PURVA JAVALE

SOFTWARE ENGINEER | BIG DATA | MACHINE LEARNING

United States | +1(502) 650-1425 | purvajavale@gmail.com | LinkedIn | Portfolio | GitHub | Medium

Work Authorization: H4 EAD – No sponsorship required.

PROFESSIONAL SUMMARY

Software Engineer with 6+ years of experience building large-scale Big Data platforms, ETL pipelines, and audit-ready reporting solutions for U.S. banks and fintechs. Skilled in Java, Spark, Hadoop, AWS, Databricks, and Power BI, with exposure to Machine learning workflows. Specialized in automating regulatory reporting with a strong focus on data lineage, controls, data quality, and compliance accuracy, enabling reliable and scalable enterprise data workflows.

During 2022–2025, focused on upskilling in Big Data, Cloud, Databricks, and Machine Learning, completing certifications and applied projects to stay aligned with industry trends.

CORE SKILLS

Big Data: Hadoop, Spark, Hive, HBase, Kafka, Airflow
Database: MySQL, SQL Server
Formats: TXT, CSV, XML, JSON, AVRO, ORC, PARQUET
Machine Learning: Data Preprocessing, Model-Ready Pipelines
Visualization: Power BI, Dashboard Design

Programming: Java, Python, SQL, .NET Core
Cloud & DevOps: AWS, Databricks, Maven, Jenkins
Database: MySQL, SQL Server
Legacy: IBM Mainframe (z/OS), WebSphere
Methodologies: Agile, SDLC, Data Governance, Compliance

Tools: Team Foundation Server, GitHub, Eclipse, Visual Studio, PyCharm, Microsoft Project, Jira, Microsoft Office

---

PROFESSIONAL EXPERIENCE

Software Engineer | Hexanika – Pune, India | Feb 2015 – Jan 2022

- Scaled Hexanika's regulatory data platform from inception to enterprise-level adoption, automating compliance reporting for U.S. banks.
- Built ETL pipelines using Java, Hadoop, Spark, and AWS, improving performance by 30% and reducing manual effort by 40%.
- Designed Spark-based transformation and validation frameworks with schema checks, rule versioning, lineage tracking, and audit controls.
- Modernized legacy workflows by deploying Spark on IBM z/OS mainframe for parallel rule execution and faster compliance processing.
- Integrated processed datasets with Power BI dashboards for real-time regulatory insights and operational analytics.
- Implemented CI/CD pipelines using Jenkins and Maven for Spark ETL jobs deployed on AWS EMR.
- Led Agile delivery, mentoring engineers on Spark optimization, AWS data lake design, and data quality best practices.

Platforms Delivered

Smart Join - *Regulatory Data Preparation*

- Designed a data cleaning framework with standardized mapping, boosting data accuracy and reducing preprocessing workload by 40%.
- Implemented Spark-based transformation and validation engine with audit checks, schema validation, and governed rule execution.

Smart Reg - *Regulatory Reporting*

- Developed a Spark SQL rules engine covering FRY-9C, Call Reports, HMDA-LAR, and AML-SAR regulatory logic at scale.
- Integrated predictive ML models for rule applicability scoring, reducing false positives and improving compliance accuracy.
- Provisioned and maintained multi-node Hadoop/Spark clusters and AWS infrastructure (S3, IAM, EC2, EMR) with secure access controls.
- Built SOAP APIs and Power BI dashboards to monitor pipeline performance, controls, and end-to-end data lineage.

EDUCATION

Pune University, India

Bachelor of Engineering (B.E.) - Information Technology | June 2010 - June 2014

CERTIFICATIONS

Stanford University (Coursera) - *Machine Learning Specialization | 2025*
Google (Coursera) - *Agile Project Management | 2024*
TrendyTech - *Big Data Masters Program | 2025*
IBM (Coursera) - *IBM Mainframe | 2024*
Databricks - *Accreditation Databricks Fundamentals | 2025, Accreditation Databricks Platform Architect (AWS) | 2025*

PERSONAL PROJECTS

Data Cleaning: Data cleaning pipeline in Python for AWS S3, handling missing values, duplicates, outliers, and type corrections.
Data Professionals Survey: Built interactive Power BI dashboards to visualize survey insights on trends, tools, salaries, and skill gaps.
Data Engineering Project using Databricks, AWS S3, and Power BI: End-to-end pipeline for data ingestion, transformation, and visualization.
Clustering Engine: Python tool for fuzzy matching and K-Means clustering to clean, normalize, and uncover patterns in data.
AI-Powered Movie Review Sentiment Classifier: A Sentiment analysis classifying reviews with 85%+ accuracy using Python and Streamlit.

LEADERSHIP & EXTRAS

- Wore multiple hats across engineering, client delivery, and hiring.
- Delivered guest lectures on Big Data at MIT Pune to mentor aspiring data engineers.
- Recognized as 'Star Performer of the Year' and 'Well-Done Achiever' for leading key innovations.
- Secured 2nd place in Hexanika's FinTech Hackathon for creative data engineering solutions.