**Lalitha Abburi**

**Eamil:** [**sudheer.r@tekprosol.com**](mailto:sudheer.r@tekprosol.com)

**Contact: 609 981 8255**

**Sample project in GitHub -** [**https://github.com/labburu/Renewable-Energy-ML-project**](https://github.com/labburu/Renewable-Energy-ML-project)

Professional Summary

* ~15 years of experience in architecting and implementing real-time applications, bigdata engineering & Machine learning solutions to support operations across industries.
* Involved in the entire project life cycle including technical architecture design, infrastructure building, Cleansing, Manipulation, Engineering, Modeling, Optimization (performance tuning), QA and data pipeline (ETL) to Data warehouse Deployment on premise & on Cloud (Azure, AWS, GCP).
* Extensive experience in designing/developing/tuning/managing batch and real-time data pipelines that ingests data from source systems like APIs, legacy systems, Databases in JSON/ XML/avro/parquet/csv file formats, processed them and loaded into NoSQL and SQL databases.
* Experience in creating workflows scheduler to manage jobs by DAG of actions and control flow, performed debugging, scheduling & monitoring airflow jobs.
* Built Rest APIs to perform CRUD operations using python for different service functionalities. Deployed these microservices in docker containers and Kubernetes to automate the deployment, scaling, and management of containerized applications.
* Instantiated, created and maintained CI/CD pipelines to apply code changes to higher environments, worked on automation tools like GIT, Terraform for Infrastructure as Code, Kubernetes for containerized applications etc.
* Experience in building data modeling (start & snowflake schema) , transactional modeling & slowly changing dimension (SCD), also designed and created data warehouse(Kimball), views, indexes, stored procedures, functions for enterprise data product.
* Created dashboards and reports from databased using visualization tools like tableau, power BI, SSRS, QlikView etc.
* As a lead engineer, initiated project goals, designed end-to-end architecture, collaborated with cross-functional teams comprised of engineers, UI designers and user experience experts in the development of three new analytical & predictive tools and a host of new capabilities and present them to stakeholders and Clients.

Skills

|  |  |
| --- | --- |
|  |  |
| * **Statistical Tools**: Python (pandas, scikit-learn, TensorFlow, PyTorch, matplotlib etc.), R, SAS, SPSS | * **Data Visualization Tools**: PowerBI, Tableau, Qlikview, Looker, Advanced MS Excel (macros, arrays, pivots, lookups) |
| * **Database & Big data tech**: PostgreSQL, MySQL, SAP HANA studio, Hadoop, MapReduce, Apache Spark, Beam, Hive, Kafka, SQL server (SQL queries, PL/SQL, creating Views, triggers, stored procedures, functions), Vector database for ML/AI features, NoSQL database (MongoDB, Cassandra, DynamoDB), Vector DB(SingleStoreDB) | * **Programming Languages**: Python, FastAPI, C#, C++, Groovy, UNIX shell scripting, T-SQL, Scala(beginner), Presto |
| * **BI tools**: SSIS, SSRS, SSAS (DAX, MDX), Informatica ETL, Qlik Replicate | * **Other applications**: Salesforce CRM, SAP ERP systems, Agile (Scrum) methodology, Jira, Confluence, Alteryx, Git Repo, Django rest framework, API management (Azure) |
| * **Cloud technologies** - AWS (Redshift, EC2, kinesis, EMR, S3, Glue, RDS, SQS, Lambda function, ECS), Snowflake(Snowpipe, SnowQL), Azure (ADF, databricks, devops, Visual studio, synapse, data lake, application insights, purview, log analytics), SAAS based applications, GCP services(Cloud functions, PubSub, CloudSQL CircleCI, BigQuery, GCS, Dataproc etc), snowflake data warehouse | * **Data science skillsets**: Natural language processing, LLM models forß generative AI, Transformer Models, XGBoost, Hypothesis Testing, predictive modeling, Time-series analysis, Bayesian modeling, regression and decision trees, neural networks, Clustering, PCA |
| * **Build, CI/CD & Other tools**: Snowflake/Star schema modeling, Jenkins, Airflow, Zookeeper, Kubernetes, Terraform, Docker, Cloudwatch, Data dog, GraphQL, Denodo | * **Web services**: HTTP/HTTPS, SOAP, Rest |

Work History

Sr Data Architect/ ML Engineer Lead Consultant (Vancouver Airport Authority, Amalgamated Bank, PwC, Carenet Healthcare, Auto-collision and shipping industry, Suncor Inc, Microsoft) - Azure & GCP Cloud platforms - 09/2020-current

* Design and implementation of end-to-end data solution for optimal extraction, transformation, and loading of data from a wide variety of data sources using SQL, Azure cloud services (e.g., Azure Databricks with unity catalog, Cosmos DB, Data Factory, Azure Data Lake Storage), Hadoop, and Apache Spark, as well as building data products with analytics tools that utilize enterprise datasets to unlock actionable insights to enhance passenger experience and increase operational efficiencies.
* Utilized Apache Kafka's pub/sub model to capture real-time data from various airport sources, including flight schedules, gate changes, and delay notifications, Integrated data streams with Azure Event Hubs for seamless connectivity with other Azure services and applications, built interactive, real-time flight display dashboards using Power BI.
* Created REST APIs using Azure Functions to serve processed flight data to various airport applications, including mobile apps, geospatial applications and digital signage systems.
* Developed real-time operational application that is used by Airport staff/stake holders for airport management. Ingested data from 3rd part source Flight data, API endpoints, application logs from GCP & AWS cloud as a part of live data sharing.
* Identified, designed, and implemented internal process improvements by automating manual processes, optimizing data delivery, and ensuring data infrastructures, pipelines, and processes are scalable, resilient, and performant.
* Constructed robust and efficient data streams with Kafka on the Azure cloud, employing a toolkit that includes Databricks, Python, Spark, YAML/UNIX bash scripts, TERRAFORM code for DevOps CI/CD pipelines. Utilized event hubs for event listening, application insights for Databricks job logs, SQL warehouse for data consumption, and ADLS Gen2 for storage. Managed orchestration and containerized applications using Kubernetes for scalability, while also implementing workflow scheduling tools like Airflow to ensure job sequencing for arrival/departure passenger movements.
* utilized Python based Apache Beam libraries to develop data flow stream pipelines at scale, leveraging its autoscaling infrastructure to process large datasets seamlessly.
* Built end-to-end cloud data solutions for Automobile industry that deals with collisions, repairs and maintenance. Converted daily collision repair order XML files to a structured data format that can be used by industry experts take business decisions.
* Created value added solutions like creating ML models (LLMs, deep learning, NLP) for vehicle collision industry based on road accidents per day in various geographical locations that can predict flow of collision vehicles for repair order management to the nearest client shops, also developed bot provided real-time updates to customers on the status of their claims, ensuring transparency and improving the overall customer experience.
* Designed snowflake schema to create Dim and Fact tables in synapse SQL pool & snowflake which in turn uses Power BI to create dashboards that uncover insights into customer acquisitions, operational efficiency & other metrics.
* Due to multiple Data warehousing solutions, migrated data pipelines, from azure to snowflake multi-cluster data warehouse as it handles semi-structured data which are in turn connected to analytical engines for reporting.
* Involved in building unit tests and integration tests for automated data quality checks before deployment. Continuously improved data processes to increase scalability, security, reliability, and efficiency, implemented centralized data governance policies using azure Purview to maintain compliance across business.
* Implemented other ML models (LLM) for passenger flow demand, sentiment analysis using social media data mining, passenger communication and baggage volume forecast model, airports can optimize their operational efficiency, enhance passenger experience, and improve resource utilization.
* Developed a POC for Operations and airport flow dashboard, commercial cargo volume prediction models (both short and long term) to manage airport and customs staff availability based on volume.
* Utilized Spark-SQL & PySpark based models on Azure Databricks ML workspace, create databricks catalog, external tables using delta lake, built & deployed ML models at scale for suspicious phishing, covid-19 impact, customer activity models, ecosystem, Retention/stickiness for Microsoft products.
* Leveraged Apache Kafka for real-time data streaming ensuring low latency, developed webhooks and REST APIs to enable seamless real-time data sharing between Azure, GCP, and AWS platforms used by other airline companies and airports.
* Designed and implemented robust data models using Power BI, created interactive dashboards, Leveraged DAX (Data Analysis Expressions) to create complex calculations, measures, and custom visuals to meet business requirements.

Senior Data Architect/ML engineer - Lead role - Contractor (Clients – Uplight, Coast Capital Savings, RGF Wealth, Metrie) – AWS & GCP Cloud – 07/2019-10/2021

* Worked with Engineering Pillar to build and improve our platforms (especially our API layers and data processing) to deliver flexible and creative solutions to our Energy partners using AWS S3, kinesis, GLUE, lambda function, EC2, sns/sqs,redshift.
* Developed complex real-time energy data streams using big data-based tools such as Kafka, Apache spark, python, databricks, Scala, SQL, Airflow for logical execution flow in python, and NOSQL databases on AWS cloud along with machine learning models, deploying containerized applications (Building, testing & deployment), application insights for logging, scaling the software to handle huge data and maintaining the code base in Github for collaboration with other developers.
* Created data governance templates and standards for the project like maintaining data consistency, catalog, lineage, dictionary, profiling, mapping, quality, data life cycle, privacy and security using aws cloud security.
* Worked end-to-end on creating cloud infrastructure automation using Terraform, aws Code pipeline & jenkins, cloud formation, CircleCI and migration of client’s system from AWS to GCP in a robust way.
* Built and maintained batch and real-time data pipelines using Apache Beam pipeline in a Google Cloud Platform architecture (BigQuery, BigTable, Dataproc, Firestore, etc.) for new clients along with dashboard reports creation.
* Developed event-driven microservices with AWS Lambda and Amazon EventBridge to process real-time financial transactions and fraud detection alerts using twilio communication API.
* Utilized Python, Spark on AWS Elastic search, sagemaker to develop & execute Analytics & Machine learning models for fraud detection, credit risk modelling, forecasting, covid-19 impact, customer churn models.
* Collaborated closely with Product and other stakeholders across the company to implement a ML design that addresses energy optimization in energy sector such as natural gas, electricity and renewal energy to meet demand, minimize waste and environmental impact.
* Implemented a data warehousing solution on AWS using Amazon Redshift. This project aims to create fact and dimension tables within Redshift to support efficient data storage, querying, and analytics for banking data.
* Also, built custom ML data models or attributes for customer data that were built from banking customer data and store them in vector database. These data models were used by Data scientists for solving banking related problems especially in capital risks and customer credit default scores.
* Built & maintained product Pricing database, migrated supply chain & finance SAP HANA Database that included data ingestion using Kinesis into S3, data processing using databricks pyspark, schema development (Snowflake), ETL, creating OLAP data cubes on premise and AWS Cloud using Redshift.
* Collaborated with the team on creating real-time Pricing and Analytics data streaming containerized applications using AWS lambda exposed using API gateway, utilized kafka on AWS EC2 to fully build, train and deploy Machine learning models at scale managed on Kubernetes clusters and productionized web application for customer usage in B2B ecosystem.
* Implemented data and application security features though IAM policies, encryption, CloudWatch/data dog for log monitoring, AWS Guard duty for malicious activity, Automated Data backup, disaster recovery etc.
* For ML/AI models, built custom data (set of features) developed from multiple sources of data based on feature extraction algorithms which would set the base for AI based applications in energy consumption sector, maintained feature dataset in a vector database.
* Implemented performance improvement on Redshift using indexing, table design and distribution keys and other means, wrote automated data quality checks, CloudWatch for data pipeline monitoring, I AM policies for data security.
* Designed and automated tableau dashboards and reports to support several key projects, working closely with project teams, stakeholders and partners in Data Management which was deployed as a part of data pipeline refreshing reports on cloud.
* Migrated data from various sources using data pipelines, that transformed and loaded into snowflake Data warehouse tables, performed unit tests for data accuracy. Implemented data governance processes, compliance frameworks & data quality.
* Implemented version control system and maintaining CI/CD pipelines using Jenkins for deploying and maintaining multi-tiered infrastructure and web applications using Docker and Kubernetes handle loads.

Data science Engineer (Eli Lilly and Company, Blue Ocean, ANZ Bank, TE Connectivity, IQVIA, Genpact) - 08/2008 to 06/2019

* Was responsible for delivery of big data engineering solutions, real-time analytics & ML solutions against healthcare datasets at operational, clinical, financial, marketing & other business functions using SAS, SQL and python, loading & integrating large volumes of healthcare data to support seamless delivery on azure cloud.
* Developed event-driven applications using Azure Functions and Azure Event Grid to process and analyze large datasets, accelerating drug discovery processes.
* Built platforms integrating multiple data sources(databricks), including genetic information, lifestyle history, and patient health records, using Azure Data Factory and Azure Synapse Analytics, encrypted sensitive data and being HIPAA compliant.
* Developed and deployed ML models to predict patient outcomes, optimize treatment plans, and identify potential health risks. Models included predictive analytics for patient readmission rates and chronic disease progression.
* Extracted raw data from third-party companies like IQVIA, developed data statistical models/features for supply chain and logistics, persisted in vector database to create forecast models to fill gaps in supply-demand in Auto-immune, Cancer & Diabetes segment.
* Conducted data exploration and pre-processing and analysis for resource allotment, product placing in store, brand performance, promotion impact, consumer behavior, and ROI analysis, Inventory levels forecasting and analysis.
* Performed Bayesian, Market-mix modelling for a major brand, studied Impact of celebrity endorsement on consumer base by acquiring data from primary and secondary sources.
* Developed live predictive systems using Deep learning AI to combat online payment frauds thereby saving approx. ~$200k in terms of cost and helped mitigating risks using PySpark.
* Analyzed data of about 100,000 applicants’ information using Artificial Neural Network method to create credit risk model that predicts borrower's estimate of the probability of default.
* Was involved in building financial front-end application intensive processing of over 1 billion transactions from 25 data warehouse tables on Hadoop distributed system, transformed and aggregated data, which showcased different KPI metrics & charts at product, customer and demographic levels using power BI.
* Implemented enterprise-wide logical data model and design for health insurance transaction processing and information management system, predictive models & analytics for patient-level data.
* Responsible for processing CPG POS data (between UNIX & Mainframes), Data Cleaning, validation, developing ETL pipelines (Snowflake), transforming & loading OLAP data for reporting using SAS.
* Conducted consumer research, in-market studies and retail landscape evaluation to develop predictive models as part of a concentrated effort to improve customer reach and product placement, created ML models incorporating social media data into customer analytics, determine price elasticity for various product offerings.
* Created Automation jobs and scripts on mainframes by connecting DB2 to SAS that flagged any outliers and data issues. Also modified the C++ scripts with code corrections on testing environment.
* Support senior modelers in the development of statistical credit risk models from large datasets, including data analysis, writing and testing code, monitoring model performance, writing technical documentation, etc.
* Other ADHOC projects - Anomaly Detection & security in a Smart Industrial Machinery Plant Using IoT and Machine Learning systems to identify patterns, A/B testing project for a startup e-commerce site, educational institution database creation using PostgreSQL & Python as a freelance project.

Education

* Master of Science: Bioinformatics 09/2007

Orissa University of Agriculture & Technology - Bhubaneshwar, India

* Bachelor of Science: Biosciences (Botany Major) 11/2004

Sathya Sai University - Anantapur, India