# Teja Govindula – Data Engineer

# tejagovindula@gmail.com, (330)945-0629.

# Master’s in computers and Information Sciences, Kent State University with 3.78 GPA.

**Professional Summary:**

* Versatile Data Engineer with over 7 years of experience specializing in Cloud, AWS services, big data technologies, ETL, data modeling, databases, and other key techniques.
* Extensive AWS expertise, utilizing more than 25 services to build efficient data pipelines by integrating Lambda, Step Functions, Glue, Kinesis, EMR, and Kafka.
* Strong practical experience with Azure Cloud Services (PaaS & IaaS), including Azure Synapse Analytics, SQL Azure, Data Factory, Azure Analysis Services, Application Insights, Azure Monitoring, Key Vault, and Azure Data Lake.
* In-depth knowledge of Dimensional Data Modeling, Relational Data Modeling, Star and Snowflake schemas, FACT & Dimensions Tables and both Physical & Logical Data Modeling and Data Analysis.
* Proven skills in database design and development for business intelligence, employing OLAP cubes, Star and Snowflake schemas, SQL Server Integration Services (SSIS), and SQL Server Analysis Services (SSAS).
* Proficient in Python for data extraction and manipulation with extensive use of libraries such as NumPy, Pandas, and Matplotlib for data analysis.
* Experienced in developing web applications using Python, Django, C++, XML, CSS, HTML, JavaScript, and jQuery.
* Significant expertise in Talend, executing ETL methodologies including Data Profiling, Migration, Extraction, Transformation, and Loading.
* Hands-on experience in Data Engineering disciplines, including Data Lakes, Data Warehouses, Reporting and Analytics with a focus on Data Storage, Data Ingestion, Batch Processing, Stream Processing, and Real-Time Message Ingestion.
* Successfully developed Spark tasks to execute various data transformations on source data using Spark DataFrame and Spark SQL APIs.
* Proficient in creating reports and dashboards using Tableau and Power BI visualization tools.
* Familiar with data warehousing and ETL tools, including Informatica.
* Skilled in writing complex SQL and PL/SQL for table, view, index, stored procedure, and function design, with experience in databases like Teradata and Oracle.
* Experienced with version control tools such as Bitbucket, Git, and SVN.
* Knowledgeable in NoSQL databases like MongoDB, Cassandra, and HBase.
* Experienced in creating relational databases, including Oracle, DB2, MySQL, and MSSQL Server, and writing SQL queries, stored procedures, functions, packages, tables, views, and triggers.
* Experienced in developing Hadoop-based applications using HDFS, MapReduce, Spark, Hive, Sqoop, HBase, and Oozie.
* Knowledgeable in real-time data streaming technologies such as Spark Streaming and Kafka.
* Experienced in CI/CD processes using containerization technologies like Docker and Kubernetes.
* Understanding of both Waterfall project management and Agile/Scrum development methodologies.
* Proficient in various operating systems, including Windows, Linux, UNIX, and Ubuntu.

**Technical Skills:**

**ETL Tools:** AWS Glue, Azure Data Factory, Airflow, Spark, Sqoop, Flume, Apache Kafka, Spark Streaming, Informatica, Talend

**Programming and Scripting:** Spark Scala, Python, Java, MySQL, PostgreSQL, Shell Scripting, Pig, HiveQL

**NoSQL Databases:** MongoDB, Cassandra, Amazon DynamoDB, HBase

**Data Warehouse:** AWS RedShift, Google Cloud Storage, Snowflake, Teradata

**SQL Databases:** Oracle DB, Microsoft SQL Server, IBM DB2, PostgreSQL, Teradata, Amazon RDS

**Monitoring Tools:** Splunk, Chef, Nagios, ELK

**Containerization:** Docker, Kubernetes, OpenShift

**Hadoop Distribution:** Cloudera, Hortonworks, Map R, AWS EMR

**AWS:** EC2, S3, Glacier, Redshift, RDS, EMR, Lambda, Glue, CloudWatch, Recognition, Kinesis, CloudFront, Route53, DynamoDB, Code Pipeline, EKS, Athena, Quick Sight

**Hadoop Tools:** HDFS, HBase, Hive, YARN, MapReduce, Pig, HIVE, Apache Storm, Sqoop, Oozie, Zookeeper, Spark

**IDE’s**: IntelliJ, Eclipse, Spyder, Jupyter.

**Professional Experience:**

**Client: Flex Ltd., Buffalo grove IL.**

**Role: AWS DATA ENGINEER**  July 2018 – Present

**Responsibilities:**

* Worked on setting up and configuring AWS's EMR Clusters and used Amazon IAM to give users granular access to AWS resources. Used AWS Data Pipeline to schedule an Amazon EMR cluster to clean and process web server logs stored in an Amazon S3 bucket.
* Design and Develop ETL Processes in AWS Glue to migrate data from external sources like S3 and Parquet/Text Files into AWS Redshift.
* Created Python AWS Lambda functions for AWS's Lambda service, which runs Python scripts on massive data sets in EMR clusters to conduct various transformations and analyses.
* Made Snowflake Schemas by constructing a Sub Dimension called Demographic as a subset of the Customer Dimension and properly normalizing the dimension tables.
* Involved in migrating 5 data pipelines to use data bricks, even building CI/CD (push flows) to integrate with the existing Jenkins pipeline.
* Involved in end-to-end development and automation of ETL pipelines using SQL and Python.
* Supported Tableau reporting for 250+ ESG reporting metrics and scores for business-critical decisions.
* Made of PySpark to create data processing jobs, such as receiving data from external sources, merging data, doing data enrichment, and loading into target data destinations.
* Developed PySpark code that creates data frames from raw layers in Avro format and writes them to internal tables of the data service layer in orc format using Spark SQL.
* Created a logical and physical data model for Snowflake and specified virtual warehouse sizing for Snowflake for various workload types.
* Built Enterprise ingestion Spark framework to ingest data from different sources (Salesforce, Excel, SFTP, FTP, and JDBC Databases), which is 100% metadata driven and 100% code reuse which lets Junior developers concentrate on core business logic rather than spark/Scala coding.
* Codified Teradata BTEQ scripts to load, transform and clean up duplicate data, as well as remedy errors like SCD-2 data chaining.
* Used Informatica power center to Extract, Transform and Load data into Netezza Data Warehouse from various sources like Oracle and flat files.
* ETL jobs were designed and created to extract data from the Salesforce replica and load it into the Redshift data mart.
* Developed Airflow DAGs in python by importing the Airflow libraries.
* Created a separate topic for reading data from Kafka and used it for real-time data ingestion.
* Utilized Jira as project management methodology and Git for version control to build the program.

**Environment:** AWS EMR, S3, Redshift, Lambda, Boto3, Dynamo DB, Amazon Sage Maker, Apache Spark, Apache Kafka, RDBMS, Python, SQL, Snowflake, ETL, PySpark, Python, Tableau

**Client*:* Anthem Inc., Chicago IL.**

**Role: Data Engineer Dec 2017 – June 2018**

**Responsibilities:**

* Implementing Infrastructure as a Code (IaaC) using CloudFormation templates and configuring and integrating the necessary AWS services in accordance with business requirements.
* Processed batch and streaming data load pipeline utilizing Snow Pipe and Matillion from data lake's Confidential AWS S3 bucket while working on Snowflake Schemas and Data Warehousing.
* Built Informatica Snowflake pipelines to ingest data from SQL servers for running transactions
* Created ETL-type SCD-2 Informatica mappings to load dimensional data from SAP S&D to EDW.
* Designed and built ETL methods to handle the progressively varying Type 1 and Type 2 logic for dimension support while loading data into fact and dimension tables.
* Involved in ingesting financial data from 20+ vendors like FactSet, Bloomberg, MSCI, S&P, etc., into data lake by integrating Vendor API with Data Lake Infra for batch, real-time processing.
* Designed 3NF data models for OLAP and dimensional data using star and snowflake Schemas.
* Created AWS Glue jobs and optimized PySpark jobs for various data jobs.
* Developed conceptual, logical, and physical models for Star/Snowflake schema implementations in OLTP, Data Warehouse, Data Vault, and Data Mart.
* Performed Data Cleaning, features scaling, and features engineering using pandas and NumPy packages in python.
* Worked on ETL Migration services by developing and deploying AWS Lambda functions to generate a serverless data pipeline that can be written to Glue Catalog and queried from Athena.
* Undergone Databricks training to migrate existing spark jobs in EMR to Databricks by fine-tuning jobs to leverage Advanced data bricks capabilities such as delta lake and delta engine etc.
* Used Snow pipelines to leverage claims and enrollment data in near real-time reporting by consuming it from suppliers (Institutional, Financial, and independent).
* Involved heavily in data modeling and warehousing with advanced patterns such as cohesive data models and data harmonization patterns.
* Monitored and optimized the usage of warehouses, automatic clustering, and Snow pipes based on business needs and reduced the cost by 15%.
* Performed data blending, data preparation for Tableau consumption using Alteryx and SQL and publishing data sources to Tableau server.
* Worked on Creating Airflow Dag’s, YAML Parameters/Scripts to set up new CI/CD pipeline
* Environment: Snowflake, AWS S3, Lambda, SAP data, Tableau, Airflow, Teradata, Stored Procedures, Matillion, Python, Snowpipe, Snow SQL, Informatica Cloud (IICS), Teradata, Oracle, Power BI, MS SQL Server.

**Environment:** Hadoop (HDFS, MapReduce), Databricks, Spark, Talend, Impala, Hive, PostgreSQL, Jenkins, NiFi, Scala, Mongo DB, Cassandra, Python, Pig, Sqoop, Hibernate, spring, Oozie, Autoscaling, Scala, Azure, DynamoDB, UNIX Shell Scripting

**Client*:* Johnson Controls, Chicago, IL.**

**Role: Data Engineer**  **Dec 2016 – Nov 17**

**Responsibilities:**

* Involved in the creation of a dependable and scalable data pipeline as well as the requirements gathering, design, analysis, and testing of client specifications during all stages of the software development life cycle (SDLC).
* Created data pipelines utilizing Linked Services/Datasets/Pipeline in Azure Data Bricks and Azure Data Factory to extract, transform, and load data from a variety of sources, including Azure SQL, Blob Storage, and Azure SQL Data warehouse.
* Azure Databricks data processing and input into one or more Azure Services (Azure Data Lake, Azure Storage, Azure DW).
* Working with Data Governance and Data Quality to design various models and processes.
* Streaming analytics are performed in Databricks using Spark Streaming on ingested data in mini batches that have undergone RDD transformations.
* Designed and implemented Scala programs that transform and act on input data using Spark data frames and RDDs.
* Automated the data processing with Oozie to automate data loading into the Hadoop Distributed File System.
* Designing and creating Oracle PL/SQL and Shell scripts, data conversions, data cleaning, and import/export functions.
* Participated in all phases of the project and its scope, using MDM as a reference, and produced a Data Dictionary and a Mapping from Sources to targets in the MDM Data Model.
* Maintain and work with our data pipeline that transfers and processes several terabytes of data using Spark, Scala, Python, Apache Kafka, Pig/Hive & Impala.
* Writing Pig scripts to create Map Reduce jobs and carrying out ETL operations on the HDFS data
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, and Scala.
* Performed data analysis with Cassandra using Hive External tables.
* Established CI/CD tools such as Jenkins and Git Bucket for code repository, build, and deployment of the python code base.
* Used GitHub version control tool to push and pull functions to get the updated code from the repository.
* Configured Spark Streaming to receive real-time data from Kafka and store the stream data to HDFS.
* Designing and Developing Apache NiFi jobs to get the files from transaction systems into data lake raw zone.
* Created action filters, parameters, and calculated sets for preparing dashboards and worksheets using PowerBI.

**Environment:** Hadoop (HDFS, MapReduce), Databricks, Spark, Talend, Impala, Hive, PostgreSQL, Jenkins, NiFi, Scala, Mongo DB, Cassandra, Python, Pig, Sqoop, Hibernate, spring, Oozie, Autoscaling, Scala, Azure, DynamoDB, UNIX Shell Scripting

**Client: Applab Systems Inc., Edison NJ.**

**Role: Automation Engineer**  **Feb 2016 – Nov 2016**

**Responsibilities:**

* Performed requirement-based testing, user acceptance testing (UAT), negative testing, GUI testing, system testing, integration testing, performance testing, stress testing, end-to-end testing, and back-end testing on multiple operating systems. Used tools such as JIRA, JMETER, LOADRUNNER, VSTS, and SoapUI.
* Worked as an ETL Tester, responsible for requirements/ETL analysis, ETL testing, and designing flow and logic for data warehouse projects.
* Designed and developed automation scripts using Selenium WebDriver, including UAT regression automation scripts and functional testing frameworks. Proficient in data-driven and XPath locator approaches.