**Pavan Kumar Tammineedi**

**Sr. Data Engineer**

**Mobile: 469 300 4907**

**victor@narveetech.com**

**Professional Summary:**

* With approximately 10 years of expertise in the fields of cloud computing, data warehousing, data engineering, feature engineering, Hadoop big data, ETL/ELT, and business intelligence, I specialize as a Cloud Data Architect and Engineer.
* Experienced in managing and implementing data lakes and data warehouses, particularly within the AWS, Snowflake ecosystem. Proven expertise in leveraging AWS services to design, develop, and optimize data storage and analytics solutions. Skilled in ensuring data integrity, scalability, and efficient data processing.
* Developing ETL pipelines in and out of data warehouses using a combination of Python and Snowflakes Snow SQL Writing SQL queries against Snowflake.
* Proficient in the Hadoop ecosystem, cloud-based data engineering on AWS, data visualization, reporting, and ensuring data quality solutions.
* Extensive experience with Amazon Web Services, utilizing tools such as S3, IAM, EC2, EMR, Kinesis, VPC, DynamoDB, Redshift, Amazon RDS, Lambda, Athena, Glue, DMS, QuickSight, Amazon Elastic Load Balancing, Auto Scaling, CloudWatch, SNS, and SQS among other AWS offerings.
* Experience with different ETL tool environments like SSIS, Informatica, and reporting tool environments like SQL Server Reporting Services, and Business Objects.
* Skilled in applying data analytics services including Athena, Glue, Data Catalog, and QuickSight to real-world data challenges.
* Managing Database, Azure Data Platform services (Azure Data Lake (ADLS), talendtory (ADF), Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Data bricks, NoSQL DB), SQL Server, Oracle, Data Warehouse etc. Build multiple Data Lakes
* Demonstrated expertise in managing AWS databases, including RDS (Aurora), Redshift, DynamoDB, and Elastic Cache (using both Memcached and Redis).
* Developed Hadoop-based applications leveraging technologies like HDFS, MapReduce, Spark, Hive, Sqoop, HBase, and Oozie.
* Skilled in architecting and executing legacy data migration projects from on-premises systems to the AWS cloud.
* Authored and deployed Python-based AWS Lambda functions to facilitate complex data transformations and analytics on large datasets within EMR clusters.
* Engineered, designed, and implemented robust AWS solutions for a variety of applications, ensuring high availability, fault tolerance, and scalability across services including EC2, Glue, Lambda, SNS, S3, RDS, CloudWatch, SQS, and IAM.
* Conducted migrations of Cassandra and Hadoop clusters to AWS, establishing tailored read/write strategies to optimize data access across different geographic regions.
* Experience in building and optimizing AWS data pipelines, architectures, and data sets.
* Extensive experience in agile software development methodology.
* Proficient in programming with Python and Java, demonstrating strong capabilities in developing complex, efficient, and scalable software solutions. Skilled in leveraging the advanced features of both languages to create high-performance applications.
* Proficient in automating and managing cloud resources using AWS CloudFormation templates and the AWS Cloud Development Kit (CDK).
* Very capable at using Amazon Web Services utilities such as EMR, S3 and CloudWatch to run and monitor Hadoop/Spark jobs on AWS.
* Written PySpark job in AWS Glue to merge data from multiple tables and in Utilizing Crawler to populate AWS Glue data Catalog with metadata table definitions.
* Generated a script in AWS Glue to transfer the data and utilized AWS Glue to run ETL jobs and run aggregation on PySpark code.
* Implemented and managed AWS Storage Gateway for efficient and secure data integration, facilitating seamless hybrid storage between on-premises environments and AWS cloud.
* Proficient in the management and optimization of Amazon Neptune, specializing in configuration, performance tuning, and efficient data handling for enhanced database performance and reliability.
* Skilled in leveraging AWS services like Data Pipeline, Glue, and DMS (Database Migration Service) to facilitate smooth, reliable, and efficient data migration, ensuring data integrity and optimizing for graph database performance.
* Worked on different file formats like JSON, XML, CSV, ORC, Paraquet. Experience in processing both structured and semi structured Data with the given file formats.
* Worked on Apache Spark performing the Actions, Transformations on RDDs, Data Frames & Datasets using spark SQL and Spark streaming contexts.
* Creating Pipelines in ADF using Linked Services/Datasets/Pipeline/ to Extract, Transform, and load data from different sources like Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool and backwards.
* Files extracted from Hadoop and dropped on daily hourly basis into S3. Working with Data governance and Data quality to design various models and processes.
* Experience managing Azure Data Lakes (ADLS) and Data Lake Analytics and an understanding of how to integrate with other Azure Services. Knowledge of USQL
* Worked on Java EE 7 and 8. Developed ETL\Hadoop related java codes, created RESTful APIs using Spring Boot Framework, developed web apps using Spring MVC and JavaScript, developed coding framework, etc.
* Having good experience in spark core, spark SQL and spark streaming.
* Having good experience in writing Python Lambda functions and calling the API’s.
* Hands on experience with Confidential components like EC2, EMR, S3, and Elastic Search.
* Good knowledge in Kafka and Flume.
* Experience in Java, Java EE (J2ee) technologies and proficiency in Core Java, Servlets, JSP, EJB, JDBC, XML, and spring, Struts and Hibernate and RESTful Webservices.
* Well-versed in relational database management systems (RDBMS) including Oracle, MS SQL Server, MYSQL.
* Successfully leveraged the ETL tool Matillion to efficiently migrate data from AWS DynamoDB to Redshift.
* Familiar with Git for version control, adept at managing code repositories, and proficient in using branches and merging for collaborative development.
* Quick learner and vibrant team player with ability to work independently under pressure and meet demanding deadlines.
* Agile and Waterfall techniques are well-understood. Agile scrum review and planning sessions are executed.
* Strong analytical and problem-solving ability, as well as troubleshooting talents. Excellent technical and user communication and interpersonal abilities.

**Skills Summary:**

|  |  |
| --- | --- |
| **Cloud Skills** | Snowflake, Storage Gateway, S3, EC2, Lambda, Kinesis, CloudWatch, Step function, IAM, Glue, Athena, SNS, DynamoDB, Redshift, Sage Maker, EMR, Neptune |
| **Programming Languages** | C, Java, Python, R, JavaScript, TypeScript, Linux, Shell scripting. |
| **Database**  | MySQL, SQL Server, DynamoDB, MongoDB,  |
| **Hadoop Components** | HDFS, MapReduce, Spark, Azure, Hive, Sqoop, HBase, and Oozie. |
| **Reporting and ETL Tools** | AWS CloudWatch Dashboards, AWS Glue, Azure, SQL Server Reporting Services, Informatica, Tableau, Matillion. |
| **CI/CD Tools** | Git, Cloud Formation, Cloud Development Kit (CDK).  |
| **Methodologies** | Data Modeling, OOP, CI/CD, Agile/Scrum, Waterfall.  |
| **Operating Systems** | Windows, Linux, Ubuntu. |
| **Development Tools** | IntelliJ, Visual Studio, Eclipse, Maven. |

**Work Experience:**

**Amazon Web Services |**Data Engineer **|** Boston, MA | Oct 2022 – May 2024

**Responsibilities:**

* Ensured seamless cloud backups and archival processes on the AWS Storage Gateway team.
* Developed Java code to manage and mitigate S3 throttling, enhancing data transfer efficiency during high-traffic periods.
* Published key metrics and logs to AWS CloudWatch to improve system monitoring and transparency.
* Used CloudWatch analytics to detect trends and anomalies for proactive system troubleshooting and enhanced reliability.
* Developed sophisticated Python-based dashboards for real-time insights via AWS CloudWatch, enhancing operational decision-making.
* Hosted Python dashboards on a Wiki platform to facilitate easy access and promote collaborative environment.
* Managing Database, Azure Data Platform services (Azure Data Lake (ADLS), talendtory (ADF), Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Data bricks, NoSQL DB), SQL Server, Oracle, Data Warehouse etc. Build multiple Data Lakes
* Led the team's transition to a serverless architecture, optimizing scalability and cost efficiency using AWS serverless components.
* Authored infrastructure code with AWS Cloud Development Kit (CDK) in TypeScript, applying infrastructure as code (IaC) principles.
* Created CloudFormation templates to automate and manage CI/CD pipelines efficiently within AWS.
* Managed serverless functions and orchestrated workflows with AWS Lambda and Step Functions, improving application efficiency.
* Used AWS IAM to manage secure access and permissions for AWS Step Functions and other services.
* Orchestrated complex workflows with AWS Step Functions, integrating services like DynamoDB, Lambda, SNS, and Athena for efficient solutions.
* Set up an AWS Lambda function that runs every 15 minutes to check for repository changes and publishes a notification to an Amazon SNS topic.
* Employed Gremlin and SPARQL for graph database queries, demonstrating proficiency in graph data structures and query optimization.
* Integrated graph databases with AWS services like Amazon Neptune, AWS Lambda, and AWS Glue for robust data management solutions.
* Spearheaded the adoption of modern monitoring practices by implementing automated alerting and response systems using AWS Lambda and Amazon SNS, reducing system downtime, and improving response efficiency.
* Enhanced team productivity and project agility by introducing DevOps best practices, which included automated testing and deployment strategies using AWS Code Pipeline and AWS Code Build, thus ensuring faster delivery cycles and higher quality outputs.
* These points emphasize further advancements in system monitoring, operational efficiency, and deployment processes, reinforcing a commitment to continuous improvement and technological leadership.
* Developed Python-based modules for machine learning and predictive analytics in a Hadoop environment on AWS, including distributed algorithms.
* Provided on-call support, ensuring high system availability and reliability.
* Managed code versions and collaborated using Git for efficient team development.
* Actively engaged in Agile methodologies to drive team planning and project management.

**Environment:**  Java, Python, Storage Gateway, CloudWatch, DynamoDB, S3, SNS, Lambda, Step Function, Athena, CloudFormation, CDK, IAM, Git.

**Plymouth Rock Assurance (Contract)|** Sr. Data Engineer | Boston, MA | Sep 2019 – Oct 2022

**Responsibilities:**

* Designed and implemented cloud-native applications on AWS, focusing on optimal performance and cost efficiency.
* Managed, wrote, and maintained ETL processes feeding into all management dashboards and interfaces.
* Developed an interactive dashboard application using Python, PySpark, and MySQL for dynamic data visualizations, enhancing accessibility for sales and management teams.
* Engineered a high-availability ad server on AWS using EC2, Elastic Load Balancing, and Autoscaling Groups, handling over a million requests daily.
* Designed and developed ETL processes in AWS Glue to transfer campaign data from formats like ORC, Parquet, and text files into AWS Redshift and S3.
* Utilized AWS EMR to efficiently transfer large volumes of data into and out of AWS data stores such as Amazon S3 and DynamoDB.
* Architected and developed an automated ETL pipeline for regular ingestion of partner data, utilizing REST APIs and screen scraping as needed.
* Executed data extraction, aggregation, and consolidation within AWS Glue using PySpark, specifically for Adobe data.
* Developed frameworks and processes to analyze unstructured information. Assisted in Azure Power BI architecture design
* Designing and Developing Oracle PL/SQL and Shell Scripts, Data Import/Export, Data Conversions and Data Cleansing.
* Architect & implement medium to large scale BI solutions on Azure using Azure Data Platform services (Azure Data Lake, Data Factory, Data Lake Analytics, Stream Analytics
* Developed a real-time data pipeline processing thousands of data points per second with AWS Kinesis, Apache Zookeeper, and Storm, using Redis for data storage and real-time display on dashboards.
* Managed a wide array of AWS services including VPC, Lambda, API Gateway, Directory Service, EBS, ELB, RDS, IAM, Route53, CloudFront, Kinesis, Kafka, Kubernetes, DynamoDB, CloudWatch, SQS, SNS, along with Terraform and Docker.
* Collaborated with business users to gather and analyze requirements and designed tailored solutions.
* Extensively utilized SQL, PL/SQL, and Oracle Database tools including Import/Export, SQL\*Loader, and SQL\*PLUS.
* Engaged in business analysis, gathered ETL requirements, and contributed to both physical and logical data modeling and documentation.
* Analyzed SQL scripts and engineered solutions for implementation with PySpark.
* Used Sqoop to import data from various sources into HDFS.
* Managed data importing, transformation with Hive and MapReduce, and loading into HDFS.
* Transferred data from MySQL to HDFS using Sqoop.
* Automated deployment processes using YAML scripts for large-scale builds and releases.
* Utilized a suite of technologies including Apache Hive, Pig, HBase, Spark, Zookeeper, Flume, Kafka, and Sqoop for data management tasks.
* Worked on AWS Data Pipeline to configure data loads from S3 to into Redshift.
* Used JSON schema to define table and column mapping from S3 data to Redshift.
* Leveraged AWS Glue to catalog ingested data, making it easy to discover and access datasets for analysis.
* Successfully migrated on-premises Informatica ETL processes to AWS cloud and Snowflake, harnessing the power of AWS services for data processing and storage in a scalable environment.
* Utilized PL/SQL for data validation and cleansing within the ETL process, ensuring data accuracy and integrity during extraction and transformation phases.
* Created and maintained PL/SQL stored procedures within the Oracle database to handle complex data transformations and business logic.

**Environment:**  Python, MySQL, PySpark, EC2, Glue, RedShift, S3, DynamoDB, PL/SQL, Git, CloudWatch, Hive, HBase, Spark, Kafka, Sqoop.

**State of Virginia (Contract)|** Data Engineer | Virginia | Jan 2017 – Sep 2019

**Responsibilities:**

* Served as a Big Data Engineer, resolving issues on the firm's proprietary platform, analyzing data, and developing enhancements.
* Worked with large datasets, particularly security logs, for analysis and processing.
* Loaded data from relational databases to HIVE using Sqoop and handled various flat files in text, EDI, and XML formats from different vendors.
* Developed MapReduce jobs for data cleaning and manipulation.
* Conducted Cassandra migration from Oracle and SQL Server RDBMS to Hadoop for enhanced data processing.
* Developed and managed ETL processes in Snowflake, including data extraction, transformation, and loading from various sources into Snowflake data warehouses.
* Collaborated with stakeholders to gather requirements and ensure the optimal design and implementation of Snowflake data pipelines.
* Designed and implemented Snowflake stages, pipelines, and tasks to facilitate efficient data movement and transformation.
* Transformed Hive/SQL queries into Spark transformations using RDDs, Python, and Scala.
* Employed Flume to aggregate and store web log data from various sources into HDFS.
* Developed Spark Streaming jobs, utilizing Spark RDDs and DataFrame with Spark SQL.
* Wrote Hive queries for data retrieval from multiple tables and managed output via multiple MapReduce jobs.
* Managed star schema in HBase, handling data access with various HBase commands.
* Designed Hive tables, both internal and external, for operational efficiency and implemented partitioning and bucketing for performance enhancement.
* Developed Spark programs in Scala and Python for data quality checks.
* Created tables in S3, SNS, SQS, and HBase for storing various data formats, including PII.
* Optimized Hive with techniques like Map side joins, bucketing, and partitioning for performance and storage improvements.
* Leveraged Hive for analyzing partitioned and bucketed data to compute metrics for reporting.
* Integrated HBase with Spark for data imports and CRUD operations on HBase.
* Developed MapReduce programs to process files from Hive queries, generating key-value pairs for upload to the NoSQL database HBase.
* Leveraged Matillion's cloud-native capabilities for efficient ETL processes, optimizing data transformations directly into Amazon Redshift for enhanced performance and scalability.
* Implemented real-time data processing in Matillion, utilizing CDC and streaming data integrations to ensure continuous data ingestion and up-to-date analytics, driving timely business decisions.

**Environment**: Python, Scala, Hadoop, Agile, Snowflake S3, SNS, SQS, Redshift, Cassandra, Sqoop, HDFS, NoSQL, HBase, Hive, Spark, Matillion.

**Motivity Labs |** Data Engineer | Hyderabad | May 2014 – Sep 2016

**Responsibilities:**

* Developed MapReduce programs to parse raw data, populate staging tables, and store refined data in partitioned tables in the Enterprise Data Warehouse (EDW).
* Created HBase tables to handle large datasets from UNIX, NoSQL, and various portfolios, including structured, semi-structured, and unstructured data.
* Developed Java-based MapReduce jobs for data cleaning and preprocessing.
* Prepared detailed Business Requirement Specifications (BRS) documents to outline project requirements.
* Designed and established an Enterprise Data Lake to support analytics, data processing, storage, and reporting for rapidly changing and voluminous data.
* Maintained high-quality reference data in source systems through cleaning, transformation, and integrity checks, collaborating with stakeholders and solution architects.
* Designed a Security Framework for fine-grained access control to AWS S3 objects using AWS Lambda and DynamoDB.
* Creating Pipelines in ADF using Linked Services/Datasets/Pipeline/ to Extract, Transform, and load data from different sources like Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool and backwards.
* Files extracted from Hadoop and dropped on daily hourly basis into S3. Working with Data governance and Data quality to design various models and processes.
* Experience managing Azure Data Lakes (ADLS) and Data Lake Analytics and an understanding of how to integrate with other Azure Services. Knowledge of USQL
* Configured Kerberos authentication to secure network communications and facilitate new user access to HDFS, Hive, Pig, and MapReduce clusters.
* Developed ETL scripts for ad-hoc data retrieval requests from analytical platforms.
* Configured, deployed, and maintained multi-node development and testing Kafka Clusters.
* Conducted comprehensive architecture and implementation assessments for various AWS services, including Amazon EMR, Redshift, and S3.
* Implemented machine learning algorithms using Python on AWS to predict and automatically suggest product quantities for users, utilizing Kinesis Firehose and S3 data lake.
* Utilized AWS EMR to transfer large data volumes between AWS data stores and databases, such as Amazon S3 and DynamoDB.
* Applied Spark SQL in Scala and Python to transform RDD case classes into schema RDDs.
* Imported data from various sources like HDFS and HBase into Spark RDDs and performed computations using PySpark to produce output responses.
* Demonstrated a strong understanding of database objects and adeptly handled issue triaging.
* Using Tableau, designing dashboard to show operational metrics.
* Reviewed and modified PL/SQL code to meet new development requirements.

**Environment:**  AWS EMR, S3, RDS, Redshift, Lambda, Boto3, DynamoDB, Apache Spark, HBase, Apache Kafka, HIVE, SQOOP, Map Reduce, Apache, Pig, Python, Java, PL/SQL, SSRS.

**Education:**

* Bachelor of Sciences in Electronics and Communications, Jawaharlal Nehru Technological University, Kakinada.