Manoj Nani

Senior Cloud Data Engineer

Address Dallas, TX 77058

Phone +1 (682) 232-4719

E-mail manojasnani@outlook.com

Professional Summary:

A Senior Data Engineer with over 7 years of committed expertise in data engineering, also an expert in designing and delivering data solutions that support business intelligence and analytics efforts. From data intake using ETL tools like **Apache Kafka** and **AWS Glue** to data warehousing via technologies like **Snowflake** and **Amazon Redshift**, my expertise covers the whole data ecosystem. Have experience in creating and improving data pipelines for processing massive amounts of data using tools like **Apache Spark, Hadoop, and Flink**, also created and administered data lakes and data warehouses on cloud platforms like **AWS** and **Azure**, assuring scalability and high availability. Possess hands-on experience in connecting **Azure Databricks** with **Azure Synapse Analytics** to conduct data analysis utilizing **Apache Spark** and then visualizing the results using **Power BI**. Worked in different domains like Manufacturing, Retail, Finance and Health care. Have a lot of expertise with **agile methodologies** and am involved in requirements gathering, designing business use cases, building end user reports, deriving metrics and analysis from Business Data. An expertise of working with cross-functional teams to develop data solutions that enable businesses to extract useful insights and drive data-driven decision-making.

Skills

Azure Cloud Platform: ADF, ADLS2, Azure SQL DB, Azure Synapse Analytics, Databricks, Azure Cosmos DB, Azure Stream Analytics, Event Grid, Azure DevOps.

AWS Cloud Platform: Amazon EC2, Amazon S3, Amazon RDS, DynamoDB, AWS Lambda, Cloud Watch, SNS, SQS, Amazon Redshift, Amazon EMR, AWS Glue, AWS Step Functions.

Programming Languages: Python, Java, Shell Scripting, SQL, PySpark.

Big data Technologies: Hadoop, HDFS, Map Reduce, HIVE, PIG, HBase, Sqoop, Apache Spark, Apache Kafka.

Data Storage & Databases: Oracle, Microsoft SQL Server, MySQL, MongoDB, DynamoDB, Cosmos DB, PostgreSQL, SQL databases.

Visualization Tools: Power BI, Amazon Quicksight, Tableau.

Cloud Stack: AWS, Azure, Snowflake.

Methodologies: Waterfall, Agile/Scrum, Kanban, SDLC.

Orchestration: Apache Airflow, Oozie.

Operating systems: Linux, Unix, Windows.

Certifications

AWS Certified: Data Analytics – Specialty

Microsoft Certified: Azure Data Engineer Associate

Work History

Azure Data Engineer

Dec 2022 – Current

BEHR Corporation, Santa Ana, California

Project Description:

- **BEHR Corporation**, a leading manufacturer and supplier in the world of paints and coatings, faced a significant challenge in managing and reporting sales and revenue data effectively.
- To achieve this, developed ETL pipelines that extracted sales data from various regions and stores, transformed and aggregated this data, and generated daily, weekly, and monthly sales reports for management, ensuring accuracy and timeliness.

Responsibilities:

- Collaborated with business/user groups to understand the business process, gather requirements, analyze, design, development, and implementation according to client requirements.
- Developed **data models** that streamlined data processing pipelines in the **Azure environment**, resulting in an increase of **25%** in productivity.
- Designed and implemented data ingestion pipelines to collect sales and revenue data from databases, APIs, including point-of-sale systems, online sales channels, and distributor transactions.
- Utilized **Azure Data Factory** to connect to data sources in the cloud and ingest data from these sources into a data warehouse, **Azure Synapse Analytics** (formerly SQL Data Warehouse).
- Used **Azure Databricks** for data transformation tasks, job scheduling, and sending notifications as part of data processing workflows.
- Set up and managed Azure data storage solution, **Azure Data Lake Storage** to store historical sales and revenue data.
- Used **Apache Airflow** to define and orchestrate the entire ETL workflow, create directed acyclic graphs (**DAG**s) that represent the sequence of tasks required for data extraction, transformation, and loading.
- Implemented **Flask API** to provide a **RESTful interface** for accessing and retrieving the analyzed data, allowing seamless integration with other applications and services.
- Optimized database schemas and SQL queries in **Azure Synapse Analytics** for efficient data retrieval and analysis, created data models, and developed interactive reports and dashboards using **Power BI.**
- Generated visually compelling reports and dynamic dashboards in the form of bar charts, pie charts, and tables to represent sales and revenue insights using **Power BI** to stakeholders.
- Used Azure pipelines in **Azure DevOps** for **CI/CD** to automate deployment and updates of data engineering solutions.
- Proficient in applying **Agile** principles and practices, particularly within the **Scrum** framework.

- Continuously monitored the health and performance of data pipelines, databases, and storage using **Azure Monitor**.
- Generated, maintained, and analyzed Azure monitoring **dashboards, reports, and trends**, minimizing customer pain points by **30%**.
- Used **Splunk** for the real-time monitoring capability to keep a close eye on data streams as they are ingested.
- Optimized data pipelines and workflows for scalability and responsiveness, leveraging **Azure Resource Manager**.
- Implemented security measures and access controls in Azure, including **Azure Active Directory and Azure Key Vault** to ensure the protection of sensitive sales and revenue data and maintain regulatory compliance.
- Collaborated with business analysts and stakeholders to understand reporting requirements and ensured that data solutions aligned with business goals.
- Utilized **SharePoint** for the documentation to describe how daily, weekly, and monthly sales reports are generated from the transformed and aggregated data.

Environment: Azure Data Factory, Azure Synapse Analytics, Databricks, Azure Data Lake Storage, PowerBI, Azure Active Directory, Azure Key Vault, Azure Monitor, Resource Manager, Azure DevOps, Agile/SCRUM, PowerShell, Azure Cloud Shell, SharePoint**.**

Azure Data Engineer

Oct 2021 – Nov 2022

NCR Corporation, Addison, Texas

Project Description:

- NCR Corporation is a global payment technology company that invests in IT to facilitate secure electronic transactions worldwide.
- Due to significant surge in the volume of online transactions, the existing data pipelines and systems struggled to efficiently handle the increasing data loads and user demands, potentially lead to extended processing times, and an overall degradation of service quality.
- To achieve this, developed ETL pipeline that enabled efficient processing of large datasets across multiple nodes, reduced processing times and enhanced scalability.

Responsibilities:

- Collaborated with Business Analysts to gather business requirements and identify workable items for further development.
- Designed and Implemented ETL processes to extract transaction data from APIs, databases, log files, transformed it into a standardized format, and loaded it into data storage systems.
- Used **REST APIs** to fetch transaction data from these sources into the ETL pipeline.
- This involved using **Azure Data Factory** or custom scripts to ensure data consistency and accuracy.
- Created and maintained a data warehousing solution on Azure, **Azure Synapse Analytics** (formerly SQL Data Warehouse), to store and manage transaction history, user profiles, and payment data.
- This includes optimizing database schemas and queries for efficient data retrieval and analysis.
- Implemented real-time data streaming using **Azure Stream Analytics** to divide streaming data into batches as an input to **Azure Databricks** for batch processing to capture and process transaction data as it occurs.
- Utilized **PySpark** in Azure **Databricks** to extract and load data and perform SQL queries using **Spark SQL**.

- Performed Data Analysis, Data Migration, Data Cleansing, Transformation, Integration, Data Import, and Data Export through **Python.**
- Involved in designing and deploying multi-tier applications using Azure services like **Azure Virtual Machines**, **Azure Blob Storage**, **Azure Functions**, **Azure SQL Database, Azure Key Vault** focusing on high-availability, fault tolerance, and auto-scaling in **Azure Resource Manager**.
- Responsible for building the data ingestion pipelines using Hadoop and **Azure Databricks** with Scala as the Data Processing Engine and **Azure Data Lake Storage** as the Consumption Layer
- Supporting Continuous storage in Azure using **Azure Managed Disks, Azure Blob Storage, Azure Archive Storage,** and created Storage Accounts and configured Snapshots for **Azure Virtual Machines.**
- Used **Apache Flink** for real-time data processing to handle incoming online transactions as they occur.
- Designed, developed, and implemented pipelines using Python API (**PySpark**) of **Apache Spark**.
- Generated reports on predictive analytics using Python and **Power BI**, including visualizing model performance and prediction results.
- Collaborated with data analysts to create interactive dashboards and reports using **Power BI.**
- Monitored and optimized the performance of the **ETL pipeline** to reduce processing times and ensure data is available for analytics and reporting in a timely manner.
- Provided insights into user behavior, transaction trends, and app performance to support data-driven decision-making.
- Utilized **Kusto Query Language** in **Azure Log Analytics** for querying and analyzing data, which is used for monitoring, diagnostics, and log analysis.
- Used Azure pipelines in **Azure DevOps** for **CI/CD** and **Azure DevTest Labs** for environment provisioning and management.
- Utilized **Scrum** methodology for team and project management.
- Established monitoring and alerting systems to proactively detect and address issues in the ETL pipeline. **Azure Monitor** can be valuable tools for this purpose.
- Architected an automated environment using **PowerShell** and **Azure Cloud Shell** to deploy Azure data solutions, driving cost savings of **30%**.
- Documented the **ETL pipeline** design, configuration, and processes for knowledge sharing using **Confluence** and collaborating with cross-functional teams, including analysts, and business stakeholders.
- Regularly maintained and updated the ETL pipeline to accommodate changes in data sources, business requirements, and **Azure services**.

Environment: Azure Data Factory, Azure Synapse Analytics, Databricks, Python, Virtual Machines, Azure Blob Storage, Azure Functions, SQL Database, Azure Key Vault, Azure Stream Analytics, Azure Data Lake Storage, Managed Disks, REST APIs, Flink, Power BI, Scrum Methodology, Azure Monitor, Azure DevOps, Confluence.

Big Data Engineer **Oct 2018 – Sept 2021**

DBS, Hyderabad, India

Project Description:

- DBS Bank aims to enhance customer support by providing 24/7 assistance for routine queries, such as balance inquiries, transaction history, and account updates.

- There was a challenge in streamlining customer support for routine queries to improve efficiency and reduce response times. However, incomplete, or inaccurate data from various sources led to incorrect responses from the chatbot.
- Implemented robust data integration processes, including data cleansing and validation, and ensured that data from different sources is accurate and complete.

Responsibilities:

- Extracted customer data, including account details, transaction history, and contact information, from DBS core banking system, CRM databases, Transaction logs.
- Established data pipelines to ingest customer data, historical queries, and chatbot interactions for analysis and model training.
- Selected and configured an appropriate database system **Amazon RDS**, for storing customer data securely and efficiently.
- Used **AWS EMR** to implement robust data integration processes, including data cleansing and validation, to ensure that data from different sources is accurate and complete.
- Developed **ETL** (Extract, Transform, Load) processes to preprocess and clean data before storing it in the data warehouse, **Amazon Redshift**.
- Ensured data accuracy and completeness by cleansing and validating incoming data.
- Used **AWS Glue** for data cleansing and transformation, leveraged **AWS Lambda** for custom data validation and cleansing scripts.
- Designed and implemented a data warehousing solution **Amazon Redshift** to store and consolidate customer interaction data for reporting and analytics.
- Integrated the chatbot with customer data and the data warehouse to enable personalized responses and maintain a history of customer interactions.
- Implemented data visualization tool, **Amazon QuickSight** to create dashboards and reports for monitoring chatbot performance and customer interaction trends.
- Used **Jenkins** for CI/CD pipeline for the chatbot application, allowing for automated testing and deployment of updates to ensure continuous improvement and stability.
- Worked within **Scrum**, collaborating with cross-functional teams to prioritize and deliver chatbot features iteratively. Participated in sprint planning, daily stand-ups, and retrospectives.
- Implemented monitoring and alerting systems, **Amazon CloudWatch** to track the health and performance of the chatbot, database, and data pipeline.
- Optimized the performance of AWS-hosted applications with **CloudWatch** monitoring resulting in a **10%** decrease in error rates.
- Implemented data backup to minimize data loss and downtime, continuously monitored and optimized costs to ensure cost-effectiveness.
- Ensured that customer data is handled securely and in compliance with data privacy regulations.
- Designed the architecture to be scalable to accommodate increasing chatbot usage and ensure high availability to provide 24/7 support.
- Created comprehensive documentation for the architecture, data processes, and deployment procedures. Share knowledge and provide training to team members.
- Collaborated closely with business stakeholders, customer support teams, and developers to understand requirements, gather feedback, and ensure that the chatbot meets customer needs.

Environment: Amazon RDS, AWS ETL Pipeline, AWS Glue, AWS EMR, Amazon Redshift, Amazon QuickSight, CloudWatch, AWS Lambda, Jenkins, Agile/Scrum.

Hadoop Developer **Oct 2016 – Sept 2018**

Optum, Hyderabad, India

Project Description:

- In the healthcare industry, especially within Optum Healthcare, there was a critical need for more robust and comprehensive analytics and reporting tools.
- The current limitations in data analysis and reporting capabilities presented several challenges like Data Complexity, Cost Control, Quality Improvement.
- Developed advanced analytics tools to analyze patient data, track healthcare outcomes, and generated insights for better decision-making.

Responsibilities:

- Developed a strategy for migrating existing patient data, historical records, and relevant healthcare information to the new analytics platform. Ensured data integrity and security during the migration process.
- Developed and maintained data ingestion processes to acquire diverse healthcare data sources, including **EHR**s, claims data, and patient-generated data.
- Utilized **Hadoop t**echnologies like **Sqoop** for efficient data ingestion.
- Implemented **ETL processes** to cleanse, preprocess, and transform raw healthcare data into a structured format suitable for analysis.
- Leveraged **Hadoop tools** such as **Apache Spark** for data processing and transformation.
- Integrated and consolidated healthcare data from various sources into the **Hadoop Distributed File System (HDFS).**
- Used **HDFS** to store vast amounts of healthcare data, including patient records, medical images, and administrative data.
- Utilized **MapReduce** to perform data transformations, aggregations, and calculations on patient data to derive meaningful insights.
- Used **Hive** which provides a SQL-like interface to query and analyze data stored in Hadoop and to create structured tables and perform ad-hoc queries for reporting and analytics.
- Used **Pig** for handling unstructured or semi-structured data often found in healthcare, and **HBase** to storing and managing semi-structured healthcare data, such as patient records with varying attributes.
- Used **Sqoop** to import and export data between **HDFS and relational databases**, which are commonly used in healthcare systems, helps in integrating structured data sources with Hadoop-based analytics.
- Optimized algorithms for distributed processing on **Hadoop clusters**.
- Implemented real-time data processing solutions using **Apache Kafka** and **Spark** Streaming to support real-time analytics requirements.
- Implemented data security measures and access controls within the **Hadoop ecosystem** to ensure the protection of sensitive patient data.
- Worked with data visualization developers to integrate Hadoop data into interactive dashboards and reporting tools.
- Ensured that healthcare professionals can access and explore insights easily.

- Collaborated closely with business analysts, and healthcare professionals to understand analytics requirements and deliver valuable insights.
- Ensured that the **Hadoop infrastructure** is scalable to handle growing volumes of healthcare data.
- Managed cluster resources efficiently to optimize costs.

Environment: Hadoop, Sqoop, ETL (Extract, Transform, Load), Apache Spark, HDFS, MapReduce, Hive, Pig, HBase, Apache Kafka, Data Security Measures, Resource Management.

Education

Master of Science: Computer Science
University of Houston-Clear Lake - Houston, Texas

Bachelor of Technology: Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering & Technology - Hyderabad, India