Data engineer
vinodreddy@gmail.com
Vinod Reddy
(513) 428-5051

PROFESSIONAL SUMMARY:

- Overall 5+ years of hands-on experience as a Data Engineer in Data Warehousing, Data Integration, and Data Modelling, including design, development, coding, testing, bug fixing, and production support of data warehousing applications.
- Experienced in all stages of the software development life cycle (SDLC), as well as Agile Software Methodology.
- Skilled with T-SQL, SQL, PL/SQL packages, functions, stored procedures, triggers, and materialized views to implement business logic in the Oracle database.
- Experienced in Software/Application Development using Python, Scala, C, and SQL, and in-depth understanding of Distributed Systems Architecture and Parallel Processing Frameworks.
- Excellent experience with Python libraries such as Requests, NumPy, Matplotlib, SciPy, PySpark, and Pandas during the development lifecycle, as well as experience developing APIs for the application using Python, Django, MongoDB, Express, ReactJS, and NodeJS.
- Experienced in gathering requirements, analysing requirements, designing, developing, testing, and implementing business intelligence solutions using data warehouse/data mart design/data modeling, ETL, OLAP, OLTP, BI, and client/server applications.
- Well-versed in Big Data installation, configuration, support, and management, as well as the underlying infrastructure of the Hadoop Cluster.
- Experience in manipulating and analyzing large datasets, as well as identifying patterns and insights in structured and unstructured data.
- Excellent working knowledge of Star Schema, Snowflake, and Sqoop for importing data from RDBMS into HDFS.
- Expertise in working complex DAX functions in Power BI.
- Worked on calculated columns and measures queries in Power BI desktop to show good data analysis techniques.
- Able to administer and monitor Snowflake computing platform.
- Deep understanding of MapReduce with Hadoop and Spark. Good knowledge of Big Data ecosystems like Hadoop 2.0 (HDFS, Hive, Pig, Impala), and Spark (Spark SQL, Spark MLlib, Spark Streaming).
- Expert in data validation against files and technical data quality checks to certify source and target/business usage.
- Proficient in leveraging Azure Databricks and Spark for distributed data processing and transformation tasks.
- Skilled in ensuring data quality and integrity through validation, cleansing, and transformation operations.
- Experienced in Azure Data Factory and preparing CI/CD scripts for the deployment.
- Solid experience on building ETL ingestion flows using Azure Data Factory.
- Experience in building ETL (Azure Data Bricks) data pipelines leveraging PySpark, Spark SQL.
- Hands on experience in implementing data pipeline solutions using Hadoop, azure, ADF, Synapse, Pyspark, Map-Reduce, Hive, Tez, Python, Scala, Azure functions, Logic apps, stream sets, ADLS Gen2 and Snowflake.
- Execution of Automated test scripts and all higher environments before PROD.
- Experience with Apache Spark, Scala, and Python to convert Hive/SQL queries into RDD transformations.
- Proficiency in writing UNIX shell scripts for Data Warehouse jobs, file operations, and data analytics.
- Strong experience with tuning and debugging existing ETL processes.
- Ensured data quality and integrity by performing data validation, cleansing, and transformation operations using Azure Data Factory and Databricks.
- Experience developing web services (WSDL, SOAP, and REST) and consuming web services with Python.
- Expertise in enabling CI/CD processes using Jenkins for the integration of all object deployments such as ETL code, SQL files, Shell scripts, and Python code base files, among others.
- Extensive experience in designing, developing, and deploying ETL processes using Talend for efficient and automated data extraction, transformation, and loading.
- Proficient in writing SQL queries specifically optimized for Snowflake, ensuring efficient data ingestion into Snowflake environments.
- Proficient in Microsoft BI platform technologies including SQL Server, SSIS, SSRS, Azure Data Factory, and Power BI.
- Extensively used Informatica client tools Source Analyzer, Warehouse Designer, Mapping Designer, Mapplet Designer, Transformations, Informatica Repository Manager, and Informatica Server Manager.

TECHNICAL SKILLS:

Programming Languages	Python, SQL, PL/SQL, T-SQL, Unix, Shell, YAML, PySpark
Cloud Technologies	Azure, Amazon Web Services, Snowflake

Big Data Technologies	Hadoop, HDFS, HDInsight, Map Reduce, YARN, Pig, HBase, Spark, Zookeeper, Hive, Oozie, Sqoop, Flume, Kafka, Scala
Schedulers	Airflow, Oozie, TIDAL
IDEs	Sublime Text, PyCharm, Eclipse, and NetBeans, Visual Studio.
Version Controls	GitHub and Bitbucket
Databases	Snowflake, Teradata, Oracle & MySQL, SQL Server, MongoDB, Azure Synapse, MS Excel
Operating Systems	Windows, Linux, Unix, Centos, Ubuntu
Azure Services	Azure Data Factory, Azure Data Bricks, snowflake, Logic Apps, Functional App

WORK EXPERIENCE:

Molina Healthcare, Long Beach, CA **Aug 2021 – Present**
Senior Data Engineer

Responsibilities:

- Worked on all the Azure data factory pipeline with different cases i.e. Truncate load, Incremental load, Insert Update load and automate them as per the business requirements.
- Data Ingestion on premise databases to cloud migration and processing the data in Azure Databricks.
- Involved in the development and testing phases of SDLC.
- Imported data from file-based systems and relational databases into the azure datalake storage in standard file formats such as Apache Parquet using Azure Data Factory and Azure Databricks.
- Hands-on coding - Write and test the code for the Ingest automation process - Full and Incremental Loads.
- Design the solution and develop the scripts for data ingestion using PySpark & Spark SQL in Azure Databricks and orchestrate them by using the data factory pipelines.
- Extract Transform and Load data from Sources system to Azure Data Storage Services using a combination of Azure Data Factory, T-SQL, Spark SQL.
- Worked on migration and conversion of data using Pyspark and Spark SQL for data extraction, transformation and aggregation from multiple file formats for analyzing and transforming from DataBricks Notebooks using Python.
- Enabling monitoring and azure log analytics to alert support team on usage and stats of the daily runs.
- Designed and developed Azure logic apps to trigger emails whenever a pipeline failure occurs in the azure data factory pipelines.
- Developed various automated scripts for DI (Data Ingestion) and DL (Data Loading) using PySpark.
- Hands on Experience on Unified Data Analytics with Data Bricks, Databricks workspace user interface, Managing Databricks Notebooks, Delta Lake With Python, Delta Lake with Spark SQL
- Experience working with Spark SQL and creating RDD's using Pyspark Spark Context & Spark Session.
- Exposed transformed data in Azure Spark Databricks platform to Apache Parquet, and Delta file formats for efficient data storage.
- Developed Json Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data using the Cosmos Activity.
- Create and maintain optimal data pipeline architecture in cloud Microsoft Azure using Data Factory and Azure Databricks.
- Used advanced features of T-SQL in order to design and tune T-SQL to interface with the Database and other applications in the most efficient manner and created stored Procedures for the business logic using T-SQL.
- Experience in Developing Spark applications using Spark - SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
- Worked on migration for large amount of data from OLTP to OLAP by using ETLPackages.
- Developing custom stored procedures for delta loads, functions, triggers using SQL, T-SQL on cloud SQL server/Azure Synapse.
- Performing research to identify source and nature of data required for ETL solutions using Azure Databricks.
- Migrate the entire CRM database present on the IBM DB2 servers, Informix to cloud based data warehouse called Snowflake using ETL DataStage.
- Identified data quality issues and design processes and procedures to improve/maintain quality.
- Performance tuning of SQL queries, Data Pipelines, Tableau and Power BI Dashboards.
- Maintaining version control of code using Azure Devops and GIT repository.
- Performing Code release from one environment to other environment using release management in Azure Devops.

Environment: Azure Data Factory, Azure Data Lake, Azure Synapse Analytics(DW), Azure Devops, Snowflake, PowerBI, SharePoint, Spark, PySpark, Hadoop, Hive, HDFS, PyTest, Kafka, MySQL, Eclipse, Jira, GitHub, Jenkins, PyCharm.

OSF Healthcare, Burbank, IL **Jan 2019- July 2021**
Azure Snowflake Data Engineer

Responsibilities:

- Implemented end-to-end data pipelines using Azure Data Factory to extract, transform, and load (ETL) data from diverse sources into Snowflake.
- Designed and implemented data processing workflows using Azure Databricks, leveraging Spark for large-scale data transformations.
- Built scalable and optimized Snowflake schemas, tables, and views to support complex analytics queries and reporting requirements.
- Working on Azure Synapse Analytics for implementing Pyspark Notebooks.
- Developed data ingestion pipelines using Azure Event Hubs and Azure Functions to enable real-time data streaming into Snowflake.
- Experience in identifying critical information in process automation, supply chain analytics, building vendor relationships, manufacturing losses reduction, marketing analytics, and building claims forecasting models.
- Implemented a CI/CD framework for data pipelines using the Azure Devops, enabling efficient automation and deployment.
- Collaborated on ETL tasks, maintaining data integrity and verifying pipeline stability.
- Designed and implemented scalable and distributed data solutions using Azure Cosmos DB, a globally distributed, NoSQL database service, to efficiently store semi-structured, and unstructured data.
- Designing and architecting Log Analytics solutions to meet specific business and technical requirements.
- Leveraged Azure Data Lake Storage as a data lake for storing raw and processed data, implementing data partitioning and data retention strategies.
- Integrated Azure Data Factory with Azure Logic Apps for orchestrating complex data workflows and triggering actions based on specific events.
- Complete responsibility of code deployments via CICD to test, SIT, UAT, PSUP and PROD.
- Implemented data governance practices and data quality checks using Azure Data Factory and Snowflake, ensuring data accuracy and consistency.
- Implemented data replication and synchronization strategies between Snowflake and other data platforms using Azure Data Factory and Change Data Capture techniques.
- Developed and deployed Azure Functions for data preprocessing, data enrichment, and data validation tasks in data pipelines.
- Integrated Azure Logic Apps with other Azure services such as Azure Functions, Azure Service Bus, and Azure Storage, leveraging their capabilities to enhance data processing, message handling, and storage within the workflows.
- Monitoring and optimizing the performance of Log Analytics queries and data retrieval processes.
- Involved in converting Hive/SQL queries into transformations using Python.
- Automated and created templates for deployment of internal applications to Dev, Test and Production environments.
- Designed and implemented data archiving and data retention strategies using Azure Blob Storage and Snowflake's Time Travel feature.
- Collaborated in Agile Scrum Methodology, actively participating in daily stand-up meetings. Proficiently utilized Visual SourceSafe for Visual Studio 2010 and effectively tracked projects using Trello.
- Integrated Snowflake with Power BI and Azure Analysis Services for creating interactive dashboards and reports, enabling self-service analytics for business users.
- Optimized data pipelines and Spark jobs in Azure Databricks for improved performance, including tuning of Spark configurations, caching, and leveraging data partitioning techniques.
- Developed CI/CD framework for data pipelines using Jenkins tool.
- Collaborated with DevOps engineers to develop automated CI/CD and test-driven development pipeline using Azure as per the client requirement.
- Hands on programming experience in scripting languages like Python and Scala.
- Involved in running all the Hive scripts through Hive on Spark and some through SparkSQL
- Developed a data pipeline using Kafka, Spark, and Hive to ingest, transform, and analyze data.
- Developed Spark core and Spark SQL scripts using Scala for faster data processing.
- Working with JIRA to report on Projects, and creating sub tasks for Development, QA, and Partner validation.

Environment: Azure Databricks, Data Factory, Logic Apps, Snowflake, Functional App, MS SQL, T-SQL, Oracle, HDFS, MapReduce, Spark, Hive, SQL, Python, Scala, PySpark, shall scripting, GIT, JIRA, Jenkins, Kafka, ADF Pipeline, Power Bi.

Flexera, Itasca, IL **Jan 2018 – Dec 2019**
Data Engineer

Responsibilities:

- Participated in architecture and requirement analysis talks.
- Collaborated with internal and external business partners to gather requirements.
- Involved in designing MVP plan for data migration from Hadoop Ecosystem to AWS cloud.
- Deployed AWS Data Sync agent near to HDFS cluster hosted on Azure VM to test out the data copy process from HDFS to AWS S3.
- Developed technical design documents, target to source mapping document, and mapping specification document based on business requirements.
- Extensively worked on Informatica PowerCenter.
- Unpacked the existing mapping logics in informatica PowerCenter to recreate the mappings in IICS.
- Worked on Informatica PowerCenter - Source Analyzer, Mapping Designer, Workflow Manager, Mapplet Designer and Transformation Developer.
- Created Complex mappings using Unconnected and connected Lookups and Aggregate and Router transformations for populating target table in efficient manner.
- Created mapplets, worklets and reused in mappings and workflows accordingly.
- Developed python 3 scripts to trigger the informatica workflows.
- Used Hue to interact with the Hadoop Ecosystem.
- Experienced in using Hive to process and analyze data.
- Adept in following Agile methodology and successful in meeting the fast packed bi-weekly sprint deadlines.

Environment: Python3, DB2, AWS s3, AWS Redshift, AWS Data Sync, AWS EC2, Control M, Linux, shell, HDFS, Hue, Hive, Azure VM, Informatica PowerCenter 10.2, IICS, SQL, JIRA, Cognos, Tableau, PyCharm, Jupiter Notebooks.

Education:

Masters in Computer Science from University of Dayton