|  |  |
| --- | --- |
| Varma V  Sr Data Engineer | horizontal line  *9033265270* [*vvarma1516@gmail.com*](about:blank)[*www.linkedin.com/in/vvarma1516*](http://www.linkedin.com/in/vvarma1516) Senior Data Engineer with around 7 years of data solutions experience, combining technical skills with organizational objectives. Proven ability to lead complex projects, collaborate across teams, and deliver actionable insights. Experienced in ETL development, data warehousing, and big data technologies, with cloud platforms like AWS and Azure. Strong problem-solving skills and commitment to data quality and security standards. |
| **ㅡ** Technical Skills | horizontal line  **ETL Tools -** Informatica, Talend, PowerCenter, SSIS  **Big Data Technologies -** HDFS, Apache NIFI, Map Reduce, Sqoop, Flume, Pig, Hive, Oozie, Impala, Zookeeper, Ambari, Storm, Spark, and Kafka, Horton Works, Cloudera, MapReduce.  **Databases -** Oracle SQL server, MY SQL, MS SQL Server, Vertica, Teradata, Snowflake, HBase, Cassandra, MongoDB, PostgreSQL.  **Cloud Technologies -**  **AWS -** Amazon EC2, Amazon S3, Amazon SimpleDB, Amazon MQ, Amazon ECS, Amazon Lambdas, Amazon Sagemaker, Amazon RDS, Amazon Elastic Load Balancing, Elastic Search, Amazon SQS, AWS Identity and access management, AWS Cloud Watch, Amazon EBS and Amazon CloudFormation.  **Azure -** Azure Data Lake, Data factory, Azure Databricks, Azure SQL database, Azure SQL Datawarehouse, Azure Functions, Azure Synapse, Azure HD Insights, Azure Blob Storage, Azure Event Hub, Azure streaming Analytics.  **GCP -** BigQuery, Cloud DataProc, GCS.  **Data Orchestration Tools -** Apache Airflow, Prefect, Oozie, Dagster, Maze  Data Modelling Tools: Erwin Data Modeler, ER/Studio, IBM InfoSphere datastage, Datagrip, Luna Modeler.  **File Formats -** csv, json, xml, parquet, avro, txt, Apache iceberg.  **Monitoring/Reporting -** Tableau, Power BI, Cognos, Microsoft SSIS, SSAS and SSRS.  **Version Control Software(VCS) -** Git, SVN, Jira.  **IDE -** PyCharm, Visual Studio Code, SSMS, Data Studio, IntelliJ.  **Operating Systems -** Linux, Unix, Windows 8, Windows 7, Windows Server 2008/2003.  **Programming/Markup Languages -** Java, Python, SQL, Perl, Ruby, JavaScript, UNIX Shell Scripting, Pig Latin, HiveQL. |
| ㅡCertifications | horizontal line   * Azure - Data Engineering *DP203*  Azure Databricks |
| ㅡProfessional Summary | horizontal line   * Adept Understanding on latest technologies including in data engineering, data services, data virtualization, data integration, Master Data Management. * Professional experience on building analytical applications and data systems including the data ingestion pipeline design, **Hadoop/Spark** **architecture, data modeling, data mining, advanced SQL data processing** and cloud ecosystem. * Proficient in creating complex mappings, transformations, and workflows to extract, transform, and load (ETL) data from various sources into target systems. * Skilled in developing Microservices based on Restful web service using Akka Actors and Akka-Http framework in **Scala** which handles high concurrency and high volume of traffic. * Proficient in Apache Kafka, mastering real-time data streaming and event-driven architecture to ensure seamless data flow, processing, and integration, facilitating robust and scalable data-driven applications * Experienced in Apache Airflow, adept at orchestrating complex data workflows, automating tasks, and scheduling data pipelines, enhancing operational efficiency and reliability in data processing and analytics * Excellent Knowledge of Ralph Kimball and Bill Inmon's approaches to Data Warehousing. * Experience designing warehouse architecture and common data models adhering the best practices of **Enterprise Data Warehouse**. * Proficient in constructing, manipulating, and analyzing large datasets with a variety of data types, leveraging tools such as SAS, SQL, Python, R, Minitab, Power BI, and others. Skilled in extracting valuable insights, including multi-factor interactions, to drive meaningful changes within system application architectures. * Experienced utilizing Azure cloud services like **Azure Data Factory (ADF), Azure Data Lake (ADL), Azure Data Lake Analytics, Azure SQL database, Azure Databricks, Azure Synapse, Azure SQL Data Warehouse, Azure BLOB Storage.** * Expertise in the development of various reports, and dashboards using various **Power BI,** and **Tableau Visualization.** * Proficient in leveraging GCP **BigQuery**  features and capabilities to perform complex data transformations, aggregations, and analysis on large datasets. * Snowflake ,Snowpipe, Virtual Data Warehouse, data staging from AWS or Azure * *Experience with Snowflake Multi-Cluster Warehouses. And Virtual Warehouses.* Participated in migrating objects from Teradata to Snowflake. * Expertise **in AWS services including S3, EC2, SQS, RDS, Neptune, EMR, Kinesis, Lambda, Step Functions, Terraform, Event Bridge, Glue, Redshift, Athena, DynamoDB, Elasticsearch, Service Catalog, CloudWatch, IAM.** * Experienced in Teradata SQL queries, Teradata Indexes, Utilities such as Mload, Tpump, Fast load, and Fast Export. * Experience working on development and deployment of Hadoop and bigdata ecosystems - HDFS, MapReduce, Spark, Pig, Sqoop, Hive, Oozie, Kafka, zookeeper, and HBase. * Expertise in data processing by utilizing **MapReduce**, **Spark,** and **Hive jobs** for data analysis and optimizing and performance by SQL tuning for the existing jobs. * Experienced in working with **NoSQL** databases like Cassandra and **HBase** and developed real-time read/write access to very large datasets via **HBase.** * Used various Hadoop distributions (Cloudera, Hortonworks, Amazon EMR, Microsoft Azure HDInsight) to fully implement and leverage new Hadoop features. * Extensive experience in designing, developing, and implementing data integration solutions **using Informatica PowerCenter and Informatica Intelligent Cloud Services (IICS).** * Expertise in aspects of **Agile** framework from Sprint planning, retrospective analysis, and work estimates. * Strong experience in **CI (Continuous Integration)/ CD (Continuous Delivery)** software development pipeline stages like Commit, Build, Automated Tests, and Deploy using Bogie in Jenkins. * Experienced in writing UNIX shell scripting and hands-on experience with scheduling shell scripts using Control-M. * Proficient in version control systems including Git and SVN, and adept at using Jira for efficient project tracking and collaboration. |
| **ㅡ** Experience | horizontal line Caterpillar Inc. / Sr. Data EngineerAug’21-Present, Peoria, IL (Remote) ● Collaborated with Analysts and other departments to report and review the documents for tiebreaker analysis and metadata documents.  ● Responsible for data services and data migration initiatives from downstream systems to Upstream Snowflake Environment  ● Implemented **AWS** infrastructure to process the facility shipment data from across the globe using services **Amazon S3, Lambda, Kinesis, Redshift and DynamoDB** for data movement and data processing to unified data storage unit  ● Designed a **data lake** architecture when needed to handle both streaming and batch data.  ● Built real-time analytics models using **Amazon Kinesis Analytics, Apache Kafka Streams, AWS Lambda functions, or Apache Spark Streaming** to perform continuous analysis on the streaming data and generate insights in real-time.  ● Integrated **AWS QuickSight, Tableau** data visualization tool to create interactive dashboards and reports that displayed real-time analytics results to stakeholders and leadership team.  ● Architectured and implemented the data pipelines with scalability to handle growing data volumes and optimized for performance to deliver timely insights, including **AWS Auto Scaling** where appropriate.  ● Implemented security measures, including encryption (using **AWS Key Management Service**), access control (using **AWS Identity** and **Access Management**), and compliance checks to protect sensitive data and ensured regulatory compliance.  ● Implemented data quality checks and validation using **AWS Glue DataBrew, AWS Data Quality Rules**, or custom scripts to ensure data accuracy and integrity throughout the pipeline.  ● Automated ETL workflows using **AWS Step Functions and AWS Glue** ETL jobs to reduce manual intervention and streamlined data processing.  ● Used **AWS glue** catalog with crawler to get the data from **S3** and perform SQL query operations using Crawlers and scheduled the job and crawler using workflow feature.  ● Developed Python scripts, **UDFs** using both **Data frames/SQL/Data sets** and **RDD/MapReduce** in **Spark** for Data Aggregation, queries and writing data back into **OLTP** system through **Sqoop**.  ● Developed **Spark Applications** by using **Python** and Implemented **Apache Spark** data processing project to handle data from various RDBMS and Streaming sources. Responsible in performing sort, join, aggregations, filter, and other transformations on the datasets using **Spark** Extract Real time feed using **Kafka** and **Spark Streaming** and convert it to **RDD** and process data in the form of Data Frame and save the data as **Parquet** format in **HDFS**.  ● Worked on **SnowSQL** and **Snowpipe** and converted **Talend** Joblets to support the **snowflake** functionality.  ● Created **Snowpipe** for continuous data load and used COPY to bulk load the data into **Snowflake**.  ● Created data sharing between two snowflake accounts and created internal and external stage to transform data during load stage.  ● Worked on **Apache** **NiFi** to support scalable directed graphs for data routing, transformation and automating the movement of data between disparate systems.  ● Implemented **Lambda** to configure **Dynamo DB** Autoscaling feature and implemented **Data Access Layer** to access **AWS DynamoDB** data.  ● Experience in performance SQL tuning a **Cassandra** cluster to optimize writes and reads.  ● Worked on cloud deployments using **Maven, Docker, and Jenkins**.  ● Provided on call Production Support for **Pipelines** and helped Architecture team with feedbacks for future improvements. US Bank / Sr. Data EngineerJun’19-Jun’21 / Dallas,TX ● Used **ETL** to implement the Slowly Changing Transformation, to maintain Historical data in the Data warehouse.  ● Implemented Copy activity and custom **Azure Data Factory** Pipeline Activities.  ● Built processes supporting Data Transformation, Data Structures, Metadata, Dependency, and workload management.  ● Involved in Refactoring and optimizing the existing data flows from multiple source systems to enhance the performance and processing time for daily batch data.  ● Utilized **Azure pipelines** and **python** scripts to perform data migration from different sources including **Teradata, SQL Server**, and incremental file ingestion into **Snowflake**.  ● Worked on **Cloud shell SDK** in **GCP** to configure services like **Data Proc, Big Query & GCS**.  ● Designed **DDL** statements, complex **SQL** queries, views, and indexes to create and test tables.   ● Developed a framework to generate daily adhoc reports and extracts from enterprise data using **BigQuery**.  ● Worked with Google data catalog and other **Google cloud API**’s for monitoring, queries and billing related analysis using **BigQuery**.  ● Created data transformation procedures using **Spark SQL** and **PySpark**, including complex SQL methods like Joins & Aggregations, Window Functions, and Stored Procedures.  ● Ensured data integrity standards using **Great Expectations**.  ● Conducted training sessions, created documentation, and utilized collaboration tools like **Microsoft Teams, Confluence, and SharePoint**. Provided ongoing support through **Jira** for efficient communication and issue resolution.   ● Used **Azure Monitor** for monitoring platform functionality and health, ensuring system reliability and performance.  ● Used **Azure Service Bus** for proactive alerts and efficient communication within the system.  ● **Azure Data Factory** as part of implementing copy activity and custom pipeline activities for data processing.  ● **Azure Data Warehouse** is often used in conjunction with Azure Data Factory for data storage and analytics purposes.  ● **Azure Key Vault** has been used for securing and managing cryptographic keys, secrets, and certificates.  ● Implemented **Apache Airflow** for orchestrating and automating complex data workflows, ensuring efficient scheduling, monitoring, and management of data pipelines, contributing to improved operational efficiency and data processing reliability.  ● Utilized **Pyspark** for developing data transformation procedures, incorporating advanced SQL methods like Joins & Aggregations, Window Functions, and Stored Procedures to optimize data processing and analytics.  ● Leveraged **Databricks** in conjunction with **Delta Lake** to enhance data engineering workflows, improve data quality, and enable efficient data management, ensuring data consistency and reliability throughout the project.  ● Collaborated with **Azure Synapse Analytics** (formerly known as SQL Data Warehouse) for data warehousing and analytics capabilities, supporting complex SQL queries, data structures, and analytical operations within the project's data ecosystem. United Airlines / Data EngineerJan’18-May’19, Houston, TX ● Extensively used **Erwin r9.5** to design Logical/Physical Data Models, relational database design, forward/reverse engineering, publishing Data Model to acrobat files.  ● Created **ERWIN** reports in **HTML**, **RTF** format depending upon the requirement, Published Data Model in the model mart, created naming convention files, co-coordinated with **DBAs** to apply the Data Model changes.  ● Implemented various algorithms for analytics using **Cassandra** with **Spark** and **Scala**.  ● Involved in requirement analysis, **ETL** design, and development for extracting data from the source systems like **Teradata 13.1**, **DB2**, **Sybase, Oracle 9i, flat files,** and loading into **Netezza**.  ● Involved in **Ralph Kimball** and **Bill Inmon** Methodologies (**Star Schema, Snowflake Schema**).  ● Coded using **Teradata Analytical functions**, **BTEQ SQL of TERADATA**, wrote **UNIX** shell scripts to validate, format, and execute the SQLs on the LINUX environment.  ● Created various types of reports such as drill down & drill through reports**, Matrix reports, Sub-reports, and Charts using SQL Server Reporting Services (SSRS).**  ● Performed exploratory data analysis (**EDA**) using **Python** and did Python integration with **Hadoop Map Reduce** and **spark**.  ● Developed data models for **AWS Redshift** and **Hive** from dimensional data models and designed and implemented a **Data Lake** to consolidate data from multiple sources, using **Hadoop** stack technologies like **SQOOP, HIVE/HQL**.  ● Worked in importing and cleansing of data from various sources like **Teradata, Oracle, Netezza flat files, SQL Server** with high volume data.  ● Implemented naming Standards and Warehouse Metadata for facts and dimensions of Logical and Physical Data Models.  ● Involved in ETL processing using **Pig & Hive in AWS EMR, S3, and Data Profiling, Mapping, and Integration** from multiple sources to **AWS S3**.  ● Created complex Stored Procedures and **PL/SQL** blocks with optimum performance using **Bulk Binds** (BULK COLLECT & FORALL), Inline views, Reference cursors, cursor variables, dynamic SQL, v-arrays, external tables, nested tables, etc.  ● Used **Sqoop** to import/export data between RDBMS and hive tables, incremental imports, and created **Sqoop** jobs for last saved value.  ● Designed Informatica mapping for Error handling and was involved in the preparation of the low-level design (**LLD**) documents for **Informatica Mappings**.  ● Designing and developing SQL Server Database, Tables, and Indexes, Stored procedures, Views, User Defined Functions, and other **T- SQL** statements.  ● Experience in designing and developing POCs in **Spark** using **Scala** to compare the performance of **Spark** with **Hive** and **SQL/Oracle**.  ● Assisted with data capacity planning and node forecasting and collaborated with the infrastructure, network, database, application, and BI teams to ensure data quality and availability. EY / Big Data EngineerNov’16-Dec’17, Bengaluru-India ● Evaluated the existing Hadoop cluster infrastructure, including **HDFS, YARN**, and related components, and identified areas for optimization to enhance data processing efficiency. Utilized tools like **Hadoop HDFS**, **YARN Resource Manager**, and Hadoop **MapReduce** for cluster evaluation.  ● Designed and implemented a robust data ingestion framework for acquiring real-time and batch data from various sources, such as trading platforms, market feeds, and internal databases. Employed tools like **Apache NiFi, Flume, and Kafka** for efficient data ingestion.  ● Improved the existing ETL (Extract, Transform, Load) pipelines by leveraging Hadoop tools like **Apache Spark, MapReduce**, and **Apache NiFi** to streamline data transformation and loading processes. Used Spark SQL for data transformation.  ● Implemented data quality checks, validation rules, and monitoring mechanisms to ensure data accuracy and integrity throughout the processing pipelines. Leveraged tools like **Apache NiFi** for data quality checks and **Apache Falcon** for data lineage.  ● Installed and configured **Hadoop MapReduce, HDFS** Developed multiple MapReduce jobs in java for data cleaning and preprocessing. Worked on installing, configuring, and using Hadoop ecosystem components.  ● Worked in administration, installing, upgrading, and managing **CDH3, Pig, Hive & HBase**. Importing and exporting data into **HDFS** and **Hive using Sqoop.**  ● Constructing Spark Streaming tasks using **PySpark, and Spark-Shell to create RDDs (Resilient Distributed Datasets).**  ● Extraction, Transformation, and Loading **(ETL)** data into Data Warehouses, as well as data processing tasks such as collecting, aggregating, and moving data from numerous sources using **Kafka, Power BI, Microsoft SSIS, MapReduce, PySpark, Spark-SQL, and Pig.** ● Successfully integrated **Informatica PowerCenter** with external systems, such as **Salesforce, SAP,** and **SharePoint,** enabling seamless data integration and synchronization.  ● Participated in the development/implementation of the **Cloudera Hadoop environment.** Ran Map Reduce Programs that are running on the cluster. Involved in loading data from the UNIX file system to **HDFS.**  ● Created **HBase** tables to store variable data formats of **PII data coming** from different portfolios. Implemented best income logic using **Pig scripts.** Cluster coordination services through **Zookeeper.**  ● Enhanced data security by implementing encryption methods for sensitive financial data using tools like **Hadoop KMS (Key Management Service)** in compliance with industry regulations.  ● Developed strategies for horizontal scalability of the Hadoop cluster to accommodate growing data volumes and ensure optimal performance. Used tools like **Hadoop YARN** for resource management and **Cloudera Manager** for cluster scalability.  ● Implemented robust monitoring and logging solutions using tools like **Log4j, Ambari, Prometheus, and Grafana** to proactively identify and address issues.  ● Ensured that data processing and storage complied with financial industry regulations (e.g., **SEC, FINRA**) and implemented auditing mechanisms for compliance reporting. Utilized **Apache Ranger** for security policies and Apache Atlas for metadata governance.  ● Created comprehensive documentation for data engineering processes, best practices, and configurations using tools like **Confluence** and Markdown to facilitate knowledge sharing and future maintenance.  ● Conducted training sessions for the data engineering team to ensure the adoption of optimized processes and tools. Utilized collaboration tools like **Microsoft Teams** for training sessions and Jenkins for continuous integration and deployment. KPMG / Data Engineer (Internship)Jan’15-Oct’16, Hyderabad-India ● Involved in the entire lifecycle of the projects including Design, Development, Deployment, Testing, Implementation, and support.  ● Used **Agile** application development techniques and participated in the Software Development Life Cycle phases of analysis, definition, design, implementation, and testing. Used **Python** along with other libraries such as **matplotlib** for charts and graphs, **MySQL dB** for database connectivity, **Python-twitter, Py Side, Pickle, and Pandas data frame**.  ● In **Linux/Python**, I created an embedded software data-driven test automation framework. Developed test cases and test plans. Experience in the creation of Indexes, Stored Procedures, Constraints, Cursors, Triggers, Views, and User Defined Functions.  ● On the **Django** Web Framework, I developed complete frontend and backend modules in **Python** and used **Py** **Unit**, the Python unit test framework, for all Python applications. I wrote and executed several **MySQL** database queries in Python using the **Python-MySQL** connector and **MySQL DB** package.  ● Worked on developing web-based applications using **PHP, XML, JSON, and MVC3**. Used Python scripts to update the content in the database and manipulate files.  ● Involved in building database models, **APIs**, and views utilizing Python, to build an interactive solution.  ● Well-versed in the design and development of the presentation layer of web applications using technologies like **HTML, CSS, JavaScript, jQuery, AJAX, and Bootstrap.**  ● Developed the required **XML** schema documents and implemented the framework for parsing XML documents.  ● Used **PANDAS** to handle enormous amounts of data as Data Frames Written scripts to handle missing data, analyze data, and reduce dimension. Writing unit-test cases for APIs and individual Python scripts to maintain code standards and stability, as well as debugging to provide error-free code.  ● Used python for **XML, JSON** processing, data exchange, and business logic implementation.  ● Worked on Python OpenStack APIS and used **NumPy** for numerical analysis. Worked in interfacing with third-party APIs using **REST and SOAP**.  ● Installing the necessary **Python** and **Django** packages and tools to set up the staging and test servers. Used the **Git** version control tool to organize team development.  ● Worked on **Python OOD** code for quality assurance, logging, monitoring, and debugging. **Selenium** Library was used to create a fully functional test automation process that simulated submitting different requests from multiple browsers to a web application.  ● Managed large datasets using **Pandas** data frames and MySQL. Worked with varieties of Relational Databases (**RDBMS**) like **MySQL, Oracle, and PostgreSQL**.  ● Part of the team implemented the rest of APIS in Python using micro-framework like Flask with **SQLALCHEMY** in the backend for the management of data center resources on which OpenStack would be deployed.  ● Worked in the development of applications and worked on Jenkins continuous integration tool for deployment of project and deployed the project into Jenkins using GIT version control system and involved in **Agile** Methodologies and **SCRUM** Process. |
| **ㅡ** Education | horizontal line University of Central MissouriMaster’s in Technology Specialized in Big Data, Machine Learning and Networking |
|  | horizontal line |