NIHARIKA GANJI

Big Data Engineer

Email: niha21463@gmail.com **Dallas (TX)**

Ph: +1 972-813-9772

LinkedIn https://www.linkedin.com/in/niharika-g-34b061237/

PROFESSIONAL SUMMARY

- Over 7+ years of IT experience specializing in Big Data technology using Cloudera and Hortonworks.
- Possess a strong understanding of Hadoop architecture and comprehensive knowledge of Hadoop-Daemons and its various components such as HDFS, YARN, Resource Manager, Node Manager, Name Node, Data Node, and MapReduce programming paradigm.
- Proficient in programming languages including Scala, Java, Python, SQL, T-SQL, and R.
- Skilled in configuring and installing Hadoop/Spark Ecosystem Components.
- Involved in building Data Models and Dimensional Modeling with 3NF, Star and Snowflake schemas for OLAP and Operational data store (ODS) applications and experience on Essbase.
- Hands-on experience in building and deploying enterprise applications using key Hadoop ecosystem components such as MapReduce, YARN, Hive, HBase, Flume, Sqoop, Spark MLlib, Spark GraphX, Spark SQL, and Kafka.
- Utilized Sqoop for migrating data between RDBMS, NoSQL databases, and HDFS.
- Strong background in Data Engineering, Data Pipeline Design, Development, Documentation, Deployment, and Integration as a Data Engineer/Data Developer and Data Modeler.
- Proficient in various development environments, including Eclipse, IntelliJ IDE, PyCharm IDE, Notepad++, and Visual Studio.
- In-depth knowledge of data architecture, encompassing data ingestion pipeline design, Hadoop/Spark architecture, data modeling, data mining, machine learning, and advanced data processing.
- Expertise in working with Hive data warehouse tool, including table creation, data distribution through partitioning and bucketing, as well as writing and optimizing HiveQL queries.
- Experienced in collecting real-time streaming data, creating data pipelines using Kafka, and storing data in HDFS and NoSQL databases using Spark.
- Proficient in Text Analytics and developing Statistical Machine Learning and Data Mining solutions using R, SAS, and Python. Skilled in generating data visualizations using tools such as Tableau, Matplotlib, Seaborn, ggplot2, and Plotly.
- Experienced in ETL transformations using AWS Glue and AWS Lambda to trigger and process events.
- Experience in creating REST APIs and performing CRUD operations (post, put, get) using curl.
- Well-versed in the Software Development Life Cycle (SDLC) with good knowledge of testing methodologies, disciplines, tasks, resources, and scheduling.
- Extensive experience in Shell/Python scripting and setting up production systems in UNIX/LINUX environments. Comprehensive understanding of developing MapReduce and Streaming jobs using

Scala and Java for data cleansing, filtering, and data aggregation. Possesses detailed knowledge of the MapReduce framework.

- Proficient in developing and implementing web applications using Java, J2EE, JSP, Servlets, HTML, JSON, jQuery, CSS, XML, JDBC, JNDI, and Web Services (REST, SOAP) with frameworks such as Spring and Struts.
- Experience in designing error and exception handling procedures for identifying, recording, and reporting errors.
- Proficient in requirements gathering, system analysis, and handling business and technical issues, with the ability to communicate effectively with both business and technical users.

TECHNICAL SKILLS

Big Data/Hadoop Technologies	Map Reduce, Sqoop, Hive, Oozie, Impala, Zookeeper, Ambary, Storm, Spark Streaming, Spark, Hadoop, HDFS, Flume, pig, Kafka, Pyspark, HBase, Sqoop, NiFi, Yarn, Sparklib, Dataiku
Languages	Python, SQL, Java, R, Pyspark, C, C++, PowerShell, Shell Scripting, Scala, PySpark
Hadoop Distributions	Cloudera, EMR and Horton Works
Operating Systems	UNIX, LINUX, Ubuntu, Windows Vista/7/8/10
Cloud Technologies	AWS (EC2, S3, Redshift, Lambda, RDS, EBS, cloud watch), Azure (Azure Data Factory, Data Lake,), Google Cloud Platform
IDE and Notebooks	Eclipse, IntelliJ, PyCharm, Jupiter, Databricks notebooks
Databases	Oracle, MySQL, SQLServer, Cassandra, Teradata, HBase, MongoDB, PostgreSQL.
Web/Application servers	Apache Tomcat, WebLogic, JBoss
Java Technologies	Servlets, JDBC, Spring, Hibernate, SOAP/REST services
BI tools	Power BI, Data Studio, Tableau
Web Technologies	HTML, XML, JSON, CSS, jQuery, JavaScript
Methodologies	Software Development Lifecycle (SDLC), Waterfall, Agile

PROFESSIONAL EXPERIENCE

Schlumberger, Houston, Texas, United States October 2022 - Present

Designation: Data Engineer

Responsibilities:

- Created Pipelines in AD using Linked Services/Datasets/Pipeline/ to Extract, Transform, and load data from different sources like Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool and backwards.

- Experience in building ETL (Azure Data Bricks) data pipelines leveraging PySpark, Spark SQL.
- Demonstrated expertise in Cognite Data Fusion, leveraging its capabilities to enable seamless integration, management, and analysis of industrial data.
- Experienced in developing flows, extractors and setting up automatic triggers with the help of scenarios and deploying the flows to higher environments in Dataiku.
- Experience in Developing Spark applications using Spark - SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
- Successfully implemented Cognite Data Fusion to consolidate and harmonize data from multiple sources, ensuring data quality and reliability for advanced analytics and decision-making processes.
- Orchestrated data integration pipelines in ADF using various Activities like Get Metadata, Lookup, For Each, Wait, Execute Pipeline, Set Variable, Filter, until, etc.
- Analyzed large datasets to identify operational inefficiencies and proposed data-driven solutions to improve production performance.
- Analyze, design, and build Modern data solutions using Azure PaaS service to support visualization of data. Understand current Production state of application and determine the impact of new implementation on existing business processes.
- Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL, and U-SQL Azure Data Lake Analytics.
- Data Ingestion to one or more Azure Services - (Azure Data Lake Azure Storage, Azure SQL, Azure DW) and processing the data in In Azure Databricks.
- Developed Spark applications using Pyspark and Spark-SQL for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
- Responsible for estimating the cluster size, monitoring, and troubleshooting of the Spark databricks cluster.
- Experienced in performance tuning of Spark Applications for setting right Batch Interval time, correct level of Parallelism and memory tuning.
- Used ETL to implement the Slowly Changing Transformation, to maintain Historically Data in Data warehouse.
- Performed ETL testing activities like running the Jobs, Extracting the data using necessary queries from database transform, and upload into the Data warehouse servers.
- Created ETL packages with different data sources (SQL Server, Flat Files, Excel source files, XML files) and then loaded the data into destination tables by performing different kinds of transformations using SIS/DTS packages.
- Performed Data transformations, cleaning, and build model using Dataiku.
- Utilized Cognite Data Fusion's data contextualization features to enhance data visibility, enabling efficient monitoring and analysis of industrial assets and processes.

Environment:

Azure Databricks, Azure Data Lake, MS SQL Server 2016, T-SQL, Flexible Data Modeling(FDM), SQL Server Integration Services (SSIS), Azure SQL database, Azure SQL Datawarehouse**,** SQL Server 2017, Programming Scala, Python, Spark SQL, Data Visualization**,** Data Migration**,** SQL Server programming**,** Cognite Data Fusion, Azure Databases, Azure Devops, Azure Repos, Pyspark, Delta-lake, Azure Data-warehouse, Azure Data Factory(ADF), Data Lake Storage (ADLS), Analysis Services (AAS), Databricks (DBRX), PowerBI, Azure Automation Accounts, Runbooks, Webhooks, SparkSQL.

Securian Financials, Minnesota, United States January 2020 – October 2022

Designation: Big Data/ Data Engineer

Responsibilities:

- Using AWS Redshift, extracted, transformed, and loaded data from various heterogeneous data sources and destinations.
- Extracted, transformed, and loaded data from various heterogeneous data sources and destinations like Access, Excel, CSV, Oracle, flat files using connectors, tasks and transformations provided by AWS Data Pipeline.
- Utilized Spark SQL API in PySpark to extract and load data and perform SQL queries.
- Worked on AWS Data pipeline to configure data loads from S3 into Redshift.
- Worked with AWS cloud and created EMR clusters with spark for analyzing raw data processing and access data from S3 buckets.
- Stored customer details and Transactions info to Hive for better Business analysis and Marketing.
- Maintain log data with Kafka consumes them and process using pyspark and store the historical data to Datawarehouse Hive.
- Primarily involved in Data Migration using SQL, SQL Azure, Azure Storage, and Azure Data Factory, SSIS, PowerShell.
- Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Python and Scala.
- Performed data analysis, design, and created, maintained large, complex logical and physical data models, and metadata repositories using ERWIN and MB MDR.
- Designed SSIS Packages to extract, transfer, load (ETL) existing data into SQL Server from different environments for the SSAS cubes (OLAP).
- Using SQL Server reporting services (SSRS), created & formatted Crosstab, Conditional, Drill-down, Top N, Summary, Form, OLAP, Sub reports, ad-hoc reports, parameterized reports, interactive reports & custom reports.
- Performed ETL testing activities like running the Jobs, Extracting the data using necessary queries from database transform, and uploading into the Data warehouse servers and Pre-processing is performed using Hive and Pig.
- Architect & implement medium to large scale BI solutions on Azure using Azure Data Platform services (Azure Data Lake, Data Factory, Data Lake Analytics, Stream Analytics, Azure SQL DW, HD Insight / Databricks, NoSQL DB).
- Developed a detailed project plan and helped manage the data conversion migration from the legacy system to the target snowflake database.
- Migration of on-premises data (Oracle/ SQL Server/ DB2/ MongoDB) to Azure Data Lake and Stored (ADLS) using Azure Data Factory (ADF V1/V2).

Environment:

MS SQL Server 2016, T-SQL, SQL Server Integration Services (SSIS), SQL Server Reporting Services (SSRS), SQL Server Analysis Services (SSAS), Management Studio (SSMS), Advance Excel (creating formulas, pivot tables, HLookup, VLOOKUP, Macros), Spark, Python, ETL, Power BI, Tableau, Hive/Hadoop, Snowflakes, Power BI, AWS Data Pipeline, IBM Cognos 10.1, Data Stage, Cognos Report Studio 10.1, Cognos 8 & 10 BI, Cognos Connection, Cognos office Connection, Cognos 8.2/3/4, Data stage and Quality stage 7.5, ZooKeeper, Kafka, Pyspark.

Webster Bank, Waterbury (CT), March 2019 - December 2019

Designation: Big Data Engineer

Responsibilities:

- Implemented Apache Airflow for authoring, scheduling, and monitoring Data Pipelines.
- Experience in building and architecting multiple Data pipelines, end to end ETL and ELT process for Data ingestion and transformation in GCP.
- Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
- Experience in building and architecting multiple Data pipelines, end to end ETL and ELT process for Data ingestion and transformation in GCP and coordinating tasks among the team.
- Used cloud shell SDK in GCP to configure the services Data Proc, Storage, Big Query.
- Using rest API with Python to ingest Data from and some other site to BIGQUERY.
- strong understanding of AWS components such as EC2 and S3.
- Worked with g-cloud function with Python to load Data into Big Query for on arrival csv files in GCS bucket.
- Implemented a Continuous Delivery pipeline with Docker and GitHub.
- Build a program with Python and Apache beam and execute it in cloud Dataflow to run Data validation between raw source file and Big Query tables.
- Submit spark jobs using gsutil and spark submission get it executed in the Dataproc cluster.
- Implemented UDFs, UDAFs, UDTFs in java for hive to process the data that can't be performed using Hive inbuilt functions.
- Effectively used Oozie to develop automatic workflows of Sqoop, MapReduce and Hive jobs.
- ETL transformations using pyspark and Spark SQL and store the data to Hive.
- Written Shell scripts with 2 logging features to automate jobs and scheduled with Autosys
- Deployed and extracted data using Microsoft Azure into Netezza.
- Performed regression testing for integral code releases.
- Worked on confluence and Jira skilled in data visualization like Matplotlib and seaborn library.
- Hands-on experience with big data tools like Hadoop, Spark, Hive.
- Wrote a Python program to maintain raw file archival in GCS bucket.
- Designed several DAGs (Directed Acyclic Graph) for automating ETL pipelines.
- Performed Data Analysis, Data Migration, Data Cleansing, Transformation, Integration, Data Import, and Data Export through Python.
- Devised simple and complex SQL scripts to check and validate Data Flow in various applications.
- Involved in gathering and processing raw data at scale (including writing scripts, web scraping, calling APIs, writing SQL queries, writing applications).
- Developed a near real time data pipeline using spark.
- Developed and deployed data pipelines in the cloud such as AWS and GCP.
- Process and load bound and unbound Data from Google pub/subtopic to BigQuery using cloud Dataflow with Python.
- Responsible for data services and data movement infrastructures and good experience with ETL concepts, building ETL solutions and Data modeling.
- Performed data engineering functions such as data extraction, transformation, loading, and integration in support of enterprise data infrastructures - data warehouse, operational data stores and master data management.

- Involved in gathering and processing raw data at scale (including writing scripts, web scraping, calling APIs, writing SQL queries, writing applications).
- Hands on experience on architecting the ETL transformation layers and writing spark jobs to do the processing.
- Devised PL/SQL Stored Procedures, Functions, Triggers, Views, and packages. Made use of Indexing, Aggregation and Materialized views to optimize query performance.
- Experience implementing machine learning back-end pipeline with Pandas and NumPy.

Environment: GCP, BigQuery, GCS Bucket, G-Cloud Function, Apache Beam, Cloud Dataflow, Cloud Shell, Gsutil, Docker, Kubernetes, AWS, Apache Airflow, Python, Pandas, Matplotlib, seaborn library, text mining, NumPy, Scikit-learn, Heat maps, Bar charts, Line charts, ETL workflows, linear regression, multivariate regression, Python, Scala, Spark

VNC Technologies, Hyderabad, March 2016 – October 2018

Designation: Big Data Engineer

Responsibilities:

- Performed data extraction, transformation, loading, and integration in data warehouse, operational data stores and master data management.
- Implemented Apache Airflow for authoring, scheduling, and monitoring Data Pipelines.
- Responsible for data services and data movement infrastructures.
- Performed Data Migration to GCP.
- Experienced in ETL concepts, building ETL solutions and Data modeling.
- Aggregated daily sales team updates to send reports to executives and to organize jobs running on Spark clusters and loaded application analytics data into data warehouses in regular intervals of time.
- Day to-day responsibility includes developing ETL pipelines in and out of data warehouse, develop major regulatory and financial reports using advanced SQL queries in snowflake.
- Leveraged cloud and GPU computing technologies for automated machine learning and analytics pipelines, such as AWS, GCP.
- Designed & built infrastructure for the Google Cloud environment from scratch.
- Worked on confluence and Jira.
- Implemented one time data migration of multistate level data from SQL severe to Snowflake by using python and SnowSQL.
- Experienced in fact dimensional modeling (Star schema, Snowflake schema), transactional modeling and SCD (Slowly changing dimension).
- Involved in preparing associated documentation for specifications, requirements, and testing.
- Designed and implemented configurable data delivery pipeline for scheduled updates to customer facing data stores built with Python.
- Proficient in Machine Learning techniques (Decision Trees, Linear/Logistic Regressors) and Statistical Modeling.
- Implemented a Continuous Delivery pipeline with Docker, and GitHub and AWS.
- Compiled data from various sources to perform complex analysis for actionable results.
- Participated in the full software development lifecycle with requirements, solution design, development, QA implementation, and product support using Scrum and other Agile methodologies.

- Measured Efficiency of Hadoop/Hive environment ensuring service-level agreement (SLA) is met.
- Collaborated with team members and stakeholders in design and development of data environment.
- Analyzed the system for new enhancements/functionalities and perform Impact analysis of the application for implementing ETL changes.

Environment: AWS, GCP, Big Query, GCS Bucket, G-Cloud Function, Apache Beam, Cloud Dataflow, Cloud Shell, Gsutil, Dataproc, Cloud SQL, MySQL, Postgres, SQL Server, Python, Scala, Spark, Hive, Spark -SQL.

Vedicsoft, Hyderabad, June 2015 – February 2016

Designation: Data Analyst

Responsibilities:

- Designed an ETL strategy to transfer data from source to landing, staging, and destination in the data warehouse using SSIS and DTS (Data Transformation Service).
- Involved in designing and managing schema objects such as Tables, Views, Indexes, Stored Procedures, Triggers and maintained referential integrity using SQL Server Management Studio.
- Created many complex stored procedures, triggers, functions, indexes, views with T-SQL statements in SQL Server Management Studio (SSMS).
- Processed ETL to transfer data from remote data centers to local data centers using SSIS. Cleansing, messaging of the data is done on the local database.
- Used reporting service (SSRS) created several reports based on cube and local data warehouse.
- Developed SSIS packages to export data from Excel/Access to SQL Server, automated all the SSIS packages and monitored errors using SQL Server Agent Job.
- Interacted with subject matter experts in understanding the business logic, implemented complex business requirements in backend using efficient stored procedures and flexible function.
- Stage the API or Kafka Data (in JSON file format) into snowflake DB by Flattening the same for different functional services.
- Involved in designing and managing schema objects such as Tables, Views, Indexes, Stored Procedures, Triggers and maintained referential integrity using SQL Server Management Studio.
- Prepared reports using SSRS (SQL Server Reporting Service) to declare discrepancies between user expectations and service efforts involved in scheduling the subscription reports via the subscription report wizard.

Environment: MS SQL Server Management Studio 2008 R2/2012 (SSMS), Snowflake, SQL Server Integration Service (SSIS), SQL Server Reporting Services (SSRS), MS Visual Studio 2012, SQL Server Profiler.