**BHARADWAJ NARNE**

**Data Engineer**

**Contact: 945 244 8789**

**Email:**[**bharadwaj.narne.bn@gmail.com**](mailto:bharadwaj.narne.bn@gmail.com)

**Professional Summary:**

* Demonstrated proficiency as a Data Engineer with over 8 years of experience, delivering effective solutions across diverse industries such as pharmaceuticals, entertainment, finance, and consulting.
* Proficient in developing web services using Python, processing large datasets with Spark using Scala and PySpark, and leveraging Spark Core and Spark SQL libraries for querying.
* Skilled in developing Apache Spark jobs using Scala and Python, optimizing Hadoop algorithms, and writing production-level Machine Learning models for predictive analysis.
* Experienced in managing data from Kafka sources, configuring Spark Streaming for real-time data processing, and storing data in DynamoDB using Scala.
* Proficient in dimensional data modelling, star schema/snowflake modelling, FACT & Dimensions tables, and physical & logical data modelling using tools like ERWIN, Oracle Designer, and Data Integrator.
* Experienced in Text Analytics, developing Statistical Machine Learning, and Data Mining solutions to address various business challenges.
* Created python workflows using Apache Airflow for workflow management, including task scheduling, dependency management, and SLA monitoring.
* Experienced in Agile methodologies including extreme programming, SCRUM, and Test-Driven Development (TDD) for efficient project delivery.
* Extensive expertise in Big Data Analytics, proficient in various Apache Hadoop components including HDFS, MapReduce, HIVE, HBase, PIG, SQOOP, Spark, and Flume.
* Skilled in managing and optimizing Hadoop clusters, including HDFS, Job Tracker, Task Tracker, Name Node, Data Node, and MapReduce, with experience in Hadoop HBase updates and Oozie for workflow management.
* Expertise in ETL processes, utilizing tools like Informatica Power Center and Talend to load datasets into Hive, ensure Data Quality, and implement Data Warehousing solutions.
* Well-versed in translating business requirements into efficient ETL methods, leveraging Informatica Power Center for enhanced performance and maintainability.
* Skilled in halting AWS Lambda jobs, creating automated backups in S3 buckets and EBS, and developing Python-based transformations and analytics on large datasets in EMR clusters.
* Proficient in configuring data loads from S3 into Redshift using AWS Data Pipeline, utilizing Glue catalogue for metadata retrieval, and running SQL queries for data processing.
* Experienced in Microsoft Azure services like SQL Data Warehouse, Azure SQL Server, Azure Databricks, Azure Data Lake, Azure Blob Storage, and Azure Data Factory.
* Skilled in developing JSON scripts for deploying pipelines in Azure Data Factory, integrating with SQL Activity, and executing business analytical models.
* Experienced in managing Azure Data Lakes (ADLS) and Data Lake Analytics, with a solid understanding of integration with other Azure Services and knowledge of USQL.
* Proficient in using Azure Databricks and Azure Machine Learning for identifying trends, patterns, and inconsistencies in petabytes of data.
* Experienced in frameworks like Flask, Django, and various Python packages for enhanced functionality and visualization.
* Experienced in developing AWS Lambda functions in Python for AWS Lambda that invokes Python scripts to perform various transformations and analytics on large data sets in EMR clusters.

**Technical Skills:**

|  |  |
| --- | --- |
| **Languages** | Shell scripting, SQL, PL/SQL, Python, R, PySpark, Pig, Hive QL, Scala |
| **Hadoop Distribution** | Hadoop, Spark, MapReduce, Hive QL, HDFS, Sqoop, Pig Latin, Cloudera CDH,  Horton Works HDP, Apache, AWS |
| **Big Data Ecosystem** | HDFS, MapReduce, Hive, Pig, Sqoop, Flume, Oozie, Zookeeper, Kafka,  Cassandra, Apache Spark, Spark Streaming, HBase, Flume, Impala, NiFi |
| **Python Libraries** | Beautiful Soup, NumPy, SciPy, Matplotlib, Pandas, network, urllib2, MySQL dB |
| **Frameworks** | Django, Flask, Pyramid, web2py, Apache Airflow |
| **Databases** | Oracle, Redshift, SQL Server, MySQL, Cassandra, Teradata, PostgreSQL,  MS Access, Snowflake, NoSQL Database (HBase, MongoDB). |
| **Cloud Technologies** | AWS(EC2/S3/Redshift/EMR/Lambda/Snowflake), Microsoft Azure |
| **Agile CI/CID** | Confluence, JIRA, GITHUB, GITLAB, Docker, Kaban, BitBucket |
| **Operating Systems** | Windows, Unix, Linux, AIX, Zaloni, PagerDuty |
| **Statistical Methods** | Hypothesis Testing, ANOVA, Principal Component Analysis (PCA), Time Series,  Correlation (Chi-square test, covariance), Multivariate Analysis, Bayes Law. |
| **Machine Learning** | Linear Regression, Logistic Regression, Naive Bayes, Decision Trees, Random  Forest, Support Vector Machines (SVM), K-Means Clustering, K-Nearest  Neighbours (KNN), Random Forest, Gradient Boosting Trees, Ada Boosting,  PCA, LDA, Natural Language Processing |
| **Deep Learning** | Artificial Neural Networks (ANNs), Convolutional Neural Networks(CNNs),  RNN, Deep Learning on AWS, Keras API. |
| **Other tools & technologies** | TensorFlow, Keras, AWS ML, Azure ML studio, GCP, NLTK, SpaCy, Gensim,  MS Office Suite, Google Analytics, GitHub |

**Project Experience:**

**Client: Next Health Technologies, Denver, CO Aug 2022 – till date**

**Role: Senior Data Engineer | Machine Learning Engineer**

**Responsibilities:**

* Implemented end-to-end machine learning workflows, including data gathering from AWS Snowflake, preprocessing, feature extraction, modeling, evaluation, and deployment using Python.
* Utilized various Python packages such as NumPy, Pandas, Matplotlib, Seaborn, and stats models for exploratory data analysis and machine learning tasks.
* Trained Random Forest and convolutional neural network (CNN) models on customer web activity data to predict potential customers.
* Performed feature engineering, PCA, and hyperparameter tuning to enhance model accuracy.
* Applied a range of machine learning algorithms including linear regression, logistic regression, decision trees, random forests, K-means clustering, SVMs, and XGBoost based on client requirements.
* Developed machine learning models using recurrent neural networks (RNNs) such as LSTM for time series analysis and predictive analytics.
* Fine-tuned CNN models using Google TensorFlow Keras API for classification tasks and adjusted parameters for improved performance.
* Demonstrated proficiency in image classification using pre-trained models like VGG16, VGG19, ResNet, ResNetV2, and InceptionV3, along with knowledge of OpenCV for real-time computer vision tasks.
* Applied natural language processing techniques for documentation classification, text processing, and text summarization using NLTK, SpaCy, and TextBlob.
* Automated tasks such as consuming data subject requests from AWS Snowflake and cataloging data in Alation using Python scripts.
* Integrated with Adobe Analytics API for data retrieval and developed ETL jobs for data processing.
* Implemented stored procedures in AWS Snowflake to ensure compliance with CCPA regulations by anonymizing sensitive data.
* Utilized AWS Boto3 API for making HTTP calls to AWS services like S3, Secrets Manager, and SQS.
* Developed integrations to process consumer subscription information from HBO using AWS SQS and stored metadata in Postgres tables.
* Generated reports to provide consumer information to data subjects and developed automation jobs using Python.
* Implemented AWS Lambda functions for automating tasks such as posting privacy files to Malibu data privacy endpoints.
* Utilized technologies including Python, Postgres, AWS Snowflake, Alation Data Catalog, AWS services (EC2, S3, Lambda, Secrets Manager, SQS), Adobe Analytics, Linux, and various Python libraries (Scikit-learn, NumPy, Pandas, Matplotlib, Seaborn) in an Agile/SCRUM environment.
* Engaged in all stages of data processing, from acquisition to visualization, using Python and R.
* Utilized Python libraries like NumPy, Pandas, Matplotlib, and Stats for data manipulation, cleaning, and feature engineering.
* Extracted data from AWS Snowflake and ensured its quality and integrity using Pandas and NumPy.
* Addressed imbalanced datasets using sampling techniques and handled categorical variables with one-hot encoding.
* Implemented various machine learning models like logistic regression, KNN, and gradient boosting and experimented with ensemble methods to improve accuracy.
* Collaborated with business partners to identify improvement opportunities and developed actionable plans.
* Deployed models on AWS and maintained performance reports using Tableau.
* Exposure to Spark, Spark Streaming, Spark MLlib, snowflake, Scala and Creating Data Frames handled in Spark with Scala.

**Technology Stack**: Python, Postgres, AWS snowflake, Alation data catalog tool, snowsql, AWS EC2, S3, AWS lambda, AWS secrets manager, AWS SQS, Adobe analytics, Linux, Scikit-learn, SciPy, NumPy, Pandas, Matplotlib, Seaborn, JIRA, GitHub, Agile/ SCRUM.

**Client: Fidelity, Westlake, Texas Mar 2020 – Aug 2022**

**Role: Senior Data Engineer**

**Responsibilities:**

* Conducted end-to-end Architecture & implementation assessment of various AWS services like Amazon EMR, Redshift, S3. (Achieved a notable 20% reduction in data storage costs and enhanced data processing efficiency by 25%)
* Installed and configured Hive, HDFS, and NIFI for the implemented HDP Hadoop cluster. (Successfully implemented the cluster with a 95% uptime and enhanced data processing efficiency by 30%)
* Managed various HDFS file formats like Avro, Sequence File, Nifi, JSON, and various compression formats like Snappy, bzip2. (Improved data processing efficiency by 20% and reduced data storage costs by 15%)
* Employed AWS Glue catalog with a crawler to retrieve data from S3 and execute SQL query operations. (Decreased data processing time by 20% and improved data accuracy by 15%)
* Developed a Big data web application using Agile methodology in Scala, leveraging its fusion of functional and object-oriented programming. (Completed the project within 6 months with a team of 5 developers, resulting in a significant enhancement of 30% in data processing speed)
* Utilized SQL Server Integrations Services (SSIS) for extraction, transformation, and loading data into the target system from multiple sources. (Successfully processed over 10TB of data, leading to a marked 15% improvement in data accuracy)
* Engaged in data ingestion from different RDBMS sources into Hadoop using Sqoop. (Effectively ingested over 5 million records from multiple sources with an impressive 95% accuracy rate)
* Developed routines to import data from Oracle MAINODS to Elasticsearch. (Processed over 2 million records with a remarkable 90% improvement in data accuracy)
* Collaborated on AWS Data Pipeline to configure data loads from S3 to Redshift. (Successfully processed over 3TB of data with a 25% enhancement in data processing efficiency)
* Produced reports for the BI team utilizing SQOOP to export data into HDFS and Hive. (Trimmed manual processing time by 25% and enhanced data accuracy by 15%)
* Established Hive tables to load large datasets of structured data coming from WADL after the transformation of raw data. (Improved data processing efficiency by 20% and reduced data storage costs by 15%)
* Assisted in performance tuning and monitoring and participated in loading and transforming large sets of structured data from router location to EDW using a NIFI ETL pipeline flow. (Enhanced data processing speed by 25% and reduced manual processing time by 30%)
* Developed custom NIFI processors for parsing data from XML to JSON format and filtering broken files. (Enhanced data processing efficiency by 30% and reduced manual processing time by 25%)
* Worked on Configuring Zookeeper, Kafka, and Logstash cluster for data ingestion and Elasticsearch performance and optimization and focused on Kafka for live streaming of data. (Improved data processing efficiency by 25% and reduced data storage costs by 20%)
* Developed PySpark code and Spark-SQL for expedited testing and processing of data. (Boosted data processing speed by 40% and decreased manual processing time by 20%)
* Worked on Data serialization formats for converting complex objects into sequence bits by employing Parquet, ORC, AVRO, JSON, and CSV formats. (Improved data processing efficiency by 25% and reduced data storage costs by 20%)
* Crafted Map Reduce jobs using Java API and Pig Latin.

**Tech Stack:** AWS Services, Hadoop (HDFS, Map Reduce), Kafka, Python, Flask, Django, PySpark, Beautiful Soup, Scala, Yarn, IAM, PostgreSQL, Spark, Impala, Hive, Mongo DB, Pig, DevOps, HBase, Oozie, Hue, Sqoop, Flume, Oracle, NIFI, Git, AWS Services (Lambda, EMR, Auto scaling), Airflow.

**Company: SafeWay, Pleasanton CA Jun 2019 –Mar 2020**

**Role: Big Data Engineer**

**Responsibilities:**

* Designed and setup Enterprise Data Lake to support Analytics, processing, storing, and reporting of voluminous, rapidly changing data, utilizing AWS EMR, S3, Glue, and Redshift.
* Orchestrated data ingestion pipelines using AWS Glue to automate ETL processes into the data lake, ensuring data accuracy and reliability.
* Implemented data partitioning strategies in AWS S3 to optimize data retrieval and processing efficiency, reducing query latency, and improving performance.
* Developed custom data processing algorithms in Apache Spark to handle large-scale transformations and analytics tasks, leveraging distributed computing capabilities.
* Integrated Apache Kafka with AWS services for real-time data streaming and event-driven architectures, enabling timely insights and decision-making processes.
* Designed and implemented data encryption mechanisms using AWS KMS to secure sensitive data at rest and in transit, ensuring compliance with regulatory requirements.
* Established data governance policies and procedures to govern data access, usage, and lifecycle management, ensuring integrity and compliance.
* Collaborated with data scientists to deploy machine learning models on Amazon SageMaker for predictive analytics and pattern recognition.
* Utilized AWS Step Functions to automate complex workflows and orchestrate multi-stage data processing tasks, increasing operational efficiency.
* Conducted performance tuning and optimization of AWS EMR clusters to improve resource utilization and job execution times, maximizing cost-effectiveness.
* Implemented disaster recovery strategies and backup mechanisms for the data lake infrastructure using AWS Backup and S3 Versioning.
* Coded Teradata BTEQ scripts for loading, transforming data, and fixing defects, enhancing data quality and efficiency.
* Developed reusable frameworks to automate ETL from RDBMS systems to the Data Lake, utilizing Spark Data Sources and Hive data objects.
* Conducted data blending and preparation using Alteryx and SQL for Tableau consumption, enabling effective data visualization and analysis.
* Developed Kibana Dashboards based on Logstash data and Integrated different source and target systems into Elasticsearch for near real-time log analysis.
* Integrated Apache Airflow with AWS to monitor multi-stage ML workflows with tasks running on Amazon SageMaker.
* Design and maintain databases using Python and developed Python based API (RESTful Web Service) using Flask, SQL Alchemy, and PostgreSQL
* Generated Python Django forms to record data of online users and used PyTest for writing test cases.
* Worked on AWS Lambda functions in python for AWS Lambda which invokes python scripts to perform various transformations and analytics on large data sets in EMR clusters.
* Implemented Data Quality in ETL Tool Talend and having good knowledge in Data Warehousing
* Used AWS-CLI to suspend an AWS Lambda function. Used AWS CLI to automate backups of ephemeral data-stores to S3 buckets, EBS.

**Tech Stack**: AWS EMR, S3, Glue, KMS, Apache Spark, Apache Kafka, AWS Key Management Service, Amazon SageMaker, AWS Step Functions, AWS Backup, Python, Data Governance, Data Encryption, Disaster Recovery, Alteryx, Teradata, Tableau, SQL.

**Client: Greet Labs, India Jul 2015 – Dec 2018**

**Role: Data Engineer |SQL Developer**

**Responsibilities:**

* Engaged in utilizing Python OpenStack APIs for updating database content and file manipulation tasks.
* Implemented AWS services for scaling Tableau server, ensuring its security through Amazon VPC, security groups, AWS IAM, and Direct Connect.
* Configured EC2 instances, IAM users, and roles, and orchestrated S3 data pipes via Boto API for seamless data loading.
* Developed an automated mechanism for migrating proprietary binary format data files to HDFS utilizing the Ingestion service.
* Utilized Python OpenStack APIs and various Python libraries such as wxPython, NumPy, and Matplotlib.
* Executed data transformations within HIVE, employing partitions and buckets for enhanced performance.
* Conducted data ingestion into Hadoop using Sqoop, applying transformations with Pig and HIVE, alongside Python and Django for graphic creation, XML processing, and business logic implementation.
* Utilized Git, GitHub, and Amazon EC2 for deployment purposes, leveraging Heroku for deployment, and NumPy and SciPy for mathematical operations.
* Developed server-based web traffic analysis tools using RESTful APIs and Flask, integrated with Pandas for statistical analysis.
* Leveraged the Pandas API for time series and tabular data manipulation and retrieval, while incorporating Amazon AWS S3 and RDS for hosting static/media files and databases.
* Participated in the design, construction, and deployment of NoSQL solutions like MongoDB and executed MYSQL queries from Python using Python-MySQL connector and MySQL dB package.
* Developed Python scripts for vulnerability assessments, SQL injection prevention, permission checks, performance analysis, and database migration to PostgreSQL.
* Participated in SOAP-based web service development for data exchange and interfacing with external interfaces in XML format.
* Troubleshooted and deployed bug fixes for Python applications serving both external customers and internal teams.
* Executed complex SQL queries to extract data from SQL Server and Oracle databases and evaluated Information Management System Database for data quality improvement.
* Implemented Data Governance policies in the Students Information Management Database and conducted Data Analysis and Visualization on survey data using Tableau Desktop and Python libraries.
* Developed a machine learning model for recommending friends to students based on similarities and utilized Alteryx for Data Preparation.
* Analysed university research budgets and recommended data standardization for ensuring data integrity.
* Reviewed SQL queries in Tableau Desktop, conducted statistical analysis using Python and R, and created dashboards for Global Services & Technical Services using SSRS, Oracle BI, and Excel.
* Cleaned and documented user satisfaction survey data and developed data gathering applications using C#.Net.

**Tech Stack:** Python, Hive, Oozie, Amazon AWS S3, MySQL, HTML, Python 2.7, Django, HTML5, CSS, XML, MySQL, MS SQL Server, GIT, Jenkins, JIRA, MySQL, Cassandra, Pig, Hadoop, AWS Cloud Watch, AWS Redshift, SQL, SOAP, Rest APIs, AWS EC2, XML, JavaScript, AWS, Linux, Shell Scripting, AJAX, Mongo dB

**EDUCATION**

* **Master of Science, Business Analytics (Data Science and Machine Learning), GPA 3.6**

The University of Texas at Dallas

Relevant Courses: Applied Machine Learning, Modelling for Business Analytics, Time Series and Econometric Analysis, Data Visualization, Database foundations for Data Science, Big Data, Machine Learning in Google Cloud, Advance Statistics, Predictive Analytics and Prescriptive Analytics.

* Bachelor of Engineering, BITS Pilani, India