# Divyansh Yadav

+1 (419)-378-2895 | divyanshyadav47@outlook.com | LinkedIn | GitHub | divyanshyadav.info

## SUMMARY

Experienced Data Engineer with a Master's in Computer Science and 6+ years of expertise in designing scalable data solutions. Proficient in ETL processes, cloud-based architectures, machine learning, and data pipeline development. Skilled in Python, AWS, DevOps, and Big Data technologies.

## SKILLS

**Methodologies:** SDLC, Agile, Waterfall
**Programming Language:** Python, R, SQL, SAS
**Big Data Ecosystem:** Hadoop, Hive, Apache Spark, Pig, PySpark, Databricks, Snowflake, MapReduce, Sqoop
**ETL Tools:** SSIS, Apache NiFi, Apache Kafka, Talend, Apache Airflow, Informatica
**Cloud Technologies:** AWS, Azure, GCP
**DevOps Tools:** Docker, Kubernetes, Jenkins, CI/CD
**Packages:** NumPy, Pandas, Matplotlib, SciPy, Scikit-learn, Seaborn, TensorFlow
**IDE's:** PyCharm, Jupyter Notebook
**Visualization Tools:** Tableau, Power BI, SSRS, Looker
**Databases:** MySQL, MS-SQL Server, HBase, Cassandra, DynamoDB, MongoDB, PostgreSQL
**Version Control:** Git, GitHub, GitLab
**Operating System:** Windows, Linux

## EXPERIENCE

**Flow Global Software Technologies, LLC, Bowling Green, OH | Senior Back End Developer** **Sep 2024 – Present**

- Developed backend systems in Python for sales lead generation tool, transitioned monolithic architecture to microservices, designed efficient database schemas, and integrated services for seamless data flow to AI data pipelines
- Implemented CI/CD pipelines to streamline deployment processes and enhance development efficiency

**Bowling Green State University, Bowling Green, OH | Graduate Research Assistant** **Jan 2023 – Dec 2023**

- Developed machine learning models (Linear Regression, Random Forest, Gradient Boosting, XGBoost) using Scikit-learn for price prediction model, Compared model performances, and optimized accuracy.
- Utilized Pandas and NumPy for data manipulation, and Matplotlib and Seaborn for visualization and exploratory analysis to identify key predictors.
- Implemented a Transformer machine learning model using Python, Pandas, and PyTorch for prediction of high-resolution HLA using low-resolution HLA, achieving a BLEU score of 80%.

**Impetus Technologies, Los Gatos, CA | Senior Software Engineer** **Oct 2020 - Aug 2022**

- Led Teradata warehouse migration to AWS Hive for Marsh McLennan's insurance and finance data, creating a robust AWS data lake and managing complex views and pipelines, achieving a 70% cost reduction.
- Migrated base tables with Informatica and ETL and generated derived tables using PySpark and Spark SQL by converting complex BTEQ logic to SQL, handling large datasets effectively.
- Built CI/CD pipelines with Git and StepFunctions to execute HQL files on EMR, translated complex Teradata views into Hive HQL, and automated view creation for Tableau.
- Architected a Step Function for weekly table refresh with robust failure handling, automating PySpark file execution on EMR, and integrated MongoDB for configuration data retrieval and audit logging.

**Impetus Technologies, Los Gatos, CA | Software Engineer** **Jul 2019 - Oct 2020**

- Developed a data lake accelerator for AWS with CloudFormation as an Impetus product, to expedite the creation of end-to-end big data lakes by establishing VPC and foundational infrastructure, cutting time-to-market by 70%.
- Engineered a Python tool for managing AWS CloudFormation templates, implementing Infrastructure-as-Code. Integrated AWS Glue for data cataloging and improving data quality across both raw and curated datasets.
- Utilized AWS EMR for batch data analysis using Spark and Hive, stored analyzed data in AWS S3.
- Employed Amazon SageMaker to deploy machine learning models and visualized data with AWS QuickSight.

**Impetus Technologies, Los Gatos, CA | Associate Software Engineer** **Aug 2018 - Jul 2019**

- Automated client migration processes using Python, Bash Script, and Jenkins for Concentrix, transitioning from server-oriented pipelines to server-less Kubernetes jobs, reducing migration costs by 40%
- Deployed multiple microservices on AWS by designing and implementing a CI/CD pipeline using Jenkins and Git.
- Enhanced system stability with Python Django for RESTful APIs and ReactJS dashboard for monitoring AWS instances and Kubernetes pods.

**Metasystems, India | Data Engineer** **May 2017 - Jul 2018**

- Implemented data processing frameworks with Hadoop, MapReduce, and Apache Spark, reducing data processing times by 40% for large-scale datasets.
- Designed and implemented ETL processes using Informatica, Apache NiFi, and Apache Kafka to streamline data flow across multiple systems, reducing data latency by 30%.
- Developed and maintained big data processing solutions using Pig, Sqoop, Pyspark, and Databricks, optimizing data transformation and reducing processing times by 40%.
- Developed and managed data warehouses with MySQL and PostgreSQL, enhancing data retrieval speeds and storage efficiency.
- Collaborated with cross-functional teams to integrate data pipelines and optimize workflows using GitHub for version control.

## EDUCATION

**Master of Science in Computer Science** **Apr 2024**
Bowling Green State University Bowling Green, OH

**Bachelor in Computer Science** **May 2018**
Rajiv Gandhi Proudyogiki Vishwavidyalaya (R.G.P.V) Bhopal, Madhya Pradesh

## PROJECTS

**ETL Fetch [GitHub] | Python, AWS SQS, Boto3, PostgreSQL, Docker** **Jul 2024 – Jul 2024**

- Designed ETL pipeline for Fetch Rewards, extracted streaming data from AWS SQS on the local stack, transformed the data using Python, and loaded it to PostgreSQL after data masking and transformation.
- Constructed an advanced data masking algorithm that obscures data in the database while allowing for easy readability and duplicate identification, with the capability to unmask the data if required.

**HLA Imputation [GitHub] | Python, PyTorch, NumPy, Pandas, Sklearn, NLP** **Nov 2023 – Mar 2024**

- A transformer-based machine learning model to predict HLA genotypes.
- HLA are proteins—or markers—on most cells in your body. Your immune system uses HLA to identify which cells belong in your body, it is important for organ transplant.

**Flight Data Analysis [GitHub] | Python, Spark, NumPy, Pandas, SQL, Big Data** **Sep 2023 – Nov 2023**

- Analyzed the 2015 Flight Delays and Cancellations dataset using PySpark and machine learning techniques.
- Executed querying techniques to rank airline companies and created a predictive model for flight delays with 80% accuracy.
- Used flight information to find the shortest path between two airports using Dijkstra's Algorithm.

**Knowledge Graph for ALS [GitHub] | Python, TensorFlow, Pandas, Matplotlib** **May 2023 – Aug 2023**

- Constructed a Knowledge Graph (KG) for Amyotrophic Lateral Sclerosis (ALS) using innovative ML algorithms and NLP techniques.
- Efficiently extracted entities and relationships from scholarly articles to build the Knowledge Graph.

**Data Lake Accelerator [AWS] | CloudFormation, Glue, EMR, VPC, QuickSight** **Aug 2018 – Jan 2019**

- A data lake accelerator for AWS using CloudFormation as part of an Impetus product.

## CERTIFICATIONS

**AWS Solutions Architect Associate:** Udemy
**Advanced SQL Certificate:** Hacker Rank
**Spark with Scala:** Udemy