Data Engineer

Nithya Koritala
Phone: +1 (660)-858-8082
Email: nithya.koritala07@gmail.com

Professional Summary:

- Over 6+ years of experience as a Data Engineer in Analysis, Design, Development, Testing, Implementation, Maintenance and Enhancements on various IT Projects.
- Hands on experience in installing, configuring, and using Apache Hadoop ecosystem components like Hadoop Distributed File System (HDFS), MapReduce, PIG, HIVE, HBASE.
- Experience in writing Spark applications using Scala.
- Experience in converting Hive queries into Spark transformations using Spark RDDs and Scala.
- Experienced in handling large datasets using Partitions, Spark in Memory capabilities, Broadcasts in Spark, Effective & efficient Joins, Transformations and other during ingestion process itself.
- Experience in designing, developing, scheduling reports/dashboards using Tableau.
- Experience in optimizing MapReduce Programs using combiners, partitioners, and custom counters for delivering the best results.
- Proficiently leveraged Oracle PL/SQL within data project for efficient database management, querying, and data manipulation tasks.
- Experience in creating modules for spark streaming in data into Data Lake using Spark.
- Experience with designing Python-based ETL pipelines using BigQuery, managing JSON and CSV formats for efficient data transformation.
- Experience with using VSCode for developing and optimizing Python ETL scripts, leveraging IntelliSense and version control for improved efficiency.
- Experience with automating data workflows using Apache Airflow, optimizing scheduling and execution for high-performance operations.
- Experience with managing database infrastructure with Cloud Run and Cloud SQL, integrating PostgreSQL for seamless data access.
- Experience with implementing CI/CD pipelines using Bamboo, automating build, test, and deployment for continuous delivery across environments.
- Experience in Dimensional Data Modeling Star Schema, Snow-Flake Schema, Fact and Dimensional Tables, concepts like Lambda Architecture, and Batch processing.
- Experience in Extraction, Transformation and Loading (ETL) data from various sources into Data Warehouses, as well as data processing like collecting, aggregating, and moving data from various sources using Apache Flume, Kafka, Power BI and Microsoft SSIS.
- Experience with Informatica (ETL Tool) for Data Extraction, Transformation and Loading.
- Experience with SSIS, Power BI Desktop, Power BI Services) Interactions, DAX.
- Hands on experience in setting up workflow using Apache Airflow.
- Experience in Microsoft Azure providing data movement and scheduling functionality to cloud-based technologies such as Azure Blob Storage and Azure SQL Database.
- Experience working with Azure Blob Storage, Azure Data Lake, Azure Data Factory, Azure SQL, Azure SQL Datawarehouse, Azure Analytics, PolyBase, Azure HDInsight, Azure Databricks.
- Experience in SQL and PL/SQL for development of Procedures, Functions, Packages and Triggers.
- Experience in Data Modeling with expertise in creating Star & Snow-Flake Schemas, FACT and Dimensions Tables, Physical and Logical Data Modeling using Erwin.
- Experience in using different file formats like Text files, CSV, Parquet, and JSON.
- Experience in managing MongoDB environment from availability, performance, and scalability perspectives.
- Experience in administration activities of RDBMS data bases, such as MS SQL Server.
- Extensive experience in agile software development methodology.
- Team Player as well as able to work independently with minimum supervision, innovative & efficient, good in debugging and strong desire to keep pace with latest technologies.
- Excellent Communication and presentation skills along with good experience in communicating and working with various stake holders.

Technical Skills:

Databases	Snowflake, AWS RDS, Teradata, Oracle, MySQL, Microsoft SQL, PostgreSQL, BigQuery
NoSQL Databases	MongoDB, Hadoop HBase and Apache Cassandra.
Programming Languages	Python, SQL, Scala, MATLAB.
Cloud Technologies	Azure, AWS, GCP, Docker, Cloud Run, Cloud SQL, Vertex AI
Data Formats	CSV, JSON
Querying Languages	SQL, NO SQL, PostgreSQL, MySQL, Microsoft SQL
Integration Tools	Jenkins, BitBucket, GitHub, Bamboo, SonarQube
Scalable Data Tools	Hadoop, Hive, Apache Spark, Pig, Map Reduce, Sqoop.
Operating Systems	Red Hat Linux, Unix, Windows, macOS.
Reporting & Visualization	Tableau, Matplotlib.

Professional Experience:

Fifth Third Bank | Cincinnati, Ohio **March 2023 – Present**
Data Engineer

Responsibilities:
- Spearheaded the design and optimization of Python-based ETL pipelines utilizing BigQuery, effectively managing JSON and CSV data formats to drive efficient transformation and processing of large datasets.
- Developed, tested, and optimized Python scripts for ETL pipelines and data processing workflows using VSCode, leveraging IntelliSense and version control integrations to enhance development efficiency and code quality.
- Designed and Deployed AWS Solutions using EC2, S3, EBS, Elastic Load balancer (ELB), auto-scaling groups and OpsWorks.
- Utilized Bitbucket for version control and collaborative code management while maintaining high standards of code quality through continuous integration and code review with SonarQube.
- Writing Pig and Hive scripts with UDF in MR and Python to perform ETL on AWS Cloud Services.
- Performed Tableau type conversion functions when connected to relational data sources.
- Involved in converting Map Reduce programs into Spark transformations using Spark RDD's using Scala and Python.
- Performed transformations using Python and Scala to analyze and gather the data in required format.
- Involved in making Hive tables, stacking information, composing hive inquiries, producing segments and basins for enhancement.
- Written Terraform scripts to automate AWS services which include ELB, CloudFront distribution, RDS, EC2, database security groups, Route 53, VPC, Subnets, Security Groups, and S3 Bucket and converted existing AWS infrastructure to AWS Lambda deployed via Terraform and AWS Cloud Formation.
- Imported data from AWS S3 into Spark RDD, Performed transformations and actions on RDD's.
- Created Map Reduce programs to handle semi/unstructured data like xml, json and sequence files for log files.
- Reviewed basic SQL queries and edited inner, left, and right joins in Tableau Desktop by connecting live/dynamic and static datasets.
- Utilized Oracle PL/SQL for data processing tasks, including data extraction, transformation, and loading (ETL).
- Developed PL/SQL scripts to perform complex data manipulations and transformations within the Oracle database environment.
- Analyzed the SQL scripts and designed the solution to implement using PySpark.
- Extracted files from MongoDB through Sqoop and placed in HDFS and processed.
- Collaborated with database administrators to optimize PL/SQL code for performance and scalability.
- Created scripts to read CSV, JSON, and parquet files from S3 buckets in Python and load into AWS S3, DynamoDB and Snowflake.
- Orchestrated and automated data workflows with Apache Airflow, optimizing job scheduling, dependency management, and execution to ensure high-performance data operations and reliable pipeline execution.
- Monitored and ensured data pipeline health using Dataflow, proactively identifying and resolving quality issues to guarantee seamless and reliable data delivery while optimizing performance.
- Led the management of database infrastructure leveraging Cloud Run and Cloud SQL, while integrating external data sources using PostgreSQL, ensuring high availability and accessibility of data across platforms.
- Successfully executed the migration of project data from Bamboo to GitHub as part of a proof of concept, enhancing version control, collaboration, and repository management for the project team.
- Involved in Requirement gathering phase to gather the requirements from the business users to continuously accommodate changing user requirements.

Environment: Spark, Scala, Hadoop, Python, PySpark, AWS, MapReduce, Pig, ETL, HDFS, Hive, HBase, SQL, Agile and Windows.

MetLife | New York City, NY **Oct 2022 – Feb 2023**
Data Engineer

Responsibilities:
- Involved in Requirements and Analysis: Understanding the requirements of the client and the flow of the application.
- Developed spark applications for data transformations and loading into HDFS using RDD, Data Frames and Datasets.
- Worked on customized UDF's in Spark for eliminating duplicate columns, identifying alphanumeric variables.
- Designed and Developed Scala workflows for data pull from cloud-based systems and applying transformations on it.
- Developed Spark streaming application to pull data from cloud to Hive table.
- Worked on 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, HDInsight/Databricks, NoSQL DB).
- Developed shell scripts for ingesting the data to HDFS and partitioned the data over Hive.
- Involved in designing optimizing Spark SQL queries, Data frames, import data from Data sources, perform transformations; perform read/write operations, save the results to output directory into HDFS.
- Developed Pig Latin scripts to extract the data from the web server output files to load into HDFS.
- Built tools using Tableau to allow internal and external teams to visualize and extract insights from big data platforms.

- Involved in loading the data from HDFS, have done transformations using spark SQL and Data frames and stored results back to HDFS using Spark.
- Worked on data pre-processing and cleaning the data to perform feature engineering and performed data imputation techniques for the missing values in the dataset using Python.
- Wrote, compiled, and executed programs as necessary using Apache Spark in Scala to perform ETL jobs with ingested data.
- Worked on ETL Processing which consists of data transformation, data sourcing and mapping, Conversion, and loading.
- Used various sources to pull data into Power BI such as SQL Server, Excel, Oracle, SQL Azure etc.
- Optimized MapReduce Jobs to use HDFS efficiently by using various compression mechanisms.
- Worked with building data warehouse structures, and creating facts, dimensions, aggregate tables, by dimensional modeling, Star and Snowflake schemas.
- Automated resulting scripts and workflow using Apache Airflow and shell scripting to ensure daily execution in production.
- Design and implement database solutions in Azure SQL Data Warehouse, Azure SQL.
- Performed transformations, cleaning and filtering on imported data using Hive, Map Reduce, and loaded final data into HDFS.
- Install and configure Apache Airflow for S3 bucket and Snowflake data warehouse and created DAGs to run the Airflow.
- Followed agile methodology and involved in daily SCRUM meetings, sprint planning, showcases and retrospective.
- 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 Azure Databricks.
- Worked on Dockers containers by combining them with the workflow to make them lightweight.
- Data sources are extracted, transformed, and loaded to generate CSV data files with Python programming and SQL queries.
- Worked on MongoDB by using CRUD (Create, Read, Update and Delete), Indexing, Replication and Shading features.
- Followed agile methodology for the entire project.

Environment: Spark, Scala, Azure, ETL, Kafka, Tableau, Hadoop, Python, Snowflake, HDFS, Airflow, Hive, MapReduce, PySpark, Pig, Docker, Sqoop, Teradata, JSON, MongoDB, SQL, Agile and Windows.

Solugenix | Hyd, India **April 2019 – July 2021**
Data Engineer

Responsibilities:
- Gathering business requirements, business analysis and design various data products.
- Developed spark applications for performing large scale transformations and denormalization of relational datasets.
- Used Scala to convert Hive / SQL queries into RDD transformations in Apache Spark.
- Written Programs in Spark using Scala for Data quality check.
- Developed various spark applications using Scala to perform various enrichment of these click stream data merged with user profile data.
- Developed Spark code in Python and Spark SQL environment for faster testing and processing of data and Loading the data into Spark RDD and doing In-memory computation to generate the output response with less memory usage.
- Implemented and optimized CI/CD pipelines using Bamboo, automating build, test, and deployment processes to ensure efficient integration and delivery across multiple environments, while also enhancing version control and collaborative code management.
- Applied OpenAI for advanced natural language processing capabilities to drive enhanced functionality for chatbots, improving user interaction and data understanding.
- Integrated Vertex AI Search to streamline data retrieval and facilitate efficient knowledge exploration, ensuring faster and more accurate chatbot responses.
- Containerized data processing applications and workflows using Docker, ensuring consistent deployment environments, improved scalability, and simplified management across development, testing, and production stages.
- Adopted Agile methodologies with Jira, efficiently managing project timelines, tasks, and resources, ensuring the timely and high-quality delivery of all milestones.
- Developed Simple to complex MapReduce Jobs using Hive and Pig.
- Profile structured, unstructured, and semi-structured data across various sources to identify patterns in data and Implement data quality metrics using necessary queries or python scripts based on source.
- Worked on PySpark APIs for data transformations.
- Prepared dashboards using Tableau for summarizing Configuration, Quotes, Orders, and other e - commerce data.
- Extract, transform, and load (ETL) data from multiple federated data sources (JSON, relational database, etc.) with Data Frames in Spark.
- Leveraged a comprehensive suite of GCP Cloud services to build scalable, cloud-native solutions, streamline operations, and optimize workflows, ensuring robust performance and cost-efficiency across the project lifecycle.

- Developed Spark scripts by using Python shell commands as per the requirement.
- Worked with No-SQL databases like HBase in creating HBase tables to load large sets of semi-structured data coming from various sources.
- Involved in daily SCRUM meetings to discuss the development/progress and was active in making scrum meetings more productive.

Environment: Python, PostgreSQL, ETL, GCP, CI/CD, BitBucket, BigQuery, Dataflow, Bamboo, Cloud Run, Cloud SQL, Vertex AI, SonarQube, GitHub, Apache Airflow, VSCode, Jira.

Mastercard | Hyd, India **June 2017 – March 2019**
Data Analyst

Responsibilities:
- Performed Data analysis, Data Profiling and Requirement Analysis.
- Developed automated processes for flattening the upstream data from Cassandra which in JSON format. Used Hive UDFs to flatten the JSON Data.
- Optimized MapReduce Jobs to use HDFS efficiently by using various compression mechanisms
- Involved in managing and reviewing Hadoop log files.
- Developed various Python scripts to find vulnerabilities with SQL Queries by doing SQL injection, permission checks and analysis.
- Conducted data analysis and profiling using PL/SQL queries to identify patterns and trends in large datasets.
- Created Hive tables and involved in data loading and writing Hive UDFs. Developed Hive UDFs for rating aggregation
- Implemented Map Reduce programs to handle semi/unstructured data like XML, JSON, Avro data files and sequence files for log files.
- Use SQL queries and other tools to perform data analysis and profiling.
- Worked on SQL queries in dimensional data warehouses and relational data warehouses. Performed Data Analysis and Data Profiling using Complex SQL queries on various systems.
- Designing NoSQL schemas in HBase.
- Involved in weekly walkthroughs and inspection meetings, to verify the status of the testing efforts and the project as a whole.

Environment: Spark, Scala, Hive, JSON, MapReduce, Hadoop, Python, XML, NoSQL, HBase, and Windows.