**Nayeema S**

[**nayeema1608@gmail.com**](mailto:nayeema1608@gmail.com)

**+1 (775) 387-4047‬**

**PROFESSIONAL SUMMARY:**

* Experienced **Data Engineer with over 8+ years** of strong technical expertise, business acumen, and exceptional communication skills. Proven ability to drive high-impact business outcomes through data-driven innovations and decisions.
* Proficient in all aspects of the Software Development Lifecycle (**SDLC**), including requirement analysis, design, development, testing, implementation, and maintenance.
* Extensive experience in web/application development and analytical programming using **Python, Azure, GCP, Java,** and various databases such as **Sybase, Microsoft SQL Server, Oracle,** and **Teradata**.
* Demonstrated expertise in transforming business requirements into analytical models, designing algorithms, building models, and developing **data mining** and reporting solutions that scale across large volumes of **structured and unstructured data**.
* Skilled in data manipulation using **Python**, including loading, extraction, and analysis, leveraging libraries such as **NumPy, SciPy, Matplotlib, Scikit-Learn, and Pandas.**
* Proficient in Python programming for developing **APIs, web services, machine learning models, and data engineering tasks.**
* Proficient in architecting and deploying data solutions on **Amazon** **Web** **Services** (**AWS**), utilizing services like **Amazon** **S3**, **Amazon** **Redshift, AWS lambda**, and **AWS** **Glue**.
* Designing and implementing **Data** **Lakes** using technologies like **Apache** **Hadoop**, **Amazon** **S3**, or **Azure** **Data** **Lake** **Storage**, to store and manage vast amounts of structured and unstructured data.
* Strong development skills in **Azure Data Lake, Azure Data Factory, SQL Data Warehouse, Azure Blob, and Azure Storage Explorer.**
* Experienced in architecting and implementing **ETL** and data movement solutions using **Azure Data Factory (ADF) and SSIS.**
* Seasoned in applying **machine learning** algorithms and predictive modelling techniques, including Linear Regression, Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, KNN, Neural Networks, and K-means Clustering.
* Extensive experience in building **ETL** pipelines on **AWS** using **AWS** **py**, automating data extraction, transformation, and loading processes for optimal data preparation.
* Good understanding of **Google Cloud** design considerations, limitations, and pricing implications.
* Extensive experience in the **Big Data Hadoop ecosystem**, including **data ingestion, storage, querying**, processing, and analysis using components such as **MapReduce, YARN, Hive, HBase, Flume, Sqoop, Spark MLlib, Spark GraphX, Spark SQL, and Kafka.**
* Skilled in enhancing algorithm efficiency using **Spark** and its components, including **Spark Context, Spark SQL, Spark MLlib, DataFrames, Pair RDDs, and Spark YARN.**
* Extensive experience in developing **ETL** (**Extract**, **Transform**, **Load**) pipelines to ingest and process data from various sources into the **Data** **Lake**, ensuring **data** **quality** and integrity.
* Proficient in **ETL** processes, **data integration**, and **data processing using Apache Flume, Kafka, PowerBI, and Microsoft SSIS.**
* Experienced in providing technical consulting and end-to-end delivery, including architecture, **data modelling, data governance, and solution implementation.**
* Skilled in working with various data sources, including **Oracle SE2, SQL Server, flat files, and unstructured files, for data warehousing purposes.**
* Hands-on experience with **Google Cloud Platform (GCP)** and its big data products, such as **BigQuery, Cloud Dataproc, Google Cloud Storage, and Composer (Airflow as a service)**.
* Proficient in **GCP** services, including **BigQuery**, Cloud Storage (**GCS**), **Cloud Functions, Cloud Dataflow, Pub/Sub, Cloud Shell, GSUTIL, Data Proc, and Operations Suite (Stackdriver).**
* Skilled in **data mining** solutions and generating data visualizations using **Python, R, and Tableau.**
* Strong expertise in designing and developing relational database concepts, with proficiency in multiple **RDBMS databases, including Oracle 10g, MySQL, and MS SQL Server**.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Hadoop/Big Data Technologies** | HDFS, Apache NIFI, Map Reduce, Sqoop, Flume, Pig, Hive, Oozie, Impala, Zookeeper, Ambari, Storm, Spark and Kafka |
| **Programming and Scripting** | Scala, Java, SQL, JavaScript, Shell Scripting, Python, R, Pig Latin, HiveQL |
| **Cloud Platform** | Azure, Google Cloud(GCP) ,AWS , |
| **Python Libraries** | NumPy, SciPy, Matplotlib, Scikit-Learn, Pandas |
| **Visualization Tools** | PowerBI, Tableau |
| **IDEs** | Eclipse, IntelliJ, PyCharm, Notepad ++, Visual Studio |
| **Other Tools** | Airflow, Linux, Git, REST APIs (FastAPI), Spark/PySpark |
| **Analytics Tools** | Microsoft SSIS, SSAS and SSRS |
| **Build and Deployment Tools** | Maven, Ant, Git, SVN, Jenkins |
| **Databases** | Oracle10g, MySQL, MS SQL Server, Cassandra, MongoDB, Teradata |

**PROFESSIONAL EXPERIENCE:**

**Client: Sam’s club, Bentonville, AR Apr 2021 to Present**

**Data Engineer**

* Led the construction and architectural design of multiple data pipelines, end-to-end **ETL** and **ELT** processes for data ingestion and transformation in the **Google Cloud Platform (GCP**), while effectively coordinating tasks among team members.
* Translated business problems into scalable **Big Data** solutions, defining comprehensive strategies and roadmaps for **Big** **Data** initiatives.
* Utilized the **GCP Cloud Shell SDK** to configure and manage services such as **Data Proc, Cloud Storage, and BigQuery**.
* Implemented and optimized **Databricks**-based **ETL** workflows to leverage **Azure** resources efficiently, resulting in improved performance and cost savings.
* Developed a **Kafka** producer **API** to facilitate the streaming of live data into various **Kafka** topics.
* Monitored **BigQuery, Dataproc**, and **Cloud Dataflow** jobs across different environments using **Stackdriver**.
* Utilized **Spark/Scala** and **Python** in the **Hadoop/Hive** environment, implementing regular expression (regex) projects, and working with Linux/Windows for large-scale data processing.
* Imported metadata into **Hive data warehouse** using Scala and migrated existing tables and applications to operate effectively on Hive and the **Azure cloud**.
* Successfully migrated code bases to Amazon **EMR** and evaluated various components of the Amazon ecosystem, including **Redshift**.
* Imported legacy data from **SQL** Server and Teradata into **Amazon S3**.
* Integrated **iSpark** with cloud-based storage solutions, such as **Amazon S3** and Azure Data Lake Storage, to efficiently store and retrieve large volumes of data in distributed environments.
* Designed and implemented pipelines in **Azure Data Factory** (**ADF**) using Linked Services, Datasets, and Pipelines to extract, transform, and load data from diverse sources such as **Azure SQL, Blob storage**, and **Azure SQL Data Warehouse**.
* Managed **Azure Data Lakes (ADLs)** and Data Lake Analytics, showcasing a comprehensive understanding of integration with other **Azure services**.
* Implemented automated data quality checks and validations within the **CI/CD** pipeline to ensure that incoming data is accurate, complete, and compliant with established data quality standards.
* Developed **Python**-based **RESTful web services**/**APIs** to track revenue and perform revenue analysis.
* Conducted data pre-processing and feature engineering using **Python** Pandas for subsequent predictive analytics.
* Created custom operators and hooks in **Python** to extend **Airflow's functionality**, addressing specific integration requirements with various data sources and destinations.
* Collaborated with cross-functional teams to design and implement scalable data pipelines using **iSpark**, enabling faster data processing and analysis.
* Prototyped customer **data analysis** and joining using **Spark** in **Scala**, processing it in **HDFS**.
* Developed database solutions using **Azure SQL Data Warehouse** and Azure SQL.
* Implemented **iSpark** optimizations to enhance **Spark job performance**, reducing processing times and resource utilization for data processing tasks.
* Led the migration initiative from **AWS** to **Azure** for a large-scale data engineering project, ensuring a seamless transition of **Databricks** workloads and **Unity Catalog** and **Deltalog** data.
* Designed and implemented migration strategies for traditional systems on **Azure**, employing techniques such as Lift and Shift and **Azure Migrate**, alongside other third-party tools.
* Designed and implemented end-to-end data solutions encompassing storage, integration, processing, and visualization in **Azure**.
* Successfully designed and optimized **data** **warehousing** solutions on **AWS**, leveraging Amazon **Redshift's** columnar storage for efficient data querying and analysis.
* Proposed architectures considering cost optimization in Azure, providing recommendations for right-sizing **data infrastructure**.
* Implemented **Sqoop** for incremental jobs, facilitating the extraction of data from **DB2** and loading it into **Hive tables**, further connecting it to **Tableau** for generating interactive reports via **Hive Server2**.
* Contributed to the entire project lifecycle, from defining the reference data approach to master data management (**MDM**), to creating a comprehensive data dictionary and mapping from sources to the **MDM data model.**
* Led cross-functional training sessions to educate colleagues on **AWS** and **GCP** data engineering best practices, fostering knowledge sharing.
* Designed, developed, and tested dimensional **data models** using **Star and Snowflake schema methodologies** followingthe **Kimball method.**
* Extracted, transformed, and loaded data sources, generating **CSV** data files using **Python** programming and **SQL** **queries**.
* Developed data processing applications in **Scala** using **Spark** **RDDs** and **SparkSQL** **APIs** for efficient data manipulation and analysis.
* Developed and maintained infrastructure code using **Terraform** for **enterprise-level** projects.
* Created **Tableau** reports with complex calculations and conducted ad-hoc reporting using **PowerBI**.
* Followed the **Scrum** methodology and worked in a continuous integration and continuous deployment (**CI/CD**) environment utilizing Jenkins.
* Provided day-to-day **GIT** support for various projects, overseeing the design and maintenance of **GIT** repositories and access control strategies.

**Environment:** Python, Spark, Scala, Redshift, Azure, GCP, BigQuery, Stackdriver, SparkSQL, SparkRDD, Databricks, Teradata, Kafka, Azure SQL, Ispark , Hadoop, Hive, Scala, HDFS, Sqoop, Shell Scripting, Power BI, Tableau, Jenkins, GIT, Azure SQL Data Warehouse

**Client: Molina healthcare, Bothell, WA Feb 2019 to Mar 2021**

**Data Engineer**

* Developed **Python** scripts to efficiently parse **XML** documents and load the extracted data into **databases**.
* Generated comprehensive reports on predictive analytics utilizing **Python** and **Tableau**, including visualizations of model performance and prediction results.
* Implemented **iSpark** optimizations to enhance the fault tolerance and resilience of data processing tasks, reducing the impact of node failures on job completion.
* Utilized **iSpark's** machine learning libraries to implement predictive analytics and recommendations, adding advanced analytical capabilities to data engineering workflows.
* Utilized **Airflow's** task dependencies and scheduling features to manage the execution order of tasks, ensuring proper sequencing and minimizing data processing bottlenecks.
* Demonstrated ability to integrate **AWS** Lambda functions for real-time data processing within **AWS** data pipelines, enhancing data insights and decision-making.
* Utilized **GCP's** **BigQuery**, **Dataflow**, and Pub/Sub to create real-time data pipelines, enabling faster decision-making for business insights.
* Worked with **Terraform Enterprise features** like remote state management and collaboration.
* Skilled in implementing **data** **governance** and compliance measures on **AWS**, ensuring data security and adhering to industry standards and regulations.
* Played a key role in data migration projects, leveraging **SQL**, **SQL** **Azure**, **Azure** **Storage**, **Azure** **Data** **Factory**, **SSIS**, and **PowerShell**. Successfully migrated data from on-premise sources (**Oracle**, **SQL** **Server**, **DB2**, **MongoDB**) to **Azure** **Data** **Lake** and **Azure** **Data** **Store** using **Azure** **Data** **Factory**.
* Implemented medium to large-scale business intelligence (**BI**) solutions on **Azure**, utilizing various **Azure** **Data** **Platform** services such as **Azure** **Data** **Lake**, **Data** **Factory**, **Data** **Lake** **Analytics**, **Stream** **Analytics**, **Azure** **SQL** **Data** **Warehouse**, **HDInsight**, and **Databricks**, as well as **NoSQL** **databases**.
* Developed ad-hoc analysis solutions using **Azure** **Data** **Lake** **Analytics**/Store and **HDInsight**, enabling efficient processing of both structured and unstructured data from multiple sources.
* Created and published multiple interactive dashboards and reports using **Tableau** **Server**, incorporating advanced analytics techniques such as text analytics, Naive Bayes, sentiment analysis, word clouds, and data retrieval from social networking platforms, including Twitter.
* Automated data cleaning processes using **Python** **scripts** for datasets consisting of a combination of unstructured and structured data from diverse sources.
* Proficient in implementing data security and access controls on **GCP**, ensuring data privacy and compliance with industry regulations like **GDPR.**
* Integrated **Airflow** with **GIT** to manage and track changes to DAG configurations, promoting collaboration and reproducibility.
* Proficient in using **AWS** Step Functions for orchestrating and coordinating complex **data** **workflows**, ensuring seamless data processing and transformation.
* Improved fraud prediction performance through feature selection using random forest and gradient boosting techniques within the **Python** Scikit-learn framework.
* Collaborated with multidisciplinary teams to deploy data solutions on **AWS**, providing technical expertise in data engineering and cloud architecture.
* Implemented near-real-time **data** **pipelines** using a framework based on **Kafka** and **Spark**, enabling efficient **data** **processing** and analysis.
* Developed **Spark** programs using Scala to assess and compare performance between **Spark** and **Hive**, as well as **SparkSQL**.
* Collaborated with cross-functional teams to deploy and manage data solutions on **GCP**, providing end-to-end data engineering expertise for various business needs.
* Developed **PySpark** scripts to encrypt raw data using hashing algorithms based on client-specified columns.
* Ingested data from relational **databases**, performed necessary transformations, and exported the transformed data to **Cassandra** in alignment with specific business requirements.
* Implemented the **ELK** (**Elasticsearch**, **Logstash**, **Kibana**) stack to collect and analyze logs produced by the **Spark** **cluster**, facilitating effective monitoring and troubleshooting.
* Successfully executed **data** **migration** projects to **AWS**, optimizing data storage, accessibility, and scalability while minimizing downtime and disruptions.
* Migrated existing computational code from **SQL** to **PySpark**, improving efficiency and scalability.
* Created action filters, parameters, and calculated sets to prepare interactive dashboards and worksheets using **PowerBI**.
* Integrated version control systems (e.g., Git) with **CI/CD** tools to ensure data pipeline code and configurations are tracked, tested, and deployed consistently.

**Environment:** Python, Scikit-Learn, Machine Learning, Azure, Hadoop, Hive, MapReduce, Sqoop, Kafka, Spark, Yarn, Pig, PySpark, Cassandra, Oozie, Shell Scripting, Ispark , HBase, Scala, GCP, Maven, Java, JUnit, SQL, Hortonworks, Tableau, Teradata, MySQL, Azure Cloud, Azure SQL, Azure Storage, and Azure Data Factory, Azure Data Lake

**Client: Silicon Valley Bank, CA Dec 2017 to Jan 2019**

**Data Engineer**

* Developed **Spark** applications using **Spark** **SQL** in **Databricks** to efficiently extract, transform, and aggregate data from multiple file formats. Leveraged the power of **Spark** to analyse and transform data, revealing valuable insights into customer usage patterns.
* Played a key role in project planning and data conversion migration, ensuring a smooth transition from the legacy system to the target **Snowflake** database.
* Created robust applications utilizing **Kafka** to monitor consumer lag within Apache **Kafka** clusters, enabling efficient **data** **ingestion** in real time.
* Assumed responsibility for importing real-time data from various sources to **Kafka** **clusters**, facilitating continuous **data** **flow** and processing.
* Utilized **Spark** Streaming **APIs** to perform necessary transformations and actions on the data obtained from **Kafka**, ensuring **data** **integrity** and reliability.
* Conducted in-depth analysis of Hadoop clusters and various **big** **data** **analytic** tools, including Pig and **Hive**, to derive meaningful insights from large-scale datasets.
* Migrated **MapReduce** jobs to **Spark** jobs, leveraging **SparkSQL** and **DataFrames** **API** to seamlessly load structured data into **Spark** **clusters**.
* Utilized **PowerBI** for direct querying, enabling comprehensive comparison between legacy and current data. Generated insightful reports, stored data, and created visually appealing dashboards.
* Played a pivotal role in building multiple data pipelines and end-to-end **ETL**/**ELT** processes for efficient data ingestion and transformation in the **Google** **Cloud** **Platform** (**GCP**). Effectively coordinated tasks among team members to ensure project success.
* Successfully migrated on-premises **Hadoop** systems to the **Google** **Cloud** **Platform** (**GCP**), leveraging its robust infrastructure and services.
* Ensured infrastructure reproducibility and consistency through **Terraform** best practices.
* Transferred previously written cron jobs to Airflow/Composer in **GCP**, enhancing workflow management and task scheduling capabilities.
* Provided ongoing support for existing **GCP** **Data** **Management** implementations, ensuring smooth operations and system optimization.
* Created authorized views in **GCP** **BigQuery**, enabling row-level security and controlled access to data for various teams.
* Designed and developed **RESTful APIs** for multiple projects, adhering to best practices.
* Implemented **ETL** processes, including Slowly Changing Transformation, to maintain historical **data** **integrity** within the **data** **warehouse**.
* Conducted thorough **ETL** testing activities, including job execution, data extraction, transformation, and upload into **data** **warehouse** servers.
* Designed and developed **SSIS** packages to extract, transfer, and load (**ETL**) data from various environments into **SQL** Server for **SSAS** cubes (**OLAP**) and **SQL** Server Reporting Services (**SSRS**).
* Created and formatted diverse report types, including Crosstab, Conditional, Drill-down, Top N, Summary, Form, **OLAP**, Sub reports, ad-hoc reports, parameterized reports, interactive reports, and custom reports.
* Leveraged **SQL** joins among Hive tables in **Hive Data Warehouse** to generate input for **Spark** batch processes, ensuring efficient **data** **processing** and **analysis**.
* Collaborated closely with the data science team to build statistical models using **Spark** **MLLib** and **PySpark**, harnessing the power of advanced machine learning techniques.
* Utilized version control tools like **GitHub** to facilitate seamless code sharing and collaboration among team members.

**Environment:** Python, SSIS, PowerShell, SparkSQL, Kafka, Spark, MapReduce, Power BI, Tableau, SSRS, PySpark, Spark MLlib, SQL, GCP

**Client: Yana Software Private Limited Hyderabad, India Jun 2016 to Aug 2017**

**Data Engineer (Azure)**

* Developed comprehensive mapping documents to establish clear column-level mappings from source to target systems, ensuring accurate data transformation and integration.
* Created efficient **Azure Data Factory (ADF)** pipelines utilizing **Azure Polybase** and **Azure Blob storage**, facilitating seamless data movement and transformation.
* Leveraged **Azure Databricks for ETL** processes, successfully migrating on-premises **Oracle ETL** workflows to **Azure Synapse Analytics** for enhanced scalability and performance.
* Automated script generation using **Python scripting**, streamlining development processes, and improving efficiency. Implemented data curation techniques within **Azure Databricks** to ensure data quality and consistency.
* Leveraged advanced features of **Azure Data Factory**, such as stored procedures, lookup, execute pipeline, data flow, copy data, and Azure functions, to orchestrate complex data workflows and transformations.
* Optimized performance of **Informatica** mappings and sessions, identifying and eliminating bottlenecks to improve overall **ETL process** efficiency.
* Developed and optimized complex **SQL queries** and **PL/SQL procedures**, seamlessly converting them into efficient **ETL tasks**.
* Worked with **PowerShell** and **UNIX** scripts for file transfer, email automation, and various file-related tasks, ensuring smooth data operations.
* Demonstrated expertise in utilizing **Azure Data Lake Storage Gen2** to store diverse file formats, such as Excel and Parquet files, and efficiently retrieve data using **Blob APIs**.
* Utilized Azure Databricks for **ETL** processes, leveraging workspace, **Hive**, **RDDs**, and mapping and reducing functions to perform data transformation and processing tasks.
* Designed and implemented **star schema** models for efficient data drilling and analysis. Developed **PySpark** procedures, functions, and packages to load and manipulate data.
* Configured **Azure Traffic Manager** to establish routing for user traffic, optimizing application performance and availability.
* Prepared comprehensive capacity and architecture plans for creating **Azure cloud environments**, ensuring seamless hosting of migrated **IaaS VMs** and **PaaS role** instances.
* Implemented highly available (HA) deployment models using **Azure Classic** and **Azure Resource Manager**, ensuring high availability and fault tolerance for critical systems.
* Configured **Azure Active Directory** and effectively managed users and groups, ensuring secure and streamlined **access management.**
* Configured continuous integration from source control, setting up build definitions within **Visual Studio Team Services (VSTS)**, and enabling continuous delivery for automated deployment of **ASP.NET MVC** applications to Azure web apps.
* Deployed **Desired State Configuration (DSC)** from **Azure Automation** to both on-premises and cloud environments, ensuring consistent configuration management across the infrastructure.

**Environment:** SQL Server, Azure, Pyspark, ETL, PowerShell,DataBricks, UNIX, ASP.NET MVC, Visual Studio, T-SQL

**Client: Careator Technologies Pvt Ltd Hyderabad**  **Apr 2015 - May 2016**

**Data Engineer**

* Successfully built a robust reporting data warehouse by extracting and integrating data from **ERP system modules**, including **Order Management**, Invoice, and Service Contracts.
* Acted as the **subject matter expert (SME**) for **Data Warehouse**-related processes, providing valuable insights and guidance to the team.
* Conducted thorough data analysis to inform the development of the **Reporting Data Mart**, ensuring accurate and meaningful reporting capabilities.
* Collaborated closely with Reporting developers to oversee the **implementation of report** and **universe designs**, ensuring alignment with business requirements.
* Managed deployments from the development environment to **user acceptance testing (UAT)**, and ultimately to production, ensuring smooth and seamless transitions.
* Designed and developed various metrics using **statistical tools** such as **3 sigma** controlcharts**, P-Charts,** and **Box plots** in **QlikView**. Implemented **dynamic subgrouping** techniques to analyze shifts in sensor data.
* Automated the deployment of **GCP resources** and services using **Terraform**.
* Utilized the Cloud Shell SDK in **Google Cloud Platform (GCP)** to configure essential services such as **Data Proc, Storage, and BigQuery**, contributing to successful GCP proof-of-concept **(POC)** initiatives.
* Played a key role in the migration of data and applications from on-premises environments to Google Cloud, showcasing expertise in **GCP technologies** and **IAM (Identity and Access Management)** roles.
* Applied the statistical data cleansing model to production jobs using **Alteryx**, automating the data cleansing process for improved efficiency and accuracy.
* Collaborated with various upstream and downstream systems and global customers to integrate systems, perform **data extractions**, implement **ETL** processes, and meet diverse analytics and reporting needs.
* Utilized **SQL Server Integration Services (SSIS)** to seamlessly integrate and analyze data from heterogeneous information sources, enabling comprehensive data insights.
* Developed customized reports using **SQL Server Reporting Services (SSRS)** to review scorecards and analyze business trends based on data from different business and market locations.
* Leveraged expertise in handling semi-structured sensor data, implementing **statistical methodologies** to assess **data quality** and **address outliers.**

**Environment:** SQL Server, SSIS, SSRS, Alteryx, QlikView, ERP, SQL Server, GCP