|  |  |
| --- | --- |
| **V Deepthi**  **848-600-1656**  [**deepthiv1969@gmail.com**](mailto:deepthiv1969@gmail.com)  [**www.linkedin.com/in/deepthi-vemula-27986717a**](http://www.linkedin.com/in/deepthi-vemula-27986717a) | |
|  | **Sr. Data Engineer** |
|  |
| |  |  | | --- | --- | | **Key Skills** | **Profile Summary** | | **Languages:** Python, Scala, Pyspark, PowerShell  **Cloud Platforms:** AWS, Azure  **Big Data Technologies:** Apache Spark, Hadoop, NiFi, Hadoop Ecosystem Tools (Sqoop, Hive, Pig), Apache Beam, Kafka  **Database Management:** Oracle, MySQL, SQL Server, MongoDB, Azure CosmosDB, Cassandra  **Data Formats:** XML, JSON  **Data Warehousing:** Snowflake, Azure DW  **Version Control:** Git, GitHub, SVN, BitBucket  **CI/CD and Automation Tools:** Jenkins, Docker, Kubernetes, Ansible, Azure DevOps  **IaC Tools:** Cloud Formation, Terraform  **Monitoring and Visualization Tools:** Nagios, AWS CloudWatch, Prometheus, Grafana, Tableau, Power BI, AWS Quicksight  **Collaboration and Project Management:** Jira, Confluence, SharePoint, Bugzilla  **Security and Access Management:** AWS IAM, Azure Key Vault  **Programming Libraries:** NumPy, Pandas  **Microsoft Office:** Excel (Pivot Tables, VBA, Macros)  **Operating Systems:** Windows, MacOS, Linux | * Results-driven and highly skilled Data Engineer with over 10 years of hands-on experience in the end-to-end management of complex data engineering projects. * Proficient in utilizing a diverse set of programming languages, including Python, Scala, and PowerShell, to implement robust solutions across various platforms such as AWS, Azure, and Google Cloud Platform. * Expertise in big data technologies, encompassing Apache Spark, Hadoop, NiFi, and Hadoop Ecosystem Tools (Sqoop, Hive, Pig, Apache Beam, Kafka), ensuring efficient data processing and transformation. * Adept at working with many database management systems, including Oracle, MySQL, SQL Server, MongoDB, Snowflake, and Azure CosmosDB. * Skilled in handling XML and JSON data formats and implementing effective ETL processes. * Extensive experience in monitoring and visualization tools such as Nagios, AWS CloudWatch, Prometheus, Grafana, Tableau, Power BI, and AWS Quicksight. * Proven proficiency in cloud infrastructure management using Cloud Formation, and Terraform, and expertise in IAM policies for AWS and Azure through Azure Key Vault. * Well-versed in containerization and orchestration tools like Docker, Kubernetes, Ansible, and Jenkins, ensuring streamlined and automated deployment processes. * Strong version control skills with Git, GitHub, SVN, and BitBucket. * Demonstrated success in Agile and Scrum development environments, utilizing project management tools like Jira, Confluence, SharePoint, and Bugzilla. * An effective collaborator with excellent communication skills, dedicated to delivering scalable and maintainable data engineering solutions that meet and exceed business objectives. | | |

**Work Experience**

**Finish Line, Boulder, CO (Remote) | Sr. Data Engineer | Mar 2023 - Present**

**Roles & Responsibilities:**

* Led end-to-end data engineering projects through the Software Development Life Cycle (SDLC), ensuring the design, development, and deployment of robust and scalable solutions.
* Implemented Python scripts for data extraction, transformation, and loading (ETL), integrating diverse data sources seamlessly.
* Implemented Azure Data Factory (ADF) extensively for ingesting data from different source systems like relational and unstructured data to meet business functional requirements
* Orchestrated ETL workflows using AWS Glue and AWS Data Pipeline, automating the movement and transformation of data between various sources and AWS services.
* Managed and optimized relational databases, including AWS RDS (Relational Database Service), Oracle, and NoSQL databases like MongoDB.
* Created numerous pipelines in Azure using Azure Data Factory v2 to get the data from disparate source systems by using different Azure Activities like Move &amp; Transform, Copy, filter, for each, Databricks, etc.
* Maintained and provided support for optimal pipelines, data flows and complex data transformations and manipulations using ADF and PySpark with Databricks.
* Deployed and managed Apache Spark and Hadoop clusters using AWS EMR (Elastic MapReduce), optimizing distributed data processing tasks.
* Ingested, processed, and analyzed real-time streaming data using Amazon Kinesis, ensuring timely insights for data-driven decision-making.
* Used Polybase to load tables in Azure Synapse
* Implemented Azure, self-hosted integration runtime in ADF
* Utilized AWS Lambda for serverless data processing, executing scalable functions to handle various data engineering tasks.
* Stored and managed large volumes of data in Amazon S3, supporting data lakes and enabling efficient data storage and retrieval.
* Define and maintain EC2 instance configurations, security groups, and networking settings as code for repeatability and consistency.
* Scheduled, automated business processes and workflows using Azure Logic Apps.
* Developed robust and maintainable Java code for implementing data pipelines, ETL processes, and data integration solutions.
* Orchestrated serverless workflows using AWS Step Functions, ensuring efficient and reliable execution of data processing tasks.
* Implemented event-driven architectures with Amazon SNS, facilitating communication between different components of data systems.
* Implemented data processing logic using Java frameworks such as Apache Beam, Apache Flink, or Spring Batch.
* Used Azure DevOps & amp Jenkins pipelines to build and deploy different resources(Code and Infrastructure)in Azure.
* Worked with diverse data formats, including XML and JSON, ensuring interoperability and efficient data exchange.
* Collaborated on projects involving traditional data warehouses (Snowflake), optimizing storage and retrieval of structured data.
* Experience in managing Azure Data Lakes (ADLS) and Data Lake Analytics and an understanding of how to integrate with other Azure Services.
* Migration of on-premise data (Oracle/ Teradata) to Azure Data Lake Store(ADLS) using Azure DataFactory(ADF V1/V2).
* Implemented partitioning, indexing, and compression strategies to manage and access data efficiently in VLDB environments.
* Implemented and optimized data workflows using Hadoop ecosystem tools such as Spark, Hive, Pig, and Sqoop.
* Utilized Apache NiFi for data integration and data flow management, ensuring seamless movement of data between systems.
* Automated deployment processes using Jenkins, Docker, Kubernetes, Cloud Formation, and Ansible, ensuring efficient and reproducible infrastructure.
* Monitored and gained insights into system performance using AWS CloudWatch, supporting proactive management and issue resolution.
* Developed interactive and insightful dashboards and reports using AWS Quicksight, providing stakeholders with meaningful visualizations of data.
* Implemented RESTful APIs or SOAP APIs to expose data services and functionality to other applications and users.
* Created a Linked service to land the data from different sources to Azure Data Factory.
* Developed data import scripts or pipelines to transform and load structured and unstructured data into Neo4j.
* Designed and implemented data storage solutions using the Parquet file format for efficient columnar storage.
* Managed project workflows and issues using Jira, contributing to Scrum and Agile development practices.
* Implemented version control for codebase using BitBucket, collaborating with cross-functional teams in an Agile development environment.

**Tech Stack:** Python, AWS, AWS Glue, AWS Data Pipeline, Databricks, RDS, Azure Data Factory (ADF v2), Azure SQL Database, Azure functions Apps, Azure Data Lake, EMR, Kinesis, Lambda, S3, EC2, Step Functions, SNS, Oracle, MongoDB, REST API, XML, JSON, Hadoop, Spark, Hive, Pig, Sqoop, Apache NiFi, Snowflake, Jenkins, Docker, Kubernetes, Cloud Formation, Cloud Watch, Ansible, Quicksight, Jira, Scrum, Agile, BitBucket.

**GGK Technologies, India | Data Engineer | Nov 2019 – Jun 2022**

**Roles & Responsibilities:**

* Participated in designing and implementing data engineering solutions through the Software Development Life Cycle (SDLC), ensuring robustness, scalability, and maintainability.
* Implemented Scala scripts for data extraction, transformation, and loading (ETL), integrating diverse data sources seamlessly.
* Implemented data transformation and cleansing logic using Java libraries like Apache Spark or Apache Hadoop.
* Orchestrated end-to-end data workflows in Azure using Azure Data Factory (ADF), ensuring efficient and automated data movement and transformation.
* Utilized Azure Databricks for big data processing, implementing scalable Apache Spark clusters for large-scale analytics and machine learning tasks.
* Managed and optimized Azure SQL Data Warehouse (Azure DW) and MySQL databases, ensuring high-performance storage for structured data.
* Implemented NoSQL solutions using Azure CosmosDB, supporting flexible and scalable storage for semi-structured and unstructured data.
* Integrated Java applications with Azure services like Azure Data Lake Storage, Azure SQL Database, and Azure Cosmos DB for data storage and retrieval.
* Developed custom data processing algorithms and analytics applications in Java for complex data manipulation and analysis tasks.
* Developed and maintained data processing jobs using Scala and PySpark in Azure HDInsight, optimizing distributed data processing tasks.
* Implemented and optimized Hadoop-based data workflows, utilizing technologies like Sqoop, HDFS, Pig, and Hive for efficient data transfer and processing.
* Automated data engineering tasks and infrastructure provisioning using PowerShell and Terraform, enhancing efficiency and repeatability.
* Integrated security features such as encryption and authentication into Java-based data solutions to protect sensitive data.
* Written Cypher queries and scripts to retrieve, analyze, and visualize graph data stored in Neo4j.
* Implemented and optimized data processing using Data Lake Analytics, enabling on-demand and parallel processing of big data stored in Azure Data Lake Storage.
* Managed secrets and cryptographic keys securely using Azure Key Vault, ensuring data security requirements in the cloud environment.
* Handled XML data formats in data engineering workflows, ensuring interoperability with systems using this format.
* Configured SSIS logging and monitoring features to track package execution, performance metrics, and data flow statistics for troubleshooting and performance optimization purposes.
* Developed interactive and insightful dashboards using Tableau, providing stakeholders with meaningful data visualizations.
* Implemented CI/CD pipelines for data engineering solutions using Azure DevOps, ensuring smooth and automated deployment processes.
* Ingested, processed, and analyzed real-time streaming data using Azure Stream Analytics, providing immediate insights for decision-making.
* Monitored and troubleshoot performance issues in Java-based data solutions deployed on Azure.
* Monitored and gained insights into the performance of data engineering solutions using Prometheus, ensuring optimal system health.
* Implemented data governance policies within Smartsheet Datamesh to ensure compliance and security.
* Implemented version control for codebase using Git, collaborating with cross-functional teams in an Agile development environment.
* Managed project workflows and issues using JIRA, ensuring efficient tracking, prioritization, and resolution of tasks.

**Tech Stack:** Python, Scala, Azure, Azure AD, ADF, Databricks, Azure DW, MySQL, CosmosDB, Scala, HDInsight, Spark, PySpark, Hadoop, Sqoop, HDFS, Pig, Hive, PowerShell, Terraform, XML, Tableau, Azure DevOps, Prometheus, Git, Agile, JIRA.

**Byteridge Software, India | Data Engineer | June 2017 –Oct 2019**

**Roles & Responsibilities:**

* Utilized Amazon S3 as a scalable and durable storage solution for storing raw and processed data, log files, and intermediate results generated by the analytics platform.
* Secured S3 buckets and objects using IAM policies, bucket policies, and access control lists (ACLs), ensuring data confidentiality, integrity, and availability.
* Designed and implemented ETL processes using AWS Glue to extract, transform, and load data from various sources into data warehouses like Amazon Redshift.
* Configured AWS Glue crawlers to automatically discover and catalog new data sources, ensuring the metadata is up-to-date and accessible for queries.
* Designed the data warehouse scheme on Amazon Redshift, designing optimized tables, distribution keys, and sort keys to support efficient querying and analytics
* Optimized Spark jobs for parallelism, data partitioning, and caching to improve processing efficiency and reduce job execution time on EMR clusters.
* Deployed AWS Lambda functions to perform real-time data enrichment, validation, and aggregation as part of the analytics pipeline, processing streaming events triggered by Kinesis data streams.
* Developed and optimized data processing workflows on Amazon EMR for large-scale data analytics and processing tasks.
* Monitored Lambda function executions, logs, and error metrics using Amazon CloudWatch, configuring alarms and notifications to detect and respond to performance issues or failures.
* Configured Kinesis data analytics applications to analyze streaming data in real time, detecting anomalies, identifying trends, and triggering automated responses or alerts based on predefined thresholds.
* Optimized SQL queries to extract actionable insights from semi-structured or unstructured data formats such as JSON, CSV, or Parquet stored in S3 buckets.
* Monitored pipeline execution and performance metrics using AWS CloudWatch, logging pipeline activities and alerts for operational visibility and troubleshooting.
* Utilized Kafka Streams DSL or Kafka Streams API to build stateful and stateless stream processing applications, handling windowed aggregations, joins, and event-time processing
* Implemented machine learning algorithms and predictive models using Spark MLlib for customer segmentation, product recommendation, and sales forecasting.
* Ensured data integrity and consistency within databases through proper Oracle PL/SQL programming techniques.
* Developed real-time data processing solutions with Pub/Sub for seamless ingestion and distribution of streaming data.
* Designed and implemented datamarts to serve as specialized subsets of data warehouses for specific business units or departments.
* Scheduled Sqoop jobs as part of batch processing or streaming data pipelines to ingest, process, and analyze data in real-time or batch mode, depending on business requirements and SLAs.
* Implemented data loading pipelines to ingest and load data into Snowflake from various sources such as on-premises databases, cloud storage, SaaS applications, and streaming platforms.
* Integrated and transformed data between SQL Server, CDH, and other data storage systems using SSIS, ETL, and PL/SQL.
* Designed and queried NoSQL databases such as Amazon DynamoDB or MongoDB for storing and accessing semi-structured or unstructured data formats, supporting flexible and scalable data models
* Configured data ingestion and integration pipelines using Apache NiFi or ETL tools like Talend and Informatica to extract, transform, and load data from external sources into the analytics platform.
* Managed metadata for Iceberg tables, including schema evolution, statistics, and partition information.
* Implemented continuous integration and continuous deployment (CI/CD) pipelines using Jenkins or AWS CodePipeline to automate software development workflows, including code builds, tests, and deployments.
* Developed interactive dashboards and visualizations using Tableau, Power BI, or Amazon QuickSight to visualize and communicate insights from the analytics platform.
* Managed codebase and version control using Git, ensuring code quality, maintainability, and collaboration among team members.
* Implemented version control using Git and GitHub, ensuring collaborative and organized development practices.

**Tech Stack:** Python, Scala, Pyspark, SSIS, ETL, PL/SQL, SQL Server, CDH, Hadoop, Sqoop, Spark, Snowflake, Apache Beam, Hive, Kafka, JSON, IAM, Power BI, GIT, GitHub, Jenkins, Git, AWS Glue, Amazon Redshift, Amazon EMR, Amazon S3, AWS Lambda, Amazon Kinesis, Amazon Athena, Amazon Dynamo DB, MongoDB, Apache NiFi, Talend, Informatica, CI/CD Pipelines, Tableau, Power BI, Amazon quick sight

**Smac-X Inno Labs Private Limited, India | Data Analyst/Engineer | April 2014 – May 2017**

**Roles & Responsibilities:**

* Designed and implemented a data pipeline using Scala and Apache Spark to process and analyze large-scale datasets.
* Used Pandas and NumPy libraries for data manipulation and preprocessing tasks within the pipeline.
* Orchestrated ETL workflows using Apache Airflow to automate data integration and processing tasks.
* Used Talend for data extraction, transformation, and loading into target systems, ensuring data consistency and accuracy.
* Used data modeling tools and techniques to create entity-relationship diagrams, dimensional models, and data flow diagrams.
* Extracted data from relational databases using SQL queries, performing joins, aggregations and filtering operations as needed
* Transformed data using SQL statements to meet specific business requirements or data quality standards.
* Integrated data from MongoDB NoSQL databases into the data pipeline for processing unstructured or semi-structured data.
* Optimized datamart schemas and structures for querying and reporting performance.
* Configured and managed Kafka clusters to ingest high volumes of streaming data from various sources such as web applications, IoT devices, and sensors.
* Used PySpark to interface with MongoDB collections and perform data transformations before loading into the target data warehouse.
* Applied Hadoop ecosystem tools such as Sqoop for efficient data transfer between Hadoop and relational databases, Spark for distributed data processing, and Hive for querying and managing structured data.
* Created interactive visualizations and dashboards using Tableau to communicate insights from processed data effectively.
* Used Tableau's features to customize visualizations, add filters, and create dynamic presentations for stakeholders
* Managed project codebase and collaborated with team members using Git for version control.
* Utilized Git branches for feature development, bug fixes, and code reviews to maintain code quality and project integrity.
* Partitioned data files using Parquet's partitioning capabilities to improve query performance and enable efficient data pruning.
* Maintained and enhanced existing Scala codebase, adhering to best practices and coding standards.
* Used Bugzilla for tracking and managing software bugs and issues encountered during project development.
* Integrated Snowflake cloud data warehouse into the data pipeline for scalable storage and analytics capabilities.
* Used Snowflake's features for data warehousing, data sharing, and advanced analytics to support decision-making processes.
* Transferred data between Hadoop and relational databases using Apache Sqoop to leverage the strengths of both environments.
* Configured Sqoop jobs to import/export data between Hadoop Distributed File System (HDFS) and relational database management systems (RDBMS).
* Basic knowledge of Apache Beam for building data processing pipelines.
* Performed data manipulation and analysis tasks in Excel using advanced functions, formulas, and pivot tables.
* Collaborated with team members and stakeholders by sharing project documents and updates through SharePoint

**Tech Stack:** Python, Pyspark, SQL, Hadoop, Spark, Kafka, Relational Database, No SQL Databases(MongoDB, Cassandra), Apache Airflow, Git, Hive, Snowflake, ETL, Talend, Numpy, Pandas, Excel, Tableau, Bugzilla, Sharepoint, Beam, Sqoop

**Educational Details:** Master’s in Computer Science