A picture containing text, sign, blue

Description automatically generatedA picture containing text, sign, outdoor, blue

Description automatically generated**SAI THOTA**

**Cloud Data Engineer**

**Employer-details:[Globexus963@gmail.com](mailto:Globexus963@gmail.com)Mobile :** 732-427-8156

**SUMMARY:**

* Cloud Data Engineer with 8+ years of experience in building data intensive applications, tackling challenging architectural and scalability problems. Expertise in Spark, Databricks, Azure ecosystem, Snowflake, Kafka, and Airflow.
* Experience in leading the integration of AI technologies like GENAI with SharePoint on Azure, showcasing skills in organizing large-scale data lakes in Azure Blob Storage.
* Proficient in automating data extraction and transformation using Python, regular expressions, and Spark data frames, resulting in enhanced AI processing accuracy and efficiency.
* Development work experience in technology areas of Python, Cloud environment. Proficient in Spark, Databricks, Azure ecosystem, Snowflake, Kafka, and Airflow.
* Extensive experience in implementing Microsoft BI/Azure BI solutions like Azure Data Factory, Azure Databricks, Azure Analysis Services, SQL Server Integration Services, SQL Server Reporting Services and Tableau.
* Experience in designing Azure Cloud Architecture and Implementation plans for hosting complex application workloads on MS Azure.
* Experience as Azure Data Engineer in Azure Cloud, Azure Data Factory, Azure Data Lake Storage, Azure Synapse Analytics, Azure Analytical services, Azure Cosmos, NO SQL DB, Azure HDInsight, Big Data Technologies (Hadoop and Apache Spark) and Databricks.
* Developed data pipelines using Snowflake, Kafka, and Airflow for real-time data ingestion and processing.
* Used Apache Airflow to schedule and manage data pipeline workflows, including data quality checks, data validation, and error handling.
* Utilized Snowflake's built-in support for semi-structured data to handle JSON and Avro data formats, which were ingested by Apache Kafka
* Utilized Snowflake's SQL to perform complex data analysis and reporting on the data stored in Snowflake.
* Utilized Snowflake as the data warehousing solution for storing and querying large amounts of data.
* Experienced in optimizing Hive queries by tuning configuration parameters.
* Expertise in all facets of Business Intelligence applications with a strong background in Data extraction, data visualization, report generation, infographics, and information visualization.
* Experienced in Logical and Physical data modelling.
* Worked with Snowflake's security features, such as row-level security and data masking, to ensure data governance and compliance.
* Competent in the creation of solutions utilizing Azure technologies including Azure Blob Storage, Event Hubs, and Data Lake Analytics.
* transformation and cleansing, data quality checks, and error handling.
* Proficiency in SQL and Python for data analysis and manipulation, including experience with popular libraries such as Spark and PySpark.
* Experience with data visualization tools such as PowerBI, including experience creating interactive dashboards and reports.
* Familiarity with data governance and security best practices, including experience implementing data encryption and access controls.
* Ability to work independently and as part of a team, with strong communication skills and a collaborative approach to problem-solving.
* Developed and maintained Continuous integration and Continuous Deployment of Data pipeline using Gitlab and Jenkins.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| *Programming Languages:* | Python, JavaScript, SQL, Scala, C# |
| *Parallel & Syncing cloud outlineComputing:* | Spark, Databricks, Snowflake, Redshift, Kubernetes, Apache Hive, Apache Airflow, ML Flow, Teradata, AWS, Azure, GCP, Kafka, Jenkins |
| *SQL & NoSQL:* | MongoDB, Dynamo DB, PostgreSQL, MySQL, SQL Server, Oracle, FIAS |
| *Data Visualization:* | Tableau, Power BI, Cognos, Business Objects, Hyperion, Web Focus |
| *Data Models:* | JSON, XML, Data Vault, Subject Area modeling, ACORD. |
| *Tools:* | Lang Chain, AWS Glue, Informatica, Data Stage, APIGEE, Flask, Fast API, AWS API Gateway GITLAB, JIRA, AWS Code build, Splunk, Visio |

**Certifications**

* **Microsoft Certified: Azure Data Engineer Associate**
* **Data Bricks Accreditation: Generative AI**
* **Microsoft Certified: Power BI Data Analyst Associate**
* **Microsoft Certified: Power Platform Fundamentals**
* **Microsoft Certified: Azure AI Fundamentals**
* **Microsoft Certified: Azure Data Fundamentals**
* **Microsoft Certified: Azure Fundamentals**
* **AWS certified solutions architect associate**

**PROFESSIONAL EXPERIENCE**

GEN AI Data Engineer May 2023 to Present

E&Y – Columbus, oh

* Led the integration of GENAI with SharePoint on Azure, successfully linking the bot to the company’s internal database, which included over 10,000 documents into Azure Blob Storage, forming an organized and scalable data lake.
* Employed Python scripts, regular expressions, and Spark data frames to automate the extraction and organization of data from intricate policy documents. Transformed this data into a structured format compatible with AI processing, significantly improving GENAI’s response accuracy.
* Orchestrated data ingestion using Azure Data Factory, integrating data from SharePoint, Bloomberg, Reuters, and social media.
* Leveraged Azure Databricks for intensive data processing, utilizing Python within Databricks notebooks to run Spark jobs, transforming raw data into structured, analysis-ready formats.
* Conducted ETL (Extract, Transform, Load) operations in Azure Databricks, employing regular expressions and Spark data frames to parse and structure complex policy documents.
* Established a data warehouse using Azure Synapse Analytics, seamlessly integrating with Databricks and Data Factory for efficient storage and quick data retrieval.
* Integrated GENAI with Azure infrastructure, enhancing its NLP capabilities through Azure Cognitive Services, thus improving response accuracy and efficiency.
* Implemented effective caching strategies and query optimization in Databricks, enhancing GENAI's response time by 25%.
* Ensured GDPR compliance and data security within the Azure environment, implementing encryption, access controls, and conducting regular audits.
* Developed interactive dashboards and automated reporting tools in Power BI, connected to Azure Synapse Analytics, providing real-time insights and analytics.
* Addressed complex data integration challenges, focusing on inconsistencies in data formats. Utilized Databricks for seamless integration and processing, thereby improving GENAI's reliability and performance.
* Optimized data retrieval processes using Spark SQL and Databricks caching mechanisms, achieving a 25% enhancement in GENAI's response times.
* Facilitated end-to-end data encryption both in transit and at rest, utilizing Azure's security features. This ensured that all data handled within the GENAI project adhered to stringent security standards.
* Implemented a dynamic data partitioning strategy in Databricks, optimizing data storage and access patterns, which led to a significant reduction in query execution times and resource utilization.
* Regularly updated and refined the data processing and AI algorithms in response to evolving data patterns and business requirements, ensuring that GENAI remained adaptive and relevant to current market trends.

**Environment**: Azure Data Factory, Azure Blob Storage, Azure Databricks, Python, Spark, Azure Synapse Analytics, Azure Cognitive Services, Power BI, Azure Stream Analytics, Azure Monitor, Azure DevOps, Machine Learning Algorithms, Third-Party APIs.

Cloud Data Engineer Mar 2021 to APR 2023

AT&T － Dallas, TX

* Involved in Architecting and developing Azure data factory pipelines by creating datasets, source and destination linked services to move the data from oracle database to Azure Data Lake Store Raw Zone
* Used Azure Databricks for performing data transformations and load data into Azure Data Lake
* Worked on migration of data from On-prem SQL server to Cloud databases (Azure Synapse Analytics (DW) & Azure SQL DB)
* Have good experience working with Azure BLOB and Data Lake storage, loading data into Azure SQL Synapse analytics (DW)
* Designed and implemented a wide range of ETL pipelines using Azure Databricks to extract, transform, and load data from various sources, including flat files, databases, and APIs.
* Designing and maintaining reports in Power BI, built on top of Azure Synapse/Azure Data Warehouse, Azure Data Lake, Azure SQL
* Utilized Snowflake as the data warehousing solution for storing and querying large amounts of data.
* Implemented Kafka for real-time data streaming and handling high volume, high throughput, and low latency data.
* Worked with Snowflake's security features, such as row-level security and data masking, to ensure data governance and compliance.
* As a Power BI admin creating workspaces, designing security including row level security for various reports.
* Working with users to gather requirements and training users.
* Utilized Azure Cloud Technologies to store and manage data, including Azure Data Lake and Azure SQL Database
* Developed and maintained data pipelines to support business intelligence and reporting needs, including creating/scheduling jobs, integrating with Azure DevOps, and utilizing notebooks and clusters.
* Created pipelines in Azure Data Factory to execute Databricks notebooks, design robust pipelines to handle unexpected scenarios, create dependencies between activities and pipelines, and schedule pipelines using triggers.
* Collaborated with cross-functional teams to understand data requirements and deliver solutions, including working with data analysts and business users to define requirements and scope.
* Created dashboards and reports using Power BI to visualize and communicate data insights, including interactive visualizations and reports for stakeholders.
* Mounted Azure Storage in Databricks using service principle, redacted secrets stored in Azure Key Vault
* Worked closely with data analysts and business users to understand data needs and requirements, provided technical guidance and support as needed.
* Participated in data modelling and data architecture design discussions. provided recommendations on data storage and management solutions.
* Conducted data quality checks and audits to ensure data accuracy and completeness, implemented processes for ongoing data monitoring and maintenance.
* Assisted with the design and implementation of data governance and security policies and procedures, including data encryption and access controls.
* Utilized and worked on Source/Version control Tools using perforce, validate the change sets code changes, Check-in/Out and versioning and Maintained Lower Environment CI/CD pipeline.

**Environment**: SQL Server, Azure SQL database, Snowflake, Kafka, airflow, Azure Delta, Azure Data warehouse, Azure Synapse Analytics, Azure Data Factory, Databricks, Azure Analysis Services, Azure cloud services, Azure Data Lake, Data Visualization, Agile, Jira, Python, GITLAB, JENKINS

Cloud Data Engineer

Wabtec Corp-－ Dallas, TX Jan 2020 to Feb 2021

* Design and implement end-to-end data solutions in Azure, including experience with Databricks for data processing and analysis, Azure Data Lake, Data Factory, Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Databricks, and NoSQL DB
* Design and implement end-to-end data solutions in Azure, including experience with Databricks for data processing and analysis, Azure Data Lake, Data Factory, Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Databricks, and NoSQL DB
* Develop Python programs for manipulating data from various sources and converting them into CSV files.
* Design and implement database solutions in Azure SQL Data Warehouse and Azure SQL
* Use Spark to encode and decode Json objects, create and modify data frames in Apache Spark within Databricks notebooks.
* Configure Spark Streaming to receive ongoing data from Kafka and store the stream information in HDFS.
* Perform statistical data analysis and data visualization using Python and R within Databricks notebooks.
* Used Azure API Manager to maintain on-premises API services with policies.
* Create data models in Splunk using pivot tables by analysing vast amounts of data and extracting key information to suit various business requirements.
* Develop ETL data workflows and code for new data integrations and maintain/enhance existing workflows and code using Databricks.
* Build pipelines to move hashed and un-hashed data from Azure Blob to Data Lake using Databricks.
* Recreate existing application logic and functionality in the Azure Data Lake, Data Factory, SQL Database, and SQL Data warehouse environment.
* Set up Azure Repo and Pipelines for CI/CD deployment of objects.
* Create Hive queries within Databricks to help market analysts spot emerging trends by comparing incremental data with Teradata reference tables and historical metrics.
* Develop normalized Logical and Physical database models for designing an OLTP application.
* Work with senior management to plan, define, and clarify dashboard goals, objectives, and requirements using Databricks for data analysis and visualization.
* Communicate daily with management and internal organizations regarding the status of all assigned projects and tasks using insights and data analysed in Databricks.

**Environment**: Azure Databricks, Python, Teradata, Azure SQL Data Warehouse, Azure SQL, Azure Data Platform, Azure Data Lake, Data Factory, Analytics, NoSQL Kafka, HDFS, Spark, Splunk

Data Engineer

AT technologies pvt ltd －India Jan 2018 to Aug 2019

* Built and maintained a wide range of ETL pipelines to extract, transform, and load data from various sources, including flat files, databases, and APIs.
* Developed ETL as per mapping requirements to load data into staging from multiple sources such as csv, xml, xlsx, etc.
* Utilized SQL and Python for data manipulation and analysis, including experience with popular libraries such as Spark and PySpark
* Implemented data governance and security best practices to ensure data integrity and compliance.
* Worked on developing ETL pipelines on S3 parquet files on data lake using AWS Glue.
* Wrote lambda functions using python for custom data transformation and ETL.
* Updated data mapping documents based on the actual development.
* Worked with AWS Cloud Formation templates for reusable code deployment across environments.
* Collaborated with cross-functional teams to understand data requirements and deliver solutions.
* Created dashboards and reports using Power BI to visualize and communicate data insights.
* Worked closely with data analysts and business users to understand data needs and requirements and provided technical guidance and support as needed.
* Participated in data modeling and data architecture design discussions and provided.
* Devised simple and complex SQL scripts to check and validate Dataflow in various applications.
* Performed Data Analysis, Data Migration, Data Cleansing, Transformation, Integration, Data Import, and Data Export
* Devised PL/SQL Stored Procedures, Functions, Triggers, Views, and packages.
* Made use of Indexing, Aggregation and Materialized views to optimize query performance.
* Involved in developing and documenting the ETL (Extract, Transformation and Load) strategy to populate the Data Warehouse from various source systems.
* Developed logistic regression models (using R programming and Python) to predict subscription response rate based on customer's variables like past transactions, response to prior mailings, promotions, demographics, interests, and hobbies, etc.
* Created Tableau dashboards/reports for data visualization, Reporting and Analysis and presented it to Business.
* Design and develop spark job with Scala to implement end to end data pipeline for batch processing.
* Created Data Connections, Published on Tableau Server for usage with Operational or Monitoring Dashboards

**Environment**: AWS, Spark, Python, PowerBi, Glue, S3, SQL, ETL, EC2, Cloud formation, Shell Scripting, GITLAB.

SQL, PL/SQL, R programming, Python, ETL (Extract, Transform, Load), Data Warehouse, Tableau, Spark

**Education and Training**

Master of Computer Science **Dec 2020**

University of Dayton

Bachelor of Technology**,** Computer Science and Game Development **May 2016**

Jawaharlal Nehru Architecture and Fine Arts University (JNTU **INDIA)**