NIHARIKA

 (**732-338-8659)**| p.s.niharika07@gmail.com |

# SUMMARY

Senior Data Engineer with 8+ years of experience building and optimizing large-scale data pipelines across cloud platforms including Azure, AWS, and GCP. Proficient in developing batch and real-time ETL workflows using Apache Spark (Java, PySpark), Kafka, and cloud-native orchestration tools. Skilled in managing large datasets, writing complex SQL queries, and deploying scalable analytics solutions using CI/CD pipelines. Recent experience includes working with Cloud Composer equivalents, schema governance, and implementing monitoring strategies for production-grade pipelines.

* Expertise in PySpark, Spark SQL, DataFrames, and Python-based data transformations
* Hands-on with GCP services (Cloud Functions, Cloud Storage, BigQuery, PUB/SUB, GKE) through recent project alignments and migration experience
* Strong understanding of monitoring and alerting, using tools like Grafana, CloudWatch, and custom pytest-based frameworks
* Proficient in building CI/CD pipelines using Git, Jenkins, and GitHub Actions
* Familiar with modern data modeling, pipeline lifecycle management, and cloud security best practices
* Familiar with orchestration, CI/CD, data lineage, and monitoring best practices in the cloud
* Familiar with Hadoop, Hive, Airflow, Git, Jenkins, CloudFormation, TerraformExpertise in **Agile** software development methodologies with experience managing production support, incident management, and system troubleshooting

# SKILLS

**Languages:** Java, Python, SQL, Shell **Big Data:** Apache Spark (Java, PySpark), Kafka, Hadoop, Hive, Spark SQL, DataFrames **GCP:** BigQuery, Cloud Storage, Cloud Functions, Pub/Sub, Dataflow, Composer (familiar via Airflow), GKE (familiar with containerized deployment)
**Cloud:** Azure (ADF, Event Hubs, Blob Storage), AWS (S3, Glue, Lambda, Redshift, Athena) **DevOps/CI-CD:** Git, Jenkins, GitHub Actions, Terraform, CloudFormation **Monitoring/Testing:** Grafana, CloudWatch, PyTest, JSON/Protobuf validation, alerting scripts
**Data Formats:** JSON, Avro, Parquet, Protobuf **Tools:** Power BI, Tableau, Redash, Airflow, Splunk **Methodologies:** Agile, Scrum, CI/CD, Data Governance, Production Support

# EXPERIENCE

## Bluescape, CA | Data Engineer | March 2023 – Present

* + Built large-scale Spark-based batch and streaming pipelines integrated with Azure and AWS; now exploring GCP migration patterns
	+ Engineered data ingestion using Kafka + Spark Streaming, designed for cloud-agnostic architecture (GCP-compliant)
	+ Implemented monitoring and alerting with Grafana, CloudWatch, and PyTest for data quality and pipeline reliability
	+ Delivered schema governance, data validation (JSON/Protobuf), and lineage tracking for audit complianceDeveloped multi-cloud ingestion and analytics pipelines integrated with AWS Glue and Azure Data Lake
	+ Partnered with DevOps for CI/CD automation using Git, Jenkins, and GitHub Actions
	+ Participated in internal POCs evaluating BigQuery and Pub/Sub as potential replacements for Redshift and Kafka
	+ Optimized PySpark DataFrame operations and Spark SQL logic for cost efficiency and faster compute times
	+ Supported data modeling tasks aligned with GCP architecture principles
	+ Implemented data versioning and lineage tracking mechanisms for auditability and compliance
	+ Supported multi-environment deployments and coordinated with DevOps to manage **CI/CD** pipeline using **Git**

and **Jenkins**

* + Led migration of legacy **ETL** workflows to **distributed Spark jobs** using AWS Glue, reducing job runtime by 50%
	+ Collaborated with data scientists to integrate **pre-processed feature sets** into model training pipelines using S3 and Athena
	+ Created monitoring dashboards in **CloudWatch and Grafana** for Airflow DAGs and batch job health
	+ Standardized data schema definitions and enforced **data governance** through schema registry and versioning

## Amazon Web Services, Boston | SDE | July 2022 – March 2023

* + Developed Spark-based ETL pipelines for regulatory compliance and analytics using Java, PySpark, and SQL
	+ Created event-driven ingestion layers with S3 + Lambda + SNS, analogous to GCP’s Cloud Functions + Pub/Sub
	+ Designed templated CloudFormation stacks, similar to Terraform for GCP deployments
	+ Led performance optimization for Redshift queries, with exposure to BigQuery performance tuning strategies
	+ Contributed to internal analytics pipelines leveraging Athena and S3 – applicable to BigQuery and GCS patterns
	+ Automated snapshot and backup logic (WORM) relevant to data lifecycle best practices in GCP
	+ Designed and optimized **Lambda functions** for data processing and automated alerting
	+ Supported multi-region deployments for compliance and developed tools for **WORM-compliant backup policies**
	+ Enabled **performance tuning** of SQL queries in Redshift using query plans, EXPLAIN output, and compression encodings
	+ Designed and implemented **event-driven Lambda architectures** triggered by S3 events and SNS notifications
	+ Collaborated on building a data pipeline for internal usage analytics using **Redshift Spectrum and S3 partitioned data**
	+ Created templated CloudFormation stacks for deploying snapshot management infrastructure across environments

Authored extensive documentation and internal wikis for **CI/CD automation**, pipeline architecture, and compliance auditing

## Accenture Solutions, India | SE | Nov 2018 – May 2020

* + Migrated SSIS workflows to Azure Data Factory; learned and applied design patterns translatable to Dataflow and Cloud Composer
	+ Built automated validation scripts in Python and PyTest for data quality and regression testing
	+ Created DAX/SQL-driven dashboards in Power BI; experience applicable to BigQuery BI integrations
	+ Participated in Agile release cycles, production cutovers, and CI/CD deployment pipelines
	+ Developed reusable Python-based logging/alerting framework, integrated with Splunk, similar to GCP log-based alertingCollaborated in Agile teams for sprint planning, UAT, and production releases.
	+ Developed **test automation frameworks** in Python to validate data ingestion and processing for BAU pipelines
	+ Scheduled ETL and reporting jobs using **cron jobs** and automated logs/metrics collection
	+ Worked closely with stakeholders to conduct impact analysis, create test plans, and support UAT
	+ Facilitated production readiness and supported cutovers during high-priority release cycles
	+ Developed a reusable Python module for **log parsing and alerting**, integrated with Splunk and internal monitoring systems
	+ Used **SSIS and Azure Data Factory** to modernize legacy ETL jobs for cloud compatibility
	+ Conducted peer code reviews and contributed to team’s best practices for Git branching and version control
	+ Automated server health checks and ETL job validation scripts for pre-deployment testing phases

## Xceed Technologies, India | DE | September 2017 - Nov 2018

* + Built real-time ingestion pipelines using Apache Kafka and Spark (Java)
	+ Created complex transformation logic using Spark SQL and custom UDFs
	+ Improved data processing speeds by 40% through optimized transformation and partitioning strategies
	+ Integrated Git and Jenkins for CI workflows and developed validation scripts for data quality
	+ Implemented monitoring and alerting for data quality using log-based validation.
	+ Created custom Spark UDFs to implement complex transformation logic not natively available in Spark
	+ Maintained **Git** repositories for version control and collaborated with DevOps to integrate builds into **Jenkins**

pipelines

* + Implemented alerting scripts and job retry strategies for pipeline resilience and data quality enforcement
	+ Created documentation for onboarding, including architecture diagrams, data flow maps, and glossary of ETL terms
	+ Engaged in continuous integration practices and enhanced **ETL monitoring** via **dashboards** and alerting tools

# EDUCATION

* + - **Master of Science in Information Technology & Analytics –** Rutgers University, Newark, New Jersey, USA

### Graduate Assistant:

* Automated data workflows using **Python** and **SQL**, reducing manual data processing time by a few hours/week.
* Developed **Tableau** dashboards for financial data analysis, enabling non-technical stakeholders to track KPIs.
* Collaborated with faculty to design academic content using Adobe Illustrator and **MS Office Suite**.
* **Bachelor of Technology in Computer Science and Technology –** SNDT University, Mumbai, Maharashtra, India

### CERTIFICATION

* SQL Gold (Hackerrank), AWS CCP, Business Analysis Fundamentals (Udemy)