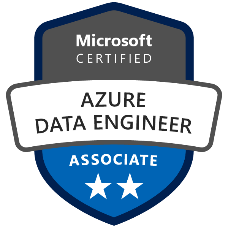
[](https://www.credly.com/badges/14e5b968-7a1c-482e-8572-15f4c026636d/public_url)

**VYSHNAVI P**

(732) 366-1404 | [vyshnavi.p508@gmail.com](mailto:vyshnavi.p508@gmail.com) | [LinkedIn](https://www.linkedin.com/in/vyshnavi-p/) | [GitHub](https://github.com/vyshnavi-pvr)

**PROFESSIONAL SUMMARY**

* **Around 7 years of progressive experience in** **Data Science, ML Engineer, Data Engineer** - interpreting and analyzing disparate data sets to drive business, performing statistical analysis, building, and evaluating machine learning models, creating a dashboard to provide actionable insights for decision making
* Implemented a **Confluence** workflow to track the progress of different projects
* Proficient with HPC applications and frameworks like **MPI**, **OpenMP**, and **TensorFlow**, and ability to optimize them for Kubernetes clusters
* Skilled in **Data Manipulation** and **Data Preparation** techniques using **Python (Pandas/NumPy)**
* Experience with **Web Scraping** using REST APIs and **Automation** using **Python** and **VBA**
* Experience in creating Python scripts to extract and load data from flat files like **CSV, JSON** to **Redshift/S3** and also moving data from different applications using APIs
* Experienced with cloud platforms like Amazon Web Services, Azure, Databricks (both on Azure as well as AWS integration of **Databricks**)
* Experience working with HPC scheduling systems such as **SLURM**, **PBS**, and integrating them with **Kubernetes**
* Hands-on experience in creating **Docker containers** leveraging existing Linux Containers and AMI's in addition to creating Docker containers from scratch. As well as worked on **Docker hub**, creating **Docker images** and handling multiple images primarily for middleware installations and domain configurations.
* Proficient in **designing relational database models** and writing **SQL** queries to extract, combine, transform, and aggregate data
* Experience with **Git** and a strong understanding of pull request workflows
* Hands-on experience with **Snowflake** utilities, SnowSQL, SnowPipe, Big Data model techniques using **Python**
* Understanding of the Oracle Cloud Infrastructure and ability to perform integrations and customizations
* Integration of Maven/Nexus, Jenkins, GIT, Confluence and Jira
* Experience in using different Hadoop ecosystem components such as HDFS, YARN, MapReduce, Spark, Pig, Sqoop, Hive, Impala, Hbase, Kafka, and Crontab tools
* Defined and deployed monitoring, metrics, and logging systems on **AWS**
* Hands on experience on Google Cloud Platform (GCP) in all the bigdata products BigQuery, Cloud Data Proc, Google Cloud
* Good experience with Dynamics 365, AX, GP, and/or NAV
* Proficient in SAS programming, including knowledge of advanced SAS features such as SAS/STAT, SAS/GRAPH, and SAS/ETS
* Good experience in deploying, managing and developing with **MongoDB** clusters
* **Docker** container orchestration using ECS, ALB and **Lambda**
* Proficient in scripting and automation using tools such as **Bash**, **Python**, or **Ansible** to manage and deploy Kubernetes clusters and workloads.
* Experience with Unix/Linux systems with scripting experience and building data pipelines.
* Experience in **AWS** services like **EC2**, **Lambda**, and **Step** functions
* Automated end-to-end data science project using **Jenkins**
* Strong experience working in all phases of **Software Development Life Cycle** (**SDLC**) in both **Waterfall** and **Agile** environment
* Ability to communicate cross-functionally, derive requirements and architect shared datasets; ability to simplify and explain complex problems to both technical and non-technical stakeholders
* Ability to thrive in a dynamic environment - willing to jump in and take on critical challenges

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| Languages & Scripting | Python, SQL, Java, R, Shell Scripting, C++ |
| Reporting/Analytics  Tools | Tableau, Power BI, Looker, Google Analytics, MS Excel, Erwin, SSAS, Informatica |
| Databases | Oracle 11g, MySQL, Microsoft SQL Server, MS Access |
| Statistical Modeling | Time Series, A/B Testing, Hypothesis Testing |
| Project Management | Agile, Waterfall, Kanban, Trello, MS Visio , LucidChart |
| Big-Data | Hadoop (HDFS, MapReduce, Hive), PySpark, Impala, Apache Spark, MongoDB, Kafka |
| Cloud-Computing | AWS (EC2, S3, RDS, Redshift), Azure ML , GCP |
| Machine Learning | Logistic Regression, Linear Regression, Decision Trees, Random Forests, K-Nearest Neighbors, Gradient Boost, Naïve Bayes, K-Means Clustering, Hierarchical Clustering, Density-Based Clustering, Gaussian Mixtures, Principal Component Analysis (PCA), XGBoost, SVM |
| Deep Learning | Convolutional Neural Networks, Multi-Layer Perceptron, Recurrent Neural Networks, LSTM |
| Libraries | NumPy, SciPy, Pandas, Scikit-Learn, TensorFlow, Keras, PyTorch, Statsmodel, NLTK, Seaborn, Matplotlib, H2O, Beautiful Soup, Auto ARIMA, Genism, Selenium, PyMongo, Ploty, GGPlot |
| IDE | Jupyter Notebook, VS Code, RStudio, NetBeans, Eclipse, Notepad++ |
| Web Services/Server | REST, Apache Tomcat, SOAP |
| Containers and orchestration | Docker, Kubernetes |
| Data Warehouse | Snowflake, AWS Redshift |
| Version Control | GitHub, Jira |
| Automation | Jenkins, Airflow |
| Operating Systems | Windows, Linux, Ubuntu, Unix |

**EDUCATION**

Northeastern University  
**Master of Science (Software Engineering Systems)**

**PROFESSIONAL EXPERIENCE**

**Data Scientist/Data Engineer**  May 2022 –Present

*AAA Life* *Boston, MA*

* Written ETL jobs in Data Net and **Informatica** to process data from different sources to transform data to multiple targets
* Implemented ETL pipelines, querying **Snowflake** DB using SQL, and developing scripts using Unix, Python, etc. to load, extract, and transform data into the DB
* Implemented various ETL and ELT flows to move data from different sources to **S3** and **Redshift**
* Evaluated multiple Machine Learning models in **Python (SVMs, Logistic Regression, Neural Networks, XGBoost, Gradient Boost)** to classify delinquent customers with a 62 Gini Index
* Used **Grafana** to monitor and analyze **Kubernetes** clusters and microservices
* Worked on real time streaming, performed transformations on the data using **Kafka** and **Spark** **Streaming**
* Implemented Spark using Pyspark, Spark streaming for real time and SparkSQL for faster testing and processing of data
* Integrated the ML model with the front-end using the **Flask** framework. Stored input data and intermediate results in **AWS S3** buckets. Deployed the model as **AWS Lambda** and **Step** **functions**.
* Integrated applications with various services such as **Redis, RabbitMQ,** and **Kafka**
* Used Hive to retrieve data from the **Hadoop** cluster and SQL to retrieve data from Oracle DB. Created scripts using **SQL** and **Hive** to validate successful data load and quality
* Used **Terraform** configuration files for creating AWS infrastructure which consist of API Gateway, Elastic Load balancers, IAM roles, DNS using Route 53 services, Multiple ECS Clusters.
* Create, configure and support the **Redshift scalable clusters** that supports the data warehouse.
* Created Data pipeline process using python to load data from **S3** to **Redshift** and transfer of files between different applications
* Created monitors, alarms, notifications and logs for AWS Lambda functions, Glue Jobs, EC2 hosts using **Cloudwatch**
* Developed RESTful APIs with **FastAPI**, including endpoints, request and response models, authentication and authorization mechanisms, and error handling
* Integrated **Swagger** with other tools, such as **Docker**, **Kubernetes**, and **Jenkins**, to automate API deployment, testing, and continuous integration/continuous deployment (CI/CD) pipelines
* Designed E2E **Azure** cloud-based analytics including **Power BI** dashboards
* Implemented Spark using **Pyspark**, Spark streaming for real time and SparkSQL for faster testing and processing of data
* Worked on transferring data from **Kafka topic** into **AWS S3** storage
* Used **AWS Glue** for the data transformation, validate and data cleansing
* Used Python **Boto3** to configure the services **AWS glue, EC2, S3**
* Create custom BI reports using Data Module and Analysis for HCM and financial modules and responsible to identify the data with **Fusion Cloud database architecture**
* Processed the output from **PIG, Hive** and formatted it before sending to the **Hadoop** output file
* Performed CRUD operations on large datasets in **MongoDB** using the **PyMongo** module
* Transformed data with Data frame and Data set APIs in **Spark** and **Scala**. Recorded results in **JSON** on a real-time basis and sent to business owners across the organization through **AWS SNS**
* Involved in managing and reviewing **Hadoop** log files
* Worked in an **Agile** environment using **JIRA** end-user stories

**Environmen**t: Python (Scikit-learn/ NumPy/ Pandas/ Matplotlib/ NLTK/ PySpark), Machine Learning (Logistic Regression/Support Vector Machine/Gradient Boosting/Random Forest), Spark 1.6, Scala, FastAPI, Swagger, AWS (EMR, S3, SNS, Glue, Redshift),Grafana, Cloudwatch, Boto3, Oracle DB, Terraform, Kafka, Informatica, Git, Flask, Hadoop Cluster, R,Snowflake, PowerBI

**Data Scientist/Data Engineer (Remote)** April 2020 –Dec 2021

*Verizon Plano, TX*

* Leveraged **NLP** techniques like Word2Vec, BOW (Bag of Words), Avg Weighted Word2Vec, and TF-IDF to create word embeddings
* Integrated **FastAPI** with various tools and platforms, such as **Kubernetes**, **Docker**, AWS,, to enable scalable and secure deployment of web applications
* Worked with **Kafka** to integrate data from multiple topics to database. Managed **RESTful API** to integrate with Streamsets to move data.
* Performed end-to-end delivery of **PySpark** ETL pipelines on **Azure databricks** to perform the transformation of data orchestrated via **Azure Data Factory (ADF)** scheduled through Azure automation accounts and trigger them using **Tidal Schedular**
* Designed and implemented secure data pipelines into a **Snowflake** data warehouse from on premise and cloud data sources
* Worked on **HCM Data Loader (HDL), File Based Loader (FBL)** for loading Oracle HR
* Experience with data pipeline and workflow management tools such as Luigi, **Airflow**, etc.
* Setup and benchmarked **Hadoop/HBase** clusters for internal use
* Stored the log files in **AWS S3**. Used versioning **in S3** buckets where the highly sensitive information is stored
* Led the design and implementation of a data warehousing solution using **Oracle EBS** as the primary data source
* Developed ETL processes to extract data from various EBS modules, transform the data to conform to a common schema, and load it into the **Redshift**
* Create, configure and support the **Redshift scalable clusters** that supports the data warehouse.
* Transformed data with Data frame and Data set APIs in **Spark** and **Scala**.
* Recorded results in **JSON** daily on a real-time basis and sent to business owners across the organization through **AWS SNS**
* Utilized **Istio** and **Ambassador** to manage and secure microservices in **Kubernetes** clusters
* Integrated **FastAPI** with various tools and platforms, such as **Kubernetes**, **Docker**, AWS,, to enable scalable and secure deployment of web applications
* Created scripts using **SQL** and **Hive** to validate successful data load and quality
* Developed **MapReduce** pipeline for feature extraction using Hive. Used Hive to analyze partitioned and bucketed data
* Implemented Data Lake in **Azure Blob Storage**, Azure Data Lake, Azure Analytics, Data bricks Data load to Azure SQL Data warehouse using **Polybase**, **Azure Data Factory**
* Processed the output from **PIG, Hive** and formatted it before sending to the **Hadoop** output file
* Weekly presentation to the business users about the **Power BI** reports and their changes
* Implemented **CI/CD** data pipelines for various machine learning models via **Jenkins**

**Environment:** Python (TensorFlow, Pandas, Numpy), Machine Learning (Logistic Regression/Support Vector Machine/Gradient Boosting/Random Forest), Azure databricks, Kafka, MongoDB, MySQL (Oracle), Kubernetes, PySpark, Snowflake, Jenkins, PySpark, Git, Airflow, Docker

**Data Scientist/Data Analyst (Offshore)** April 2019 – March 2020

*Excelra Boston, MA*

* Gather business requirements, definition, and design of the data sourcing, working with the data warehouse architect on the development of logical data models
* Implemented various statistical techniques to manipulate the data (missing data imputation, principal component analysis, and sampling)
* Designed and developed ETL processes using **Talend** to extract, transform, and load data from various sources into a **Snowflake** data warehouse, resulting in improved data accuracy and consistency
* Developed a **PySpark** application to pull data from multiple vendor sources to the Data Lake using Sqoop, performed transformations on raw data using **Spark SQL** to build data extracts according to business needs, and store them in **Hive** tables
* Deployed and managed applications on **GCP** (Google Cloud Platform) using tools such as **App Engine, Cloud Functions**, and **Kubernetes**
* Identified outliers and inconsistencies in data by conducting exploratory data analysis (EDA) using python **NumPy** and **Seaborn** to see the insights of data and validate each feature
* Performed univariate and multivariate analysis on the data to identify any underlying pattern in the data and associations between the variables
* Developed and maintained custom **PL/SQL** scripts to automate data processing tasks, such as data validation, cleansing, and aggregation
* Used **Power BI**, Power Pivot to develop data analysis prototype, and used Power View and Power Map to visualize reports
* Used **Oracle Fusion Functional** Setup Manager to perform the setup tasks required to complete the implementation
* Designed Data Marts by following Star Schema and **Snowflake** Schema Methodology, using industry leading Data modeling tools like **ER Studio**
* Experience with using **FBDI** templates to import data into **Oracle E-Business Suite** (EBS)
* Worked with containerization technologies such as **Docker** and **Kubernetes** to deploy deep learning models in a High Performance Compute (HPC) cluster
* Implemented Installation and configuration of **multi-node cluster** on Cloud using Amazon Web Services (**AWS**) on **EC2**
* Integrated **AWS Dynamo DB** using **AWS lambda** to store the values of items and backup the **DynamoDB** streams
* Involved in Unit testing and Integration testing of the code using **PyTest**
* Used clustering techniques like **DBSCAN, K-means,** and **Hierarchical clustering** for customer profiling to design insurance plans according to their behavior pattern
* Used **Grid Search** and **Random Search** to evaluate the best hyper-parameters for my model and **K-fold cross-validation** technique to train my model for best results
* Worked with Customer Churn Models including **Random forest regression, lasso regression** along with pre-processing of the data
* Designed rich data visualizations to model data into human-readable form with **Matplotlib**

**Environment:** Python (Scikit-learn/ NumPy/ Pandas/ Matplotlib), Machine Learning (Logistic Regression/Naïve Bayes/ KNN/ Support Vector Machine /Random Forest), Kubernetes, Power BI, , PyTest , Agile, SQL Server, Docker, PySpark, Snowflake, GCP, ETL, UNIX, Neural Networks

**ML Engineer/Data Analyst** July 2018 – March 2019

*Onward Health Hyd,* *India*

* Communicated and coordinated with the end client for collecting data and performed **ETL** to define the uniform standard format. Queried and retrieved data from **Oracle database servers** to get the dataset
* Developed and maintained scalable data pipelines for processing large datasets on Google Cloud Platform **(GCP)** using tools such as Cloud Storage, BigQuery, BigTable, Data Proc (Spark), and Pub/Sub
* In the data exploration stage, used correlation analysis and graphical techniques in **Matplotlib** and **Seaborn** to get some insights about the patient admission and discharge data
* Experimented with predictive models including **Logistic Regression, generalized linear models, Support Vector Machine (SVM), Gradient Boosting** and **Random Forest** using Python **Scikit-learn** to predict whether a patient might be readmitted
* Utilized **PyTorch's GPU** acceleration to improve the training and inference speed of deep learning models using Distributed data model
* Implementation of Web testing tools such as **Postman** to test the calls to the web services and check if correct data is returned
* Utilized **CUDA** and **cuDNN** to accelerate deep learning operations on NVIDIA GPUs
* Implemented, tuned and tested the model on **AWS EC2** with the best performing algorithm and parameters
* Implemented **Spring Boot** microservices to process the messages into the **Kafka** cluster setup
* Worked with **Oracle Cloud Infrastructure** and ability to perform integrations and customizations
* Designed Data Marts by following Star Schema and **Snowflake** Schema Methodology, using industry leading Data modeling tools like **ER Studio**
* Designed and implemented effective Analytics solutions and models with **Snowflake**
* Deployed the model on **AWS Lambda**. Collected the feedback after deployment, retrained the model and tweaked the parameters to improve the performance
* Installed **Ansible Registry** for local upload and download of **Docker** **images** and even from **Docker hub**
* Used **Agile** methodology and **Scrum** process for project developing

**Environment:** AWS EC2, AWS Lambda, AWS S3, Oracle DB, Linux, Python (Scikit-learn/ NumPy/ Pandas/ Matplotlib), Machine Learning (Logistic Regression/Support Vector Machine/Gradient Boosting/Random Forest), Snowflake, GCP, Docker and Tableau

**Jr. Data Analyst (Unpaid Internship)** January 2016 – June 2018

*NinjaCart Hyd, India*

* Involved in reviewing and understanding the Business requirements
* Wrote scripts to Import and Export data **to CSV, EXCEL** formats from different environments using **Python**
* Developed remote integration with third party platforms by using **RESTful web services**
* Developed web services that make database calls for **inserts, updates, and select queries**
* Developed Unit, Integration and Performance Test Cases using **Junit, Selenium** and **Rational Function Tester**
* Created and managed automated **ETL** processes to extract, transform, and load data from various data sources into **GCP** (Google Cloud Platform) for analysis and reporting
* Reviewed basic **SQL** queries and edited inner, left, and right joins in **Tableau Desktop** by connecting live/dynamic and static datasets
* Worked with Oracle databases and the ability to write efficient and optimized **PL/SQL** code
* Created complex **SQL** codes, functions, and procedures. Built data pipeline in **Kafka** to fetch the data from OLTP and store in warehouse.
* Improved code reuse and performance by making effective use of various **design patterns**.
* Used Python library **Beautiful Soup** for **web scrapping**
* Fetched twitter feeds for certain important keyword using python-twitter library
* Created action, filters, and calculated sets for dashboards and worksheets in **Tableau**
* Managed large datasets using **Pandas data frames** and **SQL**
* Wrote and executed various **MySQL** database queries from python using Python-MySQL connector and MySQL DB package
* Test-Driven Development approach was used and developed Merge jobs in Python to extract and load data into MySQL database. Worked on **SQL Server Integration Services (SSIS)** and **SQL Lite**
* Developed methods for **Create, Read, Update and Delete (CRUD)** in Active Record
* Used **Django** evolution and manual **SQL** modifications able to modify Django models while retaining all data, while site was in production mode

**Environment**: Python, Celery, Django, Angular, Beautiful Soup, Apache Kafka, JQuery, JavaScript, AJAX, XML, MySQL server, T-SQL, GCP, Tableau, Jasper, GIT, Reports, RESTful, Linux