Avinash Varma

AWS Data Engineer

Email: avinashvarma555@gmail.com

Phone: +1 (857)-204-6355

aws certified
Solutions Architect
Associate

SUMMARY:

- Accumulated over 7 years of IT experience spanning Database Development, ETL Development, Data Modeling, Report Development, and Big Data Technologies across diverse domains including Health Care, Pharmaceuticals, Finance, and Retail.
- Proficiently utilized cloud platforms such as Azure and AWS, leveraging services like Data Lake, Databricks, S3, EC2, EMR, RedShift, and Snowflake to architect and manage scalable, secure data storage and processing systems.
- Demonstrated expertise in data warehousing technologies like Amazon Redshift and Snowflake, as well as proficient handling of data storage and retrieval for analytics applications.
- Utilized ETL solutions such as Informatica Intelligent Cloud Services and Talend to streamline data integration and transformation operations.
- Specialized in Data warehousing, Data modeling, Data integration, Data Migration, ETL process, and Business Intelligence, with expertise in SSIS, Informatica ETL, and reporting tools.
- Engineered advanced data ingestion and transformation pipelines, automated ETL processes, and enhanced data quality and reliability using Python, PySpark, and Scala.
- Developed reliable, scalable data processing pipelines utilizing a variety of programming languages and technologies including Python, PySpark, R, Scala, SQL, and shell scripting.
- Proficient in Python OOP concepts and experienced in developing Spark applications using Spark Core, Spark SQL, and Spark Streaming APIs.
- Demonstrated expertise in Big Data technologies such as Hadoop Ecosystem, Spark, Hive, and Kafka for processing massive datasets efficiently.
- Exhibited data visualization and reporting skills using technologies like Power BI, Tableau, and QuickSight to facilitate informed decision-making.
- Proficient in job scheduling and Apache Airflow, enhancing automation and operational efficiency.
- Possessed advanced knowledge of database management systems including Oracle and MS SQL Server, ensuring data integrity, security, and performance.
- Implemented data validation and integrity constraints within Oracle databases using PL/SQL to ensure data accuracy and consistency.
- Designed and implemented NoSQL database solutions to efficiently meet specific data requirements, including document, key-value, or column-family databases.
- Managed and maintained MongoDB databases, ensuring seamless integration with data pipelines, and overseeing day-to-day database operations.
- Experienced in designing and implementing innovative data solutions on AWS, including data lakes and analytics, ensuring scalability and accessibility of data for analytics and decision-making.
- Led significant migration projects, focusing on security and compliance, utilizing encryption mechanisms, and adhering to regulatory standards.
- Spearheaded the development and maintenance of data warehousing solutions, optimizing for performance and scalability, and implementing robust data backup and recovery strategies.
- Collaborated closely with data scientists and business analysts, translating complex data analysis requirements into actionable data models and insights.
- Practiced version control and CI/CD practices using Git, GitLab, Azure DevOps, and Jenkins to ensure code integrity and streamlined deployment processes.

- Managed and optimized CI/CD pipelines, facilitating efficient code integration and deployment across various environments.
- Proficient in Agile project management methodologies, utilizing tools like Jira and Kanban to enhance team productivity and project visibility.
- Leveraged Confluence macros and plugins to enhance documentation functionality and visualization, improving the overall user experience.
- Instrumental in re-architecting and re-platforming legacy data warehouses to modern data platforms on the cloud, employing AWS services and adopting best practices in data governance and quality.

TECHNICAL SKILLS:

Big Data Tools	Kafka, Cassandra, Apache Spark, Spark Streaming, HBase, Impala, HDFS, MapReduce, Hive, Pig, Sqoop, Flume, Oozie, Zookeeper
Hadoop Distribution	Cloudera CDH, Apache, AWS, Horton Works HDP
Programming Languages	SQL, PL/SQL, Python, UNIX, Pyspark, Pig, HiveQL, Scala, Shell Scripting
Spark Components	RDD, Spark SQL, Spark Streaming
Data Modeling Tools	Erwin Data Modeler, ER Studio
Methodologies	RAD, JAD, System Development Life Cycle (SDLC), Agile
Cloud Platform	AWS, Azure, Google Cloud.
Cloud Management	Amazon Web Services (AWS)- EC2, EMR, S3, SNS, SQS, Redshift, EMR, Lambda, Athena
Databases	Oracle, MySQL, DB2
NoSQL Databases	MongoDB, HBase, DynamoDB
OLAP Tools	Tableau, SSAS, Business Objects, and Crystal Reports 9
ETL/Datawarehouse Tools	Informatica, and Tableau.
Build Tools	Maven SBT
Containerization Tools	Kubernetes, Docker, Docker Swarm
Version Control	CVS, SVN, Clear Case, Git
Operating System	Windows, Unix, Sun Solaris

PROFESSIONAL EXPERIENCE:

Mass Mutual, Springfield, MA **Jan 2021 – Present**

AWS Data Engineer

Responsibilities:

- Contributed to Apache Spark data processing project, involving data processing from RDBMS and various data streaming sources, and developed Spark apps in Python on AWS EMR.
- Designed and deployed multi-tier applications on AWS Cloud Formation utilizing AWS services like EC2, Route 53, S3, RDS, DynamoDB, focusing on high availability, fault tolerance, and auto-scaling.
- Connected and integrated various data sources such as S3, Amazon Redshift, RDS, Athena, and third-party databases with Quicksight for data analysis and visualization.
- Implemented ETL pipelines within and outside of a data warehouse using Python and Snowflake's Snow SQL.
- Developed database design and reporting design based on business intelligence and reporting requirements and Utilized AWS EC2 instances to run Spark tasks on AWS Elastic Map Reduce (EMR).
- Leveraged Spark Data Frames, Spark SQL, Spark File formats, and Spark RDDs for data manipulations.
- Used Python scripts in Spark to transform data from various sources (Text, CSV, JSON).
- Loaded data from various sources like RDBMS (MySQL, Teradata) using Sqoop jobs.
- Handled JSON datasets using Spark by creating customized Python functions to parse JSON data.
- Created a preprocessing task to flatten JSON documents into flat files using Spark Data Frames.
- Optimized existing methods using Spark to improve the cluster's performance.
- Worked with Parquet files and Impala using PySpark, as well as Spark Streaming with RDDs and Data Frames.

- Consolidated log data from multiple servers using Apache Kafka and improved Kafka performance while incorporating security measures. Developing batch and streaming processing apps for functional pipeline requirements utilizing Spark APIs.
- Automated data storage from streaming sources to AWS data lakes like S3, Redshift, and RDS using AWS Kinesis (Data Firehose).
- Analyzed streaming data in real-time utilizing AWS Kinesis (Data Streams) integration capabilities.
- Cleaned and treated missing values in data using backward-forward filling methods and applied feature engineering using Python and Scikit-learn preprocessing techniques.
- Stored data into different AWS S3 tiers based on business needs and data access frequency.
- Worked with Informatica Data Quality (IDQ) toolkit for data analysis, cleansing, matching, conversion, exception handling, reporting, and monitoring.
- Utilized Informatica Data Quality for initial data profiling and removing duplicate data.
- Performed reporting analyses on data from AWS stack using BI tools like Tableau and Power BI.
- Worked on SQL optimization for databases such as Oracle, MySQL, and MS SQL, collaborating with the database administration team.
- Assisted in configuring and implementing MongoDB cluster nodes on AWS EC2 instances.
- Monitored Spark apps through Spark UI to identify executor failures, data skewness, and runtime issues.
- Established model versioning and CI/CD practices using tools like Kubeflow and Cloud Build.
- Automated deployments and periodic activities using UNIX Shell Scripting.
- Collaborated with the Data Science team to develop machine learning models on Spark EMR cluster to meet business data needs.
- Maintained version control using GitLab, ensuring traceability and accountability in the development lifecycle. Collaborated with cross-functional teams to implement GitLab CI/CD pipelines.
- Collaborated in an agile environment to implement projects and upgrades using weekly SCRUMs.

Hill Corp Energy, Houston, TX Sep 2018 – Jan 2021

Big Data Engineer
Responsibilities:

- Created S3 buckets in AWS to store files, serving static content for web applications as needed.
- Developed data normalization jobs for Redshift ingestion and managed AWS Data Pipeline for S3 to Redshift data loads.
- Constructed AWS Redshift ETL pipelines tailored to business requirements, streamlining data flow from diverse sources to target systems.
- Automated ETL processes and integrated various data sources into Snowflake and Redshift using Snowpipe, AWS Data Pipeline, and AWS Glue for efficient data management and analytics.
- Developed and optimized Spark jobs for data ingestion into Hive from various sources, enhancing data processing with Spark Core, Streaming, and SQL.
- Conducted Proof of Concepts (POCs) with Spark, benchmarking performance on YARN clusters against Hive and SQL/Teradata solutions.
- Transitioned SQL Server Stored Procedures to Redshift PostgreSQL, integrating them within the Python panda's ecosystem for advanced data manipulation.
- Initiated migration of legacy databases to Snowflake, employing best practices for ETL transformation, data validation, and minimizing downtime during transition.
- Developed and maintained Snowflake SQL scripts, User-Defined Functions (UDFs), and stored procedures for complex data transformation, aggregation, and analysis tasks.
- Transformed complex Hive/SQL queries into efficient Spark transformations using RDDs, Python, and Scala for scalable data processing.

- Facilitated data migration between HDFS, Hive, and Amazon S3, using Sqoop for seamless import/export operations.
- Initiated migration of SQL databases to Amazon's data services like Data Lake, Data Lake Analytics, RDS, and Redshift, ensuring optimal data governance and accessibility.
- Streamlined data workflows by automating imports with Sqoop, scheduling with Oozie, and managing job execution across various environments using Control-M.
- Managed AWS Redshift data analysis, S3 integration, and optimized Hive storage with Sequence files, ORC, bucketing, and partitioning.
- Crafted Python and Shell scripts for consistent process scheduling and developed custom Hive UDFs, ensuring seamless execution within HDFS.

Ryder, Fort Worth, TX **Mar 2017 – Sep 2018**

Data Engineer/Analyst
Responsibilities:

- Worked on various transformations including Expression, Aggregator, Stored Procedure, Java, Lookup, Filter, Joiner, Rank, Router, and Update Strategy, developing reusable Mapplets and Transformations.
- Participated in Design, analysis, Implementation, Testing, and support of ETL processes for Stage, ODS, and Mart, defining project scope, gathering business requirements, and performing GAP analysis.
- Designed the Architecture for Data Lake and Implemented it using Hadoop architecture.
- Generated ad-hoc SQL queries using joins, database connections, and transformation rules to fetch data from legacy DB2 and SQL Server database systems.
- Utilized Erwin for reverse engineering to connect to existing databases and ODS, creating graphical representations in the form of Entity Relationships.
- Developed Data Mapping, Data Governance, Transformation, and cleansing rules for the Master Data Management Architecture involving OLTP, ODS, and OLAP.
- Collaborated with ETL, BI, and DBA teams to analyze and provide solutions to data issues and other challenges during the OLAP model implementation.
- Designed and developed Informatica's Mappings and Sessions based on business user requirements and rules to load data from source flat files and oracle tables to target tables.
- Maintained a data dictionary to create metadata reports for technical and business purposes.
- Analyzed data for data-conversion, including data mapping from source to target database schemas, and automated primary keys creation using PL/SQL triggers and master tables.
- Worked with different data formats such as Flat files, SQL files, Databases, XML schema, CSV files.
- Contributed to the project cycle plan for the data warehouse, source data analysis, data extraction process, transformation, and ETL loading strategy designing.
- Generated various reports using Power BI and Tableau based on Client specifications.
- Responsible for generating actionable insights from complex data to drive real business results for various application teams, working extensively in Agile Methodology projects.

EDUCATION:

Masters in Data Analytics at Northeastern University:2023