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: Foundations of Data Analytics, Deterministic Operations Research, Computation and Visualization, Data Mining

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

TECHNICAL SKILLS

Tools	Hadoop, Hive, Impala, Spark, Sqoop, Snowflake, MySQL, Control M, AWS, IBM Cloud, Heroku, Tableau
Languages	Scala, Python, R, SQL, Unix Shell scripting
Packages	Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, Folium,
DevOps	Agile, Confluence, JIRA, Git, Bitbucket, GitHub, Bamboo, Maven
Data Science	Supervised learning, Recommender Systems, Data Visualization, Natural Language Processing
Certifications	IBM Certified Data Science Professional

EXPERIENCE

Legato Health Technologies, Anthem Inc. Bangalore, India
Senior Data Engineer Jun 2018 - Aug 2021

- Built data pipelines to provide clinical investigative insights in AWS using S3, Athena, Step functions, and EMR
- 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 runtime by 50% which lead to $7000 quarterly savings by refactoring Spark Scala ETL code
- Developed and managed enhancements, code migration, release management, production loads, and continuous integration and continuous deployment (CI/CD) pipelines for 4 projects using Bamboo, Maven, Git, and Shell scripting
- Proficient in stakeholder interaction, requirements gathering, data analysis, design documents, performance tuning, and enhancements

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 1.5 TB of data daily from ingestion to reporting layer using Shell scripting, Hadoop & Spark
- Implemented Sqoop for dataset transfer of 26 TB between the Hadoop and MySQL RDBMS.
- Performed performance tuning, analysis, and response time reduction techniques in Spark, SQL, and Sqoop
- Redesigned the Hadoop ecosystem to handle different file formats such as CSV, Parquet, and snappy compressed files

PROJECTS

Predicting Hits on Spotify Feb 2022 - Present

- Analyzing over 40000 songs of 6 different decades to predict hit songs on Spotify using various classification Machine learning models and Python.

Retro Movies Recommender Mar 2021 - May 2021

- Built a unsupervised content based recommendation engine API for 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

Treatment Costs Prediction 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 a million movies to draw useful insights on viewer engagement using Spark and Scala, featured at #3 on Data Machina issue #130.