**JAINISH SAVALIYA**

(315) 395-0245 | jainish7.js@gmail.com | [Linkedin](https://www.linkedin.com/in/jainish10/) | [GitHub](https://github.com/jainish77)

# EDUCATION

**Syracuse University, School of Information, Syracuse, New York August 2022 - May 2024**

Master of Science | Information Systems | Advanced Certification in Data Science GPA: 3.7/4.0

**Coursework:** Applied Machine Learning, Data Warehousing, Data Analytics & Decision Making, Database Management Systems

# Dharmsinh Desai University, College of Engineering, India July 2017 - May 2021

Bachelor of Technology | Instrumentation and Control Engineering GPA: 3.2/4.0

**Coursework**: Microprocessor & Micro-Controller, Mathematics, Advanced C Programming, Robotics Engineering

# EXPERIENCE

**Graduate Research Assistant, School of Information, C4 Lab, Syracuse, New York November 2023 – Present**

Reddit Data Transformation and Analysis (Ongoing)

* Designing scalable data pipeline to extract and transform 500GB Reddit dataset (100M+ comments/submissions), aiming for 20% faster processing using zstandard compression and parallel processing techniques.
* Developing custom Python scripts to parse 150M JSON records with 98% accuracy, ensuring 95% overall data quality and integrity.
* Engineering efficient data storage solution by converting data to Apache Parquet format, optimizing for 15% better compression and 20% faster partition pruning to enable high-performance queries on 16-core server.
* Validating transformed Parquet data by running 20+ SQL queries, confirming 98% data integrity and 90% queryability for downstream analysis and using structured 500GB dataset to build ML models forecasting Reddit user behavior and engagement.
* Presenting initial findings and 500GB Parquet dataset to professor, demonstrating suitability for predictive models with 80%+ accuracy.

**Programmer Analyst, Saeculum Solutions Pvt Ltd, Ahmedabad, India June 2021 – April 2022**

* Built D3.js & React front-end, Driving15% increase in user engagement &10% improvement in data-driven decisions.
* Conducted 10 A/B tests on web features, resulting in a 10% improvement in key metrics such as conversion rate and bounce rate.
* Collaborated cross-functional to define data requirements, design analytics pipelines, and improve data-driven decisions by 20%.
* Presented 20 data analysis findings to stakeholders, influencing 50% of product roadmap decisions and marketing strategies.

# SKILLS

**Programming Languages & IDEs:** Python, R, SQL, PySpark, Scala, Visual Studio Code, PyCharm

**Packages/Tools:** NumPy, Pandas, Matplotlib, SciPy, TensorFlow, PyTorch, scikit-learn, Spark, Hadoop, Hive, Seaborn, Ggplot2

**ML& Statistical Techniques:** Linear Regression, Logistic Regression, Decision Trees, Classification, Time Series Forecasting, Hypothesis Testing, K Means, KNN, Naïve Bayes, Random Forests, XGBoost, SVM, CNN, LLM

**ML-Ops:** EC2, Lambda, IAM, Elastic Beanstalk, RDS, S3, VPC, CloudWatch, Docker, Databricks, CI/CD, Apache Airflow, Git, Linux

**Databases:** MySQL, PostgreSQL, MongoDB, Cassandra, Oracle

**Visualization Tools & Methodologies:** Tableau, Power BI, Alteryx, Plotly, SDLC, Agile, Waterfall

**Certifications:** AWS Solutions Architect - Associate**,** Snowflake Hands on Essentials - Data Warehouse**,** Excel Associate (2019)

# PROJECTS

**Business Intelligence Solutions for Fudge Mart Inc. (Snowflake, DBT, Tableau) November 2023**

# Performed data integration with ELT for 1M+ records, streamlining data accessibility and enhancing order fulfillment efficiency by 24%.

# Streamlined data transformations and pipeline automation within Snowflake using DBT for fast and accurate data delivery.

# Leveraged DBT for Snowflake data modeling, created Tableau dashboards visualizing KPIs to optimize order fulfillment.

# Loan Eligibility Prediction (Pandas, NumPy, scikit-learn, Matplotlib) April 2023

* Developed loan prediction model using Python, achieving 84% accuracy and 20% improvement over previous model.
* Implemented data science techniques such as feature engineering and ensemble models to reduce bad loans by 15%, increase approved loans by 10%, and cut processing time by 25%.Top of FormBottom of Form

# Spotify ETL Data Pipeline using Airflow January 2023

* Orchestrated a scalable Spotify end-to-end data pipeline using Apache Airflow, leveraging the Spotify API to extract and transform approximately 100MB JSON to CSV daily, enhancing downstream analytics accessibility.
* Deployed on EC2 via Directed Acyclic Graphs (DAGs) for reliable data delivery, streamlining complex pipelines and optimizing processing time by 15%.

# Healthcare Cost Analysis and Prediction (R, Shiny, ggplot2) November 2022

* Analyzed healthcare costs using R and statistics, visualized findings for Health Management Organization and reduced costs by 10-15%.
* Formulated regression model accurately predicting individual healthcare costs, enabled high-cost patient management.
* Utilized advanced analytics techniques and built a predictive model that reduced hospital readmissions by 3-5%, and designed an interactive Shiny app presenting insights, which empowered data-driven decisions and improved operational efficiency by 10-15%.