Vasant Kumar Desai

Senior Software Engineer
+91 9916799178 | vpdesai2020@gmail.com | Bengaluru, Karnataka, India | Open to Relocate |
www.linkedin.com/in/vpdesai/ | www.vpdesai.in

SUMMARY

Data Engineer & Data Scientist with 7+ years of experience building large-scale data pipelines and advanced analytics solutions across healthcare, pharma, and enterprise domains. Skilled in AWS Glue, PySpark, and Iceberg for data lakes and ETL frameworks, combined with ML/NLP expertise using SageMaker, scikit-learn, and spaCy. Proven ability to design predictive models, automate unstructured document intelligence, and deliver governed data platforms with compliance, traceability, and business impact. Strong at bridging engineering with applied data science to drive actionable insights.

SKILLS

Programming & ML: Python, SQL, PySpark, R, scikit-learn, FBProphet, ARIMA

Big Data & Cloud: AWS Glue, EMR, S3, Redshift, Athena, Snowflake, Apache Iceberg, Databricks

ML & NLP Pipelines: SageMaker, spaCy, Textract, NLTK, Regex, BERT, Selenium-based scraping

ETL & Data Warehousing: Data Lakes, Medallion Architecture, CDC, Parquet/ORC

Workflow & Automation: Airflow (MWAA), Control-M, GitHub Actions, Snakemake, Docker, CI/CD

Analytics & Visualization: Tableau, Power BI, MySQL

Governance & Compliance: HIPAA, GDPR, GxP, ALCOA+, FAIR Data Principles

CERTIFICATIONS

AWS Certified Data Engineer Associate – 2024
AWS Certified Machine Learning Engineer – Associate - 2025
Google Business Intelligence Specialization – Coursera
AutoML and BERT NLP – Coursera/LinkedIn
Data Governance with Databricks – Coursera

WORK EXPERIENCE

TEKsystems Global Solution (Client: American Automobile Association)

Senior Software Engineer(Data Engineering) | Technical Lead , Bengaluru, India | Oct 2024 — Present

- Architected end-to-end, scalable **ETL pipelines** using **AWS Glue**, **PySpark**, and **Apache Iceberg**, enabling schema-evolving data lakes and efficient downstream processing.
- Designed and maintained a **Medallion Architecture** (Bronze, Silver, Gold) on **S3**, **Databricks**, and **Redshift**, ensuring clean data flow and model-driven analytics.
- Built **automated validation frameworks** on **EMR** and **AWS Lambda**, achieving **99.8% data quality** through rule-based checks and ingestion error detection.
- Developed **dbt-powered Snowflake data models and dashboards**, enhancing cross-functional reporting and metric traceability.
- Migrated **100+ TB of structured data** from on-prem **Hadoop** to **AWS Athena** and **S3**, reducing query latency by 50% and accelerating insight generation 3x.
- Integrated heterogeneous data sources (**Salesforce**, **SQL Server**, **Teradata**, **Hadoop**) to unify operational datasets across legacy and modern platforms.
- Worked on **Qlik Replicate-based ingestion pipelines**, replicating change data from **SQL Server to S3**, and processing them using **EMR merge logic**.
- Orchestrated multi-step batch pipelines using **Amazon MWAA (Airflow)** and **Control-M**, enabling seamless cross-system data coordination.
- Enhanced ETL performance with **Spark tuning**, **memory optimization**, and **Snappy compression**, reducing processing times significantly.

Labcorp Laboratories India

Data Delivery & Automation Specialist | Bengaluru, India | Apr 2021 – Sep 2024

- Built and deployed predictive models in **AWS SageMaker** (using **FBProphet, ARIMA, and scikit-learn**) to forecast clinical trial milestones, dropout risk, and site enrollment patterns—achieving **80–95% accuracy**, enabling proactive intervention and risk mitigation.
- Applied **NLP techniques** (SpaCy, AWS Textract, regex, NLTK) to extract metadata, entities, and context from **10,000+ informed consent forms/month**, enhancing trial document indexing, traceability, and **regulatory compliance (HIPAA/GDPR)**.
- Engineered scalable **document intelligence pipelines** using Lambda, Textract, and Redshift to automate ingestion of 20,000+ structured/unstructured files monthly, reducing manual effort by 70% and standardizing metadata enrichment.
- Collaborated closely with clinical operations, regulatory affairs, and compliance teams to **align ML pipelines with GxP and ALCOA+** principles, ensuring audit-readiness and ethical data use.
- Developed **automated access control models** for clinical datasets using rule-based RBAC, enabling secure and trackable data sharing across teams and study sites.
- Built and published 10+ **interactive dashboards** in **Power BI and Tableau** to monitor trial KPIs, adverse event trends, and protocol deviations—enabling 25% faster decision-making by clinical leads and sponsors.
- Integrated real-world datasets from S3, SFTP, and third-party APIs using Python and SQL, and applied feature engineering and transformation logic for machine learning–ready formats.

- Conducted exploratory data analysis (EDA), statistical validation, and visualization to support hypothesis generation and data interpretation for internal R&D teams.
- Contributed to design of a **centralized data warehouse** for clinical trial data with schema enforcement and metadata versioning, cutting compliance review time by 300+ hours monthly.

COE RVCE, Bengaluru

Junior Research Fellow | Bengaluru, India | Aug 2018 – Mar 2021

- Developed a PyMOL plugin (https://pymolwiki.org/index.php/PICv) in collaboration with PDB Europe to visualize residue-level protein interactions, aiding literature-linked structural biology research and antibody design.
- Automated NGS workflows using **Snakemake** and **Bash**, streamlining high-throughput genomic data processing with minimal manual intervention.
- Integrated public genomic databases like **Ensembl** and **dbSNP** into internal variant annotation workflows, enhancing the traceability and scientific reproducibility of results.
- Built Python-based **text mining automation** using **Selenium** to extract and download protein structure data from scientific literature and open databases.
- Designed and executed in silico models using Schrödinger tools to study drug transport proteins (P-gp), contributing to early-stage drug discovery.
- Managed and maintained clinical and genomic research databases, improving data accessibility, version control, and audit readiness.
- Wrote Python scripts for data cleaning, statistical summaries, and visualization, enabling reproducible analysis across multiple research teams.
- Collaborated with scientists and bioinformaticians on validation and improvement of genomic workflows, aligning deliverables with project and publication goals.

EDUCATION

M.Tech in Bioinformatics – R.V. College of Engineering, Bengaluru **Jun 2018**
B.E in Biotechnology – M.S. Ramaiah Institute of Technology, Bengaluru **Jun 2015**

AWARDS & ACHIEVEMENTS

Spot Award June 2025 | TEKsystems Global Solution
Ace Award Level 4 – Automation & Process Optimization | Labcorp Laboratories India
Ace Award Level 3 – Vendor Analytics Tool Development | Labcorp Laboratories India
Outstanding Performance Rating – 2024 | Labcorp Laboratories India

PUBLICATIONS

- **Desai, V.**, *Novelty of Data Mining Techniques for Bioinformatics Approach*, Journal of Environmental Research and Development, June 2017. [Focus: Application of data mining in biological datasets]
- **Desai, V.**, *Identification of Banana Heat Responsive Long Non-Coding RNAs and Their Expression Analysis*, 3Biotech Journal. [Focus: Bioinformatics analysis of stress-responsive lncRNAs in plants]