**Arun Kokkula**

**AWS Lead Data Engineer**

[**arunkokkula5@mail.com**](mailto:arunkokkula5@mail.com)

**Contact : +1(469) 415 9783**

**Linkedin :**   <https://www.linkedin.com/in/arun-kokkula-7b1940159/>

**Work Authorization : H1B (DE)**

**Professional Summary**

* Data Engineer / Spark Developer with 9+ years of IT Industry experience in designing, developing and maintaining large-scale systems/applications with significant expertise in Python 2.7, 3.2 to 312.
* Experienced with full software development life-cycle, architecting scalable platforms, object-oriented programming, database design and agile methodologies
* Expert knowledge of and experience in Object Oriented Design and Programming concepts.
* Experience in writing data processing frameworks for large scale applications
* Hands on experience in using Spark ecosystem components like S3, RDS, Snowflake, REST, Kafka
* Experience in closely working with data analysis and Data scientists for converting POCs into production grade software.
* Experience in consuming data from Kafka into spark micro batches.
* Worked on spark application tuning and resource allocations based on use case
* Experience in Shell Scripting, SQL Server, UNIX and Linux.
* Experience in building applications in AWS infrastructure; Cloud Formation Templates, Cloud Watch Alarms, S3, RDS, Security Groups, VPCs, EC2.
* Familiar with JSON based REST Web services and Amazon Web services.
* Automated the continuous integration and deployments using Jenkins and AWS Cloud Templates deployment services (Lambda).
* Experience with web scraping using beautiful soup; Created own flight logistic application using information from scraped data.
* Experience in project deployment using Jenkins and using web services like Amazon Web Services (AWS) EC2, Cloud Formation Templates, AWS S3 and Cloud watch.
* Exposure to ML/DL ecosystem, algorithmic approaches along with design and performance constraints
* Experience in tools in java as part of personal development.
* Exposure to Tensor flow and PyToch frameworks & hardware for AI.

**Professional skills**

**Programming Languages:** Python, Java, JavaScript

**Web services:** RESTful

**Data bases:** Oracle 10/11g, MySQL, SQL Server

**IDE’s and tools:** Eclipse, Pycharm, NetBeans

**OS & Environment:** XP, windows, Linux, Unix, Ubuntu

**Unix Shell Scripting:** Unix Shell Scripting

**Version control:** GitHub

**Development Methodologies:** Agile, Scrum

**Hadoop:** HDFS, MapReduce, Spark

**Machine Learning:** KNN, Gradient Descent, Back Propagation.

**Professional Experience**

**Vanguard, Philadelphia, PA 04/2022 to Till Date**

**Data Engineer**

Worked on implementing accelerationism; e/acc principles on software development by deriving and maintaining new metrics to track development activities. This involves taking metrics from multiple data sources including remote version control, workday and Jira.

* Responsible for design, build and deploying of data pipeline for telemetry data.
* Created ECS application for capturing development metrics
* Migrated development metrics presented in tableau to aws ecs service
* Created Tableau dashboards for presenting various development metrics.
* Worked on internal CI/CD pipeline bamboo eco system to build different services on the aws platform
* Worked on integrating snowflake data source into pipelines and configure the data source using SQL
* Implemented and optimized large-scale data processing using EMR, demonstrating proficiency in managing and scaling clusters for distributed computing workloads.
* Developed and deployed custom applications on EMR clusters to process, analyze, and visualize big data sets efficiently.
* Created snowflake tables and recovery mechanism for LTC data metrics and designed the SQL queries for LTC metrics.
* Provided support on snowflake data tables and ongoing data quality issues on snowflake tables and SQL scripts.
* Implemented performance tuning strategies to enhance the overall efficiency and speed of data processing jobs on EMR.
* Worked on creating Tableau dashboards for development metrics
* Worked on converting Tableau dashboards to AWS services
* Developed and maintained ETL jobs using AWS Glue for seamless data integration across various data sources.
* Designed and implemented data transformation scripts and jobs to ensure data quality and consistency.
* Automated data cleansing and enrichment processes using AWS Glue to improve data accuracy and usability.
* Build and maintain code repositories and code migrations using CI/CD strategies
* Built modules to consume Honeycomb API data to gather telemetry data using aws lambda
* Preprocessed Call volume data for presentation and exploration at Tableau

**Environment: PSQL, Python, JSON, SNOWFLAKE, AWS, CI/CD, RISK management, REST, Bit Bucket, Agile, Honeycomb, SQL, Tableau, Open Telemetry, AWS Glue, Pyspark, BitBucket.**

**Capital One, Richmond VA 08/2019 to 04/2022**

**Lead Data Engineer**

**Develop and maintain data loading frameworks built in python, Pyspark, scala, shell scripts and SQL. Build tools and systems to handle outages from the upstream and from the internal systems.**

* Design and develop features to the data loading frameworks in PySpark and python.
* Create automation script for data load validation and stability of the system
* Worked on infrastructure maintenance and enhancement
* Review the legacy frameworks and streamline the code base using the parameters
* Resolve infrastructure and design document issues in the job
* Maintain and monitor all the Snowflake data loads across Capital One, and assist in meeting SLA for critical loads.
* Create stand alone snowflake scripts to serve data analysts and independent applications.
* Provided support for 3000 snowflake tables and served the Data analysis department for their data quality issues on SQL.
* Created multiple data processing frameworks using Bash, python and pyspark and loaded data to Snowflake SQL.
* Created a development and production environment by designing tables and data flow for various snowflake SQL.
* Created base tables and view table mechanism for card data on Snowflake tables.
* Implemented snowflake’s time travel features in applications for data transfer and data loading activities.
* Designed and implemented scalable and high-performance data models in DynamoDB, showcasing expertise in NoSQL database design.
* Developed and optimized DynamoDB queries to ensure fast and efficient access to data for various applications.
* Implemented DynamoDB Streams for real-time data processing and integration with other AWS services.
* Provide zero knowledge proofs on user’s Data Quality requests

**Environment:** Linux, Data Lake, Python, Pyspark, JSON, AWS, CI/CD, IOT, REST, GitHub, Snowflake, AWS.

**Caterpillar Innovation Center, Champaign IL 10/2018 to 08/2019**

**Data Engineer II**

Worked on an initiative to build a data lake and an ecosystem of data intensive applications to enable real time insight into telemetric data in manufacturing and on field machines. Build pipelines to transform time series feature engineer time series data for visualization.

* Review design documents on echo system for data-lake and test alternatives for empirical proofs
* Develop python modules to support and develop in built ETL tool for ETL process visitation and development
* Maintain existing linear regression model on telemetric data for pressure sensors using IOT data
* Managed and optimized Redshift data warehouse clusters for high-performance analytics, ensuring efficient query execution and data retrieval.
* Designed and implemented data partitioning and distribution strategies to enhance query performance on Redshift.
* Implemented and maintained ETL processes to load and transform data into Redshift for analytical reporting.

**Environment:** Linux, Rel7, Data Lake, Python, JSON, AWS, CI/CD, IOT, React JS, Seaborne, REST, GitHub, Notebooks, pytorch.

**Capital One, Richmond VA 12/2017 to 09/2018**

**Data Engineer**

Capital One maintains an ecosystem of internal data processing applications designed to prevent fraud of various types including transactional and application fraud with dynamic fraud patterns. Ecosystem of highly available multiple spark clusters running rule on kafka streams and batch data.

* Developed multiplatform PySpark framework used to create spark jobs and provide SQL like interface for data analysts. This application can be used for both creating spark jobs and productionize them.
* Responsible for migrating findings from Data Scientist and fraud investigators into existing fraud defenses. And over seeing Data Science Approaches so that functionalities can be converted to production environments.
* Designed data flow for new business needs and participated in architectural and workflow discussions.
* Responsible for maintaining allocating and tuning resources for faster performance for heavy data loads.
* Resolved Big Data small files issue in the organization using spark Hadoop configurations and dependency configurations.
* Designed infrastructure configurations for spark resource management suited for the data processing needs.
* As part of Capital One Fraud Prevention team, worked on maintaining old and developing new rule engines along with application resiliency strategies. Created new Fraud defenses using existing data processing patterns
* Responsible for developmental and production data security aspects along with upgrading the system with new software and infrastructural features
* Converted exiting traditional Teradata fraud defenses into cloud architecture and spark SQL
* Create QA data by performing data analysis using Databricks notebooks with maintained entropy.
* Configured Apache Arrow for columnar data processing at processors in pyspark 2.3
* Migrate existing fraud defenses from Teradata to PySpark environment. Tune existing spark jobs for performance.
* Created modules for switching cluster stack (active and inactive) using Jenkins, AWS lambda to EC2 cluster.
* Converted monolithic application into a pip installable package and incorporated it to schedule Jupyter Notebooks using papermill
* Built CI/CD pipeline using Jenkins and AWS lambda. Built a secret management system. Built redundancy catch up in fraud case creation.
* Upgraded Spark Hadoop Version and participated in building Custom Assume Role Credential Provider in jar for spark session assume role auto renewal.

**Environment:** Linux, Rel7, PySpark, Data Lake, Hadoop, Teradata, PSQL, Python, JSON, AWS, CI/CD, RISK management, Apache Arrow, Credit Card Fraud Prevention, REST, GitHub, Notebooks, pytorch.

**Delta Air Lines 01/2017 to 12/2017**

**Python Developer**

Built web applications for internal HR management applications. Provided insights into data for Small to Mid-range businesses and present regular findings from data.

* Responsible for the complete SDLC process in gathering requirements, system analysis, design, development, testing and deployment.
* Developed web applications and RESTful web services and APIs using Python, Django and PHP.
* Worked on HTML5, CSS3, JavaScript, AngularJS, NodeJS, Git, REST API, Mongo DB, IntelliJ IDEA.
* Design and Setting up of the environment of MongoDB with shards and replica sets. (Dev/Test and Production)
* Developed and tested many features for the dashboard using Python, Bootstrap and JavaScript.
* Spearheaded adoption of Responsive web design principles and converted existing websites to Responsive websites by using Bootstrap.
* Worked with the team on creating an Ansible script to develop Sonar on an Amazon instance.
* Worked on Python Open stack API and used NoSQL as database and followed Python test-driven development techniques.
* Develop SQL code to be used with automated processes to identify revenue opportunities and financial issues.
* Data mine large datasets in Relational Databases to find emerging issues and root-cause in provisioning, marketing and billing systems
* Drive timely resolution of marketing issues for a seamless customer experience.
* Proactively monitor daily processes and results to ensure consistent coverage.
* Create REST APIs for web services using python Flask, Django frameworks.

**Environment:** Ubuntu, Python, SQL, JSON, AWS, REST, Node.js, MongoDB, Jupyter Notebooks, GitHub.

**HDFC Bank Limited, Mumbai, India**

**Java Developer 01/2013 to 07/2015**

Responsibilities:

Responsible and active in the analysis, definition, design, implementation, management, and

deployment of full software development life cycle of project.

* Developing User Interface and implementing business processes using JSP and Servlets.
* Development of the application that was used to create dynamic JSPs, given input from a database
* table containing information about the contents of the JSP being developed.
* Involved in writing the presentation layer in Java Server Pages (JSP)
* Wrote Servlets programming and JSP scripting for the communication between web browser and
* server.
* Responsible for coding SQL Statements and Stored procedures for back end communication using
* JDBC
* Developed an API to write XML documents from a database. Utilized XML and XSL Transformation for
* dynamic web-content and database connectivity.
* Coded different deployment descriptors using XML. Generated Jar files are deployed on Apache
* Tomcat Server.
* Involved in the development of the presentation layer and GUI framework in JSP. Client-Side validations
* were done using JavaScript. ‘
* Involved in code reviews and mentored the team in resolving issues.
* Participated in weekly design reviews and walkthroughs with project manager and development
* teams.
* Provide technical guidance to business analysts, gather the requirements and convert them into
* technical specifications/artifacts for developers to start.

Environment: HTML, JSP, Servlets, JDBC, JavaScript, Tomcat, Eclipse IDE, XML, XSL, Web logic, Git,

Linux.

**EDUCATION**

* Master of Science in Internet and Web Design, Wilmington University, Wilmington DE, 2017