THOMAS GEORGE THOMAS

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

EDUCATION

Northeastern University, Boston, MA — Expected Aug 2023
Master of Science in Data Analytics Engineering
Courses: 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

Data Engineering	Hadoop, Hive, Impala, Spark, Sqoop, Kafka, Snowflake, MySQL
Data Science	Clustering, Recommender Systems, Linear Regression, Data Visualization, Natural Language Processing
Cloud	AWS, IBM Cloud, Heroku
Languages	SQL, Shell scripting, Scala, Python, R
Agile	Confluence, JIRA
DevOps	Git, Bitbucket, GitHub, Bamboo, Maven
Scheduler	Control M
Distributions	Cloudera, Hortonworks
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 nappy compressed files

PROJECTS

Social Media Analytics in R: Analyzed 1.6 million twitter user's data and visualized useful and interesting insights using techniques including text mining, sentimental analysis, probability, and hierarchical clustering in R. — Dec 2021

Olympic History Analytics in R: Discovered and visualized various distinctive trends after analyzing 120 years of Olympic history using R — Oct 2021

Retro Movies Recommender API: Built a content-based recommendation engine API for movies of the 1900s using NLP, Flask, Heroku, and Python — May 2021

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 — Aug 2020

Treatment Costs Prediction: Predicted the cost of healthcare and insurance using Python and Linear Regression Machine Learning model with 80% accuracy — Jul 2020

Movies Analytics: Analyzed a million movies to draw useful insights on viewer engagement using Spark and Scala, featured at #3 on Data Machina issue #130. — May 2020

Covid-19 Tweet Data Scraping: Streamed & ingested live tweets about Covid-19 between high-performance tuned Kafka and Elasticsearch — Apr 2020