**Vikram N**

**Sr. Data Engineer**

**PROFESSIONAL SUMMARY**

* I have accumulated a total of **9 Years** of professional experience in IT, encompassing a wide range of responsibilities such as analysis, design, development, documentation, deployment, and integration. My expertise lies in SQL and Big Data technologies.
* Throughout my career, I have successfully implemented various solutions, including **Big Data analytics**, **cloud data engineering**, **data warehousing**, **data marts**, **data visualization**, **reporting**, **data qualityassurance**, and **data virtualization**.
* I have a proven track record of serving as a Data Engineer on **Amazon Web Services** (AWS) cloud platforms, as well as working with **Big Data** and **Hadoop applications** and product development. Additionally, I am proficient in coding with **Python** and possess a strong foundation in computer science principles, including data structure and algorithm design.
* I am experienced in conceptualizing, designing, and implementing the entire data modeling lifecycle, including conceptual, logical, and physical data modeling, using tools like **Erwin** and **E/R Studio**.
* My expertise extends to integrating data from diverse sources, including **SQL Server**, **Oracle**, **Teradata SQL Assistant**, **MySQL**, **flat files**, **XML**, and **XSDs**.
* I am well-versed in utilizing **Teradata tools** and utilities such as **BTEQ**, **Fast Load,Multiload, FastExport**, and **TPUMP**.
* I have a thorough understanding of AWS services, particularly **AWS Glue cloud services**, including **EC2**, **S3**, **Glue**, **Athena**, **DynamoDB**, and **RedShift**. Moreover, I am skilled in job/workflow scheduling and monitoring tools like **Oozie**, **AWS Data Pipeline**, and **Autosys**.
* I have expertise in setting up monitoring, metrics, and logging systems on AWS and have experience in creating and managing Docker images with multiple microservices, along with container orchestration using **ECS**, **ALB**, and **Lambda**.
* I possess substantial experience in deploying, managing, and developing with **MongoDB clusters**, and I am highly proficient in **JAVA/J2EE**, **Oracle**, and **MySQL** technologies. I am experienced in all phases of the software development lifecycle, including analysis, design, development, and testing.
* I am an expert in building both supervised and unsupervised machine learning experiments using Microsoft Azure, utilizing various algorithms for predictive analytics and developing web service models for diverse data types.
* My machine learning skills include Linear & Logistic Regression, Classification Modeling, Decision Trees, **Principal Component Analysis (PCA),** and Cluster and Segmentation analyses, with a history of contributing to scholarly articles using these techniques.
* I have hands-on experience with **Data Warehousing ETL** using **Informatica,IBM Infosphere/Websphere DataStage**, and **Ascential DataStage**, covering various stages such as Transformer, Aggregator, Lookup, Join, Sort, Copy, Merge, Funnel, CDC, Change Apply, and Filter.
* Additionally, I have extensive experience with parallel stages like **Datasets**, **Change Data Capture,Row Generator**, and other stages for ETL coding. I have developed automated scripts for **Data Ingestion (DI)** and **Data Loading (DL)** using **Python** and **Java MapReduce**.
* My experience includes Unix/Linux systems with scripting expertise, allowing me to build efficient data pipelines. I have also been responsible for migrating on-premises applications to the Azure cloud.
* I am skilled in detailed system design, employing use case analysis, functional analysis, and **UML modeling techniques** with class and sequence diagrams, activity diagrams, and state diagrams using tools like Rational Rose.
* I have created ETL pipelines using **Hive** and **Spark** for data ingestion from multiple sources and utilized **Hive SQL**, **Presto SQL**, and **Spark SQL** for **ETL processes**, ensuring the right technology is applied to achieve optimal results.
* My experience extends to working with **Cloud Databases** and **Data Warehouses**, particularly SQL Azure and Confidential Redshift/RDS.
* Comprehensive understanding of data governance principles and compliance with data privacy regulations such as GDPR and CCPA.
* Committed to staying updated with emerging technologies and best practices in data engineering and data analysis to drive innovation and efficiency.
* Proficient in working with various databases including **MongoDB**, **Cassandra**, **MySQL**, **ORACLE**, and **MS SQL Server**.
* Played a pivotal role in migrating **Cassandra** and **Hadoop** clusters to **AWS Glue**, defining different read/write strategies.
* Demonstrated expertise in **SQL** and **PL/SQL tuning** as well as application tuning, utilizing tools such as **EXPLAIN PLAN**, **SQL TRACE**, **TKPROF**, **and AUTOTRACE**.
* Hands-on experience with a wide range of **AWS services**, including **EC2**, **S3**, **ELB**, **RDS**, **SQS**, **EBS**, **VPC**, **AMI**, **SNS**, **CloudWatch**, **CloudTrail**, **CloudFormation**, **GCP Config**, **Autoscaling**, **CloudFront**, IAM, and **Route 53**.
* Utilized Chef for configuration management of hosted instances within **Google Cloud Platform** (GCP) and handled configuring and networking of **Virtual Private Cloud (VPC).**
* Proficient in working with event-driven and scheduled AWS Lambda functions to trigger various AWS resources.
* Integrated Lambda with **SQS** and **DynamoDB** using step functions to process lists of messages and update status information in **DynamoDB** tables.
* Possesses strong **SQL** development skills, including the ability to write Stored Procedures, Triggers, Views, and User-Defined Functions.
* Designed and developed Data Warehouses using both Star Schema and Snowflake Schema, depending on specific business requirements.
* Expertise in creating **SSIS/DTS** Packages for extracting, transforming, and loading (ETL) data into data warehouses/data marts from heterogeneous sources.
* Responsible for defining rules and workflows to govern the creation, modification, storage, and deletion of data in compliance with the Master Data Governance program.
* Created **ETL pipelines** into and out of data warehouses using a combination of Python and Snowflake's SnowSQL, as well as writing **SQL queries** against **Snowflake**.
* Possesses a solid understanding of software development methodologies, including Agile (Scrum).
* Proficient in developing various reports and dashboards using a range of Tableau visualizations.
* Experienced in multiple programming languages, including Java, Python, R, and SAS.
* Well-versed in using various components of the Hadoop ecosystem, including **HDFS,YARN**, **MapReduce**, **Spark**, **Pig**, **Sqoop**, **Hive**, **Impala**, **Hbase**, **Kafka,** and **Crontab** tools.
* Expert in creating **HIVE UDFs** using Java to analyze data sets for complex aggregate requirements.
* Proficient in designing, developing, testing, implementing, and supporting Data Warehousing ETL processes using Talend.
* Involved in the development and creation of **PL/SQL** stored procedures and functions.
* Extensive experience in developing **ETL applications** for processing large volumes of data using tools such as **MapReduce**, **Spark-Scala**, **PySpark**, **Spark-SQL**, and **Pig**.
* Skilled in using **SQOOP** for importing and exporting data between **RDBMS** and **HDFS/Hive**.
* Familiar with **MS SQL Server**, including **SSRS** (SQL Server Reporting Services), **SSIS** (SQL Server Integration Services), and **T-SQL**.

**TECHNICAL SKILLS**

**Big Data Tools**: Hadoop Ecosystem: Map Reduce, Spark 2.3, Airflow 1.10.8, Nifi 2, HBase 1.2, Hive 2.3, Pig 0.17 Sqoop 1.4, Kafka 1.0.1, Oozie 4.3, Hadoop 3.0

**BI Tools**: SSIS, SSRS, SSAS. Alation Data Cataloging

**Data Modeling Tools**: Erwin Data Modeler, ER Studio v17

**Programming Languages**: SQL, PL/SQL, and UNIX.

**Methodologies:** RAD, JAD, System Development Life Cycle (SDLC), Agile

**Cloud Platform:** AWS, Azure, Google Cloud.

**Cloud Management:** Amazon Web Services (AWS)- EC2, EMR, S3, Redshift, EMR, Lambda, Athena

**Databases:** Oracle, Teradata R15/R14.

**OLAP Tools:** Tableau, SSAS, Business Objects, and Crystal Reports 9

**ETL/Data warehouse Tools:** Informatica 9.6/9.1, and Tableau.

**Operating System:** Windows, Unix, Sun Solaris

**PROFESSIONAL EXPERIENCE:**

**Sr. Data Engineer**

**Santander Bank – Denver, Colorado Dec 2020 to Present**

**Responsibilities:**

* Attended daily scrum meetings and sprint planning sessions as part of the Scrum process. Collaborated with business teams to gather requirements. Worked closely with development teams on project deliverables. Contributed to finalizing technical design with various teams.
* Designed a new data architecture from the ground up.
* Implemented Bitbucket repositories in Bamboo for code deployment.
* Utilized serverless architecture, including API Gateway, Lambda, and DynamoDB, deploying AWS Lambda code from Amazon S3 buckets and configuring a Lambda Deployment function to receive S3 bucket events.
* Designed data models for data-intensive AWS Lambda applications, focused on complex analysis and generating analytical reports for end-to-end traceability and key business element definitions from Aurora.
* Created a secure, on-demand EMR launcher with custom Spark submit steps using S3 Events, SNS, KMS, and Lambda functions.
* Developed Python scripts for invoking Lambda functions using Secrets Manager in AWS to build a pipeline.
* Built a pipeline using Lambda to extract data from PostgreSQL and DynamoDB into an AWS S3 bucket.
* Employed CloudFormation templates to create an S3 bucket with subfolders.
* Leveraged AWS Glue catalog with a crawler to retrieve data from S3 and perform SQL query operations.
* Utilized various Python libraries, including Pandas, NumPy, Seaborn, SciPy, Matplotlib, Scikit-learn, and NLTK for developing machine learning algorithms.
* Worked with diverse data formats such as JSON and XML for machine learning tasks.
* Managed Lambda functions and Glue Jobs, EC2 hosts by setting up monitors, alarms, notifications, and logs using CloudWatch.
* Performed data transformation, validation, and cleansing using AWS Glue.
* Established storage and data analysis tools in Amazon Web Services cloud infrastructure.
* Developed a Python utility using packages like SciPy, NumPy, and Pandas.
* Implemented supervised classification algorithms like Logistic Regression, Decision Trees, KNN, and Naive Bayes. Analyzed data from various services, including Service Now, Snow, Splunk, and flight tower, using Amazon Comprehend.
* Applied Topic Modeling with two-column input and provided accuracy predictions for models.
* Conducted training and testing accuracy assessments with Topic Modeling in Amazon Comprehend for large daily data volumes.
* Created Hive UDFs and Pig UDFs in Microsoft HDInsight environments using Python.
* Developed a Python wrapper for multi-threaded applications.
* Established a production bucket in AWS Sage Maker and ran machine learning algorithms using Python scripts to predict model accuracy.
* Produced bar graphs and Pareto graphs for Amazon Comprehend models in AWS Sage Maker for a client feedback system.
* Implemented a Recommender system using AWS Sage Maker with machine learning algorithms.
* Designed monitoring and trace/track dashboards in Splunk and CloudWatch using data metrics.
* Demonstrated active participation in multiple assigned projects, excelling at multitasking.
* Created UNIX shell scripts for database connectivity and parallel job execution.
* Utilized SQL \*Loader and TOAD's import utility for data warehouse table population.
* Demonstrated proficiency in mapping source-to-target data using IBM Data Stage 8.x.
* Implemented multi-node configuration for performance enhancement.
* Performed performance tuning on Targets, Sources, Mappings, and Sessions for optimal efficiency.

**Environment**: AWS Services, Hadoop, Map Reduce, Hive, Pig, Flume, Sqoop, Spark, Teradata, Parquet, Oracle 12c, SQL Plus, TOAD, SQL Loader, SQL Developer, PL/SQL, Informatica Power Center 10.2.0, Designer, Workflow Manager, Workflow Monitor, Repository Manager, Kafka, GitHub, Shell Scripts, UNIX, Windows XP, Splunk.

**Sr. Data Engineer**

**CSC Service works – Irving, Texas August 2018 – Nov 2020**

**Responsibilities:**

* Created migration scripts and indexing strategy for Redshift from SQL Server and MySQL.
* Enhanced legacy systems for Spark and Hadoop on Azure Cloud.
* Used Pig for ETL tasks like data transformations and event joins before storing in HDFS.
* Installed Apache Airflow, developed Python workflows for automation. Crafted marketing materials for data governance.
* Configured GCP servers, including EC2, RDS, security groups, and load balancers.
* Developed Python test framework.
* Assisted in ERP implementations, focusing on master data.
* Developed Python ETL processes for Hadoop and SQL data warehouse.
* Conducted relational and dimensional data modeling using ER Studio.
* Managed databases and storage in GCP, including S3 backups.
* Utilized Teradata utilities for data loading. Proposed process improvements, integrated Airflow, and migrated to Azure Cloud.
* Designed MongoDB sharding and indexing strategies.
* Worked with Oracle, RedShift, and Snowflakes.
* Developed modules using Java, JSP, Casper, and Struts.
* Optimized Pig scripts and used Hive for data analysis.
* Managed data import/export with XML and Ziplined and designed databases based on business requirements.
* Oversaw on-premises application migration to Azure.
* Used SSIS for multi-dimensional cube creation.
* Implemented indexing and data distribution strategies for fast query response.
* Built a statistical model with neural networks for student ranking. Leveraged Informatica and various transformations for mapping development.
* Worked with Hadoop, Hive, Java, Python, Scala, and Struts.
* Utilized Core Java and J2EE for web-based solutions.
* Expertise in Business Intelligence, data extraction, and visualization.
* Used Power BI and SSRS for report generation and visualization.
* Designed Data Marts with Star and Snowflake Schemas.
* Created Talend jobs for file transfers and managed mapping documents.
* Implemented Spark, Hive, and ETL for data ingestion and cleaning.
* Employed robotic process automation in Python for data preparation.
* Used SQL, Presto SQL, and Spark SQL for ETL tasks.
* Prepared and uploaded SSRS reports with database management.
* Developed SQL queries for SSRS and Power BI.
* Built analytical dashboards for student record tracking.
* Worked with deep learning frameworks like MX Net, Caffe 2, TensorFlow, and more.
* Demonstrated programming skills in Python, C/C++, Java, and distributed systems.
* Participated in requirements meetings and data mapping sessions.

**Environment**: Data security, GCP, MDM, Kafka, ER Studio, Airflow, AWS, OLTP, Teradata, Informatica, Sqoop, MongoDB, MySQL, HDFS, Linux, Shell, scripts, SSIS, SSAS, HBase, Azure, MDM.

**Data Engineer/SSRS Developer**

**UKG – Miami, Florida May 2017 to July 2018**

**Responsibilities:**

* designed and constructed a multi-terabyte Data Warehouse infrastructure entirely from the ground up on Confidential Redshift, geared for handling millions of daily records at a large scale.
* My responsibilities included Extracting, Transforming, and Loading data from Source Systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL, and U-SQL in Azure Data Lake Analytics. This data was then ingested into one or more Azure Services, such as Azure Data Lake, Azure Storage, Azure SQL, and Azure DW, and subsequently processed within Azure Databricks.
* Within Azure Data Factory (ADF), I crafted Pipelines employing Linked Services, Datasets, and Pipelines to effectively Extract, Transform, and Load data from various sources, including Azure SQL, Blob storage, Azure SQL Data warehouse, as well as performing data write-back tasks.
* I developed Spark applications using PySpark and Spark-SQL to carry out data extraction, transformation, and aggregation from diverse file formats. This analytical approach allowed me to uncover valuable insights into customer usage patterns.
* I took charge of estimating cluster sizes, monitoring, and troubleshooting the Spark Databricks cluster. My experience encompassed fine-tuning Spark Applications by determining appropriate Batch Interval times, configuring the correct level of Parallelism, and optimizing memory usage.
* My work extended to the realm of Big Data Integration and Analytics, utilizing technologies such as Hadoop, SOLR, Spark, Kafka, Storm, and web Methods.
* I actively defined requirements for ensuring data quality, establishing data lineage, reinforcing data security, maintaining data privacy, and addressing other pertinent aspects related to business terms and data assets.
* Initiating improvements in data management was a key aspect of my role. I set guidelines and standards for data quality, data security, data privacy, and data issue remediation.
* To meet specific business needs, I authored UDFs (User-Defined Functions) using Scala and PySpark.
* For streamlined data processing, I developed JSON Scripts that facilitated the deployment of Pipelines in Azure Data Factory (ADF) through the utilization of the SQL Activity.
* Automation was a significant focus, encompassing the deployment and management of cloud infrastructure using Chef and Python.
* My expertise spanned Java and Python programming languages.
* I created SSRS (SQL Server Reporting Services) reports and SSIS (SQL Server Integration Services) packages for extracting, transforming, and loading data from a variety of source systems.
* I managed the implementation and oversight of ETL (Extract, Transform, Load) solutions while also automating operational processes.
* My role involved optimizing and fine-tuning the Redshift environment to enable queries to deliver up to 100 times faster performance, particularly benefiting tools like Tableau and SAS Visual Analytics.
* I effectively utilized Teradata utilities such as Fast load, Multiload, and Tpump for data loading.
* I crafted, tested, and implemented Teradata Fast load, Multiload, and Bteq scripts, encompassing tasks related to DML (Data Manipulation Language) and DDL (Data Definition Language).
* Tuning and optimizing complex SQL queries was part of my responsibilities, making use of Teradata Explain for analysis.
* My role also encompassed defining facts and dimensions and designing data marts following Ralph Kimball's Dimensional Data Mart modeling methodology, using Erwin.
* My responsibilities included creating Entity Relationship Diagrams (ERD), Functional diagrams, Data flow diagrams, and enforcing referential integrity constraints, as well as crafting logical and physical models using Erwin.
* I had extensive experience working with various databases, including Teradata, PostgreSQL, H2 Database, Oracle, MS SQL Server, AS400, DB2, and MS Access. I also utilized SQL Editors like Teradata SQL Assistant, H2 Console, TOAD, SQL PLUS, and SQL Analyzer.
* I was involved in implementing the CI/CD (Continuous Integration/Continuous Deployment) process for Apache binary and DGA's (Data Governance Applications). This involved building Airflow Docker images through Docker Compose and uploading them to JFrog Artifactory, followed by deployment through AWS ECS cluster.
* My experience included integrating various data sources, ranging from Teradata and Oracle to PostgreSQL, S3, HDFS, MS SQL Server, Flat Files, and XML Definitions.
* I focused on enhancing the performance of Airflow DAGs (Directed Acyclic Graphs) and task instances.
* I played a role in configuring Airflow schedulers and worker settings within the airflow’s file.
* Ad hoc queries and reports were created by me to support business decisions through SQL Server Reporting Services (SSRS).
* I was responsible for developing complex ETL Packages using SQL Server 2008 Integration Services, facilitating data extraction from various sources like Oracle, SQL Server, and DB2 into a Staging Database and onward to the Data Warehouse.
* I conducted thorough analysis of existing application programs and fine-tuned SQL queries using tools such as execution plans, query analyzers, SQL Profiler, and the database engine tuning advisor to optimize performance.
* I conducted source data analysis using Informatica PowerCenter Designer, extracting and transforming data from various source systems, including Oracle 10g, DB2, SQL server, and flat files, while applying business rules using different objects and functions supported by the tool.
* I implemented multiple Data pipeline DAGs and Maintenance DAGs within the Airflow orchestration framework.
* I documented Informatica mappings in Excel spreadsheets.
* Tuning Informatica mappings to ensure optimal load performance was a part of my responsibilities.
* I provided expertise in business and data analysis, driving changes in data governance where necessary.
* My involvement extended to Forward Engineering of logical models, generating physical models using Erwin, and deploying Data Models using ERWin for subsequent integration into the Enterprise Data Warehouse.

**Environment**: Data Governance, Airflow, MDM, Kafka, Talend, SQL Server, Erwin, Oracle, Redshift, Informatica, RDS, NOSQL, Snow Flake Schema, MySQL, PostgreSQL.

**Data Analyst**

**SCL Health Corp - Atlanta, Georgia Jan 2015 – May 2017**

**Responsibilities:**

* Participated in all phases including requirement analysis, client interaction, Design, Coding, Testing, support and Documentation.
* Created dimensions and facts in the physical data model.
* Installed and configured Data stage client and server tools. Used the data stage designer to develop processes for extracting, cleansing, transforming integrating, and to load the data like FLAT Files, ORACLE, and DB2 into database.
* Extensively used SQL and PL/SQL coding in Data stage jobs for processing data.
* Developed batches and sequencers in designer to run and control set of jobs.
* Used the data stage director and its run-time engine to schedule running the job, testing and debugging its components, and monitoring the resulting executable versions.
* Used data stage manager for importing metadata from repository, new job categories and creating new data elements.
* Used parallel extender for distributing load among different processors by implementing pipe line and partitioning of data in parallel extender.
* Used quality stage to access the metadata server to obtain live access to current metadata for enterprise data.
* Used quality stage to access data generated by web sphere information analyzer.
* Used partition methods and collecting methods for implementing parallel processing.
* Used DS administrator to configure, create and maintain data stage jobs.

**Environment:** IBM data stage 11.3, Data stage director, PL/SQL, Oracle 12x, UDB/UNIX, Windows.

**ETL/ SQL Developer**

**Intelenet Global Services - Hyderabad, India June 2012 – July 2013**

**Responsibilities:**

* Involved in gathering, understanding and documenting the data mapping requirements from the user through daily calls and continuous follow-ups.
* Involved in designing and building of High level and low-level ETL technical design documents.
* Develop complex ETL mappings and packages\procedures\functions to extract data from the flat files and load it into the Oracle database.
* Error validation of the data moving from flat file to the Oracle database.
* Involved in the Unit testing, self- review and peer review.
* Work with team to achieve timely resolution of all production issues meeting or exceeding Service Level Agreements.
* Used Bulk Collections for better performance and easy retrieval of data, by reducing context switching between SQL and PL/SQL engines.
* Extensively participated in translating business needs into Business Intelligence reporting solutions by ensuring the correct selection of toolset available across the Tableau BI suite.
* Conduct code review to ensure the work delivered by the team is of high-quality standards.
* Maintain relationships with assigned customers post integration support their needs and build the relationship to encourage future growth of business with the customer.
* Used shell scripts and PMCMD commands to conduct basic ETL functionalities.

**Environment**: Informatica, Oracle 9i, Toad, UNIX KSH, Tortoise SV.

**Education:**

Bachelors - Computer Science Teegala Krishna Reddy Engineering College, Hyderabad, India Sept 2008 - May 2012

Masters - Management Science Lewis University, Illinois, USA Aug 2013 - Dec 2014