Nikhil PORIKA
Senior Data Engineer

(470) 851-4973 | LinkedIn
pnikhil1845@gmail.com

PROFESSIONAL SUMMARY:

Extensive experience in providing strategic business solutions to problems associated with big data Azure drafting solutions for better business opportunities. expertise lies in efficiently migrating on-premise ETLs to Microsoft Azure using cloud-native tools like Azure Data Factory, Databricks, and Azure Blob Storage. I designed and implemented data engineering solutions on the Azure cloud platform, harnessing Azure Data Factory, Azure Synapse Analytics, Azure Data Lake Storage (ADLS), and Azure Functions to create robust and scalable data pipelines. Proficient in various Hadoop distributions, including Cloudera and HortonWorks, I have hands-on experience in data import/export between RDBMS and HDFS using Sqoop. Skilled in Python, SQL, and ETL tools like Talend and Informatica, I excel in designing and implementing ETL workflows. I am adept at Python scripting, data visualisation with NumPy, Matplotlib, and Pandas, and microservices scaling with Kubernetes and Docker. My experience includes Test-Driven Development, Agile-Scrum methodologies, and strong communication and problem-solving skills.

TECHNICAL SKILLS:

Big Data/Hadoop Ecosystem	Apache Spark, HDFS, Map Reduce, HIVE, Sqoop, Oozie, Zookeeper, Kafka, Flume, IntelliJ
Programming Languages	Python, Scala, R, SQL, PL/SQL, Linux Shell Scripts
NoSQL Database	HBase, Cassandra, MongoDB, DynamoDB
Database	Oracle 11g/10g, MY SQL, MS-SQL Server, DB2, Teradata
Data Engineering	ETL processes, Data Warehousing, Data Pipelines, Data Modeling
Data Transformation	SQL, Python (ETL scripts), Java, Pandas, Spark
Data Visualization	Tableau, Power BI, Amazon QuickSight, Matplotlib, Seaborn
Web Technologies	HTML, XML, JDBC, JSP, CSS, JavaScript, SOAP
Tools Used	Eclipse, Putty, Winscp, NetBeans, QlikView, PowerBI
Operating Systems	Linux, Unix, Windows, Mac OS-X, CentOS, Red Hat
Methodologies	Agile/Scrum, Rational Unified Process and Waterfall
DevOps & CI/CD	Jenkins, Docker, Kubernetes, Terraform, Git
Distributed Platforms	Cloudera, Horton Works, MapR

EDUCATION:

- Master's degree from Northwest Missouri State University, Missouri
- Bachelor's degree from Kakatiya Institute of Technology and Science

CERTIFICATIONS:

- Microsoft Certified Azure Data Engineer Associate Certification Link
- Snowflake Hands-on Essentials Data Warehouse.

PROFESSIONAL EXPERIENCE:

Client: State of New Mexico, Albuquerque, New Mexico **Feb 2024 - Present**
Role: Sr. Data Engineer
Responsibilities:

- Designed ETL workflows using Azure Data Factory and Synapse pipelines to ingest raw CSVs from REST APIs and external databases into bronze tables in Azure Data Lake Storage.
- Created external tables for staging in Synapse and automated population of data from blob storage to support seamless integration.
- Applied schema evolution strategies in ADF pipelines to adapt to changing data structures without disrupting downstream processes.
- Developed bronze-silver-gold layer architecture.
- Bronze layer stored raw files with delta load support and version tracking.
- Silver layer Applied cleansing, null handling, and formatting using PySpark and SQL.
- Gold layer Performed aggregations and joins for curated datasets optimized for Power BI and SSRS reporting.
- Experience on Synapse Spark pools and PySpark to deduplicate household and child-level data using window functions and filtering logic.
- Implemented GDPR-compliant transformations, including data anonymization at the silver layer for sensitive fields.
- Developed census ETL workflows in Databricks to analyze child population trends across years using notebook-based pipelines with parameterized execution logic.
- Wrote PySpark transformations for household size, income level grouping, and demographic segmentations for yearly census snapshots.
- Worked on complex SQL queries in Spark SQL to join birth records, housing data, and school enrollments for population estimations.
- Used Python scripting to validate data integrity across multiple census feeds and send alerts on record count mismatches.
- Experience on reusable Databricks widgets and parameterized notebooks for batch-processing multiple counties and census years.
- Used SQL stored procedures and advanced T-SQL window functions for gold-layer analytics including demographic trend analysis and KPI calculations.
- Built and monitored scalable pipelines for child census records, enabling region-level segmentation and demographic analysis.

- Streamed real-time census updates through Kafka topics, using structured Kafka consumers for ingestion into the data lake.
- Migrated transformed datasets to Snowflake, tuning micro-partitions for faster performance in education-focused dashboards.
- Designed Synapse schemas with partitioning and indexing to support hierarchical census data models and boost query performance.
- Integrated alerting and error-handling modules into ingestion pipelines using Azure Monitor and Splunk for operational visibility.
- Created Power BI and Tableau dashboards visualizing population distribution, education levels, and birthrate trends.
- Built SSRS reports on SQL Server using T-SQL to support state compliance tracking and weekly census reporting.
- Extracted and ingested complex JSON data from government APIs using Python scripts and integrated validation checks using data quality rules.
- Implemented interactive dashboards using Power BI with semantic models and real-time connectivity to gold-layer datasets.
- Developed reusable parameterized pipelines for flexible ingestion by fiscal year and data type.
- Managed CI/CD using GitHub and Azure DevOps, automating build and release pipelines with YAML and DACPAC deployments.
- Maintained centralized metadata documentation and data dictionaries in Confluence to ensure governance and cross-team collaboration.
- Designed a Lakehouse architecture leveraging Delta Lake on Azure Databricks to unify structured and semi-structured data storage.

Environment: Azure, Azure Synapse Analytics, Azure Data Factory, Azure Data Lake Storage (ADLS), Azure Monitor, Azure Blob, Azure DevOps, Azure SQL, SQL Server, Spark, PySpark, Databricks, Delta Lake, Kafka, Snowflake, Git, GitHub, CI/CD, SSRS, Power BI, Tableau, Python, SQL, RDBMS, Kubernetes, Splunk, scikit-learn, Deep Learning.

Client: Crowe, Chicago, Illinois **Aug 2022 - Feb 2024**
Role: Sr. Data Engineer
Responsibilities:

- Extracted patient data from healthcare APIs and ingested nested JSON structures into Azure Data Lake using ADF pipelines for unified data processing.
- Designed and developed ETL pipelines in Azure Data Factory and Synapse Data Flows for efficient transformation and loading of petabyte-scale healthcare datasets.
- Built Spark-based transformations and leveraged PySpark for scalable processing of patient demographics, treatments, and diagnosis histories in Databricks.
- Implemented Lakehouse architecture using Delta Lake to store structured/unstructured healthcare data and enable seamless analytics in Azure Synapse.
- Experience on Azure Synapse SQL Pools with partitioning and clustering strategies to enhance performance of large-scale healthcare data queries.
- Integrated diverse on-prem and cloud healthcare data sources via Azure ADLS Gen2 and Azure Storage, ensuring secure and scalable storage.

- Migrated legacy healthcare data systems to Snowflake, time travel and automatic clustering for version control and query optimization.
- Built Power BI dashboards to visualize hospital KPIs, patient treatment patterns, and demographic trends, driving data-driven decisions for healthcare stakeholders.
- Developed and deployed CI/CD pipelines using Azure DevOps, integrating with Synapse and Databricks notebooks for real-time patient data workflows.
- Configured Azure Event Grid for real-time alerting on critical healthcare updates and integrated Azure Functions for trigger-based automation.
- Applied data masking in Azure Synapse for compliance with HIPAA regulations, securing sensitive patient identifiers in analytics workloads.
- Implemented Spark jobs on Databricks for real-time analysis of genomic data, supporting personalized medicine initiatives.
- Worked on Databricks workflows to orchestrate ingestion and transformation of multi-state Medicaid claims into clean analytical layers.
- Experience on PySpark window functions to identify gaps in patient coverage periods and categorize claim types efficiently.
- Worked on optimized SQL-based views in Delta Lake to enable real-time tracking of high-cost healthcare utilizations.
- Wrote Python scripts for handling encrypted PHI data, using hashing and masking libraries to ensure HIPAA compliance.
- Automated audit trail logging in Databricks using custom Python classes to monitor claim ingestion status and anomalies.
- Integrated Kafka with Databricks to build streaming pipelines, enabling ingestion of HL7/FHIR healthcare messages for live analytics.
- Worked on Kubernetes to deploy and orchestrate containerized Spark jobs, ensuring scalable and reliable data workflows.
- Designed and maintained Databricks Delta Lake for CDC (Change Data Capture), enabling incremental updates across patient care records.
- Developed Python and Scala-based ETL scripts for ingestion, cleansing, and validation of large volumes of healthcare data across platforms.
- Experience on Azure Synapse's integration with Azure Data Catalog and Purview for data lineage, traceability, and compliance tracking.
- Built and monitored Talend workflows for automated healthcare data integration, ensuring daily ingestion accuracy with minimal manual effort.
- Employed Splunk for monitoring data pipeline logs and performance metrics, ensuring system reliability and quick issue resolution.
- Designed interactive Tableau reports for executive summaries and patient care optimization across multiple healthcare facilities.

Environment: Azure, Azure Synapse Analytics, Azure Data Factory, ADLS Gen2, Azure DevOps, Azure Functions, Azure Event Grid, Databricks, Delta Lake, Power BI, Tableau, Python, PySpark, Scala, Spark, Snowflake, SQL Server, SSRS, Kafka, Git/GitHub, Kubernetes, CI/CD, RDBMS, Talend, Splunk, Deep Learning, scikit-learn

Client: Merck & Co, New Jersey, United States **Sep 2019 - Aug 2021**

Role: Sr. Data Engineer
Responsibilities:

- Developed real-time data pipelines for pharmaceutical research, ensuring instant delivery of critical insights for clinical decision-making.
- Extracted and processed complex nested healthcare data from IoT devices using REST APIs, enabling continuous patient monitoring.
- Built end-to-end data workflows using Azure Data Factory, ingesting and transforming healthcare data into Azure Data Lake for downstream analytics.
- Implemented Spark-based ETL jobs for processing real-time healthcare data streams, improving scalability and responsiveness of analytics pipelines.
- Maintained versioned healthcare datasets using Delta Lake to ensure compliance, traceability, and support for audit requirements.
- Migrated healthcare data into Snowflake for centralized storage and advanced analytics, optimizing data access across care teams.
- Designed and deployed CI/CD pipelines using Azure DevOps to automate deployment of healthcare analytics solutions across environments.
- Created Power BI dashboards visualizing patient health trends, treatment effectiveness, and clinical KPIs to support care quality initiatives.
- Integrated Azure Event Grid with Spark streaming for real-time ingestion and processing of IoT-generated patient data.
- Implemented dynamic schema evolution in ADF pipelines to support rapidly changing data formats in healthcare research datasets.
- Worked on Databricks Delta tables to unify patient surveys, EMRs, and appointment logs for holistic experience analysis.
- Developed PySpark pipelines to cleanse raw feedback data, tokenize free-text fields, and classify sentiment labels.
- Employed advanced SQL CTEs for aligning patient encounters with survey feedback across different providers and facilities.
- Created Python-based dashboards using plotly for visualizing real-time NPS and patient wait-time metrics.
- Experience in a dynamic environment switches in Databricks notebooks using Python-based config files for DEV/UAT/PROD workflows.
- Developed Azure Synapse SQL pools and optimized resource allocation to meet SLAs in a high-throughput healthcare transaction system.
- Integrated Azure Synapse with Azure Monitor and Key Vault for secure, monitored, and scalable healthcare data operations.
- Applied data lineage tracking and metadata documentation across Azure services to ensure compliance in pharmaceutical clinical trial data management.
- Defined and enforced data retention policies in Synapse dedicated SQL pools to automate archival and meet regulatory standards.
- Worked on Spark queries with data partitioning and enhanced them using Python UDFs for domain-specific healthcare transformations.

- Performed data profiling and validation in Spark workflows using Python, improving healthcare data quality and reliability.
- Experience on Spark-based data pipelines for scalable deployment on Kubernetes, enabling seamless management of resources.
- Configured Spark on Kubernetes to support dynamic scaling and handle high-velocity healthcare event streams.
- Developed secure data encryption mechanisms and integrated compliance checks for protected health information (PHI) using Scala and Spark.
- Worked on parameter-driven pipeline executions in ADF, enabling flexible data refreshes based on clinical study requirements.

Environment:Azure Data Factory, Azure Synapse Analytics, Azure Data Lake, Azure Monitor, Azure Key Vault, Azure DevOps, Power BI, Databricks, Apache Spark, Spark Streaming, Delta Lake, Scala, Python, Kubernetes, Snowflake, REST APIs, SQL, Git, Linux.

Client: The Home Depot, Georgia, United States **Dec 2017 - Aug 2019**
Role: Data Engineer
Responsibilities:

- Designed and deployed scalable retail data solutions using Azure PaaS for end-to-end data integration and analytics.
- Ingested and processed high-volume retail sales data from e-commerce platforms via REST APIs, transforming complex JSON structures into structured formats using Azure Data Factory and Azure Data Lake.
- Designed ETL pipelines using Azure Data Factory to integrate data from multiple retail systems, enabling unified sales and inventory tracking.
- Created CI/CD pipelines with Azure DevOps for automated deployment of retail analytics workflows.
- Implemented Delta Lake to manage structured and unstructured retail data, ensuring data consistency and reliability.
- Migrated retail data to Snowflake for centralized storage and conducted advanced analytics for customer behavior analysis.
- Applied data partitioning and indexing in Azure Synapse Analytics to optimize performance for retail sales forecasting.
- Developed custom UDFs in Scala within Databricks for specialized retail data transformations.
- Processed large-scale retail sales and transaction data using Apache Spark to generate insights into customer segmentation and sales trends.
- Developed interactive Power BI dashboards to visualize real-time sales performance, inventory updates, and customer demographics.
- Configured Azure Event Grid to trigger real-time inventory and sales updates for enhanced operational visibility.
- Migrated legacy ETL into Databricks notebooks to process point-of-sale data, integrating with Azure Data Lake Gen2.
- I worked on the PySpark code to calculate store-wise inventory turnover and real-time promotion effectiveness.

- Rewrote existing SQL Server queries into Spark SQL for faster execution over large parquet-based datasets.
- Designed custom Python modules for Azure Blob file metadata tracking, validation, and logging during batch uploads.
- Worked on configuring cluster autoscaling and error handling logic in Databricks for cost-effective pipeline execution.
- Worked on complex data workflows on Databricks combining batch and streaming data for fraud detection and promotional analysis.
- Used Spark on Kubernetes for dynamic scaling and containerized data pipelines for consistent and scalable deployment.
- Estimated Spark cluster sizing, monitored performance, and resolved bottlenecks in Databricks pipelines.
- Implemented Databricks Jobs for scheduled pipeline automation and monitoring.
- Experience on Azure Synapse with Databricks for cross-platform analytics and query optimization.
- Applied column-level security in Azure Synapse to restrict access based on roles in retail reporting scenarios.
- Implemented row-level security in Power BI for user-specific sales and revenue data visibility.
- Integrated Azure Key Vault with ADF for secure credential management in sensitive retail data pipelines.
- Developed data profiling and quality checks using Python within Spark to validate product and transaction data.
- Established automated data quality checks in ADF to ensure data reliability in promotional campaign reporting.
- Designed complex data transformations in ADF mapping data flows for customer segmentation and loyalty analytics.
- Worked on Snowflake storage and cost, ensuring continuous data availability in analytics use cases.
- Engineered Snowflake's zero-copy cloning for rapid creation of test environments.
- Integrated Snowflake external functions to extend analytics capabilities with third-party services.
- Migrated legacy SQL databases to Azure Data Lake, SQL Database, and Databricks for modernized retail reporting.
- Designed and maintained Teradata views and macros for efficient access and automation of reporting logic.
- Tuned Teradata queries using Explain plans and applied Axes-based routing for efficient data flow in IoT-based retail use cases.

Environment: Azure Data Factory, Azure Synapse Analytics, Azure Event Grid, Azure Key Vault, Azure DevOps, Azure SQL Database, Azure Data Lake, Databricks, Apache Spark, Scala, Python, Power BI, Snowflake, Kubernetes, Teradata, REST APIs, Delta Lake, ADF Mapping Data Flows, Axes, CI/CD.

Client: Unisys, Bangalore **July 2016 - Nov 2017**
Role: Data Engineer
Responsibilities:

- Migration of data marts to the centralised warehouse, employing SQL and SSIS for integration, and implementing Kafka, HDFS, and AWS.
- Developed efficient ETL processes using Python scripts, ensuring accuracy through unit testing and leveraging Spark, Hive, and Sqoop for data ingestion and analysis.
- I configured secure Kafka clusters with Kerberos and SSL and engineered custom Spark, Hive, and Sqoop adapters for diverse data sources, optimising data transfer to HDFS.
- Employed Sqoop, Flume, and Spark Streaming API for loading data from Web servers and Teradata, reducing patient expenses by 40%.
- Enhanced data processing efficiency with Informatica, creating mappings, sessions, worklets, and workflows.
- Implemented machine learning models (Logistic Regression, KNN, Gradient Boosting) using Python libraries, executing various algorithms based on client requirements.
- Analysed and optimised software tools expenses, achieving a 30% cost reduction.
- Innovated customer complaint chatbot, providing estimated resolution times and enhancing satisfaction.
- Successfully managed small projects, leading a team of 5, planning milestones, and ensuring project deliverables for a data mart migration project.

Environment: Python, AWS EMR, Apache Spark, Hadoop ecosystem (MapReduce, HDFS, Hive), Scala, LogRhythm, Openvas, Informatica.

Client: KPIT Technologies Ltd, Bangalore **March 2014 - Jun 2016**
Role: ETL Developer
Responsibilities:

- Designed databases with ER diagrams, applied normalisation, and ensured relational integrity.
- Developed SQL Server stored procedures, optimised queries, and created user-defined functions and views.
- I executed jobs in SAP Data Services, implemented IDOC in BODS, and facilitated data loading via LSMW and custom tools.
- Created Technical Design Documents (TDDs) for comprehensive code tracking and documentation.
- Implemented triggers for referential integrity, addressed exceptional scenarios, and wrote complex SQL queries for Crystal Reports.
- Automated jobs, tuned SQL queries using execution plans and profilers, and developed controller components using Servlets and EJBs.
- Established schedules, documented development efforts, and analysed system requirements for efficient project management.
- Developed dynamic interfaces with HTML, JavaScript, JSP, and Servlets, incorporating JMS elements for message handling.
- Created ETL packages with SSIS/DTS for diverse data sources, monitored and deployed packages, and maintained alerting.
- Conducted performance tuning exercises, including index and table rebuilding, and managed database backup and recovery.
- Contributed to system testing and user acceptance tests, mapping requirements to test cases in the Quality Center

Environment: MS SQL Server, SSRS, SSIS, SSAS, DB2, HTML, XML, JSP, Servlet, JavaScript, EJB, JMS, MS Excel, MS Word.