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. **GPA: 4.0** Expected Aug 2023
Master of Science in Data Analytics Engineering
Courses: Computation and Visualization, Data Mining, Foundations of Data Analytics, Deterministic Operations Research

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

TECHNICAL SKILLS

Languages	Python, Scala, R, SQL, Unix Shell scripting
Data Engineering	Hadoop, Hive, Impala, Spark, Sqoop, IBM Cloud, AWS; S3, Athena, Step functions, EMR, RDS
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
DevOps	Git, Bitbucket, GitHub, Bamboo, Maven, Agile, Confluence, Jira
Tools	Snowflake, MySQL, Control M, Heroku, Google Colab, Jupyter Notebook
Certifications	IBM Certified Data Science Professional

EXPERIENCE

Legato Health Technologies, Anthem Inc. 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 Apr 2022 - May 2022

- Created a data-driven storyboard visualizing the impact of global plastic pollution on the environment; Land and Ocean and the recycling rates of the different countries on Tableau.

Retro Movies Recommender Mar 2021 - May 2021

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

Clustering Paris and London Jul 2020 - Aug 2020

- 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

Predicting Healthcare Costs Jul 2020 - Aug 2020

- Predicted the cost of healthcare and insurance using Python and Linear Regression Machine Learning model with 80% accuracy

Movies Analytics Mar 2020 - May 2020

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