**Prathik R**

**Email Id:** rprathik0045@gmail.com

**Contact: +1(513) 916 4544**

**Data Engineer**

**SUMMARY**

* Over **9** years of experience in Data Engineering, Data Pipeline Design, Development and Implementation as a Sr. Data Engineer across various industries including Hospitality& Entertainment, Healthcare, Retail, Financial etc.
* Strong experience in Big Data technologies and Hadoop ecosystem components like HDFS, Hive, Impala, Apache Pig, Sqoop, Oozie, Zookeeper, YARN, Apache Spark, Python, Spark, AWS (Amazon Redshift, Amazon Athena, Lambda, and Kinesis.
* Proficient in applying the SLDC software development process to establish a business analysis methodology. Strong hands-on experience implementing dashboarding, data visualizations, and analytics using Tableau.
* Expertise in working with Hive data warehouse infrastructure-creating tables, data distribution by implementing Partitioning and Bucketing, developing and tuning the HQL queries.
* I have expertise in data transformations through Hive and MapReduce, followed by data loading into HDFS. I also manage Sqoop jobs for incremental loading into Hive external tables and import streaming data into HDFS with Flume sources and sinks, utilizing Flume interceptors for data transformation.
* Experience with Amazon EC2, Amazon S3, Amazon RDS, VPC, IAM, Amazon Elastic Load Balancing, Auto Scaling, CloudWatch, SNS, SES, SQS, Lambda, EMR and other services of the AWS family.
* Experience in building pipelines using Azure Data Factory and moving the data into Azure Data Lake Store.
* Involved in developing Spark scripts using Python on Azure HDInsight for Data Aggregation, Validation and verified its performance over MR jobs.
* Experience in implementing OLAP multi-dimensional cube functionality using Azure SQL Data Warehouse.
* Experience of designing, building, and deploying data pipelines using tools using BigQuery, Cloud Dataflow, Cloud Proc, Cloud Pub/Sub, Cloud Composer, Google Data Studio, Google Cloud Storage
* Experienced in handling various types of data (Structured, Semi-Structured, Unstructured) coming from various sources with both batch and real time data pipelines.
* Expertise in Creating, Debugging, Scheduling, and Monitoring jobs using Airflow and 0ozie.
* Experienced with using most common Operators in Airflow - Python Operator, Bash Operator Google Cloud Storage Download Operator, Google Cloud Storage Object Sensor, GoogleCloudStorageToS3Operator.
* Expert in designing ETL data flows using creating mappings/workflows to extract data from SQL Server and Data Migration and Transformation from Oracle/Access/Excel Sheets using SQL Server 5515.
* Expert in designing Parallel jobs using various stages like Join, Merge, Lookup, remove duplicates,
* Filter, Dataset, Lookup file set, Complex flat file, Modify, Aggregator, XML.
* Experience working on various file formats including delimited text files, clickstream log files, Apache log files, Parquet files, Avro files, JSON files, XML files and others.
* Involved in Development and Created PL/SQL stored procedures, and functions.
* Worked withSnowflake features like clustering, time travel, cloning, logical data warehouse, caching etc.
* Experienced in python data manipulation for loading and extraction as well as with python libraries such as NumPy, SciPy and Pandas for data analysis and numerical computations
* Developed mapping spreadsheets for (ETL) team with source-to-target data mapping with physical naming standards, data types, volumetric, domain definitions, and corporate meta-data definitions.
* Designed ETL process using Informatica Designer to load the data from various source databases to target data warehouse in Vertica.
* Good knowledge and experience on NoSQL databases like HBase, Cassandra, and MongoDB and SQL databases like Teradata, Oracle, PostgreSQL and SQL Server.
* Extensive knowledge of utilizing all types of SQL Server Constraints, SQL Server Database design and maintenance principles, developing Transact- SQL queries, stored procedures, and triggers.
* Cassandra, Jenkins Pipelines, Kubernetes, Kibana, Rancher, GITHUB, Rancher, Kibana, Hadoop HDFS, Hive, IntelliJ, Cassandra, SQL server etc.
* Experience working in Data warehouse Design, methodologies Extraction, Transformation and Loading, Data Analysis, Data migration, Data preparation, Data Validation, Data Quality, Data Profiling Reporting and Documentation.
* Proficient knowledge of Data Analytics, Predictive Modeling, Natural Language Processing (NLP), and Deep Learning algorithms.
* A Data Science enthusiast with strong Problem solving, Debugging and Analytical capabilities, who actively engages in understanding and delivering business requirements.

**TECHNICAL sKILLS**

|  |  |
| --- | --- |
| **Operating Systems** | Unix, Linux, Mac OS, Windows |
| **Database programming** | T-SQL, Dynamic SQL, MDX, DAX |
| **Cloud Computing**  **Tools** | AWS, AZURE, Snowflake |
| **NoSQL Databases** | HBase, Cassandra, MongoDB, Apache, Hadoop HBase |
| **Programming Languages** | Python (Jupiter, PyCharm IDE), R, Java, Scala, SQL, PL/SQL, SAS |
| **Development Tools** | BIDS, SSDT, SSMS |
| **Databases** | Snowflake Cloud DB, Oracle, MySQL, SQL Server 2000/2005/2008/2012 |
| **SQL Server Tools** | SQL Server Management studio, Enterprise Manager, Query Analyzer, Profiler, Export, and Import (DTS) |
| **IDE** | IntelliJ, Eclipse, Visual Studio, IDLE |
| **Methodologies** | Agile, Scrum, Waterfall and Spiral |
| **Reporting / BI Tools** | SSRS, Tableau Desktop, Power BI, Excel Power BI |
| **Data Engineer/Big Data Tools/ Cloud** | Databricks, HDFS, Hive, Pig, Sqoop, MapReduce, Spring Boot, Flume Hortonworks, Cloudera, Mahout, Oozie, Zookeeper, etc. AWS, Azure Databricks, Azure Data Explorer, Azure HDInsight, Salesforce, NI-FI, Google Shell, Linux, Big Query, Tableau, Power BI, SAS, Crystal Reports. |

**EDUCATION DETAILS**

* Bachelors in Computer Science Engineering, JNTU, Hyderabad, India, 2015

**PROFESSIONAL EXPERIENCE**

**Crowd Strike, Austin, TX April 2023 - Current**

**Role: Sr. Data Engineer**

**Responsibilities:**

* Worked along with project managers to gather requirements for the data synchronization task, such as data scope, and data mapping.
* Validate the data flow going to the Snowflake DW cloud platform from the source systems.
* Develop Nifi workflow to pick up the data from rest API server, from data lake as well as from SFTP server and send that to Kafka broker.
* Implement Spark Kafka streaming to pick up the data from Kafka and send to Spark pipeline.
* Experience in working with different join patterns and implemented both Map and Reduce Side Joins.
* Wrote Flume configuration files for importing streaming log data into HBase with Flume.
* Imported several transactional logs from web servers with Flume to ingest the data into HDFS.
* Consulting on Snowflake Data Platform Solution Architecture, Design, Development and deployment focused to bring the data driven culture across enterprises.
* Involved in data ingestion into HDFS using Sqoop for full load and Flume for incremental load on variety of sources like web server, RDBMS, and Data API’s.
* Built the infrastructure required for optimal extraction, transformation, and loading of data from a wide variety of data sources using SQL and ‘big data’ technologies like Hadoop Hive, Azure Data Lake storage.
* Worked extensively on Azure data factory including data transformations, Integration Runtimes, Azure Key Vaults, Triggers and migrating data factory pipelines to higher environments using ARM Templates.
* Built the pipelines in Azure Data Factory that will transfer data from one location to another (ADF V1)
* 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.
* Created batch tasks in Amanda, and wrote stored procedures in SQL, database functions, and database triggers.
* Worked on the project to migrate data from on-premises databases to confidential Redshift, RDS, and S3.
* Written SQL queries against Snowflake, and created DWH, databases, schemas, and tables.
* Worked on creating tabular models on Azure analysis services for meeting business reporting requirements.
* Used error handing techniques like try catch and error functions to create robust T-SQL code.
* Utilized dynamic SQL for user-customizable queries to be answered by the OLTP Server.
* Identify and write the best possible code in case of new deployments or when rewriting code when migrating to a newer version of SQL Server.
* Implemented the machine learning algorithms using python to predict the quantity a user might want to order for a specific item so we can automatically suggest using kinesis firehose and 53 data lake.
* Engaged in requirements gathering, Agile PLM System design, and development.
* Worked with database objects like stored procedures, user-defined functions, triggers, and indexes using T-SQL to create complex scripts and batches.
* Used Azure Synapse to manage processing workloads and served data for BI and prediction needs.
* Used Spark SQL for Scala & amp, Python interface that automatically converts RDD case classes to schema RDD.
* Import the data from different sources like HDFS/HBase into Spark RDD and perform computations using PySpark to generate the output response.
* Evaluated Snowflake Design considerations for any change in the application.
* Developed a reusable framework to be leveraged for future migrations that automate ETL from RDBMS systems to the Data Lake utilizing Spark Data Sources and Hive data objects.
* Worked with Azure Stream Analytics and Azure Databricks for near real-time analytics and visualized it with
* Implemented dynamic SQL to develop customizable queries, answerable by the OLTP Server.
* Designed complex T-SQL queries and user-defined functions in SQL server.

**Environment: SQL Server, T-SQL, SQL, ETL, SSIS, AZURE, SAS, SQL Profiler, Erwin, SSMS, SSDT, Excel, DAX, Power BI, Power BI Desktop, TFS, S3, RDS.**

**Pitney Bowes Inc, CA Dec 2021 – Mar 2023**

**Role: Data Engineer**

**Responsibilities:**

* Responsible for building Confidential data cube using SPARK framework by writing Spark SQL queries in Python so as to improve efficiency of data processing and reporting query response time.
* Developed spark programming code in Python Data bricks workbooks.
* Performance tuning of SQOOP, Hive and Spark jobs.
* Implemented proof of concept to analyze the streaming data using Apache Spark with Python; Used Maven/SBT for build and deploy the Spark programs.
* Responsible for modification of ETL data load scripts, scheduling automated jobs and resolving production issues (if any) on time.
* Created Application Interface Document for the downstream to create new interface to transfer and receive the files through Azure Data Share.
* Creating pipelines, data flows and complex data transformations and manipulations using ADF and PySpark with Databricks.
* Developed PySpark scripts using Python on Azure HDInsight for Data Aggregation, Validation, verifying its performance over MR jobs and to extract the data from the web server output files to load into HDFS.
* Developed Scala Library on Databricks by setting up new Jobs and configuring Automated Cluster to Install Library by adding dependency.
* Ingested data in mini-batches and performs RDD transformations on those mini-batches of data by using Spark Streaming to perform streaming analytics in Data bricks.
* Created, provisioned different Databricks clusters needed for batch and continuous streaming data processing and installed the required libraries for the clusters.
* Improved performance by optimizing computing time to process the streaming data and saved cost to company by optimizing the cluster run time.
* Implemented Spark using Scala and Spark SQL for faster testing and processing of data.
* Perform ongoing monitoring, automation and refinement of data engineering solutions
* Prepare complex SQL views, stored procs in azure SQL DW and Hyperscale
* Loaded different files from ADLS by using U-SQL scripts into target Azure Data warehouse.
* Worked on complex U-SQL for the data transformation and loading table and report generation.
* Designed and developed a new solution to process the NRT data by using Azure stream analytics, Azure Event Hub and Service Bus Queue.
* Involved in developing of Confidential Data Lake and in building Confidential Data Cube on Microsoft Azure HDINSIGHT cluster.
* Create / update various database/scripts to support teams for effective daily activities in Apache Data Bricks.
* Worked on HIVE Query Optimization by implementing SPARK concepts.
* Created Linked service to land the data from Caesars SFTP location to Azure Datalake.
* Created numerous pipelines in Azure using Azure Data Factory v2 to get the data from different source databases Informix, Sybase etc by using different Azure Activities like Move &Transform, Copy, filter, for each, Databricks etc.
* Utilized Spark Core, Spark Streaming and Spark SQL API for faster processing of data instead of using MapReduce in Java.
* Using Git to update existing versions of the Hadoop Py-Spark script to its new model.
* Used Spark-SQL to Load JSON data and create Schema RDD and loaded it into Hive Tables and handled structured data using SparkSQL.

**Environment: Spark, Python, ETL HDFS, Hive, Spark, Spark-Streaming, Spark SQL, Sqoop, Azure, Azure Data Bricks, PySpark, Git**

**AIS Healthcare, Ridgeland MS Jan 2021 – Nov 2021**

**Role: Data Engineer**

**Responsibilities:**

* Worked along with project managers to gather system specification requirements for the project via HLD sessions.
* Used Python scripts to update content in the database and manipulate files.
* Gathered the various reporting requirements from the business analysts.
* Used Jira software from Atlassian to undermine the delivery of Agile teams.
* Designed a new data mart from the existing data warehouse utilizing a top-down approach.
* Modified SQL database model per changes in the business requirements.
* Used data analysis to validate business rules and identify low-quality missing data in the data warehouse.
* Performed data analysis to find anomalies and ensure data integrity before consolidating data.
* Fine-tuned SQL queries using show plans and execution plans for better performance.
* Performed quality testing of converted data, identifying the root cause of issues and designing/documenting proposed solutions.
* Responsible for designing logical and physical data modeling for various data sources on confidential Redshift.
* Migrated on-premises database structure to confidential Redshift data warehouse.
* Worked with DevOps methodologies and cloud infrastructure automation tools like AWS by using Terraform.
* Used Bash and Python including Boto3 to supplement automation provided by Ansible and Terraform for tasks such as encrypting EBS volumes backing AMIs and scheduling Lambda functions for routine AWS tasks.
* Implemented and maintained the monitoring and alerting of production and corporate servers/storage using AWS Cloud watch.
* Worked on AWS pipeline to configure data loads from S3 to Redshift.
* Architect AWS solutions for cloud migration of custom and enterprise applications.
* Develop Architecture Diagrams of a server migration and articulate the value of proposed AWS Solutions.
* Utilized dynamic SQL, procedures, views, and database triggers.
* Used data analysis techniques to validate business rules and identify low-quality missing data in the existing Enterprise Data Warehouse.
* Used JSON schema to define table and column mapping from S3 data to Redshift.
* Performed data analysis on test execution & defect data across projects to identify gaps and potential areas of improvement – exploratory data analysis, factor analysis, sentiment analysis, and text mining.
* Performed root cause analysis on failures and resolved performance tuning-related issues on queries with SQL profiler and DTA.
* Used SAS/ACCESS to read datasets from databases, Flat files, and read and write other PC file formats.
* Generated highly customized reports in the form of listing HTML, RTF, and PDF using SAS ODS, PROC SQL, PROC REPORT, PROC TABULATE.
* Developed code with respect to security vulnerabilities such as SQL injections.
* Implemented SQL tuning advisor for monitoring queries to work efficiently on high-volume scenarios.
* Monitored and resolved deadlocks in SQL Server databases with locks, isolation levels, and SQL Profiler.
* Transformed stored procedures with SQL profiler to refine report latency and increase delivery access.
* Incorporated error redirection during ETL loads in SSIS Packages.
* Created various transformation procedures by using SAS 9.1, SAS Data Integration Studio, and SAS Enterprise Guide.
* Scripted T-SQL queries, ranking functions, derived tables, common table expressions (CTEs), stored procedures, views, user-defined functions (UDF), constraints & database triggers.

**Environment:** **SQL Server 2016, SSDT, T-SQL, SSIS, ETL, SQL Profiler, Erwin, SSMS, SSDT, Power BI, TFS, AWS, SAS/ACCESS, S3, Redshift.**

**Amigos Software Solutions, Hyderabad, India** **Aug 2018 - Dec 2020**

**Role: Hadoop Consultant**

**Responsibilities:**

* Written MapReduce code for processing and parsing the data from various sources and storing parsed data into HBase and Hive using HBase-Hive Integration.
* Worked with HBase and Hive scripts to extract, transform and load data into HBase and Hive.
* Worked on moving all log files generated from various sources to HDFS for further processing.
* Developed workflows using custom MapReduce, Pig, Hive, and Sqoop.
* Developed workflow in Oozie to automate the tasks of loading the data into HDFS and pre-processing with PIG.
* Tuned the cluster for optimal performance to process these large data sets.
* Worked hands on with ETL process. Handled importing data from various data sources, performed transformations.
* Built reusable Hive UDF libraries for business requirements which enabled users to use these UDFs in Hive Querying.
* Loaded the created HFiles into HBase for faster access of large customer base without taking performance hit.
* Written Hive UDF to sort Structure fields and return complex data type.
* Responsible for loading data from UNIX file system to HDFS.
* Involved in writing HQL queries and SQL queries.
* Developed suit of Unit Test Cases for Mapper, Reducer and Driver classes using MR Testing library.
* Design job using Talend platform to extract, transform, and load data into Data Warehouse.
* Used Maven extensively for building jar files of MapReduce programs and deployed to Cluster.
* Modelled Hive partitions extensively for data separation and faster data processing and followed.
* Worked on the Ad hoc queries, Indexing, Replication, Load balancing, Aggregation in MongoDB.
* Work with network and Linux system engineers to define optimum network configurations, server hardware and operating system.
* Evaluate and propose new tools and technologies to meet the needs of the organization.
* Production support responsibilities include cluster maintenance.
* Pig and Hive best practices for tuning.

**Environment**: CDH, Hive, MySQL, HBase, HDFS, HIVE, Eclipse, Hadoop, Oracle, PL/SQL, SQL\*PLUS, Toad 9.6, Flume, PIG, Sqoop, UNIX.

**Hudda Infotech Private Limited Hyderabad, India** **Apr 2016 - Jul 2018**

**Role: Hadoop Developer**

**Responsibilities:**

* Written MapReduce code for processing and parsing the data from various sources and storing parsed data into HBase and Hive using HBase-Hive Integration.
* Worked with HBase and Hive scripts to extract, transform and load data into HBase and Hive.
* Installed and configured Hadoop MapReduce, HDFS and developed multiple MapReduce jobs in Java for data cleansing and preprocessing.
* Involved in loading data from UNIX file system to HDFS.
* Installed and configured Hive and written Hive UDFs.
* Importing and exporting data into HDFS and Hive using Sqoop
* Used Cassandra CQL and Java APIs to retrieve data from Cassandra Table.
* Responsible for cluster maintenance, adding and removing cluster nodes, cluster monitoring and troubleshooting, manage and review data backups, manage and review Hadoop log files.
* Worked hands on with ETL process.
* Handled importing of data from various data sources, performed transformations using Hive, MapReduce, and loaded data into HDFS.
* Worked with development teams to deploy Oozie workflow jobs to run multiple Hive and Pig jobs which run independently with time and data availability.
* Worked on trouble shooting the Hadoop cluster issues and fixing the cluster issues.
* Extracted the data from Teradata into HDFS using Sqoop.
* Analyzed the data by performing Hive queries and running Pig scripts to know user behavior.
* Exported the patterns analyzed back into Teradata using Sqoop.
* Continuous monitoring and managing the Hadoop cluster through Cloudera Manager.
* Installed Oozie workflow engine to run multiple Hive.
* Involved in creating Hive tables, loading with data, and writing hive queries which will run internally in map reduce way.
* Developed Hive queries to process the data and generate the data cubes for visualizing.

**Environment: Hadoop, MapReduce, HDFS, UNIX, Hive, Sqoop, Cassandra, ETL, Pig Script, Cloudera**.

**Analytix Data Services, India** **Aug 2015 - Mar 2016**

**Role: ETL Developer**

**Responsibilities:**

* Written MapReduce code for processing and parsing the data from various sources and storing parsed data into HBase and Hive using HBase-Hive Integration.
* Worked with HBase and Hive scripts to extract, transform and load data into HBase and Hive.
* Responsible for developing, support and maintenance for the ETL (Extract, Transform and Load) processes using Informatica PowerCenter.
* Designed and developed several ETL scripts using Informatica, UNIX shell scripts.
* Analyzed the source data coming from Oracle, Flat Files, and MS Excel coordinated with data warehouse team in developing Dimensional Model.
* Created FTP, ODBC, Relational connections for the sources and targets.
* Performed Backup/Restoration of repositories.
* Implemented Slowly Changing Dimension Type 2 methodology for accessing the full history of accounts and transaction information.
* Scheduled Sessions and Batches on the Informatica Server using Informatica Server Manager.
* Experience with writing and executing test cases for data transformations in Informatica.
* Created JIL scripts and scheduled workflows using CA Autosys.
* Developed Shell Scripts to automate file manipulation and data loading procedures.
* Developed UNIX shell scripts to control the process flow for Informatica workflows to handle high volume data.
* Prepared Test cases based on Functional Requirements Document.
* Involved in Unit, System, Performance, Concurrent and QA testing.

**Environment**: **Informatica Power Center 9.x, Oracle 11g, SQL plus, PL/SQL, Oracle, SQL Developer, UNIX.**