Santhosh Raj Murugesan

San Jose, CA | https://www.linkedin.com/in/santhosh-raj-murugesan | santhoshraj296000@gmail.com | 3194005132

Education

Master in Computer Science, The University of Iowa, **GPA: 3.67** *Dec 2021*

Skills

Programming Languages: Python(5 yrs), Pyspark(3 yrs), Javascript, C
Databases: SQL, Postgres, Hive, Elastic Search, Redis
Cloud: Microsoft Azure(3y), AWS(3y), Databricks(1y)
CI / CD: Jenkins, Azure devops
Others: Spark, Databricks, Delta lake, Data orchestration and ETL pipeline, Data lifecycle management, Apache Airflow, Docker, Hadoop, HDFS, Kafka
Github: https://github.com/santhoshraj2960

Experience *(5 years)*

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

Project: Oozie Data ETL

- Maintaining current data ETL pipelines and adding new ones to move data from Hadoop to Azure blobs

Adobe, San Jose - *Data Engineer Intern* *May 2021 - Aug 2021*

Project: Azure Data Lake Storage (ADLS) - Delta lake ETL 1

- Optimized ETL run time by **82%** by replacing full data load with incremental data reload
- Leveraged delta's merge upsert and selective insertion for updating and appending new data
- Achieved **50%** optimization in the run time of sql table join queries by leveraging delta's dynamic file pruning

Project: Azure Data Lake Storage (ADLS) - Delta lake ETL 2

- Designed and orchestrated data ETL pipeline using Azure databricks and airflow
- Copied bronze data in CSV format from Azure blob storage to Delta lake on Azure data lake gen 2 storage
- Optimized ETL run time by **40%** by fine tuning spark's shuffle partitions and delta configurations

Project: Azure Data Lake Storage (ADLS) - Data integrity

- Integrated deequ for data quality and integrity checks in ETL pipeline
- Performed these checks after ingestion to delta table
- Time travelled back and restored a stable version of the table if integrity failed
- Achieved **85%** optimization in run time by replacing integrity checks on CSV (before ingestion), to Delta (after ingestion)

Project: Azure Data Lake Storage (ADLS) - Selective deletion (GDPR)

- Developed a python script that deletes a particular customer's information from all the tables in delta lake
- Achieved **300x** faster execution of queries through optimal partitioning, z-ordering and activating file pruning
- Reduced manual effort for code deployment to **nil** by facilitating CI / CD using Azure devops and Jenkins

Project: Azure Data Lake Storage (ADLS) - Data lifecycle management

- Developed a python script that uses Azure's python SDK to move data across different access tiers (Hot, cool, archive and deleted)
- Cut down costs by **15%** by moving data to less expensive access tiers

Principal Financial Group, Des Moines - *Data Engineer Intern* *June 2020 - Aug 2020*

- Developed a model using AWS sagemaker that predicted the financial advisors, the sales team should go after, which in turn increased the conversion rate by **15%**

DrumUp, India - *Software Engineer (Full Stack)* https://drumup.io *Dec 2015 - Jul 2019*

- Automated payment tracking and subscription management with the help of two checkout api and efficiently handled payments of up to 2 million dollars
- Reduced db access time by **50%** by normalizing tables and eliminating redundancies
- Migrated tables that required text search to ElasticSearch and reduced db query time by **50%**

Project: DrumUp Chrome Extension Github | Chrome Web Store Link

- Suggested articles relevant to the contents of a web page the user is viewing and enabled sharing it on their social media accounts
- Suggested content **60%** more relevant to users using Python by predicting keywords from the content retrieved from billions of websites.
- Achieved this accuracy by introducing RAKE for keyword extraction and ElasticSearch as database

- ➢ This resulted in an increase of free to paid user conversion rate by **15%**

Project: Celery-Rabbitmq
- ➢ Achieved **70x** times faster execution of background tasks by integrating celery-rabbitmq
- ➢ Distributed millions of tasks across different machines and executed them parallely

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

Project: Task Manager (Django) Github
- ➢ Online task management system for creating and managing projects/tasks within a company

Certification: *Fire Extinguishing Robot / Robex* Certificate 1 | Certificate 2 | Github
- ➢ Presented a paper titled "Fire Extinguishing Robot" that won **second place** in graVITas'12, an **international** tech event at VIT, Vellore. Later, built a functional version of this prototype and secured **first place** at Protocol 15e

Project: Protein Visualizer (Pandas, Matplotlib) Github
- ➢ Tool for analytics and graphical visualisation of protein data