**SREENATH REDDY VEMIREDDY**

Senior Big Data Engineer | Cloud Data Engineer

**Phone:** [+1 (469) 793-3883](tel:+14697933883)

**Email:** [vemisreenathreddy@gmail.com](mailto:vemisreenathreddy@gmail.com)**LinkedIn:** <https://www.linkedin.com/in/sreenath-09bb91183>**Portfolio:** [vemireddysreenath.github.io](https://vemireddysreenath.github.io/)

**Professional Summary:**

Experienced **Senior Big Data** and **Azure Data Engineer** with over **10 years** of experience in building, automating, and optimizing large-scale data engineering solutions. Adept at leveraging open-source Big Data technologies and cloud-native platforms like **Azure** and **AWS** to solve complex business challenges across finance, insurance, and healthcare domains. Strong experience of managing end-to-end data migrations, regulatory compliance pipelines, analytics platforms from scratch, cloud integration and regulatory reporting**.**

* **Extensive Data Pipeline Development:**  
  Built and maintained scalable, high-throughput data pipelines using Hadoop, Spark, PySpark, and Hive to support critical reporting and analytics use cases. Designed data models to handle structured and semi-structured data at scale.

Extensive hands-on experience with Cloudera and Hortonworks distributions of Hadoop for enterprise-grade data processing, deployment, and cluster management.

* **Advanced Data Lake & Cloud Integration:**  
  Successfully migrated terabytes of data from on-prem and hybrid sources (e.g., Oracle, Vertica, S3) into Azure Data Lake Storage and SQL databases, enhancing data availability and performance.Designed, developed, and deployed scalable cloud-native data platforms for regulatory, risk, and compliance workloads in financial services using Azure Data Factory, Databricks, and SQL Server.
* **Metadata-Driven Automation & Frameworks:**  
  Designed and implemented custom metadata-driven frameworks for schema validation, lineage tracking, alerting, and audit logging—reducing manual effort and operational overhead
* **Regulatory Reporting Expertise:**  
  Delivered critical data solutions for MAS 637 and PILLAR3 reporting using PySpark, ADA Framework, and Airflow, ensuring strict compliance with regulatory bodies like the Monetary Authority of Singapore.
* **ETL Workflow Orchestration:**  
  Skilled in designing complex DAGs using Apache Airflow and Zena for ETL orchestration. Automated failure alerts, error handling, retries, and monitoring across the data lifecycle.
* **Cloud Security and Governance:**  
  Integrated with Azure Key Vault for secure access to credentials and secrets. Defined access controls and lineage using Collibra, ensuring data security and governance.
* **Cross-functional Collaboration:**  
  Worked closely with compliance teams, analysts, QA engineers, and business stakeholders to deliver reliable and actionable data products. Led cross-functional teams and mentored junior engineers on DevOps practices, Databricks clusters, API integrations, and Unity Catalog for cloud-native data platforms.  
  Contributed to key projects with clients such as DBS Bank, Country Financial, PayPal, ICICI Bank, AbbVie, Apple, and Nokia.
* **Performance Optimization & Troubleshooting:**  
  Continuously monitored and improved and optimized Spark jobs, SQL queries, and ADF pipelines for better performance. Diagnosed bottlenecks and applied best practices for code and system-level improvements.
* **Agile & DevOps Methodologies:**  
  Familiar with Agile delivery, version control using Git, CI/CD pipelines using Azure DevOps, and automated testing in Big Data environments.
* **Mentoring and Leadership:**  
  Mentored junior developers and data engineers by providing architectural guidance, code reviews, and training. Led small teams in project planning, task allocation, and delivery assurance.
* **Domain Versatility:**  
  Applied big data solutions across multiple industries:
  + **Finance:** Risk analysis, customer profiling, compliance reporting
  + **Healthcare:** Patient data unification and insights
  + **Insurance:** Policy and claims data transformation
  + **Banking:** Regulatory audit pipelines and trend analysis
  + **Geospatial:** Map data ingestion and visualization pipeline support

### **Technical Skills**

* **Big Data Technologies:** Hadoop, Apache Spark, PySpark, Hive, HDFS, Impala, Pig, Sqoop, Cloudera, Hortonworks, TEZ,
* **Cloud Platforms:** Microsoft Azure (Azure Data Factory, Azure Blob Storage, Azure SQL Database, Azure Key Vault), AWS, S3, Databricks, Data Lake, Delta Lake, CasmosDB, Snowflake, Unity Catalog.
* **Data Analytics & BI Tools:** Adobe Analytics, Power BI, Apache Superset, SAS, Collibra, Click Stream Data.

**Programming & Scripting Languages:** Python, Scala, Java, Unix/Shell Script, SQL/PSQL

* **Orchestration & Workflow Tools:** Apache Airflow, Zena, Autosys (JIL scripting)
* **Development Tools & IDEs:** Git, IntelliJ, Eclipse, PyCharm, Jupyter Notebook, Azure DevOps
* **Databases:** MySQL, Oracle, SQL Server, Vertica, SAP HANA, MongoDB
* **Frameworks (Internal / Custom):** ADA Framework, Metadata-Driven Python Frameworks
* **Other Skills:** Data Modeling, ETL Design, Data Migration, Regulatory Reporting, Metadata Management, Data Governance, Performance Optimization, Agile Delivery, Data Engineer, Data Platform Engineer, ETL Developer, Cloud Data Engineer.

**Professional Experience**

**Cloud Data Engineer | Client: Country Financial, USA(Remote)**  
*Jan 2024 – Present*

**Project:** Comm-Agg End-to-End Data Migration Solution  
**Description:** Designed and developed a metadata-driven, end-to-end data migration and analytics platform to ingest data from Guidewire S3 into Azure SQL, Data Lake, and Databricks Delta Lake, enabling scalable analytics, historical tracking, and regulatory compliance.

* Built and maintained cloud-native pipelines using Azure Data Factory to ingest from Guidewire S3 into Azure Data Lake Gen2 and Azure SQL, enabling financial, risk, and compliance analytics.
* Integrated Delta Lake and Databricks for performant, ACID-compliant, and schema-evolving data pipelines—supporting both batch and near-real-time workloads.
* Implemented Slowly Changing Dimension Type 2 (SCD2) strategy using Databricks to ensure accurate historical versioning of customer and policy data.
* Designed metadata-driven pipelines for automated schema validation, change capture, and ingestion, improving maintainability and deployment velocity by 70%.
* Engineered Python scripts for data quality checks, error handling, email alerting, and automated comparison reporting between ODS and Hive layers.
* Enabled cross-platform data movement from Azure SQL to Hadoop, staging data in landing/core/active zones, ensuring high data quality and accessibility.
* Used Databricks notebooks for on-demand data profiling, transformation, and exploratory analysis for business teams and QA.
* Played a key role in on-prem to Azure cloud migration initiatives, supporting data governance, audit, and finance platforms across banking and insurance clients.
* Orchestrated pipeline execution via Zena, automating job control, audit logging, and recovery logic, reducing manual intervention.

**Technologies Used:** Azure Data Factory (ADF), Azure Data Lake Storage Gen2, Delta Lake, Databricks, Blob Storage, Azure SQL, Python, Hadoop, Hive, Impala, Zena, Shell Scripting

**Consultant (Sr Big Data Engineer) | Client: DBS Bank, Singapore(On-Site)**  
*Jun 2022 – Jan 2024*

**Project:** MAS 637 and PILLAR3 Reporting  
**Description:** The MAS 637 project involved generating reports for the Monetary Authority of Singapore (MAS) to ensure regulatory compliance. This project analyzed pool statuses across various user types and tracked the changes over a 12-month period. PILLAR3 reporting focused on risk management, using predictive analytics to evaluate user behavior and default risks based on historical data.

* Developed and optimized PySpark scripts within the ADA (Advanced DBS Analytics) framework to analyze user pool statuses across different time periods.
* Automated trend analysis and default risk evaluations for users, improving regulatory compliance and decision-making.
* Provided insights into financial transactions and default status for users by integrating various data sources within the bank’s system.
* Delivered high-quality, compliant reports to the Monetary Authority of Singapore, ensuring that data and analysis met regulatory standards.
* Applied predictive analytics techniques to evaluate potential default risks, improving the bank's risk management strategies.
* Enhanced the reporting process by implementing Presto for faster data retrieval and analysis.
* Created and managed metadata structures within Collibra to support proper data governance, lineage tracking, and regulatory compliance.

**Technologies Used:** ADA (In-house framework), PySpark, Spark SQL, Presto, Hive, Hadoop, Airflow, Jupyter, Collibra, Python, Shell Scripting

**Project:** SAS Exit  
**Description:** The SAS Exit project was designed to migrate legacy SAS scripts into DBS Bank’s ADA platform. The migration involved converting complex SAS reports into PySpark-based solutions, improving the performance and scalability of the reporting system.

* Led the migration of complex SAS reporting scripts to PySpark within the ADA framework, significantly improving the performance of data processing.
* Designed a robust architecture for the ADA platform that met the requirements of the SAS reports and ensured seamless integration with existing systems.
* Developed and maintained metadata structures in Collibra, improving data governance and ensuring compliance.
* Automated the execution of PySpark jobs using Apache Airflow, ensuring timely report generation and reducing manual intervention.
* Collaborated with key stakeholders to understand business requirements and ensure the migration of reports aligned with their expectations.
* Collaborated with compliance and data governance teams to define domain-specific policies, ensuring consistent rule enforcement across systems.
* Integrated Collibra with bank-wide data pipelines to automate lineage tracking and data quality profiling.

**Technologies Used:** SAS, ADA (In-house framework), PySpark, Airflow, Collibra, Python, Spark SQL, Presto, Hive, Hadoop

**Consultant – Sr Big Data Engineer | Clients: ICICI, India(On-Site)**  
*Jan 2021 – Jun 2022*

**Project:** CRM Data Migration  
**Description:** This project involved migrating CRM data from multiple on-premises and cloud environments to an Azure-based platform. The goal was to ensure smooth data transfer while generating insightful reports for the business.

* Led the migration of CRM data to Azure SQL using Azure Data Factory (ADF), ensuring reliable and efficient data processing.
* Developed scalable ADF pipelines to handle data ingestion from multiple sources, including Vertica and Oracle databases.
* Managed a team of junior engineers, providing guidance on best practices for cloud migration and data governance.
* Automated the scheduling of data transfer tasks using Azure Logic Apps, improving overall efficiency and reducing delays in data processing.
* Liaised with business and compliance teams to define requirements for CRM data reporting and ensured pipeline design met audit and security protocols.

**Technologies Used:** Azure Data Factory (ADF), Blob Storage, SQL Database, Oracle, Key-Vault, Vertica, Azure Logic Apps

**Consultant – Sr Big Data Engineer | Clients: PayPal, USA(Remote)**  
*Jul 2020 – Jan 2022*

**Project:** H2H (HANA to Hadoop)  
**Description:** The H2H project involved migrating SAP HANA reports to a Hadoop-based environment, ensuring improved performance and scalability for reporting and analytics.

* Led the design and migration of SAP HANA reports to the Hadoop ecosystem, focusing on Spark and Hive for data processing.
* Developed end-to-end Hadoop data loading pipelines using Hive, Impala, and Tez for high-volume batch ingestion and reporting.
* Built complex SQL queries for analytics and integrated Python scripts to validate, clean, and transform data pre-load.
* Developed Spark and Hive-based reports to improve data accessibility and reporting performance.
* Utilized Hadoop, HDFS, and MongoDB to store and manage large volumes of data, improving data storage and processing efficiency.
* Collaborated with business teams to deliver reports through various channels, including email, web UI, and dashboards.
* Optimized the data migration process to ensure accurate and timely delivery of reports to end-users.

**Technologies Used:** PySpark, HDFS, Hive, MongoDB, GIT, Shell, Custom PayPal Frameworks

**Software Engineer – Data Engineer| Client: AbbVie, Chicago, USA(Remote)**  
*May 2019 – Jul 2020*

**Project:** IAP (Integrated Analytics Platform)  
**Description:** The IAP project was designed to centralize patient data from various applications into Hadoop for analysis. The goal was to process and analyze patient activity data, allowing healthcare providers to make informed decisions.

* Ingested Click stream data from various source applications into Hadoop, ensuring seamless integration and data processing.
* Developed Spark applications to analyze patient activity data, providing a comprehensive view of patient behavior and outcomes.
* Collaborated with business users to translate data transformation needs into scalable Hive/Impala jobs.
* Delivered performance-optimized ETL jobs using Hadoop ecosystem tools to support downstream analytics.
* Built efficient Hive and Impala queries to transform and analyze the data, enabling the creation of actionable insights for healthcare providers.
* Automated batch data processing using Autosys, ensuring timely updates and reducing manual interventions.
* Managed the end-to-end pipeline for data ingestion, processing, and reporting.

**Technologies Used:** Click Stream Data, Adobe Analytics, Apache Spark, Hadoop, Hive, Scala, Impala, Shell Scripting, Autosys

**Engineer – Geospatial Data Solutions | Client: Apple, USA(Remote)**  
*Mar 2017 – Apr 2019*

**Project:** SDC (Spatial Data Collaborator)  
**Description:** This project involved working with geospatial data sources to support the development of Apple’s map features. We processed and integrated data from multiple sources to create accurate and up-to-date maps.

* Ingested geospatial data from multiple sources (USGS, TIGER) into Hadoop for processing and integration into Apple’s mapping systems.
* Developed Hadoop-based pipelines to process large geospatial datasets, enabling more accurate and comprehensive map features.
* Worked with Apple’s mapping team to ensure the geospatial data was accurately merged, improving map development.
* Automated the data ingestion process from SFTP servers, reducing manual data processing time.

**Technologies Used:** Hadoop, Hive, HDFS, Python, Shell Scripting, MapReduce, USGS, TIGER Data Sources

**Trainee Engineer – Data Engineering | Client: Nokia, Finland(Remote)**  
*Jul 2015 – Mar 2017*

**Project:** Mapping Data Integration  
**Description:** This project focused on integrating Nokia's geospatial data into Hadoop for large-scale analytics, enabling the creation of advanced mapping and navigation solutions.

* Assisted in the integration of Nokia's geospatial data into Hadoop for large-scale analysis.
* Worked on developing data ingestion pipelines to process geospatial data from various sources, improving data quality and consistency.
* Contributed to developing tools to automate the ingestion process, reducing manual effort and improving operational efficiency.

**Technologies Used:** Hadoop, Hive, HDFS, Python, Shell, Spark

**Education**:

**Bachelor's degree in engineering** *June 2011 – May 2015*  
Jawaharlal Nehru Technological University – Hyderabad