**Senior Data Engineer**

**Email:**  dineshn1408@gmail.com

**Phone:** 4702315325

**SUMMARY**

* Around 8+ years of experience in the IT industry specializing in Data Engineering, Big Data Environment, and Hadoop ecosystem. Experience with Leading cloud systems such as AWS and Azure.
* Experienced in data manipulation and analysis using **Python, PySpark**, and libraries such as **Pandas, and**

**Sklearn**

* Demonstrated proficiency in optimizing **Snowflake's** unique features for dynamic data integration, storage, and advanced analytics, and excelled in data management with systems like **CTMS and Rave**
* Advanced proficiency in**, Scala** for developing high-performance data processing tasks, executing complex **SQL queries** to manipulate and analyze vast datasets efficiently,

* Proficient in leveraging **AWS services** such as **EC2, S3, and RDS, EMR, VPC, IAM, DynamoDB, Redshift** and Implemented serverless data processing workflows using **AWS Lambda,** and other AWS services for various data engineering tasks.
* Integrated **AWS Glue Catalog** with Snowflake for unified metadata management, enabling streamlined data discovery and governance.
* Developed and optimized **PySpark batch jobs** with parameterized configurations, improving resource utilization and processing efficiency in AWS EMR and Databricks environments.
* Extensive use of **Hadoop** and **Spark** for large-scale data storage, processing, and analytics, improving application performance and data management Efficiency
* Proven ability to optimize **partitioning strategies** and **caching mechanisms** in PySpark workflows to reduce job execution time for high-volume transactional datasets.
* Implemented a microservices architecture for a complex data processing application, involving automated deployment and monitoring using **AWS CloudWatch** and **Jenkins** for **CI/CD**,
* Skilled in data visualization and reporting using **Power BI and Tableau**, and advanced knowledge in **T-SQL AND c** for insightful analytics and decision Support
* Strong background in **Agile/Scrum Development**, Scaled Agile Framework, and Continuous Integration/Continuous Delivery (**CI/CD**).
* Proficient in implementing data migration and processing workflows utilizing **Azure cloud services**, including **Azure Blob Storage, Azure Data Lake, Azure Data Factory, Azure SQL, Azure SQL Data Warehouse, Azure Analytics, and Azure Databricks**..
* Proficient in implementing and managing significant data ecosystems using tools like **Hadoop MapReduce, Apache Spark, NoSQL, and Informatica Cloud Services.**
* Good knowledge of NoSQL databases like **HBase, Cassandra, and MongoDB,** with good knowledge of **COSMOS DB** and **SQL databases like Teradata, Oracle, PostgreSQL, MySQL, and SQL Server.**
* Showcased proficiency in managing various file formats, including Parquet, **XML, CSV**, and **ORC**, to optimize data management processes for structured and semi-structured datasets.
* Strong leadership experience, having managed data engineering teams and effectively handled onsite-offshore project coordination in service-based company settings.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
|  **Title** |  **Tools Used** |
| **Cloud Platforms** | AWS (EC2, S3, RDS, EMR, VPC, DynamoDB, Redshift, Lambda, CloudWatch), Azure (Blob Storage, Data Lake, Data Factory, SQL, SQL Data Warehouse, Analytics, Databricks). GCP (Big Query, Google Cloud Composer, Cloud Dataflow, Cloud SQL, Cloud Bigtable, Firestore Deployment Manager Google Kubernetes Engine) |
| **Programming Languages** | Python, Pyspark ,Scala, SQL |
| **Big Data Technologies** | Hadoop (HDFS, YARN, MapReduce), Spark (SQL, Scala Streaming), Hive, Pig, HBase, Apache Kafka, Apache NiFi |
| **Data Warehousing** | Informatica Cloud, Five Tran, Amazon Redshift, Azure SQL Data Warehouse, Snowflake |
| **Databases** |  Oracle, MySQL, COSMOS DB, PostgreSQL, Neo4j, Graph DB, Teradata, MongoDB, Cassandra, HBase, DynamoDB, Azure SQL |
| **ETL Tools** | Informatica PowerCenter, SQL Server Integration Service (SSIS), AWS Glue, Talend, Azure Data Factory |
| **Project Management & Collaboration** | Salesforce, ITIL, SDLC, JIRA, Rally, GitHub, Docker |
| **Scheduling & Workflow Automation** | Apache Airflow, Autosys, Windows Scheduler |
| **Data Visualization** | Power BI, Tableau |

 ***Client:* Fannie Mae Washington, DC |DEC 2022 - till now *| Role: Sr. AWS Data Engineer***

* Designed and implemented an automated data pipeline that streamlined fraud detection to reduce processing time, used Python for scripting and Apache Airflow for workflow automation to integrate data from several sources.
* Employed AWS Lambda to perform on the-fly data validation checks whenever new transaction data is ingested from Salesforce and SQL databases.
* Utilized AWS CloudFormation to automate infrastructure deployment, ensuring consistency and scalability.
* Designed Snowflake schemas tailored for transactional and analytical use cases, ensuring efficient storage and retrieval of fraud detection data.
* Tuned Snowflake queries by implementing clustering keys, partition pruning, and result caching, significantly improving query performance for large datasets.
* Utilized stored procedures and complex SQL queries for fraud detection analysis: Analyzed historical transaction data to identify patterns.
* Built PySpark workflows to validate data integrity, identify anomalies, and ensure completeness before loading into Snowflake.
* Developed and implemented a Neo4j-based solution to model and analyze complex transactional relationships, enhancing fraud detection with graph database technology.
* Utilized Docker for containerizing data processing applications, ensuring consistent and scalable deployment across different environments.
* Transition to using Snowflake as the centralized platform for consolidating transactional data from diverse systems, leveraging its ETL capabilities to provide a unified view of customer and financial data.
* Used Amazon S3 to construct a scalable data and processing infrastructure for cost-effective data
* Deployed Amazon EC2 instances to dynamically scale computational resources for processing data and running complex fraud detection algorithms.
* Integrated CI/CD pipelines using Git for version control and streamlined deployment processes, boosting productivity and efficiency.
* Utilize Spark, integrated within Snowflake, to enhance the ETL pipeline, applying parallel processing and in-memory computations to handle transactional data effectively.
* Spearheaded the development of a real-time analytics platform using Spark Streaming, enabling processing and analyzing streaming data from online transactions and analytics tools to monitor transactions, detect fraud, and generate warnings utilizing technologies such as Apache Kafka.
* Automated Kafka cluster operations using Terraform, improving deployment efficiency and consistency.
* Established advanced monitoring and alerting systems for ETL processes, using tools like AWS CloudWatch and custom Python scripts.
* Leveraged Python to create machine learning models Directly in Snowflake, including logistic regression and random forests, to analyze transactional behaviors and accurately predict fraudulent activities.
* Building data standards to provide consistency in data formats, naming conventions, and metadata handling throughout the organization.
* Implemented data governance and compliance strategies, offering compliance with regulatory standards like GDPR and CCP.

**Environment**: Python, Apache Airflow, Lambda, Redshift Neo4j,S3 Buckets , Snowflake, AWS Redshift, EC2, PySpark, SQL, Apache Kafka, ML.

***Client: The Cigna Group*  Bloomfield, CT *|* AUG 2021-NOV 2022 *| Role: Data Engineer***

* Implemented Azure Data Factory to extract and load data sources into a streamlined workflow to improve efficiency, including CTMS, Rave, and data throughput.
* Implemented API Gateway to create, publish, maintain, monitor, and secure APIs at any scale.
* Utilized Complex SQL Queries and Led the deployment of data migration and processing workflows within Azure data factory pipelines, ensuring high performance in processing large healthcare datasets.
* Employed Python Programming and Scala for advanced data pipeline operations such as cleansing, aggregation, and transformation, leveraging the computational power of Azure Databricks for execution.
* Integrated Data Vault modeling to provide a scalable, flexible, and robust framework for evolving healthcare data needs.
* Configured Kafka Connect for seamless integration with various data sources, ensuring data consistency and integrity.
* Implement Azure Security Center to enforce robust data privacy and protection measures within Azure Data Lake and all data processes.
* Implemented data privacy and protection strategies. Ensured that all healthcare data was encrypted, access-controlled, and compliant with healthcare regulations like HIPAA and PII data from unauthorized access.
* Led the integration of healthcare data systems, including CTMS and Rave, aligning with EMR/EHR functionalities to streamline data workflows, enhance patient data management, and support clinical decision-making.
* Leveraged Informatica PowerCenter for complex data integration tasks, especially where deep transformations are required or when integrating with legacy systems not fully supported by ADF.
* Leveraged Pig for processing and analyzing large data sets, ensuring data is optimally prepared for analytics and machine learning models.
* Utilized PySpark within Azure Databricks for large-scale data processing, enabling real-time analytics and insights that support critical decision-making.
* Integrated Five Tran technology to automate and refine data integration processes, enhancing the accuracy and efficiency of reporting.
* Administered and optimized Azure SQL Databases, optimizing them for performance, security, and reliability to support healthcare applications and analytics platforms.
* Developed Predictive Models in Azure Databricks using Python, applying machine learning models to predict patient outcomes and operational efficiencies.
* Integrated Azure Cosmos DB to manage real-time data storage and querying, utilizing its global distribution, scalability, and low latency capabilities.
* Incorporated Power BI for dynamic data visualization and reporting, enabling actionable insights and decision-making through interactive dashboards tailored to healthcare analytics needs

**Environment**: Azure Data Factory, SQL, Python, Kafka, Neo4j, PostgreSQL, Azure Security Center, Informatica Power Center, Azure Databricks, Cosmos DB, Azure SQL, Power BI, PySpark, Hive, Pig, Fivetran.

***Client:* Kroger Atlanta, Georgia *|* JAN 2020 – MAY 2021*| Role: Data Engineer***

* Spearheaded the creation of sophisticated data processing workflows using Talend and Big Query, integrating Python for data transformations and SQL for optimization, ensuring high scalability and data integrity.
* Enhanced data retrieval speeds and reduced storage costs through SQL stored procedures, partitioning, and clustering, driving performance improvements.
* Utilized Talend to integrate multiple data sources, including Cloud SQL and Firestore, into a cohesive data ecosystem.
* Utilized BigQuery alongside Google Cloud services like Cloud Dataflow and Cloud Composer to create a robust, scalable data architecture that supports real-time data processing and analytics
* Conducted meticulous validation and testing of ETL processes and data models with Talend, ensuring the highest data quality standards.
* Advocated for and implemented DevOps practices, including the use of Git for source control and CI/CD pipelines for continuous integration and delivery.
* Managed comprehensive ETL pipeline orchestration using Google Cloud Composer, integrating Python with Cloud Dataflow for efficient data handling.
* Established a robust monitoring and logging system with Google Cloud Operations for proactive ETL operations management in Big Query.
* Developed Python-based automation for ETL pipelines, enhancing the efficiency of processing diverse data se
* Employed Google Cloud Bigtable for the scalable management of extensive data volumes, ensuring high performance and reliability.
* Developed custom functions and stored procedures in BigQuery to streamline complex data transformations and aggregations, significantly reducing processing time and improving data query performance.
* Leveraged Google Cloud Deployment Manager for efficient provisioning and management of cloud resources, ensuring consistent environments.
* Directed the deployment of scalable data applications on Google Kubernetes Engine, optimizing resource utilization with Python automation.
* Coordinated project delivery using Jira, applying agile methodologies for timely and efficient execution.
* Leveraged data visualization tools such as Tableau to create insightful dashboards and reports to make complex datasets accessible and actionable for stakeholders.

**Environment**: Python, SQL, Talend, Google Cloud Composer, Cosmos DB, cloud data flow, Big Query, Cloud SQL, Firestore, Google Cloud Bigtable, Google Cloud Deployment Manager, Git, JIRA, Tableau.

***Client: TechMojo Solutions, India* |JUNE 2017 – JULY 2019*| Role: Data Engineer***

* Spearheaded implementing and managing Hadoop clusters, leveraging HDFS for scalable data storage and YARN for efficient resource management.
* Oversaw and updated the Hadoop Distributed File System (HDFS), ensuring reliable and scalable storage solutions for large datasets.
* Utilized Sqoop for efficient data transfers between Hadoop clusters and relational databases, maintaining data consistency and integrity.
* Leveraged Python libraries such as Pandas and NumPy for data cleaning and preprocessing, ensuring the high quality of data for analytics and reporting.
* Designed and executed ETL processes using Python and AWS Glue, transforming structured and unstructured data into a centralized data warehouse.
* Integrated SQL Server Integration Services (SSIS) for comprehensive ETL capabilities, facilitating data extraction, transformation, and loading across different databases and file formats.
* Employed MapReduce to analyze, filter, and aggregate data, deriving more profound insights into market trends, property valuations, and customer behavior.
* Utilized Apache Hive to manage structured real estate data, facilitating real-time analytics and reporting needs with HiveQL.
* Implemented Scala to interact with NoSQL databases like MongoDB and enhanced structured data management with Apache Hive.
* Leveraged AWS services, including AWS Lambda for event-driven data processing and Amazon S3 for scalable storage solutions.
* Expertly managed and optimized relational databases using PL/SQL, improving critical data storage and retrieval processes.
* Developed and maintained Tableau dashboards for visual data presentation, assisting stakeholders in making informed decisions.

**Environment**: HDFS, YARN, HIVE, AWS Glue, SSIS, AWS Lambda, Amazon S3, Sqoop, MongoDB, XML, PL/SQL. Tableau.

 ***Client: Netenrich Technologies Inc, India* */ August 2015 -May 2017 / ROLE: Data Analyst***

* Developed and optimized PL/SQL scripts for managing and querying Extensive datasets, significantly enhancing data retrieval processes in a software development Company.
* Partnered with cross-functional teams using Erwin Data Modeler to create and refine effective data models to support software development initiatives, enhancing data structure and utility. .
* Developed Python scripts for thorough data cleansing, transformation, and enrichment, improving data quality and facilitating more accurate data analysis.
* Utilized Apache NiFi to process and analyze data in JSON format to extract vital information for software development initiatives, ensuring seamless integration.
* Provided expertise in data modeling best practices for efficient and scalable data management in software development initiatives.
* Implemented and leveraged big data technologies such as Hadoop for processing large-scale datasets, enabling more efficient data analysis and insights extraction.
* Designed and deployed interactive Tableau dashboards and reports, translating complex data into accessible insight.
* Documented data processes and models thoroughly, sharing knowledge and best practices within the team to foster a collaborative and informed working environment.
* Managed and optimized on-premises cloud infrastructures, primarily Linux-based, to support robust data storage, processing, and analysis platforms.

**Environment**: PL/SQL, Python, Apache NiFi, Erwin Data Modeler, Tableau, Linux

Top of Form