**Bindu Sr.Big Data Engineer**

**Email: Bindu.arik1991@gmali.com**

**Phone: 1 469 846 8345**

**PROFESSIONAL SUMMARY:**

* 8+ years of experience in Big Data using Hadoop, AWS, Databricks, GCP, Scala, design and development of web applications using JAVA, Spring boot and data base and data warehousing development using My SQL, Oracle.
* Experience in developing Spark applications using Spark-SQL in Databricks for data extraction, transformation and applying aggregations
* Strong working experience on Big Data Analytics with hands on experience in installing, configuring and using ecosystem components like Hadoop Map reduce, HDFS, HBase, Zookeeper, Hive, Sqoop, Pig, Flume, Cassandra, Kafka and Spark, Oozie, Airflow, NiFi(ETL).
* Good Understanding of Hadoop architecture and Hands on experience with Hadoop components such as Job Tracker, Task Tracker, Name Node, Data Node and Map Reduce concepts and HDFS Framework.
* Expertise in preparing Interactive Data Visualization's using Tableau Software from different sources
* Extensive experience with performance tuning applications on GCP and configuring Hadoop/GCP systems to maximize performance
* Excellent knowledge on HadoopArchitecture and ecosystems such as HDFS, JobTracker, TaskTracker, Name Node, Data Node and MapReduce programming paradigm, Hive, Sqoop, Kafka, HBase, Cassandra, MongoDB, Oozie, Zookeeper, Flume, Impala, Spark with Scala and PySpark.
* Knowledge in installing, configuring, and using Hadoopecosystem components like Hadoop Map Reduce, HDFS, HBase, Oozie, Hive, Sqoop, Pig, Zookeeper and Flume.
* Experience in analyzing data using HiveQL, HBase and custom MapReduce programs in Java.
* Experience in importing and exporting data using Sqoop from HDFS to Relational Database Systems and vice-versa
* Expertise in implementing Ad-hoc queries using HiveQL and good knowledge in creating Hive tables and loading and analyzing data using hive queries
* Strong experience in analyzing large amounts of data sets writing PySpark scripts and Hive queries.
* For the app developing project, I implemented applications with Scala along with Akka and Play framework and implemented Restful services in Spring.
* Running of Apache Hadoop, CDH and Map-R distros, dubbed Elastic MapReduce(EMR) on (EC2)
* Implemented advanced procedures like text analytics and processing using the in-memory computing capabilities like Apache Spark written in Scala.
* Experienced with cloud: Hadoop-on-Azure, AWS/EMR, Cloudera Manager (also direct-Hadoop-EC2 (non EMR)).
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDD and Pyspark concepts.
* Extensively worked on AWS services like EC2, S3, EMR, FSx, Lambda, Cloud watch, RDS, Auto scaling, Cloud Formation, SQS, ECS, EFS, DynamoDB, Route53, Glue etc.
* Hands on experience in VPN Putty and WinSCP, CI/CD(Jenkins).
* Experience in Data load management, importing & exporting data using SQOOP&FLUME.
* Experience in analyzing data using Hive, Pig and custom MR programs in Java.
* Experienced in writing Map Reduce programs & UDF's for both Pig & Hive in java.
* Experience in dealing with log files to extract data and to copy into HDFS using flume.
* Experience in integrating Hive and Hbase for effective operations.
* Experience in Impala, Solr, MongoDB, HBase and Spark, Kubernetes.
* Hands on knowledge of writing code in Scala, Core Javaand also with R.
* Expertise in Waterfall and Agile - SCRUM methodologies.
* Experienced with code versioning and dependency management systems such as Git, SVT, and Maven.
* Experience in Testing and documenting software for client applications.
* Writing code to create single-threaded, multi-threaded or user interface event driven applications, either stand-alone and those which access servers or services.
* Good experience in using Data Modelling techniques to find the results based on SQL and PL/SQL queries.
* Experience working with different databases, such as Oracle, SQL Server, MySQL and writing stored procedures, functions, joins, and triggers for different Data Models.
* Great team player and quick learner with effective communication, motivation, and organizational skills combined with attention to details and business improvements.
* Experienced in handling different file formats like Text file, Avro data files, Sequence files, Xml and Json files
* Developed ApacheSpark jobs using Scala in test environment for faster data processing and used SparkSQL for querying.
* Experienced in SparkCore, SparkRDD and Spark Deployment Architectures.
* Experienced working with SparkStreaming, SparkSQL and Kafka for real-time data processing.
* Used SparkData Frames API over Cloudera platform to perform analytics on Hive data.
* Developed ApacheSpark jobs using Scala in test environment for faster data processing.
* Worked on loading CSV/TXT/AVRO/PARQUET files using Scala/Java language in SparkFramework and process the data by creating SparkDataframe and RDD and save the file in parquet format in HDFS
* Imported the data from different sources like AWSS3, Local file system into SparkRDD.
* Experienced with performing real time analytics on NoSQL data bases like HBase and Cassandra.
* Experience in Data Modeling and working with Cassandra Query Language (CQL).
* Extracted data from MongoDB through Sqoop and placed in HDFS and wrote Queries to generate reports.
* Created dataflow between SQLServer and Hadoopclusters using ApacheNifi

**TECHNICAL SKILLS**:

|  |  |
| --- | --- |
| Big Data Technologies | HDFS, Hive, MapReduce, Pig, Sqoop, Flume, Oozie, Hadoop distribution, and Hbase, Spark, Spark Streaming, Yarn, Zookeeper, Kafka, ETL.(Nifi, Talend etc.) |
| Programming languages | Core Java, Spring Boot, R, Scala, Terraform, Angular. |
| Databases | MySQL, MS-SQL Server 20012/16, Oracle 10g/11g/12c |
| Scripting/Web Languages | HTML5, CSS3, XML, SQL, Shell/Unix, Perl, Python. |
| Databases | Cassandra, HBASE, mongoDB, Oracle, MS SQL, Teradata. |
| Operating Systems | Linux, Windows XP/7/8/10, Mac. |
| Software Life Cycle | SDLC, Waterfall and Agile models. |
| Utilities/Tools | Eclipse, Tomcat, NetBeans, JUnit, SQL, SVN, Log4j, SOAP UI, ANT, Maven, Alteryx, Visio, Jenkins, Jira, Intellij. |
| Data Visualization Tolls | Tableau, SSRS, Cloud Health. |
| Cloud Services | AWS(EC2, S3, EMR, RDS, Lambda, Cloudwatch, FSx, Auto scaling, Redshift,Cloud Formation, Glueetc.) , Azure Databricks, GCP. |

**PROFESSIONAL EXPERIENCE:**

**CapitalOne, TX July 2021– Current**

**Sr. Big data Engineer**

* Involved in Requirement gathering, Business Analysis and translated business requirements into Technical design in Hadoop and Big Data
* Involved in SQOOP implementation which helps in loading data from various RDBMS sources to Hadoop systems
* Good understanding of Spark Architecture with Databricks, Structured Streaming. Setting Up AWS and Microsoft Azure with Databricks, Databricks Workspace for Business Analytics, Manage Clusters In Databricks, Managing the Machine Learning Lifecycle
* Developed Python scripts to extract the data from the web server output files to load into HDFS.
* Involved in HBASE setup and storing data into HBASE, which will be used for further analysis.
* Worked on Written a python script which automates to launch the EMR cluster and configures the Hadoop applications using boto3.
* Created various data pipelines usingSpark, Scala and SparkSQL for faster processing of data.
* Written Spark-SQL and embedded the SQL in SCALA files to generate jar files for submission onto the Hadoop cluster
* Created Data Quality Scripts using SQL and Hive to validate successful das ta load and quality of the data. Created various types of data visualizations using Python and Tableau.
* Extensively worked with Avro and Parquet, XML, JSON files and converted the data from either format Parsed Semi Structured JSON data and converted to Parquet using Data Frames in PySpark.
* Create Pyspark frame to bring data from DB2 to Amazon S3
* Experienced in working with spark eco system using Spark SQLand Scala queries on different formats like Text file, CSV file,transformation in GCP
* Experience in developing Spark applications using Spark-SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats for Analyzing& transforming the data to uncover insights into the customer usage patterns
* Provide guidance to development team working on PySpark as ETL platform.
* Developed a Python Script to load the CSV files into the S3 buckets and created AWS S3buckets, performed folder management in each bucket, managed logs and objects within each bucket.
* Analysed the sql scripts and designed it by using PySpark SQL for faster performance.
* Involved in Analyzing system failures, identifying root causes, and recommended course of actions, Documented the systems processes and procedures for future references.
* Involved in Configuring Hadoop cluster and load balancing across the nodes.
* Involved in Hadoop installation, Commissioning, Decommissioning, Balancing, Troubleshooting, Monitoring and, debugging Configuration of multiple nodes using Hortonworks platform.
* Involved in working with Spark on top of Yarn/MRv2 for interactive and Batch Analysis
* Worked with a team to migrate from Legacy/On prem environment into AWS.
* Created Dockerized backend cloud applications with exposed Application Program Interface (API) interfaces and deployed on Kubernetes.
* Stored and retrieved data from data-warehouses using Amazon Redshift.
* Experienced in analyzing and Optimizing RDD's by controlling partitions for the given data
* Experienced in writing live Real-time Processing using Spark Streaming with Kafka
* Used HiveQL to analyze the partitioned and bucketed data and compute various metrics for reporting
* Worked with querying data using SparkSQL on top of Spark engine.
* Involved in managing and monitoring Hadoop cluster using Cloudera Manager.
* Used Python and Shell scripting to build pipelines.
* Developed data pipeline using SQOOP, HQL, Spark and Kafka to ingest Enterprise message delivery data into HDFS.
* Developed workflow in Oozie to automate the tasks of loading data into HDFS and pre-processing with Pig and Hive.
* Developed and executed a migration strategy to move Data Warehouse from an Oracle platform to AWS Redshift.
* Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift.
* Assisted in creating and maintaining Technical documentation to launching HADOOP Clusters and even for executing Hive queries and Pig Scripts.
* Assisted in Cluster maintenance, cluster monitoring, adding and removing cluster nodes and Installed and configured Hadoop, Map Reduce, HDFS, developed multiple Map Reduce jobs in java for data cleaning and pre-processing.
* Involved in file movements between HDFS and AWS S3 and extensively worked with S3 bucket in AWS.
* Automated and monitored complete AWS infrastructure with terraform.
* Created data partitions on large data sets in S3 and DDL on partitioned data.
* Converted all Hadoop jobs to run in EMR by configuring the cluster according to the data size.

**Environment**: HDFS, Hive, Scala, Sqoop, Spark,Tableau, Yarn, Cloudera, SQL,Terraform, Splunk, RDBMS, Elastic search, Kerberos, Jira, Confluence, Shell/Perl Scripting, Zookeeper, AWS(EC2, S3, EMR, Redshift, ECS,Glue, S3, VPC, RDS etc.), Ranger, Git, Kafka, CI/CD(Jenkins), Kubernetes, Azure Databricks.

**Alcon Laboratories, Fort Worth, TX Feb 2020 – Jun 2021**

**Big data Engineer**

**Responsibilities:**

* Import data from sources like HDFS/HBase into Spark RDD.
* Usage of Spark Streaming and Spark SQL API to process the files.
* Migrated their Big Data Platform from on-premise Hadoop to Google Cloud Platform (GCP) to one of the projects that we are working
* Worked extensively with Sqoop for importing and exporting the data from HDFS to Relational Database systems/mainframe and vice-versa loading data into HDFS
* Hands on experience on Unified Data Analytics with Databricks, Databricks Workspace User Interface, Managing Databricks Notebooks, Delta Lake with Python, Delta Lake with Spark SQL
* Stored data in AWS S3 like HDFS and performed EMR programs on data stored in S3.
* Worked on Big Data Hadoop cluster implementation and data integration in developing large-scale system software
* Developing UDFs in java for hive and pig and worked on reading multiple data formats on HDFS using Scala.
* Developed workflow in Oozie to automate the tasks of loading data into HDFS and pre-processing with Hive.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs and Scala.
* Involved in Migrating the platform from Cloudera to EMR platform.
* Developed analytical component using Scala, Spark and Spark Streaming.
* Collecting and aggregating large amounts of log data using Apache Flume and staging data in HDFS for further analysis
* Involved in creating Hive Tables, loading with data and writing Hive queries which will invoke and run MapReduce jobs in the backend.
* Extensively involved in developing Restful API using JSON library of Play framework.
* Developed Storm topology to ingest data from various source into Hadoop Data Lake.
* Developed web application using HBase and Hive API to compare schema between HBase and Hive tables.
* Played a vital role in Scala/Akka framework for web based applications
* Connected to AWS s3 using SSH and ran spark-submit jobs
* Developed Python Script to import data SQL Server into HDFS & created Hive views on data in HDFS using Spark.
* Expert in Troubleshooting MapReduce Jobs.
* Created scripts to append data from temporary HBase table to target HBase table in Spark.
* Developed complex and Multi-step data pipeline using Spark.
* Worked on Big Data Integration and Analytics based on Hadoop, SOLR, Spark, Kafka, Storm and web Methods technologies.
* Populated HDFS and Cassandra with huge amounts of data using Apache Kafka.
* Monitoring YARN applications. Troubleshoot and resolve cluster related system problems.
* Upgrading the Hadoop Cluster from CDH3 to CDH4, setting up High Availability Cluster and integrating HIVE with existing applications.
* Collecting and aggregating large amounts of log data using Apache Flume and staging data in HDFS for further analysis.
* Involved in creating ETL flow using Pig, loading with data and writing Pig Latin queries which will run internally in Map Reduce way.
* Involved in writing Unix/Linux Shell Scripting for scheduling jobs and for writing pig scripts and hive QL.
* Involved in creating Hive Tables, loading with data and writing Hive queries which will invoke and run MapReduce jobs in the backend.
* Assisted in exporting data into Cassandra and writing column families to provide fast listing outputs.
* Used Zookeeper for providing coordinating services to the cluster.
* Worked with Hue UI in scheduling jobs with ease and File browsing, Job browsing, Metastore management.
* Developed and designed system to collect data from multiple portal using kafka and then process it using spark.

**Environment**: Hadoop, HDFS, Hive, Core Java, Sqoop, Spark, Scala, Hive, Cloudera CDH4, Oracle, Elastic search, Kerberos, SFTP, Databricks,Impala, Jira, Wiki, Alteryx, Teradata, Shell/Perl Scripting, Kafka, AWS EC2, S3, EMR, Cloudera.

**PWC, Tampa, FL Dec 2017 – Jan 2020**

**Big data Engineer**

**Responsibilities:**

* Installed and configured multi-nodes fully distributed Hadoop cluster for large numbers of nodes.
* Experience on Hortonworks and Cloudera Manager.
* Implemented multiple Spark Jobs in Scala and Sql for data cleaning and Transformation.
* Involved in testing HDFS, Hive, Pig and Map Reduce access for the new users.
* Cluster maintenance as well as creation and removal of nodes using Apache Ambari.
* Configured Zookeeper to implement node coordination in clustering support.
* Creating snapshots and restoring snapshots.
* Worked on setting up Hadoop cluster for the Production Environment.
* Used Impala to read, write and query the Hadoop data in HDFS.
* Tested Mesos frameworks such as Kafka and tested isolation.
* Used impala for optimization of query performance instead of HIVE.
* Data processing using spark.
* Involved installation and configuration of Tableau server.
* Experience in understanding the security requirements for Hadoop and Integrating with Kerberos authentication infrastructure-KDC server setup, creating realm/domain.
* Building massively scalable multi-threaded application for bulk data processing primarily with Apache Spark and PIG on Hadoop.
* Developed Scripts and Batch job to schedule various Hadoop program.
* Involved in cluster capacity Planning, Hardware Planning, Installation, Performance tuning of the Hadoop Cluster.
* Load log data into HDFS using Flume.
* MapReduce jobs to power data to search and aggregation.
* Enable HA for Name node, Resource Manager, Yarn Configuration and Hive Megastore.
* Run the benchmark tools to test the cluster performance.
* Provided support to users for diagnosis, reproducing and fixing Hadoop related issues.
* Ensure the critical user issues are addressed quickly and effectively.
* Configured Rack Awareness on HDP clusters.

**Environment:**Cloudera Manager, Core Java, Hortonworks, HDFS, Yarn, Spark, Hive, Pig, HBase, MapReduce, Sqoop, Flume, Kerberos, Zookeeper, RHEL

**T. Rowe Price, Owing Mills, MD Jan 2015 – Nov 2017**

**Big data/Scala Developer**

**Responsibilities:**

* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs and Scala.
* Developed multiple POCs using Scala and deployed on the Yarn cluster, compared the performance of Spark, with Hive and SQL/Teradata.
* Analysed the SQL scripts and designed the solution to implement using Scala.
* Developed analytical component using Scala, Spark and Spark Stream.
* Developing UDFs in java for hive and pig and worked on reading multiple data formats on HDFS using Scala.
* Used Scala collection framework to store and process the complex consumer information.
* Used Scala functional programming concepts to develop business logic.
* Designed and implemented Apache Spark Application (Cloudera)
* Importing and exporting data into HDFS Sqoop and Flume and Kafka.
* Troubleshoot and debug Hadoop ecosystem run-time issues.
* Worked with Play framework and Akka parallel processing.
* Hands on experience in Multithreaded programming using akka actors.
* Developed Microservices based on Restful web service using Akka Actors and Akka-Http framework in Scala which handles high concurrency and high volume of traffic
* Used Oozie Scheduler systems to automate the pipeline workflow and orchestrate the map reduce jobs that extract
* Worked with Hue GUI in scheduling jobs with ease and File browsing, Job browsing, Metastore management.
* Worked with BI team in the area of Big Data Hadoop cluster implementation and data integration in developing large-scale system software.
* Installed/Configured/Maintained Apache Hadoop clusters for application development and Hadoop tools like Hive, Pig, HBase, Flume, Oozie Zookeeper and Sqoop.
* Responsible for cluster maintenance, adding and removing cluster nodes, cluster monitoring and troubleshoot managing and reviewing data backups and Hadoop log files.
* Continuous monitoring and managing the Hadoop cluster through Cloudera Manager.
* Extensively involved in Installation and configuration of Cloudera distribution Hadoop, NameNode, Secondary NameNode, JobTracker, TaskTrackers and DataNodes.
* Created POC to store Server Log data in MongoDB to identify System Alert Metrics.
* Monitored Hadoop cluster job performance, performed capacity planning and managed nodes on Hadoop cluster.
* Worked with application teams to install operating system, Hadoop updates, patches, version upgrades as required.
* Loaded data into the cluster from dynamically generated files using Flume and from relational database management systems using Sqoop.
* Performed analysis on the unused user navigation data by loading into HDFS and writing MapReduce jobs. The analysis provided inputs to the new APM front end developers and lucent team.
* Wrote MapReduce jobs using Java API and Pig Latin.
* Wrote Pig scripts to run ETL jobs on the data in HDFS and further do testing.
* Used Hive to do analysis on the data and identify different correlations.
* Involved in HDFS maintenance and administering it through Hadoop-Java API.
* Written Hive queries for data analysis to meet the business requirements.
* Automated all the jobs, for pulling data from FTP server to load data into Hive tables, using Oozie workflows.
* Involved in creating Hive tables & working on them using HiveQL and perform data analysis using Hive and Pig.
* Used Qlikview and D3 for visualization of query required by BI team.
* Defined UDFs using PIG and Hive in order to capture customer behavior.
* Design and implement MapReduce jobs to support distributed processing using java, Hive and Apache Pig.
* Create Hive external tables on the MapReduce output before partitioning, bucketing is applied on it.
* Loaded the load ready files from mainframes to Hadoop and files were converted to ASCII format.
* Configured Hive Server (HS2) to enable analytical tools like Tableau, Qlikview and SAS to interact with Hive tables**.**

**Environment:** Hadoop, MapReduce, HDFS, Hive, Java, SQL, Cloudera Manager, Pig, Sqoop, ZooKeeper, Teradata, PL/SQL, MySQL, Hbase, ETL(Informatica/SSIS).