Neha Golekar

Sr.Data Engineer

**Phone: +12167723762**

**Email: nehagolekar7@gmail.com**

Professional Summary

* Accomplished Senior Data Engineer with 10+ years of comprehensive IT experience, including 5+ years specializing in Data Engineering and 4+ years in Data Warehousing.
* Proven expertise in architecting and implementing enterprise-scale data solutions across cloud and on-premises environments.
* Expert in designing end-to-end ETL/ELT data pipelines with emphasis on scalability, reliability, and performance optimization.
* Proficient in real-time and batch data processing using Apache Spark, Kafka, and Azure Data Factory.
* Experienced in building robust data lake architectures and implementing modern data mesh principles.
* Advanced expertise in Microsoft Azure ecosystem (Data Factory, Databricks, Synapse, Logic Apps, Functions)
* Skilled in Snowflake data platform for cloud data warehousing and analytics.
* Proficient with Hadoop ecosystem (HDFS, Hive, HBase, MapReduce, YARN) and distributed computing frameworks.
* Implemented comprehensive data quality frameworks and automated validation processes.
* Experienced in data governance, lineage tracking, and metadata management.
* Expert in handling Slowly Changing Dimensions, change data capture, and data modeling best practices.
* Advanced proficiency in Python, Scala, SQL, PySpark, and Spark SQL for data processing and analytics.
* Skilled in developing reusable frameworks and optimizing code for high-performance data operations.
* Experience with Infrastructure as Code (IaC) and containerization technologies.
* Proficient in CI/CD pipeline implementation using Jenkins, Azure DevOps, and Git for automated deployments.
* Experienced in monitoring, logging, and troubleshooting complex data systems in production environments.
* Strong background in Agile methodologies, including Scrum ceremonies and cross-functional team collaboration.
* Stream processing and real-time analytics implementation.
* Data security and compliance frameworks (GDPR, SOX).
* API development and microservices architecture for data services.
* Performance tuning and cost optimization for cloud-based data solutions.
* Machine Learning pipeline integration and MLOps practices.
* Docker containerization and Kubernetes orchestration for data applications.
* Power BI dashboard development and advanced data visualization.
* Self-service analytics platform design and implementation.
* Cross-functional stakeholder management and requirements gathering.

Education

* Bachelors in Electronics and Communication Engineering from JNTUK, India.

| **Technical Skills** |  |
| --- | --- |
| **Category** | **Technologies** |
| Azure Services | Azure Data Factory, Airflow, Azure Databricks, Logic Apps, Functional App, Snowflake, Azure DevOps |
| AWS Services | AWS Glue, AWS EMR, AWS Lambda, AWS S3, AWS Redshift, AWS Kinesis, AWS CloudFormation, AWS IAM, AWS EC2, RDS |
| Big Data Technologies | MapReduce, Hive, Python, PySpark, Scala, Kafka, Spark Streaming, Oozie, Sqoop, Zookeeper |
| Hadoop Distribution | Cloudera, Horton Works |
| Languages | Java, SQL, PL/SQL, Python, HiveQL, Scala |
| Operating Systems | Windows (XP/7/8/10), UNIX, LINUX, UBUNTU, CENTOS |
| Build Automation Tools | Ant, Maven |
| Version Control | GIT, GitHub |
| IDE & Build Tools | Eclipse, Visual Studio |
| Databases | MS SQL Server 2016/2014/2012, Azure SQL DB, Azure Synapse, MS Excel, MS Access, Oracle 11g/12c, Cosmos DB |

Work Experience

**Azure Data Engineer | Jan 2023 - Till Now**

**Fulton Bank, Plano, TX**

**Responsibilities:**

* Managed end-to-end operations of ETL data pipelines, ensuring scalability and smooth functioning.
* Implemented optimized query techniques and indexing strategies to enhance data fetching efficiency.
* Utilized SQL queries, including DDL, DML, and various database objects (indexes, triggers, views, stored procedures, functions, and packages) for data manipulation and retrieval.
* Integrated on-premises (MySQL, Cassandra) and cloud-based (Blob storage, Azure SQL DB) data using Azure Data Factory, applying transformations and loading data into Snowflake.
* Orchestrated seamless data movement into SQL databases using Data Factory's data pipelines.
* Developed data warehousing techniques, data cleansing, Slowly Changing Dimension (SCD) handling, surrogate key assignment, and change data capture for Snowflake modelling.
* Designed and implemented scalable data ingestion pipelines using tools such as Apache Kafka, Apache Flume, and Apache Nifi to collect and process large volumes of data from various sources.
* Developed and maintained ETL/ELT workflows using technologies like Apache Spark, Apache Beam, or Apache Airflow, enabling efficient data extraction, transformation, and loading processes.
* Implemented data quality checks and data cleansing techniques to ensure the accuracy and integrity of the data throughout the pipeline.
* Built and optimized data models and schemas using technologies like Apache Hive, Apache HBase, or Snowflake to support efficient data storage and retrieval for analytics and reporting purposes.
* Developed ELT/ETL pipelines using Python and Snowflake Snow SQL to facilitate data movement to and from Snowflake data store.
* Created ETL transformations and validations using Spark-SQL/Spark Data Frames with Azure Databricks and Azure Data Factory.
* Collaborated with Azure Logic Apps administrators to monitor and resolve issues related to process automation and data processing pipelines.
* Optimized code for Azure Functions to extract, transform, and load data from diverse sources, including databases, APIs, and file systems.
* Designed, built, and maintained data integration programs within Hadoop and RDBMS environments.
* Implemented a CI/CD framework for data pipelines using the Jenkins tool, enabling efficient automation and deployment.
* Collaborated with DevOps engineers to establish automated CI/CD and test-driven development pipelines using Azure, aligning with client requirements.
* Demonstrated proficiency in scripting languages like Python and Scala for efficient data processing.
* Executed Hive scripts through Hive on Spark and SparkSQL to address diverse data processing needs.
* Collaborated on ETL tasks, ensuring data integrity and maintaining stable data pipelines.
* Utilized Kafka, Spark Streaming, and Hive to process streaming data, developing a robust data pipeline for ingestion, transformation, and analysis.
* Utilized Spark Core and Spark SQL scripts using Scala to accelerate data processing capabilities.
* Utilized JIRA for project reporting, creating subtasks for development, QA, and partner validation.
* Actively participated in Agile ceremonies, including daily stand-ups and internationally coordinated PI Planning, ensuring efficient project management and execution.

**Environment**: Azure Databricks, Data Factory, Logic Apps, Functional App, Snowflake, MS SQL, Oracle, HDFS, MapReduce, YARN, Spark, Hive, SQL, Python, Scala, PySpark, Spark Performance, data integration, data modeling, data pipelines, production support, Shell scripting, GIT, JIRA, Jenkins, Kafka, ADF Pipeline, Power Bi.

**Azure Data Engineer | Sep 2021 – Jan 2023**

**Medical Guardian, Philadelphia, PA**

**Responsibilities:**

* Enhanced Spark performance by optimizing data processing algorithms, leveraging techniques such as partitioning, caching, and broadcast variables.
* Implemented efficient data integration solutions to seamlessly ingest and integrate data from diverse sources, including databases, APIs, and file systems, using tools like Apache Kafka, Apache NiFi, and Azure Data Factory.
* Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in Azure Databricks.
* Worked on Microsoft Azure services like HDInsight Clusters, BLOB, Data Factory and Logic Apps and also done POC on Azure Data Bricks.
* Perform ETL using Azure Data Bricks, Migrated on premise Oracle ETL process to azure synapse analytics.
* Worked on Migrating SQL database to Azure data lake, Azure data lake analytics, Azure SQL Database, Data Bricks and Azure SQL Data warehouse
* controlling and granting database access and Migrating on Premise databases to azure data lake store using Azure Data Factory
* Data transfer using azure synapse and Polybase.
* Deployed and optimized Python web applications to Azure DevOps CI/CD to focus on development.
* Developed enterprise level solution using batch processing and streaming framework (using Spark Streaming, apache Kafka.
* Designed and implemented robust data models and schemas to support efficient data storage, retrieval, and analysis using technologies like Apache Hive, Apache Parquet, or Snowflake.
* Developed and maintained end-to-end data pipelines using Apache Spark, Apache Airflow, or Azure Data Factory, ensuring reliable and timely data processing and delivery.
* Collaborated with cross-functional teams to gather requirements, design data integration workflows, and implement scalable data solutions.
* Provided production support and troubleshooting for data pipelines, identifying and resolving performance bottlenecks, data quality issues, and system failures.
* Processed the schema oriented and non-schema-oriented data using Scala and Spark.
* Created Partitions, Buckets based on State to further process using Bucket based Hive joins.
* Created Hive Generic UDF's to process business logic that varies based on policy.
* Worked with Data Lakes and big data ecosystems (Hadoop, Spark, Hortonworks, Cloudera).
* Load and transform large sets of structured, semi structured, and unstructured data.
* Written Hive queries for data analysis to meet the Business requirements.
* 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 a data pipeline using Kafka, Spark, and Hive to ingest, transform and analyzing data.
* Worked on RDD’s & Data frames (SparkSQL) using PySpark for analyzing and processing the data.
* Implemented Spark Scripts using Scala, Spark SQL to access hive tables into spark for faster processing of data
* Used Spark Streaming to divide streaming data into batches as an input to Spark engine for batch processing
* Implemented CICD pipelines to build and deploy the projects in Hadoop environment.
* Using JIRA to manage the issues/project workflow.
* Worked on Spark using Python (PySpark) and Spark SQL for faster testing and processing of data.
* Used Git as version control tools to maintain the code repository.

**Environment**: Azure Databricks, Data Factory, Logic Apps, Functional App, Snowflake, MS SQL, Oracle, HDFS, MapReduce, YARN, Spark, Hive, SQL, Python, Scala, PySpark, Spark Performance, data integration, data modeling, data pipelines, production support, Shell scripting, GIT, JIRA, Jenkins, Kafka, ADF Pipeline, Power Bi.

**Data Engineer | Jun 2019 – Aug 2021**

**Costco, Dallas, TX**

**Responsibilities:**

* Designed and setup Enterprise Data Lake to provide support for various uses cases including Analytics, processing, storing and Reporting of voluminous, rapidly changing data.
* Responsible for maintaining quality reference data in source by performing operations such as cleaning, transformation and ensuring Integrity in a relational environment by working closely with the stakeholders & solution architect.
* Worked on creating tabular models on Azure analytic services for meeting business reporting requirements.
* Data Ingestion to one or more cloud Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and cloud migration processing the data in Azure Databricks.
* Creating pipelines, data flows and complex data transformations and manipulations using ADF and PySpark with Databricks.
* Working with Azure BLOB and Data Lake storage and loading data into Azure SQL Synapse analytics (DW).
* Developed Python, PySpark, Bash scripts logs to Transform, and Load data across on premise and cloud platform.
* Worked on Apache Spark Utilizing the Spark, SQL, and Streaming components to support the intraday
* and real-time data processing.
* Set up and worked on Kerberos authentication principals to establish secure network communication on cluster and testing of HDFS, Hive, Pig and Map Reduce to access cluster for new users.
* Used Spark SQL for Scala & amp, Python interface that automatically converts RDD case classes to schema RDD.
* Import the data from different sources like HDFS/HBase into Spark RDD and perform computations using PySpark to generate the output response.
* Implementing different performance optimization techniques such as using distributed cache for small datasets, partitioning, and bucketing in hive, doing map side joins etc.
* Good knowledge on Spark platform parameters like memory, cores and executors
* Developed reusable framework to be leveraged for future migrations that automates ETL from RDBMS systems to the Data Lake utilizing Spark Data Sources and Hive data objects.
* Importing & exporting database using SQL Server Integrations Services (SSIS) and Data Transformation Services (DTS Packages).

**Environment**: Azure, Azure Data Factory, Databricks, PySpark, Python, Apache Spark, HBase, HIVE, SQOOP, Snowflake, Python, SSRS, Tableau.

**Data Engineer | Jun 2017 – Jun 2019**

**Bank of America, Dallas, TX**

**Responsibilities:**

* Designed and developed the applications on the data lake to transform the data according business users to perform analytics.
* In depth understanding/ knowledge of Hadoop architecture and various components such as HDFS, application manager, node master, resource manager name node, data node and map reduce concepts.
* Involved in developing a Map Reduce framework that filters bad and unnecessary records.
* Involved heavily in setting up the CI/CD pipeline using Jenkins, Maven, Nexus, GitHub, and AWS.
* Developed data pipeline using flume, SQOOP, pig and map reduce to ingest customer behavioural data and purchase histories into HDFS for analysis.
* Used Spark-SQL to load JSON data and create schema RDD and loaded it into Hive tables handled structured data using Spark SQL
* Used HIVE to do transformations, event joins and some pre-aggregations before storing the data onto HDFS.
* The Hive tables created as per requirement were internal or external tables defined with appropriate static and dynamic partitions, intended for efficiency.
* Implemented the workflows using Apache OOZIE framework to automate tasks.
* Developing design documents considering all possible approaches and identifying best of them.
* Written Map Reduce code that will take input as log files and parse the and structures them in tabular format to facilitate effective querying on the log data.
* Developed scripts and automated data management from end to end and sync up b/w all the Clusters.
* Implemented Fair schedulers on the Job Tracker to share the resources of the cluster for the Map Reduce jobs given by the users.

**Environment**: Cloudera CDH 3/4, Hadoop, HDFS, MapReduce, Hive, Oozie, Pig, Shell Scripting, MySQL.

**Data warehouse Developer | Oct 2012 – Feb 2015**

**Global Logic, Hyderabad, India**

**Responsibilities:**

* Create and maintain database for Server Inventory, Performance Inventory.
* Worked in Agile Scrum Methodology with daily stand up meetings, great knowledge working with Visual SourceSafe for Visual studio 2010 and tracking the projects using Trello.
* Generated Drill through and Drill down reports with Drop down menu option, sorting the data, and defining subtotals in Power BI.
* Used Data warehouse for developing Data Mart which for feeding downstream reports, development of User Access Tool using which users can create ad-hoc reports and run queries to analyze data in the proposed Cube.
* Deployed the SSIS Packages and created jobs for efficient running of the packages.
* Expertise in creating ETL packages using SSIS to extract data from heterogeneous database and then transform and load into the data mart.
* Involved in creating SSIS jobs to automate the reports generation, cube refresh packages.
* Great Expertise in Deploying SSIS Package to Production and used different types of Package configurations to export various package properties to make package environment independent.
* Experienced with SQL Server Reporting Services (SSRS) to author, manage, and deliver both paper-based and interactive Web-based reports.
* Developed stored procedures and triggers to facilitate consistent data entry into the database.
* Shared data outside using Snowflake to quickly set up to share data without transferring or developing pipelines.

**Environment**: Windows server, MS SQL Server 2014, SSIS, SSAS, SSRS, SQL Profiler, Power BI, C#, Performance Point Server, MS Office, SharePoint.