**Pradeep Pal**

(402)-591-0374 | pradeeppal0072@gmail.com

https://www.linkedin.com/in/pradeep-pal-17b76a46/

**PROFESSIONAL SUMMARY:**

* Over 10+ years of work experience in IT industry working with Advanced Analytics using Big Data Technologies (Pyspark, AWS, Hadoop and SnowFlake).
* Expertise in designing and developing Big-data solutions using Amazon Web Services (S3,Lambda Redshift, EC2, EMR,AWS Athena, AWS Glue, AWS Lambda).
* Strong Hands-on Experience in AWS Glue with Python(PySpark), SSIS, Apache Nifi, Sqoop, Hive.
* Expertise in supporting data analysis projects by using RedShift, DynamoDB and EMR on the Amazon Web Services (AWS) cloud. Performed Export and import of data into S3.
* Hands on experience with Snowflake(Snow SQL and Snowpark API).
* Expertise in ETL/ELT operations using Airflow.
* Extensive experience on Python-BOTO3 to automate AWS connectivity with ETL jobs and frameworks.
* Experience with AWS Cloud Formation, deploying and developing Spark jobs in EMR Cluster.
* Worked and experienced a great deal from Amazon Web Services (AWS) Cloud services like EC2, S3 and VPC.
* Automated the complex workflows using the Airflow workflow handler.
* Expertise in developing Big Data applications and identify various data patterns using HDFS, PySpark, Spark SQL, Hive, Sqoop, YARN, Zookeeper, Kafka.
* Worked with various formats of files like delimited text, CSV, Parquet files, click stream log files, Apache log files, Parquet files, JSON files, XML Files.
* In depth understanding of Monitoring/Auditing tools in AWS such as CloudWatch and Cloud Trail.
* Good understanding of NOSQL databases- DynamoDB, MongoDB and relational databases- SQL server & Oracle.
* Gain expertise in Apache Spark & Hive Optimization Techniques on Big data platforms.
* Experience in using CRON, Shell Scripting and version control tools like SVN and Github.

**TECHNICAL SKILLS:**

**Software Methodologies/tools:** SDLC, Agile, Scrum, Jira, Slack.

**Programming Languages:** Python, PySpark, C#, T-SQL, Unix shell Scripting.

**ETL:** SSIS, Informatica, Apache NiFi, AWS Glue, MicroStrategy, Tableau, Power BI.

**Relational Databases & NoSQL:** MS SQL Server, Oracle, Snowflake, Redshift, DynamoDB, MongoDB.

**Tools:** PyCharm, Eclipse, GitHub, Visual Studio, Bitbucket, Bamboo, Jerkins.

**Cloud Services:** AWS RDS, EC2, S3, CloudFormation, AWS Glue, Athena, AWS EMR, Redshift.

**EDUCATION & CERTIFICATION:**

* Master of Science, Management Information System University of Nebraska at Omaha - May 2018
* Bachelor of Technology in Metallurgical Engineering National Institute of Technology, Raipur, INDIA - May 2011
* Data Science, R Programming, Regression Model, Machine Learning

**PROFESSIONAL EXPERIENCE:**

**July ‘23 – Present** *AWS Data Engineer,* **Vanguard, Malvern, PA  
Project Descriptions - Automated data pipeline-** Aim of the project is to collect Bank Invoice csv file from the AWS S3 by triggering AWS Glue job from Lambda using S3 event and populated output table in Athena platform.

* Designed and implemented AWS Lambda functions to orchestrate and automate data processing workflows, improving operational efficiency, reducing processing time by 40% and enhancing data pipeline efficiency.
* Utilized AWS Glue crawlers to automate data discovery and cataloging processes, ensuring data lineage and metadata accuracy, resulting in streamlined data management and improved governance.
* Managed IAM roles effectively to enforce least privilege access principles, enhancing data security and compliance with industry regulations and organizational policies.
* Integrated AWS Lambda with other AWS services, such as S3, DynamoDB, and API Gateway, to build end-to-end serverless solutions.
* Integrated PySpark with AWS Glue ETL service, streamlining data ingestion, transformation, and loading processes.
* Skilled in using PySpark to perform complex data transformations, data cleansing, and data enrichment to ensure data quality and accuracy.
* Proven track record of optimizing PySpark jobs for performance by fine-tuning configurations, partitioning data, and utilizing cluster resources effectively on AWS.
* Proficient in Hive Query Language (HQL) and experienced in seamlessly transitioning HQL skills to AWS Athena's Presto-based SQL for querying and analyzing data stored in Amazon S3.
* Adept at crafting and optimizing complex SQL queries in Athena, leveraging its distributed query engine to achieve high performance and efficient data processing.
* Successfully integrated AWS Athena with data lakes in Amazon S3, enabling on-demand, serverless querying of vast datasets without the need for complex infrastructure management.
* Proficiently utilized AWS Glue Data Catalog to automatically discover and catalog metadata, facilitating schema management and enhancing data accessibility for analysis in Athena.
* Designed DAGs for Apache Airflow to schedule jobs for daily execution for QA and Production environment.
* Implemented query performance tuning techniques in Athena, such as partition pruning, predicate pushdown, and query optimization, resulting in significant reductions in query execution times.
* Designed and implemented fully operational solutions on Snowflake Data Warehouse, including in-depth knowledge of Snowflake internals and externals.

**Environment:** AWS, Lambda, AWS Glue, S3, DynamoDB, IAM, PySpark, ETL, AWS Athena, Data Lakes, HQL, SQL.

**Mar ‘22 – Jun ‘23**  *Data Engineer*, **Jewelry Television, Knoxville, TN**

**Project Description: Market Refer Visitor for JTV-** Aim of the project is to collect data from various source like JTV e-commerce website Using Tealium IQ Tag management, Facebook, Instagram, Pinterest, and Adobe campaign and perform ELT on it and load data into DWH for further analysis using MicroStrategy, Athena & Quick sight.

* Collaborate with business stakeholders, team members, and customers to assist them with data-related technical issues and support their data infrastructure needs.
* Involved in the full development cycle of Planning, Analyzing, Design, Development, Testing, and Implementation of orchestrated data pipelines for real-time and batch data processing in Agile Scrum Methodology.
* Provisioned key AWS Cloud services and configured them for scalability, flexibility, and cost optimization.
* Developed data ingestion modules using AWS Glue for loading data in various layers in S3 and reporting with Athena and Quick sight.
* 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 tableau dashboards.
* Load the data into EMR from various sources S3 process it using Hive Scripts and results are stored in Redshift.
* Developed various SQL scripts in Redshift to analyze data and prepared Design patterns and performance tuning documents for Redshift.
* Created real-time/batch streaming of large variety of data using Apache Nifi- Kafka dynamically loaded into HDFS.
* Built Apache Nifi workflow DAGs for ingesting various external data sources such as Facebook, Pinterest, Bambuser and Tealium iQ Tag Management (AWS S3) and move to HDFS (NoSQL database MapR-DB) on daily basis.
* Developed Spark & Drill engine-based Applications by using Python and implement data processing project to handle raw data from various RDBMS sources and populated tables into HDFS (MapR-DB).
* Used python scripts to write test-cases against the development code to ensure bug-free code deployed.
* Designed DAGs for Apache Airflow to schedule jobs for daily execution for QA and Production environment.
* Developed MicroStrategy Dynamics Dashboard for monitoring real-time analysis reports and User activity.
* Used Apache Drill/ dbshell for querying NoSQL database MapR-DB for data validation and QA testing.
* Designed and Developed ETL Processes in AWS Glue to migrate Campaign data from external source to AWS Redshift.
* Responsible for creating on-demands tables on S3 files AWS Glue using Pyspark.
* Used AWS Glue for the data transformation, validate and data cleansing.
* Built and maintained data pipelines using AWS Glue and Snowpipe
* Developed and executed data transformation scripts using Python and SnowSQL
* Worked with data scientists to build and deploy machine learning models in Snowflake for sentiment analysis.
* Implemented Partitioning and Bucketing on Hive tables and enhanced the performance by up to 30%.
* Optimized HDFS storage efficiency using compression techniques and file formats such as ORC, Parquet, Avro.

**Environment:** ETL, Apache Nifi- Kafka, Pyspark, HDFS, Hive, MapR, AWS, Redshift, Glue, S3, NO SQL, SnowFlakes.

**Mar ‘20 – Mar ‘22** *Data Engineer/Risk Analyst,* **USAA, San Antonio, TX**

**Project Description: Fraud Pattern Recognition -** Utilizing big data tools like Spark, PySpark, and cloud technologies, this project aims to detect and prevent fraudulent activities within a leading investment bank, USAA. Leveraging AWS S3 and Apache Airflow, the project focuses on enhancing fraud pattern recognition for improved security and risk mitigation.

* Designed and developed a security framework to provide fine-grained access to AWS S3 objects using AWS Lambda and DynamoDB.
* Used AWS Athena to ingest structured data from S3 into other systems, such as RedShift, and to produce reports.
* Designed and deployed ETL pipelines on S3 parquet files in a data lake using AWS Glue.
* Created a cloud formation template in JSON format for cross-region replication and content delivery through AWS Virtual Private Cloud.
* Used AWS Code Commit Repository to save programming logic and scripts, replicating them to new clusters.
* Performed data transformation tasks like data cleansing, data enrichment, aggregation, and data format conversions using AWS Glue activities and Data Flow transformations.
* Implemented columnar data storage, advanced compression, and massive parallel processing using Multi-Node Redshift technology.
* Transferred a quality monitoring program from AWS EC2 to AWS Lambda and created logical datasets to administrate quality monitoring on snowflake warehouses.
* Responsible for data mapping, ETL data structures and analysis, traceability of data entities structures to system capabilities, and aligning business operational activities to the data business rules to drive the data model design development.
* Proven expertise in designing and implementing fully operational solutions on Snowflake Data Warehouse, including in-depth knowledge of Snowflake internals.
* Hands-on experience with Snowflake utilities, administration, and data loading from cloud and APIs, especially AWS.
* Demonstrated proficiency in Snowflake internals and seamlessly integrated Snowflake with other data processing and reporting technologies.
* Managed Snowflake utilities, such as SnowSQL and SnowPipe, while also overseeing Snowflake administration tasks.
* Led the creation of Data models for Snowflake, enhancing data transformation and analysis capabilities.

**Environment:** AWS, Elastic Map Reduce cluster, EC2s, CloudFormation, AWS S3, AWS Redshift, Dynamo DB, Cloud Watch, SnowFlake, Hive, Python, Apache Spark, Spark SQL, Shell Scripting, Tableau, Cloudera.

**Mar ‘19 - Feb ‘20** *Data Engineer/BI Analyst,* **Midcontinent Communication, Sioux Falls, SD**

**Project Description: Data pipeline application for ingesting daily-data from the Point-of-Sale system (POS)-** End-to-end DWH scope of loading data into model tables where all the channel partners, campaign team, sales team and operation teams have access to database to check the Opt-out information of the customer.

* Implemented full life cycle which involves Gap analysis, Requirement analysis, Design Data model, ETL Design, Testing, Documentation and Reporting.
* Gather requirements from the product owner to develop user stories using Jira and ensure solutions meet business needs and requirements at the end of the Agile sprint.
* Loading data from multiple sources on AWS S3 cloud storage.
* Creating S3 buckets and managing policies for S3 buckets and Utilized S3 bucket and Glacier for storage and backup on AWS.
* Load the data into EMR from various sources S3 process it using Hive/Pyspark Scripts and results are stored in Redshift.
* Developed a Python Script to load the CSV files into the S3 buckets and created AWS S3buckets, performed folder management in each bucket, managed logs and objects within each bucket.
* Created data partitions on large data sets in S3 and DDL on partitioned data.
* Developed Spark eco system using Spark SQL queries on different formats like Text file, CSV file.
* Loading data into Snowflake tables from the internal stage using SnowSQL.
* Involved in writing Spark applications for Data validation, cleansing, transformations and custom aggregations.
* Converted Hive queries into Spark transformations using Spark RDDs.
* Imported data from AWS S3 into Spark RDD, Performed transformations and actions on RDD's.
* Experienced in querying data using SparkSQL on top of Spark engine for faster data sets processing.
* Worked on Spark SQL for joining multi hive tables and write them to a final hive table and stored them on S3.
* Experience in hive partitioning, bucketing and perform joins on hive tables and utilizing hive SerDes like REGEX, JSON and Parquet.
* Fine Tuned Spark Applications to optimize the performance of long running jobs.
* Imported and exported databases using Microsoft Integration Services (SSIS) and Data Transformation Services (DTS Packages).
* Developed Complex ETL Packages using SSIS to extract data from staging tables to ODS and then to data warehouse tables using incremental load using various source systems.

**Environment:** AWS S3, Pyspark , SnowFlake, Redshift, AWS EMR, Hive, Star Schema, SQL, BI, Tableau, SSIS.

**Jul ‘18 - Dec ‘18** *Market Data Engineer/Analyst Intern,* **Valmont Industries, Inc., Omaha, NE**

* Responsible for broad range of market research analysis projects using advanced statistical methods.
* Performed the data cleaning and transformation using python and segregated the integration (Client and product performance) data by North American regions for predicting sales.
* Created data visualization dashboard for stocks market trends using Tableau and R (shiny).
* Analyzed complex data sets using SQL and performed ad-hoc analysis and data manipulation.
* Converted existing SQL queries into HiveQL queries.
* Responsible for troubleshooting issues in the execution of Mapreduce jobs by inspecting and reviewing log files.
* Developed various Python scripts to find vulnerabilities with SQL Queries by doing SQL injection, permission checks and performance analysis.
* Analyzed the weblog data using the HiveQL, integrated Oozie with the rest of the Hadoop stack.
* Used Oozie Scheduler system to automate the pipeline workflow and orchestrate the Map Reduces jobs that extract the data on a timely manner. Responsible for loading data from UNIX file system to HDFS.

**Environment:** Tableau, R, SQL, Oozie, HiveQL

**Jul ‘11 - Oct ‘15** *Operation Data Analyst,* **Monnet Ispat & Energy Ltd, (MIEL), India**

**Key Performance Index of Iron and Steel production analysis:** This project aims to collect data from the FTP server and create data-pipeline to load data into in-premises data warehouse and build visualization dashboard on top of it for further analysis.

* Facilitated cross-functional communication with Product Owner, SMEs, Software developer, Solution architect, and other teams to gain valuable insights that will lead to business decisions and opportunities.
* Built data pipelines using Microsoft SQL Server (SSIS) which includes extracting survey data from 3rd party vendor using REST APIs in Script task using C# programming.
* Proficient in Normalization (3rd normal form), De-normalization (Star Schema) for Data Warehousing.
* Hands-on experience in Performance Tuning and Query Optimization using Clustered and Non-Clustered Indexes
* to keep up the SQL server performance.
* Hands-on experience of creating Stored Procedure and Views to support BI solution using Tableau.
* Proficient in TFS to add new business logic, fixed defects using C# Script, to check-in those changes.
* For scheduling solutions using MS SQL Agent to create, schedule SSIS jobs and sending e-mails/Alerts.
* Utilized Alteryx, Excel, and SQL to create dashboards to track employee productivity, performance metrics, and KPI’s and Visualizations.
* Generated quarterly reports of finding to help management with making key decisions.
* Increased annual revenue by 7% by recommending improvements in operations efficiency.
* Performed CRUD operations in MS SQL Server.
* Implementing end-to-end database solutions using SSMS and SQL to optimize business processes.
* Built SQL Server Reporting Services (SSRS) for identifying trends, tracking Weekly/Monthly Key performance index (KPI) of production.

**Environment:** SSIS, SSRS, SSMS, SQL, Alteryx, Tableau.