Thomas George Thomas

thomasgeorgethomases@gmail.com | GitHub: Thomas-George-T | www.thomasgeorgethomas.ml

PROFILE SUMMARY

- Big Data Engineer with **5+** years of experience in Design, Architecture, Development, and Deployment of **Hadoop, Spark & Big Data Technologies** with work experience in the Middle East and India.
- Skilled in Hadoop, Spark, Hive, Shell scripting, Sqoop, SQL, Scala, and Kafka
- Good experience with **Agile DevOps** lifecycle tools to build and maintain code quality, continuous integration, and continuous deployment (CI/CD) pipelines.
- Having exposure to **NoSQL** databases like Splice Machine, HBase, and DynamoDB.
- Having Operational knowledge of **IBM Cloud** and **AWS**.
- Holds domain knowledge about Healthcare and Pharmaceuticals.
- Having exposure to Operational Analytics and Reports.
- Possess strong debugging, Critical thinking skills, and problem-solving abilities.
- IBM Certified **Data Science** Professional

TECHNICAL SKILLS

Big Data	Hadoop, Hive, Impala, HUE, Spark, Sqoop, Kafka, Zookeeper
Cloud	IBM Cloud, AWS, Heroku, Snowflake
Languages	SQL, UNIX Shell scripting, Scala
Agile	Confluence, JIRA
DevOps	Git, Bitbucket, GitHub, Bamboo, Maven
Scheduler	Control M
Distributions	Cloudera, Hortonworks
Operating Systems	Microsoft Windows, Linux-Ubuntu, OpenSUSE

EDUCATION

Manipal Institute of Technology Bachelor of Technology in Computer Science & Engineering	2016

EXPERIENCE

Legato Health Technologies, Anthem Inc.   Bangalore, India.
Senior Software Engineer - Niche   Nov 2020 - Present
Software Engineer – Big Data   June 2018 – Oct 2020
Domain: Healthcare

CII: Clinical Investigation data are aggregated and transformed to provide Insights into patient healthcare.
- Transformed segmentation data from S3, Athena to the reporting layer in Snowflake.
- Built data pipelines in AWS leveraging services S3, RDS, Athena, Step functions, and EMR.
- Migrated data from the on-premise Hadoop cluster to AWS and Snowflake.

Technologies: AWS: RDS, S3, EMR, Glue, Spark, Scala, Hive, Hadoop, Snowflake, Git, Bitbucket, Maven

COCA: Dollar amounts and utilization metrics of the services rendered by providers for their members are computed based on the monthly and quarterly data. These are then used to generate Tableau reports sent to the providers for their incentive assessment.
- Engaged primarily in developing spark Scala code involving RDD's, dataframes, and SparkSQL.
- Developed shell scripts to process 1.5 TB CSV, Parquet data from inbound to the outbound layer for generating Tableau reports.
- Developed fully automated CI/CD pipelines using Bamboo to migrate Unix items and ETL jars into pre-prod and prod environments removing any manual effort.

Innovations & Enhancements:
- Automated validation reports post-migration bringing down costs by 90%.
- Improved runtime from 18 hours to 9.5 hours by refactoring Spark Scala ETL code.
- Refactored tables to use parquet formats, snappy compressions, and include partitions.

Technologies: Spark, Scala, Hive, Impala, Hadoop, Unix, Shell scripting, Control M, Bamboo, Git, Bitbucket, Maven, Eclipse, Cloudera distribution

CAHMO, CCQP, HPIP, NDW: Healthcare data across the US is collected, cleansed, and stored into a hive data lake where further processing is based on transformation logic according to business requirements using spark to meet the business requirements and sent to Blue Cross Blue Shield & providers.
- **Leadership:** Interim Lead for a team size of 4 spanning 4 projects.
- Primarily developed and managed enhancements, code migration, release management, production loads, and continuous integration and deployment (CI/CD) for 4 projects.
- Efficient in stakeholder interaction, requirements gathering, data analysis, design documents creation, solutions, performance tuning, and enhancements.
- **Innovation:** Improved efficiency and turnaround time from 6 hours to 1.5 hours by automating data quality and validity checks between Hive and SQL server loads.
- **Innovation:** Developed SIT automation scripts in Spark Scala to assess quality, validity, counts of inbound data files & tables to remove manual effort and intervention.

Technologies: Spark, Scala, Hive, Sqoop, MS SQL Server, Shell scripting, Control M, Git, Maven, Eclipse, Cloudera Distr.

Middle East Management Consultancy and Marketing   Muscat, Sultanate of Oman.
Software Engineer – Big Data   June 2017 – May 2018
Associate Software Engineer – Big Data   June 2016 – May 2017

Hadoop Enterprise Data Warehouse: A top pharmaceutical organization in Oman needed a Hadoop cluster for data analytics. The Consumer product details are collected and loaded using SQOOP from heterogeneous database sources into Hadoop. The data cleansing has been done using MapReduce jobs and queried using Hive. The Data transformed to generate reports. For the first time, this project was implemented as part of the business solution and marketing strategy. This pilot project turned into one of the top-level projects of the firm.
- Shipped and delivered product end to end.
- Implemented SQOOP for massive dataset transfer between the Hadoop file system and RDBMS.
- Involved in the design and creation of partitioned table DDLs in Hive.
- Worked on performance tuning, analysis, and response time reduction techniques in SQL and Sqoop.
- Worked with different file formats such as CSV, Parquet, and snappy compressed files.
- Processed delimited data using Spark SQL to build pipelines from landing zone to outbound layer.

Domain: Pharmaceuticals
Technologies: Hadoop, Sqoop, Hive, Shell scripting, MySQL, Spark, Scala, Hortonworks Distr.

PASSION PROJECTS
- **Retro Reel Recommender:** A content-based recommendation engine API for movies of the 1900's built using NLP, Flask and Heroku. [GitHub | Demo]
- **A Tale of Two Cities:** Clustering the Neighborhoods of Paris and London using K Means Machine Learning model. [GitHub | Medium]
- **Movies Analytics in Spark and Scala:** Analyzing a Million Movies to draw useful insights with Spark and Scala, featuring on Data Machina issue #130 at number 3. [GitHub]
- **Covid-19 Tweet Data Collection:** Streaming & collecting tweets about Covid-19 using high-performance Kafka and Elasticsearch. [GitHub]
- **File Processing Comparative Analytics:** Data Analysis to find the quickest and slowest file processing capacity among different languages and execution engines. [GitHub | Medium]
- **Regression on Personal Health Data:** Predicting the cost of treatment and insurance using Linear Regression Machine Learning model. [GitHub | Medium]

AWARDS
- Anthem IMPACT: Awarded for going above and beyond in 2020.
- 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 for 2019 – 2020.
- GitHub Arctic Code Vault Contributor: Awarded for OSS contributions towards the GitHub Archive program.

COURSEWORK & CERTIFICATIONS

- IBM Certified Data Science Professional.
- AWS Solutions Architect Associate
- Apache Kafka
- Network Management
- Python for Data Science and AI
- Machine Learning with Python