Pranjal Ranjan

+1 (540) 558 5918 pranjalranjan@vt.edu /in/pranjalranjan /pranjalranjan299

Education

Virginia Tech (VT) **Aug 2022 – Present (Exp. Aug 2024)**

Master of Science - Computer Engineering (Specialization - Data Science) *Virginia, USA*

Coursework: Big Data Engineering, Computer Vision, Natural Language Processing, Advanced ML

Awarded Pratt Fellowship - GPA: 4.0/4.0

Experience

Propulsion AI **June 2024 – Present**

Data Engineer (Intern) *San Francisco, CA*

- Engineered a robust data processing pipeline using Apache Spark, Pandas, and Dask for distributed computing, capable of efficiently handling diverse user-uploaded datasets (CSV, JSON, Parquet, Avro) up to 500GB.
- Developed a flexible LLM integration framework supporting multiple open-source models (including Llama 3, Gemma, Mixtral, and Phi-2), implementing efficient fine-tuning techniques such as LoRA, QLoRA, and P-tuning v2.
- Architected a scalable model serving solution using FastAPI and Docker, deploying fine-tuned LLMs as RESTful APIs on Kubernetes. Implemented model versioning, caching mechanisms, and automatic scaling based on inference demands.

UBS **Aug 2020 – Jul 2022**

Data Engineer *Mumbai, India*

- Architected a highly scalable data lake using Azure Data Lake Storage (ADLS) Gen2, Azure Databricks, & Azure Data Factory, efficiently storing, processing, and managing petabyte-scale datasets from databases like Oracle & SQL Server.
- Engineered high-performance data pipelines utilizing Apache Spark (PySpark and Scala), and Azure Synapse Analytics, optimizing data processing, transformation, and analysis through advanced partitioning and compression techniques.
- Automated data pipeline workflows using Azure Data Factory and Bash scripting on Azure Virtual Machines. Implemented data validation and cleaning routines using Pandas and PySpark to ensure data quality and consistency.
- Designed a real-time data streaming architecture using Azure Event Hubs for high-throughput ingestion, Stream Analytics for complex event processing, and Apache Kafka for scalable event streaming.
- Developed secure and scalable data APIs using Azure Functions, Azure API Management, Flask, and SQLAlchemy, implementing robust authentication and authorization mechanisms with Azure Active Directory and OAuth2.
- Created visually compelling and interactive dashboards and reports using Tableau and Power BI, integrating with Azure Synapse Analytics, Azure Databricks, and Azure SQL Database to enable self-service analytics for business users.
- Implemented enterprise-grade data governance, lineage tracking, and security measures using Azure Purview, Azure Key Vault, and Azure Policy, ensuring compliance with industry regulations and data privacy standards.

Projects

Sales Reporting System on Azure | *Azure Data Factory, Azure Databricks, Azure Synapse Analytics, Power BI*

- Engineered a cloud-native sales reporting ecosystem leveraging Azure SQL Database for structured data storage, Azure Blob Storage for file management, and Flask-based RESTful APIs for efficient data processing and metric calculations.
- Implemented an AI-powered natural language querying system using the Together API, coupled with dynamic data visualization through Chart.js and Power BI Embedded, enabling business users to derive actionable insights.

Text-to-SQL Generator | *Python, PyTorch, Hugging Face, Large Language Models*

- Developed and fine-tuned large language models such as T5 and BART for converting natural language to SQL queries.
- Implemented strategic data preprocessing techniques to improve query generation performance, including token preprocessing and adding table context to the input as a knowledge source.

Certifications

- Microsoft Certified: Azure Data Engineer Associate
- TensorFlow Developer Professional Certificate, Coursera
- TensorFlow Data & Deployment Specialization, Coursera
- Deep Reinforcement Learning Nanodegree, Udacity
- Machine Learning Engineer Nanodegree, Udacity
- Microsoft Certified: Azure Data Fundamentals
- Microsoft Certified: Azure AI Fundamentals
- Microsoft Certified: Azure Fundamentals

Technical Skills

Programming : Python | Java | Scala | SQL | PySpark | R | C++ | Bash | JavaScript
Big Data : Apache Spark | PySpark | Apache Kafka | Apache Hadoop | Apache Flink | Apache NiFi | Apache Beam | Databricks
Databases : PostgreSQL | Hive | Snowflake | AWS Redshift | Google BigQuery | MySQL | MongoDB | Cassandra
Cloud : Microsoft Azure | AWS | GCP | Azure Data Factory | Azure Data Lake Storage | Azure Synapse Analytics
CI/CD : GitLab | Azure DevOps | Jenkins | AutoSys | Airflow | Docker | Kubernetes
Data Science : Keras | TensorFlow | PyTorch | Scikit-Learn | Pandas | NumPy | Matplotlib | Seaborn | Jupyter