

THOMAS GEORGE THOMAS

Boston, MA • +1-857-891-3705 • thomasgeorgethomases@gmail.com • [LinkedIn](#) • [GitHub](#) • [Portfolio](#)

SKILLS

Programming Languages: Python, Scala, SQL, Unix shell scripting, Cypher
Big Data Technologies: Hadoop, Apache Spark, Hive, Impala, Sqoop, API, Streamlit, Control M, Heroku, Linux
Amazon Web Services: S3, Athena, Glue, EMR, EC2, Lambda, Step Functions, Batch, SQS, Redshift, Boto3
Databases & Data Warehouses: Snowflake, Microsoft SQL Server, MySQL, MariaDB, SQLite, MongoDB, Neo4j, PostgreSQL
Packages: Pandas, NumPy, Scikit-learn, Matplotlib, Requests, BeautifulSoup, Multiprocess, Pytest, SQLAlchemy, Plotly, Anaconda
Data Visualization Tools: Tableau, Flourish, Data Wrapper, DBeaver, Jupyter Notebooks
DevOps & CI/CD: Agile, Git, Bitbucket, GitHub, Atlassian Bamboo, CRON, Maven, Confluence, Jira
Certifications: IBM Certified Data Science Professional

EXPERIENCE

Data Analyst February 2023 - Present

Northeastern University, Boston, USA

- Building a Dashboard for 20 Smart Homes showcasing uptime analysis, energy prediction, and KPIs on Plotly Dash and MariaDB

Data Engineer July 2022 - December 2022

Montai Health, Cambridge, USA

- Developed AWS ETL pipelines using Redshift, SQS, Lambda, EMR, EC2, PySpark, Athena, and Glue to transform 100 TB data
- Established 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 daily
- Implemented CI/CD, test-driven development (TDD), and test automation on GitHub actions, increasing code quality by 100%

Senior Big Data Engineer June 2018 - August 2021

Legato Health Technologies, Bangalore, India

- Constructed data pipelines for 5 initiatives, including providing Clinical Investigative Insights in AWS, Hadoop, and Apache Spark
- 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 - Hadoop Developer & Big Data Engineer June 2016 - May 2018

Middle East Management Consultancy and Marketing, Muscat, Oman

- Shipped and delivered analytics product, leading to an increase in pharmaceutical finance sales by 12% annually
- Designed scalable data processing pipeline from data ingestion to reporting layer using Shell scripting & Spark worth 1.5 TB/day
- Implemented data sets 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 Data Lake design to utilize Parquet and Snappy compression to cut storage and compute costs by 30%

EDUCATION

Master of Science, Data Analytics Engineering

Northeastern University, Boston, Massachusetts, USA

Expected December 2023

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

PROJECTS

Appliances Energy Prediction March 2023 - May 2023

- Predicted the energy consumed by appliances using custom-coded Machine Learning models and Algorithms like PCA, Neural Networks, Lasso, Ridge, and Linear Regression from scratch in Python with 80% confidence

YouTube Analytics Dashboard December 2022 - March 2023

- Created a real-time dashboard on Streamlit operating Python and YouTube API, picturing Sentiments and 5 KPIs for any video