Bharath Kumar

bharathk0307@gmail.com | PH: 913-730-0365 |LinkedIn: bharathkumar1528

Data Engineer

PROFESSIONAL SUMMARY

- Data engineering professional with extensive skills and a track record of successful implementation across diverse platforms, driven by self-motivation and a strong sense of personal accountability in both individual and team contexts.
- With over 8 years of experience as a Senior Data Engineer/Data Developer and Modeler, focusing on Data Engineering, Pipeline Design, Development, and Implementation.
- Extensive experience in Software Development Life Cycle (SDLC), encompassing Requirements Analysis, Design Specification, and Testing in both Waterfall and Agile methodologies.
- Strong experience in scripting with Python, PySpark, and Spark APIs for data analysis.
- Extensively used Python Libraries PySpark, Pytest, Pymongo, cxOracle, PyExcel, Boto3, Psycopg, embedPy, NumPy and Beautiful Soup.
- Experience in Google Cloud components, Google container builders and GCP client libraries and cloud SDK's
- Hands-on use of Spark and Scala API's to compare the performance of Spark with Hive and SQL, and Spark SQL to manipulate Data Frames in Scala.
- Proficient in Python and Scala, developing user-defined functions for Hive and Pig using Python.
- Experience in developing Map Reduce Programs using Apache Hadoop for analyzing the big data as per the requirement.
- Experience in working with Flume and NiFi for loading log files into Hadoop.
- Experience in working with NoSQL databases like HBase and Cassandra.
- Experienced in creating shell scripts to push data loads from various sources from the edge nodes onto the HDFS.
- Extensive experience in orchestrating data pipelines using both Oozie and Airflow, ensuring smooth data flow and task coordination.
- Worked with Cloudera and Hortonworks distributions.
- Expertise in developing SSIS/DTS packages for ETL processes, extracting, transforming, and loading data from various sources into data warehouses and data marts.
- Good working knowledge of Amazon Web Services (AWS) Cloud Platform which includes services like EC2, S3, VPC, ELB, IAM, DynamoDB, Cloud Front, Cloud Watch, Route 53, Elastic Beanstalk (EBS), Auto Scaling, Security Groups, EC2 Container Service (ECS), Code Commit, Code Pipeline, Code Build, Code Deploy, Dynamo DB, Auto Scaling, Security Groups, Red shift, CloudWatch, CloudFormation, CloudTrail, Ops Works, Kinesis, IAM, SQS, SNS, SES.
- Hands-on experience in Data Analysis, Profiling, Integration, Migration, governance, Metadata Management, Master Data Management, and Configuration Management.
- Experience in developing customized UDF's in Python to extend Hive and Pig Latin functionality.
- Skilled in designing complex mappings and proficient in performance tuning, particularly with slowly changing dimension tables and fact tables.

- Experienced in developing Automation Regression scripts to validate ETL processes across multiple databases including Oracle, SQL Server, Hive, and MongoDB using Python.
- Proficient in SQL across multiple dialects, including MySQL, PostgreSQL, Redshift, SQL Server, and Oracle.
- Expert in building Enterprise Data Warehouse or Data warehouse appliances from scratch using both Kimball and Inmon's Approach.
- Experience in designing star schema, Snowflake schema for Data Warehouse, ODS architecture.
- Skilled in System Analysis, E-R/dimensional Data Modeling, Database Design, and implementing RDBMS-specific features.
- Experienced in Normalization and De-Normalization techniques for optimizing performance in relational and dimensional database environments.
- Good knowledge of Data Marts, OLAP, and Dimensional Data Modeling using Ralph Kimball Methodology, including Star Schema Modeling and Snowflake Modeling with Analysis Services.

PROFESSIONAL EXPERIENCE

Sr. DATA ENGINEER

AIG, Houston, TX **May 2022 to Present**

RESPONSIBILITIES:

- Worked closely with Business Analysts and Subject Matter Experts (SMEs) from various departments to collect business needs and pinpoint actionable tasks for future progress.
- Teamed up with ETL developers to ensure data cleanliness and the timely updating of the data warehouse for reporting needs, using Pig.
- Selected and exported data into CSV files, saving them onto AWS S3 via AWS EC2. Afterward, I organized and stored the data in AWS Redshift.
- Processed some basic statistical analysis for data profiling, examining metrics such as cancellation rates, variance, skewness, kurtosis of trades, and stock performance over different time intervals like 1 minute, 5 minutes, and 15 minutes.
- Used PySpark and Pandas to compute the moving average and RSI score for the stocks, then stored the results in the data warehouse.
- I was part of integrating the Hadoop cluster with the Spark engine to handle both BATCH and GRAPHX operations.
- Performed data and feature engineering with Python Pandas to set the stage for predictive analytics.
- Developed and validated machine learning models including Ridge and Lasso regression for predicting total amount of trade.
- Boosted regression model performance by doing polynomial transformations and feature selection, then applied those methods to pick stocks.
- Generated report on predictive analytics using Python and Tableau including visualizing model performance and prediction results.
- Utilized Agile and Scrum methodology for team and project management.
- Used Git for version control with colleagues.

ENVIRONMENT:

Spark (PySpark, SparkSQL, Sparklib), Python 3.x(Scikit-learn, Numpy, Pandas), Tableau 10.1, GitHub, AWS EMR/EC2/S3/Redshift, and Pig.

AWS DATA ENGINEER

AMFAM, Madison, Wisconsin **Sep 2020 to April 2022**

RESPONSIBILITIES:

- Migrate data from on-premises to AWS storage buckets
- Developed a python script to transfer data from on-premises to AWS S3
- Developed a python script to hit REST API's and extract data to AWS S3
- Worked on Ingesting data by going through cleansing and transformations and leveraging AWS Lambda, AWS Glue and Step Functions
- Crafted YAML files for each data source, incorporating Glue table stack creation.
- Worked on a Python script to extract data from Netezza databases and shuttle it over to AWS S3
- Developed Lambda functions and set up IAM roles to execute Python scripts, triggered by SQS, EventBridge, and SNS.
- Created a Lambda Deployment function, and configured it to receive events from S3 buckets.
- Writing UNIX shell scripts to automate tasks and scheduled cron jobs using Crontab for job automation.
- Developed various Mappings with the collection of all Sources, Targets, and Transformations using Informatica Designer.
- Developed Mappings using Transformations like Expression, Filter, Joiner, and Lookups for better data messaging and to migrate clean and consistent data.
- Created Python Spark - AWS Glue ETL jobs for processing raw and aggregated data in Parquet format and push the output to S3.
- Leveraged AWS Lambda for serverless data operations, facilitating efficient data transformation tasks.

ENVIRONMENT:

Python 3.6, AWS (Glue, Lambda, Step Functions, SQS, Code Build, Code Pipeline, EventBridge, Athena), Unix/Linux Shell Scripting, PyCharm, Informatica PowerCenter, Code Build, Code Pipeline, EventBridge, Athena), Linux Shell Scripting, Informatica PowerCenter.

BIG DATA ENGINEER

Truist Bank, Charlotte, NC **May 2018 to Aug 2020**

RESPONSIBILITIES:

- Created and executed Hadoop Ecosystem installation and document configuration scripts on Google Cloud Platform.
- Transformed batch data from various sources like SQL Server, MySQL, PostgreSQL, and CSV files into data frames using PySpark.
- Researched and downloaded jars for Spark-avro programming.
- Developed a PySpark program that writes dataframes to HDFS as avro files.
- Utilized Spark's parallel processing capabilities to ingest data.
- Created and executed HQL scripts that creates external tables in a raw layer database in Hive.
- Developed a Script that copies avro formatted data from HDFS to External tables in raw layer.
- Created PySpark code that uses Spark SQL to generate dataframes from avro formatted raw layer and writes them to data service layer internal tables as orc format.

- In charge of PySpark code, creating dataframes from tables in data service layer and writing them to a Hive data warehouse.
- Installed Airflow and created a database in PostgreSQL to store metadata from Airflow.
- Configured documents which allow Airflow to communicate to its PostgreSQL database.
- Developed Airflow DAGs in python by importing the Airflow libraries.
- Utilized Airflow to schedule automatically trigger and execute data ingestion pipeline.

ENVIRONMENT:

Google Cloud Platform, Hadoop ecosystem, Hive for data warehousing, and Airflow for workflow orchestration, ensuring efficient data processing, storage, and management.

DATA INTEGRATION ENGINEER

IBing Software Solutions Private Limited, Hyderabad, India **Oct 2016 to Feb 2018**

RESPONSIBILITIES:

- Developed data ingestion pipelines using Talend ETL tool and bash scripting with big data technologies like Hive, Impala, Spark, Kafka, and Talend.
- Built scalable and secure data pipelines for large datasets.
- Gathered requirements for new data source ingestion, including lifecycle, data quality checks, transformations, and metadata enrichment.
- Supported data quality management by implementing proper checks in data pipelines.
- Provided data engineer services to Data Scientists, including data exploration and ad-hoc ingestions, using big data technologies.
- Created machine learning models using PySpark and MLlib to demonstrate Big data capabilities.
- Improved Data Ingestion Framework by enhancing data pipelines' robustness and security.
- Implemented data streaming using Kafka and Talend for multiple data sources.
- Worked with various storage formats (Avro, Parquet) and databases (Hive, Impala, Kudu).
- Managed S3 Data Lake, handling inbound and outbound data requests through the big data platform.
- Had working knowledge of cluster security components like Kerberos, Sentry, SSL/TLS.
- Developed agile and iterative data modeling patterns to provide flexibility.
- Implemented JILs for job automation in production clusters.
- Troubleshot user analysis issues (JIRA and IRIS Ticket).
- Collaborated with SCRUM team to deliver user stories on time for every Sprint.
- Analyzed and resolved production job failures in various scenarios.
- Implemented UNIX scripts to define workflow and process data files, automating jobs.

ENVIRONMENT:

Spark, Redshift, Python, HDFS, Hive, Pig, Sqoop, Scala, Kafka, Shell scripting, Linux, Jenkins, Eclipse, Git, Oozie, Talend, Agile Methodology.

DATA ENGINEER

Dhruvsoft Services Private Limited, Hyderabad, India **June 2015 to Sep 2016**

RESPONSIBILITIES:

- Migrated data from file system to Snowflake within the organization.
- Imported legacy data from SQL Server and Teradata into Amazon S3.
- Created consumption views to optimize complex query performance.
- Exported data into Snowflake using staging tables to load various file data from Amazon S3.
- Conducted leaf-level data comparison across databases during data transformation or loading to ensure data quality and integrity.
- Wrote SQL scripts for data mismatch analysis and historical data loading from Teradata SQL to Snowflake as part of data migration.
- Developed SQL scripts to handle sensitive data (e.g., National Provider Identifier Data) in Teradata, SQL Server Management Studio, and Snowflake databases.
- Used Spark commands to retrieve data from file system to S3.
- Managed S3 buckets, configured policies, and utilized S3 bucket and Glacier for storage and backup on AWS.
- Established Metric tables and End user views in Snowflake to supply data for Tableau refresh.
- Crafted Custom SQL queries to validate dependencies for daily, weekly, and monthly jobs.
- Utilized Nebula Metadata to register Business and Technical Datasets for respective SQL scripts.
- Proficient in Spark ecosystem, employing Spark SQL and Scala queries across various formats like text and CSV files.
- Developed Spark code and Spark-SQL/streaming for expedited data testing and processing.
- Actively engaged in scheduling daily, monthly jobs with Precondition/Postcondition based on specific requirements.
- Monitored daily, weekly, and monthly jobs, providing support in case of failures or issues.

ENVIRONMENT:

Snowflake, AWS S3, GitHub, Service Now, HP Service Manager, EMR, Nebula, Teradata, SQL Server, Apache Spark, Sqoop.

PROFESSIONAL SKILLS:

Data Modeling Tools : Erwin Data Modeler, ER Studio v17.
Programming Languages : PL/SQL UNIX,spar PYTHON, Spark, Scala, SHELL SCRIPTING, SQL, PySpark.
Methodologies : RAD, JAD, System Development Life Cycle (SDLC), Agile.
Cloud Platform : AWS, Azure, Google Cloud.
Databases : Oracle 12c/11g, Teradata R15/R14.
OLAP Tools : Tableau, SSAS, Business Objects, and Crystal Reports 9
ETL/Data warehouse Tools: Informatica 9.6/9.1, and Tableau.
Operating System : Windows, Unix, Sun Solaris.
Big Data Tools : Hadoop Ecosystem, Map Reduce.
Databases : Hive, MYSQL, NETEZZA, SQL Server, Redshift, Snowflake.
Cloud Computing Tools :AWS (EC2, S3, Lambda, Athena, Glue, Redshift), GCP, Snowflake.