**Vamshi**

Mail:[Vamsivitta2810@gmail.com](mailto:Vamsivitta2810@gmail.com)  **Ph: 9137124282**

**Senior Big Data Engineer**

**Professional Summary:**

* Over 7+ years of professional experience in IT including comprehensive experience in **Big Data** development primarily using **Hadoop** and **Spark** Ecosystems.
* Excellent hands-on experience in developing **Hadoop** Architecture in both **Windows** and **Linux** platforms.
* Experience in design, development, and implementation of big data applications using **Hadoop Ecosystem** frameworks and tools like **HDFS**, **MapReduce**, **Yarn**, **Pig**, **Hive**, **Sqoop**, **Spark**, **Storm, Kafka**, **Flume**, **NiFi**, **Impala**, **Oozie**, **Zookeeper**, **Airflow** etc.
* Hands-on experience in **GIT** and **Maven.**
* Good working knowledge of **AWS** cloud services like **EMR**, **S3**, **Redshift**, **EMR**, **Lambda**, **Glue**, **Data Pipeline**, and **Athena** for **Big Data** development.
* Have good experience working with **Azure BLOB and Data Lake storage** and loading data into Azure SQL Synapse Analytics (DW).
* Proficient in Python and its libraries such as **NumPy, Pandas, Matplotlib, and Seaborn**.
* Expert in prepossessing data in Pandas using visualization, data cleaning, and engineering methods such as looking for Correlations, Imputations, Scaling, and Handling Categories.
* Well-versed with **Cloudera** and **Hortonworks** distributions.
* Experienced in performing code reviews, and involved closely in smoke testing sessions, and retrospective sessions.
* Having a good experience with **Python** scripting language to use in **Spark** development life cycle.
* Worked on **Spark RDDS** to do the transformations on the data.
* Worked on data processing and transformations and actions in spark by using **Python** (**Pyspark**).
* Implemented **Spark** using **python** and utilizing **SparkCore**, **SparkStreaming,** and **SparkSQL** API for faster processing of data instead of MapReduce in Java.
* Used **Spark-SQL** to Load **JSON** data and create Schema RDD and loaded it into Hive Tables and handled structured data using Spark SQL.
* Designed a Spark model for the existing MapReduce model and Migrated MapReduce models to Spark Models using **Python**.
* Experience in importing data from different sources like **HDFS**/**HBase** into **Spark RDD**.
* Wrote ETL scripts in **Python/SQL** for extraction and validating the data.
* Having hands-on experience **on the DevOps CI/CD tools GIT**, **JIRA, Bitbucket, Maven and Jenkins**.
* Exposure to **project deliveries using agile methodologies** and technical discussions with clients and daily communication calls for project analysis specs and development aspects.
* Experience in various domains like Health Insurance, Supply Chain Management and Logistics.
* Developed wrapper scripts using python and **Shell Scripts.**
* Involved in creating **Hivetables**, loading with data and writing **hivequeries** that will run internally in **MapReduce** way.
* Implemented the workflows using **Apache Oozie** framework to automate tasks.
* Written MapReduce code that will take input as log files and parse the logs and structure them in tabular format to facilitate effective querying on the log data.
* Developed **PigLatin** scripts to extract the data from the web server output files to load into HDFS.
* Used **HIVE** to do transformations, event joins and some pre-aggregations before storing the data onto HDFS.
* Used **Hive** to analyze the **partitioned** and **bucketed** data and compute various metrics for reporting.
* Used **hive optimization techniques** during joins and best practices in writing hive scripts using HiveQL.
* Importing and exporting data into **HDFS** and **Hive** using **Sqoop**.
  + Very keen in knowing newer techno stack that Google Cloud platform (GCP) adds.
* Can work parallelly in both GCP and Azure Clouds coherently.

**Technical Skills:**

|  |  |
| --- | --- |
| **Hadoop/Big Data/ETL Technologies** | HDFS, Map Reduce, Sqoop, Pig, Hive, Hbase, Oozie, Impala, Zookeeper, Ambari, Storm, Spark, and Kafka, hue |
| **No SQL Database** | HBase, Cassandra |
| **Monitoring and Reporting** | Tableau, Shell Scripts |
| **Hadoop Distribution** | Horton Works, Cloudera |
| **Build and Deployment Tools** | Maven, GitHub, SVN, Jenkins |
| **Programming and Scripting** | SQL, Linux Shell Scripting, Python, Pig Latin, HiveQL, PySpark |
| **Databases** | Snowflake, Oracle, MY SQL, MS SQL Server, Vertica, Teradata, Sybase and DB2 |
| **Analytics Tools** | Tableau, SAP BO |
| **IDE Dev. Tools** | Eclipse, IntelliJ, PyCharm, Oracle, Ant, Maven |
| **Operating Systems** | Linux, Unix, Windows 8/10, Windows Server, MacOS |
| **Cloud** | AWS, Azure, GCP Cloud Storage |
| **AWS Services** | EC2, EMR, S3, Redshift, EMR, Lambda, Glue, Data Pipeline, Athena, AWS MAWAA |
| **Network protocols** | TCP/IP, UDP, HTTP, DNS, DHCP |
| **Project &Trouble reporting tools** | Jira |

**Project Experience:**

**Client: CITRIX, Fort Lauderdale, FL Apr 2021 – till date**

**Role: Senior Big Data Engineer**

**Responsibilities:**

* Developed Apache **Spark**applications by using **Scala**for data processing from various streaming sources.
* Configured **Spark**Streaming to receive real time data from the **Apache Kafka**and store the stream data to **DynamoDB**using **Scala.**
* Created a set of data classifiers that reads from **DynamoDB** and classifies the features into bins and stores them in **DynamoDB**.
* Collected data using Spark Streaming from **AWS S3 bucket** in near-real-time and performs necessary Transformations and Aggregation on the fly to build the common learner data model and persists the data in **HDFS.**
* Used different stages of Datastage Designer like Lookup, Join, Merge, Funnel, Filter, Copy, Aggregator, and Sort etc.
* Developed and maintained batch data flow using **HiveQL** and **Unix scripting**
* Involved in converting **Hive/SQL queries** into **Spark** transformations using **Spark RDD, Scala** and **Python.**
* Worked with developer teams on **NiFi** workflow to pick up the data from **rest API server,** from data lake as well as from SFTP server and send that to **Kafka.**
* Created functions and assigned roles in **AWS Lambda** to run python scripts, and **AWS Lambda** using java to perform event driven processing.
* Expertise in analyzing data using **Pig scripting, Hive Queries, Sparks (python**) and **Impala.**
* Developed a **python script** to transfer data, **REST API’s** and extract data from on-premises to **AWS S3.** Implemented **Micro Services** based Cloud Architecture using **Spring Boot.**
* Worked on Ingesting data by going through cleansing and transformations and leveraging **AWS Lambda, AWSGlue** and Step Functions.
* Developed applications using Java that reads data from MSK (kafka) and writes it to **Dynamo DB.**
* Developed applications that **leverages step functions**and**cloud watch event triggers** to fetch data and generate features from that data.
* Involved in creating **research data-lake**by extracting customer's data from various data sources to **S3**which include data from Excel, databases, and log data from servers
* Developed business logic using **Kafka** Direct Stream in Spark Streaming and implemented business transformations.
* Accessed the endpoints to call the model in real time to generate the insight
* Created **Hive schemas** using performance techniques like partitioning and bucketing.
* Used **Hadoop YARN** to perform analytics on data in **Hive.**
* Created **python-based**lambda functions for feature extraction.

**Environment:** AWS**,** Lambda, MSK, KMS, Spark, Hadoop Yarn, Hive, Scala, Pig, MapReduce, Impala, SQL Server 2016, DB2, DynamoDB, CloudWatch, Spring boot, Micro Services, Tableau, Python, Pyspark, SNS, Nifi, Oozie, step functions, MongoDB, Hbase, Cassandra, Zookeeper, UNIX

**Client: The Bridge Corp, Philadelphia, PA Jun 2019 – Mar 2021**

**Role: Big Data Engineer**

**Responsibilities:**

* Stored data in **AWS S3** like **HDFS** and performed **EMR** programs on data stored.
* Used the **AWS**-**CLI** to suspend an **AWS Lambda** function. Used **AWS CLI** to automate backups of ephemeral data-stores to **S3 buckets, EBS.**
* Performing **ETL testing** activities like running the Jobs, Extracting the data using necessary queries from database transform, and upload into the Data warehouse servers.
* Developed **HIVE UDFs** to incorporate external business logic into **Hivescript** and Developed join data set scripts using **HIVE** join operations.
* Developed custom **Kafka** producer and consumer for different publishing and subscribing to **Kafka** topics.
* Migrated **Map reduce** jobs to **Spark** jobs to achieve better performance.
* Working on designing the **MapReduce** and Yarn flow and writing **MapReduce** scripts, performance tuning and debugging.
* Implemented **Composite** server for the **data virtualization** needs and created multiples views for restricted data access using a **REST API**.
* 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.**
* 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.
* Designed the data models to be used in data intensive **AWS Lambda** applications which are aimed to do complex analysis creating analytical reports for end-to-end traceability, lineage, definition of Key Business elements from Aurora.
* Developed **Spark** code using Scala and **Spark-SQL**/Streaming for faster testing and processing of data.
* Implemented **AWS** provides a variety of computing and networking services to meet the needs of applications
* Writing **HiveQL** as per the requirements and Processing data in **Spark** engine and store in **Hive** tables.
* Importing existing datasets from Oracle to Hadoop system using **SQOOP**.
* Brought data from various sources into Hadoop and Cassandra using **Kafka**.
* Exported the analyzed data to the relational databases using Sqoop for visualization and to generate reports for the BI team Using **Tableau.**
* Used **Spark**-**Streaming APIs** to perform necessary transformations and actions on the data got from **Kafka**.
* Developed workflow in **Oozie** to automate the tasks of loading the data into **Nifi** and pre-processing with **Pig**.
* Worked on **Apache NIFI** to decompress and move **JSON** files from local to **HDFS**.
* Like Access, Excel, CSV, Oracle, flat files using connectors, tasks and transformations provided by **AWS** Data Pipeline.

**Environment:** Hadoop (HDFS, MapReduce), Scala, Spark, Impala, Hive, MongoDB, HBase, Oozie, Hue, Sqoop, Flume, Oracle, AWS Services, Mysql,Sql Server, Python, Scala, Spark, Hive, Spark-Sql.

**Client: Ditech, Fort Washington, PA. Jan 2017 – May 2019**

**Role: Big Data Engineer**

**Responsibilities:**

* Created pipelines to create a processing pipeline including transformations, estimations, evaluation of analytical models.
* Performed tuning of **SQL queries** and Stored Procedure for speedy extraction of data to resolve and troubleshoot issues in **OLTP** environment.
* Used cloud shell SDK in GCP to configure the services Data Proc, Storage, BigQuery.
* Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Experience in GCP Dataproc, GCS, Cloud functions, BigQuery.
* Worked on **Oozie workflow** engine for job scheduling.
* Worked with **Avro Data** Serialization system to work with JSON data formats.
* Worked on various performance optimizations like using distributed cache for small datasets, partition and bucketing in **Hive**, doing map side joins etc.
* Performed pre-processing on a dataset prior to training, including standardization, normalization.
* Worked with heterogeneous source to Extracted data from **Oracle database, XML** and flat files and loaded to a relational **Oracle warehouse**.
* **Hive Context**, with transformations and actions (**map, flat Map, filter, reduce, reduce** by Key).
* Developed **PIG scripts** for the analysis of semi structured data.
* Migrated **ETL jobs** to **Pig scripts** to do Transformations, even joins and some pre-aggregations before storing the data onto **HDFS**.
* Worked on different file formats like Sequence files, XML files and Map files using **MapReduce** Programs.
* Developed **PIG UDF'S**for manipulating the data according to Business Requirements and worked on developing custom **PIG Loaders.**
* Written **Hive** jobs to parse the logs and structure them in tabular format to facilitate effective querying on the log data.
* Evaluated model accuracy by dividing data into training and test datasets and computing metrics using evaluators.

**Environment:** Hive, HDFS, Oozie, GCP, Map Reduce, Oracle 10g, SQL, OLTP, Windows, MS Office

**Client: Thomson Reuters, India Aug 2015 – Nov 2016**

**Role: Data Engineer**

**Responsibilities:**

* Worked in **Azure** environment for development and deployment of Custom **Hadoop** Applications.
* Responsible to manage data coming from different sources through **Kafka**.
* Used **Spark Data Frames** Operations to perform required Validations in the data and to perform analytics on the **Hive data.** Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of **Azure** Data Factory, T-SQL, **Spark SQL** and U-SQL **Azure** Data Lake Analytics. Data Ingestion to one or more Azure Services - (Azure Data Lake, **Azure** Storage, Azure SQL, **Azure** DW) and processing the data in in **Azure** Data bricks.
* Developed workflow in Oozie to manage and schedule jobs on Hadoop cluster to trigger daily, weekly and monthly batch cycles.
* Configured Hadoop tools like **Hive**, Pig, **Zookeeper**, Flume, Impala and **Sqoop.**
* Deployed the initial **Azure** components like **Azure** Virtual Networks, **Azure** Application Gateway, **Azure** Storage and Affinity groups.
* Working in big data technologies like **spark, Scala**, **Hive**, **Hadoop** cluster (Cloudera platform).
* Making a data pipelining with help Data Fabric job **SQOOP**, **SPARK**, Scala and **KAFKA**. Parallel working in data side oracle and **MYSQL** server for data designing to source to target.
* Used Cloudera Manager continuous monitoring and managing of the **Hadoop** cluster for working application teams to install operating system**, Hadoop** updates, patches, version upgrades as required.
* Developed data pipelines using **Sqoop**, **Pig** and **Hive** to ingest customer member data, clinical, biometrics, lab and claims data into HDFS to perform data analytics.
* Analyzed Teradata procedure and imported all the data from Teradata to **My SQL** Database for **Hive QL** queries information for developing **Hive** Queries which consist of UDF’s where we don’t have some of the default functions in **Hive**.
* Provided design recommendations and thought leadership to sponsors/stakeholders that improved review processes and resolved technical problems. Managed and reviewed **Hadoop log files.**
* 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.
* Used Scala function, dictionary and data structure (array, list, map) for better code reusability
* Based on Development, we need to do the Unit Testing.
* Written multiple **MapReduce** Jobs using Java API, **Pig** and **Hive** for data extraction, transformation and aggregation from multiple file formats including Parquet, Avro, **XML, JSON, CSV**, ORCFILE and other compressed file formats Codecs like gZip, Snappy, Lzo.
* Utilized **Spark SQL** API in PySpark to extract and load data and perform **SQL** queries.
* Primarily involved in Data Migration process using **Azure** by integrating with GitHub repository and Jenkins.

**Environment:** SPARK, Kafka, Map Reduce, , Python, Hadoop, Hive, Pig, Spark, PySpark, SparkSQL, Azure SQL DW, Data brick, Azure Synapse, Azure Data lake, ARM, Azure HDInsight, Blob storage, Apache Spark, Oracle 12c, Cassandra, Git, Zookeeper, Oozie.