**Rohith Dhane**

**SR Data Engineer**

**Email :** **rohith.dhane69@gmail.com** **|| Contact .No: +1 (667) 577-0261**

**Professional Summary:**

* Around 12+ years of hands-on experience, I've refined my expertise in crafting and executing comprehensive data solutions and architectures, with a specialization in Data Engineering.
* Extensive experience in Data Warehousing and Decision Support Systems, with a robust track record in managing full lifecycle projects.
* Skilled in utilizing GCP services such as Compute Engine, Cloud Load Balancing, Cloud Storage, Cloud Dataproc, Cloud Functions, Cloud Pub/Sub, Cloud Shell, Cloud SQL, BigQuery, Cloud Data Fusion, Cloud Dataflow, Stack Driver Monitoring, And Cloud Deployment Manager.
* Hands-on experience in data analysis using HiveQL, custom MapReduce programs (Java/Python), and complex HiveQL queries for data extraction, along with developing Hive User Defined Functions (UDFs) as necessary. Proven track record of utilizing Snowflake database and Python for development and optimization tasks, ensuring high proficiency and efficiency in database-related coding and implementation.
* Worked actively in configuring and maintaining Oracle Physical Standby/Active Data Guard Database oracle 11g.
* Skilled in developing and deploying Python-based Lambda functions for real-time data processing, log parsing, alerting, and lightweight transformation tasks.
* Profound understanding and practical application of Hadoop/Big Data technologies, encompassing storage, querying, processing, and analysis.
* Engaged in software development on cloud computing platforms, including Amazon Web Services (AWS), Azure, and Google Cloud Platform (GCP).
* Deep knowledge of Hadoop's architecture and ecosystem, including key components like HDFS, MapReduce, and various ecosystem tools.
* Proficient in Python and Apache Beam for data validation and processing in Google Cloud Dataflow.
* Expertise in setting up cluster environment for load balancing and high availability of Windows servers.
* Performing Fault Injection Test Cases for verifying the High Availability at each layer of RAC stack like port failover, instance failover, ASM failover and Node failover.
* Implemented custom-built input adapters using Spark, Hive, and Sqoop to ingest data for analytics from various sources (Snowflake, MS SQL, MongoDB) into HDFS. Imported data from web servers and Teradata using Sqoop, Flume, and Spark Streaming API.
* Skilled in installing, configuring, and utilizing Hadoop ecosystem components like Hadoop MapReduce, HDFS, Hive, and Sqoop.
* Experience in managing Microsoft Windows server infrastructure and data - center operations by effectively planning, installing, configuring and optimize the IT infrastructure to achieve high availability and performance.
* Experience in working with Windows Server 2008/2008 R2 Active Directory and Administration.
* Expertise in data analysis using HiveQL, HBase, and custom MapReduce programs.
* Proficient in importing/exporting data using Sqoop between HDFS and Relational Database Systems like Teradata, Oracle, and SQL Server.
* Proficient in developing ETL workflows and data transformation logic using PySpark DataFrames, Spark SQL, and RDDs, optimized for performance and scalability.
* Experienced in designing and implementing migration strategies for traditional systems on Azure, utilizing services like Azure SQL Database and Azure Data Factory.
* Hands-on experience in GCP, particularly in BigQuery, Cloud Dataflow, and Data Proc.
* Developed complex data mappings and executed Proof of Concepts (POC) for transitioning MapReduce jobs into Spark transformations.
* Proficient in integrating Lambda functions with AWS services such as S3, Kinesis, DynamoDB, Glue, and SNS to build event-driven data pipelines.
* Managed Lambda deployments using Terraform and AWS SAM, enabling version-controlled infrastructure as code and CI/CD automation.
* Skilled in developing Apache Spark jobs using Python and Spark SQL for efficient data processing.
* Strong understanding of statistics and experience in developing machine learning models, including usage of Databricks.
* (EBS (ERP), Database and RAC), patches, SR support, application and database tuning, concurrent manager administration, cloning, upgrades.
* Skilled in integrating PySpark with cloud platforms such as AWS (S3, EMR, Glue) and Azure (Databricks, Data Lake, Synapse) to develop cost-effective data solutions.
* Proficient in Python, Scala, and core computer science concepts like data structures and algorithms.
* Configuration and Management of Data Guard on RHEL, OEL.
* Worked with various monitoring tools like AppDynamics, Datadog, ELK, Dynatrace etc. Deployed monitoring modules to new system components sing AppDynamics.
* Developed applications for data processing tasks using various databases including Teradata, Snowflake, and Postgres.
* Built ETL pipelines, visualizations, and analytics-based solutions using AWS, Azure Databricks, and open-source frameworks.
* Well-versed in Hadoop distributions like Cloudera, Hortonworks, and AWS EMR.
* Deep understanding of AWS services and cloud architectures across AWS, Azure, and GCP.
* Implemented schedulers using tools like Oozie, Airflow, and shell scripts.

**Technical Skills:**

|  |  |
| --- | --- |
| Big Data Technologies | HDFS, YARN, MapReduce, Hive, Pig, Impala, Sqoop, Storm, Flume, Spark, Pyspark, Apache Kafka, Zookeeper, Ambari, Oozie, MongoDB, Cassandra, Mahout, Puppet, Avro, Parquet, Snappy, Falcon. |
| NO SQL Databases | Postgres, HBase, Cassandra, MongoDB, Amazon DynamoDB, Redis |
| Hadoop Distributions  | Cloudera (CDH3, CDH4, and CDH5), Hortonworks, MapR, and Apache. |
| Lanuages | Scala, Python, R, XML, XHTML, HTML, AJAX, CSS, SQL, PL/SQL, HiveQL, Unix, Shell Scripting  |
| Source Code Control | GitHub, CVS, SVN, ClearCase |
| Cloud Computing Tools | Amazon AWS, (S3, EMR, EC2, Lambda, VPC, Route 53, Cloud Watch, CloudFront), Microsoft Azure, GCP |
| Databases | Teradata Snowflake, Microsoft SQL Server, MySQL, DB2 |
| DB languages | MySQL, PL/SQL, PostgreSQL & Oracle |
| Build Tools | Jenkins, Maven, Ant, Log4j |
| Business Intelligence Tools | Tableau, Power BI |
| Development Tools | Eclipse, IntelliJ, Microsoft SQL Studio, Toad, NetBeans |
| ETL Tools | Talend, Pentaho, Informatica, Ab Initio, SSIS |
| Development Methodologies | Agile, Scrum, Waterfall, V model, Spiral, UML |

**Professional Experience:**

**Client: CareFirst, MD Nov 2023 to Present**

**Role: SR Data Engineer**

**Responsibilities:**

* Involved in complete Big Data flow of the application starting from data ingestion upstream to HDFS, processing the data in HDFS and analyzing the data and involved.
* Good Experience on importing and exporting the data from HDFS and Hive into Relational Database Systems like MySQL and vice versa using Sqoop.
* Implemented sentiment analysis and text analytics on Twitter social media feeds and market news using Scala and Python.
* Designed and implemented distributed ETL pipelines using PySpark to process over of data daily from multiple sources
* Worked on Big data on AWS cloud services i.e. EC2, S3, EMR and DynamoDB
* Assisted in upgrading, configuration and maintenance of various Hadoop infrastructures like Pig, Hive, and Hbase.
* Experienced in leveraging AWS Lambda for serverless compute to automate data ingestion, transformation, and orchestration in cloud-native ETL workflows.
* Performed advanced procedures like text analytics and processing, using the in-memory computing capabilities of Spark using Scala.
* Defined and deployed monitoring, metrics, and logging systems on AWS.
* Developed Map Reduce Programs for data analysis and data cleaning
* Developed Oozie workflow engine to run multiple Hive, Pig, Tealeaf, Mongo DB, Git, Sqoop and Spark jobs.
* Experience in Job management using Fair scheduler and Developed job processing scripts using Oozie workflow to run multiple Spark Jobs in sequence for processing data
* Processed the Web server logs by developing Multi-hop flume agents by using Avro Sink and loaded into MongoDB for further analysis, also extracted files from MongoDB through Flume and processed.
* Developed complex data transformation logic using PySpark DataFrames and Spark SQL to support downstream analytics and reporting.
* 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.
* Installed Oozie workflow engine to run multiple Hive.
* Worked as a Hadoop Developer on Hadoop eco-systems including Hive, Zookeeper, Spark Streaming with MapR distribution.
* Supported Map Reduce Programs those are running on the cluster. Involved in loading data from UNIX file system to HDFS.
* Used Spark streaming to receive real time data from the Kafka and store the stream data to HDFS using Scala and NoSQL databases such as HBase and Cassandra.
* Installed and configured Hive, Pig, Sqoop, Flume and Oozie on the Hadoop cluster.
* Involved in installation, configuration, supporting and managing Hadoop clusters, Hadoop cluster administration.
* Good knowledge in Cluster coordination services through Zookeeper and Kafka.
* Involved in creating Hive tables, loading the data and writing hive queries that will run internally in a map reduce way.
* Used Spark Data frames, Spark-SQL, Spark MLLib extensively and developing and designing POC's using Scala, Spark SQL and MLlib libraries.
* Applied resource and timeout optimizations in Lambda to prevent cold starts and reduce execution costs in high-throughput environments.
* Documented the requirements including the available code which should be implemented using Spark, Hive, HDFS, HBase and Elastic Search.
* Used AWS Glue for the data transformation, validate and data cleansing.
* Developed customized UDF’s in Python to extend Hive and Pig Latin functionality.
* Worked with team of developers designed, developed and implement BI solutions for multiple projects
* Integrated PySpark with AWS services like S3, Glue, and Redshift to support scalable and cost-efficient data processing in the cloud.
* Expertise in Python and Scala, user-defined functions (UDF) for Hive and Pig using Python.
* Used Zookeeper to provide coordination services to the cluster.
* Created complex dashboard using parameters, sets, groups, and calculations to drill down and drill up in worksheets and customization using filters and actions.
* Created a Kafka producer to connect to different external sources and bring the data to a Kafka broker.
* Created and executed HQL scripts that create external tables in a raw layer database in Hive.

**Environment:** Hadoop, Hive, Spark, Hbase, MapReduce, PL/SQL, Kafka, Unix, Cucumber JVM, Mongo DB, GitHub, BitBucket, SQL, Oracle 12c, NoSQL database, API, Java, Jenkins, AWS.

**CBRE, Dallas, TX Aug 2021 to Oct 2023**

**Role: Sr. Data engineer**

**Responsibilities:**

* Analyzed business needs and wrote detailed specs while following project rules for writing programs.
* Oversaw research, planning, Proof of Concept (POC), and design for Big Data projects.
* Using Hadoop and Big Data ideas, loaded and changed large datasets that contained organized, semi-structured, and unstructured data.
* Set up and installed Apache Hadoop groups for building apps and using Hadoop tools.
* Setting up Hive, writing Hive User-Defined Functions (UDFs), and using a library of UDFs for Pig Latin.
* Built a data pipeline with Pig and Sqoop to bring ship data and user records into HDFS so they can be analyzed.
* Moved code that was already on-premises to an AWS EMR stack.
* Installed and set up Cloudera Manager and parts of the Hadoop ecosystem using the CDH package.
* Modeled the Dialog process and business processes and wrote code for business objects, Query Mapper, and JUnit files.
* Set up automatic workflows in AWS Code Pipeline to use S3 to launch Docker containers in AWS ECS.
* Used the HBase NoSQL Database to access huge amounts of data in real time and read and write to them.
* Used Spark streaming to get the real-time feed, then changed it to RDD, processed the data into a Data Frame, and put it into HBase.
* Built an AWS Lambda function that will run the Glue job as soon as a new file arrives in the Inbound S3 bucket.
* I made Spark jobs that will clean and check new source files in the incoming bucket and send rejected records to the reject-data S3 bucket.
* Created AWS CloudFormation templates and set up Auto Scaling for EC2 instances. Used Jenkins to help with the automatic setup of the AWS cloud environment.
* Set up HBase tables to load big amounts of semi-structured data from different sources.
* In charge of using the REST API to load customer data and event logs from Kafka into HBase.
* In AWS Redshift, I made tables with sort and distribution keys.
* Made shell scripts and Python scripts to handle daily jobs, even ones that need to be done for production.
* Managed Kafka Queues by adding, changing, and removing items as needed.
* Used Hadoop, Hive, and Informatica to pull the data into the HDFS system as part of the Big Data approach.
* Built apps that store and handle data using Angular6 and Java lambda expressions.
* To analyze data in Hive, we used the PySpark API over Hortonworks Hadoop YARN.
* Used Spark data frames, Scala, and Python to turn Hive/SQL searches into Spark transformations.
* Using HiveQL and Unix scripts, built and kept batch data flow.
* Designed and built an application that does real-time processing using Spark, Kafka, Scala, and Hive to do streaming ETL and machine learning.
* Implemented unit and integration tests for PySpark jobs using PyTest and handled deployment via CI/CD pipelines.
* Added different data formatter features and released to several Kafka Topics.
* Created AWS CloudFormation templates and set up Auto Scaling for EC2 instances. Used Jenkins to help with the automatic setup of the AWS cloud environment.
* Used HBase command-line tools to make automatic HBase test cases for checking the quality of data.
* Running load/endurance tests using Vugen, ALM and controller, server monitoring, analysis using Dynatrace, UNIX putty, SQL logs and other tools and reporting the performance. Analyzing errors and exceptions using putty logs (UNIX), etc.
* Helped create the Hadoop system and make the speed of multi-node Hadoop clusters better.
* Created and set up Apache NiFi in a few different settings, writing Python tools for quality assurance that kept track of files.

**Environments:** Hadoop 3.0, MapReduce, Hive 3.0, Sqoop, Agile, Apache Airflow, HBase 1.2, SQL, AWS, ETL, EC2, Kafka, Angular6, Java lambda, EMR, S3, UDFs, REST, POC, CDH, NOSQL, API, Pig 0.17, HDFS, Java 11, Hortonworks, HiveQL, PySpark, Scala, RDD, Hive, PL/SQL, Python, Jenkins.

**Merck Pharma, Branchburg, NJ Oct 2019 to July 2021**

**Role: Data Engineer**

**Responsibilities:**

* Engineered data pipelines in Apache Airflow on Google Cloud Platform (GCP) to efficiently handle ETL tasks using a diverse set of Airflow operators.
* Explored Spark to optimize performance and refine existing algorithms on Hadoop, leveraging Spark's robust features including Spark Context, Spark SQL, DataFrames, and Spark YARN.
* Utilized Spark Streaming to seamlessly ingest data into an in-house ingestion platform.
* Developed RESTful APIs using Python, employing Flask and Django frameworks, and seamlessly integrated various data sources such as Java, JDBC, RDBMS, Shell Scripting, Spreadsheets, and Text files.
* Provided technical support and troubleshooting for Python-based implementations in Athena and Teradata, resolving issues promptly to maintain continuous operation of real-time data analytics pipelines.
* Designing and implementing data processing systems on GCP using services such as Big Query, Dataflow, and Datapost.
* Build data pipelines in airflow in GCP for ETL-related jobs using different airflow operators. Experience in GCP Data proc, GCS, Cloud functions, Big Query.
* Building and managing data warehouses and data lakes on GCP, ensuring data integrity and security.
* Monitored performance alerts in production using IDERA monitoring tool.
* Integrated Python scripts with other data sources and analytics tools to enable comprehensive real-time data analysis and reporting capabilities across Athena and Teradata platforms.
* Integrating external data sources and APIs into GCP data solutions, ensuring data quality and consistency.
* Building data transformation pipelines using GCP services like Dataflow or Apache Beam to cleanse, normalize, and enrich data.
* Assisted with migration and upgrade of Jira instances, ensuring a seamless transition with minimal disruption to projects. Facilitated communication between development, operations, and business teams through Jira's collaboration features.
* Crafted PySpark scripts to automate the processing and transfer of files to third-party vendors. Established robust data pipelines in GCP using Apache Airflow, leveraging diverse Airflow operators to streamline ETL-related tasks.
* Demonstrated proficiency in GCP services such as Dataproc, Google Cloud Storage (GCS), Cloud Functions, and BigQuery.
* Managed continuous data transfers using Snowpipe and crafted SnowSQL queries for thorough data analysis.
* Operated within a Scrum/Agile environment, utilizing effective project management tools such as JIRA.
* Implemented Spark using Scala and SparkSQL to expedite data processing and streamline testing procedures.
* Translated complex Hive/SQL queries into Spark transformations using Spark RDDs in tandem with Scala.
* Developed Python scripts to efficiently extract data from datasets and store them in HDFS using PySpark.
* Designed Python solutions to retrieve data from HBase and effectively implemented them using PySpark.

**Environment:** GCP, GCP Dataproc, Apache Beam, Airflow, Hadoop, Teradata, Teradata Spark, EMR, S3, Hive, Python, Airflow, Teradata, IDERA, Hive, Spark SQL, SQL.

**Dollar Tree, Chesapeake, VA Dec 2016 to Sep 2019**

**Role: Data Engineer**

**Responsibilities:**

* Strategically designed and configured relational servers and databases on the Azure Cloud, meticulously assessing both current and future business requirements.
* Played a pivotal role in the seamless migration of data from on-premises SQL servers to Cloud databases, specifically Azure Synapse Analytics (DW) and Azure SQL DB.
* Working in base lining, capacity planning, Data Center Infrastructure & Network Designing, Windows Server Migrations.
* Demonstrated extensive proficiency in creating pipeline jobs, scheduling triggers, and mapping data flows using Azure Data Factory (V2), ensuring secure storage of credentials in Key Vaults.
* Skillfully created elastic pool databases and scheduled elastic jobs to execute T-SQL procedures, optimizing resource allocation.
* Developed ETL jobs to efficiently load, serve, and transport data into buckets, facilitating the transfer of S3 data to the Data Warehouse.
* Leveraged Kusto Explorer for log analytics and enhanced query response times, crafting alerts using Kusto query language.
* Successfully designed tabular models within Azure Analysis Services to fulfill specific business reporting requirements.
* Efficiently worked with Azure BLOB and Data Lake storage, seamlessly loading data into Azure SQL Synapse Analytics (DW).
* Addressed complex business queries involving multiple tables from different databases by crafting both correlated and non-correlated sub-queries.
* Designed and implemented business intelligence solutions using SQL Server Data Tools 2015 and 2017 versions, effectively loading data into both SQL and Azure Cloud databases.
* Conducted comprehensive analysis of data quality and enforced business rules at all stages of the data extraction, transformation, and loading process.
* Managed the testing and migration of ETL Informatica workflows and mappings, ensuring smooth transitions.
* Monitored performance alerts in production using IDERA monitoring tool.
* Oversaw validation and verification of software across various testing phases including Functional Testing, System Integration Testing, End to End Testing, Regression Testing, and more.
* Developed batch processing pipelines for data processing using Python and Airflow, while also scheduling Spark jobs using Airflow for streamlined operations.
* Actively contributed to writing, testing, and running MapReduce pipelines using Apache Crunch for efficient data processing.
* Effectively managed Hadoop log files, analyzed SQL scripts, and designed solutions for processes using Spark.
* Crafted insightful reports in TABLEAU for data visualization and thoroughly tested native Drill, Impala, and Spark connectors for data exploration.
* Developed diverse Python scripts for vulnerability assessment, including SQL injection, permission checks, and performance analysis.
* Proficiently orchestrated the import of data from various sources into HDFS using Sqoop, executed transformations using Hive and MapReduce, and subsequently loaded data into HDFS for further processing.

**Environment:** Microsoft SQL Server 2012, 2016, SSDT-2012 &2015, Azure Synapse Analytics, Azure Data Lake & BLOB, IDERA, Azure SQL, Azure data factory, Azure analysis services, BIDS.

**Concentrix Inc, Hyderabad, India Nov 2013 to Oct 2016**

**Role: Data Analyst**

**Responsibilities:**

* Applied a range of data transformations, encompassing Lookup, Aggregate, Sort, Multicasting, Conditional Split, Derived Column, and more.
* Developed Mappings, Sessions, and Workflows to efficiently extract, validate, and transform data in compliance with business rules using Informatica.
* Tailored target tables based on reporting team requirements and formulated Extraction, Transformation, and Loading (ETL) processes utilizing Talend.
* Utilized Netezza SQL scripts to seamlessly transfer data between Netezza tables.
* Scheduled Talend Jobs using Job Conductor, a scheduling tool within the Talend ecosystem.
* Took charge of querying, stored procedure creation, crafting complex queries, and leveraging T-SQL joins to address varied reporting operations and handle ad-hoc data requests.
* Prioritized performance monitoring and optimized indexes using tools like Performance Monitor, SQL Profiler, Database Tuning Advisor, and Index Tuning Wizard.
* Acted as the primary contact for resolving locking, blocking, and performance-related issues. Authored scripts and devised indexing strategies for migrating data to Amazon Redshift from SQL Server and MySQL databases.
* Managed Lambda deployments using Terraform and AWS SAM, enabling version-controlled infrastructure as code and CI/CD automation.
* Utilized AWS Data Pipeline to configure seamless data loads from S3 into Redshift.
* Employed JSON schema to define table and column mappings from S3 data to Redshift, devising indexing and data distribution strategies optimized for sub-second query response.
* Possessed hands-on experience with Dell Boomi Connectors, covering FTP, Mail, Database, Salesforce, Web Services Listener, HTTP Client, Web Services SOAP Client, Success Factors, and Trading Partner.
* Developed Database/Flat-file/JSON/XML profiles, Boomi Mappings, and Processes utilizing various connectors and logic shapes between application profiles and different Trading partners within Dell Boomi.

**Environment:** Amazon Redshift, AWS Data Pipeline, Talend Platform for Big Data MS SQL Server 2008R2/2012, Oracle 10g/9i, Dell Boomi, Netezza Mako 7.2, S3, SQL Server Reporting Services (SSRS), SQL Server Integration Services (SSIS), Share Point, TFS, MS Project, MS Access and Informatica.