

Venkata Subba Rao Y

**Summary:**

* Total 11+ years of Experience in Data Engineering/Data Architect, Python, Spark, Scala, PySpark, Java, Data Science, cloud-based enterprise-wide applications including data streaming, optimal data delivery, Data Analytics that includes Deep Learning/ Machine Learning, Data mining with large datasets of Structured and unstructured data, Data Acquisition, Data Validation, Predictive modelling, Data Visualization. Extensive experience in Data Analytics, developing different Statistical Machine Learning models, Data Mining solutions to various business problems and generating data Visualizations using Python.
* Experience working with large data sets and pipelines using tools and libraries such as Spark, PySPark, Spark SQL, Scala, HDFS, YARN, Hive and working knowledge of distributed/cluster computing; multi-threaded applications; Concurrency, Parallelism, Locking Strategies and Merging Datasets. Memory Management, Garbage Collection & Performance Tuning.
* Experience on Full stack development and Microservices Architecture, implementing distributed applications in a container environment
* Data Architect & implement medium to large scale BI solutions on Azure using Azure Data Platform services (Azure Data Lake, Data Factory, Data Lake Analytics, Stream Analytics, Azure SQL DW, Databricks).
* (Docker/Kubernetes) along with considerable Software engineering hands on experience designing Enterprise Cloud Applications with Go Lang, Python, Java, API development experience (APIs like REST / RESTful APIs).
* Experience with Hadoop distributed frameworks while handling large amount of big data using Spark and Hadoop Ecosystems.
* Experience in writing the HTTP RESTful Web services and SOAP API's in Golang. Designed configurable REST APIs on demand using Go. Worked on numerous go frameworks like gin, revel, beego.Developed new RESTful API services that work as a middleware between our application and third-party APIs that we will used using Golang.
* Produced and Consumed Restful Web Services using Postman Rest Client, Swagger UI, to interact with different applications from third-party tools. Using GO, developed a micro service for reading large volume of data(millions) from PostgreSQL database. Experience writing data APIs and multi-server applications to meet product needs using Golang.
* Hands on experience of Data Science libraries in Python such as Pandas, NumPy, SciPy, scikit-learn, Matplotlib, Seaborn, BeautifulSoup, Orange, Rpy2, LibSVM, neurolab, NLTK. Developed MapReduce/Spark Python modules for machine learning & predictive analytics in Hadoop
* Demonstrated experience in design and implementation of Statistical models, Predictive models, enterprise data model, metadata solution and data lifecycle management in both databases, Big Data environments. Expertise in building statistical models using algorithms like Regression, Random Forest, Decision Trees, Market Basket Analysis, K-Means and CART. Used PyTorch to train a neural network on a large fashion dataset. Used Python to build out data pipelines.

**Skills:**

|  |  |
| --- | --- |
| Languages | **GoLang, Python, R, Java, Scala** |
| ML Algorithms & Methods | **Linear regression, Logistic regression, Decision tree, RandomForest, K nearest neighbor, K mean, NLP, ARIMA, APRIORI, ETS, SVM, AdaBoost, Deep Learning,** |
| Packages | **Matplotlib, Seaborn, BeautifulSoup, Orange, Rpy2, LibSVM, neurolab, Pandas, NumPy, SciPy, Scikit-learn, Stats models, NLTK, Plotly, Matplotlib, Seaborn, plyR, dplyR, data.table and sqldf, tidyR, Hibernate 4.3, Log4j, Spring Boot, Spring 3.x/4.0, Spring Data, JEE Design Patterns, JAX-WS, JAX-RS** |
| BI Tools | **Tableau, Power BI from Microsoft, Amazon Redshift** |
| Cloud Computing & Devops | **EC2, VPC, RDS, S3, IAM, and Redshift, Redis, EMR, SNS, SQS, Cloudwatch, Step Functions,Google Cloud Platform, JSON, Docker, Jenkins, Kubernetes, GoCD,), BigQuery, Cloud Functions, Google Compute Engine, Azure Databricks, Azure MLFlow (Model Serving, Model Enabling), Azure Data Lake Storage, Azure SQL Server, Azure Blog Storage.** |
| BigData, Data Process & Modeling tools | **Apache Spark, Apache Kafka, Hadoop, Amazon SQS, SNS, Scala-Spark/Hive/Hadoop CLI** |
| Methodologies | **AGILE/SCRUM** |

**Certification:**

* Microsoft Certified AZ-900 Azure Fundamentals.
* Microsoft Certified AZ-104 Azure Administrator.

**Employment History & Work Experience:**

**eNcloud Services LLC** February 2023 to Till Date

**Data Engineer**

Katy, TX

**Roles & Responsibilities:**

* Analysis of Technical and Functional requirements
* Load and transform data into S3 from different sources (SQL Server, semi structure &unstructured data) using AWS DMS.
* Developed AWS EMR ETL jobs (PySpark) to make analytical data and store it to Redshift EDW / S3.
* Prepared AWS Lambda functions Service for event driven operations in pipe line.
* Prepared execution plans using AWS step functions.
* Created queries in Athena that helped business analysts spot emerging trends by comparing fresh data with EDW reference tables and historical metrics.
* Involved in performance tuning of the SQL’s.
* Coordination with the client and managing day - to - day activities.

**Environment:** Tableau Desktop, Tableau Server, VS Code, EMR Notebooks, AWS Services (Glue, Athena, S3, Lake Formation, Cloud Formation, Step Functions, EMR, EC2, IAM), Bit Bucket, Source Tree, Apache Spark, Terraform, PySpark, Python.

**AT& T** September 2021 to February 2023

**Data Engineer**

Hyderabad

**Roles & Responsibilities:**

* Design, development, and implementation of ETL pipelines using Python API (PySpark) of Apache Spark on AWS EMR.
* Implemented CI/CD process on Apache Airflow and DAGs by building Airflow Docker images by Docker-Compose and deploying on AWS ECS Cluster
* Worked on Airflow performance tuning of the DAG’s and task instance
* Worked on Airflow schedular (celery) and worker setting in airflow.cfg file
* Created Hooks and custom operator, operator will sense trigger files in S3 and start the data pipeline process.
* Automated resulting scripts and workflow using Apache Airflow and shell scripting to ensure cron jobs in productions.
* Used Airflow for orchestration and scheduling of the ingestion scripts
* Implemented Multiple Data Pipeline DAG’s and Maintenance DAGs in Airflow Orchestration.
* Designed and Developed Real time Stream processing Application using Spark, Kafka, Scala, and Hive to perform Streaming ETL and apply Machine Learning Model Serving
* Configuring and setting up Airflow dags as per the requirement to run our pyspark commands in airflow parallel and sequential.
* Lead a team of consultants to develop data engineering applications
* Developed and deployed a machine learning model for a service clients, resulting in a 25% reduction in fraudulent transactions and saving the client.
* Creating job flow using Airflow in python language and automating the jobs. Airflow will have separate stack for deploying DAG’s on and will run jobs on EMR on EC2 Cluster.
* Create, develop and test environments of different applications by provisioning kubernetes clusters on AWS using Docker, and Terraform
* Integration of data storage solutions in spark – especially with AWS S3 object storage.
* Experience with AWS cloud services: EMR, S3, Athena and Glue, Step Functions, AWS CDK,
* Worked on developing ETL pipelines on S3 parquet files on data lake using AWS Glue
* Worked on ETL Migration services by developing and deploying AWS Lambda functions for generating a serverless data pipeline that can be written to Glue Catalog and can be queried from Athena.
* Created dashboards for Churn Analytics using Tableau Desktop and deployed in Tableau Server
* Initiating alarms in CloudWatch service for monitoring the server's performance, CPU Utilization, disk usage etc. to take recommended actions for better performance.
* Design and Develop ETL processes in AWS Glue to migrate data from external sources like S3, parquet/text files into AWS Redshit
* Write terraform scripts from scratch for buiding Dev, staging production environment
* Data Extraction, aggregations and consolidation of data within AWS Glue using PySpark
* Create external tables with partitions using Hive, AWS Athena and Redshit.
* Worked on JIRA for defect/issues logging & tracking and documented all my work using Confluence
* Implemented complex, strongly-typed Spark workloads in Azure Databricks along with dependency management and Git integration
* Transformed PySpark using AWS Glue dynamic frames; cataloged the transformed the data using crawlers and scheduled the job and crawler using workflow features.
* Created scripts in Python (Boto) which integrated with Amazon API to control instance operations.
* Responsible for Design and maintenance of databases using Python. Developed Python based APIs (RESTful Web Services) by using Flask, SQL Alchemy and Postgres SQL.

**Environment:** Tableau Desktop, Tableau Server, VS Code, EMR Notebooks, AWS Services (Glue, Athena, S3, Lake Formation, Cloud Formation, Step Functions, EMR, EC2, IAM), Bit Bucket, Source Tree, Apache Spark, Terraform, PySpark, Python.

**ICICI BANK** December 2016 to September 2021

**Data Engineer**

Hyderabad

**Roles & Responsibilities:**

* Worked on automated fraud detection solutions (rules, batches, process flows) and manually handle policies to target issues.
* Designed AWS glue catalog with crawler to get the data from S3 and perform sql query operations
* Designed and developed automation scripts and batch jobs to create data pipelines between multiple data sources, pySpark based analytics platform (Databricks),
* Developed fraud data infrastructure for managing fraud risk, verify customer’s activity and provide frictionless customer experience.
* Improved data mining processes, resulting in a 20% decrease in time needed to infer insights from customer data used to develop marketing strategies. Hands on experience in working with the large volume of data with more than 10M customer records with multiple features.
* Incorporated Exploratory Data Analysis to identify the correlation between variables, multicollinearity, and hidden patterns, trends and seasonality using NumPy and Pandas libraries to perform data analysis
* Worked with product teams, implementing fraud monitoring tools and rule-based recommendations.
* Used predictive analytics such as machine learning and data mining techniques to forecast company sales of new products with a 95% accuracy rate.
* Created json files using lambda functions and AWS Glue catalog using Python and PySpark
* Created monitors, notifications and logs for lambda functions and AWS Glue jobs.
* Used clustering algorithms like k-means and hierarchical clustering to cluster the data and Devised recommendation models using different machine learning algorithms and validated the models. Used an ensemble of models to improve the prediction accuracy up to 75%
* Developed Natural Language Processing to automate the classification of customer incident queries into levels of classes to improve the customer services using databricks.
* Managed and build day-to-day analytics processes including activities such as data identification & collection, analytics design, marketing mix model development(MMM), implementation & maintenance, interpretation of findings, and coordination with stakeholders
* Used Scikit-learn in modeling various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means.
* Developed application to handle concurrent data collection and processing with Goroutines and channels using Golang.
* Used messaging format protobuf for high performance implemented different Golang libraries like go-fmt, go-ve, go-dep and also worked on race conditions which are resulted by implementing the race flag.
* Implementing the routing technique in Golang using Gorilla mux as request router and dispatcher for matching incoming requests to their respective handler.
* Created RPC-based service for reading large volume of data for PostgreSQL using Go
* Created Tableau scorecards, dashboards using stack bars, bar graphs, scattered plots, geographical maps, Gantt charts using show me functionality and publishing them to tableau server and tableau Online. Generated multiple reports using Microsoft PowerBi and SSRA and
* Work in close collaboration with marketing effectiveness team members and other internal partners to leverage expertise in data management, advanced marketing mix modeling(MMM), scoring etc., in order to develop key insights
* Implemented a machine learning model for customer sentiment pattern to better assess the heartbeat of the customer trend.
* Conducting studies, rapid plots and using advanced data mining and statistical modeling techniques to build a solution that optimizes the quality and performance of data.
* Experience in providing highly available and fault tolerant applications utilizing orchestration technologies like Kubernetes and Apache Mesos on Google Cloud Platform.Created test harness to enable comprehensive testing utilizing Python.

**Environment:** Python, NumPy, Pandas, Scikit-Learn, Pandas, Apache Spark, Py Spark, Scala, Kafka, GO, Goroutine, Microservices, Channels, cli, Gorilla Mux, AWS, EC2, S3, ELB, Autoscaling, Lambda, Kinesis, Elastic Filesystem, Databricks, RDS, DMS, VPC, Docker, Google Kubernetes Engine (GKE), GoLang, Google App Engine, Big Query, Cloud Data Flow, Cloud Functions, Google Compute Engine, Cloud Storage, Cloud SQL, Spark, NLTK, Map Reduce, Jupyter, Google Cloud Natural API.

**Dr. Reddy’s Pvt Ltd** December 2012 to November 2016

**Data Engineer**

Hyderabad

**Roles & Responsibilities:**

* Conducted data cleaning, transformation, merging and statistical analysis and generated Jupyter notebook analysis reports for the association study and copy numbers (Python 2.7, pandas, matplotlib, seaborn, numpy, scipy, statsmodels and Jupyter notebook)
* developed predictive models, forecasts and analyses to turn data into actionable solutions.
* Led data discovery, handling structured and unstructureddata, cleaning and performing descriptive analysis, and storing as normalized tables for dashboards.
* Implemented classification using supervised algorithms like Logistic Regression, Decision trees, KNN, and NaiveBayes. Utilized various new supervised and unsupervised machine learning algorithms/software to perform NLP tasks and compare performances.
* Used R to manipulate data for data loads, extracts, statistical analysis, modeling, and data munging.
* Created instances in AWS as well as migrated data to AWS from data Center using AWS migration services including Kinesis firehose, AWS Snowball, S3 Transfer acceleration, etc.
* Working on Multiple AWS instances, set the security groups, Elastic Load Balancer and AMIs.
* Created Macros, to generate reports daily, monthly basis and moving files from Test to Production.
* Experience in end to end BI and DWH implementation for the raw data acquisition till dashboard and KPI measurements using tools like Tableau.
* Evaluate the performance of various algorithms/models/strategies based on the data sets.
* Analyzed and interpreted large datasets with advanced statistical techniques (regression, trees, neural networks, principal components, factor analysis, conjoint procedures, time series analyses and others) to extract actionable business and consumer insights.
* Interaction with Business Analyst and other Data Architects to understand Business needs and functionality for various project solutions.

**Environment:** AWS Snowball, S3 Transfer acceleration, SQL Assistant, Python, BI, Birst, Elastic Load Balancer, A/B testing, Tableau, Machine learning, Spark, HDFS, Hive, Linux, Python (Scikit-Learn/SciPy/Numpy/Pandas), R, MySQL, Eclipse, PL/SQL, Tableau, Visual Studio

**Education:**

* Bachelors in Information Technology, JNTUH, Hyderabad in 2012.