

Jeffrey Michael Jackovich

jeffrey.jackovich@gmail.com • github.com/JeffreyJackovich
Brooklyn, NY • jeffreyjackovich.com • linkedin.com/in/jeffrey-jackovich

Summary

Published senior data engineer with 10+ years of professional health-technology experience using Python, PySpark, and SQL to engineer scalable data pipelines. Passionate about building local and cloud-based (Databricks and AWS) ETL pipelines. Strong engineering experience analyzing business requirements to process structured and unstructured healthcare data.

Skills

Languages: Python, SQL, Java

Statistics/Machine Learning: Statistical Analysis, NLP, Classification, Clustering, Regression

Tools and Libraries: Git, Jupyter, Pandas, Numpy, Scikit-learn, DBT, NLTK, Spacy, Matplotlib, Selenium, Cloudera (Apache Impala)

Frameworks: Apache Spark, Delta Lake, pytest

Cloud: Databricks, AWS

Databases: PostgreSQL, MSSQL, MySQL

AWS: Redshift, Lambda, S3, Glue, Comprehend

Experience

Senior Professional Data Engineer, Gainwell Technologies (prev. HMS Holdings), New York, NY **05/2021 - Present**

- Assisted build a PySpark-based machine learning pre-processing and model deployment library using Abstract Base Classes and MLFlow.
- Completed proof of concept NLP machine learning pipeline to increase APR and MS claim finding rate by engineering an end-to-end machine learning data pipeline with a model output F1-score score was ~88%.

Senior Professional Data Engineer, HMS Holdings (acquired by Gainwell Technologies), New York, NY **08/2020 - 05/2021**

- Worked on a four-person agile team engineering a Python based contract modeling proof of concept.
- Identified 7 market competitor (e.g. cotiviti, etc.) opportunities by engineering an end-to-end Python ETL from unstructured medical claims using fuzzy string matching and pandas.
- Classified concepts in unstructured medical claim notes with a ~94% F1-score using Random Forests.
- Engineered an ETL to extract billing concepts from 1M+ SQL Server unstructured medical claims.
- Prioritized claims to reprice with a Python algorithm processing 400M+ SQL Server record.

Software Engineering Senior Advisor / Big Data Engineer, Cigna, New York, NY **11/2019 - 03/2020**

- Worked on a seven-person agile team to build a data pipeline from HDFS on-premises to AWS for a contract modeling application.
- Engineered a prototype ETL pipeline with Glue, PySpark, and Lambda to transform text files (15M+ records) to Parquet.
- Performed data quality analysis within Cloudera Impala using Hive/SQL and migrate masked PHI (Protected Health Information) data from HDFS on-premises to S3.
- Identified columns in HDFS tables that contained PHI with a 99% recall score by re-engineering an open-source NLP library.

Python Technical Analyst, Remedy Partners, New York, NY **03/2019 - 09/2019**

- Worked on an agile team with four data engineers to perform data quality analysis with SQL and Python.
- Performed data quality analysis with pandas and SQL to identify and categorize duplicate records in Redshift and MySQL.
- Increased duplicate record detection by 11%, compared to SQL efficacy, by engineering a NLP algorithm that identified 10,000+ partial duplicates in 20M+ total Redshift records.

Machine Learning Engineer / Co-Author, Packt Publishing, New York, NY **05/2018 - 02/2019**

- Translated Packt's authoring request for the title '*Machine learning with AWS*' to both a book about natural language processing (NLP) algorithms and a Python-based NLP data pipeline.
- Architected a NLP data pipeline in Python with S3, Lambda, and Comprehend to output sentiment, key phrases, and entities from text documents.
- Developed 40+ Python programming examples to augment NLP concepts including topic modeling, latent dirichlet allocation, theme extraction, sentiment analysis, and entity detection.

Data Engineer, Freelance, New York, NY

01/2017 - 05/2018

- Privately contracted by a VP of Physician Development to engineer a lead generation data pipeline to identify, filter, and extract information on dermatology practices for acquisition.
- Enabled a client to meet a 72-hour lead-generation deadline by web-scraping a JavaScript website with Selenium and Python.
- Engineered a Python algorithm to integrate disparate lead generation data from social media APIs (e.g. Yelp, Twitter, etc.).

Data Engineer / Data Analyst, One Medical, New York, NY

09/2010 - 01/2017

- Programmed a ranking algorithm on 2M+ records with Python and pandas to prioritize lead generation that resulted in a 10.0% of total outreach, prospect attendance rate at M&A events.
- Created a valuation model from multiple data sources (e.g. medical records, claims, and tax forms) with pandas to identify risk factors in M&A deals of independent practice-owning physicians.
- Used pandas to join and transform disparate lead generation data (e.g. CSV, JSON, HTML, and XML) from open source APIs (e.g. Yelp, LinkedIn, etc.) with a custom ETL into a PostgreSQL data warehouse for lead generation analysis.

Education

MS, Computer Information Systems, Boston University, Boston, MA

08/2016 - 01/2019

BS, Biology, Bradley University, Peoria, IL

08/2001 - 05/2006

- Minors: Psychology and Philosophy

Certification

- AWS Technical Professional

01/2017

Publication

Jackovich, J., & Richards, R. (2018). *Machine Learning with AWS*. Birmingham, UK: Packt Publishing. ISBN-13: 978-1789806199 [<https://amzn.to/2UuSZru>]

Hackathon

Data Engineer, Mount Sinai Health Hackathon, New York, NY

10/2019

- Built an app for diabetic patients to speak their blood glucose levels to Alexa and receive dietary and exercise suggestions.
- In thirty-six hours, worked on an 11 person team from ideation to a MVP and competed against 20 teams.
- Built a Naive Bayes classification model from anonymized patient's time series blood glucose data to predict hypoglycemic outcomes.
- Technologies: Python, Node.js, Alexa, MongoDB, Heroku

Presentation

- Jackovich, J. (March 27, 2015). *Analyze Public Medical Data*. Presentation at the New York Machine Learning Workshop, New York, NY. [<https://www.meetup.com/New-York-ML-Workshop/events/221103883/>]