THOMAS GEORGE THOMAS

Boston, MA 02119 | thomasgeorgethomases@gmail.com | +1 857 891 3705 | www.thomasgeorgethomas.ml

EDUCATION

Northeastern University, Boston, MA — 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

EXPERIENCE

Legato Health Technologies, Anthem Inc. — Bangalore, India
Senior Software Engineer Niche — Jun 2018 - Aug 2021
Domain: Healthcare

- Built data pipelines to provide clinical investigative insights in AWS using S3, Athena, Step functions, and EMR
- 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 $6000 quarterly savings by refactoring Spark Scala ETL code
- Migrated 112 TB of data from the on-premises Hadoop cluster to AWS and Snowflake
- 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
Domain: Pharmaceuticals

- Shipped and delivered analytics dashboard which led to 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
- 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

PASSION PROJECTS

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: Compared the neighborhoods of Paris and London to draw useful insights on city setup using Python 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 — Jul 2020
Movies Analytics: Analyzed a million movies to draw useful insights with Spark and Scala, featured on Data Machina issue #130 at number 3 — May 2020
Covid-19 Tweet Data Scraping: Streamed & ingested tweets about Covid-19 using high-performance Kafka and Elasticsearch — Apr 2020

CERTIFICATIONS & AWARDS

- IBM Certified Data Science Professional
- Anthem Go Above IMPACT Award: Honored for going above and beyond in 2021
- Legato Iron Man of Technology 2: Awarded for being a standout performer for Q4 of 2019
- Legato Technology 2 Annual Team Innovation: Awarded for innovations delivered during 2019 - 2020
- GitHub Arctic Code Vault Contributor: Awarded for open-source contributions towards the GitHub Archive program