Discovery hosted by AWS.
- Contributed to Open-Source projects for AWS Genomics, GATK-SV by The Broad Institute and GATK-SV-AWS deployment.
- AWS Solutions Architect Associate
- Awarded Einstein Award at MIT for implementing Apache Ranger and enhancing security for user EMR clusters.
- Received Mother Teresa Award at MIT for helping team members in their tasks.
- Received SPOT and 'Star of the Month' awards for implementing and supporting multiple initiatives.
- Teradata 12 Certified Professional.
- Informatica - PowerCenter 8 Architecture and Administration.
- Six Sigma and Information Technology Infrastructure Library (ITIL) Foundation examination.

Tool & Technologies

- Cloud Computing:
  - Amazon Web Services (AWS): Batch, EFS, FSx, SNS, SES, SQS, IAM, Glue, Lambda, API Gateway, OpenSearch, S3, EC2, Elastic Map Reduce (EMR), RDS, Athena, Step Functions, Secrets Manager, Systems Manager, EventBridge, CloudWatch, CloudTrail, Lake Formation, Sagemaker AI, ECR, Fargate, Comprehend Medical etc.
  - Google Cloud Platform (GCP): BigQuery, Storage and GCR.
- Programming Languages: Python, UNIX shell scripting, SQL
- DevOps: BitBucket, GitHub, Gitlab, Docker, CloudFormation, Cloud Development Kit (CDK), Terraform and Serverless
- Database Systems: Teradata, DB2, Oracle, Aurora RDS and Serverless, Postgres, MySQL, Athena, Hive, DuckDB, Redshift
- Genomics: Expertise in DNA Seq Alignment & QC (WGS/WES/LPWGS), Joint-calling, NF-Core, GATK-SV, with tools like Omero, Hail, FastQC, IGV, and formats FASTQ, BAM, BED, OMOP, H5, GVCF.

Work Experience

Director, Data Tech Lead – Prudential Financial, *Newark, NJ* May'2023 – Present

- Led enterprise-wide AWS cost optimization, achieving $800K+ in annual savings by redesigning data pipelines, storage strategies, and compute resource allocations, ensuring long-term sustainability and efficiency.
- Designed and implemented multiple enterprise architectures, including the Market Data Repository (MDR), Disaster Recovery (DR), API infrastructure, and large-scale ingestion frameworks, enhancing data accessibility, scalability, and resilience across Enterprise Data Platform (EDP), Financial Management (FM), and analytics platforms.
- Spearheaded the Knowledge Management Architecture, building a foundation for AI-driven automation, integrating LLM-based summarization, keyword extraction, and cross-functional document indexing for enterprise-wide insights.
- Developed and executed a robust Disaster Recovery (DR) framework, ensuring compliance, security, and business continuity for mission-critical data and analytics workloads.
- Drove containerization and cloud-native modernization, leading the adoption of Docker, ECS, and serverless computing, optimizing performance, deployment agility, and cloud cost efficiency.
- Established and scaled API-driven data ecosystems, integrating high-performance data ingestion pipelines, query optimization strategies, and automated exception handling, reducing operational bottlenecks and improving analytics readiness.
- Built and led a high-performance engineering team, driving talent acquisition, mentorship, and upskilling initiatives—including AWS Cloud Practitioner training for 22+ engineers—to foster a culture of technical excellence.
- Partnered with senior leadership and business units to align data strategy with enterprise goals, improving operational efficiency, decision-making, and scalability across multiple business domains.

Principal Engineer – Vesalius Therapeutics, *Cambridge, MA* July'2022 – April'2023

- Cloud Cost Optimizations by evaluating, enhancing, cleaning and archiving storage(S3/EFS/FSx) and pipelines using Batch/EC2 instances and bringing down the spend by 15K+ USD per month.
- Design, create and manage the pipeline to create plate mappings and uploading metadata to Benchling.
- Build an ingestion pipeline for images from lab instruments and make it available for analysis through Omero.
- Collaborate with the team for creating/automating/enhancing the data pipelines for COJO/LDSC/PPMI data/UKBB clinical data etc.
- Automate multiple functionalities/repetitive tasks and provide APIs for FSx creation, EC2 actions, S3 operations, Lambda/Batch invocations, Athena queries, Inventory creation, lifecycle policies etc.
- Create and contribute to best practices, knowledge sharing and technical project documentation.

Senior Data Engineer – Goldfinch Bio, *Cambridge, MA* Nov'2019 – July'2022

- Automated and optimized genomics pipelines (DNA Seq, GWAS/RVAS, variant analysis) on AWS for large-scale data processing.
- Designed and maintained ETL pipelines ingesting clinical trials, ELN assays, compound registration, and imaging data.
- Developed a clinical data harmonization pipeline, standardizing labs, demographics, medications (RxNorm), and provider data (OMOP).
- Built APIs to efficiently query massive datasets (119B patient alleles, 774B LD records) using Athena, Elasticsearch, and API Gateway.
- Led AWS cost-saving initiatives, optimizing storage (S3, archival strategies) and reducing cloud spend by $30K+ per month.
- Created and automated CI/CD pipelines, enhancing deployment efficiency across multiple repositories.
- Developed internal Python packages to streamline AWS interactions (S3, Lambda, Athena, SNS, SQS, Batch).
- Executed multiple technology POCs, including Quilt Data Catalog (FAIR data), IGV WebApp, AWS Aurora Serverless, and AWS HealthLake integration.

AWS & Big Data Engineer - Global Atlantic Financial Group, *Brighton, MA* Apr'2019– Nov'2019

- Develop, maintain, and enhance the Data Ingestion and Validation framework for loading data from different sources like Sybase, Oracle, MySQL etc. to the Hadoop Data Lake platform.
- Develop re-usable components and utilities for the Enterprise Data Hub platform for data cleanup, archiving, batch/job status, job auto-recovery, load summary etc.
- Create tableau dashboards for getting real-time batch status reports.

AWS & Big Data Engineer - Massachusetts Institute of Technology (MIT), *Cambridge, MA* July'2018 –Apr'2019

- Develop, implement, and maintain reusable data ingestion framework using python, spark, and shell to ingest and transform data from different sources like solar and weather sensors, facilities, utilities, parking, MBTA, employee, student, Airgas cylinders etc. into the AWS Data Lake platform.
- Upgrade the Hive metastore and EMR versions for better performance, leveraging latest features and cost savings.
- Implement Active Directory and Apache Ranger for enhanced security of user EMR clusters.
- Develop bootstrapping utility for setup of self-terminating AWS EMR cluster launch scripts for ingestion process and user clusters.
- Create an inhouse scheduling tool for triggering ingestion code and users random/generic scripts on single EC2 instance or EMR clusters.
- Generate EMR Cluster utilization report consisting of YARN application and ssh (login) details of user clusters.
- Implement data reconciliation scripts and reporting for all ingested data and various data sources.

Client: Morgan Stanley, New York Mar'2010 – July'2018

Big Data Technical Lead - Hadoop Data Lake Setup, Governance, and Ingestion May'2014 – July'2018

- Designing, planning, on boarding and reviewing the new/existing application setup for using big-data tools.
- Design, develop and maintain data ingestion utility for ingesting TBs of data from Teradata and Mainframe source to Hadoop Data Lake.
- Understanding and implementing the Disaster Recovery plans for the Hadoop Clusters.
- Setup Guidelines for access control of HDFS components via Sentry, scheduling process on Edge Nodes.
- Setup Informatica Cloud Secure Agents & BDE on Hadoop Clusters for easy interaction with Salesforce clouds.
- Create reusable components and utilities for cluster, data, and application management such as data sync up between clusters, job wrappers, table structure & row counts comparison, small file checks and compaction, space/job failure/cluster/service alerts etc.
- Reduce Teradata and Data warehouse footprint by offloading the data to data lake.
- Data Validation between Teradata and Hadoop Data Lake for ensuring data integrity.

Datawarehouse Lead – Risk Assessment Remediation Apr'2013 – May'2014

- Migration of 60+ Applications from Local Unix FS to AFS and maintaining version via GIT/Stash.
- Remediation of scripts/jobs across Unix, Informatica and SSIS to use secured connections i.e., Connect Direct/LFTP for Mainframe, Kerberos, and query band connection to various databases like DB2/SQL/Teradata etc. through shell script and Informatica connectors.
- Design and developed centralized reusable utilities for Mainframe Transfers, Informatica workflow invocation, File transfer to external servers/locations, and invoke all the jobs from TWS across all applications to streamline and maintain metadata at a single place.
- Automate change management process for Informatica workflow recovery (task level), OS profile, connection creations etc.

Support Engineer – Data Warehouse Operations Mar'2010 – Apr'2013

- Data warehouse batch monitoring, stabilization, failures/outage/incident management and resolutions.
- Informatica and Business Objects Products Administrative tasks such as installation, creating Repository, extracting information from repository database tables and Domain database, command line tools, Integration services, copying repositories, backing up repositories/domains, user/group management, security domain management, auditing etc.
- Automations for End-to-End Informatica/Unix code promotion/migration, space alerts, CPU alerts, Status Reports, Job Execution Comparison etc.

Education

2005-2009

Bachelors in Electronics Engineering from Shri Ramdeobaba Kamla Nehru Engineering College (SRKNEC), INDIA