**Pradeep**

**Senior Data Engineer**

**Email:** [**pradeepby3@gmail.com**](mailto:pradeepby3@gmail.com) **| Phone Number: 704-559-9227 |LinkedIn:** [**Pradeep**](https://www.linkedin.com/in/pradeep-reddy-baireddy-45213j1992)

**PROFESSIONAL SUMMARY:**

* Over 10 years of proven excellence in the IT sector, specializing in Big Data and cutting-edge Cloud technologies.
* Extensive hands-on experience with Azure Cloud, including components like Azure Data Factory, Azure Databricks, Logic Apps, Azure Function Apps, Snowflake, and Azure DevOps services.
* Proficient in data movement within the Azure stack, showcasing expertise in moving data from Data Lake to Azure Blob Storage.
* Strong background in Data Load/Integration using Azure Data Factory (ADF) and building ETL data pipelines leveraging PySpark and Spark SQL within Azure Databricks.
* Skilled in developing data pipelines in Spark using Scala and PySpark, with additional expertise in Azure Logic App Integration tool.
* Successfully implemented Azure functions, Azure storage, and Service Bus queries for large enterprise-level ERP integration systems.
* In-depth experience in writing in-house UNIX shell scripts for Hadoop Big Data Development.
* Proficient in creating and managing Azure DevOps tools for continuous integration and deployment (CI/CD) pipelines.
* Expertise in Data Pipeline Development, Data Modeling, and the development of data ingestion workflows for efficient storage and retrieval.
* Comprehensive working knowledge of Hadoop, Java, HDFS, Map-Reduce, Hive, Tez, Python, and PySpark.
* Hands-on experience in developing large-scale data pipelines using Spark and Hive, optimizing query performance in Hive using bucketing and partitioning techniques.
* Well-versed in Apache Sqoop for importing and exporting data to and from HDFS and Hive.
* Experienced in setting up workflows using Apache Oozie for managing and scheduling Hadoop jobs.
* Proficient in optimizing data ingestion, data modeling, data encryption, and performance tuning of ETL workflows.
* Extensive hands-on experience tuning Spark Jobs and developing Spark scripts using Scala shell commands.
* Proven experience in real-time streaming with Kafka as a data pipeline using the Spark Streaming module.
* Expertise with Apache Kafka and Azure Event Hubs for messaging and streaming applications using Scala.
* Well-versed in using ETL methodology, particularly with Informatica, to support corporate-wide solutions.
* Worked on data serialization formats, converting complex objects into sequence bits using Parquet, ORC, AVRO, JSON, and CSV formats.
* Optimized Spark jobs and workflows by fine-tuning Spark configurations, partitioning, and memory allocation settings.
* Extensive experience in developing, maintaining, and implementing Enterprise Data Warehouses (EDW), Data Marts, Operational Data Stores (ODS), and Data Warehouses with Star and Snowflake schemas.
* Hands-on experience with GitHub for version control, ensuring efficient code maintenance.
* Comprehensive knowledge of the Software Development Life Cycle (SDLC) and experience working within Agile Methodology.

**TECHNICAL SKILLS**

| Big Data Technologies | MapReduce, Hive, Teg, Python, PySpark, Scala, Kafka, Spark streaming, Oozie, Sqoop, Zookeeper. |
| --- | --- |
| Hadoop Distribution | Cloudera, Horton Works. |
| Azure Services | Azure data Factory, Azure Data Bricks, Logic Apps, Functional App, Snowflake, Azure DevOps |
| AWS Services: | S3, Redshift, EMR, SNS, SQS, Athena, Glue, CloudWatch, Kinesis, Route 53, IAM, EC2, Lambda, Quick Sight, Macie, MSK. |
| Languages | Java, SQL, PL/SQL, PySpark, Python, HiveQL, Scala, Shell Scripting, Regular Expressions. |
| Web Technologies | HTML, CSS, JavaScript, XML, JSP, Restful, SOAP. |
| Operating Systems | Windows, UNIX, LINUX, CENTOS. |
| Version Control | GIT, GitHub, Bit Bucket, AWS Code Commit. |
| Reporting Tools | Power BI, Tableau. |
| Databases | MS SQL Server, MySQL, Azure SQL DB, Azure Synapse, PostgreSQL, MongoDB, DynamoDB, Oracle DB, Cosmos DB |
| Containerization Tools | Kubernetes, Docker. |
| Development Methodologies | Agile, Waterfall. |

**PROFESSIONAL EXPERIENCE:**

**Equality Health - Phoenix, Arizona Jul 2022– Present**

**Sr Data Engineer**

**Responsibilities:**

* Orchestrated all facets of the software engineering lifecycle, encompassing requirements analysis, application design, and meticulous code development and testing.
* Spearheaded the development and maintenance of end-to-end operations for ETL data pipelines, handling vast datasets within Azure Data Factory.
* Elevated data fetching efficiency through adept query optimization and indexing strategies.
* Proficiently crafted SQL queries, spanning DDL, DML, indexes, triggers, views, stored procedures, functions, and packages.
* Leveraged Azure Data Factory to seamlessly integrate on-premises (MYSQL, Cassandra) and cloud (Blob storage, Azure SQL DB) data, applying transformative processes before loading into Snowflake.
* Deployed Data Factory to establish robust data pipelines, orchestrating the flow of data into SQL databases.
* Applied advanced Snowflake modeling techniques, employing data warehousing principles, data cleansing, Slowly Changing Dimension (SCD) handling, surrogate key assignment, and change data capture.
* Demonstrated an analytical problem-solving approach, leveraging technologies like Azure Data Factory, Data Lake, and Azure Synapse to address complex business challenges.
* Engineered ELT/ETL pipelines, seamlessly moving data to and from Snowflake using a combination of Python and Snowflake Snow SQL.
* Designed and executed ETL workflows using DataStage, ensuring the integration of data from diverse sources into a comprehensive data warehouse with a focus on data quality and completeness.
* Proficiently crafted intricate SQL queries to extract, manipulate, and manage data within relational databases, including Oracle, MySQL, and SQL Server.
* Showcased expertise in creating and managing robust database structures, encompassing tables, views, indexes, and stored procedures using SQL.
* Implemented ETL transformations and validations using Spark-SQL/Spark Data Frames within Azure Databricks and Azure Data Factory.
* Collaborated with Azure Logic Apps administrators, playing a pivotal role in monitoring and troubleshooting issues related to process automation and data processing pipelines.
* Developed and fine-tuned code for Azure Functions, streamlining the extraction, transformation, and loading of data from diverse sources, including databases, APIs, and file systems.
* Designed, built, and maintained data integration programs, skillfully navigating Hadoop and RDBMS landscapes.
* Engineered a robust CI/CD framework for data pipelines using Jenkins, collaborating seamlessly with DevOps engineers to automate CI/CD and implement test-driven development pipelines in Azure.
* Demonstrated hands-on proficiency in scripting languages like Python and Scala.
* Executed Hive scripts through Hive on Spark and Spark SQL, actively participating in ETL tasks to maintain data integrity and ensure pipeline stability.
* Developed Spark core and Spark SQL scripts using Scala, contributing to expedited data processing.
* Utilized JIRA for project reporting, adeptly creating subtasks for Development, QA, and Partner validation.
* Showcased comprehensive experience in Agile ceremonies, actively participating in daily stand-ups and internationally coordinated PI Planning.

**Environment:** Azure Databricks, Data Factory, Logic Apps, Functional App, Snowflake, MS SQL, Oracle, HDFS, MapReduce, YARN, Spark, Hive, SQL, Python, Scala, Pyspark, shell scripting, GIT, JIRA, Jenkins, kafka, ADF Pipeline, Power Bi.

**Kroger - Cincinnati, Ohio Jan 2020– Jun 2022 Sr Data Engineer**

**Responsibilities:**

* Developed and automated Snowflake scripting solutions, enhancing efficiency in data pipelines, ETL processes, and data transformations.
* Architected robust data ingestion and storage solutions, leveraging AWS S3, Redshift, and Glue.
* Crafted intricate ETL workflows using AWS Glue for seamless extraction, transformation, and loading of diverse data sources into Redshift.
* Integrated AWS SNS and SQS to facilitate real-time event processing and messaging.
* Utilized AWS Step Functions for comprehensive workflow orchestration and monitoring.
* Implemented AWS Athena for ad-hoc data analysis and querying on S3 data.
* Leveraged AWS CloudWatch for vigilant resource monitoring, alarm configuration, and metric collection.
* Engineered data streaming solutions with AWS Kinesis to achieve real-time data processing capabilities.
* Managed DNS configurations and routing efficiently using AWS Route53 for streamlined application and service deployment.
* Designed and deployed Snowflake stages for optimal loading of data from various sources into Snowflake tables.
* Administered diverse Snowflake table types, including transient, temporary, and persistent tables.
* Developed and executed Python codes through AWS Lambda to invoke SNS, facilitating status emails for Glue job execution.
* Optimized Snowflake warehouses by selecting appropriate sizes and configurations for enhanced performance and cost efficiency.
* Constructed complex Snow SQL queries for extracting, transforming, and loading data from diverse sources into Snowflake.
* Deployed Databricks Delta Lake on AWS to introduce version control and ACID transactions, ensuring data consistency and reliability.
* Performed performance tuning on Databricks clusters to enhance query performance, reduce response times, and improve overall data processing efficiency.
* Implemented CI/CD practices for Databricks notebooks and pipelines, ensuring automated deployment and fostering agile development.
* Applied partitioning techniques in Snowflake to elevate query performance and streamline data retrieval.
* Configured and managed multi-cluster warehouses in Snowflake to handle high-concurrency workloads effectively.
* Developed fail-proof Python code in AWS Lambda for initiating and executing AWS Glue Jobs.
* Defined roles and access privileges in Snowflake to enforce robust data security and governance.
* Implemented Snowflake caching mechanisms to optimize query performance and reduce data transfer costs.
* Leveraged snow pipe for real-time data ingestion into Snowflake, ensuring continuous data availability and automated loading processes.
* Utilized Snowflake's time travel features for tracking and restoring historical data, enhancing auditing and analysis capabilities.
* Demonstrated proficiency in employing surrogate keys in Snowflake for unique identification and management of records in data warehousing environments.
* Implemented Change Data Capture (CDC) solutions in Snowflake for real-time tracking of data changes from source systems.
* Expertise in designing and maintaining Type 1, Type 2, and Type 3 SCDs to meet diverse data retention and tracking requirements.
* Leveraged Snowflake's time travel and versioning features to simplify SCD implementation and ensure data accuracy.
* Applied regular expressions in Snowflake for effective pattern matching and data extraction tasks.
* Designed and configured workflows for data processing and ETL pipelines, ensuring seamless operations.
* Utilized Git and GitLab for efficient version control.

**Environment:** Snowflake, AWS S3, Redshift, Glue, AWS SNS, AWS SQS, AWS Athena, AWS CloudWatch, AWS Kinesis, AWS Route53, Snow SQL, Hadoop, HDFS, Sqoop, Hive, MapReduce, Spark, IBM Tivoli, Control-M, Oozie, Airflow, Teradata, Oracle, SQL Server, Git, Git Lab, Databricks, Python, Boto3.

**Wells Fargo - Charlotte, NC Oct 2017 – Dec 2019**

**Sr Data Engineer**

**Responsibilities:**

* Imported data seamlessly from MySQL to HDFS using Sqoop, ensuring regular and efficient data loading processes.
* Performed sophisticated aggregations on extensive datasets employing Apache Spark and Scala, subsequently storing the data in Hive warehouse for in-depth analysis.
* Experienced in navigating Data Lakes and various big data ecosystems, including Hadoop, Spark, Hortonworks, and Cloudera.
* Crafted intricate Hive queries for thorough data analysis, aligning outcomes with precise business requirements.
* Leveraged HBASE integration with Hive to construct H-BASE tables in the Analytics Zone.
* Applied Kafka and Spark streaming to process streaming data in targeted use cases, enhancing real-time data capabilities.
* Developed a robust data pipeline using Flume and Sqoop to ingest and store customer behavioral data in HDFS for comprehensive analysis.
* Conducted in-depth analysis of the Hadoop cluster using various big data analytic tools such as Hive and MapReduce.
* Engineered a dynamic data pipeline integrating Kafka, Spark, and Hive for seamless data ingestion, transformation, and analysis.
* Formulated Hive queries tailored for specific business needs, creating and manipulating Hive tables to simulate MapReduce functionalities effectively.
* Implemented UNIX scripts to define workflow for use cases, facilitating data processing and job automation.
* Orchestrated deployment automation using YAML scripts for large-scale builds and releases.
* Executed the migration of existing data from RDBMS (Oracle) to Hadoop using Sqoop, ensuring efficient and accurate data processing.
* Utilized PySpark for Spark processing, employing Python for enhanced testing and data manipulation.
* Implemented Spark Streaming to partition streaming data into batches, optimizing input for batch processing by the Spark engine.
* Utilized Zookeeper for efficient coordination, synchronization, and serialization of servers within clusters.
* Operated the Oozie workflow engine for meticulous job scheduling.
* Employed Git as a version control tool for maintaining a streamlined and organized code repository.
* Conducted data analysis and processing using Spark SQL with PySpark, contributing to efficient data manipulation.
* Collaborated closely with the team to troubleshoot and resolve JVM-related issues effectively.

**Environment:** Sqoop, MYSQL, HDFS, Apache Spark Scala, Hive Hadoop, Cloudera, HBASE, Kafka, MapReduce, Zookeeper, Oozie, Data Pipelines, RDBMS, Python, PySpark, shell script, Ambari, JIRA.

**Visualsoft - Hyderabad, India Apr 2016-Sep 2017**

**Data Engineer**

**Responsibilities:**

* Worked in a dynamic role as a SQL Server Analyst, Developer, and DBA, specializing in SQL Server versions 2012, 2015, and 2016.
* Successfully designed and implemented jobs, SQL Mail Agent, Alerts, and scheduled DTS/SSIS Packages to optimize data management processes.
* Managed and updated Erwin models for Logical/Physical Data Modeling, overseeing Consolidated Data Store (CDS), Actuarial Data Mart (ADM), and Reference DB, aligning with user requirements.
* Executed the export of current Data Models into PDF from Erwin, publishing them on SharePoint for accessibility by various stakeholders.
* Proficient in writing Triggers, Stored Procedures, Functions, and Transact-SQL (TSQL) coding, ensuring the creation and maintenance of robust Physical Structures.
* Utilized GIT for version control, maintaining source code in Git and GitHub repositories.
* Developed a comprehensive ETL framework incorporating Sqoop, Pig, and Hive, facilitating seamless data extraction and availability for consumption.
* Processed HDFS data and created external tables using Hive, implementing scripts for table ingestion and repair, fostering reusability across projects.
* Designed and executed ETL jobs using Spark-Scala to migrate data from Oracle to new MySQL tables.
* Leveraged Spark-Scala and Spark-Cassandra-Connector APIs for diverse tasks such as data migration, business report generation, etc.
* Developed Spark Streaming applications for real-time sales analytics, ensuring responsiveness to dynamic business needs.
* Conducted in-depth analysis of source data, efficiently modifying data types and generating PowerBI ad-hoc reports using Excel sheets, flat files, and CSV files.
* Analyzed SQL scripts and devised solutions for implementation using PySpark, contributing to enhanced data processing efficiency.
* Extracted data from various sources into HDFS using Sqoop, ensuring seamless data integration.
* Managed the importing of data from MySQL into HDFS using Sqoop, streamlining data consolidation processes.
* Implemented automation for deployments through YAML scripts, contributing to efficient builds and releases.
* Expertise in Apache Hive, Apache Pig, HBase, Apache Spark, Zookeeper, Flume, Kafka, and Sqoop, ensuring comprehensive data ecosystem management.
* Implemented Data classification algorithms using MapReduce design patterns, contributing to advanced data processing capabilities.
* Extensively worked on creating combiners, Partitioning, and distributed cache to optimize the performance of MapReduce jobs, ensuring efficient data processing at scale.

**Environment:** SQL Server 2008/2012 Enterprise Edition, SSRS, SSIS, T-SQL, Shell script, Windows Server 2003, PerformancePoint Server 2007, Oracle 10g, Hadoop, Hive, spark, PySpark, Sqoop, Spark SQL, Shell script, Cassandra, YAML, ETL.

**Express Scripts - Bangalore, India Aug 2013-Mar 2016**

**Data Engineer**

**Responsibilities:**

* Experience in crafting intricate stored procedures, optimized triggers, essential functions, and implementing indexes and indexed views for enhanced performance.
* Demonstrated excellence in monitoring and fine-tuning SQL Server performance for optimal database operation.
* Played a key role in designing ETL data flows using SSIS, developing mappings and workflows for extracting data from SQL Server, and executing data migration and transformation from Access/Excel Sheets using SQL Server SSIS.
* Proficient in Dimensional Data Modeling for Data Mart design, adept at identifying Facts and Dimensions, and skilled in crafting fact tables and dimension tables, incorporating Slowly Changing Dimensions (SCD).
* Extensive experience in Error and Event Handling, employing tools like Precedence Constraints, Break Points, Check Points, and Logging for robust data management.
* Specialized in building Cubes and Dimensions with diverse architectures and data sources for Business Intelligence, including expertise in MDX Scripting.
* Hands-on involvement in developing SSAS Cubes, creating Aggregations, KPIs, Measures, Partitioning Cubes, Data Mining Models, and efficiently deploying and processing SSAS objects.
* Proven experience in generating Ad hoc reports and reports with intricate formulas, proficient in querying the database for Business Intelligence.
* Expertise in developing various types of reports on SSAS Cube using SSRS, including Parameterized, Chart, Graph, Linked, Dashboard, Scorecards, and reports with Drill-down, Drill-through, and Cascading features.

**Environment:** MS SQL Server 2016, Visual Studio 2017/2019, SSIS, Share point, MS Access, Team Foundation server, GIT.