Abhijit Joshi

abhijitpjoshi@gmail.com +1 201-356-7910 Milpitas, CA, USA
https://www.linkedin.com/in/abhijitpjoshi/

TECHNICAL SKILLS

- **** Apache Spark
- ***** Airflow
- ***** SQL
- **** Python/Unix/Shell
- **** NOSQL (MongoDb)
- ***** RDBMS
- ***** Data Engineering
- ***** AWS/GCP
- ***** Databricks Lakehouse
- **** Terraform
- ***** Kafka
- *** AI & ML

AREA OF EXPERTISE

Programming languages: Python, Java, C/C++, HTML/CSS, COBOL, and version control using Git, GitHub, Bitbucket, GSR.

Data engineering tools: Databricks, PySpark, Airflow, DBT Cloud, SQL, Docker, Kubernetes, Informatica, Tableau, Autosys, Data warehouse technologies, jQuery, SOAP & REST Web Services, Kafka, Pub-Sub Messaging Q, Data Modeling, Microservices

Data stores: Unity Catalog, MariaDB/MySQL, MongoDB, Redshift, BigQuery, CloudSQL, DB2, Oracle, Teradata, PostgreSQL, PL/SQL, SQLite, SQLalchemy.

Cloud computing platforms: AWS, GCP, Terraform, Databricks, PaaS, IaaS, SaaS, CRM, IAM, Cloud Networking, Salesforce

Advanced scripting: Unix/Linux (Bash, ZSH, AWK, SED), Windows, SSH/SCP/SFTP, Kerberos, LDAP, XML, JSON, Regex.

DevOps tools: Terraform, AWS CloudFormation, Jenkins, GitHub Actions

Accomplished Data Engineering Lead with over 15 years of professional experience, specializing in data engineering and infrastructure management. Proven record in leading cross-functional teams and executing strategic technological initiatives. Expert in modern data tools, and cloud platforms including AWS, GCP, Databricks, SQL, Terraform, MLOps, and CI/CD methodologies.

WORK EXPERIENCE

Oportun, San Carlos, California

Staff Data Engineer | Apr'22 - Present
Senior Data Engineer | Aug'21 - Apr'22

- Managed and architected solutions on Databricks, Airflow, DBT, and AWS, optimizing platforms and code to achieve significant cost **savings: $100K monthly on Databricks and $30K monthly** on **AWS** costs related to Databricks services, resulted in 50% savings.
- Designed and implemented hundreds of Data Pipelines using airflow, Databricks and PySpark, DBT to migrate from Minerva1.0 to Minerva2.0 for various subject areas to create Bronze, Silver, Gold and platinum Data Layers in **Delta Lakehouse**.
- Employed **Terraform** for **efficient** infrastructure management and automation, contributing to substantial cost efficiencies.
- Led 20 person cross-functional team of Data engineers, DataOps and Database admins in data and infrastructure initiatives, focusing on **cost-effective solutions and performance optimization**.
- Designed and implemented Minerva2.0 Design principals, tenets, developer best practices, cost-focused architecture converging into the cost reporting
- Diverse experience in Data Analysis, Data Warehousing, Data Governance, Data Modeling, Data Platform reporting and Visualization
- Worked with developers, DBAs and operations in elevating and automating successful Data Pipelines deployments reducing error rate by 95%
- Led PII/NPII cataloguing for Datalake, Warehouse objects to enable data access controls and efficient metadata management using Alation (Data Catalogging)
- Accelerated the design of scalable, reusable, and low maintenance ETL templates

ViacomCBS, New York City, New York

Senior Data Engineer | Aug'19 – Aug'21
Data Engineer | Sep'17 - Aug'19

- Devised the end-to-end orchestration pipeline platform for Airflow on GCP Compute Engine and later in google cloud composer. Lead a team of data engineers for implementation and best practices of various products.
- **Designed multi-repo GitHub organization** with repositories for CBS-Corp and CBS-interactive divisions. Came up with folder code structure, standards for easy implementation and reusability of data pipelines segments/objects.

CICD Tools

Docker

Jenkins

Terraform

GitHub Actions

CERTIFICATIONS

Six Sigma – Green Belt

IBM Certified Developer – DB2 9.7 SQL Procedure Developer

DB2 9.7 Application Developer

DB2 9.7 Fundamentals

Informatica university MDM Hub

Teradata TE0-141

Databricks Data engineer professional

Databricks Machine learning Associate

EDUCATION

Bachelor of Engineering (2005 – 2009)
Specialization – Information Technology and Computers,
Shivaji University

SCHOLARLY WORKS

https://dataworldfromabhijit.blogspot.com/2024/07/my-scholarly-works.html

- Implemented Kerberos-authenticated Docker setup at org level, reducing Google Cloud Platform VM costs by **$10,000 monthly**
- Designed and executed various data pipelines (Python, Airflow, SQL) for analyzing product impacts across marketing channels leading to actionable insights through Tableau, Oracle, MsSQL, DB2
- Pioneered the **end-to-end orchestration pipeline platform** for **Airflow** on GCP Compute Engine and later in **google cloud composer**. Lead a team of data engineers for implementation and best practices of various products moving from traditional data warehouse resulted saving of **50%** cost
- Spearheaded the integration of advertising sales data from various local media and network sources into centralized data warehouses.
- Collaborated with the ad sales team to analyze campaign data, identifying key metrics for improvement and reporting on ROI
- Developed advanced data models for audience segmentation, enabling more targeted and effective advertising strategies utilizing big data technologies to process and analyze large datasets, improving the accuracy of audience insights

Morgan Stanley, New York City, New York

(Consulting) Data Engineer | Feb'11 – Sep'17

- Developed a party operational data model framework using **python, SQL, AutoSys & Linux scripts** which showed potential of twice the improvement in sales/prospects per customer. Leveraged rule-based party identification engines and ensemble of technologies like Informatica MDM to predict likelihood of a customer opening new accounts.
- Performed the Data Engineering on disparate data sources in DB2, document forms like W8, W9, FATCA transactions to provide customer insights and allow informed business decisions daily essential portfolio and market analytics.
- Developed and maintained numerous batch data pipelines and scripts for critical path schedules with stringent SLAs, including rapid resolution through investigation and coordination. Automated data reconciliation with source systems for accuracy.

Union Bank, San Francisco, CA

(Consulting) Application Developer | Jun'09 – Feb'11

- Developed Informatica PowerCenter mappings and workflows, and programmed in COBOL, DB2, CICS, and JCL.
- Specialized in performance tuning and resolving bottlenecks in data sources, targets, mappings, and sessions.
- Developed, deployed, and integrated SOAP and RESTful web services, both exposed and consumed through Java.