 YUVA SRI ANUMOLU



`

.

**CONTACT**

|  |
| --- |
| Irving, TX 75063 |
| +1 469 706 0833 |
| [yuvasrianumolu84@gmail.com](mailto:yuvasrianumolu84@gmail.com) |
| [yuvasrianumolu](https://www.linkedin.com/in/yuvasrianumolu/)  https://github.com/YuvasriAnumol u/DE-portfolio |

**TOP SKILLS**

* Spark
* Hadoop
* SQL
* AWS
* Azure
* Python
* Java
* Snowflake

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**EDUCATION**

Master of Science in Computer Science

Fitchburg State University, MA

* With 7 years of IT experience, I specialize in Hadoop Administration across Cloudera (CDH), Hortonworks (HDP), Vanilla Hadoop, and MapR, alongside expertise in AWS, Kafka, Python, Elastic search, DevOps, and Linux Administration.
* I handle Hadoop Cluster installation, configuration, and support.
* My expertise extends to the Hadoop ecosystem, including HDFS, Yarn, MapReduce, Hive, HBase, Sqoop, Kafka, Spark, Oozie, NiFi, and Cassandra, with a focus on optimizing Hive and Spark SQL queries and jobs.
* Proficient in AWS services include EC2, S3, RDS, VPC, IAM, Elastic Load Balancing, Auto Scaling, CloudFront, CloudWatch, SNS, SES, and SQS, I process vast structured and unstructured data using Apache Spark, Hadoop, and Hive.
* Additionally, I have experience with NumPy, Pandas, and Matplotlib, and I'm skilled in version control systems like GIT and Bitbucket.
* In AWS, I handle data management, HDFS maintenance, and structured/unstructured data loading.
* I orchestrate Spark and Hadoop jobs, and ingest data to Hive using Oozie and AWS Glue. Proficient in Spark APIs like RDD, DataFrames, and Spark SQL, I specialize in Spark streaming, tuning, and debugging. I create technical documents and work with Hive for table creation, data distribution, and optimization.
* I'm experienced in various Hadoop file formats and creating efficient partitions in Hive tables.
* My proficiency extends to Hadoop distributions include Cloudera, Hortonworks, and in-depth understanding of Hadoop Architecture, including YARN, HDFS, Resource Manager, Node Manager, NameNode, and DataNode.
* I also work with SPARK Scala Streaming, Spark SQL, and PySpark, and migrate data from databases like MySQL and Oracle to HDFS using Sqoop.
* Utilizing Kafka, I stream, analyze, and process real-time logs from Apache application servers, storing processed log files in Amazon S3 buckets.
  + I also have hands-on experience with Jenkins for CI/CD process implementation, and I'm well-versed in SDLC methodologies includes Waterfall, Rapid Application Development, and Agile.

**Technical Proficiency:**

|  |  |
| --- | --- |
| Operating System | Red Hat Linux, Ubuntu and Windows |
| Big Data | Spark, Hadoop, YARN, Hive, HBase, Sqoop, Zookeeper, Databricks and  Oozie |
| Streaming | Kafka and Spark DStreams. |
| Cloud | AWS Glue, S3, EMR, EC2, Athena, Redshift Spectrum and Kinesis |
| Languages | Java8, Python, SQL |
| IDE/Tools | Eclipse, Intellij, SQL Developer |
| Database/D.B  Languages | Oracle 10g/11g, SQL and MySQL |
| Version Controls | Git, Bit bucket |
| Software  Methodologies | Agile Methodology Scrum and Waterfall |

**Professional Experience:**

**Centene Corporation, St. Louis, MO Dec 2022 – Present Sr. Data Engineer**

**Responsibilities:**

* + - I've developed Spark applications in Scala and Spark-SQL to analyze customer usage patterns across various data formats.
    - I optimize Spark performance, use Jupyter notebooks and Spark-Shell for development, and create tables on S3 with Lambda Functions and AWS Glue using Python and PySpark.
    - Responsibilities include loading event logs from Kafka into HBase, building data pipelines with Flume, Sqoop, and Pig, and leveraging AWS Glue for data transformation and validation.
    - Using Python Boto3, I configure AWS services like Glue, EC2, and S3, and develop workflows in Nifi for automating data loading into HDFS.
    - I optimize Hadoop algorithms with Spark Context, Spark-SQL, Data Frames, and Pair RDDs, exporting processed data to RDBMS with Sqoop.
    - In AWS, I utilize EC2, ELB, S3, RDS, and CloudWatch, participate in code reviews, and create DDLs for warehouse tables.
    - Leading process enhancements and conducting cross-team trainings, I automate workflows with Apache Airflow and shell scripting, scheduling DAGs for Hive and Pig jobs.
    - Finally, I perform Exploratory Data Analysis and Visualization with Python and Tableau, and install and configure Apache Airflow for S3 and Snowflake data warehouse.

**Environment:** Hadoop, AWS EMR, S3, Redshift, Map Reduce, Spark, Spark MLlib, Kafka, HBase, HIVE, PIG, Scala, Python, Java, Nifi, SQL, Cassandra, Oracle, MongoDB, DB2, T-SQL, PL/SQL.

**State Street, Boston, MA Apr 2021 – Nov 2022 Data Engineer**

**Responsibilities:**

* + - I develop data pipelines using AWS services like Lambda, Glue, and EMR to ingest structured data from various sources, including SaaS platforms.
    - With AWS services like EMR, Redshift, S3, Athena, Aurora Postgres RDS, and Glue, I orchestrate data pipelines using Apache Airflow.
    - I create Lambda functions to ingest API data into S3, converting nested JSON objects to Parquet files with Glue jobs.
    - Glue Crawler dynamically adjusts JSON schema changes, while Glue streaming jobs ingest Kafka data, converting AVRO to Parquet.
    - Spark on EMR performs transformations, aggregating data for Redshift staging tables.
    - I use Step Functions to process API data through Glue for reporting.
    - Great Expectations validates raw data in S3 before loading into Redshift. DDL statements create databases, tables, and views in Redshift and Aurora RDS. Automation involves transient EMR cluster creation/termination with Apache Airflow, loading data into S3 buckets, and moving curated data to Redshift with Apache Airflow JDBC operators.
    - Terraform configures AWS services in Dev, and CI/CD pipelines move DDL and Airflow DAGs from Dev to QA.
    - Lastly, I explore revenue forecasting using Azure Machine Learning Workspace. **Environment:** S3, Lambda, Glue Catalog, Glue Streaming, Redshift, Airflow, Aurora, Spark, Erwin, SQL Workbench, Kafka, Docker, Terraform, CloudFormation.

**Summit Healthcare, Fort Worth, TX Oct 2019 – Mar 2021 Data Engineer**

**Responsibilities:**

* + - I actively participated in Agile Development, engaging in daily SCRUM meetings for progress updates and task allocation.
    - My responsibilities included scripting for Oracle to Hadoop data imports, developing Kafka services with Spring Cloud Streams, and automating workflows with Oozie Scheduler.
    - I utilized HQL and Native SQL for data retrieval from Hive and databases, deploying AWS Lambda to synchronize MySQL data with the Client Portal.
    - Experience extends to real-time data processing with Spark DStreams, exploring, cleaning, engineering, and evaluating machine learning models using Python libraries like Pandas, NumPy, and Scikit-learn.
    - I developed custom functions and optimized data pipelines using Spark SQL, tuning performance and managing dependencies with Maven.
    - Hive tables were created with various formats and schemas, and I converted queries into Spark transformations using RDDs and Pyspark.
    - Additionally, I managed data ingestion from DAT files into Hive ORC tables, deploying applications with Maven and debugging with Log4J.

**Environment**: Apache Spark, Hdfs, Hive, Pyspark, AWS, Sqoop, Oozie, Oracle, Apache Kafka, Spring, SQL Developer, JUnit and Maven.

**Walking tree Technologies, Delhi, In Aug 2016 – Sep 2019 Python Developer**

**Responsibilities:**

* + - I collected and organized data from various sources, creating a structured database schema for storage.
    - Using Python, I managed different file formats like JSON, XML, CSV, Excel, and PDF, while also maintaining Linux systems through disk usage monitoring scripts.
    - Python scripts were written to retrieve data from Google Analytics and Search Console, generating alerts based on comparisons with historical data.
    - For web data extraction, I utilized Python modules such as Selenium and Beautiful Soup, while also securing Nginx web servers with SSL certificates.
    - I developed tools for bulk video generation and content engineering, integrating third- party applications and managing Django projects with Mercurial version control, deployed on AWS EC2 instances.
    - Python scripts handled disk space monitoring on AWS EC2 instances, integrated websites with analytics tools, and collected data from Amazon crowdsourcing services via APIs.
    - Furthermore, I managed system administration tasks, SEO efforts, and developed backup scripts for PostgreSQL databases, ensuring daily, weekly, and monthly backups were stored on AWS S3.

**Environment**: Python 2.7, Django 1.4, PostgreSQL, HTML, Bootstrap 3, JavaScript, Ubuntu 12.04, Nginx, XML, Google API, Amazon EC2, S3.