BHARATH SAI
Saibharath.p25@gmail.com
+1 (224) 836-1254
Sr. Data Engineer
https://www.linkedin.com/in/bharath-sai-46068926a/

PROFESSIONAL SUMMARY:

- Experienced **Data Engineer** with over 8 years of IT experience in various technologies including Data Engineering, ETL tools, Machine Learning, Data Extraction, Data Modeling, Statistical Modeling, Data Mining, and Data Visualization.
- Skilled in implementing and exploring **ML techniques**, utilizing hands-on experience in **ETL tools** and Machine Learning.
- Proficient in developing **REST APIs** for data collection from diverse data feeds to build robust **ETL mappings**.
- Extensive expertise in **Azure Cloud Services (PaaS & IaaS), Azure Synapse Analytics, SQL Azure, Data Factory, Azure Analysis Services, Application Insights, Azure Monitoring, Key Vault, and Azure Data Lake**.
- Proven ability to design cloud-based solutions in **Azure**, creating **Azure SQL databases**, setting up Elastic pool jobs, and designing tabular models in Azure Analysis Services.
- Performed **data validation, data integrity** and database testing using **SQL Queries** with multiple Databases **Oracle, SQL Server, and MySQL.**
- Hands-on experience in creating pipeline jobs, schedule triggers, and developing data processing solutions using **Azure Data Factory**.
- Developed **Scala applications** for high-volume and real-time data processing in **Hadoop and Spark SQL** environments.
- Proficient in Talend for **Data warehousing projects**, designing mappings to populate data into dimensions and fact tables.
- Experienced in working with **Snowflake** cloud data warehouse and **AWS S3** bucket for integrating data from multiple source systems.
- Proficient in **Amazon Web Services (AWS)** using **EC2** for computing and **S3** as a storage mechanism.
- Capable of using **AWS** utilities such as **EMR, S3**, and **Cloud Watch** to run and monitor **Hadoop** and **Spark** jobs on **AWS**.
- Used **snowflake** cloud data ware house database extensively to build **ETL pipelines**.
- Skilled in designing and implementing cost-effective and efficient **ETL Architecture**, providing **ETL** solutions for diverse business models.
- Proficient in batch processing solutions using **Azure Data Factory** and **Azure Databricks**, implementing clusters, notebooks, jobs, and auto-scaling.
- Knowledgeable in **Data Cleaning, Data Validation, Data Mapping, Data Analysis, Data Profiling**, feature scaling, feature engineering, statistical modeling, testing, validation, and data visualization.
- Expert in developing **visualization dashboards** using calculations, Filters, Charts, parameters, calculated fields, groups, sets and hierarchies.
- Hands-on experience with **Snowflake data warehouse**, proficient in **Machine Learning algorithms**, and Predictive Modeling including **Regression models, Decision Tree, Random Forests, Sentiment Analysis, Naive Bayes Classifier, SVM,** and **Ensemble Models**.
- Knowledgeable in **Natural Language Processing (NLP)** algorithms and Text Mining.
- Proficient in multiple programming languages such as **Java, Python,** and **R.**

EDUCATION DETAILS:

Bachelor's Degree in Computer Science.

Jawaharlal Nehru Technological University, Hyderabad (2010 - 2014)

TECHNICAL SKILLS:

Programming Languages	Python, PySpark, Spark, Scala, SQL, PySpark, C, C++
Hadoop Eco System	Hadoop, MapReduce, Spark, HDFS, Sqoop, YARN, Oozie, Hive, Apache Flume, Impala, Apache Storm, Apache Airflow, HBase
Databases	MySQL, SQL Server, Oracle 12c, MS Access
NoSQL Data Bases	MongoDB, Cassandra, HBase, KairosDB
Workflow Management tools	Oozie, Apache Airflow
Visualization & ETL tools	Tableau, BananaUI, D3.js, Informatica, Talend
Cloud Technologies	Azure, AWS
IDE's	Eclipse, Jupyter Notebook, Spyder, PyCharm, IntelliJ
Version Control Systems	Git, SVN

PROFESSIONAL EXPERIENCE:

Abbvie – Vernon Hills, Illinois | Data Engineer | Aug 2021 – Till Date

Responsibilities:

- ✓ Designed the approach for collecting business requirements, aligning it with the project scope and **SDLC methodology**.
- ✓ Installed, configured, and maintained **Data Pipelines**, developing them with **Kafka** and **Spark**.
- ✓ Created, scheduled, and monitored **Azure Data Factory pipelines** and **Spark jobs** on **Azure SQL**.
- ✓ Translated business problems into **Big Data solutions** and defined the **Big Data strategy** and **Roadmap**.
- ✓ Authored **Python (PySpark)** scripts for custom **UDFs**, facilitating row/column manipulations, merges, aggregations, stacking, data labeling, and cleaning tasks.
- ✓ Evaluated **Snowflake Design** considerations for potential changes in the application, building the Logical and Physical data model for **Snowflake** accordingly.
- ✓ Expertly developed **Json** Scripts for deploying pipelines in **Azure Data Factory** to process data effectively.
- ✓ Utilized **Azure Data Factory** for computing and handling large volumes of data.
- ✓ Designed and implemented database solutions in **Azure SQL Data Warehouse** and **Azure SQL**.
- ✓ Created pipelines in **Azure Data Factory** using Linked Services, Datasets, and other pipeline components.
- ✓ Worked on extracting data using **Azure SQL**, transformed it using **Python** libraries, and stored it in **Blob storage** and **Azure SQL Data Warehouse**.
- ✓ Utilized **Pig Scripts** to generate **MapReduce** jobs and performed **ETL** procedures on the data in **HDFS**.
- ✓ Developed solutions to leverage **ETL tools**, identifying opportunities for process improvements using **Informatica** and **Python**.
- ✓ Built and created data models in **Power BI**, dealing with data from various sources like databases, spreadsheets, and APIs. This included creating calculated columns, metrics, developing links across tables, and refining data structures for effective reporting.

- ✓ Used **Terraform** to set up and configure a cluster of machines in the cloud, specifically for running distributed data processing frame works like **Apache Hadoop** or **Apache Spark**.
- ✓ Employed **Terraform** to create and manage **databases, load balancers**, and other infrastructure components required for the **data pipeline**.
- ✓ Performed advanced procedures, such as test analysis and processing, utilizing the in-memory computing capabilities of **Spark** with **Scala**.
- ✓ Utilized **Spark** Streaming to receive real-time data from **Kafka** and stored the streaming data to using **Python**, **NoSQL** databases such as **HBase** and **Cassandra**.
- ✓ Used **Spark** for interactive queries, processing streaming data, and integrating with popular **NoSQL** databases to handle large volumes of data.
- ✓ Ensured seamless integration of **Power BI** with various data sources, regularly updating data through Power Query and other **ETL** procedures as necessary.
- ✓ Leveraged **SparkContext, Spark-SQL, Spark MLlib, Data Frame, Pair RDD**, and **Spark YARN** for effective data processing and analysis.
- ✓ Utilized **Spark Streaming APIs** to perform transformations and actions on the fly, enabling the creation of common data processing tasks.
- ✓ Developed a **Kafka** consumer **API** in Scala for efficiently consuming data from **Kafka** topics.
- ✓ Gained experience in writing live Real-time Processing and core jobs using Spark Streaming with **Kafka** as a data pipeline system, including successfully migrating an existing on-premises application to **AWS**.

Environment: Azure, ETL, MapReduce, Cloudera, Snowflake, Kafka, Spark, Azure data factory, Hadoop, Hbase, Tableau, Informatica, Python, Hive, PL/SQL, Oracle, UNIX, Shell Scripting.

Citrix – Fort Louderdale, Florida | Data Engineer | Jan 2020 – July 2021

Responsibilities:

- ✓ Exported data into **Snowflake** by creating staging tables to load data from various files from **Amazon S3.**
- ✓ Developed processes for loading data into **Snowflake** and designed **data modeling** for efficient reporting using **Tableau**.
- ✓ Utilized **Tableau** to create visually stunning and interactive dashboards and reports.
- ✓ Evaluated **Snowflake Design** considerations for any application changes and built the Logical and Physical data model accordingly.
- ✓ Redesigned Views in **Snowflake** to improve performance and conducted unit testing to validate data integrity.
- ✓ Developed solutions using **ETL too**l using **SSIS** and **Python** to identify process improvements and streamline data workflows.
- ✓ Scheduled jobs using **Airflow** scripts with **Python**, adding different tasks to **DAG** and **Lambda functions**.
- ✓ Designed and implemented an incremental job to read data from **DB2** and load it into **Hive tables**, connecting to **Tableau** for interactive reporting.
- ✓ Developed Spark applications using **PySpark** and **Spark-SQL** for data extraction, transformation, and aggregation from various file formats.
- ✓ Worked in the Production support team, maintaining **mappings, sessions,** and **workflows** for loading data into the **Data Warehouse**.
- ✓ Developed and implemented **ETL pipelines** on **S3 parquet files** in a data lake using **AWS Glue**.
- ✓ Created a cloud formation template in **JSON** format to enable content delivery with cross-region replication using **Amazon Virtual Private Cloud**.
- ✓ **Built S3 buckets** and managed policies for **S3 buckets**, utilizing **S3** and **Glacier** for storage and backup on **AWS**.

- ✓ Implemented **AWS** solutions using **EC2, S3, RDS,** and **EBS**, and used **IAM** to create new accounts, roles, and groups.
- ✓ Leveraged **AWS** services like **EC2** and **S3** for processing and storing small datasets, maintaining the **Hadoop cluster** on **AWS EMR**.
- ✓ Utilized **AWS EMR** for transforming and moving large amounts of data into and out of other **AWS** data stores and databases like Amazon Simple Storage Service and **Amazon DynamoDB**.
- ✓ Worked on Dimensional and Relational Data Modeling using **Star Schema** and **Snowflake Schemas** for **OLTP/OLAP** databases.
- ✓ Developed Automation Regressing Scripts using **Python** for validating **ETL** processes between multiple databases, including **AWS** and **SQL Server**.
- ✓ Built **ETL pipelines** for data ingestion, transformation, and validation on the **AWS cloud** service.
- ✓ Conducted sessions with **Subject Matter Experts (SME)**, stakeholders, and other management teams to finalize the User Requirement Documentation for the project.

Environment: AWS, EC2, S3, Lambda, ETL, PySpark, Snowflake, Airflow, Kafka, Spark, Hadoop, Tableau, Python, Hive, SQL, Oracle, scheduling tool, Shell scripting.

Wellcare – Tampa, Florida | Data Engineer | Jan 2018 – Dec 2019

Responsibilities:

- ✓ Developed an **ETL Pipeline** using **Spark** and **Hive** to ingest data from multiple sources, ensuring seamless data integration and transformation.
- ✓ Designed and developed **ETL** jobs to extract data from the Salesforce replica and load it into the data mart in **AWS**.
- ✓ Created an **ETL Pipeline** using **SSIS/ETL** framework from scratch, ensuring efficient data extraction, transformation, and loading processes.
- ✓ Took charge of designing logical and physical data modeling for data sources on Confidential **AWS** ensuring data integrity and efficient querying.
- ✓ Utilized **Power BI** to design multiple scorecards and dashboards, presenting relevant information to different departments and upper-level management.
- ✓ Designed data connections and extracted data from various sources like S**QL, MySQL, Excel, SharePoint**, and **Snowflake** for analysis in **Tableau**.
- ✓ Conducted **data analysis** and performed statistical calculations to identify trends, patterns, and correlations in data using tools such as **Excel** and **Power BI**.
- ✓ Extensively used for **data modeling**, creating staging and target models for the Enterprise **Data Warehouse**.
- ✓ Performed logical and physical data modeling, including reverse engineering, using the **Erwin Data Modeling tool**.
- ✓ Resolved data type inconsistencies between the source systems and the target system using mapping documents and **SQL queries**.
- ✓ Participated in performance tuning efforts, optimizing stored procedures, views, **triggers, cursors, pivot, unpivot functions,** and **CTEs.**
- ✓ Created reports using **SQL Reporting Services (SSRS)** to cater to customized and **ad-hoc queries**, enabling easy access to essential information.
- ✓ Developed stored procedures in **MS SQL** to fetch data from different servers using **FTP** and processed these files to update the tables.
- ✓ Worked extensively on **MS SQL Server**, including **SSRS, SSIS**, and **T-SQL**, ensuring efficient data handling and processing.

- ✓ Utilized **SAP - SD Module** for handling customers of the client and generating sales reports, integrating **SAP data** into the overall data ecosystem.
- ✓ Conducted **ETL testing** and used **SSIS Tester** automated tool for unit and integration testing, ensuring the quality and reliability of the **ETL** processes throughout the project.

Environment: AWS, Tableau 7, Python 2.6.8, Numpy, Pandas, Matplotlib, Scikit-Learn, MongoDB, Oracle 10g, SQL

Qualcomm Technologies Inc – Hyderabad, India | Data Analyst | Jan 2016 – Oct 2017

Responsibilities:

- ✓ Used **Python 3.X** (**numpy, scipy, pandas, scikit-learn, seaborn**) and **Spark 2.0** (**PySpark, MLlib**) to develop a variety of models and algorithms for analytic purposes.
- ✓ Developed and implemented predictive models using **machine learning algorithms** such as **linear regression, classification, multivariate regression, Naïve Bayes, random forests, K-means clustering, KNN, PCA**, and regularization for data analysis.
- ✓ Built regression models including **Lasso, Ridge, SVR,** and **XGBoost** to predict Customer Lifetime Value.
- ✓ Built classification models including Logistic **Regression, SVM, Decision Tree**, and **Random Forest** to predict Customer Churn Rate.
- ✓ Designed and developed interactive dashboards and reports using **Power BI** to enable data-driven decision-making for business stakeholders.
- ✓ Utilized **Power BI** to create various analytical dashboards that help business users quickly analyze the data.
- ✓ Performed univariate and multivariate analysis on data to identify underlying patterns and associations between variables.
- ✓ Applied clustering algorithms such as **hierarchical** and **K-means** using **Scikit-learn** and **Scipy**.
- ✓ Used evaluation metrics such as **F-Score, AUC/ROC, Confusion Matrix, MAE,** and **RMSE** to assess different model performances.
- ✓ Performed data imputation using **Scikit-learn** package in **Python**.
- ✓ Implemented **NLP techniques** to optimize Customer Satisfaction.
- ✓ Worked with data engineers and operation teams to implement the **ETL process**, wrote and optimized S**QL queries** for data extraction to meet analytical requirements.

Environment: Python 2.x, NLP, R, Machine Learning (Regressions, KNN, SVM, Decision Tree, Random Forest, XGboost, LightGBM, Collaborative filtering, Ensemble), Pandas, Numpy.

Allegis Group – Hyderabad, India | Python Developer | June 2014 – Dec 2015

Responsibilities:

- ✓ Worked on the project from gathering requirements to developing the entire application.
- ✓ Worked on the **Anaconda Python** environment, including creating, activating, and programming in the **Anaconda** environment.
- ✓ Wrote programs for performance calculations using **NumPy**.
- ✓ Developed different statistical machine learning and data mining solutions to various business problems using **R, Python,** and **Tableau.**
- ✓ Analyzed the code thoroughly and reduced code redundancy to an optimal level.
- ✓ Worked on the development of **SQL** and stored procedures in **MySQL**, executing various **MySQL** database queries using **Python MySQL** connector and **MySQL Db package**.

- ✓ Responsible for designing, developing, testing, deploying, and maintaining the web application.
- ✓ Developed **Python** routines to log into websites and fetch data for selected options.
- ✓ Worked on reading and writing data from **CSV** and Excel file formats.

Environment: Python 2.x, Anaconda, Sypder (IDE), Tableau, python libraries such as NumPy, SQL Alchemy, MySQLdb.