**Uday**

Phone: 682-377-0728

Mail: vuday6556@gmail.com

**PROFESSIONAL SUMMARY:**

* Overall, 10 years of in-depth experience in the field of data engineering, including ingestion, data lake, data warehouse, reporting, and analytics.
* Experience with all stages of the SDLC and Agile Development model right from the requirement gathering to Deployment and production support.
* Extensive working experience on Data Migration, Data Conversion, Data Quality, Data Integration and Metadata Management Services and Configuration management. Strong experience in migration to other databases to Snowflake.
* Expert in creating **ETL packages**, migrating data, cleaning data, and backing up data files, and synchronizing daily transactions using Informatica, **Alteryx**, **Autosys** etc.
* Experience in migrating data using Sqoop from HDFS and Hive to Relational Database System and vice-versa according to client's requirement. Expert in data ingestion tools like Sqoop, Flume, and Kafka.
* Hands on experience with File System and its eco-system components MapReduce, Hive, HBase, Flume, Sqoop, Zookeeper and Oozie, Storm, Spark, Kafka.
* Hands on Experience on Python programming for data processing and to handle Data integration between On-premises and Cloud DB or Data warehouse.
* Proficient in developing data engineering applications using Angular and JavaScript, leveraging their powerful front-end capabilities for data visualization and user interaction.
* Experience working with Azure Blob Storage, Azure Data Lake, Azure Data Factory, Azure SQL, Azure SQL Data warehouse, Azure Analytics, Polybase, Azure HDInsight, Azure Databricks.
* Development level experience in Microsoft Azure providing data movement and scheduling functionality to cloud-based technologies such as Azure Blob Storage and Azure SQL Database.
* Knowledge and experience on AWS services like Redshift, S3, Glue, Athena, Lambda, CloudWatch and EMRs like Hive and Presto.
* Extensive experience in data engineering within the property and casualty insurance industry, working on projects focused on data management, analytics, and reporting.
* Proficiency in Big Data frameworks and tools such as Hadoop, Spark, Kafka, or Flink.
* Experience with distributed file systems like HDFS and object storage systems like Amazon S3.
* Knowledge of NoSQL databases (e.g., Cassandra, MongoDB) or columnar databases (e.g., Parquet, ORC).
* Familiarity with data querying languages such as SQL, Hive, or Pig.
* Competence in programming languages commonly used in Big Data, such as Java, Python, Scala, or R.
* Hands on experience in setting up workflow using Apache Airflow and Oozie workflow engine for managing and scheduling Hadoop jobs.
* Experience in developing Conceptual, logical models and physical database design for Online Transactional processing (OLTP) and Online Analytical Processing (OLAP) systems.
* Good experience in Technical consulting and end-to-end delivery with data modeling, data governance.
* Strong SQL development skills including writing Stored Procedures, Triggers, Views, and User Defined functions.
* Experience working with NoSQL databases and big data tools including **Hadoop, Hive and Spark.**
* Experience in creating **Teradata SQL** scripts using **OLAP** functions like rank Over to improve the query performance while pulling the data from large tables.
* Excellent knowledge and extensively using NOSQL databases (Cassandra 3.11, MongoDB 3.6, Dynamo DB, Snowflake DB).
* Highly proficient in **Data** **Analysis**, **Design**, **Data** **Management**, **Databases,** and **Data** **Visualization**.
* Experience in **Data Warehousing** applications, responsible for the Extraction, Transformation and Loading (ETL) of data from multiple sources Into Data Warehouse.
* Experience in Data masking using **Informatica Data** masking transformation and Strong experience in using **Excel** and **MS Access** to dump the data and analyze based on business needs.
* Extensive knowledge on designing Reports, Scorecards, and Dashboards using Power BI and Tableau.
* Strong analytical skills with ability to quickly understand a client's business needs. Involved in meetings to gather information and requirements from the clients.

**TECHNICAL SKILLS:**

* **Database**: MySQL, SQL Server 2008/2012/2014.
* **Programming Languages:** C, C++, C#, Python.
* **Database tools**: SQL Server Management Studio (SSMS), MYSQL workbench, Azure Data Studio, Toad.
* **Hadoop EcoSystem:** HIVE, SQOOP, PIG, YARN, MAP REDUCE, Apache Spark, HDFS, Kafka.
* **Hive**: Hive Architecture, Tables, Partitioning, Bucketing.
* **SQOOP**: Import, Export
* **Spark:** Spark SQL, RDD’s, Data Frames, DF operations, batch processing applications, streaming applications.
* **IDE:** PyCharm, Eclipse, Jupyter Notebook, Databricks, Visual Studio Code.
* **Cloud:** Amazon Web Services (AWS), EC2 Instance, S3 bucket.
* **ETL**: Informatica, Alteryx
* **Visualizing Tools:** Tableau, Qlikview, SSRS, Seaborn, Matplotlib, ggplot2.
* **Frontend:** HTML, CSS, ADO.NET.

**WORK EXPERIENCE:**

**Client: Mercedes-Benz Financial Services - Farmington Hills, MI Oct 2021 – Current**

**Role: Sr Data Engineer**

**Responsibilities:**

* As a Sr. Data Engineer, I am responsible for development and implementation of Data Flow Pipelines.
* Involved in data migration from OnPrem to AWS Cloud using AWS DMS.
* Used AWS S3 for datalake as staging datastore.
* Used Apache Airflow to schedule the dataflow from S3 to Redshift.
* Proficient in using Glue as ETL service; used PySpark for CDC and data cleansing functions.
* Created Glue crawlers for structured schema and table definitions.
* Worked on EMR for huge unstructured data processing into Redshift.
* Used Redshift as Datawarehouse.
* Used Big Data Tools like Apache Kafka and Nifi to ingest data in batch mode for the data warehouse.
* Big Data tools like Apache Spark and flink are used to the distributed data processing and enabling tasks like data transformation.
* Apache Airflow and Nifi are used as a big data tool to design and implement data pipelines to automate data workflows.
* Designed and implemented ETL workflows using AWS Glue, leveraging its serverless architecture and visual interface to orchestrate and schedule data transformation and loading tasks
* Created and managed AWS Glue Data Catalog, a centralized metadata repository, to catalog and organize data assets across various data sources, enabling easy discovery and access.
* Designed and implemented data processing workflows using AWS EMR, leveraging its integration data frameworks like Apache Spark, Apache Hadoop, and Apache Hive.
* Deployed and managed EMR clusters, dynamically scaling resources to handle large volumes of data and processing tasks efficiently.
* Extensive experience working with Parquet, Avro, and JSON file formats for efficient storage, processing, and exchange of structured and semi-structured data.
* Developed Java-based data ingestion pipelines, leveraging Parquet, Avro, and JSON formats to efficiently handle large volumes of data.
* Implemented data transformation and manipulation logic in Java to process Parquet, Avro, and JSON files, ensuring data quality and adherence to business requirements.
* Leveraged Java libraries and frameworks (such as Apache Avro and Jackson) to read, write, and manipulate Parquet, Avro, and JSON data.
* Various Big data tools were used to schedule, monitor and manage complex data pipelines involving multiple processing steps and dependencies.
* Extracted and loaded data into Data Lake environment by using Hive which was accessed by business users.
* Proficient in leveraging AWS Redshift, a fully managed data warehousing service, to design and implement scalable and high-performance data engineering solutions.
* Designed and optimized Redshift data models, schemas, and distribution strategies to efficiently store and query large volumes of structured and semi-structured data.
* Worked on Informatica MDM to integrate data from databases and web services into a centralized MDM repository. It provides connectors and adapters for seamless data ingestion and integration.
* I synchronized master data across various systems in batch mode, as Informatica MDM can facilitate data consolidation and distribution to ensure consistent and up to data information.
* Worked on Informatica MDM to integrate with other teammates to leverage master data for advanced analytics, reporting and decision making.
* Steered the software engineering activities throughout Software Development Life Cycle (SDLC), including requirement gathering, design, development, testing, and documentation.
* Designed a common framework for REST API consumption using Spring Rest Templates. Designed application using Spring MVC, ISP, ISTL, and AlAX on the presentation layer, the business layer is built using spring and the persistent layer uses Hibernate.
* Involved in Ecommerce project Development using Spring Web Flow and Hybris and developed front- end UI with Backbone.js, Query, CSS3, Dojo and AJAX. Built cloud Microservices and developed spring-based application using Spring Boot. Developed RESTful Web Services used by several clients like Mobile, Tablet, as well Web.
* Extensive experience in leveraging JavaScript frameworks and libraries (such as D3.js, Chart.js, or Highcharts) to create visually appealing data visualizations and charts for data engineering projects.
* Utilized Angular's reactive programming paradigm to handle asynchronous data streams and real-time updates, facilitating seamless integration with backend data processing and streaming systems
* Implemented data validation and input handling mechanisms in Angular forms to ensure accurate and consistent data entry in the applications.
* Developed and implemented robust data pipelines and ETL processes to ingest, transform, and load large volumes of property and casualty insurance data from various sources, ensuring data integrity and quality.
* Designed and optimized data models and schemas specific to property and casualty insurance data, enabling efficient storage, retrieval, and analysis of policy, claims, underwriting, and financial data.
* I have worked on designing and developing ETL (Extract, Transform, Load) workflows using AWS Glue. This involves writing Glue ETL scripts in Python or Scala to transform the extracted data into the desired structure or schema, performing data cleansing, aggregation, or enrichment tasks, and handling data quality issues.
* Utilized Terraforms infrastructure state management capabilities to track and manage the lifecycle of infrastructure resources, ensuring configuration consistency and easy resource management.
* Implemented infrastructure updates and changes using Terraform, minimizing downtime and disruption to data engineering projects.
* Conducted infrastructure cost optimization exercises using Terraform, identifying and implementing resource right-sizing and cost-saving strategies.
* Loaded and transformed large sets of structured, semi structured, and unstructured data using Hadoop/ Big Data concepts.
* Created Hive External tables to stage data and then move the data from Staging to main tables.
* Implemented the Big Data solution using Hadoop, hive and Informatica to pull/load the data into the HDFS system.
* Used GIT Repositories for code checking.
* Designed and produced logical and physical data models for the financial platform and other in-house applications running on Oracle databases.
* Created dimensional models based on star schemas and designed them using Erwin.
* Worked with ETL tools to migrate data from various OLAP and OLTP databases to the AWS Cloud.
* Exported the analyzed data to the relational databases using Sqoop for visualization and to generate reports for the BI team.
* Wrote Python scripts to parse XML documents and load the data in database.
* Developed numerous MapReduce jobs in Scala for Data Cleansing and Analyzing Data in Impala.
* Used Hive to analyze data ingested into S3 Athena external tables and aggregated metrics for reporting on the dashboard.
* Optimized Hive queries to extract the customer information from HDFS.
* Analyzed data using Hive the partitioned and bucketed data and computed various metrics for reporting.
* Used python scripts to update content in the database and manipulate files.

**Environment:** AWS Services, Hive/Impala, Python, GIT

**Client: Ditech - Fort Washington, PA. March 2020 – Sept 2021**

**Senior Data Engineer**

**Responsibilities:**

* Created streams using Spark and processed real time data into RDDs & data frames and created analytics using SPARK SQL.
* AWS Infrastructure setup on EC2 and S3 API implementation for accessing S3 bucket data file.
* Written ETL jobs in using spark data pipelines to process data from different sources to transform data to multiple targets.
* Developed an end-to-end ETL pipeline using Spark-SQL on Spark engine and imported data from AWS S3 into Spark RDD, performed transformations and actions on RDDs.
* Experienced in working with the Spark ecosystem using Spark SQL queries on different formats like text file, CSV file.
* Developed multiple POC’s using PySpark module Spark SQL and in hive using Hive Query Language (HQL).
* Load and transform large sets of structured, semi structured, and unstructured data.
* Used Informatica MDM tool to monitor the master data continuously to provide reports and alerts to identify data quality issues and track data quality metrics
* Written SQL queries and stored procedures for the application to communicate with Database.
* Design and implement Spark SQL tables, Hive scripts job for scheduling and create workflow and task flow.
* Involved in the ETL phase of the project & designed and analyzed the data in oracle and migrated to Hive.
* Developed and maintained ETL pipelines to load data from various sources into Redshift, ensuring data integrity, consistency, and timeliness.
* Implemented data transformation and cleansing operations in Redshift using SQL, enabling data standardization, aggregation, and enrichment for downstream analytics and reporting.
* Collaborated with Java team to integrate Angular-based front-end components with data processing APIs and microservices, enabling end-to-end data workflows.
* Implemented data filtering, sorting, and pagination functionalities in Angular-based data tables and grids, allowing users to efficiently navigate and explore large datasets.
* Conducted performance optimization of Angular applications, leveraging techniques such as lazy loading, code splitting, and caching to improve loading times and responsiveness.
* Built an API for utilizing AWS Lambda to manage servers and run code in AWS. Developed and executed MongoDB scripts to insert and update NoSQL databases, as well as to automate data updates and reporting. Contributed to Setup Servers (Boss, Apache, and UNIX) and deployed applications in Dev, ST, and UAT environments.
* Implemented data security and access controls on EMR clusters, ensuring compliance with data governance and privacy requirements.
* Integrated EMR with data storage and analytics services like Amazon Redshift or Amazon Athena, enabling seamless data transfer and analysis between EMR and other platforms.
* Utilized AWS Glue's automatic schema discovery and data cataloging capabilities to infer and capture data schemas from diverse data sources, reducing manual efforts and improving data consistency.
* Integrated AWS Glue with other AWS services like Amazon S3, Amazon Redshift, or Amazon RDS, enabling seamless data ingestion and transformation across data warehouse and analytics platform.
* Build the code and deploy the code in the Jenkins QA server using an EC2 instance.
* Worked on the Warrants Creation Project which involved the use of Microservices Architecture, Java 8, Spring Boot, Spring Data JPA, Hibernate, and Kafka.
* Reviewed pull requests for Cloud Watch, EKS, and CI/CD changes, and managed team status using Atlassian Jira for task visibility.
* Implemented environment reproducibility by defining and managing infrastructure configurations using Terraform, ensuring consistent and scalable deployments across multiple environments.
* Collaborated with cross-functional teams to version control Terraform configurations, enabling easy collaboration, change tracking, and auditing of infrastructure changes.
* Designed and implemented data serialization and deserialization mechanisms in Java to convert data between Parquet, Avro, and JSON formats.
* Collaborated with cross-functional teams to optimize data engineering workflows, leveraging Java's parallel processing capabilities to improve performance and scalability.
* Implemented data schema evolution and compatibility strategies in Java, ensuring seamless data migration and backward compatibility when working with Parquet, Avro, and JSON data.
* Conducted performance tuning and optimization of Java-based data processing code, improving processing speed and resource utilization for Parquet, Avro, and JSON data.
* Worked with Java-based data processing frameworks (such as Apache Spark or Apache Beam) to leverage the power of distributed computing for large-scale data engineering projects involving Parquet, Avro, and JSON.
* Involving in client meetings and explaining the views to supporting and gathering requirements.
* Working in an agile methodology, understand the requirements of the user stories.
* Prepared High-level design documentation for approval.
* Performed real time analytics on transactional data using python to create statistical models for predictive and reverse product analysis.
* Used COBOL program to extract data from mainframe systems. These programs are responsible for reading and processing records from files or databases and transforming them into a suitable format for further processing.

**Environment:** PySpark, SQL, Spark SQL, Data Frames, AWS EC2 Instance, AWS S3 Buckets, Hive, RDD’s, HQL.

**Client: AT&T - Plano TX**

**Role: Azure Data Engineer July 2019 – Feb 2020**

**Responsibilities:**

* Write required DAX functions using aggregates, date, logical, and table functions.
* Created Tabular cube and published it to Azure Analysis Services.
* Designed Power BI reports and published into the Power BI report server.
* Work closely with IT and the Business group to understand business reporting requirements and analyses logical model and develop subject matter expertise in a short time.
* Developed Star and Snowflake schema models.
* Python Used to extract custom data.
* Worked on providing analytical solutions that utilizes Machine Learning.
* Utilize tools such as R Studio, Python, Jupyter notebooks, Azure Data Lake, Azure Databricks, Azure SQL database, Azure SQL Data Warehouse.
* Involved in the CI/CD pipeline management for managing the weekly releases.
* Worked with Hive, writing the scope scripts in Data Lake and HDFS to structure the Peta bytes of unstructured data stored in the Azure DataLake (Cosmos) big data system.
* Map source system data elements to target system and develop, test, and support extraction, transformation, and load processes.
* Scheduled jobs to Data Transformation tasks using COBOL like Data cleansing, normalization, aggregation, and sorting.
* COBOL is often used in batch processing scenarios where large volumes of data need to be processed in scheduled jobs. It can handle complex workflows, file handling, and record-level operations efficiently
* Create schedule automated refresh using gateway.
* Write a required DAX functions using aggregates, date, logical and table functions.
* Designed Power BI dashboards in Online by taking Data from Azure DW.
* Designed extensive reports and dashboards using Power BI Desktop and Cloud.
* Model & Architect DW design, direct and/or execute the technical characteristics of the overall strategy of the data warehouse and ETL process.
* Conversion of Stored Procedures, Functions and Sequences from Oracle to Azure SQL Database.
* Created Data Factory pipelines processing data loading from source to destinations.
* Run ETL jobs in Informatica to meet data integrity and reliability requirements and migrate the data to higher environments.
* Used Informatica MDM 360 to check data quality, completeness, and consistency. By performing data profiling to identify the data anomalies, outliners, and patterns to make informed decisions on cleaning data.
* Used Informatica MDM 360 as a central data integration hub in the project for data repository allowing multiple application integrations.
* Integrated Informatica MDM 360 to inhouse API's and other systems for data exchange and automation
* Data processed by using Poly base in Azure Synapse Analytics DW.
* Migrated all SSIS packages from on premises to Azure Data factory.
* Development of custom scripts and stored procedures for data import and manipulation
* Extensively using Azure Analysis Services to create semantic models.
* Developed Pipelines in Azure Data factory using Multiple Activities.

**Environment:** SQL Server, Synapse, Data Factory, Azure, Power Bi, Snowflake, Python, Jupyter, Oracle, SSIS

**Client Mayo Clinic - Rochester MN Jan 2018 – June 2019**

**Role: Data Engineer**

**Responsibilities:**

* Extracted the data from Teradata & MySQL into HDFS using Sqoop export/import.
* Developed Sqoop jobs with incremental load to populate Hive External tables.
* Implemented design patterns in Map Reduce to convert business data into custom format.
* Experienced with handling different compression codec's like LZO, GZIP, and Snappy.
* Expert in optimizing performance in hive using partitions and bucketing concepts.
* Experience on working hive dynamic partition to overcome hive locking mechanism.
* Developed UDFs in Java as and when necessary to use in Hive queries.
* Developed crontab for scheduling and orchestrating the ETL process.
* Implemented custom data transformations and data cleaning logic in AWS Glue using PySpark or Scala, ensuring data quality and adherence to project requirements
* Involved in indexing hive data using Solr and prepare custom tokenizer formats for querying.
* Involved in designing a real time computation engine using Kafka.
* Experienced with writing build jobs using Maven and integrate that with Jenkins.
* Experienced in building integration pipelines using Talend for ingestions purpose.
* Worked and learned a great deal from AWS Cloud services like EC2, S3.
* Conducted performance tuning and optimization of Redshift queries and workloads, leveraging techniques like query optimization, distribution key selection, and sort key design to improve query performance.
* Migrated an existing on-premises application to AWS. Used AWS services like EC2 and S3 for small data sets processing and storage, Experienced in Maintaining the Hadoop cluster on AWS EMR.
* Worked with elastic MapReduce and setup Hadoop environment in AWS EC2 Instances.
* Used Lambda, Kinesis, DynamoDB, CloudWatch from AWS.
* Actively participated in Agile development methodologies, contributing to sprint planning, code reviews, and collaborative problem-solving within data engineering teams.

**Environment**: HDP2.1, Hive 1.13, Sqoop 1.4.1, Pig0.12, Spark 1.3, crontab, Tez 0.4.0, Solr4.7.2, MapReduce, SQL, Tableau, Java, storm, python, Talend.

**Client: Accenture – India**

**Role: Big Data Analyst/ Data Engineer Aug 2014 – June 2017**

**Responsibilities:**

* Analysed and researched the problems, planed solutions, recommended software and systems, and coordinated development to meet business and system requirements.
* Created the high-level Design for the Data Ingestion and data extraction Modules.
* Developed data ingestion modules to retrieve the data and transform it into required format.
* Created Hive Generic UDF's to process business logic that varies based on values provided.
* Created ETL jobs in informatica to load the data into HDFS from different source systems.
* Cleaned and transformed the data loaded into HDFS using MapReduce.
* Created the Sqoop incremental imports, MapReduce scripts for transforming the data and load it from Teradata to Hive.
* Created and maintained cognos reports according to provided requirements.
* Migrated Teradata Database data using Sqoop into Hive Dynamic partition tables using staging tables.
* Ingested the data from several source systems in to HDFS and Hive.
* Involved in data analysis to resolve complex data missing/error issues by identifying patterns using python.
* Proficient in working with industry-standard property and casualty insurance data formats, such as ACORD standards (AL3, XML) and insurance-specific data models, including ISO and AIR.
* Implemented data integration and consolidation strategies to bring together disparate property and casualty insurance data sources, harmonizing data structures and resolving data quality issues.
* Developed and maintained data warehouses or data lakes specifically tailored for property and casualty insurance, allowing for comprehensive analysis, reporting, and business intelligence activities.
* Used Oozie workflow engine to manage interdependent Hadoop jobs and to automate several types of Hadoop jobs such as Java MapReduce, Hive and Sqoop.
* Develop complex SQL queries/scripts and similar artifacts to validate the completeness, integrity and accuracy of data within a Big Data, ETL process.
* Involved in daily agile scrum meetings, demos, monthly story reviews, release retrospectives, provided estimates for stories and tasks.
* Worked with End-Users to ensure user acceptance testing criteria and standards are met.

**Environment:** Hadoop Eco System, Hive, HDFS, Sqoop, Python, Java, Oracle 11g, Teradata, SQL, Informatica, Cognos, PL/SQL, XML, XSLT, XML SPY, MS Excel, UNIX, Putty, UNIX Shell Scripts, SQL Server, MS Visual studio

**Client: MetLife Insurance – India Jan 2012 - July 2014**

**Role: Data Analyst**

**Responsibilities:**

* Identified and defined the datasets for the report generation by writing queries and stored procedures
* Developed drill through, drill down, linked, sub and parameterized reports for better business analysis using SSRS.
* Extracted large volume of data from various data sources using SSIS packages
* Designed different types of reports using Report Designer 2008 R2 for financial analysis.
* Worked on Statistical Analysis of data for purchasing of materials and equipment.
* Worked at handling the inventory management by maintaining good safety stock levels.
* Created Auto invoice reports (shipment and backlog reports) and Yearly IT budget reports using Pivot Tables and Slicers in MS Excel by connecting to SQL server database.
* Wrote Python modules to extract/load asset data from the MySQL source database.
* Wrote and executed various MYSQL database queries from Python using Python -MySQL connector and MySQL database package.
* Utilized data profiling and data quality tools to identify and resolve data anomalies, inconsistencies, and outliers in property and casualty insurance data sets.
* Implemented data governance practices and data lineage tracking to ensure compliance with regulatory requirements and support auditability of property and casualty insurance data.
* Collaborated with data scientists and actuaries to support advanced analytics initiatives, leveraging property and casualty insurance data for predictive modeling, risk analysis, and underwriting optimization.
* Implemented data visualization solutions using tools like Tableau or Power BI to create interactive dashboards and reports, enabling stakeholders to gain insights from property and casualty insurance data.
* Understood the requirements and clarifying the conflicts with the business team.
* Analyzed the existing system, business requirements and functional Specifications.
* Followed coding standards, code versioning and quality process techniques to reduce rework from reviews.
* Tested the application, fixing the defects, and documenting the required information.
* Ensured all quality related activities are logged and shared.
* Communicated the work progress to the onshore team.
* Participated in design discussions and assured functional specifications are delivered in all phases of SDLC in an Agile Environment

**Environment:** SSIS Packages, MS Excel, SQL Server, MySQL, Python, Agile