# Name: VenkataRamana R

# Email: venkataramanaravuri800@gmail.com

# Mobile: 972-428-3834

# PROFESSIONAL SUMMARY:

* I have around 9 years of professional IT experience in the Analysis, Development, Integration, and Maintenance of web-based and Client/Server applications using **Java** and **Big Data technologies**.
* 4+ years of relevant experience in using **Hadoop Ecosystem tools** and their architecture (HDFS, MapReduce, YARN, Spark, Pig, Hive, HBase, Sqoop, Flume, Oozie).
* Strong experience working with **Spark** (Dataframes and Spark SQL) for high-performance data processing and data preparation.
* Experience in real-time analytics with **Apache Spark Streaming** and **Kafka.**
* Ingested real-time streaming events to **Kafka topics** using Kafka producers API.
* Good hands-on experience working with various Hadoop distributions mainly **Cloudera** **(CDH),** **Hortonworks** **(HDP),** and **Amazon EMR.**
* Expertise in developing production-ready **Spark applications** utilizing Spark-Core, Dataframes, Spark-SQL, Spark-ML and Spark-Streaming API's.
* Design and develop **Hadoop MapReduce** programs and algorithms to analyze cloud-scale classified data stored in **Cassandra.**
* Involved in designing and deploying multi-tier applications using all the **AWS services** like (EC2, Route53, S3, RDS, Dynamo DB, SNS, SQS, and IAM) focusing on high-availability, fault tolerance, and auto-scaling in AWS Cloud Formation.
* Experience in using **D-Streams** in Spark streaming, accumulators, Broadcast variables, various levels of caching, and optimization techniques in **Spark.**
* Worked extensively on **Hive** for building complex data analytical applications.
* Strong experience writing complex map-reduce jobs including development of custom Input Formats and custom Record Readers.
* Sound Knowledge in map-side join reduce-side join, shuffle & and sort, distributed cache, compression techniques, and multiple Hadoop Input & output formats.
* Worked extensively on **Sqoop** for performing bulk and incremental ingestion of large datasets from **Teradata** to **HDFS.**
* Having Experience in monitoring and managing the **Hadoop cluster** using **Cloudera Manager**.
* Experience in job workflow designing and scheduling using **Oozie**.
* Having good knowledge of **Scala** programming concepts.
* Expertise in distributed and web environments focused on **Core Java technologies** like Collections, Multithreading, IO, Exception Handling, and Memory Management.
* Expertise in the development of Web applications using **J2EE technologies** like Servlets, JSP, Web Services, Spring, Hibernate, HTML5, JavaScript, jQuery, AJAX, etc.,
* Knowledge of standard build and deployment tools such as Eclipse, Scala IDE, and Maven.
* Extensive knowledge in **Software Development Lifecycle** (SDLC) using Waterfall and agile **methodologies**.
* Facilitate Sprint planning, daily scrums, retrospectives, stakeholder meetings, and software demonstrations.
* Strong experience working with **Spark** (Dataframes and Spark SQL), and **Splunk** for high-data performance processing and data preparation.
* Implemented spark application with **SCALA Programming Language**.
* Worked on **GCP Cloud technologies** Like **Data Proc, Data Flow, and BIG QUERY.**
* Experience in working with tools like Git Hub, Git Lab, and SVN for code repository.
* Experience with **Agile Methodology**, **Scrum Methodology,** and release management.
* Experience in analyzing data using HiveQL, HBase, and custom MapReduce programs in Java.
* Implemented **Sqoop jobs** for large sets of structured and semi-structured data migration.
* between **HDFS** and/or other data storage like **Hive** or **RDBMS.**
* Hands-on experience in the **Hadoop ecosystem** including Spark, Kafka, HBase, Pig, Impala, Sqoop, Oozie, Flume, Mahout, Storm, Tableau, and Talend big data technologies.
* Involved in converting **Hive/SQL queries** into **Spark transformations** using Spark RDD and Pyspark concepts.
* Experience in developing and scheduling **ETL workflows** in Hadoop using **Oozie.**
* Worked on data warehouse product Amazon Redshift, which is a part of **AWS.**
* Set up full **CI/CD** pipelines so that each commits a developer makes will go through the standard process of the software lifecycle and get tested well enough before it can make it to production.
* Helped individual teams set up their repositories in bit bucket and maintain their code and helped them set up jobs that can make use of the **CI/CD** environment.
* Implemented solutions utilizing **Advanced AWS Components**: EMR, EC2, etc integrated with Big Data/Hadoop Distribution Frameworks: Zookeeper, Yarn, Spark, Scala, NiFi, etc.
* Good understanding of Spark Architecture with **Databricks,** and Structured Streaming. Setting Up AWS and **Microsoft Azure with Databricks**, Databricks Workspace for Business Analytics, Managing Clusters in **Databricks,** and Managing the Machine Learning Lifecycle.
* Extracted files from **Cassandra** and **MongoDB** through Sqoop and placed in HDFS and processed.

# TECHNICAL SKILLS:

|  |  |
| --- | --- |
| **Languages** | Java, Scala, SQL, PL/SQL, Pig Latin, Python, Hive QL |
| **Web Technologies** | JEE (JDBC, JSP, SERVLET, JSF, JSTL), AJAX, JavaScript |
| **Big Data Systems** | Hadoop, HDFS, MapReduce, YARN, Pig, Hive, Sqoop, Flume, Oozie, Impala, Spark, Kafka, Splunk, Cloudera CDH4, CDH5, HortonWorks, Solr, and Ranger… |
| **RDBMS** | Oracle 10g/11g, MySQL, SQL Server 2005/2008 R2, PostgreSQL, DB2, Teradata |
| **NoSQL Databases** | HBase, MongoDB, Cassandra |
| **Cloud Technologies** | AWS, AZURE, GCP |
| **App/Web Servers** | Apache Tomcat, WebLogic |
| **SOA** | Web services, SOAP, REST |
| **Frameworks** | Struts 2, Hibernate, Spring 3. x |
| **Version Control** | GIT, CVS, SVN |
| **IDEs** | Eclipse, Scala IDE, NetBeans, IntelliJ IDEA |
| **Operating Systems** | UNIX, Linux, Windows |

# PROFESSIONAL EXPERIENCE:

# Client: Abercrombie & Fitch Jan 2023 to Till Date

# Role: AWS Big Data Engineer

# Responsibilities:

* Developed **AWS Glue ETL environment** by deploying AWS cloud formation template from scratch.
* Developed the **Glue ETL scripts** in **Scala** and **Python** for data transformation, reading, and writing the **parquet files** from **S3.**
* Implemented **Oozie** workflow engine to run multiple **Hive** and **Python jobs.**
* Experience in cloud data migration using **AWS**.
* Involved in **data migration** to **Snowflake** using **AWS S3 buckets**.
* Responsible for creating the design documents, establishing specific solutions, and creating the **Test Cases**.
* Analyzed the **SQL scripts** and designed them by using **PySpark SQL** for faster performance.
* Configured and used Query Surge tool to connect with **HBase** using Apache Phoenix for Data Validation.
* Developed **Spark applications** in **Python** (**PySpark**) on a distributed environment to load a huge number of CSV files with different schema into **Hive ORC tables**.
* Mastering/Leading in the development of applications/tools using **Python**.
* Wrapper developed in **Python** for instantiating multi-threaded applications and running with other applications.
* Worked on reading and writing multiple data formats like **JSON, ORC, and Parquet** on **HDFS** using **PySpark.**
* Created different **Hive** RAW and Standardized tables for data validation and Analysis with Partition and bucket.
* Experience in elastic search Engine index-based search, ELK log analytics tool, Elastic Search, and Logstash.
* Designed **NoSQL databases** schema to help migrate legacy applications datastore to Elasticsearch.
* Involved in loading and transforming large sets of structured and semi-structured from multiple data sources to **Raw Data Zone** (HDFS) using **Sqoop imports** and **Spark jobs**.
* Developed **Hive queries** for data sampling and analysis to the analysts.
* Developed **Spark Data Frames** and Dynamic data frames for structured data processing.
* Creating **AWS EMR Data processing** cluster for Ingestion of the data from **On-prem** to **S3 bucket.**
* Experience in handling the **EMR** configurations for handling memory issues for **Spark jobs.**
* Scheduling the **Ad-hoc** job in **AWS EMR** for the transformation of data.
* Having experience in **AWS EMR**, and Jupyter Notebook for analyzing the data and transformations.
* Production support for the EMR cluster and mainly troubleshooting memory and spark job application issues.
* Moving data in and out of an instance using import sets and transform maps and also auto imports of data into **ServiceNow.**
* Ingested the on-prem **CSV files** to **AWS S3** bucket Daily incremental using **AWS**.
* Hands-on experience on **Oozie workflows** for Ingest the data hourly basis.
* Importing data using **Sqoop** to load data from **Oracle/Linux server** to **AWS S3/HDFS** on a regular basis.
* Developing **Spark Data Frame** Operations to perform required Validations in the data and to perform analytics on the **Hive data.**
* Using **Hive** to do transformations, joins, filters, and some pre-aggregations after storing the data in **HDFS.**
* Monitoring **Control-M job** for scheduling the process and Managing Elastic MapReduce cluster through the AWS console.
* Developed **AWS Lambda** function for monitoring the EMR cluster status updates and monitoring the jobs.
* Experience in monitoring **CloudWatch logs** for EMR Bootstrap actions and Step logs etc.
* Experience in configuring and implementing AWS tools such as **CloudWatch**.
* **CloudTrail** and direct system logs for monitoring.

**Environment:** Amazon EMR, Hadoop, Hive, Jupyter Notebook, Impala, Oracle, Spark, Sqoop, Oozie, Map Reduce, GIT, HDFS, Linux, Bamboo, cucumber, and Jira.

# Client: Dematic, SD May 2020 to Dec 2022

# Role: Azure Big Data Engineer

# Responsibilities:

# Orchestrated the Azure-based Big Data engineering initiatives at Dematic, focusing on healthcare automation and logistics processes.

# Implemented scalable distributed data solutions on Azure, handling terabytes of data with Sqoop and Kafka for efficient real-time processing.

# Developed Spark programs in Python for data quality checks, ensuring data accuracy and reliability.

# Led the migration of platform architecture from Cloudera Distribution to Azure, leveraging Snowflake and Databricks.

# Utilized Azure Data Factory, T-SQL, Spark SQL, and U-SQL in Azure Data Lake Analytics for ETL processes from source systems to Azure Data Storage.

# Expertise in Splunk DB Connect for importing data from Oracle to Splunk on Azure, enhancing data analytics and monitoring capabilities.

# Orchestrated workflows with Apache Oozie, automating tasks in coordination with Zookeeper in the Azure environment.

# Leveraged Azure Databricks for processing data stored in Azure Data Lake, optimizing performance and improving overall efficiency.

# Utilized Spark-Streaming APIs to perform on-the-fly transformations from Kafka, persisting data into MongoDB on Azure.

# Successfully migrated platform architecture to Azure to meet growing system resource demands and enable more efficient data access.

# Collaborated with the SCRUM team at Dematic, ensuring timely delivery of agreed user stories for every sprint.

# Client: Amex Feb 2019 to Apr 2020

# Role: Big Data Engineer

# Responsibilities:

* Hands-on experience in working on **Spark SQL** queries and data **frames**, importing data from Data sources, performing transformations, performing read/write operations, save the results to the output directory into **HDFS**.
* 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 Data Quality Scripts using **SQL** and **Hive** to validate successful data load and quality of the data. Created various types of data validation using **Python** in **Spark.**
* Involved in converting **Hive**/**SQL** queries into **Spark Transformations** using **Spark RDDs** and **Apache PySpark**.
* Involved in Configuring spark-streaming to get ongoing information from **Kafka** and store the stream information in **Google Cloud Storage.**
* Experience in creating views and external tables in **Big Query** on **GCS Buckets.**
* Experience in using **Apache Hudi** using **spark** to perform incremental load and store data in **GCS.**
* Debugging the failure issues by capturing the array and register dumps using **Python scripts**, and traces and performing several experiments by interacting with the **design team**.
* Launching harasser across several threads along with memory stress test using the **Python** wrapper.
* Involved in analyzing system failures, identifying root causes, and recommending courses of action.
* Designed and prepared interactive and intuitive year-end dashboards and reports using IBM **Cognos.**
* Involved in Migrating from **GCP** Cloud break cluster to **GCP Dataproc** clusters.
* Developing **Yamls** with **Sqoop** to move data from different **RDBMS** to **GCS.**
* Managed the imported data from different data sources, performed transformation using **Hive**, **Spark,** and **Map-Reduce,** and loaded data in **HDFS.**
* Executed **Oozie** workflow engine to run multiple **Hive** and **spark** jobs, which run independently with time and data availability, and developed **Oozie** workflow to run jobs onto data availability of transactions.
* Developed **KafkaScripts** to extract the data from the **FTP** server output files to load into **HDFS**.
* Implemented custom UDF for Confidential Kudu then Developed **Hive UDF** to pre-process the data for analysis and Develop **Spark** for the analysts.
* Developed workflow in **Oozie** to automate the tasks of loading the data into **HDFS** and pre-processing with **Spark**.
* Collected the logs data from web servers and integrated them into **HDFS** using **Kafka.**
* Implemented Fair schedulers on the Job tracker to share the resources of the Cluster for the **Map Reduce jobs** in **Python** given by the users.
* Managed and reviewed **Hadoop** log files and Spark to analyze point-of-sale data and coupon usage.
* Exported the analyzed data to the relational databases using **Sqoop** for visualization and to generate reports.
* Worked with highly engaged Informatics, Scientific Information Management, and enterprise IT teams.

**Environment:** Hadoop, HBase, HDFS, Hive, Spark, Spark SQL, Pig, Zookeeper, Oozie, Impala, Kafka.

**Client: Caterpillar Sep 2017 to Jan 2019**

**Role: Hadoop Developer/ Spark**

**Responsibilities:**

* Created **data pipelines** to ingest data from various sources including Devices, Apps, Survey data, **S3 Buckets, and databases** (**SAP Hana, MySQL, SQL Server, Amazon Redshift, PostgreSQL).**
* Collected metrics about **Data pipelines**, stored indexes in Elasticsearch & and created **Kibana**

dashboards which facilitated us in quickly identifying data loss & and anomalies.

* Responsible for maintaining the **General Data Protection Regulation** (GDPR) provided the legal team with user’s data upon request in the desired format, and deleted user’s data from various schemas from the **Hadoop Distributed File System**.
* Worked on custom **Pyspark** libraries to push columns from various formats of data to a **Data Governance tool** called **Collibra.**
* Successfully migrated **Data Pipeline jobs** from **Oozie** to **Airflow.**
* Worked on an Anonymization Project to anonymize PII information present in various schemas within Data Lake with a provision to maintain a mapping between de-identified hash values and original values in separate **RDBMS Databases.**
* Worked on **AWS EMR Migration**, successfully moved resource-intense **ETL** to run in **AWS EMR**, and created custom libraries.
* To meet the growing demand for system resources & to accommodate more efficient data access for end users we migrated our platform architecture leveraging **Snowflake**, and **Databricks** from **Cloudera** **Distribution**. Started migrating **data pipelines** & data from **CDH** to the new system.
* Worked on a POC to explore the features of Apache Pulsar.
* Worked on building **NIFI** flows (Processors relating to **Kafka**, JoltTransform**JSON**, Files, **S3**, **HD**FS, etc.) as per the needs.
* Imported data using **Sqoop** to load data from **MYSQL** to **HDFS** on regular basis.
* Written **Hive** queries for data analysis to meet the business requirements.
* Importing and exporting data into **HDFS** and **Hive** using **Sqoop**.

**Environment:** Apache Hadoop, Cloudera, AWS EC2, S3, EMR, Glue, Kafka, Databricks, Linux, Java, MapReduce, HBase, Hive, Sqoop, Oozie and SQL, Spark, Elasticsearch, Kibana, Snowflake.

**Client: Vignan Soft Tech, Hyderabad- INDIA Jan 2014 to June 2016**

**Role: Hadoop Developer**

**Responsibilities:**

* Built a scalable distributed data solution using **Hadoop** on a 30-node cluster using **AWS cloud** to run analysis on 25+ Terabytes of data.
* Developed several new **MapReduce** and **Spark programs** to analyze and transform the data to uncover insights into customer usage patterns.
* Used **MapReduce** to Index a large amount of data to easily access specific records.
* Performed **ETL** using **Pig, Hive, and MapReduce** to transform transactional data to de-normalized form.
* Configured periodic incremental imports of data from **DB2** into **HDFS** using **Sq**oop.
* Worked extensively with importing metadata into **Hive** using **Java** and migrated existing tables and applications to work on **Hive** and **AWS cloud**.
* Wrote **Pig** and **Hive UDFs** to analyze the complex data to find specific user behavior.
* Used **Kafka** and **Solr workflow** engine to schedule multiple recurring and ad-hoc **Hive** and **Pig jobs**.
* Responsible for maintaining and implementing code versioning techniques using **Cassandra** for the entire project.
* Created **HBase** tables to store various data formats coming from different portfolios.
* Utilized cluster coordination services through **ZooKeeper**.
* Assisted the team responsible for cluster maintenance, adding and removing cluster nodes, cluster monitoring and troubleshooting, and managing & reviewing data backups and **Hadoop** log files.
* Worked with teams in various locations nationwide and internationally to understand and accumulate data from different sources.
* Worked with the testing teams to fix bugs and ensure smooth and error-free code.

**Environment:** Hadoop, MapReduce, HDFS, Hive, Java, SQL, Cloudera Manager, Pig, Sqoop, Oozie, HBase, ZooKeeper, PL/SQL, MySQL, DB2, Teradata.