PRANEETH REDDY POREDDY

AWS Data Engineer

Phone: +1 (972) 439-5728 **| Email:** praneethr042@gmail.com

PROFESSIONAL SUMMARY:

- Around 9+ years of experience as a Data Engineer, Data Analyst using Python, AWS, Snowflake, Hadoop, MongoDB, Cassandra.
- Experience in creating separate virtual data warehouses with difference size classes in AWS Snowflake.
- Experience in working with AWS S3 and Snowflake cloud Data warehouse.
- Result oriented and highly skilled professional with expertise in Snowflake, AWS, and Big Data technologies.
- Skilled in utilizing AWS EMR for big data processing, including technologies like Hadoop, Spark, Hive, MapReduce, and PySpark.
- Implemented AWS Cloud platform and its features which includes EC2, VPC, EBS, AMI, SNS, RDS, EBS, Cloud Watch, Cloud Trail, Cloud Formation AWS Config, Autos calling, CloudFront, IAM, S3, R53.
- Hands-on experience with Amazon EC2, Amazon S3, Amazon RDS, VPC, IAM, Amazon Elastic Load Balancing and other services of the AWS family.
- Provisioned the available EC2 Instances using Terraform and cloud formation and wrote new plugins to support new functionality in oracle form.
- Setup/Managing CDN on Amazon CloudFront to improve site performance.
- Expertise on working with MongoDB, Apache Cassandra.
- Experience in UNIX shell scripting for processing large volumes of data from varied sources and loading into databases like AWS Redshift, Snowflake.
- Well versed with Big data on AWS cloud services i.e., EC2, S3, Glue, Lambda Functions, Athena, EMR and RedShift.
- Proficient in using SnowSQL for complex data manipulation tasks and developing efficient data pipelines.
- Experienced in partitioning strategies and multi-cluster warehouses in Snowflake to ensure optimal query performance and scalability.
- Skilled in designing roles, views, and implementing performance tuning techniques to enhance Snowflake system performance.
- Proficient in utilizing virtual warehouses, caching, and Snowpipe for real-time data ingestion and processing in Snowflake.
- Strong knowledge of Snowflake's time travel feature for auditing and analyzing historical data.
- Extensive experience in leveraging window functions, Snowflake arrays, regular expressions, and JSON parsing for advanced data analysis and manipulation.
- Highly proficient in Snowflake scripting to automate ETL processes, data transformations, and data pipelines.
- Expertise in AWS S3 for scalable and cost-effective data storage and retrieval.
- Experienced in utilizing AWS Glue for ETL workflows, enabling efficient data extraction, transformation, and loading.
- Strong knowledge of AWS CloudWatch for monitoring and managing AWS resources, setting up alarms, and collecting metrics.
- Extensive experience in working with HDFS, Sqoop, PySpark, Hive, MapReduce, and HBase for big data processing and analytics.
- Proficient in developing and optimizing Spark and Spark Streaming applications for real-time data processing and analytics.
- Experienced in scheduling and workflow management using IBM Tivoli, Control-M, Oozie, and Airflow for efficient job orchestration.

- Implemented query performance in Hive using bucketing and partitioning techniques, and have extensive hands-on experience tuning spark Jobs.
- Implemented scheduling Hadoop jobs using Apache Oozie, Importing and exporting the data using SQOOP from HDFS to Relational Database systems.
- Strong database development skills in Teradata, Oracle, SQL Server, including the development of stored procedures, triggers, and cursors.
- Proficient in version control systems like Git, GitLab, and VSS for code repository management and collaboration.

TECHNICAL SKILLS:

AWS Services	AWS s3 , redshift, EC2, EMR, SNS, SQS, Athena, glue, CloudWatch, kinesis, route53, IAM, Lambda.
Big Data Technologies	HDFS, SQOOP, PySpark, hive, MapReduce, spark, spark streaming, HBASE, kafka, Zookeeper.
Hadoop Distribution	Cloudera, Horton Works
Languages	Java, SQL, PL/SQL, Python, HiveQL, Scala.
Operating Systems	Windows (XP/7/8/10), UNIX, LINUX, UBUNTU, CENTOS.
Database	Teradata, oracle, SQL server, Snowflake, MYSQL, MongoDB, Cassandra, PostgreSQL.
Scheduling	IBM Tivoli,control-m,oozie,airflow
Version Control	GIT, GitHub, VSS, Bitbucket.
Methodology	Agile, Scrum.
IDE & Build Tools, Design	Eclipse, Visual Studio.

EDUCATION:

- Bachelors from KL University, India.

WORK EXPERIENCE:

Role: AWS Data Engineer | Apr 2022 – Till Now

Client: Walmart, Dallas, Tx

Responsibilities:

- Developing ETL pipelines to move on-prem data (data sources that include Flat Files, Mainframe Files, and Databases) to AWS S3 using PySpark. Created and embedded python modules in the ETL pipeline to automatically migrate data from S3 to Redshift using AWS Glue.
- Collected data using Spark Streaming from AWS S3 bucket in near-real-time and performs necessary Transformations and Aggregation on the fly to build the common learner data model and persists the data in HDFS.
- Used AWS cloud product suites (S3, EMR, SQS, Redshift), Spark SQL
- Operated on AWS cloud RDS, Athena, Cloud Watch, EC2, IAM policies
- Designed and architected solutions to load multipart files which can't rely on a scheduled run and must be event driven, leveraging AWS SNS, SQS and Glue.
- Created Lambda function to run the AWS Glue job based on the defined Amazon S3 event.
- Optimize the Glue jobs to run on EMR Cluster for faster data processing.
- Performed data preprocessing and feature engineering for further predictive analytics using Python Pandas.
- Explored the usage of Spark for improving the performance and optimization of the existing algorithms in Big Data using Spark Context, Spark SQL and Spark Yarn.
- Involved in converting Hive/SQL queries into Spark Transformations using Spark RDDs
- Used Oozie workflow engine to manage independent Hadoop jobs and to automate several types of Hadoop such as Hive and Sqoop as well as system specific jobs

- Working with Amazon Web Services (AWS), AWS Cloud Formation, AWS CloudFront and using containers like Docker and familiar with Jenkins.
- 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.
- Implemented a 'server less' architecture using API Gateway, Lambda, and DynamoDB and deployed AWS Lambda code from Amazon S3 buckets.
- Worked on Cognos modernization which involves migrating the Cognos data source to AWS data lake
- Built Data blocks in data lake using the raw tables as data source
- Git is used as a version control tool and Jenkins as Continuous Integration (CI) tool.
- Published the dashboard reports to Tableau Server for navigating the developed dashboards in web.
- Followed agile methodology and actively participated in daily scrum meetings.

Environment: AWS, AWS S3, redshift, EMR, SNS, SQS, aetna, glue, cloudwatch, kinesis, route53, IAM, Sqoop, MYSQL, HDFS, Apache Spark, Python, Hive, Cloudera, Kafka, Zookeeper, Oozie, PySpark, Ambari, JIRA, IBM Tivoli, control-m, OOZIE, airflow, Teradata, oracle, SQL

Role: AWS Snowflake Engineer | Nov 2020 – Mar 2022

Client: MUFG, Phoenix, Arizona

Responsibilities:

- Developed Talend Big Data jobs to load heavy volume of data into S3 data lake and then into Snowflake.
- Developed snow pipes for continuous injection of data using event handlers from AWS (S3 bucket).
- Developed Snow Sql scripts to deploy new objects and update changes into Snowflake.
- Developed a Python script to integrate DDL changes between on-prem Talend warehouse and snowflake.
- Working with AWS stack S3, EC2, Snowball, EMR, Athena, Glue, Redshift, DynamoDB, RDS, Aurora, IAM, Firehose, and Lambda.
- Designing and implementing new HIVE tables, views, schema and storing data optimally.
- Performing Sqoop jobs to land data on HDFS and running validations.
- Configuring Oozie Scheduler Jobs to run the Extract jobs and queries in an automated way.
- Querying data by optimizing the query and increasing the query performance.
- Designing and creating SQL Server tables, views, stored procedures, and functions.
- Performing ETL operations using Apache Spark, also using Ad-Hoc queries and implementing Machine Learning techniques.
- Worked on configuring CICD for CaaS deployments (k8's).
- Involved in migrating master-data from Hadoop to AWS.
- Worked with Spark for improving performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frames, Pair RDD's.
- Developed preprocessing job using Spark Data frames to transform JSON documents to flat file
- Loaded D-Stream data into Spark RDD and did in-memory data computation to generate output response
- Processing with Amazon EMR big data across a Hadoop cluster of virtual servers on Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3).
- Imported data from AWS S3 into Spark RDD, performed transformations and actions on RDD's.
- Worked on Big Data infrastructure for batch processing and real-time processing using Apache Spark
- Developed Apache Spark applications by using Scala for data processing from various streaming sources
- Processed the Web server logs by developing Multi-Hop Flume agents by using Avro Sink and loaded into Cassandra for further analysis, Extracted files from Cassandra through Flume
- Responsible for design and development of Spark SQL Scripts based on Functional Specifications
- Worked on the large-scale Hadoop YARN cluster for distributed data processing and analysis using Spark, Hive, and Cassandra
- Involved in converting Cassandra/Hive/SQL queries into Spark transformations using RDD's and Scala
- Implemented Spark Scripts using Scala, Spark SQL to access hive tables to spark for faster processing of data.
- Developed Some Helper class for abstracting Cassandra cluster connection act as core toolkit
- Involved in creating Data Lake by extracting customer's data from various data sources to HDFS which include data from Excel, databases, and log data from servers

- Moved data from HDFS to Cassandra using Map Reduce and Bulk Output Format class.
- Extracted files from Cassandra through Sqoop and placed in HDFS and processed it using Hive
- Writing MapReduce (Hadoop) programs to convert text files into AVRO and loading into Hive table
- Experienced in writing real-time processing and core jobs using Spark Streaming with Kafka as a data pipeline system
- Extending HIVE/PIG core functionality by using custom User Defined Functions (UDF), User Defined Table-Generating Functions (UDTF) and User Defined Aggregate Functions (UDAF) for Hive and Pig
- Involved in loading data from rest endpoints to Kafka producers and transferring the data to Kafka brokers
- Used Apache Kafka functionalities like distribution, partition, replicated commit log service for messaging
- Partitioning Data streams using Kafka. Designed and configured Kafka cluster to accommodate heavy throughput.
- Exported the analyzed data to the relational databases using Sqoop for visualization and to generate reports for the BI team
- Used Apache Oozie for scheduling and managing multiple Hive Jobs. Knowledge of HCatalog for Hadoop based storage management
- Migrated an existing on-premises application to Amazon Web Services (AWS) and used its services like EC2 and S3 for small data sets processing and storage, experienced in maintaining the Hadoop cluster on AWS EMR
- Developed solutions to pre-process large sets of structured, semi-structured data, with different file formats like Text, Avro, Sequence, XML, JSON, and Parquet
- Generated various kinds of reports using Pentaho and Tableau based on Client specification
- Have come across new tools like Jenkins, Chef and RabbitMQ.
- Worked with SCRUM team in delivering agreed user stories on time for every Sprint

Environment: AWS, AWS S3, redshift, EMR, SNS, SQS, Athena, glue, cloudwatch, kinesis, route53, IAM, Sqoop, MYSQL, HDFS, Apache Spark, Python, Hive, Cloudera, Kafka, Zookeeper, Oozie, PySpark, Ambari, JIRA, IBM Tivoli, control-m, OOZIE, airflow, Teradata, oracle, SQL

Role: Big Data Developer | Aug 2019 – Oct 2020

Client: Aetna Inc., Hartford, CT

Responsibilities:

- Imported data from MySQL to HDFS on a regular basis using Sqoop for efficient data loading.
- Performed aggregations on large volumes of data using Apache Spark and Scala, and stored the results in the Hive data warehouse for further analysis.
- Worked extensively with Data Lakes and big data ecosystems, including Hadoop, Spark, Hortonworks, and Cloudera.
- Loaded and transformed structured, semi-structured, and unstructured data sets efficiently.
- Developed Hive queries to analyze data and meet specific business requirements.
- Leveraged HBASE integration with Hive to build HBASE tables in the Analytics Zone.
- Utilized Kafka and Spark Streaming to process streaming data for specific use cases.
- Developed data pipelines using Flume and Sqoop to ingest customer behavioral data into HDFS for analysis.
- Utilized various big data analytic tools, such as Hive and MapReduce, to analyze Hadoop clusters.
- Implemented a data pipeline using Kafka, Spark, and Hive for ingestion, transformation, and analysis of data.
- Wrote Hive queries and used Hive QL to simulate MapReduce functionalities for data analysis and processing.
- Migrated data from RDBMS (Oracle) to Hadoop using Sqoop for efficient data processing.
- Developed custom scripts and tools using Oracle's PL/SQL language to automate data validation, cleansing, and transformation processes.
- Implemented CI/CD pipelines for building and deploying projects in the Hadoop environment.
- Utilized JIRA for issue and project workflow management.
- Utilized PySpark and Spark SQL for faster testing and processing of data in Spark.
- Used Spark Streaming to process streaming data in batches for efficient batch processing.
- Leveraged Zookeeper to coordinate, synchronize, and serialize servers within clusters.
- Utilized the Oozie workflow engine for job scheduling in Hadoop.
- Utilized PySpark in SparkSQL for data analysis and processing.
- Used Git as a version control tool to maintain the code repository.

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

Role: Hadoop Developer | May 2018 – Jul 2019
Client: JP MORGAN CHASE, West Haven, CT
Responsibilities:

- Developed ETL jobs using Spark -Scala to migrate data from Oracle to new MySQL tables.
- Rigorously used Spark -Scala (RRD's, Data frames, Spark SQL) and Spark - Cassandra -Connector API's for various tasks (Data migration, Business report generation etc.)
- Developed Spark Streaming application for real time sales analytics
- Prepared an ETL framework with the help of sqoop, pig and hive to be able to frequently bring in data from the source and make it available for consumption
- Processed HDFS data and created external tables using Hive and developed scripts to ingest and repair tables that can be reused across the project.
- Analyzed the source data and handled efficiently by modifying the data types. Used excel sheet, flat files, CSV files to generate PowerBI ad-hoc reports
- Analyzed the SQL scripts and designed the solution to implement using PySpark
- Extracted the data from other data sources into HDFS using Sqoop
- Handled importing of data from various data sources, performed transformations using Hive, MapReduce, loaded data into HDFS.
- Extracted the data from MySQL into HDFS using Sqoop
- Implemented automation for deployments by using YAML scripts for massive builds and releases
- Apache Hive, Apache Pig, HBase, Apache Spark, Zookeeper, Flume, Kafka and Sqoop.
- Implemented Data classification algorithms using MapReduce design patterns.
- Extensively worked on creating combiners, Partitioning, distributed cache to improve the performance of MapReduce jobs.
- Worked on GIT to maintain source code in Git and GitHub repositories
- **Environment**: Hadoop, Hive, spark, PySpark, Python, Sqoop, Spark SQL, Cassandra, YAML, ETL.

Role: Data Warehouse Developer | Apr 2016 – Apr 2018
Client: Mayo Clinic, Rochester, MN
Responsibilities:

- Creating jobs, SQL Mail Agent, alerts, and scheduling DTS/SSIS packages for automated processes.
- Managing and updating Erwin models for logical/physical data modeling of Consolidated Data Store (CDS), Actuarial Data Mart (ADM), and Reference DB to meet user requirements.
- Utilizing TFS for source controlling and tracking environment-specific script deployments.
- Exporting current data models from Erwin to PDF format and publishing them on SharePoint for user access.
- Developing, administering, and managing databases such as Consolidated Data Store, Reference Database, and Actuarial Data Mart.
- Writing triggers, stored procedures, and functions using Transact-SQL (T-SQL) and maintaining physical database structures.
- Deploying scripts in different environments based on Configuration Management and Playbook requirements.
- Creating and managing files and filegroups, establishing table/index associations, and optimizing query and performance tuning.
- Tracking and closing defects using Quality Center for effective issue management.
- Maintaining users, roles, and permissions within the SQL Server environment.
- **Environment**: SQL Server 2008/2012 Enterprise Edition, SSRS, SSIS, T-SQL, Windows Server 2003, PerformancePoint Server 2007, Oracle 10g, Visual Studio 2010.

Role: Data Warehouse Developer | Feb 2014 – Mar 2016
Client: Charter Communications, Negaunee, MI

Responsibilities:

- Expert in designing ETL data flows using SSIS, creating mappings/workflows to extract data from SQL Server and Data Migration and Transformation from Access/Excel Sheets using SQL Server SSIS.
- Efficient in Dimensional Data Modeling for Data Mart design, identifying Facts and Dimensions, and developing, fact tables, dimension tables, using Slowly Changing Dimensions (SCD).
- Experience in Error and Event Handling: Precedence Constraints, Break Points, Check Points, Logging.
- Experienced in Building Cubes and Dimensions with different Architectures and Data Sources for Business Intelligence and writing MDX Scripting.
- Thorough knowledge of Features, Structure, Attributes, Hierarchies, Star and SnowFlake Schemas of Data Marts.
- Good working knowledge on Developing SSAS Cubes, Aggregation, KPIs, Measures, Partitioning Cube, Data Mining Models and Deploying and Processing SSAS objects.
- Experience in creating Ad hoc reports and reports with complex formulas and to query the database for Business Intelligence.
- Expertise in developing Parameterized, Chart, Graph, Linked, Dashboard, Scorecards, Report on SSAS Cube using Drill-down, Drill-through and Cascading reports using SSRS.
- Flexible, enthusiastic and project oriented team player with excellent written, verbal communication and leadership skills to develop creative solutions for challenging client needs.
- **Environment**: MS SQL Server 2016, Visual Studio 2017/2019, SSIS, Share point, MS Access, Team Foundation server, Git.