Jaya Deva
Data Engineer
(469) 956-9755 **jayadeva5391@gmail.com**

PROFESSIONAL SUMMARY

"Accomplished **Azure Data Engineer** with 5+ of experience in The IT field, including 5+ years specializing in Data Engineering and 4 years in Data Warehousing. Proficient in **Python** and**SQL**, with expertise in **Data Factory**, **Synapse Analytics**, and **Databricks** for batch and real-time data processing. Skilled in data streaming platforms like **Kafka** and **Azure Event Hubs**, alongside **Spark notebooks**, **Azure Data Lake Storage**, **BLOB Storage**, and relational databases. Known for optimizing **ETL** processes, utilizing **Power BI** for visualization, and leading successful migration projects to **Azure** and **Snowflake**. Transforming Complex Data Challenges into Business Insights.

- 5+ Years of experience in **Data Engineering**, with expertise in designing and implementing scalable data ingestion pipelines using **Azure Data Factory**, including advanced features like **copy activity, lookup activity**, and integrating **ADLS Gen2** for efficient data storage and management.
- Orchestrated the seamless transition of on-premises data to the cloud using **Azure Data Factory**. Designed and executed robust data pipelines, ensuring efficient transfer to **Azure Data Lake Storage Gen2 (ADLS Gen2)**. Contributed to enhanced scalability, reliability, and cost-effectiveness for the organization.
- Accomplished data professional with a strong background in **ETL pipeline management**. I specialize in ensuring scalability, smooth operations, and high-speed data transfers by leveraging compression techniques for storage optimization.
- Proficient in fine-tuning query techniques and implementing effective indexing strategies to enhance data fetching efficiency. Demonstrated expertise with Vertica and Teradata for high-performance data analytics.
- Proficient in crafting SQL queries, including Data Definition Language **(DDL)** and Data Manipulation Language **(DML)**, to manipulate and retrieve data. Expertise extends to seamlessly connecting disparate data sources using linked services in **Azure Data Factory**.
- Proficient in advanced **data warehousing** methodologies, I excel in executing data cleansing, managing Slowly Changing Dimensions (**SCD**), assigning surrogate keys, and implementing change data capture (**CDC**) within **Snowflake** environments. My adept use of lookup activities ensures the utmost accuracy and integrity of data.
- Expert in the creation and deployment of expandable data ingestion pathways, leveraging **Apache Kafka**, **Apache Flume**, and **Apache Nifi**, with a strategic focus on processing via EventHub.
- Skilled in implementing data quality checks and cleansing techniques to ensure data accuracy and integrity throughout the pipeline, applying compression techniques to optimize data storage and processing.
- Experienced in building and optimizing data models and schemas using technologies like **Apache Hive**, **Apache HBase**, or **Snowflake** for efficient data storage and retrieval for analytics and reporting, integrating **ADLS Gen2** for enhanced data management.
- I specialize in using **Azure Databricks** and **PySpark** API to build powerful data processing systems that efficiently handle both batch and real-time data, ensuring fast and reliable insights.
- Configured the **ADF** jobs, and **Snow SQL** jobs triggering in Matillion using **Python**.
- Extensive experience with **Hadoop**, **HDFS**, **Map-Reduce**, **Hive**, **Python**, and **PySpark**.
- Experienced in optimizing code for **Azure Functions** to extract, transform, and load data from diverse sources, incorporating **EventHub** for data streaming capabilities.
- Good experience in **Hortonworks** and **Cloudera** for Apache **Hadoop distributions**, leveraging Event Queues for efficient data processing and management.
- Designed and created **Hive** external tables using shared meta store with Static & Dynamic partitioning, bucketing, and indexing, enhancing query performance in Vertica and Teradata environments.
- Exploring with **Spark** to improve the performance and optimization of existing algorithms in **Hadoop** using **Spark** context, **Spark SQL**, Data Frame, pair **RDDs**, and incorporating compression techniques for data efficiency.
- Proficient in implementing **CI/CD** frameworks for data pipelines using tools like Jenkins, ensuring efficient automation and deployment, with linked services facilitating continuous integration and delivery in cloud and hybrid environments.
- Proficient in implementing **Agile methodologies** to boost project delivery efficiency, emphasizing cross-functional collaboration.

TECHNICAL SKILLS:

Azure Services	Azure Data Factory, Azure Data Lake, ADLS Gen2, Azure Data ricks, Logic Apps, Functional App, Key Vault, Azure Active Directory, Azure Synapse Analytics,
Big Data Technologies	HDFS, MapReduce, Hive, Sqoop, Oozie, Zookeeper, Kafka, Apache Spark, Spark Streaming,
Databases & Data warehouses	MS SQL Server 2016/2014/2012, Azure SQL DB, Snowflake, Azure Synapse. MS Excel, MS Access, Oracle 11g/12c, Cosmos DB, Cassandra, PostgreSQL, Teradata, MongoDB, Dynamo DB.
Hadoop Distribution	Cloudera, Hortonworks
IDE &Build Tools, Design	Eclipse, Visual Studio, PyCharm, DBT.
Operating Systems	Windows (XP/7/8/10), UNIX, LINUX, UBUNTU, CENTOS.
Programming Languages	Python, PySpark, Shell script, .NET/C#, Perl script, SQL, Java.
Version Control	GIT, GitHub, Azure GitHub.
Visualization Tools	Power BI, Tableau, SSRS
Web Technologies	XML, JSP, HTML, SOAP, JavaScript.

PROFESSIONAL EXPERIENCE:

Wells Fargo, Dallas, TX **Jan 2022 – Present**

Azure Data Engineer

Responsibilities:

- Orchestrated a large-scale **data migration** from on-premises systems to the Azure cloud platform, utilizing **Snowflake** for efficient data warehousing and storage solutions.
- Designed and implemented a **self-hosted integration runtime** within **Azure Data Factory** (ADF) to facilitate a seamless data transfer pipeline from on-premises like SQL databases, Oracle databases, CSV files, and REST APIs into Azure Blob Storage.
- configured network security measures, including **VPN** and **firewall settings**, to enable direct data access from on-premises SQL Servers to Azure Databricks via **JDBC connectors**.
- Managed data ingestion, movement, and orchestration using **ADF** pipelines, leveraging **Azure Logic Apps** for process automation and **Azure Monitor** for operational insights and analytics.
- Managed structured and unstructured data using a variety of databases, including Azure **Cosmos DB** for NoSQL data and **Azure SQL Database** for relational data.
- Developed data processing workflows using Azure Databricks, leveraging **Spark** for distributed data processing and transformation tasks.
- Implemented data quality checks and data cleansing techniques to ensure the accuracy and integrity of the data throughout the pipeline, using **Azure Data Factory** and **Databricks.**
- Developed end-to-end ETL data pipelines, ensuring scalability and smooth functioning. This included extensive use of copy activity for data movement and lookup activity for data validation, leveraging linked services to connect on-premises and cloud data sources.
- Implemented optimized query techniques and indexing strategies, enhancing data fetching efficiency and scalability using **SQL** queries and **ADLS Gen2**.

- Integrated **Snowflake** with Azure cloud services to establish secure and efficient data warehousing solutions, enabling insightful reports for strategic analysis.
- Hands-on development experience with **Snowflake** features such as **Snow SQL**; **Snow Pipe**; **Python**; Tasks; Streams; Time travel; Zero Copy Cloning; Optimizer; Metadata Manager; data sharing; and stored procedures.
- Designed and implemented real-time data processing solutions using **Kafka** and **Spark Streaming**, facilitating the ingestion, transformation, and analysis of high-volume streaming data.
- Integrated **PySpark** with **Azure Data Bricks** and **Azure Blob Storage** for seamless data ingestion and processing within the **Azure ecosystem**.
- Optimized **PySpark** jobs for performance by leveraging techniques like partitioning and caching, reducing processing times and improving system efficiency.
- Conducted **performance tuning** and capacity planning exercises to ensure scalability and efficiency of data infrastructure.
- Made strategic use of **ADLS Gen2** for efficient data storage and management in **Azure Functions**, optimizing code for data extraction, transformation, and loading.
- Developed complex **SQL** queries and data models in **Azure Synapse Analytics** to integrate big data processing and analytics capabilities, enabling seamless data exploration and insights generation.
- Built and optimized data models and schemas using technologies like **Apache Hive** and **Snowflake**, with copy activity streamlining data movements.
- Created ETL transformations and validations using **Spark SQL/Spark Data** Frames with **Azure Databricks** and **Azure Data Factory**, ensuring data accuracy and consistency.
- Integrated **GitHub** repositories with **Azure services** for enhanced collaboration and automated deployment workflows within the **Azure ecosystem**.
- Designed and deployed interactive **Power BI** dashboards providing real-time insights into various business metrics, enhancing decision-making processes.
- Collaborated with **Azure DevOps** team to improve code quality and project management efficiency through **CI/CD** pipelines.
- Actively collaborated with data analysts and business stakeholders to understand reporting requirements, facilitating the development of customized reports driving strategic business decisions.

Environment: Snowflake, Azure Databricks, Azure Data Factory, Azure Logic Apps, Oracle, Functional App, Key Vault, MySQL, Azure SQL Database, HDFS, Spark, Hive, SQL, Python, Scala, PySpark, GIT, JIRA, Jenkins, Kafka, Azure ML, Power BI, HBase, Azure DevOps.

Molina Healthcare **Jun 2019- Dec 2021**

Azure Data Engineer

Responsibilities:

- Spearheaded data engineering projects within the **Azure Kubernetes** Service (**AKS**) environment at Moss & Associates, emphasizing reliability, scalability, and efficiency in data operations.
- Architected and executed end-to-end data pipelines, seamlessly integrating Azure services like **SQL Database**, **Data Lake Storage**, and **Data Factory**.
- Implemented streamlined data integration solutions for seamless ingestion and integration of data from various sources, employing tools such as **Apache Kafka**, **Apache NiFi**, and **Azure Data Factory**, with Event Hubs for real-time data streaming.
- Managed data ingestion into Azure Services (**Azure Data Lake, Azure Storage, Azure SQL, Azure DW**) and orchestrated data processing in **Azure Databricks**, incorporating Event Hubs for real-time analytics.
- Boosted Spark performance through advanced optimization techniques like partitioning, caching, and compression, enhancing data processing efficiency.
- Leveraged Microsoft Azure services such as **HDInsight Clusters**, **BLOB**, **Data Factory**, and **Logic Apps**, alongside **Event Hubs** for streamlined data collection and processing.
- Executed **ETL** operations using **Azure Databricks**, migrating on-premises **Oracle ETL** processes to **Azure Synapse Analytics** with optimized storage and querying speeds.

- Utilized **Snowflake's** versatile support for **SQL**, **Python**, and other languages for advanced analytics and data processing tasks within **AKS**.
- Oversaw migration of **SQL databases** to **Azure data lake**, **Azure SQL Database, Azure Synapse**, and integration of **Teradata** databases for seamless data synchronization and reporting.
- Managed database access and migration to **Azure data lake** store using **Azure Data Factory**, including **Teradata** databases via efficient data transfer methods.
- Implemented data transfer using **Azure Synapse** and **Polybase**, focusing on compression techniques to enhance efficiency and reduce storage costs.
- Leveraged **Snowflake's** auto-scaling features to ensure optimal performance and resource utilization.
- Deployed and optimized Python web applications to **Azure DevOps CI/CD** pipeline, integrating data from Event Queues for effective application event management.
- Developed enterprise-level solutions using batch processing and streaming frameworks like **Spark** Streaming and **Apache Kafka**, with **Event Queues** for efficient event-driven **data workflows**.
- Designed and implemented robust data models and schemas using technologies such as **Apache Hive**, **Apache Parquet**, and **Snowflake**, applying compression techniques for optimized storage.
- Managed end-to-end data pipelines using **Apache Spark**, **Apache Airflow**, or **Azure Data Factory**, ensuring reliable and timely data processing and delivery, including integration with Teradata for comprehensive data analysis.
- Collaborated with cross-functional teams to gather requirements, design data integration workflows, and implement scalable data solutions, leveraging Event Hubs for real-time event stream processing.
- Provided production support and troubleshooting for **data pipelines**, identifying and resolving performance **bottlenecks**, **data quality issues**, and system failures, with a focus on optimizing **data flows** from **Event Queues**.
- Actively participated in **Agile** ceremonies such as **Sprint Planning**, Daily Stand-ups, **Sprint Reviews**, and Retrospectives to ensure project progress and team alignment.

Environment: Hadoop Cloudera, Microsoft Azure (including Azure Databricks, ETL, and Azure Synapse Analytics), SQL databases, Data Lake Analytics, Databricks, Polybase, Python, Azure DevOps, CI/CD, Kafka, Spark, Hive, Scala, Spark SQL, Hive tables, Hive Generic UDFs, Data Lakes, Hortonworks, PySpark, RDDs, Data Frames, Spark SQL, Git, and JIRA, Data flow, Azure SQL Server.

IQVIA, Durham, NC **Nov 2018 – Jun 2019**

Big Data Engineer

Responsibilities:

- Imported medical claims data using Sqoop to transfer data from the healthcare database to the **Hadoop Distributed File System** (**HDFS**) regularly.
- Applied aggregations on extensive healthcare claims data using **Spark** and **Scala**, storing the processed data in the **Hive** warehouse for further analysis.
- Worked collaboratively within the healthcare data ecosystem, utilizing technologies such as **Hadoop**, **Spark**, and **Cloudera** for efficient data management.
- Loaded and transformed large sets of structured, semi-structured, and unstructured medical claims data.
- Constructed **HBASE** tables by integrating **HBASE** with **HIVE** in the Analytics Zone to enhance data accessibility.
- Implemented **Kafka** and **Spark** Streaming to process streaming medical claims data in specific scenarios.
- Developed a data pipeline using Flume and Sqoop to ingest customer behavioral data histories into **HDFS** for analysis.
- Implemented **DBT** to automate data transformation pipelines, resulting in a 30% reduction in processing time."
- Analysed the **Hadoop** cluster using various big data analytic tools, including **Hive** and **MapReduce**.
- Created a data pipeline using **Kafka**, **Spark**, and **Hive** to ingest, transform, and analyze medical claims data.
- Experienced in integrating streaming solutions with other **Hadoop ecosystem** components such as **HDFS**, **YARN**, and **Hive** for seamless data ingestion, storage, and analytics across batch and streaming data sources.
- Crafted Hive queries for data analysis to meet specific business requirements, creating **Hive** tables and leveraging **Hive QL** to simulate **MapReduce** functionalities.
- Implemented **UNIX scripts** to define the use case workflow and process data files, automating key jobs.
- Automated deployments using **YAML** scripts for significant builds and releases in the medical claims processing environment.
- oversaw the effective deployment and refinement of **Apache Airflow-**based data processing workflows at FRG Solutions, improving the effectiveness of **big data pipelines**.

- Migrated existing medical claims data to Hadoop from relational databases (**RDBMS**) such as **Oracle** using **Sqoop**.
- Established **Continuous Integration/Continuous Deployment (CICD)** pipelines to build and deploy medical claims processing projects in the Hadoop environment.
- Utilized **JIRA** for issue tracking and project workflow management.
- Leveraged **PySpark** and **Spark SQL** with **Python** to expedite data testing and processing within **Spark**, enhancing efficiency and performance.
- Used Spark Streaming to segment medical claims data into batches as input to the Spark engine for batch processing.
- Utilized **Zookeeper** to coordinate, synchronize, and serialize servers within the clusters.
- Implemented **Oozie** workflow engine for scheduling medical claims processing jobs.
- Employed **Git** as the version control tool to maintain the code repository.
- Automated the processes of data loading, transformation, and ingestion by implementing Apache **Airflow** as the key orchestration tool for FRG Solutions. created specialized Airflow operators and guided the building of **DAGs** to satisfy exact business requirements, saving a significant amount of time and improving data accuracy.
- Successfully implemented **Agile** practices to manage changes effectively and ensure project delivery within time and budget constraints.

Environment: Hadoop, Spark, Scala, Hive, HBase, Kafka, Flume, Sqoop, Zookeeper, Oozie, Git, JIRA, Spark, YAML, MapReduce, Python, PySpark, RDBMS, Shell Script, JIRA, Airflow, CI/CD.

Education: Masters in Wilmington university.