THOMAS GEORGE THOMAS

Boston, MA | thomasgeorgethomases@gmail.com | +1 857 891 3705 | linkedin.com/in/thomasgeorgethomas/ |
www.thomasgeorgethomas.com | 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, SQL, Unix shell scripting
Data Engineering	Hadoop, Spark, Hive, Impala, Sqoop, Snowflake, MySQL, API
AWS	S3, Athena, Glue, EMR, EC2, Step Functions, Batch, SQS, Redshift
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, Multiprocess, Pytest
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   Massachusetts, USA
Data Scientist/Data Engineer   Jul 2022 - Present
- Built a health, drug and bioinformatic data lake of 3 TB from SQL and NoSQL databases in AWS using APIs and Python
- Developed ETL pipelines using Redshift, SQS, Lambda, Batch, EMR, Athena, Glue, and S3 to transform TBs of data
- Set up CI/CD pipelines and automated code quality checks using GitHub actions to increase code quality by 80%

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 which lead to $7000 quarterly savings
- Redesigned and refactored project architecture and Spark Scala ETL code bringing down costs by 90%
- 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 5 projects

Middle East Management Consultancy and Marketing   Muscat, 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 a 60% response time reduction
- Redesigned Data Lake to use Parquet, and Snappy compression to cut 30% storage and compute costs

PROJECTS

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

Movies Analytics
- Analyzed 1 million movies to draw useful insights on viewer engagement and movie ratings using Spark and Scala

Age of Plastic
- Visualized a data-backed storyboard showcasing the adverse effects of plastic on our ecosystem and how to mitigate its effects using Tableau.

Retro Movies Recommender
- Built a content-based recommendation engine API for 50 movies of the 1900s using Natural Language Processing, and Python deployed using Flask and Heroku.