**Jeevan G Big Data Engineer** [**g1gullapalli19@gmail.com**](mailto:g1gullapalli19@gmail.com)

(551) 265-1799

* 9+ years of SAP BW experience in Analysis, Design, Development and around 5 years in Big Data technologies like Spark, Map reduce, Hive Yarn and HDFS including programming languages like Java, and Python.
* Migration of on-premise data (Oracle/SQL Server/Db2) to Azure Data Lake Store (ADLS) using Azure Data Factory (ADF)
* Designed and developed medium to large scale BI solutions on Azure using Azure Data Platform services (Azure Data Lake, Data Factory, Azure Storage Explorer, Logic Apps, Azure SQL DW, Azure Key vault, API Connections.
* Design, development, and implementation of performant ETL pipelines.
* Created 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
* Strong experience building data pipelines and performing large-scale data transformations.
* In-Depth knowledge in working with Distributed Computing Systems and parallel processing techniques to efficiently deal with Big Data.
* Firm understanding of Hadoop architecture and various components including HDFS, Yarn, Map reduce, Hive, Pig, HBase, Kafka, Oozie etc.,
* Strong experience building Spark applications using PySpark and python as programming language.
* Good experience troubleshooting and fine-tuning long running spark applications.
* Extensive hands-on experience tuning spark Jobs.
* Experienced in working with structured data using HiveQL, and optimizing Hive queries.
* Good experience working with real time streaming pipelines using Kafka and Spark-Streaming.
* Strong experience working with Hive for performing various data analysis.
* Detailed exposure with various hive concepts like Partitioning, Bucketing, Join optimizations, Ser-De’s, built-in UDF’s and custom UDF’s.
* Good experience working with Cloudera, Hortonworks and AWS big data services.
* Strong experience using and integrating various AWS cloud services like S3, EMR, Glue Metastore, Athena, and Redshift into the data pipelines.
* Strong experience of leading multiple Azure Big Data and Data transformation implementations in various domains.
* Expertise in configuring and installation of SQL, SQL advanced Server on OLTP to OLAP systems on from high end to low-end environment.
* Strong experience in performance tuning & index maintenance.
* Detailed exposure on Azure tools such as Azure Data Lake, Azure Data Bricks, Azure Data Factory, HDInsight, Azure SQL Server, and Azure DevOps.
* Experience in analyzing, designing, and developing ETL Strategies and processes, writing ETL specifications.
* Created 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.
* Developing and migrating on-premises databases to Azure Data Lake stores using Azure Data Factory.
* Proficient knowledge and hand on experience in writing shell scripts in Linux.
* Experienced in requirement analysis, application development, application migration and maintenance using Software Development Lifecycle (SDLC) and Python/Java technologies.
* Excellent technical and analytical skills with clear understanding of design goals and development for OLTP and dimensions modeling for OLAP.
* Adequate knowledge and working experience in Agile and Waterfall Methodologies.
* Defining user stories and driving the agile board in JIRA during project execution, participate in sprint demo and retrospective.
* Done POC on newly adopted technologies like Apache Airflow and Snowflake and GitLab.
* Have good interpersonal, communication skills, strong problem-solving skills, explore/adopt to new technologies with ease and be a good team member.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **Big Data Technologies** | Hadoop, HDFS, Hive, Apache Spark, Apache Kafka |
| **Databases** | Azure SQL Warehouse, Azure SQL DB, Azure Cosmos No SQL DB, Teradata, Vertica, RDBMS, MySQL, Oracle, Microsoft SQL Server  Oracle, MySQL, SQL Server, MongoDB, DynamoDB, Cassandra, Snowflake. |
| **Programming Languages** | Python, PySpark, Shell script, SQL, Java. |
| **Cloud** | ADFv2, Blob Storage, ADLS, Delta Lake, AZURE SQL DB, SQL server, Azure Synapse, Azure Analysis Services, Databricks, Dataflows, AZURE Cosmos DB, Azure Stream Analytics, Logic Apps, Event Grid, Azure DevOps, ARM Templates  AWS (EC2, EMR, Lambda, IAM, S3, Athena, Glue, Cloud Watch, RDS, Redshift) |
| **Tools** | PyCharm, Eclipse, CI/CD , Visual Studio, SQL\*Plus, SQL Developer, SQL Navigator, SQL Server Management Studio, Eclipse, Postman. |
| **Version Control** | Bit bucket, GIT |
| **Operating Systems** | Windows 10/7/XP/2000/NT/98/95, UNIX, LINUX, OS |
| **Visualization/ Reporting** | Tableau, MS Excel, Power BI |
| **Database Modeling** | Dimension Modeling |

**PROFESSIONAL EXPERIENCE:**

**Client:** Convergint, Chicago, IL Mar’22 – Present

**Sr. Data Engineer**

**Environment:** Azure Data Factory (V2), Azure Data bricks, Python, SSIS, Azure SQL, Azure Data Lake, Azure Blob Storage, Spark 2.0, Hive.

* Contributed to the development of PySpark Data Frames in Azure Data bricks to read data from Data Lake or Blob storage and utilize Spark SQL context for transformation.
* Experience in Creating, developing, and deploying high-performance ETL pipelines with Pyspark and Azure Data Factory.
* Responsible for estimating the cluster size, monitoring, and troubleshooting of the Spark data bricks cluster.
* Worked on an Azure copy to load data from an on-premises SQL server to an Azure SQL Data warehouse.
* Worked on redesigning the existing architecture and implementing it on Azure SQL.
* Experience with Azure SQL database configuration and tuning automation, vulnerability assessment, auditing, and threat detection.
* Integration of data storage solutions in spark – especially with Azure Data Lake storage and Blob snowflake storage.
* Developed ETL pipelines in and out of data warehouse using a combination of Python and Snowflakes Snow SQL Writing SQL queries against Snowflake.
* Improving the performance of Hive and Spark tasks.
* Worked on cloud point to identify the best cloud vendor based on a set of strict success criteria.
* Used Hive queries to analyze huge data sets of structured, unstructured, and semi-structured data.
* Created Hive scripts from Teradata SQL scripts for data processing on Hadoop.
* Developed Hive tables to hold processed findings, as well as Hive scripts to convert and aggregate heterogeneous data.
* Created and utilized sophisticated data types for storing and retrieving data in Hive using HQL.
* Used structured data in Hive to enhance performance using sophisticated techniques including as bucketing, partitioning, and optimizing self joins.
* Created a series of technology demos utilizing the Confidential Edison Arduino shield, Azure Event Hub, and Stream Analytics, to show the possibilities of Azure Stream Analytics.
* Leveraged Terraform modules to enable consistent deployment of Azure Synapse resources across multiple environments, ensuring configuration and compliance consistency.
* Developed custom Terraform modules to automate deployment of Azure Synapse Analytics and Azure Synapse Studio workspaces, enabling fast, repeatable deployments for data engineering and analytics teams.
* Created Terraform scripts to integrate Azure Synapse with other cloud services, including Azure Data Factory, Azure Stream Analytics, and Azure Databricks, to build end-to-end data processing pipelines.
* Utilized Terraform to enable automated scaling and provisioning of Azure Synapse resources, including scaling up and down compute resources based on workload demand.
* Developed Terraform scripts to configure and deploy Azure Synapse security features, including network security groups (NSGs), virtual network service endpoints, and Azure Private Link integration.
* Experienced in writing Terraform code to deploy and manage cloud infrastructure resources in Azure.

**Client:** Walmart, Bentonville, AR Feb’20 - Feb’22

**Sr. Data Engineer**

**Environment:** S4 HANA, SAP ECC,Azure Data Factory (ADF), Azure Databricks, Azure Storage, Azure Blobs, and Azure Data Lake Storage Gen 2, Azure Repos, Azure Storage Explorer and Azure Data Studio, SQL Server, Azure SQL and MySQL

* Participate in design and development of Big Data analytical applications.
* Design, support and continuously enhance the project code base, continuous integration pipeline, etc.
* Provide and implement data pipeline solutions / proof of concepts involving Microsoft Azure cloud services as-in Data bricks, Data factory, PySpark.
* Collaborate with Data Warehouse Implementation teams, BI Administrators, Developers and Analysts for successful development of BI reporting and analytic solutions.
* Deployment of Library on Data bricks by setting up new Jobs and configuring Automated Cluster to Install Library by adding dependency.
* Design, develop, implement and support reporting and analytics applications leveraging tools such as Congo’s, Tableau, SPSS, R, and Python to meet client objectives and requirements.
* Develop, Implement and manage dashboards, processes and procedures for cross-application and cross-function control reporting using Tableau BI tool.
* Worked on NoSQL / SQL, Micro service, RESTful API development
* Designed and created Hive external tables using shared eta store instead of derby partitioning, dynamic portioning and buckets.
* Involved in analyzing business, system and data mapping requirements and developing ETL data pipelines for real-time streaming of data using Kafka and Spark.
* Developed PySpark scripts using Python on Azure HDInsight for Data Aggregation, Validation, verifying its performance over MR jobs and to extract the data from the web server output files to load into HDFS.
* Implemented data ingestion from various source systems using Sqoop and Pyspark.
* Hands on experience implementing Spark and Hive jobs performance tuning.
* Built pipelines to move hashed and un-hashed data from Azure Blob to Data lake and wrote AZURE POWERSHELL scripts to copy or move data from local file system to HDFS Blob storage
* Involved in SQOOP implementation which helps in loading data from various RDBMS sources to Hadoop systems.
* Proficiency in using Terraform providers and modules to interact with Azure APIs
* Expertise in designing and implementing scalable and reliable infrastructure solutions using Terraform
* Understanding of cloud security best practices and ability to implement security controls using Terraform
* Knowledge of Git and experience with version control systems for managing Terraform code
* Ability to troubleshoot Terraform code and infrastructure issues
* Proficiency in using Terraform CLI for managing Terraform state, creating and updating Terraform configurations, and executing Terraform plans and applies
* Experience in using Terraform workspaces to manage multiple environments, such as development, testing, and production
* Understanding of Azure Resource Manager (ARM) templates and ability to translate ARM templates into Terraform code
* Provided automation and deployment of applications inside software containers by providing an additional layer of abstraction and automation of operating system level virtualization on Linux using Dockers, Kubernetes.
* Responsible for high performance of data architecture and design including Star Schemas, Snowflake Schemas, and Dimensional Modeling.

**Client:** Caterpillar, Peoria, IL Feb’19 – Jan’20

**Big Data Engineer**

**Environment:** Spark, Hadoop, YARN, HTML, Python, Data bricks, Kubernetes JDBC, TERADATA, NOSQL, Sqoop, MYSQL.

* Optimizing existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frames and Pair RDD.
* Used Hadoop technologies like spark and hive including using the Pyspark library to create spark data frames and converting them to normal panda’s data frames for analysis.
* Played a key role in migrating Hadoop clusters on Azure and defined different read/write strategies.
* Designed and built a Data Lake using Hadoop and its ecosystem components.
* Developed Spark, Python for regular expression (regex) projects in the Hadoop/Hive environment with Linux/Windows for big data resources.
* Worked with data investigation, discovery and mapping tools to scan every single data record from many sources.
* Imported millions of structured data from relational databases using Sqoop import to process using Spark and stored the data into HDFS in ORC format.
* Executed multiple Spark SQL queries after forming the Database to gather specific data corresponding to an image.
* Developed prototype for Big Data analysis using Spark, RDD, Data Frames and Hadoop eco system with CSV, JSON, and distributed files.
* Implemented data ingestion from various source systems using Sqoop and Pyspark.
* Hands on experience implementing Spark and Hive jobs performance tuning.
* Have knowledge on partition of Kafka messages and setting up the replication factors in Kafka Cluster and Implemented to reprocess the failure messages in Kafka using offset id.
* Reviewed Kafka cluster configurations and provided best practices to get peak performance.
* Designed and Implemented Error-Free Data Warehouse-ETL and Hadoop Integration.
* Enhancements to conventional data warehouses based on the STAR schema, data model updates, and Tableau data analytics and reporting.
* Evaluated the performance of Data bricks environment by converting complex Redshift scripts to spark SQL as part of new technology adoption project.

**Client:** Honda, Marysville, OH                                                     Oct ’17 – Jan’19

**Hadoop Engineer**

**Environment:** Hadoop, Hive, J2EE, JDBC, Pig 0.16, HBase 1.1, Sqoop, NoSQL, XML, Spark, PL/SQL, HDFS, JSON.

* Designed cutting edge technical solutions for client needs, after evaluating options, including cloud-based solutions.
* Designed a cost-effective archival platform for storing big data using Hadoop and its related technologies.
* Provided technical architecture services to client, typically in the context of solutions that have been defined.
* Created Data Lake by extracting customer's data from various data sources into HDFS. This includes data from Teradata, Mainframes, RDBMS, CSV and Excel.
* Worked on all aspects of data mining, data collection, data cleaning, model development, data validation, and data visualization.
* Worked on programming foundation DAX within Power BI / Python for ETL.
* Worked with Azure BLOB and Data Lake storage and loading data into Azure SQL Synapse analytics (DW).
* Involved in design and development of Data transformation framework components to support ETL process, which gets the Single Complete Actionable View of a Customer.
* Developed an ingestion module to ingest data into HDFS from heterogeneous data sources.
* Built distributed in memory applications using Spark and Spark SQL to do analytics efficiently on huge data sets.
* Developed and built frameworks/prototypes that integrate Big Data and advanced analytics to make business decisions.
* Worked closely with project manager to develop work plan for Data Warehouse project and keep the manager aware of any issues.
* Provided analytical network support to improve quality and standard work results.
* Supported the development of performance dashboards that encompass key metrics to be reviewed with senior leadership and sales management.

**Client:** Evoqua, Warren dale, PA                                                              Dec ’16 – Oct’ 17

**Client:** Pfizer, Collegeville, PA Nov ’14 – Nov’16

**Client:** Oakley, Lake Forest, CA Jun ’13 – Oct ‘14

**Client:** BMS, Princeton, NJ Jan ’12 – May ’13

**Client:** John Maneely Company Steel Group, Sharon, PA Aug ’10 – Dec ‘11

**Client:** GBSC /GPSG, (Johnson and Johnson), Horsham, PA Oct ’07 – Jun’10

**Education:**

Master’s in computer science – 2007, FDU, Teaneck, NJ, USA.