THOMAS GEORGE THOMAS

Boston, MA • +1-857-891-3705 • thomasgeorgethomases@gmail.com • LinkedIn • GitHub • Portfolio

Education

Master of Science, Data Analytics Engineering
Northeastern University, Boston, MA — Expected August 2023
GPA: 3.9/4.0
Relevant Coursework: Data Mining, Machine Learning, Data Management for Analytics

Bachelor of Technology, Computer Science and Engineering
Manipal Institute of Technology, Manipal University, Manipal, India — May 2016

Skills

Programming Languages: Python, Scala, SQL, Unix shell scripting
Data Engineering: Hadoop, Apache Spark, Hive, Impala, Sqoop, Snowflake, MySQL, API, Streamlit, Control M, Heroku
Amazon Web Services: S3, Athena, Glue, EMR, EC2, Lambda, Step Functions, Batch, SQS, Redshift, Boto3
Packages: Pandas, NumPy, Scikit-learn, Matplotlib, Requests, BeautifulSoup, Multiprocess, Pytest
Data Visualization: Tableau, Flourish, Data wrapper
DevOps: Agile, Git, Bitbucket, GitHub, Bamboo, Maven, Confluence, Jira
Certifications: IBM Certified Data Science Professional

Experience

Data Engineering Co-Op/Intern — July 2022 - December 2022
Montai Health, Massachusetts, USA

- Established AWS ETL pipelines using Redshift, SQS, Lambda, EMR, EC2, PySpark, Athena, and Glue to transform 100 TB data
- Developed health, drug, and bioinformatic Data Lake from RDBMS (SQL) and NoSQL databases on AWS worth 100 TB
- Created web scrapers to crawl data from CSVs, XMLs, Parquet, APIs, and FTP servers leveraging Python to collect 5 GB data daily
- Implemented CI/CD, test-driven development, and test automation on GitHub actions increasing code quality by 100%

Senior Data Engineer — June 2018 - August 2021
Legato Health Technologies - Elevance Health, Bangalore, India

- Constructed data pipelines for 5 initiatives including providing Clinical Investigative Insights in AWS, Hadoop, and Apache S park
- Migrated 112 TB of data from on-premises Hadoop cluster onto AWS Cloud and Snowflake
- Innovated and automated post-migration validation reports in Spark Scala leading to $7000 quarterly savings
- Executed continuous integration and continuous deployment (CI/CD) pipelines for 4 projects deploying DevOps
- Chaired release management and code migration for production/pre-production environments for 5 projects

Software Engineer - Big Data — June 2016 - May 2018
Middle East Management Consultancy and Marketing, Muscat, Oman

- Shipped and delivered analytics dashboard leading to an increase in pharmaceutical sales by 12% annually
- Developed pipelines to handle data of 1.5 TB/day from ingestion to reporting layer using Shell script, Hadoop & Spark
- Implemented dataset transfer of 26 TB between Hadoop and MySQL RDBMS leveraging Sqoop
- Performed performance tuning in Spark, SQL, and Sqoop resulting in a 60% response time reduction
- Restructured and redesigned Data Lake to utilize Parquet, and Snappy compression to cut 30% storage and compute costs

Projects

YouTube Analytics Dashboard — December 2022 - Present

- Created live dashboard on Streamlit operating python and YouTube API picturing Sentiments and 5 KPIs for any video

Age of Plastic — March 2022 - April 2022

- Designed data-backed dashboards showcasing adverse effects of plastic and mitigation steps on Tableau

Retro Movies Recommender — February 2021 - May 2021

- Built a recommendation engine API for 50 movies from 1900s using Natural Language Processing, Python, Flask, and Heroku

Clustering Paris and London — August 2020 - September 2020

- Analyzed cities of Paris and London to show distinct features and visualized similar neighborhoods using Folium, Python, ArcG IS, Foursquare API, and K Means Clustering Machine Learning model