# Srinivas Narukulla

# Rishitha.23g@gmail.com

# +1 469 277 7354

# Professional Summary:

* **8+** years of IT experience in a variety of industries, which includes hands on experience in Data Engineering.
* **Accomplished data engineer** with extensive experience in **architecting, developing, and optimizing data solutions.**
* Wide knowledge on **AWS and Azure cloud platforms**.
* Proficient in a wide range of cloud services including **AWS EC2, S3, Redshift, Lambda, Azure Data Lake, Synapse, Data- bricks**, and more.
* Strong expertise in **big data technologies**, leveraging **Hadoop, Spark, Hive, Pig, and Kafka** for processing and analyzing large datasets.
* Skilled in **designing and implementing ETL pipelines** using tools like **AWS Glue, Apache NiFi, and Azure Data Factory**
* to transform and move data between diverse sources.
* Proven ability to develop **data models, perform dimensional data modeling**, and implement **star and snowflake schemas** for efficient data warehousing.
* Proficient in programming languages such as **Python, Scala, and Java**, utilizing libraries like **Pandas, NumPy, and PySpark** for data manipulation and analysis.
* Experienced in **version control** using **Git, SVN, and Bitbucket**, ensuring collaborative development and code management.
* Accomplished in **streamlining workflows** using workflow managers **like Apache Airflow, Oozie, and NiFi** to orchestrate and schedule data tasks.
* Skilled in **designing and implementing data lakes**, leveraging technologies such as **HDFS and Azure Data Lake Storage** to store and manage vast amounts of raw data.
* Experienced in collaborating with **data scientists, integrating machine learning models** into data pipelines for real- time predictions and insights.
* Well-versed in **security best practices**, implementing **encryption, access controls, and compliance measures** to protect sensitive data in cloud environments.
* Strong understanding of **containerization and orchestration tools like Docker, Kubernetes, and AWS ECS** for deploying and managing applications.
* Adept at **data visualization** using **Power BI, Tableau, and Google Data Studio** to create insightful dashboards and reports for business stakeholders.
* Skilled in using **automation and configuration management tools** such as **Ansible, Jenkins, and Terraform** to stream- line deployment processes.
* Proficient in working with various **databases** including **SQL Server, Oracle, MySQL, NoSQL databases like MongoDB**, and cloud-native solutions.
* Demonstrated capability in **machine learning models and techniques**, **implementing logistic regression, decision trees, and more** for predictive analytics.
* Highly effective in **agile methodologies**, collaborating with **cross-functional teams, gathering requirements**, and de- levering successful projects.
* Dedicated to **continuous learning**, keeping up with the latest cloud and data technologies to drive innovation and solve complex challenges.

# Technical Skills:

|  |  |
| --- | --- |
| **AWS Services** | EC2, S3, ELB, EMR, Auto scaling, Cloud Front, Cloud Formation, RDS, VPC, Direct Connect, Route 53, Cloud watch, Cloud trail, IAM, SNS, S3, DynamoDB, Athena, Redshift, AWS Lambda, AWS Glue, SQS, SNS, API Gateway |
| **Azure Cloud** | Azure Data Lake, Azure Storage Gen2, Azure Synapse, Azure DevOps, Keyvaults, Event Hub and Logicapp |
| **Big Data Technologies** | Hadoop, HDFS, MapReduce, Hive, Pig, Apache Spark, Sqoop, Impala, Zookeeper, Spark, Flume |
| **Hadoop Distribution** | Cloudera, Hortonworks, MapReduce, AWS EMR |
| **Python Libraries** | NumPy, PyTables, Data Frames, Pandas, Matplotlib, SQL Alchemy |
| **Automation Tools** | Puppet, Chef, Docker, Ansible, Jenkins, Terraform, Kubernetes |
| **Visualization Tools** | Power BI and Tableau |
| **Schedulers** | Airflow, Oozie, NiFi |
| **Data Warehouse** | Azure Synapse, AWS RedShift, and Snowflake |
| **Databases** | Azure SQL Database, Amazon RDS, Oracle DB, Microsoft SQL Server, IBM DB2, Postgres, Teradata, MongoDB, Cassandra, DynamoDB |
| **Version control Tools** | GIT |
| **Data Modelling** | Dimensional Data Modeling, Star Schema, Snowflake Schema. |
| **Programming & Scripting Languages** | Python, Scala, Java, MySQL, PL/SQL, Shell Scripting, HQL, Perl, Ruby, Bash Shell scripting, YAML |

**Education:**

* Bachelors - Computer Science and Engineering (Anurag Group of Institutions) – 2015
* Masters - Master of Science in Computer & Information Science (Southern Arkansas University) – 2020

**Certifications:**

* + PL-300: Microsoft Power BI Data Analyst
	+ Graduate Certificate in Business Analytics

**Professional Experience:**

**Client: Emerson, Austin, TX**

**Senior Data Engineer August 2022- Present**

**Project Description:**

Led data engineering and cloud migration projects, focusing on designing and managing databases in Hive and MySQL, and efficiently handling large datasets using tools like Sqoop and Hadoop ecosystem components. Developed ETL pipelines and serverless data processing solutions with AWS services such as Lambda, Glue, S3, and Redshift, ensuring high-availability and scalability through AWS CloudFormation. Utilized Python, Spark SQL, and Hive for data cleaning, processing, and building data pipelines, alongside developing dashboards with Power BI and Tableau for business analytics. Implemented DevOps practices with Azure DevOps, Jenkins, and Git for continuous integration and deployment, and automated data flows using Apache NiFi and Python-based tools. Migrated on-premises applications to Azure, and orchestrated multi-tier applications leveraging a wide range of AWS services for optimal performance and reliability.

**Roles & Responsibilities:**

* Worked in designing tables in **Hive, MYSQL** using **SQOOP** and processing data like importing and exporting of databases to the **HDFS**, involved in processing large datasets of different forms including structured, semi- structured and unstructured data.
* Cleaned data and processed third party spending data into maneuverable deliverables within specific format with Excel macros and python libraries such as **NumPy, SQL Alchemy** and **matplotlib.**
* Involved in designing and deploying multi-tier applications using all the AWS services like (EC2, Route53, S3, RDS, Dynamo DB, SNS, SQS, IAM) focusing on high-availability, fault tolerance, and auto-scaling in AWS Cloud Formation.
* Transferred the data using Informatica tool from **AWS S3** to **AWS Redshift**.
* Worked on ETL Migration services by developing and deploying **AWS Lambda** functions for generating a serverless data pipeline which can be written to **Glue Catalog** and can be queried from **Athena**.
* Used **AWS Glue** for the data transformation, validate and data cleansing.
* Created monitors, alarms, notifications and logs for **Lambda functions, Glue Jobs, EC2** hosts using **CloudWatch.**
* Designed the Data Marts in dimensional data modelling using star and **snowflake** schemas.
* Programmed in **Hive, Spark SQL, and Python** to streamline the incoming data and build the data pipelines to get the useful insights, and orchestrated pipelines.
* Worked on analyzing Hadoop cluster using different big data analytic tools including Flume, Pig, Hive, HBase, Oozie, Zookeeper, Spark, and Kafka
* Developed dashboards and visualizations to help business users analyze data as well as data insight to upper management with a focus on Microsoft products like **SQL Server Reporting Services (SSRS) and Power BI.**
* Worked on **ETL pipeline** to source these tables and to deliver this calculated ratio data from **AWS.**
* **Extracted** data from the internal sources and **data-driven** workflows and developed **AWS EC2** and **S3** for processing and storage of the migrated data and ingested data from different source systems like **relational** and **non-relational** to meet business requirements.
* Built **Data Models** and **Dimensional Modeling** with Star, and **Snowflake** schemas for **OLAP.**
* Managed and automated the data using **Apache NiFi** to manage the real-time data flow.
* Used **Pandas** as API to put the data as time series and tabular format for manipulation and retrieval of data.
* Generated **Python Django** Forms to record data of online users.
* Developed **cloudformation** templates for most used AWS services like **elasticache, cloudfront, ECS, EKS, Api Gateway**, Transit Gateway,
* Created dynamic Tableau dashboards, trend charts to cater to various business scenarios with integrations from multiple sources - **Snowflake database, excel sheets** and tables.
* Implemented a module to connect and view the status of an Apache Cassandra instance using **Python.**
* Created, and provisioned different EMR clusters, notebooks, jobs, and auto scaling using the **Hadoop** applications and processed the data using PySpark.
* Actively involved in the DevOps streamlining process through **Azure DevOps Services** and Jenkins.
* Created several Spark jobs with **PySpark** to perform several table-to-table operations.
* Recorded the online users' data using Python **Django** forms and implemented test case using Pytest.
* Wrote UDF's in **Scala** and **Pyspark** to meet specific business requirements.
* Responsible for writing **OOP** code in Python for server-side scripting
* Involved in working with **Python open stock API’s.**
* Migrated on premises applications to **Azure cloud** and implemented **CICD** with Azure DevOps service.
* Developed a fully automated continuous integration system using **Git, Jenkins, MySQL,**
* and custom tools developed in **Python** and **Bash**
* Involved in building database Model, APIs, and Views utilizing Python technologies to build web-based applications.

**Environment:** AWS, Python, CI/CD, PySpark, Django, Pandas, NumPy, SQLAlchemy, Matplotlib Apache Scala, Kafka, Snowflake, Hadoop, Hive, HDFS, ETL, Git, Jenkins, Oozie, Apache, Sqoop, Zookeeper, Pytest, SQL, DB2, Flume, Power BI, MySQL, EC2, S3, AWS Lambda, DynamoDB, AWS Redshift HTML5/CSS3, OLAP, Agile

**Client: Visa, Austin, TX**

**Senior Data Engineer August 2020– July 2022**

**Project Description:**

Developed and optimized data pipelines for medical image processing and analytics using Python, Spark, and Scala, integrating real-time data with Kafka and HBase for efficient data management. Utilized AWS services such as Lambda, Glue, and CodePipeline to automate deployment and data processing workflows, and migrated applications to Azure cloud, leveraging Terraform for infrastructure automation. Implemented advanced analytics and data ingestion pipelines in Snowflake, and created visualizations with Power BI. Employed Spark SQL and Hive for big data transformations, ensuring high performance and scalability. Developed RESTful APIs with Python/Django, and maintained databases with PostgreSQL and Flask, supporting seamless data access and integration. Used Agile methodologies and managed project tasks through JIRA for efficient delivery.

**Roles & Responsibilities:**

* Developed Data pipelines using **Python** for medical image pre-processing, Training and Testing.
* Involved in the development of data ingestion, aggregation, integration, and advanced analytics using **Snowflake.**
* Developed several **Spark/Scala** scripts for data extraction from various sources and provided data insights and reports as per need.
* Wrote **Spark SQL scripts** to run over imported data and existing RDDs and implemented Spark practices partitioning, caching, and check pointing.
* Developed code from scratch using **Scala** according to the technical requirements.
* Loaded all data into **Hive** from source CSV files using **Spark.**
* Worked on extracting Real-time data using **Spark Streaming** and **Kafka, converted it to RDD, processed it into Data Frame,** and load the data into **HBase**.
* Maintained and configured the data using **Zookeeper** and tracked the nodes in the **Kafka cluster**.
* Used **Spark SQL** to read the **Parquet data** and loaded tables in the **Hive** to **Spark** using Scala.
* **Developed Spark jobs** using Scala and **Spark SQL** for faster testing and data processing.
* Worked on deriving **Structured Data** from **Unstructured** data received using **Spark.**
* Created automated pipelines in AWS Code Pipeline to deploy Docker containers in AWS ECS using services like **Cloud Formation, Code Build, Code Deploy, S3**
* Developed Cloud Formation templates, also launched AWS Elastic Beanstalk for deploying, monitoring, and scaling web applications using different platforms like **Docker, Python**
* Was responsible for creating on-demand tables on S3 files using Lambda Functions and AWS Glue using **Python** and PySpark.
* Implemented Infrastructure on **Azure** platform and migration of existing Linux and Windows-based Infrastructure to AWS and Azure Cloud platforms and automated using **Terraform** Templates.
* Migrated applications to **Azure** cloud using **Azure** migrate through Lift and shift.
* Implemented Enterprise Integration architecture using **Azure Service Bus, AppService, Active Directory, Storage, hybrid connection manager, active directory** authentication for azure SQL server and other offerings by azure.
* Used Agile methodology for rewriting existing **Python/Django** modules to deliver certain format of data and responsible for debugging the project monitored on **JIRA.**
* Worked on importing and exporting data from **snowflake, Oracle** and **DB2** into **HDFS** and **HIVE** using **Sqoop** for analysis, visualization and to generate reports.
* Managed the imported data from different data sources, performed transformation using **Hive, Pig** and **Map- Reduce** and loaded data in **HDFS.**
* Used various sources to pull data into Power BI such as SQL Server, Excel, Oracle, etc.
* Responsible for Running the Workflows for ETL process and verifying the data in target database by **ETL** process.
* Created functions in **AWS Lambda** to run python scripts and AWS Lambda using java to perform event driven processing.
* Used Python scripts to update content in the database and manipulate files and generated Python **Django** Forms.
* Designed and maintained databases using Python and developed Python-based API using **Flask** and **PostgreSQL**
* Created required shell or python scripts for data push from **Informatica/Teradata** to S3
* Worked on **big data (Hadoop)** environment along with exposure to **HIVE, Spark, Cassandra, SQL** and **ETL**
* components.
* Developed frontend and backend modules using Python on Django including Web Framework using **Git** and **GitHub**
* Developed **Restful API’s** using Python for customer care system which can be used to easily access customer and product data.
* Used **AWS Glue** for the data transformation, validate and data cleansing.

**Environment:** Python, Scala, Django, Hive, Pig, Sqoop, Map- Reduce, HDFS, Restful API, Power BI, Airflow, Azure, Data Factory, Spark, Kafka, Hive, CI/CD, AWS Lambda, S3, IAM, Dynamo DB, JIRA, Git, GitHub, Docker, Jenkins, JSON, SQL, Zookeeper, Snowflake, Tableau, SVN, Informatica, Teradata, Flask, PostgreSQL

**Discover Financial, Chicago,IL**

**Data Engineer Nov 2018 – July 2020**

**Project Description:**

Developed and managed data pipelines utilizing AWS services such as EC2, S3, EMR, and Lambda for real-time transaction data processing and storage. Implemented Kafka for real-time data ingestion and Spark Streaming for near real-time data transformations and aggregations. Created REST APIs using Python with Flask, integrating diverse data sources including RDBMS and Hadoop ecosystems like Spark, Hive, and HBase. Employed Snowflake for advanced data warehousing and performed data cleaning and feature engineering with Python libraries such as pandas and numpy. Automated infrastructure deployment using Ansible and Google Cloud Deployment Manager, and managed ETL processes with SSIS and AWS Glue. Applied Agile methodologies to streamline project development and delivery.

**Roles & Responsibilities:**

* Worked on various AWS services like **EC2, Cloud Formation, code commit, code pipeline, CloudWatch, SNS, S3 bucket, Auto scaling, Load balancer.**
* This project will download the data that was generated by real-time transaction activities. The data will be pro- cessed and stored in the **AWS EC2** and **S3** and collected into the **HDFS** system online aggregators by Kafka.
* Worked on creating **Kafka** producer and **Kafka** consumer for **Spark streaming** which gets the data from different learning systems of the patients.
* **Spark Streaming** collects this data from Kafka in **near real time** and performs necessary **transformations** and ag- gregation on the fly to build the common learner data model.
* Used **Spark Streaming** to divide streaming data into batches as an input to the Spark engine for batch processing.
* Worked with **AWS** to spin up the **EMR** cluster to process the huge data which is stored in **S3** and push it to target.
* Implemented Spark SQL to access hive tables into spark for faster processing of data.
* Developed **Rest API's** using Python with flask framework and done the integration of various data sources including Java, JDBC, RDBMS, Shell Scripting, Spreadsheets, and Text files.
* Developed Data pipeline using **Spark, Hive, Impala, and HBase** to ingest customer behavioral data and financial histories into Hadoop cluster for analysis.
* Implemented **Agile** Methodology for building an internal application.
* Expertise in **Snowflake** to create, Maintain Tables, and views.
* Performed Data Cleaning, features scaling, features engineering using pandas and numpy packages in Python.
* Created various complex **SSIS/ETL** packages to Extract, Transform and Load data.
* Used **Ansible** for configuration management of hosted Instances within AWS
* Configuring and Networking of **Virtual Private Cloud (VPC)**
* Used various python libraries such as Matplotlib for charts and graphs, **PySide, PyMongo, PySpark, Pickle, Panda’s data frame, MySQL** for database connectivity.
* Experienced with event-driven and scheduled **AWS Lambda** functions to trigger various AWS resources.
* Used **AWS Glue** catalog with crawler to get the data from **S3** and perform sql query operations.
* Automated the infrastructure in Google Cloud by using Deployment Manager Templates for various services in
* Built and engineered servers on Ubuntu and **RHEL Linux**

**Environment:** AWS, Python, Django, Restful API, SSIS, ETL, Spark, Kafka, CI/CD, AWS Lambda, AWS Glue, EC2, S3, IAM, CLI, EKS, CloudWatch, DynamoDB, JIRA, Git, GitHub, PySide, PyMongo, PySpark, Pickle, Pandas data frame, MySQL, Docker, Jenkins, Ansible, VM Ware ESX, Docker, Kubernetes, SQL, Snowflake, Tableau, SVN, Informatica, Teradata, ETL, Agile, RHEL, Linux.

**Cisco Systems, Hyderabad, India**

**Data Engineer July 2015 – July 2018**

**Project Description:**

Developed scalable distributed data solutions on Cloudera Hadoop, integrating various Hadoop components like Oozie, Pig, Hive, and Sqoop for efficient data processing and management. Collaborated with business and technology teams to redesign ETL analytics applications into Big Data solutions, implementing Hadoop security using Kerberos. Utilized Jenkins for CI/CD processes and managed data integrity via JSON-Schema. Enabled seamless data import/export between Teradata, Oracle, HDFS, S3, and Snowflake, and developed Scala and Java-based MapReduce programs for data parsing and staging. Administered Hive Metastore and HDFS using Hue WebUI, ensuring robust data infrastructure and user management.

**Roles & Responsibilities:**

* Responsible for building scalable distributed data solutions on **Cloudera distributed Hadoop**.
* Worked with **Business and technology teams** for redesigning ETL Analytics applications to **Big Data Analytics applications**.
* Implementing Hadoop security solutions Kerberos for securing **Hadoop clusters**
* Integrated **Oozie** with the rest of the Hadoop stack supporting several types of Hadoop jobs **Map Reduce, Pig, Hive and Sqoop** as well as system specific jobs such as Java programs and Shell scripts.
* The managed full data mine from the huge data volumes is exported to **MySQL using Sqoop**.
* Involved in design of **CI-CD** process using **Jules/ Jenkins** and other supporting frameworks to deploy the code into the server.
* designed and implemented data integrity via **JSON-Schema**.
* Developed solutions for import/export of data from **Teradata, Oracle to HDFS, S3 and S3 to Snowflake**.
* Developed **Scala code** using specific monad pattern for different calculations based on the requirement.
* Configured **Hive Metastore to use MySQL** database to establish multiple user connections to hive tables.
* Maintained company **SQL Server** databases to ensure integrity of data.
* Performed administration using **Hue WebUI** to create and manage user spaces in **HDFS**.
* Developed Map Reduce programs in **Java** for parsing the raw data and populating **staging Tables**.
* Designing and implementing complete end-to-end Hadoop Infrastructure including **Pig, Hive, Sqoop, Oozie,** Flume, and Zookeeper.
* Further used pig to do **transformations, event joins, elephant bird API and pre -aggregations** performed before loading JSON files format onto HDFS.

**Environment:** CDH 5.4.5, Hive1.2.1, HBase1.1.2, Flume1.5.2, Map Reduce, Sqoop1.4.6, Kanban, Spark2.1.0, Kafka, Shell Script, Oozie 4.2.0, Zookeeper 3.4.6.