Renuka Krishna Chandana

Data Engineer

A seasoned Data Engineer with 6 years of comprehensive experience in designing, implementing, and managing data pipelines and infrastructure. Committed to optimizing data workflows and leveraging advanced analytics to empower data-driven decision-making and foster organizational growth.

CONTACT

513-512-5955

rkcthota@gmail.com

EDUCATION

Masters in Computer Science from University of Cincinnati, USA

TECHNICAL SKILLS

- **Hadoop Eco System:** Hadoop, MapReduce, Spark, HDFS, Sqoop, YARN, Oozie, Hive, Impala, Apache Airflow, HBase
- **Programming Languages:** PL/SQL, SQL, Python, PySpark, Scala, and Java.
- **Data Bases:** MySQL, SQL Server, Oracle, MS Access, Teradata
- **NoSQL Data Bases:** Cassandra, HBase, Dynamo DB.
- **Workflow Management tools:** Oozie, Autosys, Apache Airflow
- **Visualization & ETL tools:** Tableau, Power BI, Informatica, Talend
- **Cloud Technologies:** Azure & AWS, GCP.
- **IDE's:** Eclipse, Jupyter notebook, Spyder, PyCharm, IntelliJ
- **Version Control Systems:** Git, SVN, Jenkins, CI/CD
- **Operating Systems:** Windows, Linux, Unix

PROFILE SUMMARY

- Around **6** Years of professional experience in full life cycle system development involving analysis, design development, testing, documentation, implementation & maintenance of application software in Web-based and Client/Server environment.
- Proficiency with **Scala, Apache HBase, Hive, Pig, Sqoop, Zookeeper, Spark, Spark SQL, Spark Streaming, Kinesis, Airflow, Yarn, and Hadoop (HDFS, MapReduce).** Designed, build and managed **ETL** data pipelines leveraging **Airflow, python,** and **GCP solutions.**
- Worked with **Spark** to improve efficiency of existing algorithms using **Spark Context, Spark SQL, Spark MLlib, Data Frame, Pair RDD's** and **Spark YARN.**
- Hands-on experience with **Azure SQL Database, Azure SQL Data Warehouse, Azure Analysis Services, HDInsight, Azure Data Lake** and **Data Factory.**
- Employed the **Agile** paradigm throughout the entire software development life cycle. Expert in creating various **Kafka** producers and consumers for seamless **data** streaming with **AWS services**.
- Experience in implementing **Azure data solutions**, provisioning storage account, **Azure Data Factory**, **SQL server**, **SQL Databases**, **SQL Data warehouse**, **Azure Data Bricks** and **Azure Cosmos DB**.
- Ability to work effectively and efficiently as a team member as well as individually with a desire to learn new skills and technology.
- Expert in designing Parallel jobs using various stages like **Join, Merge, Lookup, remove duplicates, Filter, Dataset, Lookup file set, Complex flat file, Modify, Aggregator, XML.** Good Hand-on Experience in **ETL** processing, **Migration** and data processing using **AWS** services such as **EC2, Athena, Glue, Lambda, S3, Relational Database Service (RDS)** and other data-based services of **AWS.**
- Experienced in building **Snow Pipes,** migrating **Teradata objects** into **Snowflake** environment. Designed and implemented Scalable data architecture on **AWS** using **Kubernetes**, **Terraform**, and **Snowflake**, enabling seamless data integration and processing across multiple data sources.
- Experience in using **Snowflake Clone**, **Time Travel** and **building snow pipe**.
- Experience in automating day-to-day activities by using **Windows PowerShell**.
- Proficient with Container systems like **Docker** and Container orchestration like **EC2** Container Service, **Kubernetes**, worked with **Terraform**.
- Proficient in building **CI/CD** pipelines in **Jenkins** using **pipeline syntax** and groovy libraries. Develop batch processing solutions by using **Data Factory** and **Azure Data bricks**.
- Experience in **Big Data analytics, Data manipulation**, using **Hadoop** Eco system tools **Map - Reduce, Yarn/MRv2, Pig, Hive, HDFS, HBase, Spark, Kafka, Flume, Sqoop, Flume, Oozie, Avro, Sqoop, AWS, Spark integration** with **Cassandra, Avro,** and **Zookeeper.**

WORK EXPERIENCE

Client: Great American Insurance Group, Cincinnati, Ohio, USA (Jun 2023 - Present)
Role: Azure Data Engineer

Description: Great American Insurance Group provides insurance services. I am optimizing data pipelines and queries for performance and efficiency, including identifying and resolving bottlenecks in data processing and storage.

Responsibilities:

- Part of the **Data** and reporting team creating insights and Visualization for the business to make decisions on.
- Designed and deployed a **Kubernetes**-based containerized infrastructure for **data** processing and analytics, leading to a 20% increase in **data** processing capacity. Used **Python** to write **Data** into **JSON** files for testing Django Websites, Created scripts for **data** modelling and **data** import and export.
- Design and configure **data**base, Back-end applications and programs. Managed large **data**sets using **Pandas data** frames and **SQL**. Design and build scalable **data** pipelines to ingest, translate, and analyze large sets of **data**
- Creating job flow using **Airflow** in **python** and automating the jobs. **Airflow** will have separate stack for developing DAGs on and will run jobs on **EMR** or **EC2 Cluster**.
- Responsible for Building and Testing of applications. Experience in handling **data**base issues and connections with **SQL** and **NoSQL data**bases like **MongoDB** by installing and configuring various packages in **python** (**Teradata**, **MySQL**, **MySQL connector**, **PyMongo** and **SQLAlchemy**).
- Using **Azure Cluster** services, **Azure Data Factory** V2 ingested a large amount and diversity of **data** from diverse source systems into **Azure Data Lake Gen2**. Used **Continuous Delivery Pipeline**. Deployed microservices, including provisioning **Azure** environments and developed modules using **Python** scripting and **Shell Scripting**.
- Imported real time weblogs using **Kafka** as a messaging system and ingested the **data** to **Spark** Streaming and did **data** quality checks using **Spark** Streaming and arranged bad and passable flags on the **data**.
- Worked on Big **Data** Integration & **Analytics** based on **Hadoop**, **SOLR**, **PySpark**, **Kafka**, **Storm** and **web Methods**.
- Responsible for estimating the cluster size, monitoring, and troubleshooting of the **Spark Databricks** cluster and Ability to apply the spark **Data** Frame **API** to complete **Data** manipulation within spark session.
- Instantiated, created, and maintained **CI/CD** (continuous integration & deployment) pipelines and apply automation to environments and applications. Worked on various automation tools like **GIT, Terraform, Ansible.**
- Handled importing of **data** from various **data** sources, performed transformations using **B, loaded data** into **HDFS** and Extracted the **data** from **SQL** into **HDFS** using **Sqoop**.
- Working on **data** management disciplines including **data** integration, modeling and other areas directly relevant to business intelligence/business analytics development.
- Supported development of Web portals, completed **Database Modelling** in **PostgreSQL**, front end support in **HTML**/**CSS**, **jQuery**. Developing scalable and reusable **data**base processes and integrating them.
- Performed **ETL** to move the **data** from source system to destination systems and worked on the **Data warehouse**.
- Designed and implemented Infrastructure as code using **Terraform**, enabling automated provisioning and scaling of cloud resources on **Azure**. Involved in **data** validations and reports using **PowerBI**.
- Implemented **Python** automation for Capital **Analysis** and Review, leveraging **Pandas** and **NumPy** modules to manipulate and analyze **data**, ensuring accurate reporting and streamlined decision - making.
- Looked into existing **Java**/**Scala spark** processing and maintained, enhanced the jobs.
- Developed Monitoring and notification tools using **Python**. Developed **Python Spark** modules for **Data ingestion** & analytics loading from **Parquet**, **Avro**, **JSON data** and from **database tables**.
- Developed **Spark** applications with **Azure Data Factory** and **Spark**-**SOL** for **data** extraction, transformation, and aggregation from different file formats to analyze and transform the **data** to uncover insights into customer usage patterns.
- Developed analytical components using **Scala**, **Spark**, **Apache Mesos** and **Spark** Stream and Installed **Hadoop**, **Map Reduce**, and **HDFS** and developed multiple **MapReduce** jobs in **PIG** and **Hive** for **data** cleaning and pre-processing.

Environment: Azure, Oracle, Kafka, Python, Informatica, SQL Server, Erwin, RDS, NOSQL, Snowflake Schema, MySQL, Bash, Dynamo DB, PostgreSQL, Tableau, Git Hub, Linux/Unix

Client: Procter & Gamble, Cincinnati, Ohio, USA (Nov 2022 – May 2023)
Role: AWS Data Engineer

Description: The Procter & Gamble Company (P&G) is an American multinational consumer goods corporation. I implemented and enforced data governance policies and security measures to ensure compliance with regulatory requirements (such as GDPR or HIPAA) and protect sensitive data from unauthorized access or misuse.

Responsibilities:

- Used **AWS** to create storage resources and define resource attributes, such as disk type or redundancy type, at the service level.
- The **AWS Lambda** functions were written in **Spark** with cross - functional dependencies that generated custom libraries for delivering the **Lambda** function in the cloud. Performed raw **data** ingestion into, which triggered a lambda function and put refined **data** into **ADLS**.
- Designed and setup Enterprise **Data Lake** to provide support for various uses cases including Analytics, processing, storing and Reporting of voluminous, rapidly changing data.
- Responsible for maintaining quality reference data in source by performing operations such as cleaning, transformation and ensuring Integrity in a relational environment by working closely with the stakeholders & solution architect.
- Designed and developed **Security Framework** to provide fine grained access to objects in **AWS S3** using **AWS Lambda, DynamoDB**. Set up and worked on Kerberos authentication principals to establish secure network communication on cluster and testing of **HDFS**, **Hive**, **Pig** and **MapReduce** to access cluster for new users.
- Performed end-to-end Architecture & implementation assessment of various **AWS** services like **Amazon EMR**, **Redshift**, **$3.** Implemented the machine learning algorithms using python to predict the quantity a user might want to order for a specific item so we can automatically suggest using kinesis firehose and **S3 data lake**.
- Used **AWS EMR** to transform and move large amounts of data into and out of other **AWS** data stores and databases, such as **Amazon Simple Storage Service** (**Amazon S3)** and **Amazon DynamoDB.**
- Used **Spark SQL for Scala & amp, Python** interface that automatically converts **RDD** case classes to **schema RDD**.
- Import the data from different sources like **HDFS/HBase** into **Spark RDD** and perform computations using **PySpark** to generate the output response. Creating **Lambda** functions with **Boto3** to deregister unused **AMIs** in all application regions to reduce the cost for **EC2** resources.
- Importing & exporting database using **SQL** Server Integrations Services (**SSIS**) and Data Transformation Services (**DTS** Packages). Conducted Data blending, Data preparation using **Alteryx** and **SQL** for **Tableau** consumption and publishing data sources to **Tableau server**.
- Coded **Teradata BTEQ** scripts to **load**, **transform data**, fix defects like **SCD** 2 date chaining, cleaning up duplicates.
- Developed reusable framework to be leveraged for future migrations that automates **ETL** from **RDBMS** systems to the **Data Lake** utilizing **Spark Data Sources** and **Hive** data objects.

Environment: Kafka, HBase, Docker, Kubernetes, AWS, EC2, S3, Lambda, Cloud Watch, Auto Scaling, EMR, Redshift, Jenkins, ETL, Spark, Hive, Athena, Sqoop, Pig, Oozie, Spark Streaming, Hue, Scala, Python, Databricks, GIT, Micro Services, Unix/Linux, Snowflake.

Client: Keka Technologies (ADP), Mumbai, India (Oct 2020 - Jul 2022)
Role: Application Developer/ Data Engineer

Description: ADP is a comprehensive global provider of cloud-based Human Capital Management solutions and Business Process. I involved in documenting data engineering processes, best practices, and technical specifications, and providing training and support to internal users on data tools and systems.

Responsibilities:

- Working knowledge on **Kubernetes** to deploy **scale**, **Load balance**, and **manage Docker** containers and **Open Shift** with multiple namespace versions. Presented the project to faculty and industry experts, showcasing the pipeline's effectiveness in providing real-time insights for marketing and brand management.
- Wrote and executed various MY**SQL data**base queries from **Python** using **Python-MySQL** connector and **MySQL dB** package. Performed **data** wrangling to clean, transform and reshape the **data** utilizing **panda's** library.
- Implemented airflow for workflow automation and scheduling tasks and created **DAGs** tasks.
- Storing different configs in **No SQL database Mongo DB** and manipulating the configs using **PyMongo**.
- Configured **Spark** streaming to get ongoing information from the **Kafka** and store the stream information to **DBFS**.

- Experience in creating **Kubernetes** replication controllers, **Cluster**s and label services to deployed **Microservices** in **Docker**. Involved in the entire lifecycle of the projects including Design, Development, and Deployment, Testing and Implementation, and support. Managed large **data**sets using **Panda data** frames and **SQL**.
- Consult leadership/stakeholders to share design recommendations and thoughts to identify product and technical requirements, resolve technical problems and suggest **Big Data** based analytical solutions.
- Spearheaded **HBase** setup and utilized **Spark** and **SparkSQL** to develop faster **data pipelines**, resulting in a 60% reduction in processing time and improved **data accuracy**.
- Build **Jenkins** jobs for **CI/CD** Infrastructure for **GitHub repos**. Involved in loading and transforming large sets of Structured, Semi-Structured and Unstructured **data** and analyzed them by running **Hive** queries. Processed the image **data** through the **Hadoop** distributed system by using **Map** and **Reduce** then stored into **HDFS**.
- Implemented Navigation rules for the application and page outcomes, written controllers using annotations. Worked with **AWS Terraform** templates in maintaining the infrastructure as code.
- Responsible for loading the **data** from **BDW Oracle database**, **Teradata** into **HDFS** using **Sqoop**. Implemented **AJAX**, **JSON**, and **Java script** to create interactive web screens.
- Implemented **RESTful** Web-**Services** for sending and receiving **data** between multiple systems.

Environment: AWS (EC2, S3, EBS, ELB, RDS, SNS, SQS, VPC, Redshift, Cloud formation, CloudWatch, ELK Stack), Jenkins, Ansible, Python, Shell Scripting, PowerShell, GIT, Microservice, Snowflake, Cassandra, Jira, Docker, AWS Glue, Kafka, Scrum, Git, Airflow, Control M, Tableau, Mongo DB, C#.

Client: Shree Maruti Integrated Logistics, Mumbai, India (Aug 2018 - Sep 2020)
Role: Data Engineer

Description: Shree Maruti Integrated Logistics is a Logistics Company. It provides efficient e-commerce fulfilment solutions, including streamlined shipping options, on-demand warehousing, and warehouse automation. Investigating and resolving data-related issues and providing technical support to internal users as needed.

Responsibilities:

- Extensively involved in all phases of **Data** acquisition, **data** collection, **data** cleaning, model development, model validation and visualization to deliver business needs of different teams.
- Used Django evolution and manual **SQL** modifications were able to modify Django models while retaining all **data**, while site was in production mode. Performed load testing and optimization to ensure the pipeline's scalability and efficiency in handling large volumes of **data**. Written queries in My**SQL** and Native **SQL**.
- Used **Python** based **GUI** components for the Front End functionality such as selection criteria.
- Developed business logic using **Kafka** & **Spark** Streaming and implemented business transformations. Supported Continuous storage in **ADLS** and configured Snapshots and wrote entities in spark along with named queries to interact with **data**base. Involved in various phases of Software Development Lifecycle (**SDLC**) of the application, like gathering requirements, design, development, deployment, and analysis of the application.
- Led requirement gathering, business analysis, and technical design for **Hadoop** and **Big Data** projects.
- Managed relational **data**base services in which the **Azure SQL** handles reliability, scaling, and maintenance. Integrated **data** storage solutions.
- Designed **GIT** branching strategies, merging per the needs of release frequency by implementing **GIT** flow workflow on Bit bucket. Responsible for loading the **data** from **BDW Oracle data**base, **Teradata** into **HDFS** using **Sqoop**. Implemented **AJAX**, **JSON**, and **Java script** to create interactive web screens.
- Added the Navigations and paginations and filtering columns and adding and removing the desired columns for view.
- Created **Data** tables utilizing **PyQt** to display customer and policy information and add, delete, update customer records. Integrated **Azure Data Factory** with **Blob** Storage to move **data** through **DataBricks** for processing and then to **Azure Data Lake** Storage and **Azure SQL data warehouse**.

Environment: HDFS, Hadoop, Hive, Hbase, MapReduce, Spark, Sqoop, Pandas, MySQL, SQL Server, Java, Python, Tableau, Git, Linux/Unix