**Jeevan Kumar Tirumalagiri**

**Sr. Data Engineer**

**Email:** **jeevantiru4@gmail.com**

**Ph: +1(940)268-3691**

**LinkedIn: http://www.linkedin.com/in/jeevan-kumar-t-6980912b7**

**PROFESSIONAL SUMMARY**

* Senior Data Engineer with 9 years of experience specializing in developing scalable data pipelines and ETL processes using AWS Glue, Apache Kafka, and Spark, enhancing data flow and storage solutions.
* Proficient in managing and analyzing large datasets using Amazon S3, SQL, NumPy, and Pandas, driving insightful business decisions.
* Demonstrated expertise in Spark 1.6/2.0 and PySpark for complex data processing, contributing to significant improvements in data analysis and processing speed.
* Hands-on experience with Amazon EC2, Amazon S3, Amazon RDS, VPC, IAM, Amazon Elastic Load Balancing, Auto Scaling, CloudWatch, SNS, SES, SQS, Lambda, EMR and other services of the AWS family.
* Versatile in handling data lake architectures and AWS cloud services, ensuring optimized data storage and accessibility for diverse applications.
* Expert in Python programming, automating data workflows and integrating machine learning models for predictive analytics and data insights.
* Extensive experience with Cloudera Stack, managing HBase, Hive, Impala, and Pig for robust big data ecosystem support.
* Experience of Partitions, bucketing concepts in Hive and designing both managed and External tables in Hive to optimize performance. Experience with different file formats like Avro, parquet, ORC, Json and XML
* Proficient in streamlining data flows with NiFi and Spark Streaming, enabling real-time data processing and analytics.
* Skilled in implementing ELK/Splunk for log management and data visualization, enhancing operational intelligence and data-driven strategies.
* Advanced knowledge of RESTful API, JSON, XML, and SOAP UI for efficient data integration and web services development.
* Deep understanding of database management using MySQL, Cassandra, and MongoDB, ensuring data integrity and performance.
* Experienced in cloud-based data warehousing and analytics using GCP BigQuery, Azure Data Lake, and Snowflake, providing scalable and cost-effective solutions.
* Proficient in Jenkins, Docker, and Kubernetes for CI/CD pipelines, containerization, and orchestration, enhancing deployment efficiency and scalability.
* Strong background in data visualization and reporting tools like Tableau, Power BI, and MicroStrategy, translating complex data into actionable insights.
* Experienced with GCP Dataproc, Dataflow, and Azure Data Factory for cloud-native data processing and integration services.
* Expert in NoSQL database technologies with extensive experience in designing, implementing, and managing scalable and high-performance MongoDB databases.
* Proficient in optimizing Hadoop clusters and workflows for enhanced data processing and storage efficiency.
* Demonstrated success in deploying Hadoop DistCp for large-scale data transfers, maintaining data integrity, and ensuring high availability and compliance across hybrid environments.
* Skilled in integrating NoSQL databases with various data processing and analytics tools, enhancing data-driven decision-making and operational efficiency.
* Specialized in designing and implementing real-time data processing solutions using technologies like Apache Flink, Kafka, and Spark within AWS and GCP environments.
* Developed data architectures and strategies in CDP, facilitating effective data ingest and extraction processes to meet business objectives.
* Understanding of structured data sets, data pipelines, ETL tools, data reduction, transformation and aggregation technique using DBT.
* Expertise in data security and compliance, utilizing Azure Storage, Cloud Spanner, and Cloud SQL for secure data handling and transactions.
* Proficient in system-level programming with Golang and C, specializing in developing high-performance, scalable backend services and system components for real-time data processing and analytics.
* Experienced in utilizing distributed computing frameworks like Spark or Flink for building and optimizing large-scale data processing pipelines, enabling advanced data analytics and machine learning applications.
* Adept in container orchestration with Kubernetes, managing cloud-native applications' deployment, scaling, and operational workflows, enhancing system reliability and deployment efficiency.
* Skilled in Apache Airflow and Oozie for workflow scheduling, automating and managing data pipelines for improved efficiency and reliability.
* Deep familiarity with Kafka, Kubernetes, and service mesh technologies, coupled with a robust analytical skillset for solving complex data processing challenges in dynamic environments.
* Advanced user of Terraform and Ansible for infrastructure as code and configuration management, streamlining cloud infrastructure provisioning and maintenance.
* Adept in leveraging Salesforce for CRM data integration and analytics, enhancing customer engagement and business processes through data-driven insights.
* Strong analytical and problem-solving skills, with a proven track record of improving database systems to meet the dynamic needs of businesses.

**Certifications:**

* [**AWS CERTIFIED DATA ENGINEER – ASSOCIATE**](https://cp.certmetrics.com/amazon/en/public/verify/credential/d9b93556d27145bb82d2da2b1e23e353)
* [**MICROSOFT AZURE CERTIFIED DATA ENGINEER - ASSOCIATE**](https://learn.microsoft.com/api/credentials/share/en-us/JeevanKumar-2855/C2C36DB41BCC72DA?sharingId=4928315F7ACCEFCE)

**Technical Skills:**

| **Category** | **Skills** |
| --- | --- |
| **Cloud Platforms & Services** | AWS Glue, AWS Lambda, Amazon S3, Dynamo DB, AWS SageMaker Cloudera Data Platform (CDP), GCP, GCS, Azure, Azure Data Lake, Azure Storage, Cloud Composer, Cloud Pub/Sub, Cloud Storage Transfer Service, Cloud Spanner, Cloud SQL, AWS Neptune, Azure Cosmos DB |
| **Data Processing & Analytics** | Apache Kafka, Spark, Spark 1.6 / 2.0, Flink, PySpark, SPARQL, ETL, DBT, Data Lake, NiFi, Spark Streaming, GCP DataProc, BigQuery, GCP Dataflow, Azure Data Factory, Data Flow, ETL Pipelines |
| **Database Management** | SQL Database, MongoDB, MySQL, Cassandra, Snowflake, SQL, SnowSQL |
| **Big Data Technologies** | Hadoop, Hive, Cloudera Stack, HBase, Impala, Pig, ELK/Splunk, Athena, Redshift |
| **DevOps & CI/CD** | Jenkins, Docker, Ansible, Terraform, Maven, GIT, Kubernetes, Helm, Spinnaker |
| **Programming Languages** | Python, Scala, C, Golang, R, Shell Scripting, Spark |
| **Data Visualization** | Tableau, Power BI, MicroStrategy, Quicksight, MS Office, Looker, GCP Data Studio |
| **APIs & Web Services** | RESTful API, JSON, JAXB, XML, WSDL, Soap UI |
| **Machine Learning & Statistics** | Cross Validation |
| **Software & Tools** | JMeter, ElasticSearch, Logstash, Kibana, Spring, Hibernate, Apache Airflow, Oozie, WebSphere, Splunk, Tomcat, Linux, Red Hat, Salesforce |

**PROFESSIONAL EXPERIENCE**

**Sr. Data Engineer**

**Ascena Retail Group | Pataskala, Ohio November 2022 to Present**

**Responsibilities:**

* Developed and maintained ETL processes using AWS Glue and Apache Kafka to ensure efficient data flow and storage across various platforms including Amazon S3 and Data Lake architectures.
* Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift
* Engineered and managed scalable, efficient data pipelines and ETL processes within the Cloudera Data Platform (CDP) environment, enhancing data flow and storage solutions across the organization.
* Automated data workflows using Apache Airflow and Oozie, ensuring efficient and error-free data processing cycles.
* Developed and optimized data pipelines for integrating external data sources into graph-based systems, leveraging tools like Apache Kafka and Spark for real-time and batch processing.
* Orchestrated data pipeline automation and monitoring using Jenkins, ensuring continuous integration and deployment (CI/CD) of data-driven applications.
* Engineered and optimized SQL queries and Spark scripts to perform complex data analysis, enhancing data retrieval efficiency and supporting data-driven decision-making.
* Utilized Python and NumPy for data manipulation and analysis, enabling the extraction of meaningful insights from structured and unstructured data.
* Wrote AWS Lambda functions which invokes python scripts to perform various transformations and analytics on large data sets in EMR clusters.
* Created multiple Glue ETL jobs in Glue Studio and then processed the data by using different transformations and then loaded it into S3, Redshift and RDS.
* Designed and implemented complex ETL workflows using Informatica to extract, transform, and load data from diverse sources such as SQL Server, Oracle, and flat files into AWS Redshift.
* Managed the migration of large volumes of data from legacy systems to AWS Redshift using Informatica’s optimized connectors.
* Employed in Spark Streaming for real-time data processing, enabling instant data analysis and insights.
* Developed complex data pipelines using Apache Flink integrated with AWS services (such as AWS Kinesis, S3, and EMR) to facilitate seamless data flow and storage.
* Used AWS EMR to transform and move large amounts of data into and out of other AWS data stores and databases, such as Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB.
* Creating Lambda functions with Boto3 to deregister unused AMIs in all application regions to reduce the cost for EC2 resources.
* Leveraged Java and Scala to build robust, scalable applications and services within the Apache Flink ecosystem, ensuring best practices in code quality and performance.
* Implemented stateful stream processing applications in Flink to manage real-time analytics and event-driven data processing.
* Wrote advanced SPARQL queries for extracting insights and facilitating complex analytical tasks within graph databases.
* Leveraged AWS cloud services to deploy and manage scalable data infrastructure, improving system reliability and performance.
* Successfully migrated on-premises data platforms to Cloudera Data Platform (CDP) Public Cloud in AWS, ensuring a seamless transition and optimized data processing capabilities.
* Configured and managed Cloudera Stack, including HBase, Hive, Impala, and Pig, to support big data ecosystems and analytics applications.
* Managed HDFS systems, ensuring data integrity, scalability, and accessibility in distributed computing environments.
* Used DBT to test the data (schema tests, referential integrity tests, custom tests) and ensure data quality and to debug complex chains of queries. They can be split into multiple models and macros that can be tested separately.
* Administered MySQL and Cassandra databases, optimizing data storage, retrieval, and management processes.
* Utilized Tableau, MicroStrategy, and Quicksight for data visualization, presenting complex data in an easily understandable format for business stakeholders.
* Implemented Athena and Redshift for efficient data querying and analysis in cloud environments, supporting scalable analytics solutions.
* Developed and enforced data quality frameworks using Spark and ETL processes, ensuring data accuracy and reliability.
* Optimized event time processing workflows in Apache Flink, ensuring high-throughput and low-latency data streaming capabilities.
* Optimized HBase and Impala configurations for high-performance data querying, supporting real-time analytics and decision-making.
* Developed RESTful API services using JAX-RS, Spring, and Hibernate to facilitate seamless data integration and exchange between systems.
* Managed XML, JSON, JAXB, and WSDL for data interchange and service-oriented architecture (SOA) implementations, facilitating system interoperability.
* Configured NiFi flows for efficient data routing, transformation, and system integration, enhancing operational efficiency.
* Implemented data indexing and search solutions using ElasticSearch, Logstash, and Kibana (ELK), enhancing data visibility and accessibility.
* Utilized ELK/Splunk for data logging and analysis, enhancing system monitoring and operational intelligence.
* Integrated Terraform scripts within GitLab CI/CD pipelines to automate infrastructure provisioning as part of the deployment process.

**Environment:** AWS Glue, Apache Kafka, Amazon S3, AWS Lambda, DynamoDB Data Lake, Cloudera Stack, HBase, Hive, Impala, Pig, NiFi, Spark, Spark Streaming, SnowSQL, Snowflake, DBT, Redshift, Athena, SQL, Spark 1.6/2.0, Pyspark, Apache Flink, Java, Scala, Python, NumPy, Tableau, MicroStrategy, Quicksight, Terraform, Jenkins, ELK/Splunk, MySQL, Cassandra, HDFS, RESTful API, JAX-RS, Spring, Hibernate, JSON, JAXB, XML, WSDL, ElasticSearch, Logstash, Kibana, Apache Airflow, Oozie, Golang, SPARQL.

**Sr. Data Engineer**

**Truist Bank | Charlotte, NC December 2021 to October 2022**

**Responsibilities:**

* Developed scalable data pipelines in GCP Dataflow and GCP Dataproc for real-time and batch data processing, ensuring timely and accurate financial reporting.
* Developed scalable data pipelines to support real-time processing of digital wallet transactions, leveraging GCP Dataflow and Apache Beam for efficient batch and stream data processing.
* Involved in migration of data from Hadoop environments to Google Cloud Platform, employing tools such as GCP Dataproc and Dataflow to handle real-time and batch data processing.
* Developed and maintained GCP Databricks environments for collaborative data science and engineering projects.
* Developed data integration solutions using Sqoop and Cloud Dataflow, streamlining data exchange between disparate data sources.
* Utilized Pyspark and Hadoop for processing large datasets, improving the efficiency of data analysis tasks.
* Engineered and maintained SQL Database and MongoDB systems for optimized data storage and retrieval, supporting various banking operations.
* Managed and optimized Hadoop clusters focusing on HBase, Hive, and Impala to support robust big data operations.
* Utilized Hive and Sqoop for data aggregation and transformation, facilitating seamless data integration across platforms.
* Developed Python scripts for data manipulation and analysis, automating routine data processing tasks.
* Conducted thorough performance tuning of Apache Flink jobs to optimize resource utilization and processing speed.
* Employed C programming for system-level applications, optimizing data processing routines.
* Leveraged GCP's BigQuery for fast, economical, and scalable analytics, enabling effective data-driven insights.
* Configured Cloud Spanner and Cloud SQL for highly available and scalable database services, supporting critical banking applications.
* Managed Snowflake cloud data warehouse, optimizing data storage and computation for financial analytics.
* Implemented Cloud Pub/Sub for event-driven data integration, enhancing data availability and accessibility.
* Utilized Cloud Storage Transfer Service for efficient data migration between different cloud storage services.
* Implemented GCS for secure and scalable cloud storage solutions, ensuring data availability and disaster recovery.
* Designed conceptual, logical, and physical data models to accurately represent digital wallet transactions, customer interactions, and financial data.
* Engineered financial data models using GCS and BigQuery, facilitating advanced data analysis and reporting.
* Collaborated with cross-functional teams to design and deploy scalable and resilient data architecture supporting high-volume data ingestion and analysis.
* Automated data cleansing and quality checks using GCP Dataprep, ensuring high data integrity and reliability.
* Implemented CI/CD pipelines for the digital wallet’s backend systems, using tools like Jenkins and Spinnaker to automate testing and deployment processes.
* Created dynamic visualizations using Power BI, providing actionable insights into financial trends and patterns.
* Automated financial reports generation using Power BI and GCP Data Studio, enhancing reporting efficiency and accuracy.
* Implemented Terraform for managing GCP infrastructure, automating the provisioning of resources such as BigQuery, Cloud Storage, and Dataflow.
* Managed data security and compliance within GCP and Cloud SQL environments, adhering to financial industry regulations.
* Leveraged Kafka and MQ for building scalable, fault-tolerant messaging and streaming platforms, facilitating efficient data ingestion and real-time analytics.
* Utilized Data Flow for stream and batch data processing, optimizing financial data analysis and insights.

**Environment:** GCP, Pyspark, SAS, Hive, Sqoop, Teradata, GCP DataProc, BigQuery, Hadoop, Kafka, Flink, GCS, Python, C, Snowflake, Power BI, Data Flow, SQL Database, MongoDB, Databricks, GCP Dataprep, Cloud Composer, Cloud Pub/Sub, Cloud Storage Transfer Service, Cloud Spanner, Cloud SQL, Data Catalog.

**Data Engineer**

**AbbVie | Vernon Hills, IL July 2018 to November 2021**

**Responsibilities:**

* Engineered data integration pipelines using Azure Data Factory, streamlining data flow from research datasets to Azure Data Lake, supporting immunology research projects.
* Developed and maintained robust data processing frameworks using Azure Databricks, facilitating advanced analytics on clinical trial data.
* Implemented secure data exchange and APIs with Azure Data Factory, facilitating seamless data integration across research platforms.
* Managed and optimized Azure Storage solutions for secure and scalable storage of large-scale genomic data, enhancing data accessibility for oncology research.
* Optimized data query performance and analysis using Azure Data Lake and Azure Databricks, supporting fast-paced research and development.
* Employed Terraform for infrastructure as code (IaC) management, automating cloud infrastructure provisioning and ensuring reproducibility.
* Automated environment setup and application deployment using Docker and Kubernetes, reducing setup times for data processing environments.
* Orchestrated and managed containerized applications using Kubernetes, enhancing deployment processes, scaling, and system resilience across cloud environments.
* Utilized Ansible for configuration management, ensuring consistent environments across development, testing, and production.
* Wrote and optimized Python scripts for data manipulation and analysis, extracting insights from complex biomedical data.
* Developed Shell Scripting routines for automating data processing tasks, reducing manual effort and increasing efficiency in data management.
* Utilized Golang for developing high-performance data processing tools, enhancing data throughput for large-scale data sets.
* Automated deployment and configuration processes using Ansible and Jenkins, ensuring reliable and efficient application updates within research environments.
* Managed build and deployment pipelines using Maven and GIT, streamlining code integration and version control for data engineering projects.
* Developed PowerShell scripts for automation and configuration tasks, enhancing operational efficiency in cloud and on-premises environments.
* Ensured data pipeline integrity and security through continuous integration and delivery (CI/CD) practices using Jenkins and GIT.
* Configured and maintained MySQL databases, supporting structured data storage for research findings and clinical data.
* Implemented ElasticSearch for fast and scalable search capabilities across vast repositories of research documents and data.
* Administered Linux and Red Hat servers hosting data-intensive applications, ensuring high availability and performance for data analysis tools.
* Managed WebSphere and Tomcat servers, ensuring robust hosting environments for web-based research data applications.
* Integrated Splunk for log management and analysis, monitoring data processing pipelines and ensuring system health.
* Implemented JMeter and Kafka for real-time data ingestion and processing, improving data quality and speed for gastroenterology research outcomes.
* Integrated Splunk for real-time monitoring and analytics of data operations, enhancing visibility and insights into data processing performance.
* Automated testing and validation of web services using Soap UI, ensuring data integrity and reliability in research data exchange.

**Environment**: Azure Data Factory, Azure Data Lake, Azure Storage, Azure Databricks, JMeter, Kafka, Ansible, Jenkins, Docker, Maven, Linux, Red Hat, GIT, Kubernetes, Python, Shell Scripting, MySQL, ElasticSearch, Golang, WebSphere, Splunk, Tomcat, Soap UI, Terraform, PowerShell.

**Data Analyst**

**NetEnrich Technologies Pvt. Ltd. | Hyderabad, India December 2016 to March 2018**

**Responsibilities:**

* Developed Spark applications within Databricks to extract, transform, and aggregate data from various file formats using Spark-SQL, enabling in-depth analysis of customer usage, consumption trends, and behavior.
* Demonstrated proficiency in dimensional modeling, encompassing Snowflake schema, Star schema, transactional modeling, and Slowly Changing Dimension (SCD), contributing to robust model construction.
* Engaged actively in model development by identifying, collecting, exploring, and cleansing data, ensuring its quality and relevance for modeling purposes.
* Conducted thorough data cleaning and scaling operations to bolster data quality and prepare it for further analysis.
* Developed statistical models for diagnostics, prediction, and prescriptive solutions, operating in both distributed and standalone environments.
* Applied Python libraries including NumPy, Scikit-learn, and Matplotlib for data analysis, visualization, interpretation, and reporting of key insights.
* Designed and implemented NoSQL database schemas using MongoDB, optimizing for performance, scalability, and reliability.
* Managed MongoDB clusters, ensuring high availability, efficient indexing, and optimal shard configuration for distributed data processing.
* Led the migration of legacy systems to MongoDB, ensuring seamless data transfer, integrity, and consistency across different storage systems.
* Integrated MongoDB with various data sources and applications using ETL processes, facilitating real-time data synchronization and analytics.
* Leveraged leading text mining, data mining, and analytical tools, alongside open-source software, to conduct comprehensive research.
* Developed and maintained complex data models in NoSQL environments, addressing the needs for high-speed transactions and large-scale data storage.
* Utilized MongoDB’s aggregation framework for data analysis and reporting, optimizing queries for faster response times and reduced server load.
* Optimized ETL procedures and implemented appropriate transformations to enhance data migration performance, aligning with project requirements.
* Employed Cross Validation, Log Loss Function, ROC Curves, and AUC for feature selection, ensuring rigorous evaluation of models' effectiveness.
* Generated dummy variables for specific datasets to facilitate regression analysis and improve model accuracy.
* Showcased strong data visualization skills using tools such as Matplotlib and the Seaborn package.
* Utilized Tableau to craft visually engaging data visualizations, dashboards, and comprehensive reports, effectively communicating findings to both the team and stakeholders.

**Environment**: NumPy, Pandas, Tableau, MongoDB, ETL, Cross Validation, Python.

**PalTech**

**Data Analyst | Hyderabad, India August 2014 to November 2016**

**Responsibilities:**

* Facilitated communication between IT Technical Teams and end-users, acting as a liaison to understand and convey specific needs and requirements effectively.
* Utilized advanced data analysis techniques to predict variations aligned with market demands, contributing to informed decision-making.
* Developed a deep understanding of product knowledge, enabling accurate estimation of product costs for clients.
* Interpreted and analyzed results using various techniques and tools, ensuring comprehensive comprehension of data outcomes.
* Played a pivotal role in supporting the data warehouse by aligning and revising reporting requirements.
* Conducted test runs, implemented the latest software updates, and contributed to strategic decision-making processes.
* Monitored daily activities and performance using Salesforce reports and analysis, ensuring operational efficiency.
* Expanded understanding of ETL tools, pipelining, and data warehousing to enhance overall data management capabilities.
* Automated ETL transformations and executed complex SQL queries, resulting in a 40% improvement in report generation, data preparation, and predictive analytics for business growth.
* Proactively troubleshooted database report maintenance issues, ensuring smooth and uninterrupted data operations.
* Prepared detailed and comprehensive reports using Tableau, facilitating easy comprehension of project status and outcomes.
* Created presentations and dashboards using Tableau, MS Excel, and other MS tools to effectively meet client requirements.

**Environment**: R, SQL Script, Salesforce, Tableau, ETL Pipelines, Data Warehouse, MS Office.

Education:

MLR Institute of Technology, Hyderabad, India – Bachelors of Technology in Information Technology (2010 – 2014)