THOMAS GEORGE THOMAS

Boston, MA 02119 | thomasgeorgethomases@gmail.com | +1 857 891 3705 | linkedin.com/in/thomasgeorgethomas/ |
www.thomasgeorgethomas.com | GitHub: github.com/Thomas-George-T

EDUCATION

Northeastern University, Boston, MA   Expected Aug 2023
Master of Science in Data Analytics Engineering, GPA: 3.92
Courses: Computation and Visualization, Data Mining, Foundations of Data Analytics

Manipal Institute of Technology, Manipal University, Manipal, India   May 2016
Bachelor of Technology in Computer Science & Engineering

SKILLS

Languages	Python, Scala, R, SQL, Unix shell scripting
Data Engineering	Hadoop, Spark, Hive, Impala, Sqoop, Snowflake, MySQL
AWS	S3, Athena, Glue, EMR Serverless, EC2, Step Functions
Data Visualization	Tableau, Flourish, Data wrapper, Google Data Studio
Data Science	Supervised learning, Unsupervised learning, Recommender Systems, Natural Language Processing
Packages	Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, Folium, Requests, Os
DevOps	Git, Bitbucket, GitHub, Bamboo, Maven, Agile, Confluence, Jira
Other	IBM Cloud, Control M, Heroku, Google Colab, Jupyter Lab, VS Code, PyCharm, Eclipse, Anaconda
Certifications	IBM Certified Data Science Professional

EXPERIENCE

Montai Health   Cambridge - MA, USA
Data Sciences & Engineering Co-op   Jul 2022 - Present
- Developed scripts to ingest, process and upload 10,000 data files daily into AWS S3 using API calls and Python
- Aggregated 1 TB of health data, modern bioinformatic databases using AWS EMR Serverless, Athena, Glue and S3

Legato Health Technologies - Elevance Health   Bangalore, India
Senior Data Engineer   Jun 2018 - Aug 2021
- Built data pipelines for 5 initiatives including providing Clinical Investigative Insights in AWS, and Hadoop
- Migrated 112 TB of data from the on-premises Hadoop cluster to AWS and Snowflake
- Innovated and automated post-migration validation reports in Spark Scala bringing down costs by 90% for 2 projects
- Innovated and reduced latency by 50% which lead to $7000 quarterly savings by refactoring Spark Scala ETL code
- Implemented continuous integration and continuous deployment (CI/CD) pipelines for 4 projects using DevOps
- Chaired release management and code migration for production/pre-production environments for 2 projects

Middle East Management Consultancy and Marketing   Muscat, Sultanate of Oman
Software Engineer – Big Data   Jun 2016 - May 2018
- Shipped and delivered analytics dashboard which led 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 using Sqoop
- Performed performance tuning in Spark, SQL, and Sqoop resulting in 60% response time reduction
- Redesigned Data Lake to use Parquet, and Snappy compression to cut 30% storage and compute costs

PROJECTS

Age of Plastic
- Created a data-driven storyboard visualizing the impact of global plastic pollution on Tableau.

Retro Movies Recommender
- Built an unsupervised content-based recommendation engine API for 50 movies of the 1900s using NLP, Flask, Heroku, and Python

Clustering Paris and London
- Visualized the cities of Paris and London to show distinct features of each neighborhood using Folium, Python, ArcGIS, Foursquare API, and K Means Clustering Machine Learning model

Movies Analytics
- Analyzed 1,000,000 movies to draw useful insights on viewer engagement using Spark and Scala, featured at #3 on Data Machina issue #130.