Santhosh Raj Murugesan

San Jose, CA | https://www.linkedin.com/in/santhosh-raj-murugesan | santhoshraj2960@hotmail.com | 3194005132

Summary

Data engineering leader with a passion for innovation and 6+ years of experience in the industry

Skills

Languages / Frameworks: Python, Pyspark
Databases: NoSQL (parquet, orc, hive, delta lake), SQL (postgres)
Cloud: Microsoft Azure, AWS, Databricks
CI / CD: Jenkins, Azure devops
Others: Data warehouse, ETL/ELT, Docker, LLM, Langchain, API, Airflow, Yarn, Data lifecycle management, Hadoop, Data Structures, Algorithms, Machine Learning, PowerBI, Tableau

Experience *(6.5 years)*

Adobe, San Jose - *Senior Data Engineer* *Jan 2022 - Present*

Project: Databricks ETL migration and ADLS

- ETL pipelines: Designed, built, orchestrated and optimized data ETL pipelines from scratch on Azure databricks
- ETL optimization: Optimized ETL run time and costs by 70% by applying various techniques such as smaller lookbacks, reusing job cluster to maximize parallelization, caching, lazy evaluation, and tuning spark and delta configurations
- ETL performance: Improved ETL runtime by 40% by using cache accelerated workers
- Query optimization: Achieved 30% optimization in query run time by leveraging delta's dynamic file pruning and consolidating small files
- Storage optimization: Decreased data storage costs by 70% by running periodic vacuum and deleting unused tables
- Metadata optimization: Optimized metadata operations by 90% by replacing multiple parquet tables with one delta table and using merge upsert and selective insertion

Project: ML cost optimization

- Cost optimization: Achieved 70% cost reduction in various machine learning models by choosing the appropriate hardware and tweaking the code to ensure high hardware utilization throughout the process.

Project: Firefall LLM API

- Developed and oversaw Firefall API built on Azure OpenAI, which was utilized by several teams across Adobe's product ecosystem.
- Transitioned machine learning models from asynchronous to synchronous mode, resulting in a 10% reduction in API request processing time

University of Iowa - *Teaching Assistant* *Aug 2019 - Dec 2021*

- Teaching Experience: Led discussion and lab sessions and provided office hours for various courses, including cloud computing
- Recognition: Nominated for the outstanding teaching assistant of the year award in 2021 by the professor I worked for
- Skills: Endorsed by Prof. Thamer on Linkedin for my communication and conflict resolution skills

DrumUp, India - *Lead Software Engineer* https://drumup.io *Dec 2015 - Jul 2019*

- Responsibility: Led a team of three engineers and built a social media management app from scratch
- Performance: Improved api and db query speed by 30% and 50% respectively by optimizing postgres db and ElasticSearch
- Recognition: Received Linkedin endorsements from DrumUp's CEO and CTO for leadership and technical skills

Project: DrumUp Chrome Extension Github | Chrome Web Store Link

- Content recommendation: Developed a content recommendation app that suggested relevant articles to users based on the web page they were viewing and allowed them to share them on social media
- Keyword prediction: Used RAKE algorithm to extract keywords from billions of web pages and indexed them with ElasticSearch for faster and more accurate content matching
- Conversion rate: Increased the number of paid users by 15% by providing more personalized and engaging content recommendations

Duta Software, India - *Software Engineer* *Dec 2014 - Dec 2015*

Project: Celery-Rabbitmq

- Task acceleration: Used celery and rabbitmq to speed up background tasks by 70x
- Task distribution: Parallelized millions of tasks across multiple machines

Education

Master in Computer Science, The University of Iowa, *Dec 21*
BS in Information Technology, Anna University, *May 15*
Relevant coursework: Data Structures, Design and Implementation of Algorithms, Big Data Management and Analytics

Independent Projects & Certifications

Certification: *Microsoft Azure Champion* *Microsoft Certificate*

- Certified **Microsoft Azure Champion** for exploring azure services using Microsoft's student scholarship

Project: NYC Taxi Data - ETL and report generator (Pyspark) Github | Github

- Designed an end to end ETL pipeline that generates reports and KPIs for NYC taxi trips data using Azure Databricks and Apache Airflow
- Reduced manual effort to **Nil** by enabling CI/CD using Jenkins and Azure Devops