Sampreethi Bokka

Email: sampreethi486@gmail.com
Ph#: 945-296-0831

Professional Summary:

- Over **6+ Years** of experience as a **Data Engineer**, including profound expertise and experience in statistical data analysis, such as transforming business requirements into analytical models, designing algorithms, and strategic solutions that scale across massive volumes of data.
- Experience with **Spark Core, Spark SQL, Spark MLlib, Spark Graph X and Spark Streaming** for processing and transforming complex data using in-memory computing capabilities written in **Scala.**
- Experienced in writing **Spark** Applications in **Scala** and **Python (PySpark).**
- Experience in writing **MapReduce** programs in java for data cleansing and preprocessing.
- Experience in writing scripts using **Python API, PySpark API and Spark API** for analyzing the data.
- Experienced in **ETL concepts,** building **ETL solutions** and **Data modeling.**
- Specialized in building robust ETL/ELT pipelines, data lakes, and warehousing solutions using tools like **Azure Data Factory, AWS Glue**, and **Google Dataflow.**
- Experience in **Data Visualization** with **Tableau, creating: Line and scatter plots, Bar Charts, Histograms, Pie charts, Dot charts, Box plots, Time series, Error Bars, Multiple Chart types, Multiple Axes, subplots, etc.**
- Experience in integrating Apache Kafka and creating Kafka pipelines for real-time processing.
- Experience working in **Azure Cloud, Azure DevOps, Azure Data Factory, Azure Data Lake Storage, Azure Synapse Analytics, Azure Analytical Services, Azure Cosmos NoSQL DB, Azure HD Insight Big Data Technologies (Hadoop and Apache Spark),** and **Data bricks.**
- Proficient in leveraging modern cloud platforms such as **AWS, Azure, and GCP** to build scalable, efficient, and secure data architectures.
- Proficient in **PyTorch, TensorFlow, FastAPI,** and scalable model deployment with **MLOps** tools across cloud environments.
- Strong understanding of **Generative AI, machine learning pipelines**, and integrating **AI** capabilities into software solutions.
- Machine Learning **Engineer** with experience in building and deploying end-to-end ML systems using LLMs, GANs, and transformer architectures across text, image, and multimodal domains.
- Experience developing **Airflow** workflows for scheduling and orchestrating the **ETL** process.
- Experience working with **GitHub/Git** source and version control systems.
- Experience working with **SQL, PL/SQL and NoSQL** databases like Microsoft **SQL Server, Oracle, HBase and Cassandra.**
- Experience in **Software development life cycle (SDLC)** to develop the application using **Agile and Waterfall** methodologies.
- Strong skills in analytical, presentation, communication, problem solving with the ability to work independently as well as in a team and had the ability to follow the best practices and principles defined for the team.

Technical Skills:

Databases	Snowflake, AWS RDS, Teradata, MySQL, Oracle, Microsoft SQL, PostgreSQL.
NoSQL Databases	MongoDB, Hadoop HBase and Apache Cassandra.
Machine Learning/AI	RAG, MCP, LLMS (LARGE LANGUAGE MODELS), **AI**/ML (ARTIFICIAL INTELLIGENCE), TensorFlow, PyTorch, Fast API
Languages	Python, SQL, Scala, MATLAB.
Cloud Technologies	Azure, AWS, GCP
Data Formats	**CSV, JSON**
Querying Languages	SQL, NoSQL, PostgreSQL, MySQL, Microsoft SQL
Integration Tools	Jenkins, CI/CD
Scalable Data Tools	Hadoop, Hive, Apache Spark, Pig, MapReduce, Sqoop.
Reporting & Visualization	Tableau, Matplotlib.
Operating Systems	Red Hat Linux, Unix, Windows, macOS.

Professional Experience:

Client: Jefferson Bank, San Antonio, TX. **Duration; Nov 2024 – Till Date**
Role: Data Engineer/AI/ML
Responsibilities:

- Interacted with clients to gather business and system requirements which involved documentation of processes based on the user requirements.
- Worked with **Chat GPT and Chat GPT API** to leverage its capabilities with several applications.
- Built production-grade, dynamic-response **chatbots using the LangChain** Framework, leveraging LLMs for real-time, context-aware answers across high-volume environments.
- Involved in developing end-to-end machine learning pipelines, fine-tuning Large Language Models (LLMs), and applying **GenAI** for real-world automation and insight generation.
- Developed internal **AI** guidelines and secured LLM usage for sensitive data prompts.
- Implemented **AI**-enhanced chatbot workflows for support escalation and ticket classification.
- Developed **Spark** scripts by writing custom RDDs in **Scala** for data transformations and perform actions on RDDs.
- Developing **Spark** programs with **Python**, and applied principles of functional programming to process the complex structured data sets.
- Used **MapReduce and Spark** with **Scala** for performing operations like Clickstream Analysis and to perform Analysis on batch Data.
- Developed various **Python** scripts to find vulnerabilities with SQL Queries by doing SQL injection, permission checks and analysis.
- Developed **PySpark** Data Ingestion framework to ingest source claims data into **HIVE tables** by performing Data cleansing, Aggregations and applying De-dup logic to identify updated and latest records.
- Worked on large-scale data using **Spark and MapReduce.**
- Design and develop **Tableau visualizations** which include preparing Dashboards using **calculations, parameters, calculated fields, groups, sets and hierarchies.**
- Design and development of **ETL** processes using **Informatica ETL tool** for dimension and fact file creation.
- Worked on a direct query using **Power BI** to compare legacy data with the current data and generated reports and stored and dashboards.
- Developed and Configured **Kafka brokers** to pipeline server logs data into spark streaming.
- Used **Apache Kafka** to aggregate web log data from multiple servers and make them available in Downstream systems for Data analysis and engineering type of roles.
- Extract Transform and Load data from Sources Systems to **AzureDataStorage** services using a combination **of Azure Data Factory, T-SQL, Spark SQL, and U-SQL Azure Data Lake Analytics.**
- Implemented Copy activity, Custom **Azure Data Factory** Pipeline Activities.
- Primarily involved in **Data Migration** using **SQL, SQL Azure, Azure Storage, and Azure Data Factory, SSIS, PowerShell.**
- Implemented **Apache Airflow** for authoring, scheduling and monitoring Data Pipelines.
- Created and modified several database objects such as Tables, Views, Indexes, Constraints, Stored procedures, Packages, Functions and Triggers using **SQL and PL/SQL.**
- Followed **agile** methodology and involved in daily **SCRUM** meetings, sprint planning, showcases and retrospective.
- Participated in the status meetings and status updating to the management team.

Environment: Spark, Scala, PySpark, Python, MapReduce, ETL, Tableau, Power BI, Apache Kafka, Azure, Star Schema, Snowflake, GitHub, Apache Airflow, Hive, HBase, Jira, SQL, PL/SQL, Agile and Windows.

Client: Geico Insurance, Chevy Chase, MD. **Duration: Sep 2023 – Oct 2024**
Role: Data Engineer
Responsibilities:

- Worked with the business users to gather, define business requirements and analyze the possible technical solutions.
- Performed **Spark jobs** with the **Spark core and Spark SQL** libraries for processing the data.
- Implemented **Spark** application with **Scala** Programming Language
- Worked on developing custom **MapReduce** programs and **User Defined Functions (UDFs)** in **Hive** to transform the large volumes of data with respect to business requirements.

- Performed Data Analysis, Data Migration, Data Cleansing, Transformation, Integration, Data Import, and Data Export through **Python**.
- Involved in designing and implementing scalable **ETL pipelines** to process a variety of data types (structured, unstructured), file formats (**JSON, CSV, text delimited…**).
- Developed interactive dashboards and reports using **Power BI** for day-to-day business decision making and strategic planning needs
- Implemented data ingestion and handling clusters in real-time processing using **Kafka**.
- Involved in Data Migration using **SQL, SQL Azure, Kafka, Azure Storage, Azure Data Factory, SSIS, and PowerShell.**
- Involved in converting **Hive/SQL queries** into Spark transformations using **Spark dataframe** in **Python.**
- Created **Airflow** Scheduling scripts in **Python.**
- Implemented **Apache Airflow** for authoring, scheduling, and monitoring Data Pipelines.
- Worked on **MongoDB** schema/document modeling, querying, indexing and tuning.
- Actively participated and provided feedback in a constructive and insightful manner during weekly Iterative review meetings to track the progress for each iterative cycle and figure out the issues.

Environment: Spark, Python, PySpark, MapReduce, ETL, Tableau, Azure, Star Schema, Git, Hive, HBase, Apache Airflow, MongoDB, SQL, Agile and Windows.

Company: Walbrydge Technologies Pvt Ltd, India. **Duration: Jun 2019 – Jul 2023**
Role: Software Engineer
Responsibilities:

- Gathering business requirements, business analysis, and designing various data products.
- Develop programs in **Spark** to use on applications for faster data processing than standard **MapReduce** programs.
- Developed **Apache Spark** jobs using **Scala** in the test environment for faster data processing and used **Spark SQL** for querying.
- Designed **ETL workflows** on **Tableau** and deployed data from various sources to **HDFS.**
- Developed **Tableau** data visualizations and dashboards using **Tableau Desktop.**
- Worked on **PySpark APIs** for data transformations.
- Used **AWS EMR** to transform and move large amounts of data into and out of other AWS data stores and databases, such as **S3 and DynamoDB.**
- Create data ingestion modules using **AWS Glue** for loading data in various layers in S3 and reporting using Athena and QuickSight.
- Use **Kafka, a publish-subscribe messaging system,** by creating topics using consumers and producers to ingest data into the application for Spark to process the data and create Kafka topics for application and system logs.
- Worked o**n Snowflake Schemas and Data Warehousing, and** processed batch and streaming data load pipeline using **Snow Pipe** and Matillion from the data lake Confidential AWS S3 bucket.
- Worked on SQL queries in dimensional data warehouses and relational data warehouses. Performed Data Analysis and Data Profiling using Complex **SQL** queries on various systems.
- Implemented the project under **an Agile** Project Management Environment and followed the SCRUM iterative incremental model & configured various sprints to execute.
- Worked on **data partitioning** and distribution strategies to enhance the scalability of **MapReduce** jobs.
- Led the development of scalable, real-time **data pipelines**, integrating various data sources and ensuring seamless data flow from ingestion to visualization for product-level analytics.
- Actively participated and provided feedback in a constructive and insightful manner during weekly Iterative review meetings to track the progress for each iterative cycle and figure out the issues.

Environment: Spark, PySpark, Python, MapReduce, Hive, Kafka, AWS, Star Schema, Snowflake, SQL, and Windows.

Education

The University of Texas at Dallas (UTD) May 2025
M.S. Business Analytics and Artificial Intelligence **GPA:** 3.65/4.00

Certificate: Graduate Certificate in Applied Machine Learning

Relevant Coursework: Applied Machine Learning and LLM-based solutions, Applied DEEP LEARNING, Predictive Analytics, Advanced Statistics, Applied Econometrics, developing AI solutions, Organizations for Business Analysis Platforms.

Vellore Institute of Technology (VIT) June 2016– May 2020
Bachelor of Technology in Electronics and Communications Engineering