**SARTHAK VILAS PATEL**
Data Engineer

LinkedIn: www.linkedin.com/in/sarthakvilaspatel
Website: www.sarthakvilaspatel.com

## Summary

High-performing Data Engineer with a passion for various industries like biotech, fintech, banking, insurance, and academics. 14+ years of experience impacting positive organizational outcomes through architecting, building, maintaining, optimizing, testing, and supporting highly scalable big data/data warehousing/data lake applications, infrastructures, and CI/CD pipelines. Confident in ability to collaborate with cross-functional teams to solve complex, high-stakes problems. Committed to continuous improvement and contributing to 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, ECR, Comprehend Medical etc.
    - Google Cloud Platform (GCP): BigQuery, Storage and GCR.
- Programming Languages: Python, UNIX shell scripting, SQL
- DevOps: GitHub, Gitlab, Docker, CloudFormation, Cloud Development Kit (CDK), Terraform and Serverless
- Database Systems: Teradata, DB2, Oracle, Aurora RDS and Serverless, Postgres, MySQL, Athena, Hive
- Genomics:
    - Pipelines: DNA Seq Alignment and QC for WGS/WES/LPWGS, Joint-calling, NF-Core (Hic, AtacSeq, ChipSeq), GATK-SV on AWS
    - Tools: Omero, Hail, FastQC, MULTIQC, Picard, VEP, HiC-Explorer, IGV, HiGlass, GCTA, LDSC etc.
    - Data Formats and Standards: FASTQ, BAM, BED, OMOP, COOL, H5, BEDGRAPH, GVCF, HIC
    - Orchestration: Nextflow, Cromwell and AWS Step Functions.

## Work Experience

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

- Mentor and coach, the technical team providing feedback on deliverables and project timelines including support with code reviews, domain knowledge and ownership, SLA development and application reliability.
- Collaborate with global technology partners, business partners, architecture, chief product owner and team members to drive outcomes with focus on enriched customer experience.
- Design and execute technical design and infrastructure/environments strategy, including coding best practices, continuous integration and deployment and help integrate technology strategy into consumable units of work with a Minimum Viable Product (MVP) mindset.
- Develop, maintain, and support multiple different applications and projects simultaneously with accountability.

**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

- Automate and implement various genomics pipelines to run on large datasets and scale efficiently on AWS infrastructure.
- Design, develop, maintain, and enhance the ETL/data pipeline for ingesting data from various sources for clinical trials data, ELN Assay and inventory data, compound registration, GWAS/RVAS/D-Gene Expression studies, patient alleles & zygosities, variant effects, eQTLs, allele frequencies, linkage disequilibrium (LD) & imaging data using various AWS services, docker, python and shell scripting.
- Build, maintain and update the clinical data harmonization pipeline to ingest different types of data like labs, demographics, comorbidities, family & social histories, visit observations, medications (RxNorm standards) etc. coming from various providers including OMOP format and convert it into a standard format for analysis and visualization.

- Develop complex APIs querying large datasets like 119 billion records of patient alleles, 774 billion records of LD data efficiently from S3 using Athena, Lambda, Elasticsearch and API Gateway.
- Create, maintain, and automate CI/CD pipelines for multiple projects and repositories.
- Routinely optimized cost savings on AWS including S3 storage costs, archival strategies, gather usage of services on accounts which reduced AWS spend more than 30K USD per month.
- Built an internal python package for data scientists to simplify tasks in AWS (s3 file/folder listing, download, upload, archiving, SNS, SQS, lambda invocation, batch jobs, athena queries etc.), python functions, data loading, data transfers, and logging.
- Complete multiple technology development POC's including Quilt Data Catalog as per FAIR data principles, IGV WebApp, AWS Aurora serverless, AWS Comprehend Medical, AWS HealthLake, GitLab boilerplate repository, integration of AWS alerts to Microsoft teams.

**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