Name: Nishanth Sura
Sr Data Engineer
Contact: +1 (832) 400-8698
Email: surannishanth@gmail.com
[LinkedIn]()

Microsoft CERTIFIED
AZURE DATA ENGINEER
ASSOCIATE

Professional Summary

- Proficient in designing and implementing **data processing pipelines** using various cloud platforms, including **Azure** and **AWS**.
- Skilled in **data modeling, data warehousing, data integration,** and **data analysis** using tools such as **Azure Synapse Analytics, Databricks, AWS Glue, AWS Athena, Snowflake,** and **Matillion**.
- Experienced in working with large datasets and leveraging technologies such as **Hadoop, MapReduce, Hive, Pig,** and **Spark** for distributed processing and analysis.
- Proficient in developing **machine learning applications** using tools such as **PySpark, Spark SQL, Spark MLlib,** and **Databricks AutoML.**
- Skilled in implementing automation and orchestration solutions using tools such as **Oozie, Apache Airflow,** and **Azure CI/CD.**
- Experienced in designing and implementing **real-time data processing solutions** using tools such as **Kafka** and **Spark Streaming**.
- Strong experience in developing **Unix shell scripts** for **database connectivity** and automation of **data processing** tasks.
- Proficient in working with databases such as **Dynamo DB, DocDB Cassandra, Terra data,** and **MongoDB.**
- Skilled in developing **data models, writing queries,** and **maintaining data integrity.**
- Experienced in **data profiling, data cleansing,** and **data migration** to ensure data accuracy and consistency.
- Experienced in developing **Tableau** reports for downstream users, including **designing data models, creating visualizations,** and **publishing reports** for end-users.
- Proficient in utilizing **Snowflake's virtual warehouses** to manage and scale compute resources dynamically based on workload requirements, reducing infrastructure costs and improving performance and scalability.
- Skilled in leveraging **Snowflake's automatic query optimization** and **tuning features** to improve query performance and reduce manual tuning efforts.
- Experienced in implementing **Snowflake data sharing** to securely share data with other organizations or teams without having to move or copy the data, reducing data movement costs and ensuring data security and governance.
- Highly skilled and experienced data engineer with expertise in the **Hadoop ecosystem, Google Cloud Platform, ETL tools, databases,** and **data migration** projects.
- Proven experience in designing and implementing scalable and fault-tolerant data processing pipelines using **Hadoop tools** and managing **Hadoop clusters**.
- Proficient in **Google Cloud Platform**, particularly with **BigQuery** and **Cloud Dataproc**, and experienced in designing and implementing **BigQuery data models** and using **Cloud Dataproc** for cloud-based data processing.
- Skilled in a variety of **databases**, including **Oracle, SQL, PostgreSQL,** and **Cassandra**, with expertise in designing and implementing **Cassandra data models** and using **SQL** for querying and manipulating data.
- Proficient in **Java API** and **REST API**, with experience using **Java API** to connect to **Cassandra** and designing and implementing custom **REST APIs** to expose data to downstream applications.
- Experienced in **ETL tools** such as **Informatica** and **Talend**, with experience designing and implementing complex data integration workflows and performing data quality checks and error handling.
- Expertise in implementing data cleansing and transformation techniques using **SQL**, **Hadoop tools,** and **ETL tools**, with experience in data profiling and data quality checks.
- Experienced in providing production support for **ETL** jobs, including job scheduling, monitoring, and troubleshooting, using tools such as **Control-M** and **Jenkins** for job automation.
- Skilled in data visualization and reporting using **Power BI**, collaborating with downstream data analysts to provide clean and well-structured data for analysis.

- Experienced in **Google Cloud Storage** for data storage and retrieval solutions, **Cloud Pub/Sub** for building real-time messaging and data streaming applications, **Cloud Dataflow** for building and running data processing pipelines, and **GCP**'s machine learning services such as **AutoML** and **TensorFlow** for building and deploying **machine learning** models.
- Designed and implemented data storage and retrieval solutions using **Google Cloud Storage**, built real-time messaging and data streaming applications using **Cloud Pub/Sub**, and built and ran data processing pipelines using **Cloud Dataflow**.
- Successfully migrated **50+ Terabytes** of data from **Teradata** to a new **cloud-based data warehouse** within the project timelines and budget.
- Improved data quality and accuracy by identifying and resolving data quality issues during the migration process.
- Developed and implemented new **data models** and reporting solutions that improved the accessibility and usability of data for business users.
- Contributed to the development of a scalable and **agile** data platform that enabled the organization to quickly adapt to changing business requirements.

TECHNICAL SKILLS

Cloud Services	Azure, AWS, GCP
Data Pipelines	Azure Data Factory, Azure Synapse Analytics, Databricks, AWS Glue, AWS Athena, Cloudera, Pig, Apache Flume, Sqoop, Hadoop tools, Cloud Dataproc
Databases	Azure Data Lake Storage, Matillion ETL, Kafka, DynamoDB, Cassandra, Snowflake, Teradata, MongoDB, Oracle, SQL, PostgreSQL
Data Warehousing	Azure Synapse Analytics, Snowflake, BigQuery
ETL Tools	Matillion ETL, Rivery ELT, Apache Airflow, Scala, Informatica, Talend
Programming Languages	PySpark, Spark SQL, Spark MLlib, Scala, Java, Unix shell scripting, C, SQL, Hadoop tools, C++
Machine Learning	Databricks AutoML, Spark MLlib
Data Cleansing and Transformation	SQL, Matillion ETL, Hive QL, Pig, Sqoop, Scala, SQL, Hadoop tools, ETL tools
Monitoring and Scheduling	Oozie, Apache Airflow
Data Visualization	Tableau, Power BI

Work Experience:

Client: p97networks, Houston, TX **May 2022 - Present**
Role: Sr Data Engineer

Responsibilities:

- Successfully migrated on-premises data to the cloud using **Azure Data Factory pipelines**, ensuring the security, availability, and scalability of data in the cloud. The project involved building robust and scalable **Azure Data Factory** pipelines to extract, transform, and load data from various sources into **Azure Data Lake Storage**, thus facilitating efficient and effective data processing.
- Implemented **Azure Synapse Analytics** for data warehousing and big data processing, and integrated it with **Azure Databricks** to create scalable data processing applications. This project involved utilizing **Hadoop, MapReduce,** and **Hive** with **Hive QL** to process large datasets and support data analysis needs. The project delivered robust and scalable data processing solutions that met the client's needs.

- Designed and developed data integration solutions using the **Matillion ETL tool** to connect to different data sources, transform data, and load it into target systems. The project involved working with complex data sources and delivering robust data integration solutions that met the client's business requirements.
- Performed performance tuning in **Spark** to optimize data processing and improve query performance. This project involved utilizing **PySpark, Spark SQL,** and **Spark MLlib** for building data processing applications that are scalable and efficient. The project resulted in optimized data processing and improved query performance, resulting in enhanced business value for the client.
- Developed **Unix shell scripts** for database connectivity and automation of data processing tasks. This project involved designing and developing robust automation solutions for data processing tasks, which resulted in improved efficiency and productivity.
- Utilized **Kafka with Spark Streaming** to process and analyze real-time data. This project involved leveraging real-time data processing capabilities to create data processing solutions that met the client's business requirements.
- Developed **Oozie workflows** to automate data processing and job scheduling tasks. This project involved creating robust data processing workflows that streamlined the data processing and job scheduling tasks, resulting in improved efficiency and productivity.
- Worked closely with downstream data scientists to ensure the accuracy and quality of data outputs. This project involved working collaboratively with downstream data scientists to ensure the accuracy and quality of data outputs, resulting in enhanced business value for the client.
- Implemented **Azure CI/CD** to ensure seamless deployment of code changes and updates to production environments. This project involved creating robust and scalable **CI/CD pipelines** that enabled the seamless deployment of code changes and updates to production environments, resulting in improved efficiency and productivity.
- Automated machine learning with **Databricks AutoML**, leveraging **Spark** and **MLflow** tracking to automate the model selection, hyperparameter tuning, and deployment. This project involved creating robust machine learning solutions that automated the model selection, hyperparameter tuning, and deployment, resulting in enhanced business value for the client.
- Integrated **Databricks Delta Lake** to provide ACID transactions, versioning, and data lineage capabilities for reliable and performant data processing pipelines and machine learning workflows. This project involved leveraging **Databricks Delta Lake** to create robust and reliable data processing pipelines and machine learning workflows, resulting in enhanced business value for the client.
- Utilized **Databricks Runtime**, a pre-configured **Spark** environment optimized for performance and reliability, to streamline development workflows and achieve faster time-to-value for projects. This project involved utilizing **Databricks Runtime** to streamline development workflows and achieve faster time-to-value for projects, resulting in improved efficiency and productivity.

One specific use case involved a medium-sized retail company looking to optimize its sales and marketing strategies. The company had multiple data sources, including customer demographic, transactional, and social media data. The goal was to analyze customer behavior and preferences to identify potential areas for improvement.

To address this challenge, I designed and implemented a data processing pipeline that utilized **Azure Synapse Analytics** and **Databricks** for data analysis and **machine learning**. The pipeline included data integration from multiple sources, data cleaning and transformation, and machine learning algorithms for customer segmentation and recommendation engines.

The pipeline was able to identify key customer segments based on demographic data, transactional data, and social media data. The machine learning models were used to recommend products and services based on customer preferences and behavior, resulting in a significant increase in sales and customer satisfaction.

Client: HTD Health, Newyork NY **March 2021 - May 2022**
Role: Sr Data Engineer

Responsibilities:

- Designed and implemented **AWS data pipelines** using **AWS Glue** and **AWS Athena**, ensuring that data was extracted, transformed, and loaded into the desired target systems. This involved creating and configuring data sources, creating mappings and transformations, and testing and validating the pipeline to ensure data accuracy and consistency.
- Worked with databases such as **Dynamo DB** and **Cassandra**, where I developed complex data models, wrote complex queries, and implemented strategies to maintain data integrity. This involved designing database schemas, creating indexes, and optimizing query performance for faster data retrieval.
- Extracted data from the **Hadoop Distributed File System (HDFS)** using **Cloudera** and **Pig** for data extraction. This helped in processing large datasets by splitting them into smaller chunks, parallelizing data processing, and leveraging distributed computing resources.
- Utilized **EC2** and **S3** for data processing and storage, where I worked on configuring instances, writing scripts for data processing, and managing data storage. This involved setting up **EC2 instances** and **S3 buckets**, configuring security policies, and writing scripts in **Python** or **Bash** for data processing and ETL.
- Created **Spark workflows** using **Scala** for data pull from **AWS S3**, which helped in performing distributed processing of data. This involved creating **Spark RDDs** and **DataFrames**, defining transformations and actions, and running Spark jobs on a distributed cluster.
- Integrated machine learning models for predictive analytics in the project, utilizing Python libraries such as **Scikit-learn** and **TensorFlow**. This involved developing and fine-tuning models for tasks such as patient outcome prediction and risk assessment, enhancing decision-making capabilities within the healthcare data ecosystem.
- Deployed a **natural language processing (NLP)** solution to analyze unstructured data from medical records and patient feedback, contributing to sentiment analysis and information extraction for insights on patient satisfaction and healthcare service improvements. Leveraged libraries such as **NLTK** and **spaCy**, creating pipelines that processed textual data to derive actionable insights for the organization.
- Worked on **Snowflake schemas** and **data warehousing**, where I designed and implemented data models, created tables, and loaded data into them. This involved understanding the business requirements, designing the schema and tables accordingly, and writing SQL queries to load data from various sources.
- Worked with **Snowflake** utilities like **snowsql, snowpipe,** and **Time Travel**, where I wrote scripts for data extraction and transformation. This involved setting up **Snowflake** accounts, configuring utilities, and writing SQL scripts to extract, transform, and load data from various sources.
- Utilized **Rivery ELT** to integrate data from various sources and transform it for downstream consumption. This involved setting up Rivery accounts, configuring data sources, and designing ETL pipelines using Rivery's visual interface.
- Worked with **Terra data** and **MongoDB**, where I created data models and wrote complex queries to retrieve data. This involved understanding the data structure, designing the data model accordingly, and writing complex queries in **SQL** or **NoSQL** languages to retrieve data efficiently.
- Performed data profiling, data cleansing, and data migration to ensure data accuracy and consistency. This involved identifying data quality issues, cleaning the data using various tools and techniques, and migrating the data to new systems while ensuring data integrity.
- Wrote **Scala** code to process data efficiently and effectively. This involved using **Scala's** functional programming paradigm, leveraging **Spark's APIs**, and implementing algorithms for data processing and analysis.
- Worked with **Apache Flume, Sqoop,** and **Pig** to extract and process data from various sources. This involved setting up **Flume** and **Sqoop** agents, configuring data sources, and writing **Pig** scripts to process and transform data.
- Developed and automated **Snowflake** jobs using **Apache Airflow**, which helped schedule and monitor data pipelines. This involved setting up Airflow instances, creating DAGs (Directed Acyclic Graphs) to define job workflows, and scheduling jobs for automated execution.

- Created **Tableau** reports for downstream users, which involved designing data models, creating visualizations, and publishing reports for end-users. This involved understanding the business requirements, designing the data model and visualizations accordingly, and publishing reports on Tableau server for end-users to access.
- Implemented **Snowflake data sharing** to securely share data with other organizations or teams without having to move or copy the data. This helped in reducing data movement costs and ensuring data security and governance.
- Used **Snowflake's virtual warehouses** to manage and scale compute resources dynamically based on workload requirements. This helped in reducing infrastructure costs and improving performance and scalability.
- Leveraged **Snowflake's automatic query optimization** and **tuning** features to improve query performance and reduce manual tuning efforts. This helped in optimizing query execution times and improving overall system performance.

Client: Next on top, Bangalore, India **June 2019 - Jan 2021**
Role: Big Data Engineer

Responsibilities:

- Designed and implemented scalable and fault-tolerant data processing pipelines using **Hadoop tools**, including **HDFS**, **MapReduce, Hive, Pig,** and **Oozie**. These pipelines were capable of handling large amounts of data and processing it efficiently and accurately.
- Deployed and managed **Hadoop clusters** for data processing and analysis. I ensured that these clusters were properly configured, monitored, and maintained to ensure optimal performance and reliability.
- Proficient in **Google Cloud Platform**, particularly with **BigQuery** and **Cloud Dataproc**, and designed and implemented **BigQuery** data models for data analysis and used **Cloud Dataproc** to run data processing jobs in the cloud. This allowed me to leverage the power of the cloud to process and analyze large amounts of data quickly and efficiently.
- Designed and implemented data models using a variety of **databases**, including **Oracle, SQL, PostgreSQL,** and **Cassandra**. These data models were designed to be efficient and effective for storing and retrieving data in a variety of contexts.
- Developed complex data integration workflows using **ETL tools** such as **Informatica** and **Talend**, and performed data quality checks and error handling. These workflows allowed me to integrate data from multiple sources and ensure that the data was of high quality and accuracy.
- Designed and implemented custom **REST APIs** using **Java API** to connect to **Cassandra** and expose data to downstream applications. This allowed me to provide easy access to data for downstream applications securely and efficiently.
- Implemented data cleansing and transformation techniques using **SQL, Hadoop tools,** and **ETL tools**, and conducted data profiling and data quality checks. These techniques helped me to ensure that the data being processed and analyzed was accurate and reliable.
- Provided production support for ETL jobs, including job scheduling, monitoring, and troubleshooting, using tools such as **Control-M** and **Jenkins** for job automation. This ensured that the data processing and analysis workflows were running smoothly and efficiently at all times.
- Collaborated with downstream data analysts to provide clean and well-structured data for analysis, and used **Power BI** for data visualization and reporting. This collaboration ensured that the downstream data analysts had access to high-quality data that was well-structured and easy to analyze, while the visualization and reporting provided valuable insights into the data.
- Designed and implemented data storage and retrieval solutions using **Google Cloud Storage**. This allowed me to store large amounts of data securely and efficiently, making it easily accessible for processing and analysis.
- Built real-time messaging and data streaming applications using **Cloud Pub/Sub**. This allowed me to process and analyze data in real-time, providing valuable insights into the data as it was being generated.
- Built and ran data processing pipelines using **Cloud Dataflow,** and integrated with other **GCP** services such as **BigQuery** and **Cloud Pub/Sub**. This allowed me to process and analyze large amounts of data quickly and efficiently while integrating with other **GCP** services for a seamless workflow.
- Built and deployed machine learning models using **GCP's machine learning** services such as **AutoML** and **TensorFlow,** and integrated with other **GCP** services such as **Cloud Storage** and **BigQuery**. This allowed me to leverage the power of machine

learning to gain valuable insights from the data being processed and analyzed while integrating with other **GCP** services for a seamless workflow.

Client: Casttree, Bangalore, India **Oct 2017 - June 2019**
Role: Data Analyst

Responsibilities:

- Collaborated with the business stakeholders to gather and analyze requirements for data migration from **Teradata** to a **cloud-based data warehouse**.
- Conducted a comprehensive analysis of existing data sources to identify data quality issues and data modeling requirements to ensure efficient migration to the new system.
- Developed a detailed project plan with timelines, milestones, and deliverables, and communicated the plan to stakeholders, project team members, and management.
- Coordinated with the **ETL** development team to design and implement **data pipelines** to extract, transform, and load data from Teradata to the new data warehouse.
- Conducted extensive **data validation** and **reconciliation** to ensure the accuracy and consistency of migrated data.
- Collaborated with business users and **data analysts** to develop and implement new **data models** and **reporting solutions** on the new platform.
- Conducted **data analysis** and provided **insights** to support business decisions and improve data quality.
- Developed and maintained technical documentation for the project, including **data mappings, data lineage,** and **data dictionaries**.
- Utilized **SSIS (SQL Server Integration Services)** as the primary tool for designing and deploying **ETL processes** for the migration from **Teradata** to the new **cloud-based data warehouse.**
- Configured and optimized **SSIS packages** for high performance and scalability, leveraging features such as **parallel processing, batching,** and **error handling**.
- Implemented **data integration** solutions using **SSIS** for real-time and batch processing scenarios, including integrating **third-party data sources** and **applications**.

Key Achievements:

- Successfully migrated **50+ Terabytes** of data from Teradata to the new cloud-based data warehouse within the project timelines and budget.
- Improved data quality and accuracy by identifying and resolving data quality issues during the migration process.
- Developed and implemented new data models and reporting solutions that improved the accessibility and usability of data for business users.
- Contributed to the development of a scalable and agile data platform that enabled the organization to quickly adapt to changing business requirements.

Client: Fundditt Solutions, Bangalore, India **Jan 2016 - Oct 2017**
Role: Python developer - Health Insurance Lead Prediction

Responsibilities:

- Performed **exploratory data analysis (EDA)** on a massive database of **50,000+** clients to extract meaningful insights and patterns using advanced statistical techniques such as **correlation analysis** and **feature engineering**.
- Collaborated with the project team to define the project scope, objectives, and milestones, using project management tools and methodologies such as **agile** and **scrum**.
- Designed and developed a highly accurate logistic regression model using **Python**, with a precision of **75.68%**, by utilizing **cross-validation techniques** and **hyperparameter tuning**.
- Integrated the model into a user-friendly web application with a responsive and scalable interface that allowed for the efficient input of client details and produced accurate predictions of the likelihood of a client purchasing health insurance.
- Conducted thorough testing of the web application, ensuring its performance, robustness, and scalability, and integrated it seamlessly with the existing **customer relationship management (CRM**) system.
- Provided regular updates on the project progress to the management team, stakeholders, and clients, and reported key performance indicators (KPIs) such as **accuracy, precision, recall,** and **F1 score.**
- Conducted a comprehensive post-mortem analysis of the project to identify opportunities for improvement and recommended future enhancements, such as incorporating **machine learning algorithms**, **data visualizatio**n tools, and **cloud-based technologies**.

PUBLICATIONS

Conversational Agent for Student Service and Support **May 2019**
International Journal of Research in Engineering, Science and Management **Volume-2, Issue-5**

Education Qualifications:

Stevens Institute of Technology, New Jersey Master's in Information Systems

R.V. College of Engineering, Bangalore, India B.E. in Information Technology