**Dasharathram Kota**

**Sr. Data Engineer**

**dasharathram8688@gmail.com**

**443-583-8660**

**SUMMARY:**

* Around 10 years of overall IT experience in a variety of industries, which includes hands-on experience in Big Data and Data warehouse ETL technologies.
* Have a proven track record of working as a Data Engineer on Amazon cloud services, BigData/Hadoop Applications and product development.
* Well-versed with Big data on AWS cloud services i.e. EC2, S3, Glue, Anthena, DynamoDB, and RedShift
* Hadoop Ecosystem, AWS Cloud Data Engineering, Data Visualization, Reporting, and Data Quality Solutions.
* Good experience in Amazon Web Services like S3, IAM, EC2, EMR, Kinesis, VPC, Dynamo DB, RedShift, Amazon RDS, Lambda, Athena, Glue, DMS, Quick Sight, Amazon Elastic Load Balancing, Auto Scaling, CloudWatch, SNS, SQS and other services of the AWS family.
* Hands-on experience in Data Analytics Services such as Athena, Glue, Data Catalog & Quick Sight.
* Hands-on expertise with AWS Databases such as RDS(Aurora), Redshift, DynamoDB and Elastic Cache (Memcached & Redis).
* Experience in developing Hadoop-based applications using HDFS, MapReduce, Spark, Hive, Sqoop, HBase and Oozie.
* Hands-on experience in Architecting Legacy Data Migration projects on-premises to AWS Cloud.
* Wrote AWS Lambda functions in Python for AWS's Lambda which invokes Python scripts to perform various transformations and analytics on large data sets in EMR clusters.
* Experience in building and optimizing AWS data pipelines, architectures, and data sets.
* Hands-on experience with tools like Hive for data analysis Sqoop for data ingestion and Oozie for scheduling.
* Experience in scheduling and configuring the Oozie and also having good experience in writing Oozie workflow and coordinators.
* Worked on different file formats like JSON, XML, CSV, ORC, and Paraquet. Experience in processing both structured and semi-structured Data with the given file formats.
* Worked on Apache Spark performing the Actions, Transformations on RDDs, Data Frames & Datasets using Spark SQL and Spark streaming contexts.
* Having good experience in spark core, spark SQL and spark streaming.
* Having good experience in writing Python Lambda functions and calling the API’s.
* Good knowledge of Kafka and Flume.
* Experience in building ETL(Azure Data Bricks) data pipelines leveraging PySpark, Spark SQL.
* Experience in Java, Java EE (J2ee) technologies and proficiency in Core Java, Servlets, JSP, EJB, JDBC, XML, and spring, Struts and Hibernate and RESTful Webservices.
* Proven knowledge of standards-compliant, cross-browser compatible HTML, CSS, JavaScript, and Ajax.
* Having good experience in different SDLC models including Waterfall, V-Model and Agile.
* Involved in Daily stand-ups sprint planning and review meetings in the Agile model.
* Experience in job/workflow scheduling and monitoring tools like Oozie, AWS Data pipeline & Autosys
* Defined and deployed monitoring, metrics, and logging systems on AWS.
* Experience working on creating and running Docker images with multiple micro-services.
* Docker container orchestration using ECS, ALB and lambda.
* Experience with Unix/Linux systems with scripting experience and building data pipelines
* Experience in Cloud Databases and Data warehouses ( SQL Azure and Confidential Redshift/RDS )
* Played a key role in migrating Cassandra, Hadoop clusters on AWS and defined different read/write strategies
* Strong SQL development skills including writing Stored Procedures, Triggers, Views, and User Defined functions.
* Expert in developing SSIS/DTS Packages to extract, transform and load (ETL) data into data warehouse/data marts from heterogeneous sources.
* Good understanding of software development methodologies, including Agile (Scrum).
* Expertise in development of various reports, dashboards using various Tableau Visualizations
* Hands-on experience with different programming languages such as Java, Python, R, SAS
* Experience in using different Hadoop ecosystem components such as HDFS, YARN, MapReduce, Spark, Pig, Sqoop, Hive, Impala, HBase, Kafka, and Crontab tools.
* Expert in creating HIVE UDFs using Java to analyze data sets for complex aggregate requirements.
* Experience in developing ETL applications on large volumes of data using different tools: MapReduce, Spark-Scala, PySpark, Spark-SQL, and Pig.
* Experience in using SQOOP for importing and exporting data from RDBMS to HDFS and Hive.
* Created user-friendly GUI interface and Web pages using HTML, CSS and JSP
* Experience with MS SQL Server, including SSRS, SSIS, and T-SQL.

**TECHNICAL SKILL MATRIX:**

|  |  |
| --- | --- |
| **Hadoop Components / Big Data** | PySpark, Airflow, HDFS, MapReduce, Hive, HCatalog, HBase, Sqoop, Impala, Zookeeper, Kafka, Yarn. |
| **Programming Languages** | Python, Scala, SQL, PySpark, PowerShell, T-SQL. |
| **Cloud Platform** | AWS (Lambda, S3, EC2, EMR, RDS), Microsoft Azure, (Azure Databricks, Azure Data Factory, Azure Data Explorer, Azure HDInsight, ADLS), GCP |
| **Reporting and ETL Tools** | AWS GLUE, Tableau, Power BI. |
| **Databases** | Oracle, SQL Server, MS Access, NoSQL Database (HBase, MongoDB, DynamoDB) |
| **Big Data Technologies** | Hadoop, HDFS, Hive, Oozie, Sqoop, Spark, Machine Learning, Pandas, NumPy, Seaborn, Impala, Zookeeper, Airflow, Informatica, Snowflake, Data Bricks, Kafka, Cloudera |
| **Data Analysis Libraries:** | Pandas, NumPy, SciPy, Scikit-learn, Matplotlib |
| **Containerization** | Docker, Kubernetes |
| **CI/CD Tools** | Jenkins, GitLab, Bamboo. |
| **Software Methodologies** | Agile, Scrum, Waterfall |
| **Development Tools** | Eclipse, PyCharm, IntelliJ, SSMS, Microsoft Office Suite (Word, Excel, PowerPoint) |
| **Version Control Tools** | GitHub and Azure DevOps |
| **J2SE/J2EE Technologies** | Java, J2EE, RMI, Sockets, JDBC, Servlets, JSP, JMS, Java Beans, JSTL, Jakarta Struts, JSF, EJB, Spring, Hibernate, JTA, JNDI, JPA, JMS. |

**PROFESSIONAL EXPERIENCE:**

**Client: Honeywell, Torrance, CA Apr 2023 - Till Date**

**Role: Senior Data Engineer**

**Responsibilities:**

* Implemented a 'serverless' architecture using API Gateway, Lambda, and Dynamo DB and deployed AWS Lambda code from Amazon S3 buckets. Created a Lambda function and configured it to receive events from your S3 bucket
* Designed the data models to be used in data-intensive AWS Lambda applications which are aimed at doing complex analysis, creating analytical reports for end-to-end traceability, lineage, definition of Key Business elements from Aurora.
* Involved in Migrating Objects from Teradata to Snowflake and created a Snow pipe for continuous data load.
* Created Notebooks in Azure Data Bricks and integrated it with ADF to automate the same.
* Writing code that optimizes performance of AWS services used by application teams and provides Code-level application security for clients (IAM roles, credentials, encryption, etc.)
* Creating AWS Lambda functions using Python for deployment management in AWS and designed and implemented public facing websites on Amazon Web Services and integrated it with other applications infrastructure.
* Created functions and assigned roles in AWS Lambda to run Python scripts, and AWS Lambda using Python to perform event-driven processing.
* Designed various Azure data factory pipelines to pull data from various data sources and load the data into Azure SQL database.
* Develop batch processing solutions by using Data Factory and Azure Data bricks
* Developed Spark code using Python and Spark-SQL for faster testing and processing of data.
* Worked on analyzing the AT&T Inventory, Expenses, and orders data, performed data cleansing/engineering, and provided the BRD to the business.
* Developed JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that processes the data using the SQL Activity.
* Created Databricks and scheduled a spark job to extract data from files in ADLS gen2.
* Experience in supporting data analysis projects using Elastic Map Reduce (EMR) on the AWS cloud.
* Developed batch scripts to fetch the data from ADL storage and do required transformations in PySpark using Spark framework.
* Designed and implemented highly performant data ingestion pipelines from multiple sources using Azure Data Factory and Azure Databricks
* Involved in converting Hive/SQL queries into spark transformations using Spark data frame in AWS Glue.
* Gather business requirements and design and develop data ingestion layer and presentation layer.
* Created data pipeline for different events of ingestion, aggregation, and load consumer response data in AWS S3 bucket into Hive external tables in HDFS location to serve as feed for AWS Quick sight Dashboard.
* Creating different AWS Lambda functions and API Gateways, to submit data via API Gateway that is accessible via Lambda function.
* Worked in a SAFE (Scaled Agile Framework) team with daily standups, sprint planning, quarterly planning.
* Developed JSON scripts for deploying the pipeline in Azure Data Factory that process the data using the Cosmos Activity
* Created complex ETL Azure Data Factory pipelines using mapping data flows with multiple Input/output transformations, Schema Modifier transformations, row modifier transformations using Scala Expressions
* Working on Azure Data bricks to run Spark-Python Notebooks through ADF pipelines.
* Responsible for Building Cloud Formation templates for SNS, SQS, Elastic search, Dynamo DB, Lambda, EC2, VPC, RDS, S3, IAM, Cloud Watch services implementation and integrated with Service Catalog.
* Implement Azure Data bricks clusters, notebooks, jobs and auto scaling.
* Good Understanding of other AWS services like S3, EC2 IAM, RDS Experience with Orchestration and Data Pipelines like AWS Step functions/Data Pipeline/Glue.
* Extensively worked on Azure Data Lake Analytics with the help of Azure Data bricks to implement SCD-1, SCD-2 approaches.
* Experience in writing SAM templates to deploy serverless applications on AWS cloud.
* Was responsible for creating on-demand tables on S3 files using Lambda Functions and AWS Glue using Python and PySpark.
* Coordinated with team and Developed framework to generate Daily adhoc, Reports and Extracts from enterprise data and automated using Oozie.
* Developed JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that processes the data using the SQL Activity.
* Worked closely with data engineers to define ETL processes, data transformations, and data integration workflows.

**Environment**: Teradata Data Warehouse, Data Engineering, Teradata, Snowflake, PostgreSQL, Server Integration Services (SSIS), Data Analysis, SQL Server Reporting Services (SSRS), Google Cloud Platform (GCP), Hadoop,

**Client: Lincoln Financial Group, Greensboro, NC Feb 2022 - Mar 2023**

**Role: AWS Data Engineer**

**Responsibilities:**

* Wrote scripts and indexing strategy for migration to Confidential Redshift from SQL Server and MySQL databases
* Used AWS glue catalog with crawler to get the data from S3 and perform SQL query operations
* Worked on AWS Data Pipeline to configure data loads from S3 to Redshift
* Used JSON schema to define table and column mapping from S3 data to Redshift
* Wrote indexing and data distribution strategies optimized for sub-second query response
* Developed a statistical model using artificial neural networks for ranking the students to better assist the admission process.
* Created Triggers, PowerShell scripts and the parameter JSON files for the deployments
* Implementation of end-to-end data solution on Azure using Azure Databricks, ADF, DW and PowerBI.
* Developed the PySpark code for AWS Glue jobs and EMR.
* Good Understanding of other AWS services like S3, EC2 IAM, RDS Experience with Orchestration and Data Pipelines like AWS Step functions/Data Pipeline/Glue.
* Migration of on-premises data (Oracle/ SQL Server/ DB2/ MongoDB) to Azure Data Lake Store (ADLS) using Azure Data Factory (ADF V1/V2)
* Experience in writing SAM templates to deploy serverless applications on AWS cloud.
* Created various pipelines to load the data from Azure data Bricks Lake into Staging SQLDB and followed by Azure SQL DB.
* Wrote Python scripts to process semi-structured data in formats like JSON.
* Was responsible for creating on-demand tables on S3 files using Lambda Functions and AWS Glue using Python and PySpark.
* Experienced in writing SparkApplications in Python.
* Designed and developed schema data models. Migrated an existing on-premises application to AWS. Used AWS services like EC2 and S3 for small data sets processing and storage, Experienced in Maintaining the Hadoop cluster on AWS EMR.
* Knowledge in retrieving, analyzing and presenting data using Azure Data Bricks Explorer/Kusto
* Imported data from AWS S3 into Spark RDD and performed transformations and actions on RDDs.
* Designed and deployed many ETL workflows via Azure Data Factory (ADF) and SSIS packages to extract, transform, and load data from SQL Server databases, excel and flat file sources into Data Warehouse.
* Written multiple MapReduce programs for data extraction, transformation and aggregation from multiple file-formats including XML, JSON, CSV & other compressed file formats.
* Developed automated processes for flattening the upstream data from Cassandra which in JSON format. Used Hive UDFs to flatten the JSON Data.
* Responsible for ingesting data from Data Lake to Data Warehouse using Azure services such as Azure Data Factory, and Azure Databricks.
* Performed Data cleaning and Preparation on XML files.
* Used JSON schema to define table and column mapping from S3 data to Redshift
* Used Azure Data Factory to orchestrate Databricks data preparation and load them into SQL Data warehouse.
* Robotic Process Automation of data cleaning and preparation in Python.
* Built analytical dashboards to track the student records and GPAs across the board.
* Used deep learning frameworks like MXNet, Caffe 2, Tensorflow, Theano, CNTK and Keras to help clients build Deep learning models
* Participated in requirements meetings and data mapping sessions to understand business needs.

**Environment:**Hadoop, AWS EMR, EC2, S3, Athena, Glue, DBT, Apache Spark, Airflow, Docker, PySpark, SparkSQL, Python(OOP), SQL, Kafka, HBase, HIVE, PIG, UNIX, Shell scripting, Tableau, Git, Jenkins, Jira.

**Client: Value Labs-India Sep 2018 - Jul 2021**

**Role: Data Engineer**

**Responsibilities:**

* Developing ETL pipelines in and out data warehouse using a combination of Python and Snowflake Snow SQL.
* Designed both 3NF data models for OLTP systems and dimensional data models using star and snowflake Schemas.
* Designed and implemented Kafka by configuring Topics in a new Kafka cluster in all environments.
* Working with SQL queries to QC the load on the DB with the ETL tool and managing appropriate changes to tables for query performance improvement.
* Monitored the SQL scripts using PySpark SQL
* Developed PySpark pipelines which transform the raw data from several formats to parquet files for consumption by downstream systems.
* Designing and building multi-terabyte, full end-to-end Data Warehouse infrastructure from the ground up on Confidential Redshift for large-scale data handling Millions of records every day
* Worked on Big data on AWS cloud services i.e. EC2, S3, EMR and DynamoDB
* Managed security groups on AWS, focusing on high availability, fault tolerance, and auto-scaling using Terraform templates. Along with Continuous Integration and Continuous Deployment with AWS Lambda and AWS code pipeline.
* Implementing and Managing ETL solutions and automating operational processes.
* Optimizing and tuning the Redshift environment, enabling queries to perform up to 100x faster for Tableau and SAS Visual Analytics
* Developed Python scripts to parse the Flat Files, CSV, XML, JSON files and extract the data from various sources and load the data into a data warehouse
* Wrote various data normalization jobs for new data ingested into Redshift
* Advanced knowledge of Confidential Redshift and MPP database concepts.
* Migrated on-premise database structure to Confidential Redshift data warehouse
* Was responsible for ETL and data validation using SQL Server Integration Services.
* Defined and deployed monitoring, metrics, and logging systems on AWS.
* Implemented Work Load Management (WML) in Redshift to prioritize basic dashboard queries over more complex longer-running ad hoc queries. This allowed for a more reliable and faster reporting interface, giving sub-second query responses for basic queries.
* Worked with the Version Control Systems like Team Foundation Server (TFS), and Visual Source Safe (VSS).
* Extensively worked on making REST API (application program interface) calls to get the data as JSON response and parse it
* Worked publishing interactive data visualizations dashboards, reports /workbooks on Tableau and SAS Visual Analytics.
* Expertise knowledge in Hive SQL, Presto SQL, and Spark SQL for ETL jobs and using the right technology for the job to get done.

**Environment:**Hadoop, Apache Spark, Docker, Kubernetes, PySpark, Spark SQL, Python, SQL, Git

**Client: Peri Software Solutions-India Jun 2015 - Aug 2018**

**Role: Data Engineer**

**Responsibilities:**

* Worked in an Agile environment and used a rallying tool to maintain the user stories and tasks.
* Worked with building data warehouse structures, and creating facts, dimensions, aggregate tables, by dimensional modelling, Star and Snowflake schemas.
* Created Data Quality Scripts using SQL and Hive to validate the successful das ta load and quality of the data.
* Involved in creating AWS Pipelines by extracting customer Big Data from various data sources into Hadoop HDFS and this included data from Excel, Flat Files, Oracle, SQL Server, Teradata, and log data from servers
* Developed a Python script to validate the files in one S3 with the same daily files in another bucket.
* Collaborated with cross-functional teams, including business analysts, data engineers, and data scientists, to define data requirements and architectural needs.
* Designed and implemented scalable data architectures, including data lakes and data warehouses, to support the organization's data-driven initiatives.
* Developed stored procedures in MS SQL to fetch the data from different servers using FTP and processed these files to update the tables.
* Analyze various types of raw files like JSON, CSV, XML with Python using Pandas, Numpy etc.
* Responsible for Designing Logical and Physical data modelling for various data sources on Confidential Redshift.
* Designed and Developed ETL jobs to extract data from Salesforce replica and load it in the data mart in Redshift.
* Experience with building data pipelines in Python/Pyspark/HiveSQL/Presto/BigQuery and building Python DAG in Apache Airflow.
* Used Microsoft Visual Source Safe (VSS) and Team Foundation Server (TFS) for integration, maintenance and Security of code.
* Created ETL Pipeline using Spark and Hive for ingesting data from multiple sources.
* Involved in using SAP and transactions done in SAP - SD Module for handling customers of the client and generating the sales reports.
* Coordinated with clients directly to get data from different databases.
* Worked on MS SQL Server, including SSRS, SSIS, and T-SQL.
* Designed and developed schema data models.
* Implemented data security measures, including encryption, access controls, and data masking, to protect sensitive data.
* Stayed current with industry trends and emerging technologies in data architecture, ensuring the organization's data infrastructure remained up to date.
* Acted as a technical mentor to junior team members, guiding data architecture best practices
* Documented business workflows for stakeholder review.

**Environment:** Hadoop, MapReduce, HDFS, Hive, Java, SQL DataStage, ETL (Informatica/SSIS)

**Client: Cyient Inc, India**  **Mar 2013 - May 2015**

**Role: Big Data Engineer**

**Responsibilities:**

* Involved in the implementation of new statistical algorithms and operators on Hadoop and SQL platforms and utilized optimization techniques, linear regressions, K-means clustering, Native Bayes and other approaches.
* Developed Spark batch job to automate creation/metadata update of external Hive table created on top of datasets residing in HDFS.
* Developed Data Serialization spark common module for converting Complex objects into sequence bits by using AVRO, PARQUET, JSON, and CSV formats.
* Worked on ERModeling, Dimensional Modelling (Star Schema, Snowflake Schema), Data warehousing, and OLAP tools.
* Populated HDFS and PostgreSQL with huge amounts of data using Apache Kafka.
* Design and develop Rest API (Commerce API) which provides functionality to connect to the PostgreSQL through Java services.
* Created iterative macro in Alteryx to send JSON request and download JSON response from web service and analyze the response data.
* Designed Batch Audit Process in batch\shell script to monitor each ETL job along with reporting status which includes table name, start and finish time, number of rows loaded, status, etc.
* Developed Spark jobs in Py Spark to perform ETL from SQL Server to Hadoop.
* Responsible for continuous monitoring and managing Elastic MapReduce (EMR) cluster through AWS console.
* Designed and implemented data acquisition, ingestion, Management of Hadoop infrastructure, and other Analytics tools (Splunk, Tableau).
* Working knowledge of build automation and CI/CD pipelines.
* Developed Python scripts to automate data ingestion pipeline for multiple data sources and deployed Apache Nifi in AWS.
* Design and develop Tableau visualizations which include preparing Dashboards using calculations, parameters, calculated fields, groups, sets, and hierarchies.

**Environment:** Hadoop, Map Reduce, Spark, Spark MLLib, Tableau, SQL, Excel, PIG, Hive, AWS, Postgres SQL, Python, PySpark, Flink, Kafka, SQL Server 2012, T-SQL, CI-CD, Git, XML, Tableau.