**Yeshwanth N**

LinkedIn: <https://www.linkedin.com/in/yeshwanth-n/> | M: [nyeshwanth23@gmail.com](mailto:nyeshwanth23@gmail.com)

L: Dallas, TX | D: (954) 543-0458‬ | V: H1B ‬‬‬‬‬‬‬‬

Experienced Data Engineer with 10 years of expertise in designing, developing, and maintaining robust data infrastructure. Skilled in data modeling, ETL processes, and database management. Collaborates with cross-functional teams to architect scalable solutions and deliver high-quality results.

**Professional Summary:**

* Around 10 years of IT experience as a Data Engineer in a variety of industries, which includes hands on experience in Hadoop, MapReduce, Hive, Spark (PySpark), AWS, Sqoop, Snowflake, Teradata, Oracle, RDMS, Python, Scala, ETL and Data Visualization. Adept in statistical programming languages like R and Python including Big Data technologies like Hadoop, Hive.
* Expertise in Data Migration, Data Profiling, Data Ingestion, Data Cleansing, Transformation, Data Import, and Data Export through the use of multiple ETL tools such as Informatica Power Centre.
* Proficiency in multiple databases like MongoDB, Cassandra, MySQL, ORACLE, PostgreSQL, and MS SQL Server.
* Hand-on experience in Data Architecture which includes some of the components Data modeling, Data Storage, Data Processing, Data Governance, and Data Integration.
* Extensive experience with AWS Data Services, including Lake formation, Glue ETL, EMR, S3, Glue Catalog, Athena, Kinesis, MSK, Airflow, Lambda, Step Functions, and Event Bridge, indicating a deep understanding of cloud-based data solutions.
* Experience in data warehousing applications using ETL tools and programming languages like python, SQL/PLSQL, Oracle and SQL Server databases, Informatica, and SSIS and also Utilized Cloudera's Impala and Hive for creating and managing data warehouses, optimizing queries for faster data retrieval and analysis.
* Demonstrated expertise in working with the Cloudera Hadoop ecosystem, utilizing components such as HDFS, MapReduce, Hive, Impala, and Spark for large-scale data processing.
* Experience in Text Analytics, developing different Statistical Machine Learning, Data Mining solutions to various business problems and generating Data Visualizations using Python and creating dashboards using tools like PowerBI / Qlik sense / Tableau.
* Expert in designing and implementing efficient data processing pipelines using Spark, Hadoop, and cloud based ETL tools such as AWS Glue and Azure Data Factory (ADF).
* Utilized SSIS to enable OLEDB and ODBC connections between MSSQL and Oracle Database instances,transforming and loading data between sources and destinations.
* Adept and deep understanding of Statistical modeling, Multivariate Analysis, model testing, problem analysis, model comparison, and validation.
* Excellent understanding of all stages in a typical SDLC like Requirement Analysis, Design, Programming, Project Status Review (PSR), Unit Testing, Integration Testing, Support.
* Proficiently utilize Informatica cloud products like IICS (Informatica Intelligent Cloud Services), CDQ (Cloud Data Quality), CAI (Cloud Application Integration), etc., to design, develop, and maintain robust data integration solutions.
* Proficient in designing flexible and scalable data models for a hybrid cloud environment, including Azure and Teradata, catering to data ingestion, integration, privacy, quality, and governance for clients in the beverage industry.
* Deep understanding of enterprise data strategy, landscape, and challenges, encompassing on-premises and Azure cloud environments. Hands-on experience in performing ETL operations using AbInitio environment and creating dashboard reports using Tableau.
* Hands-on experience in working with Spark SQL queries, Data frames, importing data from Data sources, performing transformations, read/write operations, and saving results to the output directory in HDFS.
* Hands-on experience in data pre-processing techniques such as Data Extraction and Data Cleaning with large sets of structured and unstructured data, Exploratory Data Analysis (EDA), Feature Engineering etc.
* Specialized in Forgerock Identity Platform, utilizing Forgerock components such as OpenAM and OpenIDM to build secure and efficient authentication and identity management solutions.
* Logical and physical database designing like Tables, Constraints, Index, etc. using Erwin, ER Studio, TOAD Modeler and SQL Modeler.
* Experienced with setting up databases in AWS using RDS including MSSQL, MYSQL, MongoDB & DynamoDB. Storage using S3 bucket and configuring instance backups to S3 bucket.
* Familiarity with cloud-based data warehousing platforms such as Amazon Redshift, Google BigQuery, or Snowflake.
* Experienced in DevOps practices, utilizing Git, CI/CD (Continuous Integration/Continuous Deployment), and JIRA for seamless collaboration and efficient software delivery.
* Experience with NoSQL data and unstructured data using MongoDB, Cassandra, Redis, ZooKeeper.
* Proficiency in Map Reduce and tools like Hadoop, Hive, Pig, Kafka, S4, Map R for data processing and analytics.
* Proficient in Kubernetes, ensuring efficient deployment and management of containerized applications.
* Experience using Jenkins and Gitlab for DevOps and build pipelines, with a strong focus on AWS technologies.
* Worked with Python and R Libraries like NumPy, Pandas, Scikit-Learn, SciPy, and TensorFlow.
* Experienced in Databricks & Big Data Technologies mainly on Data and Delta Lake Implementation.
* Proficiency in testing Cloud and On-Prem ETL using various tools like Abinitio, Informatica, SSIS, Datastage, Alteryx, and Glu.
* Working knowledge of Integration Services (SSIS), Reporting Services (SSRS), and Analysis Services (SSAS).
* Experience and have knowledge on use of Talend ETL, database, data set, HBase, HIVE, Flume, Sqoop, HDFS, PIG and HDFS components.
* Demonstrated expertise in data transformation using Matillion's extensive library of transformations, including aggregations, joins, pivots, and window functions.
* Experience includes development activities in coding with Oracle SQL, Python, Unix shell scripting and Involved in developing generic and custom AbInitio graphs for Unix environment

**Education:**

* Bachelor’s Degree in Information Technology (Graduated in Jun 2013) JNTU University - Hyderabad, India

**Technical Competencies:**

|  |  |
| --- | --- |
| **Languages** | Python, SQL, HTML, CSS, JavaScript, R, jQuery, Angular, XML, Java, React, Veu.JS, Node.JS. |
| **Python Libraries** | Pandas, NumPy, PySpark, Spark SQL, Plotly, Matplotlib, SciKit-Learn, Tensor Flow. |
| **Machine Learning** | Regression analysis, Ridge, Lasso Regression, K-NN, Decision Tree, Prognostic/Diagnostic, Support Vector Machine (SVM), Artificial Neural , NLP, Network (ANN), CNN, RNN, K Means clustering and Hierarchical clustering. |
| **Statistics** | Hypothetical Testing, A/B Testing, ANOVA, Principal Component Analysis (PCA), Cross-Validation, Correlation. |
| **Database** | MySQL, PLSQL, ORACLE, SQL Server 2008/2012/2014, PostgreSQL, CouchDB, MongoDB, DynamoDB, IBM DB2, NoSQL |
| **Big Data/Hadoop** | Databricks, Bigquery, HDFS Cluster, Hive, Unix shell scripting |
| **ETL** | Talend, Informatica, Alteryx, Teradata, Matillion |
| **Visualization** | PowerBI, Qlik, Tableau, SSRS, Seaborn, Matplotlib, ggplot2. |
| **Cloud Services** | Azure, AWS, GCP |

**Professional Experience:**

**Lennox International - Dallas, TX Aug 2022 – Present**

**Sr. Data Analyst  
*As a Senior Data Analyst at Lennox International, I have contributed my expertise in designing and implementing data solutions to drive business insights and optimize data processes. With a focus on scalability, performance, and data integrity, my key accomplishments include:***

**Roles & Responsibilities:**

* Developed and optimized ETL data pipelines in Azure Data Factory, Databricks, and AWS to efficiently extract, transform, and load large volumes of data from diverse sources into cloud-based data storage systems like Azure Data Lake and AWS S3, ensuring data quality and reliability.
* Developed and sustained stored procedures for both MSSQL and Oracle Databases, merging and converting data from stage tables imported via SSIS (For MSSQL) and SQL Loader (For Oracle).
* Designed, developed, and optimized complex ETL workflows using Informatica PowerCenter, orchestrating seamless data integration from diverse sources to target databases and data warehouses.
* Assisted Teams in implementing Machine learning models with help of data flow graphs from TensorFlow.
* Designed and implemented scalable and robust data models, schema, and architecture using tools like Erwin and advanced SQL, Python, and PySpark, facilitating complex data transformations, cleansing, and enrichment for downstream analytics and reporting. Built models using Statistical techniques like Bayesian HMM and Machine Learning classification models like XG Boost, SVM, and Random Forest.
* Utilized big data technologies like Apache Spark, Hadoop, and Hive to process and analyze vast datasets, enabling faster data processing and insights generation.
* Worked on various Statistical models like DOE, hypothesis testing, Survey testing and queuing theory.
* Leveraged Matillion's powerful ETL capabilities to orchestrate complex data integration workflows, ensuring efficient data processing, and enabling seamless transformation of data for downstream analytics
* Optimized ETL performance by utilizing Matillion's parallel processing capabilities, managing partitioning strategies, and implementing efficient data transformations.
* Used DataStage as an ETL tool to extract data from sources systems, loaded the data into the ORACLE database.
* Created DataStage jobs using different stages like Transformer, Aggregator, Sort, Join, Merge, Lookup, Data Set, Funnel, Remove Duplicates, Copy, Modify, Filter, Change Data Capture, Change Apply, Sample, Surrogate Key, Column Generator, Row Generator, Etc.
* Built automate pipelines using Jenkins and groovy scripts. (10%) Skilled in ETL (Extract, Transform, Load) processes, ideally with Informatica, to integrate and transform data from various sources into Redshift for analysis and reporting purposes.
* Automate different workflows, which are initiated manually with Python scripts and Unix shell scripting. Design and implement data models, schemas, and structures optimized for Snowflake's cloud-based architecture.
* Transformed batch data from several tables containing tens of thousands of records from SQL, Server, MySQL, PostgreSQL, and csv file datasets into data frames using PySpark.
* Created ETL pipelines in Python and PySpark to load data into Hive tables under Databricks Create automation and deployment templates for relational and NoSQL databases including MSSQL and Cosmos Develop SQL queries based on requirement documentation to extract, transform, and load data. Done UNIX shell scripting to spin clusters up and down in Bigdata.
* Develop and maintain internal frameworks such as ADF templates and scalable data processing solutions. Design and implement YAML configurations to enhance the Spark framework, optimizing performance and scalability. Developed APIs and integrate them with Spark Scala for seamless data interaction and integration with external systems.
* Leveraged data lake ingestion patterns to handle structured, semi-structured, and unstructured data formats. Designed event-driven data ingestion workflows using technologies like Apache Kafka for real-time data streaming
* Designed and implemented data transformation processes using AWS Glue, enabling automatic schema discovery, efficient data cataloging, and optimization of data processing tasks. Involved in the ETL process using AbInitio tool to setup a data extraction from several databases.
* Designed and developed data orchestration workflows using workflow management tools like Apache Airflow or Azure Data Factory, automating end-to-end data processing and integration.
* Implemented data design patterns such as OLTP (Online Transaction Processing), ODS (Operational Data Store), OLAP (Online Analytical Processing), EDW (Enterprise Data Warehouse), data marts, and Big Data systems to cater to diverse business needs.
* Executed SQL DDL/DML scripts to set up database objects in Azure SQL Database. Designed and implemented complex ETL workflows to extract data from various sources, perform data transformations, and load it into target data warehouses.
* Implemented and maintained Azure-based data solutions utilizing Databricks, ADLS Gen2, Virtual Machines, Milvus Vector DB, and Python.
* Provided MarkLogic Data Hub Service (DHS) consulting regarding both application and database architecture and implementation.
* Create automation and deployment templates for relational and NoSQL databases including MSSQL and Cosmos DBin Azure using Python
* Developed new client/server report requirements documents based on legacy mainframe reports.
* Built Data flows (effectively ETL for MarkLogic Data Hub) normalizing 100’s of sources and applying advanced Master Data Management (MDM) and Semantic enrichment
* Implement ETL processes to extract, transform, and load data from various sources into the data warehouse using tools like Apache Spark, AWS Glue, or Talend.
* Understanding the client business problems and analyzing the data by using appropriate Statistical models to generate insights.
* Implement data validation and quality checks to ensure data accuracy and integrity within MongoDB.
* Design and implement sharding strategies to distribute data across MongoDB clusters for horizontal scalability.
* Developed automated CI/CD data pipelines using Python Spark in Azure Databricks.
* Created, provisioned multiple Databricks clusters needed for batch and continuous streaming data processing and installed the required libraries for the clusters.
* Utilized AWS CloudWatch and AWS CloudTrail for monitoring and auditing data processes, ensuring compliance and identifying potential issues proactively
* Implemented data validation checks and error handling mechanisms within Matillion workflows to ensure data accuracy and reliability throughout the integration process.
* Utilized Azure Data Factory to create and manage end-to-end data integration workflows, orchestrating data movement and transformation across diverse data sources and destinations.
* Utilize Apache Spark, Python, Scala, and SQL for data transformation, enrichment, and analysis within Databricks notebooks.
* Design and implement complex ETL pipelines using Databricks, ensuring optimal data flow, reliability, and efficiency..
* Utilize SQL extensively and possess familiarity with programming languages such as Python, Scala, and AWS technologies like Athena, EMR, and Redshift.
* Ingested data to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processed the data in Azure Databricks.
* Involved in data load and unload to and from HDFS using Abinitio using Hadoop read write components and direct Hive scripts.
* Designed and implemented complex data transformation logic using ETL/ELT tools, ensuring data is transformed into the required format and structure for downstream consumption.
* Data Management (MDM), semantic ontology, PK and TLS security, schema management and scaled deployment.
* Optimize Snowflake performance through query optimization, partitioning strategies, and clustering.
* Implemented data enrichment and cleansing techniques within ETL/ELT pipelines to ensure accurate, consistent, and high-quality data for downstream analytics and reporting.
* Provided UAT (User Acceptance Testing) and QA (Quality Assurance) support, ensuring data integration solutions meet user requirements and performance expectations.
* Regularly updated JIRA with user story progress and actively drove tasks to completion, contributing to efficient project management.
* Involved in the ETL process using AbInitiotool to setup a data extraction from several databases.
* Led the automation of data extraction and preprocessing pipelines, reducing model development time by 20% and improving data consistency.
* Develop application capabilities using Oracle programming utilities (PL/SQL), triggers etc. Involved in backup activities and preparation of transition documents and ETL (Abinitio) related design and testing.
* Involved in ETL process migration from AbInitio to Hadoop environments.
* Stay up-to-date with Snowflake's latest features and capabilities and evaluate their potential applicability to the organization's needs.
* Leveraged deep expertise in Gathr to design and develop data processing solutions that align with the new ecosystem, enhancing data accessibility and analysis capabilities.
* Conducted a thorough comparison between Gathr and AbInitio, identifying key similarities and differences to inform decision-making.

**Environment: Databricks, Python, NLP, SQL, MSSQL, PL/SQL, Spark SQL, ETL, Azure Data Factory, Data Lake, API, Google Big Query, Qlik Sense, Sqoop 1.2, HIVE 2.0.0, ADF, Talend, SDLC, Agile, Netezza 7.2, Abinitio, Jira, GCP, Informatica, Unix shell scripting, MDM, Snowflake, MarkLogic.**

**Truist Financial, NC Aug 2021 – July 2022**

**Sr. Data Analyst**

***As a Data Engineer at Truist Financial, I designed and optimized data pipelines, integrated diverse data sources, and ensured data quality. Collaborating with cross-functional teams, I contributed to informed decision-making and improved data architecture. My role encompassed technical expertise, performance optimization, and data security contributions.***

**Roles & Responsibilities:**

* Evaluated and selected appropriate data integration strategies to meet business needs while ensuring scalability and maintainability.
* Used machine learning algorithms for classification such as Linear Regression, Ridge Regression, Lasso Regression, Decision Trees, Ensemble Learning and Support Vector Machines.
* Involved on configuration, development of Hadoop Environment on AWS cloud such as Lambda, S3, EC2, EMR (Electronic MapReduce), Redshift.
* Developed ETL pipelines using technologies like Apache Spark, AWS Glue, or Azure Data Factory to ingest, transform, and load data into the data lake.
* Applied MarkLogic Grove technology leveraging common client and middle-tier tools like Javascript Node.js, Java Spring Boot, React and Vue as well as MarkLogic-specific tools like server-side Javascript, MarkLogic XQuery, Template Driven Extraction (TDE), Optic API, and the Java Date Movement SDK.
* Designed data warehouses on AWS S3 using dimensional modeling techniques, and created data marts to maintain inventory and sales databases.
* Leveraged Python and R programming to preprocess and analyze data, applying statistical tests to validate assumptions and draw actionable conclusions.
* Created data governance policies and procedures for managing master data using Talend MDM, resulting in better compliance with regulatory requirements and increased data security.
* Automate different workflows, which are initiated manually with Python scripts and Unix shell scripting.
* Implemented ingestion pipelines to migrate ETL to Hadoop using Spark Streaming and Oozie workflow. Loaded unstructured data into Hadoop distributed File System (HDFS).
* Extensively used multiple libraries for data science in Python like NumPy, SciPy, Pandas and Scikit-Learn for data manipulation and predictive modeling in Amazon SageMaker.
* Developed data applications to migrate and transform large datasets into TSYS using Spark with Scala API
* Integrated Hadoop into traditional ETL, accelerating the extraction, transformation, and loading of massive semi structured and unstructured data. Used Hadoop distributed File System (HDFS) to Load unstructured data
* Engineered Change Data Capture (CDC) pipelines in StreamSets to capture and replicate incremental changes from banking databases, maintaining compliance with financial regulations.
* Designed and developed complex data pipelines using StreamSets Data Collector tailored to the unique needs of the banking domain, ensuring secure and compliant data integration.
* Responsible for converting events and cardholder data for multiple partner conversions like REI, BJs and Kohls.
* Developed Spark applications to extract data from multiple data sources with different file formats and transformed the data into TSYS format.
* Used Resampling Methods like Cross-Validation to ensure model reliability for variations in the data.
* Used AWS Kinesis for realtime data injection of telemetry data from 3D printer through OctaPi(IoT) into the data lake and queried the data using Athena in the bronze and silver layer.
* Fine-tuned Cloudera clusters for optimal performance, including configuring resource allocation, optimizing query execution, and managing memory usage.
* Developed data reconciliation jobs to validate the transformation logic between source data and transformed data in target systems.
* Created reusable job templates in DataStage to accelerate development and ensure consistency across projects.
* Responsible for publishing reconciliation and validation reports in AWS Quicksight dashboard.
* Participated in multiple mock conversion and dress rehearsal events to identify data quality issues and ensured that data conversion is successful.
* Designing and developing new batch programs using COBOL, IMS, JCL and VSAM.
* Deployed data applications in non-prod and prod environments in AWS with internal deployment pipeline.
* Responsible for maintaining 4 real time Python APIs running in Lambda which serve as backend for customer service agent portal
* Developed a batch extract job to export the data from backend RDS to internal data lake using
* Spark.
* Integrated Cloudera-based solutions with cloud platforms like AWS or Azure for enhanced scalability and resource management.
* Developed enterprise managed application deployment pipelines and deployed applications into various nonprod, staging and production environments.
* Created new staging environments in production for beta testing all the services and setup performance environments for APIs perf testing.
* Setting up systems performance monitoring and alerts using application logs in Splunk
* Utilized Java programming to integrate and customize Forgerock components, tailoring identity and access management solutions to specific business needs.
* Presented findings to stakeholders, contributing to the strategic decision to transition to Gathr for its superior performance and scalability.
* Acted as a liaison between different teams, effectively communicating technical requirements and project updates related to Gathr integration.

**Environment: Agile Scrum, MDM, MapReduce, Spark, Scala, Forgerock, Pyspark, python, Java, Kafka, JSON, Parquet, GIT, JSON, Cloudera, AWS Glue, Snowflake, Python, AWS EKS, MarkLogic, Snowflake, Redshift, Lambda, S3, Hadoop, MSSQL, IBM Streams.**

**Nfina Technologies, Alabama Aug 2020 – July 2021**

**Data Analyst/Informatica ETL Developer**

***As Data Engineer with expertise in data modeling, ETL, and cloud tech. Skilled in Python, Spark, SQL, and SAS for insightful analytics. Cloud migration and data mart development for advanced solutions. Collaborative, detail-oriented, and data-driven for optimal results.***

**Roles & Responsibilities:**

* Designed and optimized data pipelines for efficient ETL from various sources into cloud-based data storage systems.
* Created processes for downloading various Medicaid spreadsheets for input into mainframe DB2 tables, i.e., procedure codes, HCPCS, CMS. And, maintained the team analytics/metrics spreadsheets.
* Working closely with business intelligence team to understand requirements and develop ETL jobs and procedures using IBM DataStage and tools.
* Solve complex data integration challenges using PySpark, including data transformation and enrichment.
* Spearheaded the implementation of backend data and ML pipelines using Python, C++, Go, and cutting-edge technologies like Kubernetes.
* Ensured data schema consistency across datasets and documented changes to promote data uniformity and traceability.
* Developed ETL processes in Databricks, extracting weekly information from XML files and loading data into databases using Python.
* Implemented Slowly Changing Dimensions, Star schemas and 3NF data models using IBM Datastage tool.
* Demonstrated expertise in implementing complex data transformations using Informatica transformations, expressions, and scripts, enabling intricate data manipulation and enrichment.
* Created comprehensive documentation for data warehouse structures, ETL processes, and data transformation workflows, ensuring clarity and knowledge sharing among team members.
* Design and implement optimized data storage solutions using various database systems, considering data modeling techniques like normalization and denormalization.
* Develop ETL processes to extract, transform, and load data from various sources into data warehouses, ensuring data quality and consistency.
* Collaborated with cross-functional teams to define and implement a data quality strategy, using Databricks and Talend MDM, which reduced data inconsistencies and improved the overall quality of data.
* Managed database instances, including installation, configuration, and maintenance of PostgreSQL on various platforms. Leverage Databricks Delta Lake to ensure data quality, consistency, and ACID transactions in data processing.
* Utilize columnar storage formats (such as Parquet) to optimize storage and query performance in the data warehouse. Applied Python-based ML algorithms for predictive analysis and automatic suggestions using Kinesis Firehouse and S3 Data Lake.
* Employed PySpark scripts for data partitioning and optimization to improve data processing efficiency.
* Designed and implemented data models in Snowflake for efficient data storage and querying. Collaborated with Data Scientists to provide optimized data solutions for data-driven decision-making.
* Develop APIs in Spark Scala to facilitate integration with external systems and data sources.
* Developed CI/CD pipelines using AWS CodePipeline and CodeBuild for automated data engineering deployment.
* Demonstrated expertise in data encryption and security for 700 tables in a data mart. Developed GraphQL schemas for seamless data integration and retrieval.
* Responsible for developing and re-defining several complex jobs and job sequencers to process various feeds using different datastage stages, properties.
* Enhance existing frameworks using JavaScript, enabling seamless interaction with restful APIs.
* Led the migration of on-prem Hadoop systems to GCP for modernized data infrastructure. Implemented and deployed Kafka clusters for efficient data streaming. Migrated multiple MSSQL On-Premises Instances to Address data skewness and imbalance issues in PySpark jobs to ensure efficient resource allocation.
* Utilized SAS/Base and SAS/Advanced for descriptive analytics and statistical analysis.
* Developed Python scripts for efficient AWS resource management using BOTO SDK and AWS CLI.
* Managed and optimized database objects like Tables, Views, Indexes, and Stored procedures using SQL and PL/SQL.
* Validated and updated Models for process mappings, business object models, and system object models.
* Working with CICS, IBM MVS, SPUFI, ISPF, COBOL, VSAM, JCL, DB2, SQL & File-Aid, Panvalet, Xpediter, Intertest, CA/Db2 tools, TSO, FileZilla for FTP and Connect Direct etc.
* Develop and maintain data processing frameworks and libraries within the Databricks environment for reuse across projects.
* Implement advanced optimizations and tuning techniques to enhance Spark performance and reduce processing times.
* Integrated data from diverse sources into DynamoDB for data consistency. Utilized serverless processing frameworks like AWS Lambda to trigger data processing tasks directly from the data lake.
* Collaborate with cross-functional teams to understand data requirements and provide MongoDB-based solutions.
* Implemented CI/CD pipelines with Terraform and GIT for automated application deployment.
* Demonstrated expertise in AWS cloud services for efficient cloud-based data operations.
* Implemented rigorous data quality checks within StreamSets pipelines, ensuring the integrity of financial data and compliance with industry standards. Integrated Databricks with cloud platforms like AWS, or Google Cloud for leveraging cloud-based services.
* Assisted in troubleshooting and resolving ETL job failures, conducting root cause analysis and implementing corrective actions.
* Engineered multiple connectors on IBM Streams for processing Site Catalyst (Online Interaction) data.

**Environment: SDLC, Agile (SCRUM), HIVE, ER design, Python, IBM Streams, PySpark, SQL, PL/SQL, API, XML, JSON, CSV, PARQUET, Abinitio, JCL(Mainframe), AWS, Amazon EC2, Amazon S3, Amazon RDS, Lambda, EMR, Amazon Dynamo DB, Snowflake, Kinesis firehouse, GIT, Terraform, ETL, Jira, Azure, GCP,** **COBOL, JCL, DB2, CICS.**

**DXC Technologies, Hyderabad, India Oct 2017 - Aug 2019**

**Data Analyst/Report BI Analyst**

***As Data Analyst at DXC Technologies with expertise in optimizing data processes and building efficient data pipelines. Proficient in Python, Spark, SQL, and Hive for data analysis and transformation. Skilled in DynamoDB performance tuning and REST API engineering. Collaborative and detail-oriented, contributing to successful data-driven initiatives and seamless data integration.***

**Roles & Responsibilities:**

* Collaborated closely with cross-functional teams, forging strong partnerships and gathering valuable insights to provide expert recommendations on data architecture and storage solutions, aligning with business objectives.
* Develop ETL processes using Snowflake's built-in features like Snowpipe, Task, and Stored Procedures.
* Extensively used SQL, PL/ SQL, Triggers and views using IBM DB2
* Implemented data quality checks and validation mechanisms to ensure the accuracy and consistency of data within the data lake.
* Collaborated with Data Stewards to define data governance policies and enforce data quality standards across the data lake.
* Leveraged Azure DevOps extensions and integrations to enhance data engineering workflows and streamline processes.
* Streamlined and optimized complex ETL/ELT processes, data pipelines, and workflows, ensuring seamless data ingestion, processing, and querying capabilities, thereby enhancing overall efficiency and performance.
* Design, develop, and maintain MongoDB databases, collections, and schemas to accommodate large-scale data processing and storage needs.
* Designed Unit test document after the datastagedevelopment and verified results before moving it to QA.
* Utilized Data Factory to extract, transform and load (ETL) data from multiple sources into a centralized MDM system, enabling more accurate and efficient data analysis.
* Improved data governance by analyzing data lineage processes to identify vulnerable data points, control gaps, data quality issues, and overall lack of data governance.
* Used Power BI for the data visualization during the quick model construction process in Python.
* Improved data quality by working on data cleansing and standardization using the cleanse functions in Informatica MDM.
* Created workflows and mappings using Informatica ETL, working with transformations such as lookup, source qualifier, update strategy, router, sequence generator, aggregator, rank, stored procedure, filter, joiner, and sorter.
* Developed robust ETL pipelines to extract, transform, and load data from diverse sources into PostgreSQL databases.
* Develop and maintain ETL pipelines that efficiently move and transform data from source systems to the data warehouse.
* Implemented data privacy measures within StreamSets pipelines, ensuring sensitive customer information was encrypted and protected during data integration.
* Design, develop, and maintain Amazon DynamoDB databases and tables to efficiently store and retrieve data at scale.
* Implement efficient ETL processes to ingest and transform data from various sources into DynamoDB using tools like AWS Glue, Lambda, or custom scripts.
* Design and implement data models and schemas in DynamoDB that align with data access patterns and query requirements.
* Develop and manage automated workflows using Databricks Jobs and integration with orchestration tools like Apache Airflow.
* Design and implement logical and physical data models using Databricks, defining schema structures and data relationships to support efficient data storage and querying.
* Worked with AWS Kinesis for real-time data streaming and processing, enabling timely data insights and analytics.
* Implemented data ingestion from various sources into Hadoop/HDFS, using tools like Sqoop and Flume.
* Proficient in writing complex MapReduce programs to perform distributed data processing efficiently.
* Leveraged HBase and other NoSQL databases for storing and retrieving structured and semi-structured data.
* Demonstrated an exceptional command of Azure data technologies, with an expert grasp of Databricks, ADLS Gen2, and Virtual Machines, translating into well-architected and high-performing data solutions.
* Leveraged Spark optimizations and harnessed Data Frame APIs to achieve optimal data processing and analytical capabilities, enabling data-driven insights.
* Develop data pipelines using Hadoop ecosystem tools like HDFS, MapReduce, Hive, and Pig to process and analyze large volumes of data.
* Utilize Apache Spark for both batch and streaming data processing, optimizing performance and scalability.
* Implement Apache Kafka for real-time data streaming, ensuring high-throughput and low-latency data ingestion.
* Leverage stream processing frameworks like Apache Flink and Apache Beam to handle real-time data transformations and analytics.
* Applied Agile/Scrum methodologies, instilling agility and efficiency in project management, resulting in successful and timely project delivery.
* Integrated data warehouse solutions with cloud-based platforms like AWS or Google Cloud to leverage scalable compute and storage resources.
* Lead code reviews for PySpark code, providing constructive feedback and ensuring adherence to standards.

**Environment: Python, Power BI, Azkaban, PySpark SQL, Presto, Hive, Apache Crunch, Elastic Search, Spring boot, Eclipse, Abinitio, GIT Repository, Amazon S3, AWS Ec2/EMR, MDM, Spark cluster, Hadoop Framework, Sqoop, Informatica, ETL**

**ValueLabs Pvt. Ltd - Hyderabad, India Jul 2013 - Sep 2017**

**Data Engineer**

***ValueLabs helps clients derive insights from their data by offering data analytics and business intelligence services. This involves data integration, analysis, visualization, and the development of dashboards and reports to enable data-driven decision-making and improve business performance.***

**Roles & Responsibilities:**

* Developed Spark applications using Scala and Spark-SQL for data extraction, transformation, and aggregation from multiple file formats, uncovering insights into customer usage patterns.
* Migration of SDE Applications data pipeline from Hive to Pyspark using HDFS, Hive, HBase, Pyspark, Unix and Python scripts following Agile Methodology. Designing and developing new batch programs using COBOL, IMS, JCL and VSAM.
* Designed and implemented efficient PostgreSQL database schemas, considering data normalization, indexing, and constraints.
* Established data lineage and metadata management processes to track the origins and transformations of data within the data lake.
* Used DataStage Parallel Extender stages namely Datasets, Sort, Lookup, Change Capture, Funnel, Peek, SCD, and Row Generator.
* Demonstrated a rapid grasp of AI & ML concepts, tools, and workflows, continuously expanding expertise in the field.
* Developed and maintained data pipelines using AWS Data Pipeline and Apache Spark on EMR for efficient data extraction, transformation, and loading (ETL) from various sources to data warehouses.
* Integrated data from various sources into the GraphQL data layer, enabling seamless query and data retrieval for applications.
* Undertook data analysis and collaborated with downstream analytics teams to shape data according to their requirements.
* Designed and developed software products for financial institutions using Abinitio and Talend.
* Experienced in performance tuning of Spark Applications, optimizing batch interval time, parallelism, and memory usage.
* Designed and created Data Vault entities, including Hubs, Links, and Satellites, based on the Data Vault architecture.
* Contributed to the development of Informatica mappings and prepared design documents (DD), technical design documents (TDD), and unit acceptance testing (UAT) documents.
* Work with AWS services like AWS Data Pipeline and Amazon Kinesis for data movement and integration with DynamoDB.
* Integrated data lakes with cloud-based analytics services like AWS Redshift, Snowflake, or Synapse Analytics to support advanced analytics and reporting.
* Develop custom connectors and APIs to integrate MongoDB with other data storage systems and services.
* Design and implement data archiving and retention strategies to manage historical data within MongoDB.
* Utilize Databricks for data processing, leveraging SQL, Spark, and Scala for complex data transformations and analytics.
* Replaced existing MapReduce programs and Hive Queries with Spark applications using Scala.
* Deployed and tested (CI/CD) developed code using Visual Studio Team Services (VSTS). Conducted code reviews for team members to ensure proper test coverage and consistent code standards.
* Stay updated with the latest features and capabilities of Databricks and other relevant technologies to suggest and implement improvements in data engineering processes.
* Demonstrate strong expertise in RDBMS technologies, including Oracle/PLSQL, OBIEE/OAS, OODB, ODS, and Warehouse, ensuring data integrity and accuracy. Create Batch Jobs using JCL from Scratch and test the Batch data.
* Design and orchestrate complex data workflows using tools like Apache NiFi, ensuring smooth data movement and transformation and Implement workflow automation with tools like Apache Airflow, scheduling and monitoring data processes.
* Utilize Cloud AWS services for data storage and access, proficiently managing data in Cloud environments, and employing Snowflake and S3 for cloud-based data solutions.
* Developed Data Migration methodologies, specifically around DataStage to SAP BODS migration.
* Design, build, and maintain data pipelines for ingesting, transforming, and loading data into Snowflake data warehouse.
* Develop and deploy data solutions in AWS cloud environments, utilizing cloud-based technologies to optimize data operations and scalability.
* Develop and maintain ETL pipelines that efficiently move and transform data from source systems to DynamoDB.

**Environment: Python, Teradata, Netezza, Oracle 12c, PySpark, MS Office (Word, Excel, and PowerPoint), SQL Server, UML, MS Visio, Abinitio, Oracle Designer, SQL Server 2012, Cassandra, Oracle SQL, Athena, SSRS, SSIS, AWS S3, AWS Redshift, AWS EMR, AWS RDS, DynamoDB, Lambda, Hive, HDFS, Sqoop, Scala, No-SQL (Cassandra) and Tableau, Informatica, Unix Shell Scripting.**