Harsha Vardhan Sai
AI Data Engineer

Professional Summary:

- 10+ years of expertise in AI/ML systems development, specializing in Large Language Models (LLMs), Generative AI, and deep learning architectures, with extensive experience in building and deploying production-scale AI solutions.
- Advanced proficiency in developing and fine-tuning foundation models, including work with GPT-4, Claude, DALL-E 3, and Stable Diffusion XL. Expert in prompt engineering, context window optimization, and model alignment techniques.
- Deep expertise in GPU infrastructure management using NVIDIA DGX systems, A100/H100 clusters, and CUDA optimization. Experienced in distributed training across multi-GPU environments using frameworks like Horovod and DeepSpeed.
- Mastery of modern AI frameworks including PyTorch 2.0, JAX, and Transformers, with extensive experience in model quantization, pruning, and optimization techniques. Proficient in implementing RLHF (Reinforcement Learning from Human Feedback) pipelines.
- Expert in implementing production-grade RAG (Retrieval-Augmented Generation) systems using vector databases like Pinecone, Weaviate, and Milvus. Experienced in semantic search optimization and embedding model selection.
- Proficient in MLOps and AI infrastructure, utilizing tools like MLflow, Weights & Biases, and DVC for experiment tracking. Expert in implementing CI/CD pipelines for ML models using GitHub Actions and Jenkins.
- Extensive experience with AI model serving platforms including TensorRT, ONNX Runtime, and Triton Inference Server. Skilled in model optimization for both cloud and edge deployment.
- Advanced knowledge of cloud AI services across AWS (SageMaker, Bedrock), Azure (OpenAI Service), and GCP (Vertex AI). Expert in cost optimization and scaling strategies for AI workloads.
- Expert in designing end-to-end pipelines on Google Cloud Platform—leveraging Vertex AI, AI Platform, BigQuery, and Dataflow—to train, deploy, and monitor TensorFlow 2.x (TFX/TF-Serving) and PyTorch models.
- Deep expertise in building custom training pipelines for domain-specific LLMs, including data curation, tokenization strategies, and efficient fine-tuning approaches using techniques like LoRA and QLoRA.
- Proficient in developing AI-powered ETL pipelines using modern data stack including dbt, Airflow, and Dagster. Expert in implementing LLM-based data validation and transformation workflows.
- Strong background in distributed computing with Kubernetes for AI workloads, including experience with KubeFlow and Ray for distributed training and inference. Expert in container orchestration for AI services.
- Expertise in vector database optimization and management, including experience with FAISS, Chroma, and pgvector for efficient similarity search in high-dimensional spaces.
- Advanced knowledge of AI evaluation metrics and testing frameworks, including expertise in developing robust evaluation pipelines for LLM outputs using tools like ROUGE, BLEU, and custom metrics.
- Proficient in developing multimodal AI systems combining text, image, and audio modalities. Experience with vision transformers, diffusion models, and audio generation models.
- Deep understanding of AI safety and responsible AI practices, including implementation of model monitoring, bias detection, and alignment techniques in production systems.
- Expert in developing custom training datasets and data augmentation pipelines, with experience in synthetic data generation using GANs and diffusion models.
- Proficient in optimization techniques for large-scale AI training, including mixed-precision training, gradient accumulation, and efficient memory management strategies.

- Strong experience in developing AI-powered analytics solutions using modern visualization tools like Streamlit, Gradio, and Plotly, integrated with real-time AI insights.
- Expert in building conversational AI systems using state-of-the-art architectures, including implementation of context-aware memory and dynamic knowledge retrieval systems.
- Advanced knowledge of natural language processing pipelines, including expertise in modern tokenization approaches, semantic parsing, and custom vocabulary management.
- Deep expertise in AI system monitoring and observability, using tools like Prometheus, Grafana, and custom telemetry solutions for model performance tracking.
- Proficient in developing hybrid AI architectures combining rule-based systems with neural approaches for enhanced reliability and interpretability.
- Experience in implementing efficient AI-powered search systems using modern information retrieval techniques and hybrid search architectures.
- Strong background in developing custom AI solutions for specific business domains, including financial services, healthcare, and enterprise applications.
- Advanced knowledge of AI model deployment strategies, including canary deployments, A/B testing, and shadow deployment for LLMs and generative models.
- Expertise in developing AI governance frameworks, including model versioning, artifact management, and compliance monitoring systems.
- Proven track record in leading AI initiatives and mentoring teams in adopting best practices for AI development and deployment.
- Strong skills in performance optimization for AI workflows, including profiling, bottleneck identification, and implementation of efficient data pipelines.
- Experience in developing custom attention mechanisms and architectural modifications for specific use cases in transformer-based models.
- Comprehensive understanding of AI infrastructure cost optimization, including strategies for efficient resource utilization and batch processing implementations.

Technical Skills:

AI/LLM Technologies	Foundation Models: GPT-4 Turbo (128k context), GPT-3.5-turbo, Claude 3 (Opus/Sonnet/Haiku), Gemini Pro/Ultra, Llama 2, Mistral 8x7B AI Development: LangChain, LlamaIndex, Semantic Kernel, Hugging Face Transformers,Training & Fine-tuning: LoRA, QLoRA, PEFT, Accelerate Model Deployment: vLLM, TGI, Model Serving Endpoints
Workflow Tools	Workflow Management: Airflow, Dagster, Prefect 2.0,Schema Validation: Pydantic V2, Zod,LLM Workflows: LangChain Agents, Chains, Tools,Process Automation: NiFi, Luigi, Argo Workflows, Google ADK
VECTOR Database	Pinecone, Weaviate, Chroma, FAISS, Mongo DB, Postgres VectorDB
Monitoring and Reporting	Tableau, Power BI, LangChain-integrated dashboards
Hadoop Distribution	Horton Works, Cloudera
Build Tools	Docker, Kubernetes
Programming & Scripting	Python, Scala, JAVA, SQL, Shell Scripting.
Databases	Oracle, MY SQL, Snowflake, BigQuery
Machine Learning & Analytics Tools	Supervised Learning (Linear Regression, Logistic Regression, Decision Tree, Random Forest, SVM, Classification), Unsupervised Learning (Clustering, KNN, Factor Analysis, PCA), Natural Language Processing, Google Analytics Fiddler, Tableau
Version Control	Git, GitHub, SVN, CVS
Operating Systems	Linux, Unix, Mac OS-X, CentOS, Windows 10, Windows 8, Windows 7

Cloud Platform	Google Cloud Platform (GCP), Amazon Web Services (AWS), Microsoft Azure.

Professional Experience:

AAL (American Airlines), Fort Worth, TX **April 2023 to Present**
Role: Gen AI Data Engineer
Responsibilities:

- Developed and deployed SQL agents powered by Vertex AI and PaLM 2 models for natural language querying across BigQuery databases, improving data retrieval efficiency by 65%.
- Implemented RAG architectures using Vertex AI Vector Search and Cloud Storage, processing millions of documents with sub-second latency and 95% accuracy.
- Built conversational workflows using Vertex AI Conversation and BigQuery, enabling seamless query execution and automated response generation.
- Designed scalable vector search systems using Vertex AI Vector Search and Matching Engine, optimizing document retrieval for enterprise-scale operations.
- Integrated PaLM 2 and Gemini models through Vertex AI API to create multi-agent systems for natural language data pipeline interactions.
- Developed Python-based ETL pipelines using Cloud Functions and Dataflow, incorporating ML capabilities for intelligent data transformations.
- Optimized ML-enhanced data pipelines using BigQuery, Dataflow, and Pub/Sub, achieving sub-50ms latency for real-time analytics processing.
- Created custom Cloud Composer DAGs to orchestrate ML-driven ETL tasks, ensuring reliable data processing and model updates.
- Built conversational agents using Dialogflow and Cloud Run, integrating with BigQuery for complex query processing and insights generation.
- Enhanced ML operations using Cloud Functions and GKE, implementing auto-scaling and high-availability for AI applications.
- Utilized Vertex AI prompt engineering to optimize model responses, improving accuracy by 40% for business use cases.
- Implemented comprehensive performance monitoring for ML workflows using Cloud Monitoring and Cloud Logging.
- Designed enterprise-grade ML infrastructure using GKE and Cloud Load Balancing, maintaining 99.99% uptime.
- Developed and deployed TensorFlow 2.x models on Vertex AI Training, leveraging TFX pipelines to automate preprocessing, training, and model validation at scale.
- Optimized TensorFlow training jobs on GCP (using custom machine types on AI Platform) to reduce end-to-end model build time by ~30%.

Environment: Vertex AI, PaLM 2, Gemini, BigQuery, Cloud Storage, Cloud Functions, Dataflow, Pub/Sub, GKE, Cloud Run, Cloud Composer, Vertex AI Vector Search, Dialogflow, Cloud Monitoring, Cloud Logging, Python, SQL, Looker, Data Studio, Docker, Kubernetes, Terraform, Cloud Build, tensorflow2.x, Google ADK.

Cigna healthcare, Bloomfield, CT **Jun 2021 to March 2023**
Role: Sr. Data Engineer
Responsibilities:

- Established a Continuous Delivery pipeline using Docker and GitHub.
- Developed and deployed solutions with Spark and Scala code on a Hadoop cluster running on Google Cloud Platform (GCP).
- Proficient in Google Cloud components, Google Container Builders, GCP client libraries, and Cloud SDKs.

- Utilized Google Cloud Functions with Python to load data into BigQuery for incoming CSV files in GCS buckets.
- Processed and loaded both bound and unbound data from Google Pub/Sub topics to BigQuery using Cloud Dataflow with Python.
- Applied Spark and Scala APIs hands-on to compare the performance of Spark with Hive and SQL, and employed Spark SQL to manipulate Data Frames in Scala.
- Stored data efficiently in GCP BigQuery Target data warehouse, catering to different business teams based on their specific use cases.
- Devised simple and complex SQL scripts to verify Dataflow in various applications.
- Conducted Data Analysis, Migration, Cleansing, Transformation, Integration, Import, and Export using Python.
- Deployed applications to GCP using Spinnaker (RPM-based).
- Architected multiple Directed Acyclic Graphs (DAGs) for automating ETL pipelines.
- Automated feature engineering mechanisms using Python scripts and deployed them on Google Cloud Platform (GCP) and BigQuery.
- Implemented monitoring solutions in Ansible, Terraform, Docker, and Jenkins.
- Automated Datadog Dashboards through Terraform Scripts.
- Hands-on experience in architecting ETL transformation layers and writing Spark jobs for processing.
- Gathered and processed raw data at scale, employing various methods such as scripting, web scraping, API calls, SQL queries, and application development.
- Built TensorFlow-based predictive models for customer segmentation, training on BigQuery ML exports and serving via AI Platform Prediction.
- Orchestrated GCP AI Platform pipelines (Training & Serving) for continuous retraining of TensorFlow models, achieving 92% prediction accuracy on health-risk classification. Processed and loaded both bound and unbound data from Google Pub/Sub topics to BigQuery using Cloud Dataflow with Python.
- Extensive hands-on experience in GCP, BigQuery, GCS bucket, G-cloud function, Cloud Dataflow, Pub/Sub, Cloud Shell, GSUTIL, BQ command-line utilities, Data Proc, and Stack Driver.
- Implemented Apache Airflow for authoring, scheduling, and monitoring Data Pipelines.
- Proficient in machine learning techniques (Decision Trees, Linear/Logistic Regressors) and statistical modeling.
- Worked on Confluence and Jira and skilled in data visualization using Matplotlib and Seaborn libraries.
- Hands-on experience with big data tools like Hadoop, Spark, and Hive.
- Implemented machine learning back-end pipelines with Pandas and NumPy.

Environment: GCP, Bigquery, Gcs Bucket, G-Cloud Function, Apache Beam, Cloud Dataflow, Cloud Shell, Gsutil, Docker, Kubernetes, Apache Airflow, Python, Pandas, Matplotlib, seaborn library, text mining, NumPy, Scikit-learn, Heat maps, Bar charts, Line charts, ETL workflows, linear regression, multivariate regression, Python, Scala, Spark, tensorflow.

AT&T, Dallas, TX **Jan 2019 to May 2021**

Role: Data Engineer

Responsibilities:

- Extensive hands-on experience with the AWS cloud platform, including EC2, S3, EMR, Redshift, Lambda, and Glue.
- Proficient in Spark RDD, Data Frame API, Data Set API, Data Source API, Spark SQL, Spark Streaming, SQL, and MongoDB.
- Developed and deployed data pipelines in cloud environments, particularly on AWS.
- Strong understanding of AWS components, with a focus on EC2 and S3.

- Implemented Spark applications using Python and R, executing Apache Spark data processing projects to manage data from various RDBMS and streaming sources.
- Utilized Apache Spark DataFrames, Spark-SQL, and Spark MLlib extensively, designing and implementing POCs with Scala, Spark SQL, and MLlib libraries.
- Specialized in data integration, employing traditional ETL tools and methodologies to ingest, transform, and integrate structured data into a scalable data warehouse platform.
- Designed and deployed multi-tier applications on AWS, leveraging services such as EC2, Route53, S3, RDS, DynamoDB, SNS, SQS, and IAM, with a focus on high availability, fault tolerance, and auto-scaling using AWS CloudFormation.
- Expertise in Python and Scala, developing user-defined functions (UDFs) for Hive and Pig using Python.
- Extracted data from SQL Server, Teradata, Amazon S3 buckets, and internal SFTP, dumping them into the data warehouse AWS S3 bucket.
- Developed PySpark POCs and deployed them on the Yarn Cluster, comparing the performance of Spark with Hive and SQL/Teradata.
- Created Spark jobs to process data, including instance and cluster creation, and loaded the data into AWS S3 buckets, creating DataMarts.
- Utilized AWS EMR for processing and transforming data, assisting the Data Science team based on business requirements.
- Designed and developed ETL processes in AWS Glue to migrate campaign data from external sources (S3, ORC/Parquet/Text Files) into AWS Redshift.
- Worked on both batch processing and real-time data processing on Spark Streaming using the Lambda architecture.
- Developed Spark applications for cleaning and validating ingested data into the AWS cloud.
- Developed simple to complex MapReduce jobs using Java for processing and validating data.
- Processed and loaded both bound and unbound data from Google Pub/Sub topics to BigQuery using Cloud Dataflow with Python.
- Developed Python code for workflow management and automation using the Airflow tool.
- Developed scripts to load data to Hive from HDFS and ingested data into the Data Warehouse using various data loading techniques.
- Utilized Spark Streaming APIs for real-time transformations and actions.
- Developed preprocessing jobs using Spark DataFrames to flatten JSON documents into flat files.
- Loaded D-Stream data into Spark RDD, performed in-memory data computations to generate output responses.
- Used Kubernetes for the runtime environment of the CI/CD system for building, testing, and deployment.
- Collaborated with the DevOps team to implement NiFi Pipeline on EC2 nodes integrated with Spark, Kafka, Postgres, running on other instances using SSL handshakes in QA and Production Environments.
- Built Informatica mappings, sessions, and workflows, managing code changes through version control in Informatica.

Environment: Spark, Spark-Streaming, Spark SQL, AWS EMR,Scala, MapReduce, HDFS, Hive, Pig, Apache Kafka, Sqoop, Python, Pyspark, Shell scripting, Linux, MySQL Oracle Enterprise DB, Big query, SOLR, Jenkins, Eclipse, Dataflow, Oracle, Git, Oozie, Tableau, MySQL, Soap, Cassandra and Agile Methodologies.

Target, Minneapolis, MN **Jan 2017 to Dec 2018**
Role: Data Engineer
Responsibilities:

- Operated within an Agile environment, utilizing Rally tool for managing user stories and tasks.
- Implemented ad-hoc analysis solutions using Data Lake Analytics/Store and HDInsight.

- Implemented Apache Sentry to control access to Hive tables on a group level.
- Possessed expertise in MapReduce programming using Java, PIG Latin Scripting, and Distributed Application and HDFS.
- Utilized Tidal enterprise scheduler and Oozie Operational Services for coordinating the cluster and scheduling workflows.
- Integrated various Azure services such as Azure Data Factory, Azure Data Lake, Azure Data Warehouse, Azure Active Directory, Azure SQL Database, and Web Apps with AWS services to leverage the benefits of both cloud platforms and process unstructured data.
- Designed and implemented Kafka clusters by configuring Topics in all environments.
- Developed multiple Tableau dashboards catering to various business needs.
- Implemented Partitioning, Dynamic Partitions, and Buckets in HIVE for efficient data access.
- Architected and implemented medium to large-scale BI solutions on Azure using Azure Data Platform services, including Azure Data Lake, Data Factory, Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Databricks, and NoSQL DB.
- Utilized AVRO format for entire data ingestion for faster operation and less space utilization.
- Designed SSIS Packages for ETL operations, extracting, transferring, and loading existing data into SQL Server from different environments for SSAS cubes (OLAP).
- Ingested data into one or more Azure Services (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processed the data in Azure Databricks.
- Developed visualizations and dashboards using PowerBI.
- Implemented Composite server for data virtualization needs and created multiple views for restricted data access using a REST API.
- Exported analyzed data to relational databases using Sqoop for visualization and report generation for the BI team using Tableau.
- Developed Apache Spark applications for data processing from various streaming sources.
- Exposure to Spark, Spark Streaming, Spark MLlib, Snowflake, Scala, and creation of Data Frames handled in Spark with Scala.
- Developed data pipelines using Spark, Hive, Pig, Python, Impala, and HBase to ingest customer data.
- Converted Hive/SQL queries into Spark transformations using Spark RDDs, Python, and Scala.
- Queried and analyzed data from Cassandra for quick searching, sorting, and grouping through CQL.
- Joined various tables in Cassandra using Spark and Scala, running analytics on top of them.
- Brought data from various sources into Hadoop and Cassandra using Kafka.
- Migrated on-premise data (Oracle/SQL Server/DB2/MongoDB) to Azure Data Lake and Stored (ADLS) using Azure Data Factory (ADF V1/V2).
- Created action filters, parameters, and calculated sets for preparing dashboards and worksheets using PowerBI.
- Developed Spark applications using Spark-SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats to analyze and transform data for uncovering insights into customer usage patterns.

Environment: MapR, Map Reduce, HDFS, Hive, pig, Impala, Kafka, Cassandra, Spark, Scala, Azure (SQL, Databricks, Data lake, Data Storage, HDInsight), Java, SQL, Tableau, PIG, Zookeeper, Sqoop, Kafka, Teradata, Power BI.

Menlo Technologies, Hyderabad, India **Jan 2014 to October 2016**
Role: ETL Developer
Responsibilities:

- Accountable for the development, support, and maintenance of ETL (Extract, Transform, and Load) processes using Informatica PowerCenter.
- Designed and implemented numerous ETL scripts utilizing Informatica and UNIX shell scripts.
- Analyzed source data from Oracle, Flat Files, and MS Excel, collaborating with the data warehouse team to develop a Dimensional Model.
- Established FTP, ODBC, and Relational connections for sources and targets.
- Implemented Slowly Changing Dimension Type 2 methodology to access the complete history of accounts and transaction information.
- Proficient in crafting complex SQL queries, unions, multiple table joins, and experience with Views.
- Demonstrated expertise in database programming with PL/SQL, encompassing Stored Procedures, Triggers, and Packages.
- Scheduled sessions and batches on the Informatica Server using Informatica Server Manager.
- Executed and validated test cases for data transformations within Informatica.
- Created JIL scripts and scheduled workflows using CA Autosys.
- Utilized SQL scripts/queries for thorough data verification at the backend.
- Executed SQL queries, stored procedures, and performed data validation as part of backend testing.
- Utilized SQL to test various reports and ETL job loads in development, testing, and production.
- Developed UNIX shell scripts to orchestrate the process flow for Informatica workflows, handling high-volume data.
- Prepared test cases based on the Functional Requirements Document.

Environment: Informatica Power Center 9.x, Oracle 11g, SQL plus, PL/SQL, Oracle, SQL Developer, UNIX.