SANDEEP BIRUDUKOTA

6692044373 | sandeepbirudukota1@gmail.com | LinkedIn | GitHub

EDUCATION

Master of Science in Software Engineering | GPA - 3.57/4 **Jan 2022 - Dec 2023**
San Jose State University | CA, USA
Bachelor of Technology in Computer Science Engineering **Aug 2015 - Apr 2019**
S.R.K.R. Engineering College | AP, India

TECHNICAL SKILLS

- **Languages:** Python, Java, SQL, R, Shell scripting, Oracle PL/SQL
- **Databases:** MySQL, Oracle, MS SQL, MongoDB, PostgreSQL, Redis
- **Framework and Libraries:** Pandas, NumPy, NLTK, Scikit-learn, TensorFlow, Matplotlib, React.js, Flask, Express.js
- **Big Data Technologies:** Hadoop ecosystem (HDFS, Map Reduce, HIVE), Spark
- **Cloud Technologies:** Amazon Web Services (AWS), Microsoft Azure, Snowflake
- **Tools:** Power BI, Tableau, Jira, Jenkins, Git, Azure Databricks, Apache Kafka, Docker, MS Office

PROFESSIONAL EXPERIENCE

Data Engineer, San Jose State University **Jan 2022 - Dec 2023**

- Configured **AWS CloudFormation** templates to automate the creation of necessary resources, including **S3 buckets**, **IAM roles**, and **Redshift clusters**, streamlining infrastructure setup and reducing deployment time by **25%.**
- Developed an **ETL pipeline** using AWS services, extracting student data from API, transforming it with **AWS Glue**, and loading it into **AWS Redshift**, improving data processing efficiency by **30%**
- Automated data extraction and transformation with **Apache Airflow**, managing **DAG**s for seamless data flow from S3 to Redshift by creating custom **python** operators, utilizing **XComs** for cross-talk communication, reducing manual intervention by **50%.**
- Implemented **ratelimit** to handle API rate, validated data to ensure data quality and stored API keys securely using connection management. Optimized **Airflow performance** by tuning scheduler, increasing task parallelism resulting in **30%** improvement.
- Developed **PySpark** scripts, leveraging **AWS Glue**'s built-in transforms such as **ApplyMapping** to reshape the schema, rename fields, and change data types. Utilized **DynamicFrame** operations to manipulate and enrich the data.
- Integrated **Tableau** with **AWS Redshift** to enable interactive dashboards and reports. Utilized **Tableau**'s calculated fields, parameters, and filters to create dynamic and interactive visualizations which increased data-driven decisions by **35%.**

Data Engineer, TCS **Oct 2019 - Dec 2021**

- Led the design, implementation, and maintenance of a real-time **ETL (Extract, Transform, Load)** pipeline, enabling seamless data extraction from a retail client's **API** and subsequent loading into **Snowflake**, a cloud-based data warehouse.
- Implemented an **Amazon EventBridge** scheduled event to trigger an **AWS Lambda** function hourly, automating the data retrieval process from the retail API, and improving data reliability by **40%**.
- Provisioned an **Amazon DynamoDB** table with **stream** enabled, allowing capture of data modification events, and utilized the **Lambda** function to store data in DynamoDB, a **NoSQL** data store, improving data capture and storage efficiency by **60%**
- Designed a separate **AWS Lambda** function to consume the **DynamoDB stream** by polling for new records, processing the data, and loading it into an **Amazon S3** bucket, enabling efficient data storage and retrieval.
- Configured **Snowpipe** to integrate with **SQS**, generating notifications in an **SQS queue** upon successful data ingestion from the **Lambda function** into the designated **S3** bucket, providing a reliable and scalable messaging system.
- Created **Snowflake external stage** to ingest data from the **S3** bucket by receiving notifications from the **SQS queue** through **Snowpipe**, streamlining the data ingestion process into the cloud data warehouse, improving data loading speed by **40%**
- Integrated **Power BI** with the **Snowflake** data warehouse, utilizing advanced **DAX** expressions to enable visualization and analysis of retail client data, empowering data-driven decision-making through valuable insights.
- Established robust security measures through **Identity and Access Management (IAM)** protocols and **role-based access control (RBAC),** significantly reducing the risk of unauthorized data breaches by **80%**
- Implemented **CI/CD** pipelines using **GitLab** to automate the building, testing, and deployment processes of **data pipelines** and **ETL tasks**, enabling faster delivery, increased reliability, and enhanced productivity in managing large-scale data operations.

Data Engineer Intern, NR IT Software Services **Apr 2019 - Sep 2019**

- Spearheaded the development of a robust **ETL (Extract, Transform, Load)** pipeline to migrate financial data from an on-premises **SQL Server** database to **Azure Data Lake Storage**, enabling seamless data integration and centralized storage.
- Orchestrated the data ingestion process to **Azure Data Lake** using **Azure Data Factory (ADF),** leveraging **Data Flows** and Triggers to automate the extraction and storing of data in **Parquet** format, increasing data processing efficiency by **60%**.
- Performed complex data transformations using **Azure Databricks**, **PySpark,** and advanced SQL techniques, including **CTEs**, **window** functions, and **joins** according to business requirements, improving data transformation capabilities by **75%**.
- Utilized **Azure Synapse Analytics** to load the transformed data, creating optimized views for efficient querying and analysis, and designed **Power BI** dashboards, to visualize and analyze the financial data, improving data-driven decision-making by **80%.**