Sri Kumar Dundigalla

Data Engineer

FL| 813-666-7515|srikumar.8108@gmail.com|Linkedin|GitHub

Summary

- **5**+ years of experience as a Data Engineer with areas of Database Development, ETL Development, Data modeling, and Report Development.
- Proficiency in data engineering methodologies, including SDLC, Agile, and Waterfall, ensuring effective project execution.
- Knowledge of the Big Data ecosystem, including Hadoop, HBase MapReduce, Apache Spark, Hive and Pig, enabling the handling of large-scale data processing.
- Analyze or transform stored data by writing Spark Jobs using Python and Scala on AWS, Hive Scripts, Amazon AWS services ELT tool based on business requirements and having good working knowledge on Hive, Spark, Python, DBT, Snowflake, Scala Hbase, Sqoop, Oozie and Airflow for scheduling the jobs. Skilled in ETL processes, utilizing tools like Apache NiFi, Apache Kafka, Talend, and SSIS to extract, transform, and load data.
- Proficient in cloud technologies, with hands-on experience in AWS, Azure, GCP and Snowflake for data storage, processing, and used CI/CD deployment.
- Solid exposure in writing SQL queries and adept in data analysis and resolving data issues for various facts and dimensions within data warehouse and data marts.
- Working knowledge of Python including Discriminant Analysis NumPy, Pandas, SciPy, Matplotlib, Seaborn, and Scikit - learn.
- Capable of creating insightful visualizations and reports using Tableau, Power BI, and SSRS to communicate data-driven insights.
- Experienced in working with various databases, including AWS Redshift, MySQL, and Oracle and SQL Server for data storage and retrieval.

Skills

Language: Python, Scala, Java, SQL
Big Data Technologies: Hadoop, Spark, Hive, Pig, MapReduce, HBase
Data Warehousing: ETL Processes, Data Modeling, Dimensional Design
Cloud Platforms: AWS (EC2, S3, Redshift), Azure, GCP, AWS Glue, Azure Data Factory
Data Visualization: Microsoft Excel, Power BI, Tableau, Seaborn
Databases: SQL Server, PostgreSQL, MySQL, Snowflake
DevOps Tools: Git, Jenkins, Docker, Kubernetes
Data Pipelines: Apache Airflow, Luigi, Prefect
Streaming Technologies: Apache Kafka, Amazon Kinesis

Experience

JP Morgan Chase & Co., FL| October 2023 – Current |Data Engineer

- Developed various functionalities for Data Ingestion framework using **Python** resulting in a 30% reduction in data processing time.
- Creating the JSON configuration files for cleansing the data through generic Spark framework.
- Developing **Spark SQL** Jobs and processed data in **AWS S3** with complex business requirements and Reduce ETL processing time by 30% through code optimization and efficient data transfer techniques.
- Automated 80% of data pipeline tasks using Python scripts and Airflow workflows, resulting in 15% reduced operational costs.
- Improve data quality by 15% by implementing AWS Glue dynamic data filtering and deduplication capabilities, resulting in cleaner and more reliable data sets.
- Conceptualizing and executing complex **Spark SQL** Jobs, achieving a **40%** improvement in processing large-scale data sets for Data Marts.
- Executed extraction, processing, and loading of data from diverse sources to destinations using **AWS** Redshift.
- Integrate Lambda with other AWS services (API Gateway, SNS, SQS) to build event-driven data architectures, enabling real-time notifications and actions.
- Involved converting Hive/SQL queries into Spark transformations using Spark RDDs on Scala and Python.
- Orchestrated complex data pipelines with 10+ stages for data ingestion, transformation, and analysis using AWS Pipeline's visual workflow builder.

Zensar Technologies| June 2020 – July 2022 |Data Engineer

- Developed Reusable **Azure Data Factory** pipelines for data ingestion from various source systems, configuring datasets, and establishing source and destination linked services to seamlessly transfer data from **Oracle** databases to Azure Data Lake Store Raw Zone.
- Designed and executed diverse ETL pipelines using **Azure** Databricks to extract, transform, and load data from multiple sources, including flat files, databases, and APIs resulting in a 25% reduction in data processing times.
- Utilized Python on Spark to architect and develop **ETL** integration patterns and Experience in solving performance issues in spark.
- Proficiently managing **Azure BLOB** and Data Lake storage, adeptly loading data into **Azure SQL** Synapse analytics (DW).
- Leveraged the auto scaling capabilities of Azure Databricks to optimize resource utilization and reduce data processing costs by 15%.
- Proficient in writing **SQL** queries and conducting data analysis to resolve data discrepancies across diverse facts and dimensions within data warehouses and data marts, utilizing both star schema and **snowflake schema** methodologies.
- Automated the creation of resulting scripts and streamlined workflows utilizing **Apache Airflow** and shell scripting, guaranteeing daily execution in production environments.
- Integrated code versioning and continuous integration/continuous delivery (CI/CD) practices with Azure DevOps, enabling faster deployments and reduced risk of regressions.

Hexaware Technologies, India| July 2017 – May 2020 |Junior Data Engineer

- Implemented SCD2 for master reference data and developed data pipelines for daily incremental loads using Python, PySpark, and **AWS** Glue, while orchestrating end-to-end workflows with AWS Lambda, Glue, and S3 for seamless integration.
- Utilized and worked on Source/Version control Tools using Github, validated the change sets code changes, Check-in/Out and versioning and developed CI/CD pipelines using Jenkins on AWS to create, test, and deploy the code to higher environment.
- Developed **Spark SQL** solutions for transactional data transformation, established Redshift data marts for tailored datasets, and utilized Scala and Spark for large-scale data processing.
- Created Hive tables using HiveQL, then loaded the data into Hive tables and analyzed the data by developing Hive queries.
- Worked on implementing scalable infrastructure and platform for large amounts of data ingestion, aggregation, integration, analytics in Hadoop using Spark and Hive.
- Designed storage strategies using **S3** for cost-effective storage and HBase for low-latency access, including historical data.
- Collaborated cross-functionally to align data solutions with business goals, leveraging Tableau for visualization of insights and effective communication across teams.
- Conducted performance tuning and optimization of **Hive** scripts for efficient data retrieval from Hadoop data lakes, enhancing system performance.
- Hands on experience in creating real-time data streaming solutions using Apache Spark Core, Spark SQL, Data Frames and Pair RDD's.
- Skilled in **Docker**, **Kubernetes**, and **Git** enabling seamless application containerization, deployment across cloud environments, and efficient version control for collaborative development workflows.

Academic Projects

Generating Learning Outcomes & Questions Using LLM | OpenAI, LangChain, Chromdb, Pinecone, Python

- Developed methods using LLMs to streamline quiz creation for educational settings, enhancing efficiency and improving learning experiences through automated content generation. Enhanced educational content relevance using advanced prompting techniques like Zero-shot, RAG, and Dynamic-Few-shot, tailoring materials to specific student needs and objectives. Implemented k-means clustering for effective categorization of learning outcomes, optimizing content precision and ensuring comprehensive educational materials.

Classifying Human and ChatGPT Generated Answers Using Bert | Transformers, Pandas, Tiktoken, OpenAI, NLTK, PyTorch,matplotlib

- Developed a BERT-based model to distinguish between human-written and AI-generated student submissions with 99% precision, enhancing academic integrity by verifying the authenticity of work submitted. Applied the model in academic research settings to analyze and classify text submissions, effectively identifying AI-generated content and supporting investigations into the use of generative AI in educational environments.

Certifications

- Azure Fundamentals (AZ-900), Microsoft 2
- Databricks Certified Data Engineer Associate
- AWS Certified Developer - Associate

Education

Master's In Business Analytics and Information Systems
University Of South Florida, Tampa

Bachelor's In Electronics & Computer Engineering
Amrita Vishwa Vidyapeetham, India