**Venkata Pavan Kumar Upputuri**

Azure Data Engineer

**Phone: +1 919 454 0528**

**Email:** **venpavankumar@outlook.com**

**LinkedIn: http://www.linkedin.com/in/pavan-kumar0528**

**PROFESSIONAL SUMMARY**

* With over **10+ years of expertise in Data Engineering**, I specialize in architecting and implementing high-performance **Azure, Snowflake, Big Data, and ETL solutions** that optimize data workflows and drive business insights across **Banking, Finance, Healthcare, and Insurance**.
* Proficient in **Azure Data Factory, Azure Databricks, Azure Data Lake Storage Gen2, Azure Logic Apps, Azure Functions, Azure Event Hubs, Azure Synapse Analytics, Azure Active Directory, Azure SQL Database, Azure Cosmos DB, Azure Stream Analytics, Azure Blob Storage, Azure Purview, Power BI, Azure Analysis Services, Azure Machine Learning, Azure Monitor, Azure Kubernetes Service (AKS), Azure DevOps, Azure Cognitive Services** for end-to-end data solutions.
* Designed and implemented **role-based access control (RBAC)** in **Azure Synapse, ADLS Gen2, Unity Catalog, Delta Parquet files, Power BI, and Azure Active Directory**, ensuring secure and governed data access for analytics.
* Successfully migrated legacy data from **Teradata to Azure Synapse Analytics** and on-premises systems to **Azure Cloud** using **Azure Functions, SSIS, Azure Analysis Services, Microsoft Fabric, and Azure Active Directory** for seamless authentication.
* Integrated **Azure Data Lake Storage (ADLS) Gen2** with **Azure Synapse Analytics** for distributed processing, leveraging **Informatica PowerCenter and SQL Server Integration Services (SSIS)** for ETL transformations.
* Built scalable **Azure data solutions** using **Synapse Analytics, Microsoft Fabric, Databricks, Event Hubs, Delta Parquet files, and Power BI** for efficient governance, streaming, and analytics.
* Automated data pipelines with **Azure Databricks, ADF, Synapse, ETL workflows, Azure SQL Data Warehouse, and Azure Analysis Services**, ensuring seamless integration and governance via **Azure Purview**.
* Implemented **Medallion architecture** in Azure using **Synapse, ADLS Gen2, Delta Parquet, Power BI, DAX, and Purview** for scalable analytics, efficient processing, and robust governance.
* Hands-on experience with **CSV, JSON, Delta Parquet, Avro file formats** and tools like **T-SQL, Informatica, SSIS, SSRS** for data integration, transformation, and reporting.
* Integrated **ADLS Gen2 with Snowflake and Databricks** for ETL pipelines, implementing **SCD Type 2, DBT for modeling, and Power BI with DAX** for analytics.
* Orchestrated data pipelines across **Azure Databricks and Glue**, using **Azure Synapse Analytics, Redshift, Azure Logic Apps, Airflow DAGs, and Unity Catalog** for governed cloud warehousing.
* Designed **ETL workflows** with **Azure SQL Database, Synapse Analytics, Cosmos DB, and Azure Cache**, enhancing scalability via **Microsoft Fabric**.
* Migrated legacy data to **Azure Synapse and Redshift** using serverless architectures, integrating **Unity Catalog** for governance and **Power BI, DAX, T-SQL** for analytics.
* Executed **Teradata-to-Azure migrations** using **Informatica, SSIS, and Azure Functions**, ensuring serverless and scalable data solutions.
* Enforced **RBAC policies** in **Azure Synapse, Azure Key Vault** for data security and encryption in multi-cloud environments.
* Optimized query performance in **Azure Synapse** via clustering, partitioning, materialized views, and integrated **Kafka, Delta Tables, Hive, Databricks, PySpark, DataFrames, RDDs, JSON, DAGs, and APIs** for real-time processing.
* Improved **Snowflake query performance** with partitioning strategies and **SCD types** for historical data tracking.
* Connected **Spark** with **ADLS, Blob Storage, Snowflake, Delta Parquet, and Power BI**, leveraging **Microsoft Fabric** for scalable insights.
* Collaborated with **finance, healthcare, and education stakeholders** to deploy **Snowflake and Azure Databricks** solutions, using **Power BI and DAX** for real-time analytics.
* Implemented **Snowpipe** for real-time ingestion and **CDC with SCD Type 2** in Snowflake.
* Maintained **Azure ETL processes** with **Microsoft Fabric, Unity Catalog, Azure Data Catalog, and Medallion architecture**.
* Developed data pipelines using **Python (Pandas, NumPy, sci-kit-learn), Apache Spark, and Java** for large-scale transformations.
* Automated workflows with **Azure Data Factory, Event Grid, DevOps (CI/CD), and Azure Functions** for serverless event-driven processes.
* Expertise in **Big Data (Hadoop, Spark), NoSQL databases**, optimizing query performance and data transmission.
* Migrated legacy data to **Snowflake** using **DBT for data modeling and transformation, SCD Type 2 for historical tracking, Azure Data Factory** for seamless workflow orchestration, and **Power BI (DAX)** for advanced analytics.
* Partnered with **DevOps teams** using **Azure DevOps, Jenkins** for CI/CD automation.
* Agile/Scrum collaboration with cross-functional teams, ensuring **data quality, governance, and stakeholder alignment**.
* **Enhanced data quality frameworks** with **Great Expectations**, achieving **99%+ accuracy** in ETL pipelines.
* **Reduced cloud costs by 25%+** via cluster tuning, query optimization, and job orchestration.
* Integrated **hybrid cloud pipelines** across **Azure and AWS (Glue, Redshift, S3)**.
* Enforced **SOC2/HIPAA/GDPR compliance** with encryption, masking, and access controls.
* Deployed **Azure Monitor, Log Analytics, and custom alerts** for real-time pipeline observability.
* Led **4+ junior engineers**, providing mentorship in **ETL best practices and CI/CD**.
* Designed **data mesh architectures** using **domain-driven design** principles.
* Implemented **GitOps workflows** with **Azure DevOps, GitHub Actions, and Vault**.
* Delivered **POCs for Microsoft Fabric, Snowpark, and Delta Live Tables** to drive innovation.
* Architected **HA/DR strategies** with geo-redundant storage and automated failovers.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **Category** | **Skills** |
| **Azure Services** | Azure Data Factory (ADF), Azure Databricks, Azure Data Lake Storage (ADLS) Gen2, Azure Logic Apps, Azure Active Directory, Azure Functions, Azure Event Hubs, Azure ETL, Azure SQL DW, Azure Analysis Services Azure Fabric, Azure Purview, Azure Functions, Azure Synapse Analytics, Azure SQL Server, Microsoft Fabric. |
| **Big Data Technologies** | HDFS, Sqoop, Hive, MapReduce, Spark, HBase |
| **Hadoop Distribution** | Cloudera, Horton Works |
| **Languages** | Python, Java, SQL, PL/SQL, HiveQL, Scala, T-SQL, PySpark  |
| **Operating Systems** | Windows (XP/7/8/10), UNIX, LINUX, UBUNTU, CENTOS |
| **Databases** | Teradata, Oracle DB, PostgreSQL, SQL server, MSSQL, MS Access |
| **Scheduling** | IBM Tivoli, Control-M, Oozie, Airflow |
| **Version Control** | GIT, GitHub, Bitbucket, VSS |
| **Methodology** | Agile, Scrum |
| **IDE &Build Tools, Design** | Eclipse, PyCharm IDE, Visual Studio |
| **Analysis and Visualization** | Power BI, Tableau, QuickSight |

**EDUCATION:**

* **Master’s: TRINE University**
* **B. Tech: K.G Reddy College of Engineering and Technology**

**CERTIFICATIONS:**

* **Microsoft Certified: Azure Data Engineer Associate**
* **SnowPro Core Certification**
* **Databricks Certified Data Engineer Associate**

**WORK EXPERIENCE**

**Role: Azure Data Engineer | May 2022 – Till Now**

**Client: Wells Fargo, Charlotte, NC**

**Responsibilities:**

* Designed and deployed ETL workflows using Azure Data Factory (ADF), Azure Synapse, Azure SQL Data Warehouse, and Azure Analysis Services, integrating streaming data pipelines with Kafka, Event Hubs, Databricks, and Azure Active Directory to enhance fraud detection and deliver reliable analytics.
* Optimized storage efficiency by 25% by leveraging JSON, Avro, Parquet, and ORC formats in Azure Data Lake Storage (ADLS) Gen2. Implemented Unity Catalog for centralized governance, ensuring data consistency across formats. Utilized Azure Data Factory, Databricks, PySpark, and T-SQL to streamline data processing, improving storage and performance.
* Architected Azure cloud solutions, deploying ADF, Data Lake, SQL Data Warehouse, Databricks, Synapse, Spark, Python, Azure Analysis Services, DevOps, ETL pipelines, and Azure Active Directory to enhance scalability, security, flexibility, and cost efficiency.
* Enhanced Azure Synapse Analytics performance by implementing SCD Type 2 and Change Data Capture (CDC). Employed DBT for data modeling, Delta Parquet for efficient storage, Power BI with DAX for advanced analytics, and Azure Databricks with Kafka to optimize streaming workflows.
* Managed the implementation of Azure Purview adoption for enterprise-wide data governance, overseeing discovery, cataloging, and compliance. Maintained Purview’s Data Catalog, enriching assets with business glossaries, classifications, and lineage tracking to ensure transparency and regulatory adherence.
* Automated infrastructure provisioning using Terraform, integrating Azure Blob Storage, Synapse Analytics, SQL Database, ADF, and Azure Functions to streamline data engineering workflows in a serverless and scalable architecture.
* Integrated ML models into Azure Machine Learning, leveraging ADF, Databricks, and Kafka for real-time processing, driving $100K in revenue growth through enhanced data insights.
* Engineered data ingestion and storage frameworks using Azure Blob Storage, Data Lake, Synapse, ADF, SQL Data Warehouse, Databricks, Spark, Python, T-SQL, Analysis Services, and Azure DevOps, ensuring scalability, security, and access management.
* Implemented Snowflake RBAC for secure data access and utilized CDC & SCD Type 2 to maintain historical accuracy in analytics and reporting.
* Automated data transformations by integrating DBT into Terraform, deploying Azure Databricks for scalable processing. Orchestrated workflows using Apache Airflow with DAGs, optimizing task scheduling and pipeline monitoring.
* Expert in SQL databases (MSSQL, MySQL, Oracle, PostgreSQL, MongoDB) and cloud migration strategies, successfully transitioning on-premises systems to Azure, reducing costs while ensuring data integrity.
* Migrated databases to Azure using Informatica, SSIS, and Azure services, achieving 30% cost savings in on-prem infrastructure while maintaining data consistency and accessibility.
* Built multi-layered data pipelines (bronze, silver, gold): established linked services for connectivity (bronze), enabled real-time streaming via self-hosted IR (silver), and deployed advanced analytics for insights (gold). Developed APIs to enhance integration across layers.
* Optimized Azure infrastructure using Ansible playbooks, cutting manual effort by 30%. Integrated DBT, Kafka, Unity Catalog, Delta Parquet, DAX, and Power BI for efficient storage, processing, and governance.
* Designed SCD-optimized data models in Informatica, IICS, SSIS, ADF, Azure AD, and Airflow, ensuring historical accuracy and integrity.
* Deployed CI/CD pipelines with Jenkins, reducing time-to-deployment by 25% and enhancing release reliability.
* Leveraged Spark SQL Schema RDDs in Azure Databricks, reducing processing complexity by 20%. Managed pipelines with DBT & Kafka, incorporating SCDs for historical tracking and Snowflake for scalable analytics.
* Built Azure infrastructure via ARM templates, integrating Unity Catalog for governance, access control, and lineage tracking, ensuring scalability and security.
* Advanced data engineering on Azure Databricks & Microsoft Fabric, utilizing ADF, Airflow, Delta Parquet, and APIs for scalable workflows and seamless integration.
* Implemented Delta Lake & Data Lake solutions, integrating Microsoft Fabric, Power BI, DAX, and SQL to optimize pipelines and query performance.
* Automated Azure interactions using PySpark & Azure SDK, enhancing Delta Lake & Data Lake workflows for performance, cost efficiency, and reliability.
* Developed Power BI & Tableau dashboards, extracting insights from Delta Lake & ADLS Gen2, and applying SQL & DAX to improve data accessibility and visualization by 20%.
* Deployed Azure Synapse & SQL Database for ad-hoc analytics, accelerating insight extraction from ADLS Gen2 & SQL Server.
* Integrated Azure Event Hubs, Service Bus, and Microsoft Fabric with Delta Lake, Data Lake, and SQL Server in an Agile Scrum model, improving real-time processing and operational efficiency.
* Optimized Azure Synapse configurations, reducing query response times by 30% on ADLS Gen2.
* Enhanced event-driven workflows by 25% by integrating Azure Event Hubs & Service Bus for real-time processing.
* Leveraged Azure services (ADF, Databricks, ADLS Gen2, SQL DB, Event Hubs, Functions, Synapse) to extract, transform, and load data, creating tables/databases in Databricks via Glue Crawlers.
* Scaled Azure infrastructure dynamically using Terraform, optimizing resource utilization and costs. Automated workflows with T-SQL & API integrations.
* Processed real-time data with Azure Event Hubs, enhancing streaming analytics. Optimized pipelines using PySpark, Databricks, Python, and ADF.
* Deployed Kubernetes & Docker for auto-scaling and CI/CD, managing services via Kubernetes dashboard.
* Implemented Azure Monitor for resource tracking, alerting, and performance optimization.
* Automated data workflows with ADF, Oozie, Airflow, Microsoft Fabric, PySpark, Delta Parquet, and Power BI, reducing processing time by 30% and improving large-scale data operations.
* Configured ETL workflows using Microsoft Fabric & Unity Catalog for governance, access control, and lineage, with PySpark optimizing data transformations.
* Managed code repositories via Git & Azure DevOps, collaborating efficiently in PyCharm IDE for Python/PySpark development

**Environment**: Azure Data Factory (ADF), Azure Databricks, Microsoft Fabric, Azure Data Lake Storage (ADLS) Gen2, Azure Fabric, Azure Purview, Azure SQL Data Warehouse, Power BI, Azure Event Hubs, Azure Functions, Azure ETL, Azure Synapse Analytics, Delta Parquet, Snowflake, Unity Catalog, Kafka, Talend, Snow SQL, Snowpipe, Change Data Capture (CDC), Azure SQL, Snowflake Time Travel, Looker, DBT, Azure Blob Storage, Azure Data Lake Storage Gen2, Python, Azure Active Directory, Java, SQL Server, Informatica, IICS, Airflow, Terraform, Azure Analysis Services Kubernetes, Azure Monitor, Oozie, Git, Azure DevOps.

**Role: Azure Data Engineer | Aug 2019 – Apr 2022**

**Client: Truist, Raleigh, NC**

**Responsibilities:**

* Designed and implemented Snowflake data models using ADF, Data Lake, SQL Data Warehouse, Databricks, Synapse, Azure Active Directory, Spark, Python, Analysis Services, DevOps, and ETL pipelines, reducing unauthorized access by 25% through robust security integrations.
* Managed Snowflake tables for storage and processing, integrating Unity Catalog, Azure AD, and Delta Parquet files for governance. Leveraged Azure Databricks, Star Schema, Power BI, DAX, and T-SQL for advanced data modeling, transformations, and analytics.
* Engineered solutions using Snow SQL and Snowpipe, converting Talend Joblets to enhance Snowflake compatibility, boosting data processing efficiency by 20%. Developed custom data processing scripts in Python, Java, PySpark, and SQL, improving flexibility, and deployed API endpoints for seamless integration.
* Optimized Power BI reports and dashboards, refining DAX queries, data models, and refresh cycles to ensure faster load times and smoother user experiences.
* Enhanced Snowflake performance by implementing T-SQL optimizations, partitioning, and clustering. Developed ETL workflows with stored procedures and views, integrating ADF and Synapse while enforcing secure access via Azure AD.
* Built modular, reusable data models in DBT, ensuring efficient transformations and high data quality. Applied SCD Type 2 to track historical changes and align with business logic in data warehouses.
* Developed ETL pipelines with Unity Catalog, Azure AD, Delta Parquet, and Power BI for governance and analytics. Utilized Python in Azure Databricks for high-performance data processing.
* Improved Snowflake query efficiency by 30% via partitioning and multi-cluster warehouses, leveraging ADF, Data Lake, SQL DW, Azure AD, Databricks, Synapse, Spark, PySpark, Python, Analysis Services, DevOps, and ETL.
* Consulted on Snowflake Solution Architecture, specializing in design, development, and deployment. Implemented CDC in Talend to streamline delta loading into data warehouses and executed data migrations using Python and Snow SQL.
* Streamlined ETL workflows with Azure Synapse, ADF, DevOps, and T-SQL, integrating Kafka, Delta Parquet, Power BI, DAX, and APIs for real-time streaming, efficient processing, and advanced analytics.
* Integrated Snowflake with Azure Blob Storage and ADLS Gen2, enabling scalable storage and SCD time-travel capabilities for historical tracking.
* Automated real-time data ingestion using Snowpipe, reducing manual effort and improving system robustness by 20% through enhanced security strategies.
* Designed scalable CDC-enabled pipelines for real-time updates and SCD Type 2 for historical tracking. Transformed raw data into analytics-ready datasets using DBT.
* Boosted Snowflake query efficiency by 35% via caching optimizations. Migrated multi-state SQL Server data to Snowflake using Python, SnowSQL, ADF, Data Lake, Delta Parquet, SQL DW, Databricks, Synapse, Spark, Analysis Services, Power BI, DevOps, and Azure ETL.
* Optimized query performance by 30% using Performance Monitor, SQL Profiler, and Database Tuning Advisor. Leveraged ADF, Azure Databricks, and PySpark for high-performance data processing and integration.

**Environment**: Snowflake, Azure Synapse Analytics, Azure Data Factory, Azure Active Directory, Azure Databricks, DBT, Kafka, Unity Catalog, Talend, Snow SQL, Snowpipe, CDC, Snowflake Time Travel, T-SQL, Looker, Power BI, ADLS Gen2, Azure Devops, Azure Blob Storage, Azure Data Lake Storage Gen2, Azure ETL, Python, Java, Azure SQL Data Warehouse.

**Role: Big Data Developer | Jan 2017 – July 2019**

**Client: Publix, Lakeland, FL**

**Responsibilities:**

* Designed and deployed high-performance data pipelines using Flume and Sqoop, enabling efficient ingestion of customer behavioural data into HDFS with a 15% improvement in data ingestion rates. Employed advanced SQL techniques including CTEs, stored procedures, functions, indexes, and views to optimize data transformations and query performance.
* Architected real-time streaming solutions with Kafka and Spark Streaming, reducing processing latency by 20%. Developed custom data processing logic in Python and Spark for enhanced stream handling and built API integrations to connect these pipelines with downstream systems.
* Engineered and optimized data processing components including T-SQL queries, stored procedures, Python scripts, and DAX expressions to streamline ETL operations. Integrated these with Power BI for advanced analytics and developed API endpoints for seamless data exchange across platforms.
* Implemented robust SQL solutions leveraging CTEs, stored procedures, functions, indexes, and views to enhance data retrieval efficiency. Integrated these solutions with Azure ecosystem (Data Factory, Databricks, SQL Database, Active Directory, Blob Storage) to deliver scalable big data architectures.
* Optimized development workflows by implementing Maven within JIRA environments, accelerating project delivery by 20% through improved build management and dependency resolution while maintaining rigorous issue tracking.
* Leveraged Spark-Scala ecosystem (RDDs, DataFrames, Spark SQL) with Spark-Cassandra connectors and Azure services (Databricks, Synapse Analytics) for large-scale data migrations and enterprise reporting solutions.
* Streamlined data transfers from MySQL to HDFS using Sqoop and T-SQL optimizations, achieving 25% faster migrations while improving data quality by 20% through Hive and MapReduce transformations.
* Automated deployment pipelines for Hadoop environments using CI/CD principles, reducing deployment times by 30%. Implemented Jenkins, Git, Ansible with Azure integrations for cost-efficient infrastructure management.
* Developed scalable data processing frameworks for both batch (Spark, Hadoop MapReduce) and real-time (Apache Flink) workloads, enabling efficient transformation and analysis of large datasets.
* Built end-to-end data pipelines integrating Kafka, Spark, Hive, Airflow, and Azure Data Factory for comprehensive data lifecycle management, demonstrating expertise in modern data stack technologies.
* Engineered versatile ETL solutions using Apache Pig, PySpark, and Databricks, processing structured and semi-structured data through Pig scripts, XML/JSON parsing, and API integrations with secure Azure AD authentication.
* Led full SDLC implementations from requirements gathering to solution design, applying Agile methodologies to reduce development cycles by 20% while ensuring alignment with business specifications.
* Performed advanced analytics using Hive and MapReduce on Hadoop clusters, boosting analytical efficiency by 35% through optimized query patterns and cluster configurations.

**Environment**: Spark (Spark Streaming, Spark SQL, Spark-Scala, RDDs, Data Frames), Python, Power BI, Kafka, Sqoop, Apache Flume, Apache Cassandra Azure Services (Azure Data Factory, Azure Active Directory, Azure Databricks, Azure SQL Database, Azure DevOps, Azure Blob Storage), APIs, Jenkins, Git, Ansible, Maven, Apache Flink, Apache Pig, XML, Agile Methodology.

**Role: ETL Developer | Aug 2014 –Dec 2016**

**Client: CNA Insurance, Chicago, IL**

**Responsibilities:**

* Translated business requirements into technical ETL solutions using Informatica, SQL, PySpark, Shell Scripting, Bash, DB technologies, MDX, and PowerShell while strictly adhering to organizational standards and best practices.
* Designed and implemented comprehensive Informatica workflows including transformations, mappings, tasks, worklets, and workflows to orchestrate end-to-end data movement from source systems through staging to dimensional models, fact tables, and summary structures, enhanced by advanced T-SQL processing.
* Developed robust SSIS packages for seamless data integration across diverse platforms including Oracle DB, PostgreSQL, MS Access, Flat files, JSON, Excel, and SQL Server 2008 R2, while leveraging expertise in Python, SSRS, SSAS, SSIS, IICS, and Informatica for holistic data solutions.
* Automated code migration processes using Ansible templates for Informatica and Autosys components across development phases, while managing the complete BI stack (SSRS, SSAS, SSIS, IICS) for efficient data operations and reporting.
* Implemented rigorous healthcare data compliance measures including HIPAA and HL7 standards, ensuring patient data privacy and security throughout all healthcare-related projects.
* Led cross-functional ETL initiatives in collaboration with MOCS application teams, driving the full development lifecycle from analysis and design to implementation of scalable integration solutions for diverse stakeholder needs.
* Architected optimized database structures (DDL), streamlined data operations (DML), and enforced robust security protocols (DCL) through close partnership with business stakeholders.
* Accelerated ETL development by 30% by leveraging Matillion's pre-built components with Azure integration and advanced T-SQL scripting for rapid pipeline construction.
* Engineered SCD solutions using Informatica ETL and Apache Airflow to automate dimensional loading into Azure SQL Data Warehouse, enhancing historical tracking capabilities.
* Optimized database performance through query tuning and structural refinements, significantly improving processing efficiency.
* Processed large-scale datasets using Python's Pandas, NumPy and PySpark frameworks for high-performance analytics.
* Implemented comprehensive data quality frameworks including validation checks, error handling, and reconciliation processes to ensure data integrity throughout ETL pipelines.
* Mastered Informatica PowerCenter and Airflow to build complex Mappings, Mapplets, Sessions and Workflows, with performance optimization for high-volume data projects.
* Established data lineage documentation using DBT to improve transparency and governance across the entire data landscape.
* Developed sophisticated T-SQL components including complex queries, stored procedures and views to optimize data access and manipulation.
* Practiced Agile methodologies in onshore-offshore model, participating in daily scrums and sprint cycles while managing code in Visual SourceSafe and tracking projects via Trello.
* Troubleshooted and resolved data pipeline issues, minimizing downtime and ensuring continuous data availability.
* Partnered with business teams to transform requirements into technical solutions, delivering data-driven business value.
* Designed and deployed interactive reports using SSRS and Power BI, enabling data-driven decision making across the organization

**Environment**: Informatica, Unix Shell Scripting, Bash Scripts, SQL, DB Scripts, MDX scripts, PowerShell, SSIS, DBT, Python, SSRS, SSAS, IICS, HIPAA, HL7, Matillion, SQL Data Warehouse, DDL, DML, DCL, Agile Scrum, Visual SourceSafe, Trello.