SANDEEP KUMAR
AI ML Data Engineer
Contact #
E-Mail:

==========

SUMMARY:

- Experienced Sr AI/ML Data Engineer and Data Specialist with 11+ Years extensive expertise in designing, developing, and deploying scalable machine learning models and data pipelines. Skilled in end-to-end CI/CD automation using Jenkins, GitLab CI, Argo Workflows, and AWS cloud services. Proficient in building and fine-tuning transformer-based NLP models (BERT, GPT, LLaMA) leveraging PyTorch, Hugging Face, and OpenAI APIs. Adept at developing RAG pipelines and semantic search solutions with LangChain, Pinecone, and Weaviate, alongside containerized deployments using Docker, Kubernetes, and Serverless frameworks.
- Expertise in large-scale data processing with Apache Spark, Hadoop, Azure Data Factory, and Databricks. Skilled in database management and optimization with PostgreSQL, MongoDB, AWS Databricks & pySpark with Snowflake Milvus, and Chroma vector databases.
- Experienced in MLOps, cloud infrastructure management (AWS, Azure, GCP), and Agile project delivery with tools such as JIRA and Confluence. Proven ability to collaborate across teams to deliver robust AI-powered applications, including chatbots, predictive analytics, and document retrieval systems, enhancing business KPIs and operational efficiencies.
- Designed and deployed scalable end-to-end machine learning and AI solutions leveraging PyTorch, TensorFlow, and Hugging Face Transformers.
- Led cloud-native development and deployment using AWS, Azure, Docker, Kubernetes, and Serverless Framework to enhance system scalability and cost efficiency.
- Built and automated ETL and big data pipelines using Spark, Databricks, Snowflake, and Azure Data Factory, enabling reliable multi-source data ingestion and transformation.
- Developed and integrated AI-powered microservices and APIs (RESTful & GraphQL) using FastAPI, Flask, and Node.js for enterprise-grade applications.
- Implemented semantic search and embeddings pipelines with Pinecone, Weaviate, and LangChain for intelligent knowledge management and retrieval.
- Engineered and optimized MLOps pipelines with Jenkins, GitLab CI/CD, GitHub Actions, and MLflow for continuous integration, testing, and deployment of AI models.
- Designed data analytics dashboards and visualizations using Power BI, Tableau, and Seaborn to support executive decision-making.
- Managed cloud infrastructure and resources using AWS Lambda, ECS, API Gateway, and CloudFormation, improving performance and reliability.
- Oversaw project management and delivery using Agile/Scrum methodologies, including RTM, sprint planning, testing, and defect tracking.
- Mentored junior engineers and promoted best practices in AI/ML development, data governance, and CI/CD automation.

TECHNICAL SKILLS:

Machine Learning & AI	PyTorch, TensorFlow, Hugging Face Transformers, OpenAI API, BERT, GPT-2/3/4, LLaMA, T5, LoRA, PEFT, CNN, RNN, LSTM, Transformer architectures
MLOps & CI/CD	Jenkins, GitLab CI, GitHub Actions, Argo Workflows, Docker, Kubernetes, AWS ECS, Serverless Framework, AWS Lambda, AWS API Gateway, CloudFormation
Data Engineering & ETL	Apache Spark, Hadoop, Azure Data Factory (ADF), Azure Databricks, Snowflake, DBT, FiveTran, Alteryx, SSIS, Parquet files ingestion, JSON parsing

Databases & Storage	PostgreSQL, MongoDB, Oracle, Microsoft SQL Server, Redis, Pinecone, Weaviate, Milvus, Chroma, Azure Data Lake Storage
APIs & Microservices	RESTful APIs, GraphQL APIs, FastAPI, Flask, Node.js microservices
NLP & Semantic Search	LangChain, SpaCy, NLTK, Custom Tokenizers, Semantic Search (Pinecone, Weaviate), Embeddings pipelines
Cloud Platforms	AWS (EC2, Lambda, ECS, API Gateway), Azure (Data Factory, Databricks), Google Cloud Platform (GCP, Kubernetes)
Visualization & Reporting	React.js, Vue.js, Angular, Seaborn, Tableau, Power BI, Report Builder
DevOps & Infrastructure	Docker, Kubernetes, Serverless Framework, Infrastructure as Code (CloudFormation), Monitoring, Logging
Business Tools & Agile	JIRA, Confluence, SharePoint, ALM, Requirement Traceability Matrix (RTM), Agile Methodologies
CRM & ERP	Microsoft Dynamics 365, Power Platform
Other Tools	SSRS, Linked Reports, Test Strategies, Test Plans, Test Cases, Defect Tracking

PROFESSIONAL EXPERIENCE:

Client: BNY Mello, USA||Jun 2024 - Till Date
Role: AI ML Data Engineer
Responsibilities:

- Designed and implemented end-to-end CI/CD pipelines to automate training, testing, and deployment of AI/ML models using tools like Jenkins, GitLab CI, and Argo Workflows.
- Designed and implemented machine learning models for customer segmentation and predictive analytics using both supervised and unsupervised techniques.
- Utilized Apache Spark and Hadoop to process and analyze terabytes of structured and unstructured data efficiently.
- Deployed ML models in production environments using tools such as Docker and MLflow, improving business KPIs
- Designed and deployed scalable RAG pipelines using LangChain and OpenAI APIs, connected to Pinecone and Weaviate for real-time semantic search.
- Developed microservices in Node.js and Python to support AI inference and document ingestion pipelines.
- Built intuitive UIs using React and integrated with GraphQL APIs to visualize model outputs.
- Integrated GPT and BERT models via Hugging Face and OpenAI APIs for classification, Q&A, and summarization.
- Orchestrated Docker-based deployment pipelines with GitHub Actions and AWS ECS.
- Created RESTful APIs for NLP-powered applications using FastAPI and Flask.
- Built embeddings pipelines and stored vector representations in Milvus and Chroma.
- Developed a web app in Vue.js to display AI insights from LLMs, enhanced with interactive charts.
- Managed PostgreSQL and MongoDB schemas for AI data storage and processing.
- Developed and fine-tuned Transformer-based NLP models (BERT, GPT-2/3, T5) using PyTorch and Hugging Face for text classification and question answering systems.
- Deployed LLMs via FastAPI and Docker with real-time inference support on AWS EC2/GPU instances.
- Implemented fine-tuning pipelines leveraging techniques like LoRA, PEFT, and Parameter-Efficient Tuning for domain-specific datasets.
- Used optimization strategies (gradient clipping, warm-up scheduling, mixed precision training) to improve model convergence and training speed.
- Designed robust data pipelines for NLP tasks using spaCy, NLTK, and custom tokenizers.
- Contributed to the training and deployment of a multi-task LLM supporting text summarization and document retrieval.

- Developed and deployed machine learning models for classification, regression, and clustering using supervised and unsupervised learning techniques.
- Designed and implemented deep learning architectures (CNNs, RNNs, LSTMs, Transformers) for computer vision, NLP, and time-series applications.
- Built and trained custom deep learning models using PyTorch, leveraging features like torch.nn, DataLoader, and autograd for dynamic computation.
- Developed and deployed a microservices-based AI assistant using GPT-4 via OpenAI API, integrated with a RAG pipeline powered by LangChain and Pinecone.
- Designed and implemented scalable front-end interfaces in React.js and backend logic using Node.js and FastAPI.
- Created custom NLP models using BERT and fine-tuned LLaMA for domain-specific classification tasks.
- Managed cloud infrastructure on AWS and integrated CI/CD pipelines with Docker containers and GitHub Actions.
- Optimized vector similarity search using Weaviate and Milvus for real-time document retrieval.
- Built responsive web applications using Angular and Vue.js for real-time data analytics dashboards.
- Developed REST and GraphQL APIs in Python and Java; orchestrated them via Docker containers in a Kubernetes cluster on GCP.
- Implemented secure, scalable user authentication and role-based access control.
- Modeled data for PostgreSQL and MongoDB and designed high-availability NoSQL schema for Chroma vector DB integration.
- Contributed to MLOps pipelines for deploying transformer-based models using Hugging Face and TensorFlow.
- Deployed the app with Docker on AWS Lambda using Serverless Framework.
- Developed and optimized scalable ETL pipelines using Databricks on AWS to process large datasets efficiently.
- Managed and monitored Databricks clusters to ensure optimal performance and cost-effectiveness.
- Integrated Databricks with AWS services such as S3, Redshift, and Glue for seamless data ingestion and storage.
- Collaborated with data engineers and data scientists to build and deploy machine learning models using Databricks notebooks.
- Automated workflows using Databricks Jobs and Delta Live Tables for continuous data processing and analytics.
- Designed and developed distributed data processing pipelines using PySpark to transform and analyze large-scale datasets.
- Integrated PySpark workflows with other data sources including Kafka, S3, and relational databases.
- Developed and maintained data warehouse solutions on Snowflake to support business intelligence and reporting needs.
- Created and managed Snowflake stored procedures, views, and tasks for automated data workflows.
- Integrated Snowflake with ETL tools and cloud platforms to automate data ingestion and transformation.

Environment: Python, JavaScript (Node.js, React.js, Vue.js, Angular), SQL, Java, PyTorch, TensorFlow, Hugging Face Transformers, LoRA, PEFT, spaCy, NLTK, Scikit-learn, Docker, Kubernetes, Jenkins, GitLab CI, GitHub Actions, Argo Workflows, MLflow, Serverless , AWS (EC2, ECS, Lambda, API Gateway, CloudFormation, Apache Spark, Hadoop, PostgreSQL, MongoDB, Redis, Pinecone, Weaviate, Milvus, Chroma, FastAPI, Flask, GraphQL, RESTful API , React.js, Vue.js, Angular, LangChain, OpenAI APIs, GPT, BERT, Tableau, Seaborn, JIRA, Confluence, SharePoint

Client: Huntington Bank, USA||Jun 2021 - Dec 2023
Role: AI ML Data Engineer
Responsibilities:

- Designed and deployed scalable ML pipelines using PyTorch, TensorFlow, and HuggingFace Transformers.
- Implemented semantic search and embeddings pipelines with Pinecone, Weaviate, and LangChain for enterprise knowledge management.
- Led cloud-native deployments on AWS and Azure using Docker, Kubernetes, and Serverless Framework, improving infrastructure scalability and reliability.
- Built ETL workflows and big data pipelines using Spark, ADF, Databricks, and Snowflake, supporting multi-source ingestion and data quality automation.

- Developed RESTful and GraphQL APIs, microservices, and integrated them into web applications using FastAPI, Flask, Node.js, React.js, and Vue.js.
- Implemented data analytics dashboards with Power BI, Tableau, and Seaborn to provide actionable insights to business stakeholders.
- Oversaw project management activities, including RTM, Agile ceremonies, test plans, and defect tracking.
- Engineered AI models (GPT, BERT, T5) for natural language processing and predictive analytics.
- Managed cloud infrastructure using AWS Lambda, ECS, API Gateway, and CloudFormation, optimizing deployment speed and cost.
- Automated CI/CD pipelines with Jenkins, GitLab CI, and GitHub Actions, ensuring reliable software delivery.
- Designed database solutions using PostgreSQL, MongoDB, Oracle, Redis, and integrated them with cloud storage solutions.

Environment: Python, R, SQL, PyTorch, TensorFlow, Keras, Hugging Face Transformers (GPT, BERT, T5), LangChain, Pinecone, Weaviate, Apache Spark, Databricks, Azure Data Factory (ADF), Snowflake, Airflow, Kafka, Docker, Kubernetes, Serverless Framework, Jenkins, GitLab CI/CD, GitHub Actions, MLflow, AWS (Lambda, ECS, S3, API Gateway, CloudFormation, SageMaker), Azure (Machine Learning, Synapse, Data Factory), FastAPI, Flask, Node.js, React.js, Vue.js, PostgreSQL, MongoDB, Oracle, Redis, Power BI, Tableau, Seaborn, Linux, Git, Jira, Agile / Scrum

Client: Northern Trust, USA||Sep 2019 - May 2021
Role: AI ML Data Engineer
Responsibilities:

- Designed and implemented machine learning solutions for enterprise applications, improving operational efficiency
- Led AI/ML initiatives including NLP, computer vision, and speech recognition services using cloud-based AI platforms.
- Extensively used Pandas, NumPy, Seaborn, Matplotlib, Scikit-learn, SciPy and NLTK in R for developing various machine learning algorithms.
- Used R programming language for graphically critiquing the datasets and to gain insights to interpret the nature of the data.
- Researching on Deep Learning to implement NLP
- Clustering, NLP, Neural Networks. Visualized and presented the results using interactive dashboards.
- Involved in the transformation of files from GITHUB to DSX.
- Involved in the execution of CSV files in Data Science Experience.
- Built and maintained MLOps pipelines to ensure continuous integration, testing, deployment, and monitoring of models.
- Collaborated with cross-functional teams to define AI strategy and translate business requirements into technical solutions.
- Mentored junior engineers and promoted best practices in coding, testing, and deployment.
- Conducted statistical analysis and predictive modeling to support business decision-making.
- Developed AI services using Python, Spark, and R for large-scale data processing and model training.
- Integrated AI services for speech, text, and vision applications into production systems.
- Optimized data pipelines and implemented data governance practices in DataOps workflows.

Environment: Python, R, SQL, Apache Spark, TensorFlow, PyTorch, Scikit-learn, Keras, Pandas, NumPy, OpenCV, NLTK, spaCy, Hugging Face Transformers, Azure Machine Learning, AWS SageMaker, Google Cloud AI Platform, Docker, Kubernetes, MLflow, Airflow, Git, Jenkins, REST APIs, FastAPI, Flask, Power BI, Tableau, Databricks, Kafka, Hadoop, DataOps, MLOps, CI/CD, Linux

Client: Catholic Health Initiatives, USA||Sep 2017 - Aug 2019
Role: ML Data Engineer
Responsibilities:

- Converted data from PDF to XML using python script in two ways i.e. from raw xml to processed xml and from processed xml too.CSV files.
- Developing a generic script for the regulatory documents.
- Used python Element Tree(ET) to parse through the XML which is derived from PDF files.
- Data which is stored in sqlite3 datafile(DB.) were accessed using the python and extracted the metadata,tables,and data from tables and converted the tables to respective CSV tables.
- Used the XML tags and attributes to isolate headings,side-headings,and subheadings to each row in CSV file.
- Used Text Mining and NLP techniques find the sentiment about the organization.
- Deployed a spam detection model and performed sentiment analysis of customer product reviews using NLP techniques.
- Developed and implemented predictive models of user behavior data on websites, URL categorical, social network analysis, social mining and search content based on large-scale MachineLearning.
- Developed predictive models on large-scale datasets to address various business problems through leveraging advanced statistical modeling, machine learning,and deep learning.
- The major part is like being a part of the project, importing the converted CSV file to Confidential internal API which is InfoSphere Information Governance Catalog
- Used Beautiful Soup for web scraping (Parsing the data)
- Developed the code to capture the description which comes under headings of index section to the description column of CSV row.
- Used some other python libraries like PDFMiner, PyPDF2, PDFQuery, Sqlite3.
- Converted the uni-code to a nearest possible string (ASCII value) using Uni-decode module.
- Adding a column to each CSV row which gives the parent Index number of the given row.

Environment: R Studio, AWS S3, NLP, EC2, Neural networks, SVM, Decision trees, MLbase, ad-hoc, MAHOUT, NoSQL, Pl/SQL, MDM, MLLib & Git.

Client: Kroger - Cincinnati, Ohio||Oct 2013 - Aug 2017
Role: ML Data Engineer
Responsibilities:

- Designed and implemented scalable ETL/ELT pipelines using Spark + Airflow on AWS (Glue/S3/EMR).
- Built automated ML pipelines for training, validation, deployment using MLflow/SageMaker/Kubeflow.
- Developed feature engineering workflows, integrated with feature store (Feast/Tecton).
- Implemented data quality and model drift monitoring using Great Expectations & EvidentlyAI.
- Optimized data lake storage using Delta Lake / Iceberg for ACID transactions and time travel.
- Collaborated with data scientists to productionize models (batch & streaming).
- Built distributed data pipelines using Spark, Kafka, and Airflow.
- Implemented ML model deployment workflows on AWS/GCP with CI/CD automation.
- Developed scalable feature extraction pipelines for real-time predictions.
- Migrated on-prem data warehouse to cloud (Snowflake/BigQuery/Redshift).

Environment: Python, SQL, Scala, Bash, Pandas, NumPy, PySpark, Scikit-learn, TensorFlow, PyTorch, Apache Spark, Hadoop, Hive, Kafka, Delta Lake, Iceberg, AWS (S3, Glue, EMR, Lambda, Redshift, SageMaker), GCP (BigQuery, Dataflow, Vertex AI), Azure (Databricks, Data Factory, Snowflake, PostgreSQL, MySQL, MongoDB, DynamoDB