# Satyam Mishra

satyamgusmishra@gmail.com • 857-701-3819 • Boston, MA • Linkedin


## SUMMARY


2 years 3 months of Data Engineering and Data Science work experience using Python, C++, SQL, and AWS


## SKILLS

Programming: C++, R, Bash, Python (Pandas, Numpy, Matplotlib, Seaborn, NLTK, PySpark, Sklearn, Tensorflow)
Tools/Frameworks: ETL, AWS, Glue, Cloudwatch, Lambda, Vault, Spark, Hadoop, Data Modeling, Git, CI/CD, Docker
Databases: MySQL, Postgres, SQLite, NoSQL, MongoDB, Apache Cassandra, SSMS, S3, Redshift, Airflow, OLAP, OLTP
Analytics: Tableau, EDA, Hypothesis Testing, A/B Testing, Market-Basket Analysis, Statistics, Dimensional Modeling

## Experience

**Data Engineer at TransUnion** *Aug 2022 - Present*

- Implemented ETL pipelines on 9 data sources using in-house variation of C++ (codex) and AWS Redshift.
- Wrote a data pipeline while complying with GLBA and CCPR for a data source of 340 Million customers.
- Assisted in writing Python DAGs to trigger ETL operations on data pipelines every 24 hours with Airflow.
- Collaborated with data scientists on a credit scoring ML model, beating previous accuracy by 9%.
- Used envelope encryption (AES-256) on existing scripts with AWS Vault and C++ using KEK and DEK.
- Wrote and managed 4 Bash scripts in testing, development and production environments.
- Undertook the task of upgrading and testing Codex 9.2 to 9.3 while adhering to data quality checks.
- Ensured timely DDX builds every week and helped in automating distributed processing in PySpark.

**Data Science and Analytics Coop at Unum** *Jul 2021 to Dec 2021*

- Automated a pipeline to stage data from S3 to Redshift, transformed, and connected to Tableau for real-time analysis of mental health status of users majorly based on PHQ-9 and GAD-7 scores.
- Implemented data models, including ER diagrams and dimensional models on raw data sources.
- Used Kafka messaging service to improve feed flow within internal applications.
- Performed data quality checks using Pandas, then EDA in Tableau & statistical analysis on 540000 records.
- Used AWS Lambda with API gateway in Python to initiate triggers to calculate results on 3 events.

**Data Science Intern at Eletech Engineering** *Jan 2020 to Jun 2020*

- Used PyTesseract to create OCR model to digitize database, migrated it on S3 for optimal ETL with Boto3.
- Implemented dimensionality reduction with PCA and LASSO on impact metrics data and created new KPIs.
- Used K-Means clustering model for segregation of faulty manufactured components on impact metrics.
- Analyzed feedback of 10000 deliveries, followed by deploying BERT base classifier to assist marketing.

## PROJECTS

**Healthcare Data Analytics Platform on MIMIC-III Data**

- Used Hadoop for secure and scalable storage of healthcare data and EMR for parallel data processing.
- Implemented data anonymization (PII) with PySpark and Hadoop ensuring HIPAA regulations.

## EDUCATION

**Master of Science in Data Analytics Engineering,** Northeastern University *May 2022*
**Bachelor of Technology in Computer Science Engineering,** Medicaps University *Jul 2020*