**LEKHYA J**

**SR. DATA ENGINEER**

**EMAIL: Lekhya.j1901@gmail.com CONTACT: +17378008233**

**LINKEDIN:** https://www.linkedin.com/in/lekhya-j/

**PROFESSIONAL SUMMARY:**

* Demonstrated 8+ years of professional work experience in design, development, and implementation of cloud, Big Data, Spark, Scala, Hadoop, and maintenance of **data pipelines** in the position of **Data Engineer**.
* Strong experience in developing **Data Modeling**, Data migration, Design, Data Warehousing, Data Ingestion, **Data Integration**, Data Consumption, Data delivery, and integration Reporting.
* Extensive experience in Text Analytics, developing different Statistical Machine Learning, Data Mining solutions to various business problems and generating data visualizations using R, Python.
* In-depth knowledge in OLAP, OLTP, Business Intelligence and Data Warehousing concepts with emphasis on ETL and Business Reporting needs.
* Experience in Extract, Load and Transformation (ETL) process using **Informatica** **PowerCenter**, Data Stage and **SSIS**. from Sequential files, XML files, CSV files and transformed and loaded into the target Data warehouse.
* Fluent in utilizing cloud-native technologies such as **AWS**, and **Azure** to build scalable and secure data platforms.
* Proficiency in utilizing **Spark** transformations and actions to clean, transform, and pre-process data, ensuring data quality and consistency for downstream analysis.
* Experience in Designing and implementing **Kafka** based data pipelines, enabling real-time ingestion of vast data volumes from diverse sources like databases, log files, and IoT devices.
* Experience in Microsoft Azure Cloud Services SQL Data Warehouse, Azure SQL Server, **Azure** **Databricks**, **Azure Data Lake**, **Azure Blob Storage, Azure Data Factory, and Azure DevOps.**
* Good in Managing **Snowflake** accounts, warehouses, and user roles, ensuring proper resource allocation, monitoring, and maintenance of Snowflake environments.
* Experience in implementing data replication and **backup** **strategies** in Snowflake to ensure data availability and **disaster recovery** preparedness.
* Actively collaborated with data architect/engineers to establish **data governance** for MDM/security (key vault, network security schema level & row level), resource groups, integration runtime setting, integration patterns, aggregated functions for data bricks development.
* Extensive experience in storing and documentation using **NoSQL** Databases: Mongo DB, Snowflake, and HBase.
* Hands on experience in converting complex RDMS (Oracle, MySQL & Teradata) queries into Hive query language.
* Developed **Spark** applications for both batch and real-time data processing scenarios, accommodating various data velocity requirements.
* Followed **Data Governance** best practices by documenting data sources, metadata, and data lineage within Tableau to maintain data transparency and lineage tracking.
* Responsible for designing and building a **Data** **Lake** using Hadoop and its ecosystem components.
* Good knowledge in Hypothesis Testing, T-Test, Z Test, Gradient descent, Newton's Method, ANOVA test.
* Hands-on experience with **Amazon** **EC2, S3, Redshift, RDS, VPC, IAM**, **Amazon Elastic**, **Load Balancing, Auto Scaling,** and other services of the AWS family. Proficient in data mining tools like **R, SAS, Python, SQL Alchemy.**
* Experience in integrating **Talend** to transform and enrich data as it moves through the **ETL** pipeline, ensuring data quality and compatibility with downstream systems.
* Hands-on experience in setting up a workflow using **Apache Airflow** and Oozie workflow engine for managing and scheduling Hadoop jobs.
* Transformed and retrieved the data by using **Pig, Hive, Spark,** and **Map Reduce**.
* Extensive exposure to containerization technologies, including Docker, Kubernetes, and Cassandra and proficient in container orchestration using **Kubernetes**.
* Deep understanding of **Docker, Kubernetes, Jenkins, Maven, Puppet, Ansible, Terraform,** and **Git.**
* Data importing and exporting by using **Sqoop**, **Apache flume** among HDFS, Relational Database Systems, HBase, web applications, and vice versa.
* Strong experience in working with **UNIX** environments and writing **Shell Scripts**. Experience in writing SQL queries and optimizing the **Teradata**, Oracle, and SQL Server queries.
* Profound knowledge about the architecture of the **Teradata** database and experience in Teradata unloading utilities like Fast export.
* Experience in managing and maintaining high-performance **Kafka** **clusters**, ensuring data availability, reliability, and low-latency ingestion for mission-critical applications.
* Extensive experience in working with **SQL**, databases, ETL, and cloud platforms to achieve business objectives.
* Experience with visualizing the data using BI and services amp tools: **Power** **BI**, **Tableau**, **Plotly**, **SSRS**, and **Matplotlib**.
* Proven ability to effectively engage with business users and stakeholders to acquire data needs and explain design decisions Applications.
* Maintained awareness of industry trends and developing technologies in data engineering and cloud computing, as well as aggressively explored chances for skill development.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| Cloud Services | AWS Glue, S3, RedShift, EC2, S3, EMR, Dynamo DB, Data Lake, AWS Lambda,  GCP, Azure Data Factory, Azure Data Lake Storage, Azure Synapse Analytics,  Azure Analytical Services, Azure SQL Datawarehouse |
| Big Data Technologies | Hadoop, Spark, Kafka |
| Visualization Tools | Power BI, Tableau, Excel, Pivot Table, VLOOKUP, SSRS |
| ETL/ Datawarehouse Tools | Informatica PowerCenter (Mapping Designer, Source Analyzer, Warehouse  Designer, Mapplet Designer, Transformation, Developer, Workflow Manager,  Workflow Monitor) and SSIS. |
| Database Management | Backup and Recovery, Snowflake, SQL DB’s |
| Version Control &  Containerization tools | GIT, Bitbucket, Docker, and Jenkins |
| Real-Time Data Processing | Kafka, Apache NiFi |
| Programming Languages | SQL, NoSQL, PL/SQL, Python, Java, UNIX, PySpark, Pig, HiveQL, Scala,  R, Shell script |
| Databases | MySQL, DB2, Oracle, PostgreSQL, MongoDB, Cassandra |
| Methodologies | Agile, Waterfall |

**PROFESSIONAL EXPERIENCE:**

**Client: AgFirst, Columbia, SC Aug 2020 – Present Role: Sr Data Engineer Responsibilities:**

* Analysed employee **insurance** claims data to identify trends, patterns, and anomalies, providing insights into claim approval rates, common medical procedures, and cost distributions.
* Designed, developed, and maintained **AWS**-based data solutions that are **scalable, reliable**, and **cost-effective.**
* Developed and optimized **T-SQL** queries for data extraction, transformation, and analysis, ensuring efficient and high-performing data retrieval.
* Worked with **AWS** services such as **S3**, **EC2**, **RDS**, and **Redshift** to optimize **Informatica** performance.
* Written various Spark programs in Java for data extraction, transformation, & aggregation from multiple file formats, including **XML**, **JSON**, **CSV**, & other compressed file formats, & stored the refined data in partitioned tables in **EDW**.
* Utilized **Snowflake** Cloud Data Platform and **AWS** for designing, implementing, and managing scalable and secure data solutions.
* Integrated **AWS services** with on-premises data sources and third-party applications, creating seamless **data flows** between **cloud** and on-premises environments using **AWS DataSync** and **AWS Direct Connect**.
* Developed and optimized **Snow** **SQL** queries and procedures, as well as user-defined functions (**UDFs**) using **JavaScript**, **Python**, and **SQL** to ensure efficient data processing in **Snowflake**.
* Optimized the **PySpark** jobs to run on **Kubernetes** Cluster for faster data processing. Built **ETL/ELT** **pipeline** in data technologies, **PySpark**, **Hive**, **Presto,** and **Databricks**.
* Set up the **CI/CD** pipelines using **Maven**, **GitHub**, and **AWS**. Worked extensively with importing **metadata** into Hive using **Python** and migrated existing tables and applications to work on **AWS** **cloud** (**S3**).
* Successfully implemented **ETL** solutions between an **OLTP** and **OLAP** database in support of Decision Support Systems with expertise in all phases of **SDLC.**
* Developed and maintained **APIs** for accessing and manipulating data using tools such as **Fast API** or **Django REST Framework**.
* Executed **data migration projects**, including schema changes and **data transfers**, using **SQL scripts** and tools to ensure smooth transitions between **database versions** or **platforms**. Developing and maintaining data processing solutions that leverage server less architectures using tool such as **AWS Lambda**.
* Utilized data analysis techniques to identify potential instances of **insurance** fraud within employee claims, conducting anomaly detection and pattern recognition to flag suspicious activities.
* Developed a **Data** **pipeline** using **Spark**, **Hive**, **Impala**, and **HBase** to analyse customer behavioural data and financial histories in the **Hadoop** cluster.
* Created scripts to append data from temporary **HBase** tables to target **HBase** tables in **Spark** and written **Spark** **programs** in **Scala** and ran **Spark** jobs on **YARN**.
* Worked on **NoSQL** Databases such as **HBase**, and **SPARK** for real-time streaming of data into the cluster.
* Performed the **DB2** bind process for various DB2 database groups across LPARS across all the applications.
* Performed the core **DB2** **administration** activities (performed image copies of various databases, loaded /unloaded the data to/from database, created / rebuilt the indexes, validated the data, repaired the table spaces, archived the DB2 active logs and restored DB2 logs, performed the lock escalations and concurrency states and other related activities. Performed the Analysis of DB2 batch job failures.)
* Processed different kinds of data, including unstructured (logs, clickstreams, Shares, likes, topics, etc.), semi-structured (**XML, JSON**), and structured like **RDBMS.**
* Developed **Oozie Workflows** for daily incremental loads, which get data from **Teradata** and then imported into hive tables. Involved in data ingestion of log files from various servers using **NIFI**.
* Worked with Teradata utilities such as M-load, F-load, export etc. to load or extract tables.
* Used **Teradata** utilities fast load, multi load, tpump to load data.
* Wrote, tested, and implemented **Teradata** Fast load, Multiload and BTEQ scripts, DML and DDL.
* Performed complex data analysis using **T**-**SQL** to derive meaningful insights and generate reports that support decision-making processes within the organization.
* Conducted performance tuning tasks by optimizing **T**-**SQL** code, reviewing execution plans, and implementing appropriate indexing strategies to enhance query performance.
* Performed Data Analysis on the analytical data present in AWS S3, AWS Redshift, Snowflake, Teradata using SQL, Python, Spark, Databricks.
* Deployed the development application from **Docker** containers to the production Cloud Foundry environment.
* Created **Devops** procedures using Cloud Formation and **Puppet** for consistent and dependable deployment technique.
* Used **ANT** and **MAVEN** as a build tool on Java projects to create build **artifacts** from source code.
* Applied **Informatica** tools within Data Governance Frameworks, actively contributing to best practices and compliance initiatives in alignment with organizational goals.
* Used **Tableau** Desktop to analyse and obtain insights into large data sets using groups, bins, hierarchies, sorts, sets, and filters.
* Worked in **Agile** development environment and actively involved in daily **Scrum** and other design-related meetings.
* Provided strong leadership to a team of data engineering resources, fostering collaboration, setting clear goals, and ensuring effective communication within the team.
* Collaborated with **stakeholders** to identify opportunities for **data**-**driven** insights and developing solutions to address those opportunities.
* Contributed to database design and modelling activities using **T**-**SQL**, ensuring effective structures that align with data analysis requirements and performance considerations.
* Experienced in data architecture best practices, integration, and **data** **governance** solutions (Data Catalog, Data Governance frameworks, Metadata and Data Quality).

**Environment**: Git, SQL, Snowflake, Hive, Python, ETL, AWS Lambda, AWS Glacier, Apache Flink, Databricks, Confluent Tableau, Hadoop, Spark, Apache Airflow, AWS Glue, Apache Kafka, Apache NiFi, HBase, Hive, Amazon Redshift, EC2, S3, EMR, HDFS, Scala, HBase, YARN, NoSQL, PySpark, Hadoop, Apache Spark, RDBMS. MySQL, Kubernetes, Sqoop, Spark-SQL, Oracle.

**Client: Ascena Retail Group, Mahwah, NJ Nov 2018 – Jul 2020**

**Role: Data Engineer**

**Responsibilities:**

* Designed and implemented **distributed systems** using **Azure cloud technologies**. Developed **monitoring** and **alerting** systems to detect and resolve issues with data platforms and pipelines.
* Built and maintained efficient **ETL** processes for ingesting data from external sources into **Azure** Systems using a combination of **Azure Data Factory**, **T-SQL,** and **U-SQL** Azure Data Lake Analytics.
* Created and maintained **Azure SQL** databases and **Azure Data Lake** Storage accounts, ensuring data is organized, optimized, and secured. Worked on Azure Data Lake Storage for efficient data storage and retrieval.
* Designed and implemented data pipelines using **Azure** **Data** **Factory** to move and transform data from various sources to **Azure** storage solutions.
* Conducted **code reviews** and **unit tests** to ensure quality and reliability of data processing applications and pipelines.
* Utilized Azure DevOps for version control, continuous integration, and continuous deployment to streamline the development and deployment of data solutions.
* Ensured **data** **security** and compliance by implementing encryption, **authentication**, and **authorization** mechanisms within **Kafka** clusters. Collaborated with security teams to conduct regular audits and vulnerability assessments, ensuring **Kafka** clusters meet industry-specific regulatory standards.
* Implemented **Kafka** monitoring solutions, such as confluent control centre, custom monitoring scripts to proactively identify and resolve issues.
* Written **Spark** applications using **Scala** on **AZURE** **data** **bricks** to interact with the **MySQL** database using **SparkSQL** context and accessed **Hive** **tables** using **Hive** **Context**.
* Involved in **POC** to check the efficiency of **Spark** application on **Mesos** **cluster** and **Hadoop** **Yarn** **cluster**. Developed **Spark** scripts and created **data** **pipeline** with **data** **frames** to migrate data to **Hadoop** **platform**. Designed and implemented **data security** measures to protect sensitive data and prevent unauthorized access.
* Designed, implemented, and maintained data architectures to support **IT** and **software** **applications**. Developed & maintained data pipelines to move & transform **data** **across** **systems** and platforms.
* Utilized **Snowflake's** data sharing capabilities to share data with external organizations and partners, fostering collaboration and data**-driven decision-making**.
* Created logical and physical data models from scratch, according to naming conventions and metadata frameworks.
* Developed dimensional and star schema models to aid in analytics and reporting.
* Designed and implemented **data warehouses** to store and manage large volumes of **structured** and **unstructured** **data**. Developed **data models** and **schema designs** to support business requirements and data analysis.
* Optimized **Snowflake** for performance by leveraging features like automatic **clustering**, **materialized** **views**, and **query** **optimization**, **resulting** **in** **efficient** **data** **processing**.
* Provided training and support to business users, ensuring they could navigate and utilize **Tableau** dashboards and reports effectively to **derive** **insights**. Developed **data quality** checks and **data validation** processes to ensure accuracy, completeness, and consistency of data.
* Developed **data** **models** and schema designs that enabled efficient querying and analysis of data in **Azure**. Created an **Azure SQL** database, as well as monitored & restored it.
* Created Pipelines in **ADF** using Linked Services/Datasets/Pipeline/ to Extract, Transform, and load data from different sources like Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool and backwards.
* Proficiently used Azure DevOps Server & Azure DevOps Service for version control, automation, & release management.
* Implemented and maintained **CI**/**CD** pipelines within **Azure DevOps**, automating build, testing, and deployment processes, resulting in improved efficiency, quality, and faster time-to-market.
* Deployed virtual machines (VMs) and Cloud services (PaaS role instances) on Azure Infrastructure as a Service (IaaS) while ensuring secure networking within designated Nets and subnets. Managed and optimized Continuous Integration (CI) tools, specifically **Azure DevOps.**
* Migrated a Microsoft SQL server to an Azure SQL database.
* Utilized **Python** for data cleaning, handling missing values, outlier detection, and feature engineering to prepare data for analysis and **modelling**.
* Worked with large datasets using **Hadoop** and **Kafka** to support data-intensive IT and software applications.
* Worked with database systems: **MySQL**, **Oracle**, **PostgreSQL** to design, implement and maintain databases.
* Proficiently interacted with databases using **Python** libraries like **SQLAlchemy** and **PyMySQL**, performing database operations, querying, and data retrieval.
* Leveraged **Kubernetes** to deploy and manage containerized applications, improving scalability and resource utilization.
* Effective in Database administration, providing operational stability and performance enhancement.
* Adapted rapidly to changing project requirements and communicated issues to stakeholders appropriately.
* Worked with **Python** or **Scala** **programming** **languages** to develop data processing applications and data analysis tools. Designed and implemented data **APIs** to provide access to data and analytics for software applications.
* Developed **data visualization** dashboards using **Tableau**, to support data analysis and decision-making. Provided technical guidance and support to software developers and **IT** teams on data-related issues.
* Developed automated reporting solutions in **Tableau** **Server** to deliver scheduled and on-demand reports to **stakeholders**, improving data accessibility and timely **decision**-**making**.
* Managed large-scale development and integration of project, assuring **data** **integrity** and consistency.
* Developed **data governance** policies and procedures to ensure compliance with data privacy regulations and company policies. Conducted quarterly Data owner meetings to communicate upcoming Data Governance initiatives, processes, policies, and best practices.
* Managed data integration processes within the **CDP** to consolidate customer data from various sources.
* Implemented and tested disaster recovery and business continuity plans for the **IDC** to minimize data loss and downtime in the event of hardware failures or natural disasters.

**Environment:** Data architectures**,** Data pipelines, ETL, Tableau, Snowflake, Hive, Python, Scala, Azure, MySQL, Oracle, PostgreSQL, Data security, Hadoop, Spark, Kafka, Data warehouses, Data lakes.

**Client: Micron, Hyderabad, India Mar 2017 – Jul 2018**

**Role: Data Analyst**

**Responsibilities:**

* Implemented complex business rules by creating robust **Mappings**, **Mapplets**, **Sessions** and **Workflows** using **Informatica** **Power Centre**. Mapped complex data structures successfully, permitting easy data integration procedures.
* Extensive experience in Extraction, Transformation, and Loading (**ETL**) data from various data sources into Data Warehouse and Data Marts using **Informatica** **PowerCenter** tools (Repository Manager, Designer, Workflow Manager, Workflow Monitor, and **Informatica** Administration Console).
* Proficiently used **SQL** to extract, transform, and load (**ETL**) data from various sources, including **relational** **databases**, data warehouses, and external **APIs**, ensuring data quality and consistency.
* Conducted performance tuning in **PL/SQL** by optimizing **SQL** queries, indexing, and query execution plans, resulting in enhanced database performance and faster data retrieval.
* Implemented version control system **Git** for data analysis projects, ensuring efficient tracking of changes, collaboration with team members, and maintaining a historical record of data transformations and analyses.
* Integrated **Python** applications with **RESTful** **APIs** to retrieve and update data from external sources and services, enabling real-time **data** **synchronization**.
* Integrated **SQL** **databases** with other data storage systems and tools, such as **data** **lakes** and **NoSQL** databases, to facilitate seamless data access and data flow within the **data** **ecosystem**.
* Conducted thorough data validation and testing processes post-migration to validate data accuracy, integrity, and consistency between **Informatica** & **AWS** environments, ensuring data quality & compliance with business requirements.
* Properly managed and maintained master data records with Informatica **MDM**. Converted source to target mappings, ensuring a clear comprehension of data transformation requirements.
* Designed and executed a comprehensive migration strategy to transition **Informatica**-based data integration processes and workloads to **AWS,** ensuring a seamless transition while minimizing downtime and mitigating potential risks.
* Designed stage jobs to pick the data files from **S3** to **redshift** staging layer.
* Designed transformation jobs from stage table to **Redshift** target tables as per the requirement for TYPE I and Type Il
* Worked on create Python Script to start and stop the EC2 instances in **AWS**.
* Worked with **Informatica ETL** tool to develop SCD I and SCD Il mapping on Cloud in **AWS** for POC to build source to Target mapping.
* Developed **Excel** macros and VBA (Visual Basic for Applications) scripts to automate repetitive tasks, enhancing efficiency in data processing and reporting workflows.
* Utilized **VLOOKUP** to perform data validation and cleansing tasks, identifying, and handling discrepancies, duplicates, and missing data, contributing to data quality enhancement.
* Leveraged **Python** **libraries** such as **Matplotlib**, **Seaborn**, and **Plotly** to create data visualizations and reports for data analysis and presentation to stakeholders.
* Developed data models in **Power** **BI** to enable efficient data exploration, aggregation, and filtering, enhancing the overall performance of reports and dashboards.
* Created data visualizations and dashboards using tools such as **Power BI** to provide insights and facilitate decision-making.
* Implemented **Agile** testing and validation practices to ensure data accuracy, quality, and completeness, identifying and addressing issues early in the development process.

**Environment:** Python, Matplotlib, Seaborn, Plotly, Power BI, SQL, Informatica Power Centre, SQL databases, NoSQL databases, AWS, Git, Microsoft Excel, QlikView, VLOOKUP, Agile.

**Client Merck Pharma, Mumbai, India May 2015 – Feb 2017**

**Role: Data Analyst**

**Responsibilities:**

* Developed predictive models for **healthcare** applications, such as disease prediction and patient risk stratification, using machine learning frameworks like scikit-learn and TensorFlow, and evaluated model performance.
* Proficiently used **SQL** to extract and retrieve relevant healthcare data from complex databases, including Electronic Health Records (**EHR**) and Health Information Systems (**HIS).**
* Proficiently wrote **T-SQL** queries to extract and retrieve **healthcare** data from **Microsoft SQL** Server databases, including patient records, medical billing information, and clinical data.
* Designed, developed, and maintained **ETL** (Extract, Transform, Load) packages in **SSIS** to efficiently extract data from various sources, transform it to meet business requirements, and load it into **SQL** Server databases.
* Utilized **SQL** to clean and preprocess healthcare datasets, ensure data accuracy and consistency, and preparing them for analysis. Designed and maintained **T-SQL** stored procedures to automate complex data processing tasks, such as data aggregation. Migrated Oracle server database systems to **SQL** Server.
* Conducted in-depth data analysis using **SQL**, including creating complex queries and aggregations to identify trends, patterns, and anomalies in patient demographics, clinical outcomes, and medical procedures.
* Monitored and managed **EDI** transactions, ensuring timely and accurate data exchange between systems, and promptly resolved any issues or discrepancies.
* Proficiently used **Python** library Pandas to clean, reshape, and preprocess healthcare data, handling missing values, and converting data into suitable formats for.
* Conducted data analysis using **Python** to uncover insights, trends, and patterns within datasets, created visualizations with **Matplotlib** and **Seaborn** to facilitate data-driven decision-making.
* Applied **VLOOKUP** to combine data from multiple sources, matching key fields, and creating integrated datasets for analysis and reporting, streamlining data consolidation processes.
* Utilized **Excel's** advanced features to perform data analysis, including **pivot tables**, data modelling, and statistical functions, to derive meaningful insights and generate comprehensive reports.
* Designed and built interactive dashboards in **Excel** to visualize data trends and patterns, making complex information more accessible and actionable for stakeholders.
* Created, designed, and maintained interactive reports and dashboards using **SSRS** to deliver meaningful insights to stakeholders, ensuring data-driven decision-making.
* Ensured strict adherence to **HIPAA** guidelines and regulations in generating reports by handling employee insurance and claims data, implementing robust data privacy and security measures to protect sensitive health information.
* Implemented **Agile** principles and methodologies for data analysis projects, ensuring flexibility, collaboration, and responsiveness to changing business needs.

**Environment:** Python, Pandas, Matplotlib, SQL, TSQL, EDI, SQL, ETL, SSIS, SSRS, Excel, VLOOKUP, Agile.

**EDUCATION DETAILS:** Bachelor of Technology at Jawaharlal Nehru Technological University July 2011 – May 2015, Hyderabad, India.