**Sr. Data Engineer**

**Name: *Shiva Ganesh* Koluvula**

**Email:** [***shivaganeshk435@gmail.com***](mailto:shivaganeshk435@gmail.com)

**LinkedIn:** [***https://www.linkedin.com/in/shivaganesh24/***](https://www.linkedin.com/in/shivaganesh24/) **Contact Details: +17086889424**

**Education Details:**

*Bachelor of Technology in Computer Science and Engineering from JNTUH in 2013*

**Professional Summary**

* Over 10+ years of experience as Data Engineer and Data Scientist including designing, developing and implementation of data models for enterprise-level applications and systems.
* Experience in designing the Conceptual, Logical and Physical data modeling using Erwin and E/R Studio Data modeling tools.
* Strong knowledge and experience on Amazon Web Services (AWS) Cloud services like EC2, S3, EBS, RDS, VPC, and IAM.
* Proficient in Base SAS programming for data analysis and manipulation.
* Expertise in SAS Macros for automating repetitive tasks and improving efficiency.
* Advanced knowledge of SAS Programming techniques such as data merging, conditional logic, and SQL joins.
* Hands on experience on Data Analytics with Data bricks , Databricks Workspace user Interface, Managing Databricks Notebooks.
* Developed operational analytics, financial analytics, model building and enrichment, prediction engine for both batch and
* real-time using Java, Storm, Kafka, Akka, Spark MLLib, Scikit-learn
* developing and marketing enterprise software products, particularly database management systems.
* DB2 is a relational database system, which means it organizes data into tables with rows and columns. It supports SQL (Structured Query Language), the standard language for managing relational databases.
* Designed and managed public/private cloud infrastructures using Amazon Web Services (AWS) which include EC2, S3, CloudFront, Elastic File System, RDS, VPC, Direct Connect, Route53, Cloud Watch, Cloud Trail, Cloud Formation, and IAM which allowed automated operations.
* Experienced in Big Data using Hadoop framework and related technologies including HDFS, HBase, Spark, Kafka, Map-Reduce, Hive, Pig, Flume, Oozie, Sqoop, Impala, MapRDB, Drill and Zookeeper.
* Good experience in working with different ETL tool environments like SSIS, Informatica, Abinitio and reporting tool environments like SQL Server Reporting Services (SSRS), Data stage, Cognos and Business Objects.
* Proficient in Pandas and NumPy packages for data manipulation and analysis
* Deep expertise with Statistical Analysis, Data mining and Machine Learning Skills using R, Python and SQL.
* Experience in Performance Tuning and query optimization techniques in transactional and Data Warehouse Environments.
* DB2 is designed to store, manage, and retrieve large volumes of data efficiently.
* Skilled in leveraging AbInitio’s data quality and governance capabilities to ensure data accuracy, consistency, and compliance.
* Oracle Database, which is one of the most popular and widely used database management systems in the world.
* Analyze, develop, and build modern data solutions with the aw PaaS service to enable data visualization. Understand the application's current Production state and the impact of new installation on existing business processes.
* Strong knowledge of Spark for handling large data processing in the streaming process along with Scala.
* Experience in Working on NoSQL databases - HBase, Cassandra, Couchbase & MongoDB, database performance tuning &e data modeling.
* developers to write code once and run it on any device or operating system that supports Java.
* Written Terraform scripts to automate AWS services which include ELB, CloudFront distribution, RDS. EC2, database security groups, Route 53, VPC, Subnets, Security Groups.
* Worked with Different utilities in Teradata like fast export and PT export.
* Expertise in writing Hadoop Jobs to analyze data using MapReduce, Apache Crunch, Hive, Pig, and Splunk.
* Experienced in using distributed computing architectures such as AWS products (e.g. EC2, Redshift, and EMR, Elastic search), Hadoop, Python, goo and effective use of MapReduce, SQL and Cassandra to solve big data type problems.
* Partnered with cross functional teams across the organization to gather requirements, architect, and develop proof of concept for the enterprise Data Lake environments like MAPR, CLOUDERA, HORTONWORKS, AWS, and AZURE.
* Create external tables with partitions using AWS Athena and Redshift.
* Java's robustness and scalability make it a popular choice for building large-scale enterprise applications, including Customer Relationship Management (CRM) systems, Enterprise Resource Planning (ERP) software, etc.
* Developed and implemented robust data validation and cleansing processes using AbInitio components,
* Load data into Amazon Redshift and use AWS Cloud Watch to collect and monitor AWS RDS instances within Confidential.
* Skilled in using Erwin to design and maintain conceptual, logical, and physical data models for enterprise-level databases. Proficient in translating business requirements into data models and ensuring data accuracy and consistency across different systems.
* Migrated on premises MySQL to AWS using Amazon RDS and DynamoDB.
* Exposure to Both Kimball and Bill Inmon Data Warehousing Approaches.
* Experience in Working Dimensional Data modeling, Star Schema/Snowflake schema, Fact & Dimensions Tables.
* Experienced in writing Storm topology to accept the events from Kafka producer and emit into Cassandra.
* Experience in building reports using SQL Server Reporting Services and Crystal Reports.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| Big Data Tools | Cloudera Distribution, HDFS, Yarn, Data Node, Name Node, Resource Manager, Node Manager, MapReduce, PIG, SQOOP, Hbase, Hive, Flume, Cassandra, Spark, Storm, Scala, Impala |
| Operating System: | UNIX AIX 5.3, OS/390 z/OS 1.6, Windows 95/98/NT/ME/00/XP, UNIX, MS-DOS, Sun Solaris 5.8, Linux 8x |
| Languages: | Visual Basic 6.0/5.0, SQL, PL/SQL, and Transact-SQL, python |
| Databases: | Snowflake(cloud), Teradata, IBM DB2, Oracle, SQL Server, MySQL, NoSQL |
| Web Technologies: | HTML, XML |
| Version Tools: | GIT, CVS |
| Packages: | SQL\* PLUS, Toad 7.x, SQL Loader, Erwin 7.0 |
| Tools: | TOAD, SQL Developer, ANT, Log4 |
| Web Services/Cloud | WSDL, SOAP. | AWS, Azure |
| ETL/Reporting: | AbInitio GDE 3.0, Informatica, Tableau |
| Web/App Server: | UNIX server, Apache Tomcat |

**Professional Experience**

**Sr. Data Engineer Sep 2021 to Present**

**TJX, Framingham, MA**

**Responsibilities:**

* Implemented Core Framework leveraging Spark that can handle the whole pipeline in one Config.
* Designed and Implement test environment on AWS.
* Experienced in monitoring ETL pipelines, identifying and resolving data processing issues, and conducting root cause analysis to maintain data pipeline robustness and reliability.
* Proficient in Apache Kafka, a distributed streaming platform, with hands-on experience in deploying, configuring, and managing Kafka clusters for real-time data streaming and event-driven architectures.
* Conducted data quality checks and implemented data cleansing procedures to ensure data accuracy and completeness.
* Utilized Base SAS to manipulate and analyze large datasets in order to identify trends and provide insights to the business team.
* Utilized Pandas and NumPy packages in Python programming to manipulate and analyze large datasets, enabling data-driven business decision-making.
* Java's Swing and JavaFX libraries enable the development of graphical user interface (GUI) applications for desktop environments.
* DB2 is available on various platforms, including mainframes (z/OS), mid-range systems (IBM i), and distributed systems (Linux, Unix, and Windows).
* Proficient in utilizing PostgreSQL for data modeling, management, and optimization, including experience in designing and implementing complex queries, managing large datasets, and ensuring data integrity and security.
* Good Experience in Databricks using Data Pipelines to Develop, Design, Depoly.
* Java has a massive and active community of developers, along with a rich ecosystem of frameworks, libraries, and tools.
* The Oracle Database is used to store and manage large volumes of data for various applications and organizations.
* AWS Snowflake experience generating separate virtual data warehouses with differently sized classes.
* Developed scripts in Python (Pandas, Numpy) for data ingestion, analyzing and data cleaning and Data sources are extracted, transformed and loaded to generate CSV data files with Python programming and SQL queries and analyzed the SQL scripts and designed the solution to implement using Pyspark.
* Experienced in optimizing data processing performance using AbInitio.
* Written Templates for Azure Infrastructure as code using Terraform to build staging and production environments.
* Designed the data models to be used in data intensive AWS Lambda applications which are aimed to do complex analysis creating analytical reports for end-to-end traceability, lineage, definition of Key Business elements from Aurora.
* Oracle provides Enterprise Resource Planning (ERP) software to help organizations manage various business processes, including finance, supply chain, human resources, and more.
* Design and developed services to persist and read data from Hadoop, HDFS, Hive and writing Java based MapReduce batch jobs using Hortonworks Hadoop Data Platform
* Writing code that optimizes performance of AWS services used by application teams and provide Code-level application securities clients (IAM roles, credentials, encryption, etc.)
* Using SonarQube for continuous inspection of code quality and to perform automatic reviews of code to detect bugs.
* Extensively used Terraform in AWS Virtual Private Cloud to automatically setup and modify settings by interfacing with control layer.
* DB2 comes in different editions tailored for different use cases, such as DB2 for Enterprise Server Edition, DB2 Workgroup Server Edition, and DB2 Express-C, which is a free community edition for small to medium-sized businesses.
* Responsible in the maintenance and support of the Enterprise Tax Service System, in the process of 23 Tax Registers outputted by different systems from different locations, by doing the following:
* Submit jobs monthly, and pass data from remote LPARs to a common location for processing, using COBOL, SAS, SQL SERVER, JCL and CLIST.
* User profile and other unstructured data storage using Java and MongoDB.
* Collaborate with colleagues and develop a configuration management tool that transforms customers’ infrastructure into version-control codes with using Terraform.
* Responsible for developing new COBOL, SAS programs and debugging, or modifying the existing ones.
* Uploading of outputted files from SQL SERVER into mainframe, for processing and transmitting the outputted files from mainframe process into PC, thru 'MFT' and 'SFTP'.
* ntegrated Azure Log Analytics with Azure VMs for monitoring the log files, store them and track metrics and used Terraform
* Utilized AbInitio’s parallel processing, data partitioning, and load balancing techniques to improve data throughput and reduce processing time.
* Oracle is the current steward of the Java programming language and platform, which is widely used for developing enterprise and web applications.
* The reports which are in Excel are being updated to reflect the new outputted files, and sent to the users for their analysis whether there's a need for them to resend the corrected file/s.

**Environment:** Amazon S3, Databricks,EMR, Glue, ELB, Auto scaling, Oracle, Terraform, AbInitio ,Java, S3, Cloud Watch, Cloud Trail, Athena, Redshift, Spark, Hive, Kafka, SAS, Cloudera, DB2, Lambda.

**Sr. Data Engineer Feb 2020 to Aug 2021**

**FINRA, Rockville, MD**

**Responsibilities:**

* As a Data Engineer involved in the entire life cycle of the project starting from requirements gathering to end of system integration.
* Experienced in implementing event-driven architectures using Kafka's publish-subscribe (pub-sub) model, allowing for the reliable and scalable distribution of data and events among microservices and applications.
* Developed and maintained SAS programs using SAS Macros to automate data processing and reporting tasks, increasing efficiency by 30%.
* Developed and maintained Python scripts to automate data processing and reporting tasks, increasing efficiency by 40%.
* Created and presented reports to senior management that provided actionable insights for business decision-making.
* Proficient in designing and implementing data warehouses using Kimball methodology, including dimensional modeling, ETL processes, and data governance best practices
* Performed detailed analysis of business problems and technical environments and used this data in designing the solution and maintaining data architecture.
* DB2 provides features for ensuring high availability and disaster recovery, such as database mirroring, clustering, and log shipping.
* Developed Big Data solutions focused on pattern matching and predictive modeling
* Worked on Cloud computing using Microsoft Azure with various BI Technologies and exploring NoSQL options.
* Engaged in solving and supporting real business issues with your Hadoop distributed File systems and Open Source framework knowledge.
* Integrated and used various open source dev and ops tools/libraries from Netflix (Simian Army, RxJava, Hystrix, Asgard).
* Expertise in Terraform for multi cloud deployment using single configuration.
* Implemented the Big Data solution to pull the data into the HDFS system.
* Build the data pipelines that will enable faster, better, data-informed decision-making within the business.
* Responsible for analysis of massive and highly complex data sets, performing ad-hoc analysis and data manipulation for data integration.
* Developed and implemented robust data validation and cleansing processes using AbInitio components, ensuring high-quality and reliable data for downstream analysis and reporting.
* Worked on migration of data from On - prem SQL server to Cloud databases (Azure Synapse Analytics (DW) & Azure SOL DB).
* racle has a significant presence in the enterprise software market and serves a wide range of industries, including finance, manufacturing, healthcare, government, and more.
* Azure Storage, Azure SQL, Azure DW) and processing the data in Azure Databricks. Pipelines were created in Azure Data Factory utilizing Linked Services/Datasets/Pipeline/ to extract, transform, and load data from many sources such as Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool, and backwards.
* Azure Cognitive Services, Azure Databricks, etc.) Managed relational database service in which the Azure SQL handles reliability, scaling, and maintenance. Integrated data storage solutions wif Spark, particularly wif Azure Data Lake storage and Blob storage.
* DB2 offers robust security features to protect data, including access controls, encryption, and auditing capabilities.
* Used Terraform in AWS Virtual Private Cloud to automatically setup and modify settings by interfacing with control layer.
* Java's strict type-checking and exception handling mechanisms help to create more reliable and robust applications.
* Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL and U-SQL Azure Data Lake Analytics.
* Performed data collection, data cleansing, data aggregation and curation in Teradata and Abinitio. Assisted data visualization leaders to connect to customers and develop priority data visualization tools. Building connection points between datasets, enabling the creation of priority cross-functional reporting for the business.
* Developed shell script to automate job execution in AutoSys with automatic Failure detection and alert triggering there by saving around 800 hours estimated annually worth (120K dollars).
* Designed and implemented scalable Cloud Data and Analytical architecture solutions for various public and private cloud platforms using Azure.
* Implemented business logic by writing UDFs and configuring Control M Jobs.
* Identified data within different data stores, such as tables, files, folders, and documents to create a dataset in pipeline using Azure HDInsight.
* Its products are used by organizations of all sizes, from small businesses to large multinational corporations.
* Java comes with a vast standard library that provides pre-built classes and functions for common tasks, simplifying development.
* DB2 is a relational database system, which means it organizes data into tables with rows and columns. It supports SQL (Structured Query Language), the standard language for managing relational databases.
* Used data integration to manage data with speed and scalability using the Apache Spark engine in Azure Databricks.

**Environment:** Apache Spark 3.1, Zookeeper 3.7, Hive 2.3, Agile, Java, AbInitio , DB2,Terraform, Oracle, DB2, Microsoft Azure, SQL, Python 3.8, Hadoop 3.2, XML, Data Lake, Data Factory, Data Bricks

**Data Engineer Oct 2018 to Dec 2019**

**Verizon, Irving, TX**

**Responsibilities:**

* Responsible for building scalable distributed data solutions using Hadoop.
* Adept at optimizing AWS costs by implementing cost control strategies, such as Reserved Instances, spot instances, auto-scaling, and leveraging AWS Cost Explorer and AWS Trusted Advisor.
* Skilled in using Kafka for data ingestion, including streaming data from various sources into Kafka topics, and processing data with Kafka Streams or other stream processing frameworks for real-time analytics and transformations.
* Created complex SAS programs using Advanced SAS Programming techniques such as data merging, conditional logic, and SQL joins to solve business problems.
* Collaborated with cross-functional teams to develop custom data visualizations and reports using Pandas and Matplotlib libraries.
* Java is used in the development of big data processing tools, distributed systems, and cloud-based applications.
* Involved in Agile Development process (Scrum and Sprint planning).
* Handled Hadoop cluster installations in Windows environment.
* Migrated on-premise environment in GCP (Google Cloud Platform) and also Migrated data warehouses to Snowflake Data warehouse.
* Defined virtual warehouse sizing for Snowflake for different type of workloads.
* Involved in porting the existing on-premises Hive code migration to GCP (Google Cloud Platform) BigQuery.
* Involved in migration of an Oracle SQL ETL to run on Google cloud platform using cloud Dataproc &amp, BigQuery, cloud.
* Integrated with UI layer using HTML. Ajax. JavaScript
* Extracted data from data lakes, EDW to relational databases for analyzing and getting more meaningful insights using SQL Queries and PySpark.
* Responsible in the maintenance and support of the Enterprise Tax Service System, in the process of 23 Tax Registers outputted by different systems from different locations, by doing the following:
* Submit jobs monthly, and pass data from remote LPARs to a common location for processing, using COBOL, SAS, SQL SERVER, JCL and CLIST.
* DB2 is optimized for performance, and it includes features like buffer pools, multi-version concurrency control, and query optimization.
* Responsible for developing new COBOL, SAS programs and debugging, or modifying the existing ones.
* Uploading of outputted files from SQL SERVER into mainframe, for processing and transmitting the outputted files from mainframe process into PC, thru 'MFT' and 'SFTP'.
* The reports which are in Excel are being updated to reflect the new outputted files, and sent to the users for their analysis whether there's a need for them to resend the corrected file/s.
* Participated in a regular Mainframe Disaster Recovery Exercise and Validation Process
* Developed PySpark script to merge static and dynamic files and cleanse the data and Created Pyspark procedures, functions, packages to load data.
* Java has been the primary language for Android app development, allowing developers to create applications for millions of Android devices.
* Creating and managing schema objects such as tables, views, indexes, stored procedures, and triggers & maintaining Referential Integrity.
* Java can be used to build web applications through technologies like JavaServer Pages (JSP), Servlets, and frameworks like Spring and JavaServer Faces (JSF).
* Created SSIS packages to populate data from various data sources.
* Created packages using SSIS for data extraction from Flat Files, Excel Files, OLEDB to SQL Server
* Designed, developed and did maintenance of data integration programs in a Hadoop and RDBMS environment with both traditional and non-traditional source systems.
* Oracle serves a diverse range of industries, including finance, healthcare, retail, manufacturing, government, and more.
* Java has been the primary language for Android app development, allowing developers to create applications for millions of Android devices.
* Developed MapReduce programs to parse the raw data, populate staging tables and store the refined data in partitioned tables in the EDW.
* Wrote Sqoop Scripts for importing and exporting data from RDBMS to HDFS.
* Set up Data Lake in Google cloud using Google cloud storage, BigQuery and Big Table.
* Developed scripts in BigQuery and connected it to reporting tools.

**Environment:** Spark, flume 1.8, Sqoop 1.4,Hadoop 3.0, YARN, HDFS, DB2, Java, HBase 1.2, Kafka, Scala 2.12, Cassandra 3.11, Elastic Search, MapReduce, UNIX, Zookeeper 3.4, Datastage 8.7

**Big Data Engineer July 2016 to Sep 2018**

**Centene, St. Louis MO**

**Responsibilities:**

* Created Hive tables, loaded data, executed HQL queries and developed MapReduce programs to perform analytical operations on data and to generate reports.
* Proficient in setting up Kafka clusters that are highly scalable and fault-tolerant, ensuring high availability and data durability by configuring replication, partitions, and monitoring.
* Collaborated with cross-functional teams to develop data-driven solutions that improve business processes and decision-making.
* Conducted data quality checks and implemented data cleansing procedures to ensure data accuracy and completeness.
* Utilized NumPy arrays and matrices for complex mathematical operations and data modeling.
* Created Hive internal and external tables, used MySQL to store table schemas. Wrote custom UDFs in Python
* Moved data between MySQL and HDFS using Sqoop.
* Developed MapReduce jobs in Python for log analysis, analytics, and data cleaning.
* Wrote complex MapReduce programs to perform operations by extracting, transforming, and aggregating to process terabytes of data.
* Wrote many SQL, Procedures, Triggers and Views on top of Oracle.
* Involved in the implementation of the Software development life cycle (SDLC) that includes Development, Testing,
* Performed review and analysis of the detailed system specifications related to the DataStage ETL and related applications to ensure they appropriately address the business requirements.
* Evaluated impact of proposed changes on existing DataStage ETL applications, processes and configurations
* Developed MapReduce jobs, Hive & PIG scripts for Data warehouse migration project.
* loaded data into HDFS and extracted data from Teradata into HDFS using Sqoop.
* Writing the script files for processing data and loading to HDFS.
* DB2 comes in different editions tailored for different use cases, such as DB2 for Enterprise Server Edition, DB2 Workgroup Server Edition, and DB2 Express-C, which is a free community edition for small to medium-sized businesses.
* he Oracle Fusion Middleware suite includes tools for application development, integration, business process management, and content management.
* Moved all RDBMS data from Teradata into flat files generated from various channels to HDFS for further
* Supported Map Reduce Programs that are running on the cluster. Involved in loading data from the UNIX file system to HDFS.
* Developed Sqoop Jobs to load data from RDBMS, External Systems into HDFS and HIVE.
* Developed Oozie coordinators to schedule Pig and Hive scripts to create Data pipelines.

**Environment:** Hadoop, Oozie, ETL, Hive, Sqoop, Impala, Oracle, Oozie, zookeeper, HDFS, Teradata.

**Sr. ETL Consultant June 2013 to May 2016**

**Avon Technologies Pvt Ltd Hyd India**

**Responsibilities:**

* Gathering requirements for RDM project which involves implementing EDW data quality fixes and Retail data mart.
* Prepare functional and technical specification design document for building Member Data Mart according to ICDW Banking Model.
* Skilled in using ETL tools such as Apache NiFi, Talend, Informatica, Apache Spark, or custom ETL scripts to automate data extraction, transformation, and loading tasks efficiently.
* Responsible for data gathering from multiple sources like Teradata, Oracle.
* Created Hive tables to store the processed results in a tabular format.
* Written Map Reduce jobs in java to process the log data. Implemented external and managed tables using HIVE.
* Work with the Teradata analysis team using BigData technologies to gather the business requirements.
* Fixing error data, Data Reconciliation process.
* Used Partitioning and bucketing concepts for performance optimization in hive.
* Responsible for delivering the Informatica artifacts for Mart Specific Semantic Layer for subject areas like Reference, Third Party, Involved Party, Event, Customer and etc.
* Prepared and implemented successfully automated UNIX scripts to execute the end-to-end history load process.
* Managing the versioning of the mappings, scripts, documents in version-controlled tool SCM.