Name : XXXXXXXXXXXXXXXXXX
Ph.No : XXXXXXXXXXXXXX
Email ID : XXXXXXXXXXXXX

Professional Summary:

- IT professional with overall **7+ years** of experience, concentrated in **Big Data ecosystem (**Data Acquisition, Ingestion, Modeling, Storage Analysis, Integration, and Data Processing).
- Experience using **Big Data ecosystem**, **Hadoop** (HDFS, MapReduce, Yarn), **AWS, Azure, Google Cloud Platform**.
- Strong understanding of Hadoop Architecture, Hadoop Clusters, and various components such as **HDFS, Job Tracker, Task Tracker, Name Node, Data Node, Map Reduce, Spark**.
- Adept in **Python** Scripting and performed cleaned data and processed third party data into portable deliverables within specific format with **Excel macros and python libraries** such as **NumPy**, visualized using **Matplotlib** and used **Pandas** for organizing the data.
- Obtained and processed data from Enterprise applications, Clickstream events, API gateways, Application logs and database updates.
- Very keen in knowing the newer techno stack that Google Cloud platform (GCP) adds.
- In depth knowledge about **Data Warehousing** (gathering requirements, design, development, implementation, testing, and documentation), **Data Modeling** (analysis using Star Schema and **Snowflake** for FACT and Dimensions Tables), **Data Processing, Data Acquisition** and **Data Transformations** (Mapping, Cleansing, Monitoring, Debugging, Performance Tuning and Troubleshooting Hadoop clusters).
- Have strong **Problem solving, Debugging and Analytical capabilities**, and can effectively engage in understanding and conveying business requirements.
- Experienced in Providing support on **AWS Cloud** infrastructure automation with multiple tools including **Gradle**, **Chef**, **Nexus**, **Docker** and monitoring tools such as **CloudWatch.**
- Vast experience in moving hadoop platform codes such as hive, pyspark etc into appropriate GCP resources and building reliable data pipelines.
- Used Azure SQL **Data Warehouse** to control and grant database access.
- Good Knowledge on architecture and components of **Spark**, and excellent knowledge in **Spark Core, Spark SQL, Spark streaming** for interactive analysis, batch processing and stream processing.
- Shown expertise in building **PySpark** and **Scala** applications.
- Worked with various streaming **ingest services** with Batch and Real-time processing using **Spark streaming, Kafka, confluent, Storm, Flume and Sqoop.**
- Experience Cloud functions, BigQuery.in GCP Dataproc, GCS,
- Worked with tools like **Jenkins, Docker, GitHub, Slack and JIRA** to migrate legacy applications to cloud platform.
- Designed interactive dashboards, reports, performing ad-hoc analysis and visualizations using **MS Excel, Tableau and Matplotlib.**

Technical skills:

- **Hadoop Distributions:** Apache Hadoop 3.x/2.x/1.x, Cloudera
- **Cloud Computing:** Amazon AWS (EMR, EC2, EBS, RDS, S3, Glue, Elasticsearch, Lambda, Kinesis, SQS, DynamoDB, Redshift, ECS)**,** Azure HDInsight (Databricks, Data Lake, Blob Storage, Data Factory, SQL DB, SQL DWH, Synapse, Stream Analytics, Cosmos DB, Azure DevOps)
- **Programming Languages:** Python, Scala, R, Shell Scripting.
- **NoSQL Database:** Cassandra, MongoDB, Redis.
- **Database:** MySQL, Teradata, Oracle, MS SQL SERVER, PostgreSQL.
- **ETL/BI:** Informatica, Talend, SSIS, SSRS, SSAS, Power BI, Tableau.

PROFESSIONAL EXPERIENCE:

Guy Roofing, Spartanburg, SC **Jan '21- Present**

Sr. Data Engineer

Responsibilities:

- Collaboratively worked to manage the buildouts of large data clusters and real-time streaming with Spark.
- Developed ETL pipelines from web servers using Spark, Spark Streaming, Sqoop and Scala, Kafka for Open-source Hadoop applications to ingest, transform and analyze customer data.
- Launched AWS EC2 instances to execute Hadoop jobs on AWS Elastic MapReduce (EMR) to store the results in S3 buckets and used JIT servers.
- Experience in moving data between GCP and Azure using Azure Data Factory.
- Processed Extensible Markup Language (XML) messages using Kafka and the xml file using Spark Streaming to capture User Interface (UI) updates.
- Supported MapReduce Programs those are running on the cluster and involved in loading data from UNIX file system to HDFS
- Transformed and Copied data from the JSON files stored in a Data Lake Storage into an Azure Synapse Analytics table by using Azure Databricks
- Installed and configured Hive and written HiveUDFs
- Can work parallel in both GCP and Azure Clouds coherently.
- Involved in creating Hive tables, loading with data and writing hive queries that will run internally in MapReduce way
- Created HBase tables to store variable data formats of PII data coming from different portfolios
- Implemented best income logic using Pig scripts
- Tested the cluster Performance using Cassandra-stress tool to measure and improve the Read/Writes
- Exported the analyzed data to the relational databases using Sqoop for visualization and to generate reports for the BI team.
- Used cloud shell SDK in GCP to configure the services Data Proc, Storage, BigQuery
- Writing **Scala Applications** which runs on **Amazon EMR cluster** that fetches data from the Amazon S3/one lake location and queue it in the Amazon **SQS** (simple Queue Services) queue.
- Performed ETL operations using **Python, SparkSQL, S3** and **Redshift** on terabytes of data to obtain customer insights.
- Written python scripts for **internal testing** which pushes the data reading form a file into Kafka queue which in turn is consumed by the Storm application.
- Experience in GCP Dataproc, GCS, Cloud functions, Cloud SQL & BigQuery.
- Implemented the code which handles data type conversions, data value mappings and checking for required fields.
- Executed programs by using python API written in python to support Apache Spark or **PySpark**.
- Performed end to end implementation of **ETL** pipelines using **Python** and **SQL** for high volume analytics and also reviewed use cases before on boarding to HDFS.
- Built reports for monitoring data loads into GCP and drive reliability at the site level
- Worked with the Spark for improving performance and optimization of the existing algorithms in **Hadoop.**
- Experienced in writing live Real-time Processing and core jobs using **Kafka** as a Data pipe-line system and good understanding of **Cassandra architecture, replication strategy, gossip, snitches etc.**
- Utilized **Airflow DAG'S** to build ETL pipelines and create reports and dashboards.
- Worked on data analysis and reported using **Tableau** on customer usage metrics.
- Involved in writing **Python** scripts to automate the process of extracting weblogs using **Airflow** DAGs.
- Used **Git** for version control and **Jira** for project management, tracking issues and bugs.
- Worked on CICD using cloud formation templates and Terraform templates.

Environment: Hadoop, Python, Scala, SQL, RDBMS, scripting, AWS (EC2, S3, Glue, EMR, Airflow, Lambda, Redshift), Apache Kafka, Spark Streaming, Hadoop MapReduce, GIT, Jira.

Teladoc, Boston MA **Aug '19- Dec '20**

Data Engineer

Responsibilities:

- Engaged with everyday Scrum gatherings to talk about the turn of events/progress and was dynamic in making scrum gatherings more gainful.
- Strong experience with architecting highly performant databases using PostgreSQL, MySQL and Cassandra.
- Composed Advanced SQL queries over OLAP databases.
- Working on Hive Meta store backup, **Partitioning** and **bucketing** techniques in hive to improve the performance. Tuning Spark & Scala Jobs.
- Used apache airflow in GCP composer environment to build data pipelines and used various airflow operators like bash operator, Hadoop operators and python callable and branching operators.
- Created automated and highly reliable data pipelines from various RDBM systems using Sqoop and NoSQL DBs (Mongo DB, Cassandra, PostgreSQL) into HDFS using Python and Spark
- Knowledge in working with Azure cloud platform (HDInsight, Data Lake, Databricks, Blob Storage, Data Factory, Synapse, SQL, SQL DB, DWH and Data Storage Explorer)
- Experience with Oozie Workflow Engine in running workflow jobs with actions that run Hadoop Map Reduce and Pig jobs.
- Hands on experience in installing configuring and using Hadoop ecosystem components like Hadoop MapReduce HDFS HBase Hive Sqoop Pig Zookeeper, Airflow and Flume
- Experience in moving data between GCP and Azure using Azure Data Factory
- Comprehensive knowledge of system engineering, reliability life-cycle management, and reliability modelling. Played role of site reliability engineer as well.
- Knowledge on installation and administration of multi-node virtualized clusters using Cloudera Hadoop and Apache Hadoop
- Good knowledge about Hive (architecture, Thrift servers), HQLs, Beeline and other 3rdparty JBDC connectivity services to Hive.
- Developed **Scala scripts** using both **Data frames/SQL/Data sets and RDD** in Spark for Data Aggregation, queries and writing data back into OLTP system through Sqoop
- Used Hive QL to analyze the partitioned and bucketed data, Executed Hive queries on Parquet tables.
- Experience in using Apache Sqoop to import and export data to from HDFS and external RDBMS databases.
- Coordination with the production support team to find and solve issues and adapt infrastructure changes and platform updates to the current data lake capabilities.
- Performed data ingestion, data cleansing, data mining, data validation and custom aggregation.
- Developed Airflow programs to schedule PySpark Jobs and perform historical and incremental loads.
- Model complex ETL jobs that transform data visually with data flow or by using compute services Azure Databricks, and **Azure SQL Database**.
- Creating pipelines, data flows and complex data transformations and manipulations using **ADF** and Pyspark with **Databricks**.
- Worked on ADF **ARN Templates** to automate the deployments on to Azure Data Factory.
- Experience in building **PySpark**, Spark Java and Scala applications for interactive analysis, batch processing, and stream processing.
- Build data pipelines in Airflow in GCP for ETL related jobs using different airflow operators.
- Build the Logical and Physical data model for Snowflake as per the changes required
- Experience in creating separate virtual data warehouses with difference size classes in AWS Snowflake
- Define virtual warehouse sizing for **Snowflake** for different type of workloads.

Environment: Hadoop, Python, Scala, SQL, RDBMS, scripting, AWS (EC2, S3, Glue, EMR, Airflow, Lambda, Redshift), Apache Kafka, Spark Streaming, Hadoop MapReduce, GIT, Jira.

First Republic Bank Palm Beach, FL **Mar '18 to July '19**
Big Data Developer,
Responsibilities:

- Involved in loading data from UNIX file system to HDFS.
- Wrote MapReduce jobs to discover trends in data usage by users.
- Involved in managing and reviewing Hadoop log files.
- Involved in running Hadoop streaming jobs to process terabytes of text data.
- Developed HIVE queries for the analysts.
- Implemented Partitioning, Dynamic Partitions, Buckets in HIVE.
- Exported the result set from HIVE to MySQL using Shell scripts.
- Used Git for version control.
- Designed and deployed data pipelines using Data Lake, **Databricks**, and **Apache Airflow**.
- Developed Elastic pool databases and scheduled Elastic jobs to execute T-SQL procedures.
- Developed Spark applications using **PySpark** and Spark-SQL for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
- Ingested data in mini-batches and performs RDD transformations on those mini-batches of data by using Spark Streaming to perform streaming analytics in Databricks.
- Created and provisioned different Databricks clusters needed for batch and continuous streaming data processing and installed the required libraries for the clusters.
- Developed automatic job flows and ran through Oozie daily and when needed which runs MapReduce jobs internally.
- Extracted Tables and exported data from Teradata through Sqoop and placed in Cassandra.

Environment: Databricks, DataLake, CosmosDB, MySQL, Snowflake, MongoDB, Cassandra, Teradata, Tableau, PowerBI, Git, Blob Storage, Data Factory, Data Storage Explorer, Scala, Hadoop 2.x (HDFS, MapReduce, Yarn), PySpark, Airflow, Hive, Sqoop, HBase, Oozie.

Legal & General America- Frederick, MD **May '16 to Feb '18**
Data Engineer,
Responsibilities:

- Working on processing big volumes of data using different big data analytic tools including Spark Hive, SQOOP, Pig, Flume, Apache Kafka, PySpark, OOZIE, HBase, Python, Scala.
- Implementation and data integration in developing large-scale system software experiencing with Hadoop ecosystem components like HBase, Sqoop, Zookeeper, Oozie, Hive and Pig.
- Developed Hive UDF's for extended use and wrote HiveQL for sorting, joining, filtering and grouping the structure data.
- Hands on experience on Cloudera Hue to import data on the GUI.
- Worked on scalable distributed data system using Hadoop ecosystem in AWS EMR and MapR (MapR data platform).
- Performed Data Ingestion from multiple internal clients using Apache Kafka.
- Worked on integrating Apache Kafka with Spark Streaming process to consume data from external REST APIs and run custom functions.
- Used AWS glue catalog with crawler to get the data from S3and perform sql query operations.

R&L Carriers, OH **Jun '15 – Apr '16**
Software Developer
Responsibilities:

- Experience in reviewing Python code for running the troubleshooting test-cases and bug issues.
- Managed datasets using Panda data frames and **Oracle**-**SQL**, queried Oracle-SQL database queries from python using Python-Oracle-SQL connector and Oracle-SQL DB package to retrieve information.
- Design, involved in code reviews and wrote unit tests in Python and Updated site with **JavaScript**, and **SQL**.
- Developed front-end, User Interface using **HTML**, **CSS, Angular** and session validation using **Spring AOP**.

- Designed and developed web-based application using **HTML5, CSS, JavaScript, AJAX, JSP framework.**
- Experience in using various IDEs **Eclipse**, IntelliJ, and repositories SVN and Git version control systems.

Environment: Python, SQL, JavaScript, Html/Html5, CSS, XML, Angular, Eclipse, Oracle, jQuery, JSON, CSV.