**Vamshi Sudharshan Reddy**

**Email: vamshi.deait@gmail.com**

**Cell: 510-999-3644**

**Sr. Hadoop Developer/ Data Engineer**

**PROFESSIONAL SUMMARY:**

* Around **9 years** of professional IT experience in **Big** data Environment, **Hadoop** Ecosystem and good experience in **Spark**, **SQL**, **Python**, **Pyspark** Development.
* Experienced in Developing **Spark** application using **Spark Core**, **Spark SQL** and **Spark Streaming API's**
* Extensive experience with wiring **SQL** queries using **HiveQL** to perform analytics on structured data.
* Experience with Cloud Service Providers such as **Amazon AWS**, **Microsoft Azure,** and **Google GCP**
* Hands on experience working on structured, unstructured data with various file formats such as **Xml** files, **JSON** files, **Sequence** files using **MapReduce** programs.
* **Google Cloud Certified Associate Cloud Engineer**
* Implemented business logic using **Pig** scripts. Wrote custom **Pig** UDF to analyze data.
* Performed **PIG** operations, joining operations and transformations on data to aggregate and analyze data.
* Experienced in database conversion from **Oracle** and **SQL** Server to **PostgreSQL** and **MySQL**.
* Experience with **Azure** transformation projects and **Azure** architecture decision making Architect and implement **ETL** and **Data movement** solutions using **Azure Data Factory** **(ADF), SSIS**
* Hands of experience in **GCP**, **Big** **Query**, **GCS** **bucket**, **G -** **cloud** function, **Cloud dataflow**, **Pub/sub cloud shell**, **GSUTIL**, **BQ** command line utilities, **Data** **Proc**, **Stack** driver.
* Worked with cloud provisioning tools like **Terraform** and **CloudFormation**.
* Experienced with minting highly scalable and fault tolerant multilayer **AWS** environments across multiple location using **Terraform and CloudFormation.**
* Hands on experience in creating **Hive UDF's** for the requirements and to handle **Json, Xml files, CSV, Test, ORC, Parquet.**
* Delivering projects (full SDLC) using big data technologies like **Hadoop, Oozie and NoSQL**
* Have knowledge on injecting data from multiple data sources to HDFS and Hive using **NiFi** and importing data using **Nifi** tool from Linux servers.
* Hands on experience in designing and developing POCs in **Spark** to compare the performance of **Spark** with **Hive** and **SQL**/**Oracle** using **Scala**.
* Experienced working with **Pandas** and **NumPy** libraries in **Python.**
* Experienced working from other databases to **Snowflake.**
* Experienced with **Snowflake** cloud data warehouse and **AWS S3** bucketing for integrating data from multiple sources systems includes loaded nested **Json** formatted data into **Snowflake** tables.
* Exposures to various **AWS** technologies like **SQS** **Queuing**, **SNS** Notification, **S3** storage, **AWS** **Redshift**, **Data Pipeline**, **AWS Glue**, **EMR, EC2.**
* Extensive working knowledge in setting up and running **Clusters**, **Monitoring**, **Data** **analytics**, **Sentiment** analysis, **Predictive** analysis, Data presentation with **Big data** world.
* Experienced working on various secluding tools like **Tidal, Airflow, Oozie, ESP, D-series**
* Excellent interpersonal and communication skills, creative, research-minded, technically competent and result-oriented with problem solving and leadership skills
* Used **Flume** and **Kafka** to direct data from different sources to/from **HDFS.**

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Big Data Ecosystems** | MapReduce, Hive, Sqoop, Spark, Kafka, Pig, Flume, HBase, Oozie, Airflow, Tidal |
| **Streaming Technologies** | Spark Streaming, Storm |
| **Cloud Platform** | Amazon Web Services (AWS), MS Azure, GCP |
| **Scripting Languages** | Python, Bash, Java Scripting, HTML5, CSS3 |
| **Programming Languages** | Java, Scala, SQL, PL/SQL |
| **Java/J2EE Technologies** | Servlets, JSP, JSF, JUnit, Hibernate, Log4J, EJB, JDBC, JMS, JNDI |
| **Databases** | Oracle, RDBMS, NoSQL, DB2 |
| **IDEs/Tools** | Eclipse, JUnit, Maven, Ant, MS Visual Studio, Net Beans, IntelliJ |
| **Methodologies** | Agile, Waterfall |

**PROFESSIONAL EXPERIENCE:**

**CVS/Aetna, Hartford CT Sep 2020- Till Date**

**Sr. Data Engineer**

**Description**: Currently I’m working on Communication tracker project, in which our team interacts with member data flowing from outside of Aetna systems to windows share, from there we are pulling into Snow flake systems as cloud data warehousing, it involves processing of data, cleaning, enriching, by implanting several transformation enriching data by joining with existence metadata in MySQL database, using technologies like Spark, Python, AWS, Snow flake, Hive, SQL, Shell Scripting, Airflow, Sqoop, Data Frame, RDDS .

**Responsibilities**:

* Developed Job Monitoring portal (JMP) application internal tool for monitoring jobs using **Pyspark** and **Python**.
* Responsible for data extraction and data ingestion from different data sources into **AWS** by creating **AWS Redshift** **ETL** pipelines.
* Built **AWS Redshift** **ETL Pipelines**, Data **profiling**, data **modeling** for converting raw data into enterprise user data.
* Imported data from different sources windows share, **RBMS**, **S3** ingestion and **copybook** ingestion for different types of vendors into HDFS dedicated Path.
* Developed **Spark job** to load data from different sources to **hive** tables.
* Developed **Spark code** using **Python** and **Spark**-**SQL** for faster testing and processing of data.
* Created the **PySpark** programs to load the data into **Hive** databases from **PySpark** Data frames.
* Worked extensively with **Sqoop** for importing and exporting the data from data Lake **HDFS** to **Relational** Database systems like Oracle and MySQL
* Involved in converting **Hive** or **SQL** queries into **Spark transformations** using **Python**.
* Developed **Python API** involving object orientated programing **(OOPS)** for data **Analytics** of member data using **Pandas** and **NumPy**
* Worked on data analytics using **Pandas** and **NumPy** in **Python.**
* Involved in deploying systems on **Amazon Web Services (AWS)** Infrastructure services **EC2.**
* Experience in configuring, deploying the web applications on **AWS** servers using **SBT** and **Play**.
* Performed configuration, deployment and support of cloud services including **Amazon Web Services** **(AWS)**
* Working knowledge of various **AWS** technologies like **SQS** **Queuing**, **SNS** Notification, **S3** storage, **AWS Redshift**, **Data Pipeline**, **AWS Glue**, **EMR, EC2.**
* worked with **Snowflake** cloud data warehouse and **AWS S3** bucketing for integrating data from multiple sources systems includes loaded nested **Json** formatted data into **Snowflake** tables.
* Developed scalable and fault tolerant multilayer **AWS** environments across multiple location using **Terraform and CloudFormation.**
* Built the logics and physical data models for **Snowflake** as per change request.
* Exploring with the **Spark** and improving the **performance** and **optimization** of the existing algorithms in Hadoop using **Spark Context**, **Spark-SQL**, **Data Frame**, and **Pair RDD's**
* Used **Data Frame API** in **Python** for converting the distributed collection of data organized into named columns.
* Developed **Spark jobs** and **Hive Jobs** to summarize and transform data.
* Performance optimization when dealing with large datasets using **partitions**, **broadcasts** in **Spark**, effective and **efficient joins**, **transformations** during ingestion process.
* Develop and enhance **Scripts** to automate and execute various **DBA** tasks.
* Stored the data in tabular formats using **Hive** tables and **Hive** **Serdes.**
* Implemented **Partitioning**, **Dynamic** **Partitions** and **Bucketing** in **Hive** for efficient data access.
* Executed **Hive** queries that helped in analysis of trends by comparing the new data with existing **data** **warehouse** reference tables and historical data.
* Used **OOZIE** engine for creating workflow and coordinator jobs that schedule and execute various **Hadoop** jobs such as **MapReduce** Jobs, **Hive**, **Spark** and automating **Sqoop** jobs.
* Configured Oozie workflow to run multiple Hive jobs which run independently with time and data availability.

**Environment**: CDH5, Hue, Eclipse, Centos Linux, HDFS, MapReduce, Kafka, Python, Scala, Java, Hive, Sqoop, Spark, Spark-SQL, Spark-Streaming, HBase, Oracle10g, Oozie, Red Hat Linux.

**Lowe’s, Mooresville, NC March 2020- Sep 2020**

**Sr. Data Engineer**

**Description**: Worked on Store Management and warehouse management project, which Involves availability of store products in the warehouse in various regions in united states, this project helps to reduces the warehouse management products based on the customer needs, based on the Customer behavior analyzed and products are developed based on their needs. This project is developed using Spark, Python, SQL, Teradata, AWS, Hive, Sqoop, CI/CD pipelines.

**Responsibilities:**

* Developed **spark** Jobs to load data from various sources to **Hive** tables as per requirements.
* Worked with spark core, **Spark** Streaming and **spark** **SQL** modules of **Spark.**
* Developed multiple POCs using **Spark** and deployed on the **Yarn** cluster, compared the performance of **Spark**, with **Hive** and **SQL**/**Teradata.**
* Build **AWS Redshift ETL** pipelines to load data from sources to target as per the business needs.
* Involved in converting **Hive/SQL queries** into **Spark** **transformations** using **Spark** **RDDs**, **Python** and **Scala**
* Developed the code for **Importing** and **exporting** data into **HDFS** and **Hive** using **Sqoop**.
* Loaded data from different sources to **Amazon S3 buckets** to build Lowes tables.
* Involved in Migrating **SQL** database to **Amazon Data Lake,** **Amazon data lake Analytics**, **Amazon RDS Database** and **Data** warehouse and controlling and granting database access and Migrating On premise databases to **Amazon S3 buckets to** store Data.
* Automated **Sqoop** incremental imports by using **Sqoop** jobs and automated the jobs using **Oozie.**
* Responsible for writing custom Queries for analyzing data in **AWS Redshift** before loading into S3 buckets.
* Worked on **Sequence files, ORC files, bucketing, partitioning** for **Hive** performance enhancement and storage improvement.
* Implemented **Snow** pipe, Stage and file upload to **Snowflake** database using copy command.
* Defined visual warehousing size for **Snowflake** for different types of workloads.
* Involved in defining **Job flows** using **Oozie** for scheduling jobs to manage apache Hadoop jobs.
* Developed **Python** and **Shell** scripts to schedule the processes running on a regular basis.
* Developing **Hive User Defined Functions**, compiling them into jars and adding them to the HDFS and executing them with **Hive Queries**
* Tested and reported defects in an **Agile Methodology** perspective.
* Involved in importing data from **SQL** to **AWS Redshift** and **AWS S3** for analytical purpose.
* Implemented the **workflows** using **Oozie** framework to automate tasks.
* Worked on **Control-M** charters to run the code on production environment.
* Schedule the various Jobs in BMC **Control-M** for various stages of deployments like prod, pre-prod and dev.

**Environment**: Hadoop, Hue, HDFS, Spark, AWS, MapReduce, Hive, Oozie, Java, Python, NoSQL, Cloudera, Linux, MySQL, SQL

**VISA, Palo Alto, CA June 2019- March 2020**

**Sr. Hadoop Developer**

**Description**: Worked on TBL application (Transaction Balanced life cycle), in which our team updates the data pulling from Visa Payment gateway in Rocks DB and loading in DB2 and hive tables implementing servals transformations and aggregations as per the downstream requirements, this helps to reduces the time frame to understand customer behavior on their Visa card usage. This project involves technologies like Spark, Python, Scala, GCP, Cloudera, Hortonworks, DB2, Hive, ETL, Shell Scripting, ESP and D-series.

**Responsibilities**:

* Developed **Spark** code using **Python** and **Spark-SQL** for faster testing and processing of data.
* Develop and deploy the outcome using **Spark** and **Python** code in **Hadoop cluster** running on **GCP**.
* Using rest API with **Python** to ingest Data from and some other site to **BIGQUERY**.
* Involved in migration from **Cloudera** to **Hortonworks**
* Developed **Spark** scripts using **Python** on Insight for **Data** **Aggregation**, **Validation** and verified its performance over **MR** jobs.
* Worked extensively with **Sqoop** for importing and exporting the data from data Lake **HDFS** to **Relational** Database systems like **Oracle** and **DB2**
* Involved in converting **Hive** or **SQL** queries into **Spark** transformations using **Python**, Cassandra.
* Exploring with the **Spark** and improving the **performance** and optimization of the existing algorithms in Hadoop using **Spark Context**, **Spark-SQL**, **Data Frame**, and Pair **RDD's.**
* Responsible for **importing** data to **HDFS** using **Sqoop** from different **RDBMS** servers and exporting data using **Sqoop** to the **RDBMS** servers after aggregations for other **ETL** operations.
* Used Data Frame **API** in **Scala** for converting the distributed collection of data organized into named columns.
* Used Spark for interactive queries, processing of streaming data and integration with **HBase** database for huge volume of data.
* Responsible for data extraction and data ingestion from different data sources into Hadoop Data Lake by creating **ETL** pipelines using **Pig**, and **Hive**.
* Develop and deploy the outcome using **spark** and **Scala** code in Hadoop cluster running.
* Developed **Hive** queries for data sampling and analysis to the analysts.
* Executed **Hive** queries that helped in analysis of trends by comparing the new data with existing data warehouse reference tables and historical data with **NoSQL**.
* Developed **Hive User Defined Functions** in java, compiling them into jars and adding them to the HDFS and executing them with **Hive Queries**
* Create/Modify **shell** scripts for scheduling various data cleansing scripts and **ETL** load process.
* Extracted and loaded data into **Data Lake** environment (MS Azure) by using **Sqoop** which was accessed by business users.
* Involved removing **control-M** charters to make sure code running up to date in various environments.
* Performed **ETL** Performance tuning to increase the **ETL** process speed.
* Used **OOZIE**, **ESP** **D** series engine for creating **workflow** and coordinator jobs that schedule and execute various **Hadoop** jobs such as **MapReduce** **Jobs**, **Hive**, **Spark** and automating **Sqoop** jobs.

**Environment**: CDH5, Hue, HDP, Ambari, Eclipse, Centos Linux, HDFS, MapReduce, Scala, Java, NoSQL, Hive, Sqoop, Spark, SQL, Cassandra, AWS EMR, AWS S3, Spark-Streaming, Hbase, PySpark, Stream Sets, Oozie, Red Hat Linux

**Medtronic, Houston TX Jan 2017- May 2019**

**Hadoop Developer**

**Description**: Worked on Building Insights Data of Medtronic’s to understand the various behavior changes of Customers of their heath visit to the doctors and labs for their reports, this projects involves pulling the data from various vendors by writing an API call using Java and writing into Hadoop data lake and finally storing into hive tables for business use case, this project involves technologies like Java, Hive, Oracle, SQL, Linux, Sqoop, Tableau.

**Responsibilities:**

* Developing parser and loader map reduce application to retrieve data from **HDFS** and store to **HBase** and **Hive**
* Importing the unstructured data into the data lake **HDFS** using **Flume**
* Used **Oozie** to orchestrate the map reduce jobs that extract the data on a timely manner.
* Written **Map Reduce** **java** programs to analyze the log data for large-scale data sets.
* Involved in using **HBase** **Java** **API** on **Java** application.
* Automated **Sqoop** jobs for extracting the data from different Data Sources like **MySQL** to pushing the result set data to **Hadoop** Distributed File System. Implemented **Map Reduce jobs** using **Java** API and **PIG** Latin as well **HIVEQL**. Participated in the setup and deployment of Hadoop cluster.
* Hands on design and development of an application using **Hive** (UDF)
* Importing and exporting Data from **MySQL/Oracle** to **HiveQL** Using **SQOOP**
* Designed and built many applications to deal with vast amounts of data flowing through multiple Hadoop clusters, using **Pig** Latin and Java-based map-reduce.
* Developed **UDFs** in **Java** when necessary to use in PIG and **HIVE** queries.
* Responsible for maintaining BMC **control-M** to run upto date as per the business needs.
* Specifying the cluster size, allocating Resource pool, Distribution of Hadoop by writing the specification texts in **JSON** File format
* Responsible for defining the data flow within Hadoop eco system and direct the team in implement it.

**Environment**: Hadoop, HDP (Horton works), Hive, Ambari, Zookeeper, Map Reduce, Sqoop, Pig, UNIX, Java, Java API, Eclipse, Oracle, SQL Server, MySQL

**HP, Hyderabad, In July 2014 - Aug 2015**

**Java Developer**

**Description**: Worked on building payment gateway for Citi bank client, by writing an API call to pull the data and loading into SQL tables, this project I was working as data ingestion by masking customer PII information and providing support for ongoing production jobs. Using technologies like Java, SQL, Oracle, CI/CD, and Selenium.

**Responsibilities:**

* Designed the user interfaces using **JSP**. Developed Custom tags, **JSTL** to support custom User Interfaces
* Implemented Business processes such as user authentication, Account Transfer using Session **EJBs**
* Used Eclipse to write the code for **JSP**, **Servlets**, **Struts** and **EJBs**
* Deployed the applications on **Web** **Logic** Application Server
* Used **Java** Messaging Services (**JMS**) and Backend messaging for reliable and asynchronous exchange of important information such as payment status report. Developed the entire Application(s) through **Java** **API**, Eclipse
* Worked with Web Logic Application Server to deploy the Application(s)
* Developed the Ant scripts for preparing **WAR** files used to deploy **J2EE** and API components.
* Used **JDBC** for database connectivity to Oracle.
* Worked with **Oracle** Database to create tables, **procedures**, **functions** and select statements.
* Used **JUnit** **Testing**, debugging, and bug fixing.
* Used **Log4J** to capture the log that includes runtime exceptions and developed **WAR** framework to alert the client and **production** support in case of application failures.
* Performed Data driven testing using **Selenium** and **TestNG** functions which reads data from property and **XML** files. Involved in **CICD** process using **GIT**, **Jenkins** job creation, **Maven** build and publish.
* Used **Maven** to build and run the **Selenium** automation framework.
* Involved in building and deploying scripts using **Maven** to generate **WAR, EAR** and **JAR** files.

**Environment**: Eclipse, Web Sphere Application Server, JSP, Servlet, HTML, JUnit, JavaScript, CSS, EJB, Hibernate, Struts, XML, JAXP, CVS, JAX-RPC, AXIS, SOAP, TOAD, AJAX, Jenkins, Maven, Log4J, UNIX, Linux, Java, J2EE, Java API, JSP, Struts, JNDI, Oracle 10g, HTML, XML, Web Logic 8.1, Ant, CVS, Log4J, JUnit, JMS, JDBC, JavaScript, Eclipse IDE, UNIX Shell Scripting, Rational Unified Process (RUP)

**EDUCATION:**

* Master’s in information systems security at University of Cumberland’s, K.Y 2017
* Bachelors in Mechanical Engineering, JNTUH, INDIA 2014