THOMAS GEORGE THOMAS

Boston, MA | +1-857-891-3705 | thomasgeorgethomases@gmail.com | LinkedIn | GitHub

SKILLS

Programming Languages/Scripting: Python, Scala, SQL, Bash, Shell, Cypher, PowerShell
Big Data Technologies: Apache Spark, PySpark, Hadoop, Hive, Impala, HDFS, Sqoop, API, Streamlit, Heroku, UNIX, Linux, Tableau
Amazon Web Services (AWS): S3, Athena, Glue, EMR, EC2, Lambda, Step Functions, CloudWatch, Batch, SQS, SNS, Redshift, Boto3, RDS
Data warehouse & Databases: Snowflake, SQL Server, MySQL, MariaDB, MongoDB, Neo4j, PostgreSQL, Oracle, DynamoDB
Packages: Pandas, NumPy, Scikit-learn, Matplotlib, Requests, Multiprocess, Pylint, Pytest, SQLAlchemy, Plotly
CI/CD & DevOps: Agile, Git, Bitbucket, Atlassian Bamboo, CRON, Maven, Confluence, Jira, Docker, Anaconda (Condas), DBeaver
ML & MLOps: Docker, Airflow, GitHub actions, DVC, Flask, MLflow, Jupyter Notebook, TensorFlow, Regression, Clustering
Industry Knowledge: Kafka, GCP, GCS, Vertex AI, Looker, Cloud Composer, BigQuery, Classification, Statistical Analysis & Methods

EXPERIENCE

Data Engineer/Data Analyst | **ABLE Lab at Northeastern University** | Boston, Massachusetts, USA *Feb 2023 - Sep 2023*

- Engineered high-availability data streaming analytics platform, ensuring **a 98%** real-time uptime for **20** smart homes using **Python**
- Built an energy dashboard visualizing Key Performance Indicators (**KPIs**) and metrics using **SQL, MariaDB**, and **Jupyter Notebooks**
- Applied SQL query optimization and software engineering best practices, leading to a **60%** reduction in SQL query response times
- Performed in-depth Exploratory Data Analysis (**EDA**) to unveil patterns and trends to support data-driven decision-making processes

Data Engineer | **MontAI Health** | Cambridge, Massachusetts, USA *Jul 2022 - Dec 2022*

- Established collected, cleaned, and aggregated health, food, drug, biotech, and bioinformatics **Data Lake** on **AWS** of **100 TB** of data
- Developed **ETL** (Extract, Transform, Load) pipelines using **AWS** Glue, Redshift, S3, PySpark, Lambda, and SQS to process **100 TB**
- Integrated **2 ML pipelines** with Continuous Improvement/Learning and automated deployments using **Apache Airflow** and **Docker**
- Implemented Test Automation by test-driven development (**TDD**) GitHub actions workflows, improving existing code quality by **100%**
- Performed data integration of **5 GB/day** from disparate data sources, **APIs,** and file formats (**XML, CSV, JSON, Parquet, Avro, ORC**)
- Built large-scale applications and data structures for data ingestion from **Neo4j**, **MongoDB** (Graph, NoSQL OLTP) databases in **Airflow**
- Translated business objectives and Key Results (**OKRs**) into technical specifications, design documentation, and optimal solutions

Senior Big Data Engineer | **Legato Health Technologies** | Bengaluru, Karnataka, India *Jun 2018 - Aug 2021*

- Constructed and **led** scalable **ETL/ELT** data pipelines for **5 US healthcare** initiatives across Anthem and Blue Cross Blue Shield
- Optimized large-scale data transformations in **AWS Cloud** using EMR, EC2, Athena, CloudWatch, Step Functions, cutting costs by **30%**
- Automated data quality framework in **Spark Scala** for **Hive** and **SQL Server**, cutting errors, resulting in **$7000** quarterly cost savings
- Built extensible data architecture with **10 Snowflake** data marts and data models, improving Business Intelligence (BI) metrics by **20%**
- Implemented join optimization, data partitioning, and performance tuning, achieving a **40%** improvement in **Apache Spark** jobs
- **Mentored** and led a team of **4** new graduates in a training program focused on **Big Data** and **DevOps** tools (Jira, Bitbucket, Bamboo)
- Orchestrated code migration, continuous integration, and continuous deployment (**CI/CD)**, reducing deployment time by **25%**
- Unified **Agile Scrum** with unit testing, system integration testing, and code reviews, slashing post-deployment defects by **18%**

Software Engineer - Hadoop & Big Data | **Middle East Management Consultancy & Marketing** | Muscat, Oman *Jun 2016 - May 2018*

- Created high-volume robust data infrastructure on distributed computing on-premises Hadoop clusters capable of handling **15 TB**
- Designed interactive **Tableau Dashboards** to derive meaningful insights, leading to a **12%** increase in pharmaceutical finance sales
- Developed **10** end-to-end high-volume, multi-layer data processing pipelines from ingestion layers to the serving and reporting layers
- Ingested **26 TB** from Relational databases (**MySQL, Oracle, PostgreSQL**) RDBMS via Sqoop and Shell scripts, enhancing data access
- Improved enterprise data warehouse (**EDW**) scalability using facts and dimensions **data modeling** (Star Schema, SCD) by **20%**
- Established **DataOps** for version control with git, data cleaning, data management**,** governance, and data lineage tracking for datasets
- Maintained integrity and compliance with clean data sets for **4** projects, contributing to a **15%** improvement in accuracy and reliability
- Collaborated with cross-functional teams, managed stakeholders, and gathered requirements for streamlined project communication

EDUCATION

Master of Science, Data Analytics Engineering *Sep 2021 - Dec 2023*
Northeastern University, Boston, Massachusetts, USA
Relevant Coursework: Data Mining, Machine Learning, Machine Learning Operations

Bachelor of Technology, Computer Science and Engineering *May 2012 - May 2016*
Manipal Institute of Technology, Manipal University, Manipal, India

PROJECTS

E-commerce Customer Segmentation | GitHub |: Deployed and Monitored real-time Clustering machine learning models and algorithms for E-commerce customers on Google Cloud Platform (GCP) using Python, Vertex AI, BigQuery, GCS, Flask, Airflow, Docker, MLflow, TensorFlow, and Looker