Asha Dasi

**Senior Lead Engineer–Data Science, AI, ML, DL, NLP, GenAI & AWS ML**

[rohitb@interaslabs.com](mailto:rohitb@interaslabs.com)

Mob : 614-992-8709

# Summary

* Productive and focussed professional with 14 years of expertise in Software & Product Development, Solution Architecting. Skilled in various cutting-edge technologies such as Artificial Intelligence, Machine Learning, Deep Learning, Natural Language Processing(NLP), ChatGPT, LLM, Data Analysis, Predictive Analysis, Data Mining, Data Visualization, Business Intelligence, and AWS MLOps. Has comprehensive industry knowledge across Healthcare and Life Sciences, Entertainment, Retail, and Telecom sectors.
* Experienced in executing data - driven solutions to increase efficiency, accuracy, and utility of internal data processing adept at collecting, analyzing, and interpreting large datasets, developing new forecasting models.
* Strong statistical background using predictive modeling, data processing and data mining, Machine Learning, Deep Learning and NLP to solve complex business problem statements.
* Expertise in NLP tasks such as LLM, BERT, Text Generation, Sentiment Analysis, Summarization, Translation and Question Answering.
* Experience in working with geographically dispersed teams, customer interaction both onsite and offshore.
* Design and execute well-engineered, easy-to-maintain reliable and bug-free code for various company applications in collaboration with other AI engineers, data scientists, programmers and cross functional teams.
* Extensively worked on NLP libraries like NLTK, SpaCy, Scikit-Learn, GenSim, and TextBlob etc.
* Architected and Designed AWS MLOps solutions to host on AWS platform with services like VPC, IAM, EC2,EMR,Lambda ELB, S3, SageMaker.
* Implemented Data Warehousing solutions using Hadoop framework with Apache Spark/Hadoop/Hive with Python programming and Informatica Power Center/Big Data Edition.
* Deep understanding of DL architectures including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers.
* Experienced with Big Data frameworks and migration projects from on-prem to Cloud (AWS,GCP).
* Led and worked on several POC projects before implementing the solution with Big data/Snowflake architectures.
* Worked and designed data models with NoSQL databases and solutions - DynamoDb, Cassandra.
* Designed highly secured infrastructure with AWS best practices.
* Well versed in creating CI/CD Pipelines using DevOps tools like GIT, Jenkins, Docker, Kubernetes, Flask.
* Responsible for creating VPCs for Dev/Test teams, simulating a production environment for web/application servers and database traditional servers in public and private subnets respectively.
* Maintain JIRA projects, workflows, permissions, and users while working in a JIRA 6.x & 7.x environment.
* Created implementation diagrams and data flow mappings for all projects using lucid chart.
* Complete ownership of assignments, action items and issues, and remain accountable for their completion.
* Participate in the planning and design of new reporting and analysis products for internal and external use.
* Planning, execution, monitoring and resource balancing skills, attention to detail and the ability to build and lead effective teams.

# Education:

* Bachelor of Technology in Electrical Engineering, JNTU
* Master’s in Computer Science, Texas A&M University Commerce, TX

# SKILLS and COMPETENCIES:

## Operating Systems - Windows & Linux

* **Cloud Platform - AWS, Google Cloud(GCP), Azure**

## Languages - Pyspark,Python,Spark,Scala & SQL ,Unix

* **Databases - Snowflake**,**BigQuery,Redshift, PostgreSQL, DynamoDB, Cassandra,MongoDB,MySQL, Oracle**

## Reporting Tools - Google DataStudio, Microsoft Power BI, Tableau & SAP BI

* **Data Science - Machine Learning, Deep Learning, NLP & AWS SageMaker MLOps,Vertex AI,Airflow**

## DevOps - CI/CD Pipelines, GIT, Jenkins, Dockers, Kubernetes, Terraform

* **Methodology - Agile (JIRA & Confluence Software)**

# PROFESSIONAL EXPERIENCE:

**DaVita(Remote) Dec 22 - Present**

# Senior Lead AI/ML Engineer

**Environment:** AWS,GCP,Netezza,Python,Docker,Kubernetes,Minio,Airflow,Vertex AI

# Key Achievements:

* Led, implemented and deployed AI/ML models working closely with Data scientists and MLOps team.
* Applied Statistics along with end to end ML engineering including Classification, Clustering and Regression in detecting product anomalies
* Data mining and detective work to fix/find the required patient data for new Peritonitis AI/ML model
* Worked with Data Scientists to create a POC plan for existing ML jobs to run and test the data.
* Build Airflow feature,prediction and implementation dags to train ML models and run them in hybrid mode: on-prem/AWS
* During model training, the training accuracy is 95.02% and test accuracy is 85% after performing some data preprocessing the test accuracy has increased upto 92% and the model is able to detect the class of the image in less than a minute.
* Cleansed 29 million records of data using Python libraries and advanced NLP techniques.
* Created AWS CI/CD pipeline project using Code Commit, Code Build, Code Deploy, Code Artifact & Beanstalk.
* Resolved issues in Data Enrichment process using NLP libraries NLTK, BERT, LLM, SpaCy, GenSim and AWS ML techniques which helps in increasing the accuracy of classification model to 92% and standardization of item attributes
* Created ETL jobs to generate new measures for patients with ESKD and train models to generate a response variable.
* Developed ETL Solutions using IICS Data Integration and Application Integration with huge data volumes
* Use existing ETL frameworks to re-use and rebuild new Netezza SQLs and build Patient measures
* Worked on building driver report and content distribution dag to various end points.
* Dockerize ETL solutions and test them on local docker before deploying solution to the cloud

# HighmarkHealth(Remote) Mar 22 - Dec 22 Solution Architect - Lead AL/ML Engineer

**Environment**: GCP,Ascend, GCS,Teradata Studio,Hive,Pyspark,Azure Databricks,Airflow

# Key Achievements:

* Worked on data migration solutions and testing to move all the event-based clinical data from Teradata to GCP in batches
* Worked with Data Scientists to create a POC plan for existing ML jobs to run and test on the data in GCP.
* Helped the team in creating various Regression, Classification and Clustering algorithms to solve the business problem statements.
* Led a project to assign/update a unique ID for all the clinical sources(event-based) to help Highmark customers with inpatient/outpatient claims.
* The unique ECI ID helps in identifying members back to a map from each of provider subscriptions.
* Building a new dynamic ETL Pyspark Framework to process the demographics data from clinical data in batches with multiple sources
* Used Ascend to build pipelines for data orchestration, integration and observability and ingest data from on-prem into Cloud.
* Conducted daily developer meetings and led/helped the team to build the most flexible solution to be used across.
* Creating a unique MDM lookup service framework to get the ID back from the system and update the Hive database using pyspark on a daily basis and on- need for event based data.
* Worked on POC and ETL/data processing migration from hive to Databricks platform with engineering and data science teams with Airflow being the new scheduler for all clinical data jobs

# DocuSign,Seattle,CA(Remote) Sep 21 - Mar 22 Senior DataScience Engineer

**Environment:**AWS Sagemaker, Snowflake,Airflow,Dbt,Jira,Confluence **,**Python,Spark,Scala,GitHub,Agile **Key Achievements:**

* As a Data Engineer being a member of the Data Science Engineering team, contributed to the Architecture of data-driven solutions for the Data Science team.
* Design and develop solutions for conversion of existing batch processing - Big data volumes, move towards a

dynamic-automated model to serve our customers using Snowflake database**.**Performance tuned SnowSQL to process/cure billions of docusign events.

* Write Python/SQL scripts for data analysis and data matching as per the user’s needs, load into Snowflake to be used by the data science team to build and train their ML models.
* Work with Data Scientists for deeper analysis of events/new features/products offered by Docusign including Sign Rooms.
* Convert huge json blobs and flatten into csv style, using Snow SQL as execution engine to help get answers to complex business use-cases within a few minutes.
* Work on automation of daily weekly jobs using Dbt/Airflow task management tools
* Built a cross product template for event data processing at curation and aggregation levels.
* Work with deep data for lookups before aggregation of event-based data
* Created an ETL batch process with Spark and Scala to calculate the time expended in rooms,envelopes and new product features/ads.
* Work with Product manager/team to review sensitive data before processing for privacy to accommodate CCPA/GDPR requests and/or to purge or not display any PII data in the database and dashboard.

# Lacuna Technologies,Palo Alto,CA - Full Time(Remote)Nov 20 - Jul 21 Senior Data Engineer

Lacuna is a Govtech transportation firm helping cities/airports form a public-private partnership leveraging the Mobility Data Specification(MDS) along with OMF and use the geo-spatial data to know when the public right-of-way/compliance is being violated by e-scooters/bikes and other private micro-mobility operators in the cities and airports.Anticipated to serve LADOT and M-D cities/airports.

**Environment:** AWS S3, Sagemaker, Rekognition, Airflow(AWS MWAA),Aurora Postgres,Google Big Query/Data Studio,Kafka,Scala,Flink, Python,GitHub,Snowflake(POC).

# Key Achievements:

* Worked as a part of the DE team,contributing to the Data Architecture and designing a leave-behind model for customers.
* Leave-behind model is a set of ETL process/data flow and DB to be part of Open Mobility Foundation(OMF),to be used by

all cities/airports served in the future.

* Implemented Spark using Scala and utilizing Data frames/Spark SQL API/Scala API for faster processing of data.
* Led a Snowflake POC initial phases to determine the use cases and stack within Snowflake that will match Lacuna’s needs.
* Worked with AWS MWAA -Ariflow set up from scratch to automate ETL/ELT processes present to calculate aggregated metrics for LADOT and to be re-used for Miami-Dade in future.
* Worked on processing aggregate metrics automation to Google Data Studio(GDS) via data transfer to Google Big Query from Postgres DB.
* Built dashboards using GDS to show hourly trips,trip count within LA city by segment,mapped Vehicles Miles Traveled by neighborhood councils - upgraded it to show a polygon,all within LA jurisdiction boundary.
* Worked on trip-binning data process and running k-anonymization to show only trips more than 2 within a geography+hour.
* Set up external triggers to run the ELT process after event-based data refresh from API jobs - receiving from Uber/Lyft,Lime
* Helped with ELT process upgrade for event-based data in Stream(Flink) and batch to support multiple events-types within each event.
* Worked on building pipelines using AWS Rekognition for Image and text analysis object detection, content moderation within trips from different micro-mobility operators.

## Warner Bros, Burbank CA Feb 18 – Aug 20 Senior Data Engineer/Scrum Master(Consultant)

**Diginet/WWD:** Project to provide a centralized data hub for all the new and old applications or users, data scientists and reporting teams alike seamlessly and almost instantly**.** Architecture is engineered to load data dynamically (both structured and unstructured data) by data and by schedule**.** Using Apache Spark big data tool along with AWS EMR, lambda and Snowflake as final warehouse.

**Environment:** AWS EC2/EMR/RedShift/DynamoDB/S3, Snowflake, Apache Spark, Python, Scala Zeppelin, GitHub, IntelliJ IDEA, Bitbucket, Tableau, Airflow, Jenkins, Confluence, Wiki, JIRA, Agile

## Key Achievements:

* Worked in Agile fashion with daily updates and weekly retro/sprint planning..acting as Scrum Master for few of the projects/applications
* Worked with 200TB and more data volumes of user data and clicks to provide meaningful data to the data science team.
* Built flexible,performant ETL frameworks/pipelines using Python to automate the batch processing and re-use the modules in Stream processing as well.
* Built a file-triggered, automated, schema-dynamic pipeline to process files with no schedule using AWS Lambda,Python.
* Acted as product lead for the Harry Potter related app on iOS/Android coordinating with teams in the UK,US and India and successfully reached the stakeholders and users expectations.
* Implemented the solution using AWS EMR with Apache Spark(python) to process seamlessly and automatically the daily user reports for WB Consumer Intelligence team.
* Processed event files using Kafka queueing /Spark streaming,done aggregations using the window functions, utilizing watermarking feature in Spark streaming.
* Designed a pipeline using Pyspark to filter and scrub user data from Oracle Bluekai received every hour loaded into Snowflake for the data science team to build and train their ML models.
* Designed a Scala based solution to pull data using API from AWS DynamoDB contains User event data along with other App related event data in json, csv and Parquet file formats
* Wrote Apache Spark scripts using Python for data analysis and data matching for POC with – Data bricks and Qu bole.
* Involved in data warehouse transition from AWS Redshift to Snowflake for better data performance and storage.
* Worked on automation of CCPA/GDPR requests implementation to purge any PII data.

## Comtech Tel, Aliso Viejo, CA July 17 – Dec 17 Big Data Engineer(Consultant)

**Specifix:** This project’s end goal is to provide reliable data to end users on the people and business data collected. All the data is collected by web crawling,collecting, processing and analyzing the data using AWS EC2 machines along with Spark as part of the solution.

**Environment:** AWS EC2/EMR/RedShift, Apache Spark, Python, Zeppelin, GitHub, IntelliJ IDEA, Tableau, Looker,Elastic IP,Solr

## Key Achievements:

* Designed and developed solution for conversion of existing batch processing to Big data AWS cloud data processing
* Implemented the solution using AWS EMR with Apache Spark for the People and Business data with daily reports running on 350 Million data set for varied end users.
* Setup and installed Apache Hadoop/Hive/Spark software on AWS EC2 machines and optimized the memory settings for maximum performance with a given combination of Big Data stack and configuration.
* Installed Apache Hadoop/Hive on EC2 machines and did set up a 20-node cluster with different memory combinations to decide the best suited cluster for the same data set.
* Used Amazon EMR for data evaluation, conversion and for result set generation for huge people dataset with 400 attributes.
* Wrote Apache Spark scripts using Python for data analysis and data matching as per the user report needs.
* Worked with deep data, collecting people data and storing it on Elastic search machines for lookups before aggregation.
* Replaced the existing batch process to convert huge json files to csv, to use Apache Spark as execution engine and for java ETL process reducing the process time significantly.
* Used Solr as an interface for people search or analyze modules from the Portal.
* Implemented POC using Apache stack combination with Parquet, Avro, CSV file types to determine the best possible solution for the user

## University Of California, Irvine, CA Oct 16 – Jun 17 Senior Data Engineer(Consultant)

**Compass Student Success(CSS) :** The Compass Project (Comprehensive Analytics for Student Success) at the University of California, Irvine, will bring student learner data into a cohesive, comprehensive, and reportable repository to provide students, faculty, advisors, and campus student services support staff with self-service tools to support student success.

**Environment:** Hive 1.2, Horton Works, Kafka, Cassandra, Apache Spark, Python, Informatica Power Center 9.6.1/Data Quality (IDQ), XML, AWS cloud, JIRA Studio/Agile

## Key Achievements:

* Attended business meetings with end-users/UAT team and student advisor meetings/demos to understand the new front end so data can be analyzed as per changing data sources.
* Integrated data from data sources including Canvas, Registrar/Admissions data using Informatica ETL and stored procedures.
* Converted daily jobs from Tableau ETL to use Informatica Power Center which improved performance for huge fact tables.
* Worked on pulling Canvas files with huge volumes of data fact and dimensions (70 files in Star Schema) daily to local Informatica sftp server from AWS (Amazon Web Services) cloud using Python script.
* Performed data analysis on data both historic and real time using Spark Streaming and SQL using python before loading data into data storage for Compass reporting.
* Provided analysis and type 2 data through Cognos to provide interim student learner activity information to advisors and student services support staff to provide a more complete view of student progress/success.
* Created Probation, Demographic and lower division reports for future use of learning management data to provide “nudges” to students to encourage improved study habits and course materials interaction.
* Worked on Informatica IDQ mappings to write rules for completing data profiling, cleansing

## Illumina, San Diego, CA Mar 16 – Oct 16 Data Engineer(Consultant)

Illumina is aspiring to transform human health by providing products/services using sequencing and Array-based solutions for cancer/genetic research. Illumina sells several very high-throughput DNA sequencing systems known as DNA sequencers. Designed a data analytics platform which provides capabilities to ingest data from various cancer research centers across Illumina into Hive.

**Environment:** Hive 1.2, Hadoop 2.4, Horton works, Azure, Kafka, Zoo keeper, Redis, Storm, YARN, Cassandra**,** Solr**,** Informatica Power Center 9.6.1/Data Quality(IDQ), SAP Exchange, XML, AWS RedShift, Cognos Report Studio 10.2.2.0, Skybot.

## Key Achievements:

* Processed machine generated data into format consumable by Data Scientists using Hive.
* Used Kerberos authentication to connect to Hadoop from Edge node.
* Used HIVE functions like extract and explode to convert data from XML to structured column format.
* Extensively worked on Hive UDFs, external tables and Performance tuning..
* Kafka messaging queue is used to read data into Storm.Storm is used to process data and metadata and load into Cassandra database.
* Created new Informatica SAP/ABAP extractors using SAP Exchange for Informatica for new custom finance contracts (close to real time OLTP data) from source depending on Delta/Full load as needed.
* Used Informatica IDQ mappings to write rules for completing data profiling, cleansing and matching/removing duplicates.
* Dealt with General ledger/finance and product data and reports to fix the amounts and get right annual profits in reporting Worked on updating Illumina website customer and product data which is located on AWS (Amazon Web Services) cloud.

## DaVita , Brea, CA Feb 14 – Jan 16

**ETL Informatica Consultant**

**Fluid Management (FM):** DaVita is a health care/dialysis center dealing with patients who need regular dialysis done. Project FM deals with regular monitoring of patients to generate monthly and quarterly reports respectively with a list of them who are at high life risk based on body fluids and vitals. Designed solution for collecting data and processing data in Hive Warehouse. Cognos reports are built on top of the external tables in HIVE Warehouse.

**Environment**: HIVE 1.2, Hadoop 2.3, AWS, IBM Netezza, Unix shell scripts, Python, Aginity Workbench for Netezza, PowerExchange for Hadoop (Big data Edition, BDE), Informatica Power Center 9.5.1/9.1.0

## Key Achievements:

* Coordinated with business analysts and users to get requirements to build new reports for Fluid Management projects.
* Worked on developing Functional design, detailed level designs and Mapping documents.
* Integrated Hadoop into traditional ETL, accelerating the extraction, transformation and load of massive structured and unstructured data into HDFS.
* Informatica BDE for execution in converting unstructured data to structured format.
* Wrote HiveQL queries to get monthly and yearly data like count of Inpatient/Outpatient, Mode of Dialysis, Check vitals, Patient risk factor to help reporting team users.
* Hands-on experience with installing a Hadoop cluster (using command line interface) consisting of multiple data and name nodes/maintaining it on AWS EC2 server both using Windows and Unix as local OS.
* Co-ordinated with teams on site at multiple locations across US and offshore to get all the pieces of code and data required and made sure all QA/Production release is done as per the project plan.
* Performed Impact analysis on the whole warehouse when major design changes are implemented to improve performance of Fact/Dim jobs due to changing data size.
* Wrote NZSQL scripts to load fact and dimension tables in Netezza using Aginity Workbench.

## OfficeMax, Naperville, IL Feb 13 – Jan 14 ETL Production Support Engineer,Full Time

**NBD Forecasting Phase 2**:**Environment:** Informatica Power center/IDQ 8.6.1, IBM Infosphere DataStage 8.5, PL/SQL, Perl Scripting, Oracle 10g/11g, SQL Server,Toad, UC4 scheduler.

## Key Achievements:

* Adopted existing UNIX scripts and modified as per required data process and for better performance and enhancements.
* Developed and implemented the new flow of Component trades which includes loading the feed file in Perl/Unix shell, applying the business rules through stored procedures and generating the report in Perl for Adhoc reports.
* Re-designed ETL jobs for performance issues as part of Production Support on-call rotation basis.
* Developed custom item jobs to provide special prices on the products to customers with bulk orders all around the year.
* Identified and eliminated duplicates in datasets through IDQ 8.6.1 Edit Distance, Mixed Field matcher. It enables the creation of a single view of customers, helps control costs associated with mailing lists by preventing multiple pieces of mail.
* Created UNIX shell scripts to utilize the DataStage Parallel engine to handle large volumes of data

## Century Link, Monroe, LA Dec 10 – Dec 12

**College Board, Reston, VA Feb 10 – Nov 10**