Divya Shah

divya.s@protectmymails.com | +1 (628) 245-5521 | LinkedIn | GitHub | Portfolio

SUMMARY

Results-driven Data Engineer with 4+ years of experience building and optimizing data architectures, pipelines, and integrations across AWS, Azure, and multi-cloud environments. Skilled in ETL design, cloud migration, and real-time data integration for enterprise and government clients. Proven success at Morgan Stanley and Hexaware Technologies in delivering scalable data models, automated pipelines, and ML-driven analytics using PySpark, Databricks, Airflow, Kafka, and Snowflake. Recognized for improving system performance, ensuring data integrity, and enabling actionable insights through robust, compliant, and high-performing data solutions.

PROFESSIONAL EXPERIENCE

JP Morgan Chase & Co., NY | Data Engineer **May 2024 – Present**

- Led the end-to-end Data Architecture and Integration Strategy for the State of Hawaii Child Welfare Modernization Program a multi-year digital transformation initiative replacing legacy systems with the modern SaaS-based Cardinality AI platform. Acted as the technical lead bridging data engineering, integration development, and cross-functional teams across Hawaii, Georgia, and Indiana.
- Designed and implemented enterprise-grade Data Architecture Models and Compliant Management Architecture using Draw.io, optimizing star-schema designs for the data warehouse. Improved query performance and system scalability by 20%, enabling a unified single source of truth for all child welfare analytics and reporting.
- Developed and optimized complex SQL queries, stored procedures, and ETL validation frameworks to verify source-to-target data mappings during migration from legacy systems. Ensured data integrity for over 50M+ records with a 98% accuracy rate, resolving long-standing data quality issues and improving overall data reliability.
- Built and deployed automated, serverless data ingestion pipelines on AWS (S3, Lambda) to streamline daily data extracts. Implemented encryption, logging, and audit mechanisms, reducing manual intervention by 30% and providing a secure, scalable foundation for downstream analytics and integrations.
- Served as Lead Integration Engineer on the Boomi Integration Platform, architecting and developing REST and SOAP APIs for secure, real-time, bi-directional data exchange between state agencies (e.g., Georgia DHS) and the Cardinality AI ecosystem. Implemented OAuth 2.0 authentication and standardized integration patterns to enhance system interoperability and compliance.
- Provided technical leadership and solution architecture for multiple state programs (Hawaii and Georgia), ensuring successful end-to-end data migration, system integration, and compliance with state and federal data regulations. Contributed significantly to the go-live of two major Cardinality AI products Communicare DSP Invoicing and Foster Parent Invoicing by driving data and integration readiness across all environments.

Hexaware Technologies, India | Data Engineer **Apr 2021– Aug 2023**

- Engineered and maintained scalable ETL and machine learning pipelines on Azure Databricks using PySpark and Python, integrating terabyte-scale datasets (policy, claims, customer, EHR, external market) to enhance data availability for analytics by 40% and accelerate ingestion in Azure Data Factory and Data Lake Storage by 20% through schema optimization and performance tuning.
- Automated data validation, transformation, and reporting workflows with PySpark and Python, reducing manual workload by 30% and expediting actuarial and financial close cycles. Led a data governance initiative implementing rule-based quality checks, cutting data errors by 95% and cleanup time from 10 hours to 30 minutes per dataset.
- Designed and deployed an end-to-end Databricks “Email Modeling” pipeline automating feature selection, EDA, and hyperparameter tuning for 200–250 models per run, with automatic best-model promotion for deployment, boosting overall modeling efficiency and reliability.
- Applied AutoML and hyperparameter optimization to improve predictive accuracy for customer lifetime value (CLV) and churn propensity models by 15%, while implementing MLOps pipelines for automated retraining, validation, and deployment reducing time-to-production from 3 weeks to 5 days and ensuring stable model performance.
- Conducted advanced analytics and statistical modeling (Python, SQL, Cox Proportional Hazards, time series forecasting) on policyholder behavior, claims trends, and clinical outcomes to inform insurance product strategies, risk prediction models, and healthcare research initiatives.
- Developed interactive Tableau and Power BI dashboards visualizing customer segmentation, policy performance, and clinical metrics, enabling real-time insights across business and research teams and driving a 20% increase in conversion rates and improved policyholder retention.

KPIT Technologies, India | Junior Data Engineer **Dec 2020 –Apr 2021**

- Led the design, development, and optimization of large-scale, cloud-native data pipelines and analytics platforms across AWS, Snowflake, and Redshift, ensuring enterprise-grade scalability, reliability, and high performance for multi-cloud data ecosystems.
- Built and managed 15+ Apache Airflow DAGs and 50+ BMC Control-M workflows to automate end-to-end data refresh cycles, establish inter-job dependencies, and proactively monitor failures achieving a 99.8% on-time job success rate for critical data operations.
- Engineered scalable ETL pipelines using Apache Spark, Pandas, and NumPy to process petabyte-scale data; integrated Spark, Hive, and Sqoop on AWS EMR to enhance throughput, improving data reliability by 20% and reducing transformation latency by 43%.
- Executed complex integrations using SAS Enterprise Guide to extract, cleanse, and validate sensitive healthcare datasets from legacy systems, ensuring full HIPAA compliance and data accuracy prior to ingestion into enterprise data platforms.
- Designed and implemented star-schema data models consolidating 12+ source systems into unified analytical structures; integrated Amazon RDS with Snowflake to deliver a 35% boost in query performance and significantly faster reporting cycles.
- Optimized analytical queries in Redshift and Snowflake through advanced tuning (materialized views, sort/distribution keys), cutting report times from 12 minutes to under 90 seconds; automated infrastructure provisioning via Terraform and CI/CD pipelines in GitHub Actions to achieve 70% faster deployments.
- Delivered interactive Power BI dashboards powered by data from AWS S3, RDS, and Snowflake, enabling real-time KPI tracking and improving stakeholder decision-making speed by 45% across business units.

TECHNICAL SKILLS

Programming Languages:	Python | R | Java | Scala | SQL
Big Data Ecosystem:	Hadoop | Hive | HDFS | HBase | Apache Airflow | Apache Kafka | Apache Spark | Apache Flink | Databricks
Cloud Technologies:	AWS (EMR, EC2, S3, Athena, Glue, Elasticsearch, Lambda, DynamoDB, Redshift, QuickSight, Kinesis) | Azure (Data Lake, Databricks, Data Storage, Data Factory, App Service, SQL Database, Blob Storage)
Visualization & Reporting:	Tableau | Power BI | Excel
Etl & Data Integration Tools:	SSIS | SSRS | Fivetran | SAS | Informatica | PySpark | Talend | Tableau Prep
Data Processing & Analytics Packages:	Pandas | NumPy | Matplotlib | Seaborn | PySpark | Data Pipelines | DBT (Data Build Tool)
Version Control & Databases:	Git | GitHub | SQL Server | PostgreSQL | MySQL | Snowflake | Cassandra | DynamoDB
Data Management & Governance:	Data Modeling | Data Warehousing | Data Governance | Metadata Management | Data Quality | Master Data Management (MDM) | Data Cataloging

CERTIFICATES

AWS Certified Cloud Practitioner | AWS Certified Developer – Associate

EDUCATION

Master of Science in Computer Science | New York Institute of Technology, New York, USA

Bachelor of Engineering in Computer Engineering | Dharmsinh Desai University, Gujarat, India