**Yuva Sri Anumolu**

[**yuvasrianumolu84@gmail.com**](mailto:yuvasrianumolu84@gmail.com)

**Phone: +1 469 706 0833**

[**www.linkedin.com/in/yuvasrianumolu**](http://www.linkedin.com/in/yuvasrianumolu)

**Irving, Texas**

**Professional Summary:**

* Having over **7 years** of experience as Senior Data Engineer with strong technical expertise, business experience, and communication skills to drive high - impact business outcomes.
* Experience in Software Development Life Cycle (SDLC) including Requirements Analysis, Design Specification and Testing as per Cycle in both Waterfall and Agile methodologies.
* Experience with Spark Core, Spark SQL, Spark MLlib, Spark GraphX and Spark streaming for processing and transforming complex data using in-memory computing capabilities written in Scala. Worked with Spark to improve efficiency of existing algorithms using Spark Context, Spark SQL, Spark MLlib, Data Frame, Pair RDD's and Spark YARN.
* Experience in Python and Scala, user-defined functions (UDF) for Hive and Pig using Python.
* Hands-on experience with Hadoop architecture and various components such as Hadoop File System HDFS, Job Tracker, Task Tracker, Name Node, Data Node and Hadoop MapReduce programming.
* Experience with Airflow to schedule ETL jobs to extract the data from AWS data warehouse.
* Experience structural modifications using Map-Reduce, Hive and analyze data using visualization/reporting tools (Tableau).
* Experienced in fact dimensional modeling (Star schema, Snowflake schema), transactional modeling and SCD (Slowly changing dimension).
* Experience working on creating and running Docker images with multiple micro - services.
* Hands on experience working Amazon Web Services (AWS) using Elastic Map Reduce (EMR), Redshift, and EC2 for data processing.
* Experience in developing JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data using the Cosmos Activity.
* Demonstrated proficiency in designing robust data pipeline architectures to support various data processing needs.
* Developed and maintained RESTful APIs to facilitate data access and integration between multiple systems, enhancing overall data accessibility and interoperability within the organization.
* Utilized REST API endpoints for extracting, transforming, and loading (ETL) data from diverse sources into a centralized data warehouse, improving data processing efficiency and streamlining data pipelines.
* Skilled in guiding the integration of model pipelines into production settings.
* Proficient in leveraging Google Cloud Platform (GCP) services such as BigQuery, Cloud Storage, and Dataflow for scalable data processing and analytics solutions, contributing to improved data accessibility and insights for business stakeholders.
* Skilled in following established processes for alerting management when issues arise, maintaining clear communication channels.
* Reviews pipeline designs, making necessary data model design changes, and meticulously documents and reviews these changes with data science teams
* Experienced in deploying and managing data pipelines on Google Cloud Platform (GCP), utilizing services like Compute Engine and Kubernetes Engine to optimize data processing workflows and ensure efficient resource utilization within cloud environments.
* Proficient in data discovery techniques and automated ingestion processes to facilitate efficient model development.
* Skilled in conducting thorough analysis of raw data sources to assess data quality and integrity.
* Facilitates data discovery and automated ingestion for model development, conducting thorough analysis of raw data sources for quality and relevance, while aligning with business context and model development requirements.
* Hands on experience in SQL and NOSQL database such as Snowflake, HBase, Cassandra and MongoDB.
* Proficient in running model monitoring scripts to ensure the stability and performance of machine learning models.
* Extensive experience in agile software development methodology.
* Actively engages in special projects, contributing valuable insights and expertise to achieve project objectives.
* Team Player as well as able to work independently with minimum supervision, innovative & efficient, good in debugging and strong desire to keep pace with latest technologies.

**Technical Skills:**

|  |  |
| --- | --- |
| **Databases** | Snowflake, AWS RDS, Teradata, Oracle, MySQL, Microsoft SQL, Postgre SQL, Data Lakes |
| **NoSQL Databases** | MongoDB, Hadoop HBase and Apache Cassandra. |
| **Programming Languages** | Python, SQL, Scala, MATLAB, Java, Angular JS. |
| **Cloud Technologies** | AWS, Azure, GCP (Docker, Kubernetes for Containerization) |
| **Data Formats** | **CSV, JSON, Parquet, Avro.** |
| **Querying Languages** | SQL, NO SQL, PostgreSQL, MySQL, Microsoft SQL |
| **Integration Tools** | Jenkins, GitLab CI |
| **Scalable Data Tools** | Hadoop, Hive, Apache Spark, Pig, Map Reduce, Sqoop, NiFi, Kafka, DataBricks |
| **Operating Systems** | Red Hat Linux, Unix, Windows, macOS. |
| **Reporting & Visualization** | Tableau, Matplotlib, Superset, PowerBI |

**Professional Experience:**

**Client: Centene Corporation, St. Louis, MO Dec 2022 – Till date**

**Role: Senior Data Engineer**

**Responsibilities:**

* Worked with the business users to gather, define business requirements and analyze the possible technical solutions.
* Responsible for data pipeline issues identified through model monitoring alerts, ensuring smooth functioning of the pipelines.
* Developed Spark scripts by using Python and Scala shell commands as per the requirement.
* Wrote Spark jobs with RDD's, Pair RDDs, Transformations and actions, data frames for data transformations from relational sets.
* Designed and Developed Scala workflows for data pull from cloud based systems and applying transformations on it.
* Responsible for Building Scalable Distributed Data solutions using Hadoop.
* Developed highly complex Python and Scala code, which is maintainable, easy to use, and satisfies application requirements, data processing and analytics using inbuilt libraries.
* Proficient in leveraging PySpark and GlueETL for developing complex data pipelines in distributed computing environments.
* Developed PySpark script to merge static and dynamic files and cleanse the data.
* Developed ETL framework using Spark and Hive (including daily runs, error handling, and logging) to useful data.
* Developed ETL Specification Design document containing detailed information on ETL processing, mapping/workflow specifications, exception handling process, staging and data warehouse schemas, etc.
* Developed ETL code, control files, metadata, and lineage diagrams for ETL programs.
* Developed Tableau data visualization using Cross tabs, Heat maps, Box and Whisker charts, Scatter Plots, Geographic Map, Pie Charts and Bar Charts and Density Chart.
* Implemented data governance practices to ensure compliance with regulatory standards and data security requirements.
* Created tableau dashboards/reports for the business users.
* Developing End to End Analytical environment using Power BI.
* Created basic reports using confidential files as source to fetch the data in Power BI. Designed and developed Power BI graphical and visualization solutions with business requirement documents and plans for creating interactive dashboards.
* Responsible for analyzing and troubleshooting issues to identify root causes and implement effective solutions.
* Automated resulting scripts and workflow using Apache Airflow to ensure daily execution in production.
* Created airflow DAG’s to sync files from box, analyze data quality, and alert for missing files.
* Utilized AWS services with focus on big data architect /analytics / enterprise Data warehouse and business intelligence solutions.
* Designed and deployed Amazon EMR clusters to process and analyze petabytes of data from various sources.
* Implemented complex ETL pipelines using AWS Glue to transform raw data into actionable insights.
* Optimized Spark jobs on Amazon EMR clusters to improve performance and reduce processing time.
* Collaborated with data scientists to integrate machine learning models into big data workflows.
* Written Pig Scripts for sorting, joining, filtering and grouping data.
* Extracted data from Teradata database and loaded into Data warehouse using spark.
* Implemented a Continuous Delivery pipeline with Docker and GitHub.
* Worked on Snowflake Schemas and Data Warehousing and processed batch and streaming data load pipeline using Snow Pipe and Matillion from data lake Confidential AWS S3 bucket.
* Performed analysis on the unused user navigation data by loading into HDFS and writing MapReduce jobs.
* Worked on data pre-processing and cleaning the data to perform feature engineering and performed data imputation techniques for the missing values in the dataset using Python.
* Created scripts to read CSV, JSON and parquet files from S3 buckets in Python and load into AWS S3, DynamoDB and Snowflake.
* Used SQL queries and other tools to perform data analysis and profiling.
* Documented pipeline and data model designs, facilitating clear communication and collaboration across teams.
* Involved in Agile methodologies, daily scrum meetings, spring planning.
* Collaborated closely with data scientists and analysts to understand model requirements and optimize data pipelines accordingly.
* Actively participated and provided feedback in a constructive and insightful manner during weekly Iterative review meetings to track the progress for each iterative cycle and figure out the issues.

**Environment:** Spark, Scala, AWS, GLUE , EMR, ETL, Hadoop, Python, Snowflake,HDFS, Hive, MapReduce, PySpark, Pig, Docker, GitHub, Apache Spark, Teradata, JSON, PostgreSQL, MongoDB, SQL, Agile and Windows.

**Client: State Street, Boston, MA April 2021 – Nov 2022**

**Role: Data Engineer**

**Responsibilities:**

* Interacted with clients to gather business and system requirements which involved documentation of processes based on the user requirements.
* Developed data models optimized for specific use cases, improving data storage efficiency and query performance.
* Developed Spark scripts by writing custom RDDs in Scala for data transformations and perform actions on RDDs.
* Developed Spark jobs on Databricks to perform tasks like data cleansing, data validation, standardization, and then applied transformations as per the use cases.
* Developed highly complex Python and Scala code, which is maintainable, easy to use, and satisfies application requirements, data processing and analytics using inbuilt libraries.
* Designed number of partitions and replication factor for Kafka topics based on business requirements and worked on migrating MapReduce programs into Spark transformations using Spark and Scala, initially done using python (PySpark).
* Performed advanced procedures like text analytics and processing, using the in-memory computing capabilities of Spark using Scala.
* Worked on migrating MapReduce programs into Spark transformations using Scala.
* Developed ETL jobs using PySpark, Data Lineage in which the data has been transformed in multiple stages and actions like aggregations are performed.
* Implemented and maintained multiple ETL processes to synchronize data between different source systems and database.
* Preparing ETL test strategy, designs and test plans to execute test cases for ETL and BI systems.
* Involved on creating multiple kind of Report in Power BI and present it using Story Points.
* Built Ad-Hoc reports in Power BI depending on business requirements was one of the major responsibilities.
* Involved in Creating, Debugging, and Scheduling and Monitoring jobs using Airflow for ETL batch processing to load into Snowflake for analytical processes.
* Developed**PIG** scripts for the analysis of semi structured data.
* Utilized Amazon EMR to process and analyze log data for identifying patterns and anomalies.
* Developed ETL scripts using AWS Glue to clean and prepare data for analysis.
* Created and optimized SQL queries to extract relevant information from structured and semi-structured data.
* Generated reports and visualizations using tools like Amazon QuickSight to present insights to stakeholders.
* Worked on Tableau to build customized interactive reports, worksheets and dashboards.
* Used Tableau to produce the dashboards which will compare the results before using this solution and after using this solution in staging environment.
* Prepared scripts to automate the ingestion process using Python and Scala as needed through various sources such as API, AWS S3, Teradata and snowflake.
* Implemented a Continuous Delivery pipeline with **Docker** and GitHub.
* Created scripts to read CSV, JSON and parquet files from S3 buckets in Python and load into AWS S3, DynamoDB and Snowflake.
* Responsible for modifying the code, debugging, and testing the code before deploying on the production cluster.
* Worked on designing, building, deploying and maintaining Mongo DB.
* Implemented SQL, PL/SQL stored procedures.
* Worked closely with stakeholders to understand requirements and iteratively refine pipeline designs to meet evolving needs
* Involved in story-driven agile development methodology and actively participated in daily scrum meetings.

**Environment:** Spark, Scala, AWS, Amazon Glue , EMR, ETL, Kafka, Tableau, Hadoop, Python, Snowflake,HDFS, Hive, MapReduce, PySpark, Pig, Docker, Sqoop,, Teradata, JSON, MongoDB, SQL, Agile and Windows.

**Client: Summit Healthcare, Fort Worth, TX Oct 2019 – March 2021**

**Role: Data Engineer**

**Responsibilities:**

* Involved in Requirement gathering phase to gather the requirements from the business users to continuously accommodate changing user requirements.
* Developed various spark applications using Scala to perform various enrichment of these click stream data merged with user profile data.
* Designed and implemented scalable data pipelines to process and analyze large volumes of data.
* Designed and Developed Spark workflows using Scala for data pull from AWS S3 bucket and Snowflake applying transformations on it.
* Implemented robust data quality assurance measures within data pipelines to ensure the accuracy, completeness, and consistency of data.
* Developed Simple to complex MapReduce Jobs using Hive and Pig.
* Profile structured, unstructured, and semi-structured data across various sources to identify patterns in data and Implement data quality metrics using necessary query’s or python scripts based on source.
* Worked on PySpark APIs for data transformations.
* Prepared dashboards using Tableau for summarizing Configuration, Quotes, Orders and other e - commerce data.
* Extract, transform, and load (ETL) data from multiple federated data sources (JSON, relational database, etc.) with Data Frames in Spark.
* Migrated an existing on-premises application to AWS. Used AWS services like EC2 and S3 for small data sets processing and storage, Experienced in Maintaining the Hadoop cluster on AWS EMR.
* Designed highly available and fault-tolerant big data architectures using Amazon EMR as the core processing engine.
* Implemented serverless ETL workflows using AWS Glue to ingest and transform data from various sources.
* Integrated Amazon EMR with AWS services such as Amazon S3 and Amazon Redshift for seamless data integration and storage.
* Provided technical leadership and guidance to cross-functional teams in implementing best practices for big data solutions.
* Implemented AWS Elastic Container Service (ECS) scheduler to automate application deployment in the cloud using Docker Automation techniques.
* Analyzed the SQL scripts and designed the solution to implement using PySpark.
* Extracted files from MongoDB through Sqoop and placed in HDFS and processed.
* Use SQL queries and other tools to perform data analysis and profiling.
* Flexibly performs additional duties as assigned, demonstrating versatility and adaptability in fulfilling organizational needs.
* Followed agile methodology and involved in daily SCRUM meetings, sprint planning, showcases and retrospective.

**Environment:** Spark, Scala, AWS, ETL, Hadoop, Python, Snowflake,HDFS, Hive, Tableau, MapReduce, PySpark, Pig, Tableau, Teradata, Docker, JSON, XML, Apache Kafka, SQL, PL/SQL, Agile and Windows.

**Client: Walking tree Technologies, Delhi, In Aug 2016 – Sep 2019**

**Role: Python Developer**

**Responsibilities:**

* Collected data from different sources, normalized the data and designed a database schema to store the data in the database.
* Worked with different file formats JSON, XML, CSV, EXCEL, and PDF to read and structure the data using python libraries. System maintenance Tasks.
* Wrote a Shell script which monitors disk usage on Linux machines and alerts admins when the usage reaches a critical point.
* Wrote Python scripts to fetch data from Google Analytics API and Search Console on a daily basis and compares the metrics with historical data to generate alerts when required.
* Utilized python modules include Selenium, Beautiful Soup, and Urllib to extract and parse the data from multiple websites.
* Installed SSL certificates for the Nginx web server for multiple web applications.
* Developed a Video generation tool to generate bulk videos. The tool takes structured data as input and outputs a video using predefined templates.
* Engineering content generation software employing huge data sets like American community survey, building permits and insurance quotes.
* Created Django projects with Mercurial version control. Developed and maintained web applications and infrastructure to support all the sites.
* Deployed Django projects on AWS EC2 instances running with Ubuntu server 18.04 using uWSGI and NGINX.
* Integrated with third-party apps as and when needed.
* Wrote Python scripts to monitor disk space on AWS EC2 instances which sends an email to alert.
* Integrated all the websites with analytics tools, Search Console to monitor traffic and performance of web pages.
* Wrote Python scripts to collect data from crowdsourcing services like Amazon using API.
* Involved in system admin Tasks like Creating Users, Installation of Software using bash, Taking Backups & Other
* Responsible for Search Engine Optimization (SEO) to improve the visibility of the website.
* Manage day-to-day analytics and recommend new content ideas after a detailed analysis.
* Developed backup scripts using Python and bash which are being used for daily (incremental) weekly and monthly (full) backups for the PostgreSQL database and push the backup file to AWS S3.

**Environment**: Python 2.7, Django 1.4, PostgreSQL, HTML, Bootstrap 3, JavaScript, Ubuntu12.04, Nginx, XML, Google API, Amazon EC2, S3.

**Education:**

Master of Science in Computer Science

FITCHBURG STATE UNIVERSITY |FITCHBURG, MA, US