|  |
| --- |
| A logo for a company  Description automatically generated**Venkat Revanth Nalla**venkatrevanthna**@gmail.com****Phone: (216) 801-8099****Data Engineer/ ETL Developer** |

**Professional Summary:**

* Skilled Data Engineer, ETL Developer and BI Developerwith 10+ years of IT experience in Database Development, ETL solutions, and building Data Warehouse as well as proven ability to produce results in a fast-paced environment with critical deadlines.
* Extensive experience in Azure Cloud, Azure App Services, 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), Data bricks, Event Hub and Event grid.
* Experience working in reading Continuous Json data from different source system using Kafka into Databricks Delta and processing the files using Apache Structured streaming, Pyspark and creating the files in parquet format.
* Created data pipelines for both batch process, Micro-batch streaming and continuous streaming process in Databricks for high latency, low latency and ultra-low latency of data accordingly by using inbuilt Apache spark modules.
* Well versed experienced in creating pipelines in Azure Cloud ADFv2 using different activities like Move &Transform, Copy, filter, for each, Data bricks etc.
* Providing Azure technical expertise including strategic design and architectural mentorship, assessments, POCs, etc., in support of the overall sales lifecycle or consulting engagement process.
* Hands on experience in Hadoop Ecosystem components such as Hadoop, Spark, HDFS, YARN, TEZ, Hive, Sqoop, MapReduce, Pig, OOZIE, Kafka, Storm, HBASE.
* In-depth understanding of Spark Architecture including Spark Core, Spark SQL, Data Frames and Spark Streaming.
* Experience in designing and implementing data pipelines, data storage solutions, and data warehousing systems using AWS tools such as S3, RDS, DynamoDB, Redshift, and Athena.
* Experience in implementing data security and privacy policies to ensure the confidentiality and integrity of data using AWS tools such as IAM and VPC.
* Strong understanding of data architecture and design principles and the ability to develop and implement scalable data solutions using AWS services such as EC2, Glue, and Lambda.
* Ability to perform data analytics, predictive modeling, and data-driven decision-making using Azure tools such as HDInsight, Data Factory, and Synapse Analytics.
* Experience in working with data streams and real-time data processing systems using Azure tools such as Event Hubs, Stream Analytics, and Service Bus.
* Skilled in automating data migration processes using Azure Data Factory and scheduling pipelines for timely data updates.
* Strong experience in writing applications using Python using different libraries like **Pandas, NumPy**, SciPy, Matpotlib etc.
* Good understanding of NoSQL databases and hands on work experience in writing applications on NoSQL databases like CosmosDB.
* Much experience in performing Data Modelling by designing Conceptual, Logical data models and translating them to Physical data models for high volume datasets from various sources like Oracle, Teradata, Vertica, and SQL Server by using Erwin tool.
* Expert knowledge and experience in Business Intelligence Data Architecture, Data Management and Modeling to integrate multiple, complex data sources that are transactional and non-transactional, structured, and unstructured.
* Also, design and develop relational databases for collecting and storing data and build and design data input and data collection mechanisms.
* Well versed with Relational and Dimensional Modeling techniques like Star, Snowflake Schema, OLTP, OLAP, Normalization, Fact and Dimensional Tables.
* Good knowledge in creating SQL queries, collecting statistics and Teradata SQL query performance tuning techniques and Optimizer/explain plan.

**Technical Skills:**

|  |  |
| --- | --- |
| Azure Cloud Platform | ADFv2, BLOB Storage, ADLS2, Azure SQL DB, SQL server, Azure Synapse, Azure Analytic Services, Data bricks, Mapping Dataflow (MDF), Azure Cosmos DB, Azure Stream Analytics, Azure Event Hub, Azure Machine Learning, App Services, Logic Apps, Event Grid, Service Bus, Azure DevOps, GIT Repository Management, ARM Templates |
| Reporting and BI Tools | Power BI, Tableau and Cognos |
| ETL Tools:  | ADFV2, Informatica Power Center 10.x/9.x, DataStage 11.x/9.x, SSIS |
| Programming Languages | PySpark, Python, U-SQL, T-SQL, LINUX Shell Scripting, AZURE PowerShell, C#, Java |
| Big data Technologies | Hadoop, HDFS, Hive, Apache Spark, Apache Kafka, Pig, Zookeeper, Sqoop, Oozie, HBASE, YARN |
| Databases | Azure SQL Warehouse, Azure SQL DB, Azure Cosmos No SQL DB, Oracle, Microsoft SQL Server |
| IDE and Tools | Code, Eclipse, SSMS, Maven, SBT, MS-Project, GitHub, Microsoft Visual Studio |
| Cloud Stack | AWS, GCP, Azure, Snowflake |
| Methodologies | Waterfall, Agile/Scrum, SDLC |

**Professional Experience:**

**Client: Walmart, AR Aug 2021 to Present**

**Role: Azure Data Engineer**

**Responsibilities:**

* Created and deployed infrastructure templates using ARM Templates and Terraform, ensuring consistent and repeatable environment setups.
* Created and maintained PowerShell scripts to automate environment provisioning, configuration, and maintenance tasks.
* Leveraged Microsoft Fabric to integrate and manage multiple data workloads, including data engineering, data integration, and data science, through a single unified platform within Microsoft Azure.
* Utilized DBT for streamlined data transformation, simplifying workflows by allowing teams to model data directly in SQL, reducing reliance on complex ETL processes, and speeding up data readiness for analytics.
* Configured Linked Services and Datasets within Azure Data Factory, ensuring secure and optimized data connections across various Azure resources.
* Designed optimized Synapse workspaces to efficiently process, analyze, and visualize large datasets for real-time decision-making.
* Tuned Cosmos DB throughput and performance using partition keys and consistency levels, improving read/write performance by 30%.
* Designed and implemented ETL processes to extract, transform, and load different file formats data into data warehouses, ensuring data accuracy and consistency.
* Developed and deployed Shell scripts to automate routine tasks such as file backups, system monitoring, log analysis, and data extraction, reducing manual intervention.
* Automated data ingestion pipelines to bring in structured and unstructured data from various sources into Synapse using Azure Data Factory and Synapse Pipelines.
* Developed and optimized Python scripts for data extraction, transformation, and loading (ETL) processes, enhancing data pipeline performance and efficiency.
* Utilized Azure Machine Learning Studio to develop, train, and deploy machine learning models, enabling scalable solutions for predictive analytics.
* Integrated Azure Data Factory (ADF) pipelines with DevOps workflows for seamless deployment.
* Utilized Microsoft Fabric to foster collaboration between data engineers, data scientists, and business analysts, streamlining the end-to-end data processing and analytics workflow through shared workspaces.
* Designed and maintained Azure SQL databases, ensuring optimal performance, security, and scalability for transactional and analytical workloads.
* Developed custom APIs in Azure Functions for data ingestion and transformation, enhancing system integration.
* Tuned and optimized Spark workloads within Azure Databricks, improving job performance while reducing resource consumption and execution time.
* Leveraged Azure Functions and Logic Apps to create event-driven, scalable data workflows in a serverless environment.
* Implemented Cosmos DB SQL API and Gremlin API for diverse use cases like graph data and document-based storage.
* Engineered robust APIs to facilitate seamless data exchange between enterprise systems.
* Integrated Azure Databricks with Azure Data Factory, Azure Data Lake, Azure Event Hubs and Azure SQL to create end-to-end data processing pipelines, providing seamless data flow and analytics capabilities.
* Automated server configuration, system updates, and maintenance tasks using Shell scripts, improving system uptime and operational efficiency.
* Developed reusable Python patterns for Synapse integration, aggregations, change data capture, deduplication, and high watermark implementation.
* Proficient in managing Power BI Service, including publishing reports, managing datasets, setting up data refresh schedules, and sharing dashboards securely with stakeholders.
* Actively involved in the development and configuration of Unity Catalog and managed Unity Catalog to centralize access control, data goveranance, auditing, lineage, and data discovery.
* Engineered PySpark transformations and optimized processing in Azure Databricks.
* Configured triggers and schedules in ADF to ensure timely execution of data pipelines, supporting real-time and batch processing requirements.
* Optimized Databricks clusters and job configurations to improve performance and reduce operational costs, ensuring efficient resource utilization.
* Implemented custom Python connectors and APIs for seamless data integration between various systems and platforms.
* Established and enforced GIT workflows, including GitFlow and feature branching, to streamline development processes and ensure consistent code quality.
* Designed and implemented CI/CD pipelines using GitHub Actions and Azure DevOps.
* Built and optimized data pipelines with Spark and Databricks for batch and streaming data solutions.
* Leveraged DBT’s modular structure to build reusable data models, enabling scalability and quick iteration on data transformations as business needs evolved.
* Demonstrated proficiency in Databricks components: SQL, Delta Live Tables, Repos, and Task Orchestration.
* Developed complex T-SQL queries for ETL processes, ensuring data integrity.
* Worked with Oracle and SQL Server for data integration and storage requirements.
* Engineered complex PySpark transformations in Azure Databricks notebooks, optimizing processing time and resource utilization.
* Conducted performance tuning for SQL queries and Spark jobs, achieving significant throughput improvements.
* Tuned Synapse SQL pools for optimal performance in handling large-scale datasets, focusing on distribution and parallelism settings for efficient resource utilization.
* Wrote SQL stored procedures to enhance the performance of Azure Data Factory (ADF) pipeline runs, ensuring efficient data processing and reduced execution times.
* Experience embedding Power BI reports in external applications for seamless integration into business portals and providing enhanced reporting capabilities for end-users.
* Integrated Azure Data Factory with Logic Apps and Functions to automate workflows.
* Managed end-to-end data workflows, including Data Transformation activities using Stored Procedures and Azure Functions.
* Seamlessly integrated Synapse with Power BI for interactive reporting and dashboarding, enabling end-users to derive real-time insights from large datasets.
* Integrated machine learning models into production using Microsoft Fabric’s support for MLOps, automating model deployment, monitoring, and retraining processes at scale.
* Developed data validation processes within ETL pipelines to identify and rectify data anomalies proactively.
* Developed and optimized SQL views and stored procedures in Azure SQL Data Warehouse for enhanced reporting capabilities.
* Developed interactive and dynamic dashboards in PowerBI to visualize key business metrics, enabling data-driven decision-making.

**Environment:** Azure Cloud, Azure Databricks, Azure Data Factory (ADF v2), Azure functions Apps, Azure Data Lake, Blob Storage, SQL server, Windows remote desktop, Java, Unix, Azure PowerShell, Data bricks, Python, Pyspark Azure Cosmos DB, Azure Stream Analytics, Azure Event Hub, Power BI.

**Client: First National Bank- Charlotte, NC Nov 2019 to July 2021**

**Role: Azure Data Engineer**

**Responsibilities:**

* Developed and implemented robust data governance policies and procedures within the Azure ecosystem, improving data quality, accuracy, and reliability.
* Implemented globally distributed, highly available NoSQL databases with Azure Cosmos DB to support low-latency applications.
* Collaborated with stakeholders to align Azure cloud solutions with business objectives, ensuring that the cloud strategy met organizational goals.
* Utilized Azure Databricks collaborative notebooks for team-based data development, allowing efficient version control, real-time collaboration, and notebook sharing for advanced data engineering tasks.
* Deployed infrastructure as code (IaC) solutions using Terraform to manage and secure Azure resources, enabling consistent and repeatable cloud deployments.
* Managed end-to-end data pipelines within Azure, ensuring seamless data flow and transformation across the big data ecosystem.
* Integrated REST APIs into data pipelines and applications, enabling real-time data retrieval and communication between systems for efficient data processing and sharing.
* Administered and set up Azure Data Factory configurations, optimizing Linked Services, and Azure resources for seamless data integration processes.
* Engineered and managed big data processing pipelines in Azure Databricks, leveraging Apache Spark for distributed computing.
* Managed Azure Data Factory resources, including triggers, activities, and datasets, to ensure optimal pipeline performance and reliability.
* Created batch processing scripts for bulk file manipulation, data transfers, and scheduled job execution using tools like cron, enhancing the overall workflow automation.
* Leveraged Synapse data encryption and auditing to maintain data integrity and confidentiality in the BI environment.
* Provided training on Azure AI services and best practices, upskilling team members in AI model design and deployment.
* Managed source code using Git and GitFlow principles, ensuring version control and collaborative development practices.
* Designed RESTful APIs to support microservices architecture.
* Utilized Azure monitoring tools for infrastructure sizing, cost management, and proactive issue resolution through alerts and autoscaling.
* Integrated Azure Data Factory(ADF) with other Azure services like Databricks and Synapse, creating a seamless data integration pipeline.
* Developed and maintained ETL/ELT pipelines using SSIS, Spark, java, scala and SQL Stored Procedures, ensuring efficient data transformation and integration.
* Leveraged Apache Spark for advanced data processing and analytics, improving data pipeline performance.
* Integrated Event Hubs with Azure Stream Analytics and Databricks for real-time event processing and transformation.
* Integrated Python scripts into Azure Functions for serverless execution, enabling scalable and cost-effective data processing
* Designed and developed advanced Postgres queries, stored procedures, functions, and triggers to support data manipulation and business logic implementation.
* Utilized Python libraries such as Pandas, NumPy, and PySpark to perform data analysis, manipulation, and processing tasks, ensuring data accuracy and consistency.
* Utilized Azure Stream Analytics for real-time analytics and insights on streaming data.
* Implemented data pipelines and transformations to integrate data into Snowflake.
* Managed the deployment of SSIS packages and reports to production environments, ensuring seamless integration and minimal disruption.
* Configured Azure Data Factory(ADF) pipelines to process large volumes of data in real-time, reducing latency and improving data availability.
* Optimized query performance in SQL Data Warehouse by implementing partitioning, indexing, and query tuning strategies.
* Optimized predictive models on Azure using advanced Data Science and ML frameworks for increased model accuracy.
* Implemented post-migration validation and performance tuning to optimize the Azure SQL Managed Instances, ensuring they meet or exceed on-premises performance levels.
* Integrated Shell scripts within CI/CD pipelines(Azure DevOps), automating build, deployment, and monitoring processes.
* Implemented Change Data Capture (CDC) mechanisms using Azure Event Hubs and Azure Functions for seamless data integration across systems.
* Extensive experience with ETL/ELT design, development, and implementation using industry-standard practices.
* Utilized PySpark’s DataFrame API to process large-scale datasets in distributed environments, optimizing ETL workflows for data ingestion, transformation, and analytics.
* Implemented real-time data dashboards using Power BI, enabling live data streaming from multiple sources such as Azure Event Hubs and IoT devices for immediate business insights.
* Automated complex data pipelines using Azure Databricks and Azure Data Factory, enabling efficient data ingestion, transformation, and loading processes into Azure Data Lake and Synapse.
* Loaded transformed data into target systems, including databases, data warehouses, and cloud storage, employing java libraries and custom scripts.
* Utilized Azure Data Factory(ADF) Mapping Data Flows to perform data transformations, ensuring data quality and consistency across the pipeline.
* Designed and implemented an efficient Extract, Transform, Load (ETL) architecture using Azure services for seamless data transfer from source servers to the Data Warehouse.
* Actively participated in designing and developing CI/CD pipelines for data engineering within the Azure ecosystem.
* Integrated SQL Data Warehouse with Power BI for real-time data visualization and reporting, enabling data-driven decision-making.
* Configured Azure Monitor and Application Insights for proactive monitoring of AI applications, minimizing downtime.
* Optimized Power BI reports and dashboards by tuning performance, reducing report load time, and ensuring efficient data refreshes for faster analytics and reporting.
* Implemented automation from code commit to deployment using Azure DevOps.
* Managed a cloud data warehouse on Azure, facilitating batch processing and streaming.
* Managed the deployment of SSIS packages and reports to production environments, ensuring seamless integration and minimal disruption.

**Environment:** Azure Data factory, Azure Databricks, Azure Event Hubs, Azure SQL Datawarehouse, Power BI.

**Client: Repco - Hyderabad, INDIA Aug 2017 to Feb 2019**

**Big Data Engineer**

**Responsibilities:**

* Demonstrated expertise in developing and deploying custom Hadoop applications within the AWS environment, ensuring seamless integration and performance optimization.
* Designed and implemented data architectures for managing large volumes of home loans data while adhering to GDPR and CCPA data privacy and security regulations.
* Developed and automated multiple ETL jobs using Amazon EMR, facilitating seamless data transfer from HDFS to S3.
* Created batch data pipelines for extracting data from S3 and loading it into RedShift using Glue jobs.
* Utilized PySpark and Scala to automate data ingestion from various sources, including APIs, AWS S3, and Redshift.
* Developed complex ETL mappings in Informatica PowerCenter for transforming and loading large datasets.
* Provided expert support in solving real business issues by leveraging knowledge of Hadoop distributed file systems and open-source frameworks, driving operational excellence and efficiency.
* Spearheaded design, development, and maintenance of dynamic data pipelines on Snowflake, ensuring seamless integration and analytics support.
* Configured Spark streaming to store and process real-time data from Kafka.
* Leveraged AWS EMR to store structured data in Hive and unstructured data in HBase.
* Cleaned and transformed data in HDFS using MapReduce (YARN) programs for ingestion into Hive schemas.
* Automated ETL processes using SSIS and T-SQL, reducing manual intervention and improving data processing efficiency.
* Applied advanced SQL skills to fine-tune Snowflake performance and enhance query efficiency, bolstering data integrity and security across the lifecycle.
* Developed and maintained data reporting and analytics solutions to support regulatory reporting and compliance monitoring.
* Loaded and transformed large datasets using Hadoop/Big Data concepts, optimizing performance and efficiency in data processing tasks.
* Created a data lake in Snowflake using Stitch, App Testing, and Production support.
* Managed S3 buckets, implemented policies, and utilized S3 and Glacier for storage and backup on AWS.
* Generated reports for the BI team by exporting analyzed data to relational databases for visualization using Sqoop.
* Created custom User Defined Functions (UDFs) to extend Hive and Pig core functionality.
* Enabled ODBC/JDBC data connectivity to Hive tables and worked with tools like Tableau and Flink

**Environment:** AWS S3, Glue, AWS EMR, Glacier, Redshift, Snowflake, Spark SQL, Sqoop, Flink, YARN, Kafka, MapReduce, Hadoop, HDFS, Hive, Tableau, Spotfire, HBase.

**Client: Colruyt Group– Hyderabad, INDIA Jan 2014 to July 2017**

**Data Engineer**

**Responsibilities:**

* Examined claims and supporting documentation to ensure policy compliance before processing.
* Developed a strong understanding of claim processing from both client and service provider perspectives, identifying key metrics for each.
* Managed policy servicing and maintenance operations, including coverage changes, beneficiary data updates, and premium payments.
* Processed claims data efficiently through the system.
* Possess a good understanding of Electronic Health Record (EHR) systems, including their functionalities, data models, data elements, and data privacy and security regulations.
* Designed high-performance batch ETL pipelines using AWS cloud services.
* Extracted data from relational databases and APIs with AWS Data Factory to store in AWS data lake storage.
* Designed data warehousing solutions using AWS Synapse Analytics for storing and analyzing transformed data.
* Implemented and designed Python microservices in the healthcare domain.
* Monitored productivity and resources using AWS Log Analytics.
* Implemented CI/CD pipelines with AWS DevOps for automated build, test, and deployment processes.
* Utilized AWS Event Hub to capture real-time data streams and route them to the appropriate data stores.
* Monitored data pipeline performance using AWS Monitoring and Analytics tools to ensure seamless data flow and identify potential bottlenecks.
* Played a critical role in a data migration project involving EHR, ensuring accurate, efficient, and secure data migration.
* Ensured data pipeline security using AWS security features, including role-based access control and encryption, to safeguard data privacy and confidentiality.
* Managed encryption keys and passwords through AWS Key Vault.
* Utilized AWS Logic Apps to orchestrate complex business processes and workflows.
* Implemented serverless computing solutions using AWS Lambda Functions for cost-effective and scalable data processing.
* Designed visualization dashboards for data analytics using Power BI.
* Proficient in practicing Agile methodology to update workflows and manage project lifecycles and sprints.

**Environment:** AWS Data Factory, Synapse Analytics, Python, Event Hub, Logic Apps, Key Vault, Log Analytics, Scala, Power BI.

**Education:**

Bachelors in Computer Science from Acharya Institutes of technology, 2014