Nikhil Chinthalapally

Email: nikhikreddy137@gmail.com **Contact:9133337052**

Professional Summary:

- Experienced and proficient Data Engineer with 5+ years of experience in designing, building, and managing scalable data solutions on the Azure platform.
- Proven expertise in architecting data pipelines, building data lakes, and leveraging **Azure Data Factory, Azure Databricks**, **Azure Synapse Analytics, Azure Data Lake Gen2, Azure Event Hubs and Azure Blob Storage**to unlock actionable insights from data.
- Spearheaded the development and execution of robust **Data Integration and Loading** using **Azure Data Factory**, orchestrating seamless **extraction, transformation, and loading (ETL)** across diverse data sources.
- Integrated **Azure Logic Apps** into data workflow for seamless automation, allowing for event-driven data processing.
- Built ETL data pipelines using **PySpark, Spark SQL** and **Scala** to ingest, transform and load large volumes of structured and unstructured data.
- Expertise in defining roles and privileges for database access, by implementing **Identity and Access Management (IAM)** and **Role-Based Access Control (RBAC)** ensuring data security and appropriate access.
- Implemented **Azure Functions, Azure Storage,** and **Service Bus Queries** for large enterprise level ERP integration systems.
- Experience in streaming applications using **Azure Event hubs, Azure Stream Analytics, Azure Synapse Analytics.**
- Experience in creating and managing **Azure DevOps** tools for **continuous integration and deployment (CI/CD)** pipelines.
- Optimized **data ingestion, data modeling, data encryption** and performance by tuning ETL workflows.
- Implemented and managed data governance solutions using**Azure Key Vault, Azure Active Directory**, and**Azure Purview** to ensure data quality, compliance, and security.
- Developed reusable **Terraform modules** for managing Azure networking, allowing rapid configuration changes and consistent security policies.
- Skilled in creating and managing ETL/ELT workflows with **Apache Beam or Apache Airflow** to optimize data extraction, transformation and loading processes.
- Developed data ingestion workflows to read data from various sources and write it to **Avro, Parquet, Sequence, JSON, and ORC** file formats for efficient storage and retrieval.
- Expertise in Database Architecture for **OLAP and OLTP Applications**, **Database Designing**, **Data Migration**, and Data Warehousing concepts with emphasis on ETL.
- Experience building and optimizing large scale data pipelines with **Apache Hadoop, Java, HDFS, MapReduce, Hive, and Tez**.
- Experience in using **Apache Sqoop** to import and export data to and from Relational Database Systems and**HDFS**.
- Optimized Hadoop job executionusing **Oozie workflows** with **conditional branching and data dependencies**.
- Experience in optimizing query performance in **Hive using bucketing and partitioningtechniques**.
- Optimized Spark jobs and workflows by tuning Spark configurations, **partitioning and memory allocation settings**.
- Experienced in working with real-time streaming data using **Apache Kafka** as the data pipeline, and leveraging the Spark Streaming module for data processing.
- Well versed in using ETL methodology for supporting corporate-wide solution using **Informatica.**
- Expertise in using various Hadoop infrastructures such as **Map Reduce, Pig, Hive, Zookeeper, Sqoop, Oozie, Flume,**
- Proficient in querying and managing data using both **SQL and NoSQL databases** for diverse data architectures and mastered distributed database technologies like **Cassandra** for highly scalable and fault-tolerant data storage.
- Extensive experience in developing, maintaining, and implementing**Enterprise Data Warehouse** (**EDW), Data Marts, ODS** and **Data warehouse** with **Star schema and Snowflake schema.**
- Maintained and administered **Git** source code repository and **GitHub** enterprise.
- Comprehensive knowledge of **Software Development Life Cycle** and worked on **Agile and Waterfall Methodologies**.
- Collaborated seamlessly with cross-functional teams and stakeholders to implement well-aligned data models, structures, and designs.
- Dedicated to keeping up with the latest developments and industry best practices in cloud computing and data engineering technologies.

Education:

- Masters in Computer Science from University of Central Missouri

Technical Skills:

Big Data Technologies:	MapReduce, Hive, Tez, PySpark, Scala, Kafka, Spark streaming, Oozie, Sqoop, Pig, Zookeeper, HDFS
Hadoop Distribution:	Cloudera, Horton Works
Azure Services:	Azure Data Factory, Azure Data Bricks, Logic Apps, Azure Synapse Analytics, Azure Functions, Azure DevOps, Azure Event Hubs, Azure Data Lake, Polybase.
Languages:	SQL, PL/SQL, Python, HiveQL, Scala, Pyspark
Web Technologies:	HTML, CSS, JavaScript, XML, JSP, Restful, SOAP
Data Visualization Tools:	Tableau, Power BI
Operating Systems:	Windows (XP/7/8/10), UNIX, LINUX, UBUNTU, CENTOS.
File Formats:	CSV, JSON, XML, ORC, Parquet, Delta
Build Automation tools:	Ant, Maven, SBT
Version Control:	GIT, GitHub.
Methodology:	Agile, Scrum.
IDE &Build Tools, Design:	Eclipse, IntelliJ, Visual Studio.
Databases:	Azure SQL DB, MS Excel, MS Access, Oracle 11g/12c, Cosmos DB, MongoDB, HBase

Professional Experience:

Client:McKesson, Kansas City, MO **Sep 2022 to present**

Role: Data Engineer

Responsibilities:

- Analyzed data from Azure data storages using Databricks and Spark cluster capabilities, extracting valuable insights from large datasets.
- Developed and maintained end-to-end operations of ETL data pipeline and worked with large data sets in Azure Data Factory; increased data pipeline throughput by 2x.
- Developed custom activities using Azure Functions, Azure Databricks, and PowerShell scripts to perform data transformations, data cleaning, and data validation.
- Increased the efficiency of data fetching by using queries for optimizing and indexing.
- Implemented a tiered data lake architecture with Azure Data Lake Gen 2 (ADLS Gen 2) for batch and streaming data.
- Performed ETL tasks against replicated datasets to ensure quality, such as deduplication checks, null handling, and validation rules.
- Set up near real-time monitoring alerts and automated issue remediation workflows with Azure Monitor; cut incident response time by 50%.
- Worked with Azure Logic Apps administrators to monitor and troubleshoot issues related to process automation and data processing pipelines.
- Implemented Azure EventHub for real-time data ingestion, enabling efficient streaming and processing of high-volume data.
- Utilized data modeling and ETL processes to ensure data accuracy, and consistency with healthcare regulations, including HIPAA and GDPR.
- Configured and deployed Azure Purview services, including data catalog, data lineage, and data discovery capabilities.
- Enabled encryption-in-transit and at-rest using Azure Key Vault and accessed governance using Azure RBAC and Active Directory groups enabling cross-subscription data sharing while ensuring security compliance.
- Demonstrated expertise in using Azure Data Catalog to discover, and document various data assets across the organization, including databases, tables, files, and data lakes.
- Easily manage and categorize data stored in the Azure Data Lake Storage Gen2 by integrating with UnityCatalog.
- Leveraged Azure DevOps for continuous integration and deployment (CI/CD) of data pipelines and applications, streamlining the development and deployment processes.
- Created a secure network on Azure using NSGs, load balancers, autoscaling, and Availability Zones to ensure 99.95% uptime during peak data loads.
- Designed and deployed highly scalable and secure Azure data infrastructure using ARM templates, including data lakes, data warehouses, and data pipelines.
- Leveraged Power BI and Azure Analysis Services to deliver interactive dashboards and enable self-service analysis.
- Configured data pipeline orchestration using YAML pipelines in Azure DevOps, ensuring efficient and reliable execution of data workflows.
- Developed and implemented Change Data Capture (CDC) solutions using Azure Data Factory to capture changes from various sources.
- Utilized Terraform to automate infrastructure provisioning and management, ensuring consistent and reproducible deployments in an Azure environment.

- Leveraged Snowflake’s zero-copy cloning functionality to efficiently create development, testing, and production environments from a single master database, reducing storage costs and deployment time.
- Designed and implemented Directed Acyclic Graph (DAG) workflows to automate complex data processing task.
- Strong experience in developing Web Services like REST, RESTful APIsand Data Mining using Requests in Python with Jupyter notebooks.
- Collaborated closely with the data engineering team to enhance and optimize data pipelines, improving data processing speed and efficiency.
- Expertise in processing JSON, Avro, Parquet, ORC and CSV formats for efficient data ingestion transformation and storage.
- Orchestrated Docker containers for various applications, ensuring consistency across development, testing and production environments.
- Created and maintained HiveQL scripts and jobs using tools such as Apache Oozie and Apache Airflow.
- Created a Git repository and added Branching, Tagging and Release Activities on GitHub Version Control.
- Worked with JIRA to report on Projects, and creating sub tasks for Development, QA, and Partner validation.
- Experienced Agile ceremonies, from daily stand-ups to internationally coordinated PI Planning.

Environment: Azure Databricks, Data Factory, Azure Data Lake Gen 2, Logic Apps, Azure EventHub, Azure Purview,Azure Key Vault, Azure Active Directory, Azure Analysis Services,ELT/ETL, YAML, Spark Streaming, Data Pipeline, Terraform, Azure DevOps, PowerShell, Snowflake,Jenkins, Apache Oozie, Apache Airflow, Spark, Hive, SQL, Python, PySpark,PowerBI, GIT, JIRA, Agile.

Client:PNC, New York, NY **Sep 2020 to Aug2022**
Role: Big Data Engineer
Responsibilities:

- Utilized Sqoop to import data from MySQL to Hadoop Distributed File System (HDFS) on a regular basis, ensuring seamless data integration.
- Performed aggregations on petabytes of data using Apache Spark and Scala, and stored the processed data in Hive warehouse for further analysis.
- Managed Data Lakes and big data ecosystems, using Hadoop, Spark, to leverage their capabilities for efficient data processing.
- Oversaw the migration of vast volumes of data from diverse sources including Netezza, Oracle and SQL Server to Hadoop.
- Successfully loaded and transformed large sets of structured, semi-structured, and unstructured data, enabling effective analysis and insights generation.
- Developed Hive queries to analyze data and meet specific business requirements, utilizing Hive Query Language (HiveQL) to simulate MapReduce functionalities.
- Built HBase tables by leveraging HBase integration with Hive on the Analytics Zone, facilitating efficient storage and retrieval of data.
- Standardized fault-tolerant and scalable data processing solutions by leveraging technologies such as Apache Spark.
- Applied Kafka and Spark Streaming to process streaming data in specific use cases, enabling real-time data analysis and insights generation.
- Applied data visualization techniques and designed interactive dashboards using Power BI to present complex reports, bar, line, area, pie charts and graphs to team members and stakeholders.
- Developed custom scripts and tools using Oracle's PL/SQL language to automate data validation, cleansing, and transformation processes, ensuring data accuracy and quality.
- Experienced in loading logs from multiple sources into HDFS using Flume.
- Utilized JIRA and Confluence to manage project workflows, track issues, and for documentation.
- Implemented Spark using Python (PySpark) and Spark SQL for faster data testing and processing, enabling efficient data analysis and insights generation.
- Experience in integrating Apache Yet Another Resource Negotiator (YARN) with other Apache ecosystem tools and frameworks.
- Demonstrated expertise utilizing Apache Flink to create batch and real-time stream processing systems.
- Employed Spark Streaming to divide streaming data into batches as input to the Spark engine for batch processing, facilitating real-time data processing and analytics.
- Utilized Java-based frameworks such as Apache Camel to develop RESTful APIs and microservices for data integration and access.
- Utilized Zookeeper for coordination, synchronization, and serialization of servers within clusters, ensuring efficient and reliable distributed data processing.
- Spearheaded Oozie workflow engine for job scheduling, enabling seamless execution and management of data processing workflows.

- Utilized Data Analysis Expressions (DAX) to create complex calculations, measures, and calculated columns within PowerBI.
- Optimized Power Apps solutions for performance and scalability, including minimizing data latency.
- Utilized Kubernetes for container orchestration and scheduling, dynamic resource allocation, and automated deployment of data processing application.
- Engineered TensorFlow's data preprocessing functionalities to clean, transform, and prepare raw data.
- Leveraged Git as a version control tool to maintain code repositories, ensuring efficient collaboration, version tracking, and code management.
- Responsible for triggering the jobs using the Control-M.

Environment: Sqoop, PL/SQL, HDFS, Cloudera, Horton Works, Netezza, Hive Query Language, Apache Spark,Apache Flink, Apache YARN, Scala, Hive, Hadoop, HBase, Flume, Kafka, MapReduce, Zookeeper, Oozie, Java, RDBMS,DAX, Python, Power Apps,Control-M, Kubernetes, PySpark, TensorFlow, Git, JIRA, PowerBI.

Client: BNY Mellon, New York, NY **Sep 2018 to Aug 2020**
Role: Data Engineer
Responsibilities:

- Created pipelines, data flows and complex data transformations and manipulations using ADF and PySpark with Databricks which decreased the processing time by 20%.
- Established a foundation for real-time analytics with Azure IoT Hubs, Event Hubs to process millions of events per second into a data platform with near-zero latency or data loss.
- Designed and implemented data pipelines using Azure EventHub for real-time streaming and processing of terabytes of data.
- Executed advanced features of T-SQL to design and tune T-SQL to interface with Azure Synapse Analytics database and other applications in the most efficient manner and created stored procedures for the business logic using T-SQL.
- Identified and resolved bottlenecks in data pipelines and Spark jobs to improve system efficiency by 50%.
- Optimized database query performance using Python for caching, partitioning, and bucketing.
- Extensively worked on Azure Data Lake Analytics with the help of Azure Databricks to implement SCD-1, SCD-2 approaches.
- Data ingestion to Azure cloud services like Azure Data Lake, Azure Storage, Azure SQL, Azure DW, and cloud migration by processing the data in Azure Databricks.
- Designed and implemented Azure Data Lake Storage and Azure Blob Storage solutions for storing high-volume telemetry and time series data from IoT sensors.
- Utilized Azure Key Vault as central repository for maintaining secrets and referenced the secrets in Azure Data Factory and in Databricks notebooks.
- Developed Stream Analytics jobs with anomaly detection rules to identify and respond to issues pre-emptively based on changing conditions across connected systems.
- Expert in data modeling techniques, optimization of data storage and query performance in Synapse dedicated SQL pools.
- Familiarity with developing and executing SQL queries, scripts, and stored procedures within Azure Data Studio.
- Implemented data backup and disaster recovery strategies for Azure Blob Storage using Azure Backup and Azure Site Recovery.
- Used NumPy, Pandas, SciPy, and Pytables for ad-hoc analysis, data cleansing and preprocessing during development lifecycle.
- Proficient in deploying scalable ML models via Azure ML Service, including containerization, Azure Kubernetes Service (AKS), Azure Container Instances (ACI) managing versions, endpoints, and deployments.
- Experienced in using Azure Data Share to enable seamless collaboration and data exchange between cross-functional teams.
- Implemented DevOps culture utilizing GitHub workflows alongside Azure boards and blockages to increase data and site reliability engineering collaboration.
- Writing Spark and Spark SQL transformation in Azure Databricks to perform complex transformations for business rule implementation.
- Leveraged Azure Cosmos DB and Azure Blob Storage to efficiently store and manage large volumes of IoT data with high availability and scalability.
- Streamlined PolyBase in Azure environments for seamless integration and querying across diverse data sources, optimizing data engineering workflows.
- Implemented Delta Lake solutions for structured and semi-structured data in cloud data lakes, utilizing ACID transactions for data management.
- Conducted performance tuning and optimization of data pipelines and queries to improve overall system efficiency.
- Wrote Hive queries for data analysis to meet the specified business requirements by creating Hive tables and working on them using Hive QL to simulate MapReduce functionalities.
- Developed RDD's & Data frames (SparkSQL) using PySpark for analysing and processing the data.

- Leveraged Spark Streaming and Azure Functions to divide streaming data into batches as an input to Spark engine for batch processing.
- Utilized JIRA for error and issue tracking and added several options to the application to choose a particular algorithm for data and address generation.
- Used Git and Gitlab as version control tools to maintain the code repository while managing project tasks and issues.

Environment: Azure Databricks, Azure Data Factory, Azure Synapse Analytics, Spark, Azure Stream Analytics, Azure Machine Learning, Azure Functions, Poly Base, Azure Kubernetes Service, Azure Backup, Azure Site Recovery, Azure Container Instances, Azure Data Studio, Azure Key Vault, Azure IoT Hubs, Azure Data Lake, PolyBase, NumPy, PySpark, MapReduce, Spark, Hive, JIRA, Git, Gitlab.