Santosh Sawant

(+91) 7829674405

contactmeurgently@gmail.com

santoshsawant

ssawant.github.io

Experienced and results-driven technology leader with over 16+ years of expertise in Machine Learning, and Software Development, seeking a role to spearhead the development of a cutting-edge machine learning product, platform and services.

CERTIFICATION

SKILLS

Libs & Frameworks: *Pytorch, Tensorflow, Langchain, Llama index, openCV, Kubeflow, Mlflow, seldon, argoCD, airflow, spark, flink.*

Programming Language : *python, golang, Java, javascript and , c++.*

Data Science and ML Techniques :

Generative AI - Fine Tuning, Eval, RAG, caching, guardrails & data flywheel

Time series forecasting, NLP, NLU, NLG, Predictive Modeling, Clustering Models, Data and Quantitative Analysis, Object and Keypoint detection and pixel wise prediction

WORK EXPERIENCE

Sr Architect Machine Learning @ Tredence Inc. (2022 - present)

Leading a cross-functional team consisting of research, machine learning, software, DevOps, and data engineers to deliver a cutting-edge Generative AI platform and products in the Data Analytics, Healthcare and ESG domain.

- *Developed distributed cloud GPU training approaches for deep learning models, including data distribution editing, data quality improvements, and representation learning with self-supervision.*
- *Developed, tested and improved various model compression techniques for model serving.*
- *Develop and implemented various technique for RAG and foundation model evaluation under various scenarios*
- *Experience in reading papers and implementing algorithms described in papers to increase performance, quality, data management, and accuracy of AI systems.*
- *Designed lean proofs of concepts (POC) to answer targeted business questions using Gen.AI.*
- *Design, Developed and integrated various large-scale, distributed machine learning systems to deploy production grade Gen.AI products and services on multi cloud environments.*

Lead Machine Learning Engineer @ Parallel Wireless (2021 - 2022)

Lead Machine Learning engineer at parallel wireless. Responsible for developing end-to-end machine learning pipeline along with defining and executing ML workflows.

- *Developed a key performance index forecasting system using LSTM with 70% accuracy; entailing reduced network downtime with improved log-collection triggering.*
- *Develop end-to-end ml pipeline in hybrid cloud using kubeflow pipeline (model training), mlflow (experiment tracking and*

model registry), kserve (model serving), feast feature store and jenkins CICD.

- *Defined and executed specific ML workflow, which includes data collection, sampling, model building and training, metrics definition and evaluation.*

Principal Engineer @ OLA **(2015 - 2021)**

Lead engineering team at OLA. Responsible for developing and scaling various data, machine learning and application pipelines.

- *Develop services and components for deploying Airflow and flink clusters on kubernetes and onboarded/scaled various Data and BI spark and flink jobs pipelines.*
- *Develop services and components for feature store integration with existing data and ml pipeline*

SDET - 2 @ Zynga **(2011 - 2014)**

SDET - 2 @ Pengala **(2010 - 2011)**

SDET - 1 @ Dell - EMC **(2007 - 2010)**

WORK PROJECTS

Generative AI Products and Platforms for Data Analytics, Healthcare and ESG Domain **(Tredence Inc.)**

Research, design, develop and build various GenAI LLM base products and services in the following domain.

- *YODAI - Your Own data analysis intellect. Utilize LLM agents to develop various data analysis models such as churn rate analysis, next day purchase and soon.*
- *Data Whizz. Utilize LLM chains and agents to connect to DB source and execute NLP queries as well as quick insight of DB*
- *Knowledge management system using RAGs, open as well as close source LLMs system.*
- *Healthcare - Differential DX and Prior Auth.*
- *ESG (Environment, Social and Governance) reporting*

KPI Early Warning System using time series forecasting (Parallel Wireless)

- *Created an LSTM autoencoder based Anomaly detection model to predict the outlier occurrence of KPIs with 70% accuracy; optimized run-time of model with PCA dimensionality reduction*

Cell Auto configuration using Deep Reinforcement Learning (Parallel Wireless)

Cell Auto configuration uses a deep reinforcement learning model to optimize the "tilt" and "power" settings of a cell ENodeB. Develop RLOps pipeline and integrate it with ns3 simulator (digital twin) and OpenRAN systems.