**PAVAN KUMAR**

***Azure Data Engineer***

**Mobile:** 470-236-0321

**Mail:** [**Pavanka321@gmail.com**](mailto:Pavanka321@gmail.com)

**Professional Summary:**

* 10+ years of experience in Data Engineering, specializing in designing and implementing scalable data ingestion pipelines using Azure Data Factory, and proficient in SQL for data manipulation and querying.
* Skilled in ensuring data quality and integrity through validation, cleansing, and transformation operations, with a deep understanding of SQL in database operations.
* Proficient in designing scalable data processing workflows and advanced analytics solutions using Azure Databricks, leveraging machine learning algorithms to enhance clinical decision-making.
* Proficient in Snowflake Multi-Cluster Warehouses, understanding Snowflake cloud technology, and utilizing features like Snowflake Clone and Time Travel for data versioning and recovery.
* Skilled in optimizing data warehouses in Snowflake, implementing partitioning, indexing, and caching strategies to ensure high performance and efficient data retrieval.
* Implemented comprehensive data quality tools and governance policies, ensuring data integrity, consistency, and compliance with healthcare regulations such as HIPAA.
* Experienced in developing CI/CD frameworks for data pipelines using Jenkins, integrating Snowflake SQL scripts and Python-based automation to enhance data processing efficiency.
* Designed, scheduled, and monitored data pipelines in Apache Airflow for efficient workflow orchestration.
* Developed custom operators and sensors within Airflow for seamless integration with various data sources and external systems.
* Developed and optimized complex T-SQL stored procedures, functions, and triggers to support business logic and data processing tasks, ensuring efficient and maintainable code.
* Expertise in writing complex T-SQL scripts for ETL processes, conducting performance tuning, and optimizing database operations to ensure data integrity and optimal performance.
* Proven expertise in actively contributing to the development, enhancement, and upkeep of Snowflake database applications.
* Extensive expertise in optimizing Spark jobs and leveraging Azure Synapse Analytics for big data processing and analytics, incorporating SQL for querying and analysis.
* Expert in Real-Time Data Processing with strong proficiency in Apache Kafka, enabling the ingestion, processing, and streaming of large-scale datasets in real-time environments.
* Proven track record in performance optimization and capacity planning using SQL techniques to ensure scalability and efficiency.
* High proficiency in Agile methodologies, including JIRA for project management and reporting, with extensive experience in using SQL for data analysis and reporting.
* Additional expertise in using SQL for optimizing query performance in Hive using bucketing and partitioning techniques, with extensive hands-on experience tuning Spark jobs.
* Expertise in implementing Data Vault modeling for scalable and flexible data warehouse architecture, ensuring historical data tracking and auditability.
* Proficient in using Matillion for ETL processes, enabling efficient data integration and transformation workflows in cloud environments.
* Skilled in designing dimensional models for data warehouses, ensuring efficient and intuitive reporting and analytics.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Azure Services** | Azure data Factory, Azure Data Bricks, Logic Apps, Function App, Azure Active Directory, Azure Key-Vault, Azure Purview. |
| **AWS Services** | AWS Glue, Amazon S3, Amazon Lambda, Amazon Redshift, Amazon EMR, Amazon RDS. |
| **Snowflake** | SnowSQL, Snowpipe, Snowpark |
| **Big Data Technologies** | MapReduce, Hive, PySpark, Scala, Kafka, Spark streaming, Oozie, Sqoop, Zookeeper, Pig |
| **Hadoop Distribution** | Cloudera, Horton Works |
| **Languages** | SQL, PL/SQL, Python, HiveQL, Scala. |
| **Operating Systems** | Windows (XP/7/8/10), UNIX, LINUX, UBUNTU, CENTOS. |
| **Build Automation tools** | Ant, Maven |
| **Version Control** | GIT, GitHub. |
| **ETL Tools** | Informatica, SSIS, Talend, Airflow, Data Stage, DBT |
| **Databases** | MS SQL Server 2016/2014/2012, Azure SQL DB, Azure Synapse. MS Excel, MS Access, Oracle 11g/12c, Cosmos DB |

**Professional Experience:**

**Client: Wells Fargo, Charlotte, NC**  **Feb 2022 - Present**

**Role: Azure Data Engineer**

* Developed robust data ingestion pipelines using Azure Data Factory and Python to extract patient data from various sources such as Electronic Health Records (EHR) systems, medical devices, IoT sensors, and NoSQL databases into Azure Data Lake Storage.
* Developed and maintained custom connectors in Azure Data Factory to integrate with various third-party APIs, NoSQL databases, and data sources, extending the ETL capabilities and enabling seamless data ingestion.
* Integrated SQL Auto Loader in data ingestion processes to streamline and automate the detection and loading of new files, enhancing the efficiency of data pipeline workflows and reducing manual intervention.
* Integrated Azure Logic Apps into data workflows to automate and orchestrate data processing tasks across Azure services, improving workflow efficiency and reducing manual intervention.
* Integrated Azure Data Factory with Azure Key Vault to securely manage and access sensitive configuration settings and credentials, enhancing the security of ETL processes.
* Conducted performance tuning and optimization of SQL queries, indexes, and database configurations to enhance the efficiency and speed of database operations.
* Implemented data processing workflows utilizing Azure Databricks and Python to clean, transform, and enrich raw patient data, ensuring data quality and consistency for downstream analytics and reporting.
* Managed and maintained the Unity Catalog to ensure secure and organized data governance.
* Implemented role-based access control (RBAC) for data access management using Unity Catalog.
* Developed scalable data processing workflows in Azure Databricks to handle petabyte-scale datasets, ensuring efficient resource utilization and minimal processing time.
* Developed data processing workflows using Azure Databricks and Python, integrating Snowflake for distributed data processing and transformation tasks, and implemented slowly changing dimension (SCD) techniques including type 2 and type 3 dimensions.
* Developed and executed Spark SQL jobs in Python to perform complex data transformations and aggregations on large-scale datasets, enabling the population of analytical data models with enriched and cleansed data.
* Utilized Spark SQL for performing ad-hoc querying and data exploration on ingested data, ensuring the accuracy and completeness of data models designed for business intelligence and reporting.
* Designed and executed data quality assessments and audits, leveraging Ataccama to monitor data quality metrics and address discrepancies promptly.
* Collaborated with data scientists to operationalize machine learning models within Azure Databricks, facilitating real-time analytics and decision support.
* Designed and optimized data warehouses in Snowflake to support complex analytical queries, improving data retrieval times and enabling real-time analytics.
* Utilized Snowflake's Time Travel and cloning features to create point-in-time copies of data, facilitating robust data backup and recovery strategies.
* Implemented data sharing and data exchange features in Snowflake to securely share data across different business units and external partners, enhancing collaboration and data-driven decision-making.
* Integrated Azure Logic Apps for orchestrating workflows across Azure services, improving automation and streamlining data processing tasks.
* Established data governance policies and implemented security measures such as role-based access control (RBAC) and encryption to ensure compliance with healthcare regulations (e.g., HIPAA) and protect sensitive patient information stored in Azure Data Lake.
* Integrated Epic Cognito for streamlined authentication and authorization processes within data pipelines, ensuring secure access to sensitive healthcare data in compliance with industry regulations such as HIPAA.
* Created and managed insightful Power BI reports and dashboards to facilitate data-driven decision-making for clinical operations and business management.
* Developed CI/CD framework for data pipelines using Jenkins tool, integrating Snowflake SQL scripts and Python-based automation, emphasizing bulk loading and unloading data into Snowflake tables.
* Implemented Data Vault 2.0 architecture to provide a scalable and agile approach to data warehousing, ensuring historical data tracking and auditability.
* Leveraged Matillion ETL for transforming data and integrating with Snowflake, improving data pipeline efficiency and reducing processing time.
* Designed and implemented Dimensional Modeling techniques, including star and snowflake schemas, to optimize data warehousing solutions and enhance query performance.
* Developed and optimized complex T-SQL stored procedures, functions, and triggers to support business logic and data processing tasks, ensuring efficient and maintainable code.
* Implemented advanced SQL techniques for partitioning, indexing, and query optimization to ensure high performance and efficient data retrieval in Snowflake and other relational databases.

**Environment**: Azure Data Lake Storage Gen 2, Azure Data Factory, Azure Databricks, Azure Logic Apps, Azure Function Apps, Ataccama, Apache Spark, Python, Scala, SQL, Oracle, Hive, Jira, Power BI.

**Client: Express scripts pharmaceuticals Oct 2020 – Jan 2022**

**Role**: **Azure Data Engineer**

* Developed and optimized SQL queries and stored procedures within Azure Data Factory and SSIS to extract, transform, and load data, ensuring efficient processing and adherence to data warehousing best practices.
* Created T-SQL scripts for automating routine database maintenance tasks such as index rebuilding, statistics updates, and backup operations, ensuring database health and performance.
* Utilized T-SQL's advanced analytical functions such as window functions, common table expressions (CTEs), and pivoting to perform complex data analyses and generate insightful reports.
* Implemented Azure Logic Apps to automate workflows and orchestrate data integration tasks between disparate systems, enhancing operational efficiency and reducing manual effort.
* Performed performance tuning of Snowflake data warehouse using Query Profiler, caching mechanisms, and virtual data warehouse scaling to optimize query execution and enhance overall system performance.
* Developed data quality dashboards and reports to provide insights into data quality trends and issues, supporting data governance and stewardship efforts.
* Implemented data governance frameworks and policies to ensure data quality and compliance, collaborating with stakeholders to establish data standards and best practices.
* Developed and deployed Python-based Azure Functions for data preprocessing, enrichment, and validation tasks within the data pipelines, enhancing data quality and integrity.
* Integrated Python-based monitoring scripts with Azure Monitor for real-time tracking of pipeline performance and proactive issue identification, ensuring smooth operation of data processes.
* Designed and implemented PySpark jobs for data transformation tasks, leveraging its capabilities for parallel processing and in-memory computation to improve overall pipeline performance.
* Optimized Power BI reports for performance, ensuring fast load times and responsiveness by managing data granularity, using aggregations, and optimizing data models.
* Wrote complex DAX (Data Analysis Expressions) formulas to support advanced calculations and business logic within reports.
* Collaborated with data science teams to integrate PySpark-based machine learning workflows into data pipelines, enabling scalable model training and inference on large datasets.
* Optimized PySpark jobs for performance tuning parameters such as partitioning, caching, and serialization to maximize processing speed and resource utilization.
* Developed monitoring and logging solutions for PySpark jobs, integrating with existing monitoring frameworks for real-time performance tracking and issue detection.

**Environment**: Azure Databricks, Data Factory, Logic Apps, Snowflake, Functional App, Snowflake, MS SQL, Oracle, HDFS, Kafka, Spark, Hive, SQL, Python, Scala, PySpark, shell scripting, GIT, JIRA, Power BI.

**Client: State Of Arizona, Phoenix, AZ May 2017 – Sep 2020**

**Role: Big Data Engineer**

* Prepared an ETL framework using Sqoop, Pig, and Hive to bring in data from various sources and make it available for consumption.
* Processed HDFS data and created external tables using Hive along with developing scripts for table ingestion and repair for reuse across the project.
* Developed and maintained integration solutions to import and export data between SQL Server and external systems using APIs, Linked Servers, and other connectivity options.
* Created, optimized, and maintained complex stored procedures and user-defined functions to encapsulate business logic and improve application performance.
* Established comprehensive audit and logging mechanisms to track data changes, user activities, and system performance for security and compliance purposes.
* Developed ETL jobs using Spark and Scala to migrate data from Oracle to new MySQL tables.
* I analyzed source data, efficiently handled data type modifications, and used Excel sheets, flat files, and CSV files to generate Power BI ad-hoc reports.
* Extensive experience in optimizing SQL queries for improved performance.
* Designed and implemented complex stored procedures, triggers, and functions using SQL Server.
* Handled data import from various sources, performed transformations using Hive and MapReduce, and loaded data into HDFS.
* Implemented automation for deployments using YAML scripts for streamlined builds and releases.
* Implemented data classification algorithms using MapReduce design patterns.
* Extensively worked on creating combiners, partitioning, and distributing cache to enhance the performance of MapReduce jobs.
* Utilized Git and GitHub repositories to maintain the source code and enable version control.

**Environment**: Hadoop, Hive, spark, PySpark, Sqoop, Spark SQL, Shell script, Cassandra, YAML, ETL.

**Client: Global Payments, Atlanta, GA Apr 2016 – Mar 2017**

**Role: Big Data Engineer**

* Experience in developing complex store procedures, efficient triggers, required functions, creating indexes and indexed views for performance.
* Excellent Experience in monitoring SQL Server Performance tuning in SQL Server
* Expert in designing ETL data flows using SSIS, creating mappings/workflows to extract data from SQL Server and Data Migration and Transformation from Access/Excel Sheets using SQL Server SSIS.
* Efficient in Dimensional Data Modeling for Data Mart design, identifying Facts and Dimensions, and developing fact tables, dimension tables, using Slowly Changing Dimensions (SCD).
* Experienced in Building Cubes and Dimensions with different Architectures and Data Sources for Business Intelligence and writing MDX Scripting.
* Thorough knowledge of Features, Structure, Attributes, Hierarchies, Star and Snowflake Schemas of Data Marts.
* Developed DataMarts for specific business units, enhancing data accessibility and analysis.
* Utilized dimensional data modeling for DataMart design, including identifying facts and dimensions.
* Worked on the creation of Ad hoc reports and reports with complex formulas for DataMarts using SSRS.
* Expertise in developing Parameterized, Chart, Graph, Linked, Dashboard, Scorecards, Report on SSAS Cube using Drill-down, Drill-through and Cascading reports using SSRS.
* Flexible, enthusiastic and project-oriented team player with excellent written, verbal communication and leadership skills to develop creative solutions for challenging client needs.

**Environment**: MS SQL Server 2016, Visual Studio 2017/2019, SSIS, Share point, MS Access, Team Foundation server, Git.

**Client: Costco, Dallas, TX Feb 2013 – Mar 2016**

**Role: Data Warehouse Developer**

* Used BULK SQL and BULK BINDING to minimize the execution time of PL/SQL code.
* Created SQL Loader Control Files for moving data from Flat Files (Fixed Record Length) to Staging Area Tables.
* Analyzed the existing Stored Procedures, Functions, Triggers and Cursors on performance issues.
* Managing Tables, Indexes, Constraints, Views, Sequences, and stored program units.
* Used the Oracle Import/Export utilities for taking the backup of a particular partition data.
* Gathered Statistics and Analyzed Tables and Indexes for Performance tuning.
* Made use of DB LINKS for accessing data across different databases.
* Resolved various optimizing problems to maintain a high level of customer satisfaction.
* Created user’s privileges and managed user roles and grants.
* Worked on DBMS\_SCHEDULER to automate jobs.
* Created Stored Procedures, Functions, and Triggers to maintain various business rules.
* Extensively worked towards optimizing the Queries to enhance performance.
* Analyzed Indexes on regular basis.
* Loading data into the Database using SQL LOADER
* Performed debugging of the PL/SQL codes using the DBMS\_OUTPUT

**Environment**: Oracle 10.2g, HP-Unix, Oracle Enterprise Manager, VSS, SQL Loader, SQL Plus, SQL Developer 1.5.4, VB.net, Toad Data Modeler.