Devi Karri

Data Engineer

313-649-5067|Devi46k@gmail.com

PROFESSIONAL SUMMARY

- Having 4+ years of experience in IT and extensively worked in Technologies like Azure Databricks, Azure Data Factory (ADF), Azure Data Lake (ADLS Gen2), Azure Delta Lake, Azure synapse dedicated SQL pool, Spark, Scala, Python, SQL, SSIS.
- Skilled in developing Data pipelines utilizing Azure Data Factory for extraction, Azure Data Flow for transformation, and loading data from varied sources to Azure SQL Database and Azure SQL Data Warehouse.
- Excel skills at an advanced level, encompassing Power Pivot, Power Query, and mastery of advanced formulas.
- Proficiency in various AI tools, including but not limited to Python, Java, or R, coupled with expertise in machine learning frameworks like Spark, TensorFlow, or scikit-learn.
- Extensive experience in ETL methodology for performing Data Profiling, Data Migration, Extraction, Transformation and Loading using Talend and designed data.
- Good Knowledge in Amazon AWS concepts like EMR and EC2, S3, Lambda, Redshift web services which provides fast and efficient processing of Big Data.
- Good understanding of Data Modeling (Dimensional and Relational) concepts like Star-Schema Modeling, Snowflake Schema Modeling, Fact and Dimension Tables.
- Excellent analytical, problem solving and interpersonal skills. Ability to learn new concepts fast. Consistent team player with excellent communication skills.
- Expertise in using various Hadoop infrastructures such as Map Reduce, Pig, Hive, HBase, and spark for data storage and analysis.
- Experience in various cloud vendors like AWS, GCP and Azure.
- Extensively used Spark Scala APIs to build data pipelines.
- Experienced in running query - using Impala and used BI tools to run ad-hoc queries directly on Hadoop.
- Experience in designing and testing highly scalable mission-critical systems, and Spark jobs both in Scala and pySpark, Kafka.
- Design and develop ETL processes in AWS Glue to migrate campaign data from external sources like S3, ORC/Parquet/text files into AWS Redshift.
- Good understanding of Big Data concepts like Hadoop, Map - Reduce, YARN, Spark, RDD, Data frames, Datasets, Streaming.
- Involved in Snowflake utilities like Snow SQL, Snow pipe, Connectors, Data Sharing, Cloning, and creating tasks.
- Skilled on streaming data using Apache Spark, migrating the data from Oracle to Hadoop HDFS using Sqoop.
- Design and maintain scalable ETL pipelines to process and analyze large volumes of data from diverse sources.
- Hands on in troubleshooting errors in HBase Shell/API, Pig, Hive and map Reduce.
- Experience in various cloud vendors like AWS, GCP and Azure.
- Experience in managing multi-tenant Cassandra clusters on public cloud environment - Amazon Web Services (AWS)-EC2.
- Developed Reusable solutions to maintain proper coding standards across different java projects. Very good in Application Development and Maintenance of SDLC projects using different programming languages such as Java, C, Scala, SQL, and NoSQL.

- Worked in different phases of Software Development Life Cycle (SDLC) with methodologies such as Agile/Scrum methodologies, and other best practices with specific focus on the build and release of quality software.
- Utilize Azure Data Factory and Databricks for data integration and transformation tasks.
- Monitor and troubleshoot data pipelines to ensure reliability and performance.

Education

- Master of Science in Computer Science from the University of North Texas

TECHNICAL SKILLS

Languages	Java, Scala, Python, SQL, and C/C++
Big Data Ecosystem	Hadoop, MapReduce, Kafka, Spark, Apache Spark, Pig, Hive, YARN, Flume, Sqoop, Oozie, Zookeeper, Talend.
Hadoop Distribution	Cloudera Enterprise, Data Bricks, Horton Works, EMC Pivotal.
Databases	Oracle, SQL Server, PostgreSQL
Visualization Tools	Dataiku, Power BI, Tableau 9.4/9.2
Web Technologies	HTML, CSS, JSON, JavaScript, Ajax
Streaming Tools	Kafka, RabbitMQ, Kinesis
Cloud	AWS, Azure, AWS EMR, Glue, RDS, Kinesis, DynamoDB, Redshift Cluster
Testing	Hadoop Testing, Hive Testing
Application Servers	Apache Tomcat, JBOSS, WebSphere
Tools and Technologies	Servlets, JSP, Spring (Boot, MVC, Batch, Security), Web Services, Hibernate, Maven, GitHub, Bamboo.
CI/CD Tools	Jenkins, GitLab CI/CD, Docker

PROFESSIONAL EXPERIENCE

Client: Thrivent, Minneapolis, MN **Jun 2022- Present**
Role: Data Engineer

Responsibilities:

- Evaluated business requirements and prepared detailed specifications that follow project. guidelines required to develop written programs.
- Experience in developing Spark programs in Scala to perform Data Transformations, creating Datasets, Data frames, and writing spark SQL queries, spark streaming, windowed streaming application.
- Used AWS Glue catalog with crawler to get the data from S3 and perform SQL query operations.
- Extensively involved working on Hive, created the Hive tables and loaded data consuming event data from Kafka using Spark Streaming.
- Follow deployment process and CI/CD to ensure the code is properly tested before deploying to production.
- Implement data quality checks and ensure the integrity, reliability, and security of our data.
- Worked with DevOps team to Culturize NIFI Pipeline on EC2 nodes integrated with Spark, Kafka, Postgres running on other instances using SSL handshakes in QA and Production Environments.
- Developed Spark scripts, Spark SQL query for data aggregation, querying, and writing data back into RDBMS through Sqoop.
- Copy Fact Dimension and aggregate output from S3 to Redshift for Historical data analysis using Tableau and Quick sight.

- Developed Producer API and Consumer API to publish and subscribe to stream of events in one or more topics.
- Extensive hands-on experience in developing and implementing Apache Spark applications for large-scale data processing.
- Created DAG to use the Email Operator, Bash Operator, and spark Livy operator to execute and in EC2 instance.
- Created Data Quality Scripts using SQL and Hive to validate successful das ta load and quality of the data.
- Created various types of data visualizations using Python and Tableau.
- Wrote Python scripts to automate data extraction and transformation tasks
- Worked Loading and transforming sets of Structured, Semi-Structured and Unstructured data and
- Implemented CI/CD pipelines with Azure DevOps, integrating Docker images into the deployment workflow for seamless application delivery.
- Involved in performance tuning the application at various levels, Hive, Spark, etc.
- Conducting research and testing to create machine learning algorithms and predictive models.
- Developed Data pipeline using Spark, Hive and HBase to ingest data into Hadoop cluster for analysis.
- Collected data using Spark Streaming from AWS S3 bucket in batch and real time and performs necessary transformations and aggregations to build the common learner data model and persist the data in HDFS.
- Exploring with the Spark improving the performance and optimization of the existing algorithms in Hadoop using spark context, Spark SQL, Data Frame, Spark Yarn
- Designed the ETL runs performance tracking sheet in different phases of the project and shared with the production team.
- Performs quality check on the existing code to improve performance. Imported the data from different sources like AWS S3, Local file system into Spark RDD.
- Involved in converting Hive/SQL queries into Spark Transformations using Spark RDDs and python, Used Hive to analyze the partitioned and Bucketed data and compute various metrics for reporting.
- Implemented data lake solutions on AWS S3 and managed data workflows using AWS Glue.
- Involved in loading data from Linux file system to HDFS, involved in data warehousing and Business Intelligent systems, member of identifying and designing most efficient and cost-effective solution through research and evaluation of alternatives.
- Demonstrated Hadoop practices and knowledge of technical solutions, design patterns and code for medium/ complex applications deployed in Hadoop production.
- Integration of data storage solutions in spark - especially with Azure Data Lake storage and Blob snowflake storage.
- Worked on migration of data from On-prem SQL server to Cloud databases (Azure Synapse Analytics (DW) & Azure SQL DB).
- Experience in developing Spark applications using Spark-SQL in Data bricks for data extraction, transformation, and aggregation from multiple file formats for Analyzing& transforming the data to uncover insights into the customer usage patterns.
- Analyzing the Data from different sourcing using Big Data Solution Hadoop by implementing Azure Data Factory, Azure Data Lake, Azure Data Lake Analytics, HDInsight, Hive, and Sqoop.
- Designing and maintaining reports in Power BI, built on top of Azure Synapse/Azure Data Warehouse, Azure Data Lake, Azure SQL.
- Developed ETL pipelines in and out of data warehouse using a combination of Python and Snowflakes Snow SQL Writing SQL queries against Snowflake.
- Participated in code reviews and collaborated with team members to ensure best practices in data engineering.

Environment: Azure ADF, Power BI, MSBI, SQL Server, Apache Impala, Apache Spark, SQL Server Integration Services (SSIS), ETL, Microsoft Power BI, Apache Sqoop, Azure Data Lake Store, Apache Kafka, Apache Impala, Azure Logic Apps, Azure Synapse Analytics (formerly Azure SQL Data Warehouse), Apache Sqoop, Apache Hive, Microsoft SQL Server Integration.

Client: Oracle (Remote) **Sep 2017-Jan 2020**
Role: Data Engineer

Responsibilities:

- Experience in developing Spark programs in Scala to perform Data Transformations, creating Datasets, Data frames, and writing spark SQL queries, spark streaming, windowed streaming application.
- Created on demand tables on S3 files using Lambda functions and AWS Glue using Python and Pyspark.
- Developed complex Transact SQL queries and SSIS packages to load the data into warehouse.
- Create data ingestion modules using AWS Glue for loading data in various layers in S3 and reporting using Athena and Quick sight.
- Experience enhancing CI/CD build tooling in a containerized environment, from deployment pipelines (Jenkins, etc.), infrastructure as code (Terraform, CloudFormation), and configuration management via Docker and Kubernetes
- Created Data Quality Scripts using SQL and Hive to validate successful das ta load and quality of the data.
- Interacted with Data architects, Business Analysts, and users to understand business and functional needs.
- Designed SSIS packages to transfer data between the servers, load data into database and scheduled the jobs to do these tasks periodically.
- Creating Spark jobs efficiently with data cache, coalesce, repartition methods to improve performance.
- Building Jenkin pipelines as part of DevOps model. Performed pre-and post-session scripts in Informatica mappings
- Utilizes big data computation and storage tools for prototyping and dataset creation, including model training and evaluation.
- Develops and oversees internal reporting for stakeholders and senior management. Developed Rest API to serve the data generated by prediction model to serve other customers/teams.
- Developed a data quality control model to monitor business information change overtime. The model flags outdated customer information using different Apis for validation and updates it with correct data.
- Developed and deployed microservices architecture using Docker containers, ensuring efficient scalability and resource utilization.
- Worked with Business Analysts to understand the business requirements and implemented the solutions.
- Developed Spark Streaming application to consume JSON messages from Kafka and perform transformations on the data. Implemented Spark using Scala and Spark SQL for faster processing of data and loading into the data lakes.
- Implemented pySpark scripts to perform extraction of required data from the datasets and storing it on HDFS.
- Worked using Hadoop ecosystem components like HDFS, Kafka, Spark, Hive, Sqoop, Oozie. Involved in creating Hive tables, loading the data, and writing HQL queries
- Created Hive external, internal tables and implemented partitioning, dynamic partitions, and bucketing in Hive for efficient data access
- Developed AWS lambdas using Python &amp; Step functions to orchestrate data pipelines.

- Used various transformations like Source qualifier, Aggregators, lookups, Filters, Sequence generators, Routers, Update Strategy, Expression, Sorter, Normalizer, Union etc.
- Created a cleanup process for removing all the Intermediate temp files that were used prior to the loading process.
- Involved in business meetings to gather requirements, business Analysis, Design, review and Development, testing
- Worked on an Azure copy to load data from an on-premises SQL server to an Azure SQL Data warehouse.
- Worked on redesigning the existing architecture and implementing it on Azure SQL.
- Experience with Azure SQL database configuration and tuning automation, vulnerability assessment, auditing, and threat detection.

Environment: Extract, Transform, Load (ETL), ADF, Big Data, Transact-SQL (T-SQL), Python (Programming Language), Microsoft AZURE, Apache Kafka, Apache Airflow, Azure Synapse Analytics, Azure Stream Analytics, Azure Logic Apps, Apache Hive, Azure HDInsight, Apache Spark Structured Streaming, Microsoft Power Automate, Jupiter Notebook, Apache HBase, Data Bricks Delta Lake, Azure Monitor, Grafana, Kibana, and Microsoft Azure Machine Learning.