NIHARIKA

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

# SUMMARY

* 8 years of professional experience in **Data Engineering**, **Cloud Solutions**, and **Big Data Analytics** across finance, e-commerce, and technology domains
* Expertise in developing and automating **ETL pipelines**, real-time data processing, and orchestration of **data pipelines** using **Apache Spark**, **Kafka**, and **Airflow**
* Hands-on experience in **cloud computing** environments including **AWS (S3, Lambda, Redshift, EBS)** and

### Azure

* Proficient in **Python**, **SQL**, **Pandas**, **NumPy**, and using data-centric tools for large-scale **data manipulation and analysis**
* Designed and deployed **data lakes**, implemented **data warehousing** architectures, and optimized **Redshift** and

**Snowflake** queries for analytics use cases

* Developed **interactive dashboards** and self-service BI tools using **Tableau**, **Power BI**, and **Redash**, enabling timely, data-driven decision-making
* Experienced in managing **CI/CD pipelines** using **Git**, integrating with containerized deployments and automated testing

### Skilled in data modeling, feature engineering, data governance, metadata management, and cloud-native storage solutions

* Expertise in **Agile** software development methodologies with experience managing production support, incident management, and system troubleshooting

# SKILLS

* + **Programming: Python:** SQL, Java, R, C/C++
  + **Cloud & Big Data:** AWS (EBS, S3, Redshift, Lambda), Hadoop, Spark, Hive, Kafka
  + **Databases:** MySQL, MongoDB, Oracle Apex, SSIS
  + **Tools:** PyCharm, Jupyter, Git, Tableau, Power BI, SSRS
  + **Data Science:** Pandas, NumPy, Scikit-learn, TensorFlow
  + **Methodologies:** Agile, SDLC, ETL, Data Warehousin**g**
  + **Project Management Tools**: MS Project, MS Excel, MS SharePoint, Jira, Rally
  + **Analysis Skills:** Cost/Benefit, Impact, GAP, Risk, SWOT Analysis
  + **Documentation:** BRD, FRD, SRS, BPMN, UML, Use Cases, Test Plans
  + **Soft Skills:** Time Management, Leadership, Problem-Solving, Decision-Making

# EXPERIENCE

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

* + Developed end-to-end **ETL** pipelines using **Apache Spark** and **Python**, reducing batch processing latency.
  + Designed and implemented **AWS**-based data lake architecture with **S3**, **Lambda**, and **Redshift** Spectrum, supporting large-scale analytics
  + Built robust real-time data ingestion pipelines using **Kafka**, **Spark** Streaming, and AWS Glue for near real-time insights
  + Orchestrated **ETL** jobs using **Apache Airflow**, including DAG monitoring, logging, **SLA** alerts, and retries
  + Performed feature engineering and cleansing on petabyte-scale datasets using **Pandas**, **NumPy**, and custom Spark UDFs
  + Integrated machine learning models for predictive analytics using Spark MLlib and MapReduce patterns
  + Designed and published **Tableau** dashboards and KPI **visualizations** for operations and finance teams, reducing manual reporting by 60%
  + Conducted schema optimization and query tuning in **Amazon Redshift**, improving report response time by 30%
  + Designed data quality frameworks to validate batch and streaming data loads across multiple systems
  + 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

* + Built scalable SQL-based ETL workflows on **Amazon Redshift**, aggregating application logs into business-ready tables
  + Designed **Python-based automation scripts** using Boto3 for EBS snapshot creation, cross-region replication, and lifecycle archival
  + Built **data integrity validation tools** to ensure end-to-end consistency of snapshots and backups using **Python and SQL**
  + Designed **real-time dashboards in Redash**, pulling data from CloudWatch logs and Redshift views for production monitoring
  + Contributed to implementation of Git-based **CI/CD pipelines** integrated with code review and rollback automation
  + Enhanced data governance through metadata tracking and implemented data quality checks using Python assertions
  + Performed **root-cause analysis** for production incidents and authored runbooks to improve on-call response time
  + 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

* + Automated end-to-end deployment workflows across 100+ projects using **Robot Framework**, **Selenium**, and

### Shell Scripting

* + Built **PowerShell-based ETL scripts** and Azure Scheduler jobs to ingest and transform large datasets from various sources
  + Simulated event-driven test scenarios using **MySQL and Linux-based LAB servers**, enabling robust QA for mission-critical systems
  + Designed and published **Power BI dashboards** with DAX, providing clients real-time visibility into financial and operational data
  + Collaborated with cross-functional teams in Agile setup, managing incidents, testing workflows, and client demos
  + 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 robust real-time and batch **data pipelines** using **Apache Spark**, Kafka to ingest multi-source data into

**Hadoop** ecosystem

* + Developed HiveQL queries and Spark transformations to cleanse, join, and prepare data for **BI** and **ML**

workloads

* + Designed and deployed **interactive Power BI reports** powered by DAX for operations, sales, and customer success teams
  + Managed data transfers from **Oracle** and **MySQL** to **HDFS**, improving refresh cycles by 30%
  + Tuned Hadoop job configurations to optimize YARN container usage and reduce ingestion time by 25%
  + 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)