**NITHIN KUMAR PAVURALA**

(Azure Data Engineer)



**Phone: 812-657-0607**

**Email:** pavuralanithin2@gmail.com
**LinkedIn:**  [www.linkedin.com/in/nithin-pavurala-0b4b9a153](http://www.linkedin.com/in/nithin-pavurala-0b4b9a153)

**PROFESSIONAL SUMMARY**

* Having 10+ years of experience in IT, I have spent approximately 7 years as an **Azure Data Engineer** **and Big Data Engineer**, complemented by 3 years of specialized proficiency as a **Data Warehouse Developer**. My skill set encompasses a broad spectrum, including Data Processing, System Analysis, Architecture, Design, Development, and Implementation.
* Demonstrated proficiency in utilizing Azure Cloud and its various components, including **Azure Data Factory**, **Azure Data Lake Gen2**, **Azure Data Explorer**, **Event Hub**, **Azure Synapse Analytics** and **Azure Databricks**.
* Extensively developed **data ingestion patterns** using **Azure Data Factory**, incorporating Integration runtimes, **Azure Key Vaults**, **Azure Logic Apps**, **Azure Storage Containers** like Blob, ADLS Gen1 and Gen 2, Function Apps, Azure VMs, and Azure DevOps Git Repos.
* Proficient in architecting and designing comprehensive data solutions utilizing Azure services including **Azure Event Hubs**, **Azure Stream Analytics**, and **Azure Synapse Analytics**, enabling real-time data processing and analytics support.
* Proficient in designing and implementing **event streaming pipelines**, leveraging **Azure Event Hubs** to seamlessly ingest and process high volumes of real-time data from diverse sources, enhancing operational efficiency and decision-making capabilities.
* Designed and implemented **data ingestion workflows**, **data transformations**, and **data enrichment processes** within **Azure Stream Analytics** to cleanse, enrich, and aggregate streaming data.
* Designed and implemented **data models, data partitioning strategies**, and **indexing strategies** within Azure Synapse Analytics to optimize query performance and resource utilization.
* Implemented **Azure Active Directory (AAD)** solutions to manage **user identities**, **access controls**, and **authentication mechanisms** across cloud and hybrid environments.
* Proficient in designing and implementing **ETL data pipelines** using **Spark/PySpark** on Databricks, with a deep understanding of optimizing data workflows for enhanced efficiency and scalability.
* Highly experienced in architecting and overseeing the creation of high-availability **clusters** and **cluster pools** on Databricks, ensuring optimal performance and reliability for data processing tasks.
* Extensive expertise in **utilizing Azure Databricks Unity Catalog**, data sharing and auto-loader.
* Successfully implemented **Unity Catalog** within the organization, significantly enhancing data integrity, access control, and compliance.
* Experienced in connecting **Unity Catalog** with established protocols, ensuring seamless data management and compliance with regulatory standards.
* Extensive experience in utilizing **Azure DevOps** for **version control**, automated testing, and release management, contributing to the successful delivery and maintenance of complex data engineering projects over a decade-long career.
* Proficient in leveraging **Azure DevOps** methodologies and tools to streamline **CI/CD pipelines**, ensuring efficient deployment and management of data solutions across the Azure ecosystem.
* Proficient in utilizing Azure data services such as **Azure SQL Data Warehouse**, **Azure Synapse Analytics**, **Azure Data Factory**, **Azure Databricks**, and **Azure Analysis Services** to design, implement, and manage enterprise **data warehousing solutions.**
* Designed **Azure EDW** for centralized data storage, retrieval, and analysis at scale.
* Skilled in utilizing **Spark Core** and **Spark SQL** scripts with Python to enhance data processing efficiency.
* Developed ETL/ELT processes to load data into **star schema** and **snowflake schema structures** from various source systems, ensuring data quality and consistency.
* Experienced data engineer proficient in designing and implementing data lakehouses, data lakes, and data warehouses within Azure cloud environments, with a strong foundation in **dimensional data modeling**.
* Good Hands-on Knowledge on **HDFS, MapReduce, Pig, YARN, Hive, Sqoop, Oozie, Zookeeper, Kafka, and Spark**.
* Developed Spark jobs using **PySpark/Scala** and **Spark SQL** to accelerate data processing.
* Exceptionally adept at executing **Hive scripts** via Hive on **Spark** and **Spark SQL** to meet diverse data processing requirements.
* Skilled on **Hive** partitioning, bucketing, and executing various types of joins on **Hive tables**.
* Showcased adeptness in **scripting languages** **Scala** and **Python** to streamline **data processing** tasks.
* Exceptionally proficient in transferring data between **RDBMS** and **Hadoop/Hive** environments using **SQOOP**
* Good knowledge on **SQL queries** and **PL/SQL** code used to enhance performance, detecting, and addressing any performance obstacles and inefficiencies.
* Exceptionally skilled at managing vast collections of structured, semi-structured, and unstructured datasets to bolster **Big Data initiatives**.
* Utilized Azure services such as **Azure SQL Database, Azure Data Factory, or Azure Event Hubs** to capture and track changes made to databases in real-time or near real-time.
* Configured and maintained **CDC pipelines** to efficiently stream and process data changes while ensuring data integrity and minimal latency.
* Monitored the performance and health of **CDC solutions in Azure** using built-in monitoring tools and techniques such as Azure Monitor.
* Proactively identifying and address any issues or bottlenecks in the **CDC process** to ensure continuous and reliable data capture.
* Collaborated with cross-functional teams to troubleshoot and resolve any technical challenges related to **CDC implementation** or operation.
* Significant expertise in leveraging **Kafka, Spark Streaming**, and **Hive** for processing **streaming data**, constructing resilient data pipelines for ingestion, transformation, and analysis purposes.
* Considerable background in managing **NoSQL databases** like **HBase, Cosmos DB, and MongoDB**.
* A cooperative team player committed to upholding **data integrity** and maintaining robust **data pipelines** while working collaboratively on **ETL assignments**.
* Proficient in implementing **Hive** SerDes (Serializer/Deserializer) such as **JSON, ORC, XML**, and **Parquet**.
* Good knowledge on **T-SQL queries** and **PolyBase** external table definitions to extract, transform, and load data from external sources into **SQL Server databases**.
* Developed **data pipelines** for BI developers using Microsoft **Power BI** on Azure for live dashboards
* Successfully deployed the **Master Data Management platform**, adeptly managing **data loading** and **extraction** to streamline **data governance** and **integration workflows**.
* Experienced professional with a solid understanding of **Agile practices**, particularly in Scrum methodology, and a keen interest in continuous improvement. Good knowledge on **SAFe** adds value to delivering scalable **Agile solutions.**
* Developed complex mappings, transformations, and workflows in **Informatica PowerCenter** to meet business requirements and data integration objectives.
* Validated a proof of concept (POC) on recently implemented technologies, Apache Airflow and GitLab.
* Collaborated with business **stakeholders** to gather and analyze requirements, and translate them into **technical specifications** and **ETL workflows**.
* Hands on experience on employing **JIRA** for **project reporting** and **task management**, alongside version control tools such as **SVN**, **Git**, and **Bitbucket**, ensuring effective project execution within Agile methodologies.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **Cloud Technologies:**  | Azure Data Factory, Azure Blob Storage, Azure Data Lake Storage (ADLS) Gen 2, Azure Data Bricks,Unity Catalog, Azure Synapse Analytics, Azure Stream Analytics, Azure Functions, Azure Event Hubs, Logic Apps, Azure DevOps, Azure Kubernetes Cluster, Cosmos DB, Snowflake, Azure SQL Database, Azure Active Directory |
| **Big Data Technologies:**  | Apache Hadoop, Apache Spark, Apache NiFi, HDFS, Hive, Sqoop, YARN, Pig, Oozie, MapReduce, Kafka, HBase, Splunk, IaaS, PaaS. |
| **Languages:**  | Python, SQL, PL/SQL, PySpark, Linux, Unix Shell Scripting  |
| **Modeling Techniques:**  | Data Modeling, Star Schema Modeling, Snowflake Schema Modeling, FACT and Dimensions Tables, Slowly Changing Dimensions (SCD), Change Data Capture (CDC).  |
| **Databases:**  | MS SQL Server, Oracle 11g/12c, Azure SQL DB, MongoDB, Teradata, My SQL, Oracle, PostgreSQL. |
| **Business intelligence:** | Power BI, Qlik Sense, Spotfire, SAP Business Objects 11. |
| **Scheduling Tools:**  | Oozie, Apache Airflow. |
| **Version Control and CI/CD Platform:** | Git, GitHub, GitLab CI/CD, Maven, Jenkins, Docker, Kubernetes (Desktop client) |

**MICROSOFT CERTIFICATIONS**

* Azure Fundamentals (DP-900)
* Azure Data Engineer Associate (DP-203)

**EDUCATION BACKGROUND**

* Bachelors in Electrical and Electronics Engineering, Kakatiya University, Aug 2007 – May 2011
* Masters in Computer science, Chicago state university, Jan 2012 - Dec 2013.

**PROFESSIONAL EXPERIENCE**

**Client: Fidelity Investments, Dallas, Texas. Feb 2022 to Present**

**Role: Azure Data Engineer**

**Responsibilities:**

* Designed and developed medium to large-scale BI solutions on Azure utilizing Azure Data Platform services including Azure SQL Data Warehouse, Azure Data Lake, and Azure Data Factory
* Collaborated closely with business analysts and data engineers to understand the technology landscape and deliver integrated end-to-end solutions, including building a centralized Data Lake in Azure.
* Connected to various source systems like Oracle, SQL Server, and Teradata using Self-hosted Integration Runtime.
* Automated workflows in Azure Databricks, reducing financial overhead during real-time and batch processing for transactions.
* Analyzed Hive SQL scripts and implemented solutions using PySpark on Databricks for advanced analytics capabilities.
* Maximized partitioning efficiency with Azure Cosmos DB and utilized Logic Apps to trigger Azure Notification Hubs for transaction notifications.
* Developed Spark applications on Azure Databricks for data extraction, transformation, and aggregation from multiple file formats.
* Implement and optimize Azure Data Warehouse solutions to efficiently store and process large volumes of structured and unstructured data, ensuring scalability, performance, and reliability for analytical workloads.
* Lead the implementation of scalable and resilient event streaming pipelines using Azure Event Hubs to ingest and process high volumes of data from various sources in real-time.
* Designed and implemented data ingestion workflows, data transformations, and data enrichment processes within Azure Stream Analytics to cleanse, enrich, and aggregate streaming data.
* Partnered with business analysts and data engineers to understand the technology landscape and produce integrated end-to-end solution that manifest building of 3-5 TB data in a centralized Data Lake in Azure Data Lake (ADLS) Gen 2.
* Lead the design and administration of enterprise-class Windows Server systems, ensuring optimal performance and security. Proactively monitor system environments, troubleshoot issues, and provide comprehensive user support, including system application support and OS support.
* Overseen the implementation and ongoing management of Azure and Azure Active Directory domain infrastructure, ensuring seamless integration with on-premises environments and adherence to best practices for security and scalability
* Driven the design, implementation, and administration of DevOps solutions within cloud environments, utilizing tools and methodologies to automate deployment, monitoring, and management processes, thereby improving agility and efficiency
* Ensured the compliance with data governance frameworks and regulations such as HIPAA, implementing robust security measures and privacy controls to safeguard sensitive data and maintain regulatory compliance. Stay updated on evolving compliance requirements and proactively address any potential risks or vulnerabilities.
* Designed and implemented data models, data partitioning strategies, and indexing strategies within Azure Synapse Analytics to optimize query performance and resource utilization.
* Implemented Role-Based Access Control using Azure Key Vault integrated with Microsoft Entra ID for secure access to pipelines.
* Designed and implemented scalable and cost-effective data storage solutions leveraging IaaS offerings such as Azure Virtual Machines (VMs).
* Established Delta Lake in Databricks and ingested data into Azure Synapse Analytics data warehouse for live visualization dashboards using Power BI.
* Developed and maintained scalable data pipelines to ingest, transform, and deliver structured and unstructured data using PySpark, SQL, and cloud data warehouse technologies such as Azure Synapse and Databricks.
* Utilized extensive knowledge of data warehousing concepts, strategies, and methodologies to design and implement robust data solutions that meet business requirements.
* Collaborated with cross-functional teams to understand data requirements and optimize ETL/ELT processes for batch and real-time data integration using Azure Databricks and Data Factory.
* Leveraged relational data processing technologies such as MS SQL, Delta Lake, Spark SQL, and SQL Server for efficient data processing and analysis.
* Optimized fine-tune PaaS resources for performance, cost, and efficiency. Implement monitoring, logging, and alerting mechanisms to track data pipeline performance and troubleshoot issues proactively. Continuously evaluate and leverage new PaaS offerings and features to enhance data engineering capabilities and drive innovation.
* Deployed containerized data processing applications and services using Kubernetes orchestration platform.
* Implemented unified data governance solutions using Microsoft Purview Data Catalog for glossary management and metadata capture.
* Monitored data pipelines and resolved production issues quickly while enhancing preventive measures.
* Worked with Azure Functions for event-based and time-based scheduling to reduce manual efforts and minimize errors.
* Utilized strong knowledge of Integration Platform as a Service (IPAAS) offerings, such as Azure Logic Apps or Azure Integration Services, to design and deploy robust data integration solutions.
* Experienced in leveraging Azure Data Explorer for real-time analytics, enabling rapid querying and visualization of streaming data for actionable insights and decision-making.
* Developed Logic Apps and functional Apps for additional functionality like running Power BI Extracts and extracting data from SharePoint.
* Effectively leveraged Unity Catalog, a robust governance platform, to uphold data integrity, manage access, ensure security, and maintain compliance.
* Implemented fine-grained access control mechanisms within Unity Catalog to regulate data access across various organizational levels, safeguarding sensitive information.
* Expertise in Designing, developing, and maintaining PostgreSQL databases to meet the organization's data storage and processing requirements.
* Developed data pipelines for BI developers using Microsoft Power BI on Azure for live dashboards.
* Designed and developed data pipelines for provisioning and orchestration of various Data Science implementations, focusing on data quality and transaction monitoring.
* Managed source code using GIT in Azure Git and GitHub repositories for Azure ADF pipelines and Azure Databricks, participating in weekly scrum meetings.
* Took part in Agile Scrum sessions, comprising daily stand-ups and globally synchronized PI Planning sessions, to ensure efficient project management and execution.

**Environment**: Azure Data Factory, Azure Data Lake Gen2, Azure Data Explorer, Azure Event hubs, Azure Databricks, azure data warehouse, Key Vault, Azure Cosmos DB, SQL Server, Logic Apps, Azure Notification Hubs, Azure Monitoring, Oracle, Teradata, Azure Active Directory, Workload Automation Scheduler, GIT, GITHUB, Power BI, Agile.

**Client: Walmart Stores Inc, Bentonville, Arkansas. Nov 2019 to Jan 2022**

**Role: Azure Data Engineer**

**Responsibilities:**

* Developed pipelines in Azure Data Factory (ADF) to connect with various source systems including Oracle, SQL Server, Teradata, and APIs, enabling the extraction, transformation, and loading of data, established connections with third-party vendors to extract meaningful information from JSON payloads.
* Conducted data migration from on-premises servers to the cloud utilizing command line tools such as AZ Copy, Az PowerShell, and Azure CLI with Entra ID token, resulting in a tenfold acceleration in data transfer speed
* Teamed up with key stakeholders across different regions to facilitate the migration of data from an on-premises data warehouse to Microsoft Azure Cloud, enhancing operational efficiency through the development of data engineering pipelines.
* Experienced in utilizing Azure Data Lake Gen2 for efficient and scalable data storage solutions, ensuring seamless integration with many services and applications.
* Enhanced streaming pipelines by parsing and flattening diverse semi-structured data formats like JSON and XML, implementing incremental loading strategies such as "COPY INTO" for smaller datasets and Autoloader for large data files in shared locations, resulting in significant cost savings.
* Orchestrated a seamless migration from on-premises data warehouse to Azure Synapse Analytics, achieving infrastructure cost reduction, performance optimization, and substantial annual savings
* Established Delta Lake in Databricks and ingested data into Azure Synapse Analytics data warehouse for live visualization dashboards using Power BI.
* Worked closely with business analysts, project managers, and end-users to understand business requirements and translate them into UAT test scenario and Leverage Azure DevOps or similar collaboration tools to document and organize UAT test cases, ensuring alignment with business objectives and user expectations also Facilitate workshops or meetings to gather feedback and refine UAT scenarios based on stakeholder input.
* configured Azure environments specifically tailored for UAT testing purpose, Utilized Azure services such as Azure Virtual Machines, Azure Kubernetes Service, or Azure App Service to provision infrastructure components required for executing UAT test cases effectively, Implemented automation scripts or Infrastructure as Code (IaC) techniques using Azure Resource Manager (ARM) templates or Terraform to streamline environment setup and teardown processes, minimizing manual effort and ensuring consistency across UAT environments
* Integrated Azure Logic Apps and Functional Apps into data pipelines to orchestrate complex workflows, automate data processing tasks, and enable seamless integration with external systems and services, ensuring scalability, reliability, and maintainability of data solutions.
* Experienced in designing and implementing scalable event streaming pipelines using Azure Event Hubs for real-time data ingestion and processing.
* Proficient in optimizing and monitoring event hub configurations to ensure high throughput and reliability for streaming data applications.
* Experienced in utilizing PostgreSQL for robust data storage and efficient querying, contributing to streamlined data management and analysis processes.
* Proficient in designing and optimizing PostgreSQL databases to ensure scalability, reliability, and performance for data-driven applications.
* Highly skilled data engineer proficient in Teradata, adept at optimizing data warehousing solutions for efficient querying and analysis. Expertise includes designing and implementing robust ETL processes leveraging Teradata's powerful capabilities to drive actionable insights.
* Experienced data engineer proficient in Data Build Tool, adept at automating data pipeline development and deployment. Skilled in leveraging Data Build Tool to streamline ETL workflows, enhance data quality, and improve overall efficiency of data processes.

**Environment**: Azure Data Factory, Azure Data Lake Gen2, Azure Kubernetes Cluster, Azure Databricks, Azure Synapse Analytics, Azure Event Hubs, Azure SQL Database, PostgreSQL, Azure ML, Oracle, Teradata, Data Build Tool, Power BI, Agile,

**Client: Us Patent and Trademark Office,** **Alexandria, Virginia. July 2017 to Oct 2019**

**Role: Big Data Engineer**

**Responsibilities:**

* Developed startup and shutdown scripts in Python to manage Hadoop name-node, data-node, Spark worker and master servers.
* Designed and implemented iterative algorithms using Spark Streaming in Scala to enable near real-time data processing.
* Implemented Zookeeper for concurrent access to hive tables with shared and exclusive locking within the cluster.
* Lead the design, development, and maintenance of Apache NiFi data pipelines, ensuring seamless integration of varied data sources such as Mainframes, HBase, and MongoDB.
* Implemented full loads, incremental extracts, and end-to-end automation using Scala Backend, facilitating data extracts for Power BI and Workday Prism Analytics.
* Implemented performance optimization techniques including distributed cache usage, partitioning, and bucketing in Hive, resulting in a 20% reduction in processing costs.
* Utilized Sqoop to import and export dat­­­­a from SQL, DB2, and Teradata into HDFS for subsequent analysis.
* Designed and implemented data processing workflows on YARN, ensuring efficient resource utilization and optimal performance
* Executed data ingestion from various internal clients using Apache Kafka and developed Python-based publisher APIs for efficient data transfer.
* Configured Jenkins jobs to automate the building, testing, and deployment of data pipelines, ensuring reliability and consistency across environments.
* Deployed Slowly Changing Dimensions Type 2 to track account history and created shared containers and job sub-routines to log job start and end times in Audit tables.
* Utilized Avro and Parquet data formats for Apache Hive computations to meet custom business requirements.
* Enhanced performance by translating Hive/SQL queries into Spark transformations using Spark RDDs, Datagrams, and Python.
* Proficiency in deploying, configuring, and managing Kubernetes clusters, including experience with container runtime environments like Docker
* Ability to optimize cluster resources, ensure high availability, and implement best practices for security and governance.
* Conducted performance tuning of DataStage jobs with appropriate partitioning techniques, node configuration, and Teradata scripts.
* Developed custom aggregate functions using Spark SQL, created tables as per the data model, and executed interactive querying.
* Constructed ELT/ETL pipelines to facilitate the seamless extraction, transformation, and loading of data within Adobe's infrastructure.
* Extracted meaningful data from dealer CSV files, text files, and modernized legacy application mainframe files, generating Python pandas reports for data analysis.
* Utilized MapReduce for distributed data processing, enabling efficient analysis and extraction of insights from large-scale datasets.

**Environment**: Apache NiFi, HDFS, Kafka, Mainframes, Hadoop, YARN, Python, HBase, Hive, Jenkins, Sqoop, Pig, MapReduce, Scala, Zookeeper, GitHub, Spark RDDs, Linux, Spark, Unix, Teradata scripts.

**Client: Regions Bank, Hoover, Alabama. Jan 2014 to June 2017**

**Role: Data Warehouse Developer**

**Responsibilities**:

* Designed and implemented data models using dimensional modeling techniques such as star schema and snowflake schema to support the organization's data warehouse requirements.
* Developed and maintained ETL processes to extract data from various sources, transform it to fit the data warehouse schema, and load it into the warehouse using tools like Informatica.
* Integrated data from disparate sources including databases, flat files, APIs, and streaming sources into the data warehouse, ensuring data quality, consistency, and integrity
* Implemented data quality checks and validation processes to ensure the accuracy, completeness, and consistency of data stored in the warehouse.
* Hands-on experience with deploying and managing containerized applications on Kubernetes, including defining Kubernetes manifests (YAML files), managing application lifecycle with deployments, stateful sets, and daemon sets. Familiarity with Helm charts for packaging and deploying complex applications on Kubernetes.
* Developed and maintained SSRS reports and dashboards to provide insights into key business metrics and KPIs, supporting decision-making and analysis.
* Implemented efficient data integration workflows to transform and load raw data into the data warehouse using SSIS, ensuring data quality and consistency.
* Managed metadata and documentation related to data warehouse objects, including tables, columns, transformations, and data lineage using tools like Erwin or custom documentation frameworks.
* Monitored the health and performance of the data warehouse environment, diagnosed, and resolved issues, and performed routine maintenance tasks such as backups, upgrades, and patches.
* Collaborated with stakeholders such as data analysts, business users, and IT teams to understand requirements and deliver solutions. Documented ETL processes, data flows, and system configurations using tools like Confluence or SharePoint.
* Identified opportunities for process optimization, innovation, and automation to enhance the efficiency and effectiveness of data warehouse operations. Stayed abreast of industry best practices, emerging technologies, and trends in data warehousing and analytics

**Environment**: Informatica, SSRS, SSIS, Oracle 10g, Unix, Linux, visual Studio 2010, Performance Point Server, MS Office, SharePoint.