**Shanmukha**

[**Shanmukhabigdata2@gmail.com**](mailto:Shanmukhabigdata2@gmail.com)

**+1( 928) 793 2128**

**DataBase/Big data/Hadoop**

**PROFESSIONAL SUMMARY:**

* Over 6 years of IT experience in software development and support with experience in developing strategic methods for deploying **Big Data technologies** to efficiently solve Big Data processing requirement.
* Expertise in Hadoop eco system components **HDFS, Map Reduce, Yarn, HBase, Pig, Sqoop, Spark, Spark SQL, Spark Streaming** and **Hive** for scalability, distributed computing, and high performance computing.
* Experience in using **Hive Query** Language for data Analytics.
* Experienced in Installing, Maintaining and Configuring Hadoop Cluster.
* Strong knowledge on creating and monitoring Hadoop clusters on **Amazon EC2, VM, Horton works Data Platform 2.1 & 2.2, CDH3, CDH4Cloudera Manager on Linux, Ubuntu OS** etc.
* Capable of processing large sets of structured, semi-structured and unstructured data and supporting systems application architecture.
* Having Good knowledge on Single node and Multi node Cluster Configurations.
* Strong knowledge in **NOSQL** column oriented databases like **HBase, Cassandra, MongoDB** and its integration with Hadoop cluster.
* Expertise on **Scala Programming language** and **Spark Core.**
* Worked with **AWS** based data ingestion and transformations.
* Worked with **Cloud Break and Blueprint** to configure **AWS** platform.
* Good knowledge to run **TDCH** jobs to get data from Teradata to hadoop (HDFS).
* Experienced in job workflow scheduling and monitoring tools like **Oozie and Zookeeper**.
* Good knowledge on **Amazon EMR, Amazon RDS S3 Buckets, Dynamo DB, Redshift**.
* Good Hands on to run **zeppelin** notebook for running analytics on big data.
* Good experience on Kafka and Storm. Flexible with Unix/Linux and Windows Environments working with Operating Systems like **Centos 5/6, Ubuntu 13/14, Cosmos**.
* Experience working with **EDW (Enterprise Data Warehouse)** analytics.
* Experience in build scripts using **Maven** and do continuous integrations systems like **Jenkins**.
* Very Good understanding of **SQL, ETL** and **Data Warehousing Technologies**
* Knowledge of **MS SQL Server2012/2008/2005 and Oracle 11g/10g/9i** and **E-Business Suite**.
* Experience working with **Solr** for text search. Experience on using **Talend ETL tool**.
* Experience in working with job scheduler like Autosys and Maestro.
* Strong in databases like **Sybase, DB2, Oracle, MS SQL, Click stream**.
* Hands on experience with automation tools such as **Puppet, Jenkins**.

**EDUCATIONAL QUALIFICATION:**

* Bachelor of Engineering from JNTUK, India.
* MS in Computer Science from NYIT, New York , USA.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| Hadoop/Big Data Technologies | **HDFS, Map Reduce, Sqoop, Flume, Pig, Hive, Oozie, impala, Spark, Zookeeper and Cloudera Manager.** |
| NO SQL Database | **HBase, Cassandra** |
| Monitoring and Reporting | **Tableau, Custom shell scripts,** |
| Hadoop Distribution | **Horton Works, Cloudera, MapR** |
| Build Tools | **Maven, SQL Developer** |
| Programming & Scripting | **JAVA, C, SQL, Shell Scripting, Python, Scala** |
| Java Technologies | **JDBC** |
| Databases | **Oracle, MY SQL, MS SQL server, Teradata** |
| Web Dev. Technologies | **HTML, XML** |
| Version Control | **SVN, CVS, GIT** |
| Operating Systems | **Linux, Unix, Mac OS-X, Cen OS, Windows10, Windows 8, Windows 7, Windows Server 2008/2003** |

**PROFESSIONAL EXPERIENCE:**

**Neustar, Sterling, VA May 2018 – Present Hadoop/Big Data Engineer**

Neustar is one of the major leading providers of technology solutions and services on data related to communications like television, telecom, internet, iptv. While dealing with the telecom and television network data, got the data from different vendors which got deployed into Teradata source and from teradata I use to get the data into our production sandbox with in the cluster on daily and monthly basis as COIL tables and then deploy the processed data to the actual tables for the availability of BI, Programming, Analytics and Marketing teams.

• Worked on **Distributed/Cloud Computing (Map Reduce/Hadoop, Hive, Pig, HBase, Sqoop, Spark AVRO, Zookeeper etc.), Hortonworks data plotform (HDP).**

• Importing and exporting data into **HDFS, Pig, Hive and HBase using SQOOP**.

• Responsible to manage data coming from different sources.

• Flume and from relational database management systems using **SQOOP**.

• Responsible to manage data coming from different data sources.

• Developed Pig scripts for data analysis and extended its functionality by developing custom **UDF's**.Developed POC using Scala, Spark SQL and MLlib libraries along with Kafka and other tools as per requirement then deployed on the Yarn cluster.

• Experience working with **ORACLE DBA**.

• Extensive knowledge on **PIG** scripts using bags and tuples.

• Worked with query engines **Tez, Apache Phoenix**.

• Worked on Horton works Data Platform (HDP). Extract Real time feed using Kafka and Spark Streaming and convert it to RDD and process data in the form of Data Frame and save the data as Parquet format in HDFS.

• Good knowledge to run TDCH jobs to get data from **Teradata** to **hadoop (HDFS)**.

• Good Hands on to run **zeppelin** notebook for running analytics on big data.

• Developed data pipeline **using Flume, Sqoop, Pig and Java Map Reduce** to ingest customer behavioral data and financial histories into HDFS for analysis.

• Involved in collecting and aggregating large amounts of log data using **Apache Flume** and staging data in HDFS for further analysis.Experience working with **Talend** data integration for transferring data through different sources and target.

• Worked on **installing cluster, commissioning & decommissioning of Data Nodes, Name Node recovery, capacity planning, and slots** configuration.

• Developed **PIG Latin** scripts to extract the data from the web server output files to load into HDFS.

* Worked on **cloudera data plotform.**
* Experience working on **NETIZZA**, **Yellowbrick** datawarehouse
* Experience working with cloud plotform **AWS** data lake.
* Experience working with **Datastage** ETL plotform
* Worked on Mesos cluster and Marathon.
* Worked with Business Intelligent(BI) Concepts and Data Warehousing Technologies using Power BI **and R** Statistics.

**Environment:** Java 1.8, Scala 2.10.5, Apache Spark 1.6.2, Apache Zeppelin, GreenPlum 4.3 (PostgreSQL), Hive, HDFS, Sqoop 1.4.3, Flume, SOLR, Kafka, Impala, Apache Cassandra, UNIX Shell Scripting, Python 2.6, Splunk.

**Accolite, Madison, WI July 2015 – Apr 2018**

**Hadoop Developer**

**Description:** Accolite is now one of the nation's leading providers of technology solutions and services in healthcare, finance, insurance, telecom, and manufacturing. While dealing with healthcare maintained database related to insurer, policyholder, subscriber, claims, flexible spending accounts and master savings accounts details using hadoop. Enabled Online Renewals and capture request for Endorsements. Provided Online Premium Calculator for Non Registered/ Registered Users Provided Online Customer Support like Chat, Agent Locators, Branch Locators, FAQs, Best Plan Selector, to increase the likelihood of sale.

**Responsibilities:**

* Installed/Configured/Maintained Apache Hadoop clusters for Analytics, application development and Hadoop tools like **Hive, HSQL Pig, HBase, OLAP, Zookeeper, Avro, parquet** and **Sqoop** on Linux **ARCH**.
* Responsible for developing data pipeline using **Azure HD Insight, flume, Sqoop** and **pig** to extract the data from weblogs and store in HDFS.
* Installed **Oozie** workflow engine to run multiple **Hive** and **Pig** Jobs, used Sqoop to import and export data from HDFS to RDBMS and vice-versa for visualization and to generate reports.
* Involved in migration of **ETL** processes from Oracle to Hive to test the easy data manipulation.
* Worked in functional, system, and regression testing activities with agile methodology.1
* Worked on Python plugin on **MySQL** workbench to upload CSV files.
* Used Hive to analyze the partitioned and bucketed data and compute various metrics for reporting.
* Worked with HDFS Storage Formats like **Avro, Orc, Parquet, AWS** based data ingestion, transformations, **Clojure, Kafka** and **Storm** and **EMR (Elastic Map Reduce).**
* Worked with **Accumulo** to Modify server side Key Value pairs. Working experience with **shiny** and **R.**
* Good knowledge to run **TDCH** jobs to get data from teradata to hadoop (HDFS).
* Good Hands on to run **zeppelin** notebook for running analytics on big data.
* Worked with NoSQL databases like **HBase, Cassandra, DynamoDB**
* Worked on importing and exporting data from Oracle and DB2 into **HDFS** and **HIVE using Sqoop**.
* Responsible for building scalable distributed data solutions using Hadoop, and responsible for Cluster maintenance, adding and removing cluster nodes, Cluster Monitoring, and Troubleshooting, Manage and review data backups and log files.
* Developed several new Map Reduce programs to analyze and transform the data to uncover insights into the customer usage patterns.
* Worked extensively with importing metadata into Hive using Sqoop and migrated existing **ACID** tables and applications to work on **Hive**.
* Responsible for running Hadoop streaming jobs to process terabytes of xml's data, utilized cluster co-ordination services through **Zookeeper**.
* Wrote the shell scripts to monitor the health check of Hadoop daemon services and respond accordingly to any warning or failure conditions.
* Installed and configured **Hadoop, Map Reduce, HDFS** (Hadoop Distributed File System), developed multiple Map Reduce jobs in java for data cleaning.
* Developed data pipeline using **Flume, Sqoop, Pig** and **Java Map Reduce** to ingest customer behavioral data and financial histories into HDFS for analysis.
* Involved in collecting and aggregating large amounts of log data using Apache Flume and staging data in HDFS for further analysis.
* Worked on installing cluster, commissioning & decommissioning of Data Nodes, Name Node recovery, capacity planning, and slots configuration.
* Developed **PIG Latin** scripts to extract the data from the web server output files to load into HDFS.
* Worked on **Horton works Data Platform (HDP).** Worked with **SPLUNK** to analyze and visualize data.
* Worked on **Mesos cluster** and **Marathon.**
* Used Pig as **ETL** tool to do transformations, event joins and some pre-aggregations before storing the data onto **HDFS**.
* Worked with Orchestration tools like **Airflow**. Experience working with machine learning using **sparkML**.
* Write test cases, analyze and reporting test results to product teams.
* Worked with **AWS data pipeline** within the **data lake**.
* Worked with **Elastic Search, Postgres, Apache NIFI.**
* Hadoop workflow management using **Oozie, Azkaban,Hamake**.
* Worked with the automation of the streaming process using the **puppet** tool.
* Worked on the core and Spark SQL modules of Spark extensively.
* Worked on Descriptive statistics Using **R**.
* Developed Kafka producer and consumers**, HBase clients, Spark,shark,Streams** and **Hadoop MapReduce** jobs along with components on **HDFS, Hive**.
* Strong Working experience in **snowflake,Clickstream.**
* Experience working with **Elastic search** for data search and visualization.
* Experience working with data integration with Talend through different data sorces and platforms.
* Analyzed the SQL scripts and designed the solution to implement using **PySpark**.
* Experience using **Spark** with **Neo4J** where acquiring the interrelated graphical information of the insurer and to query the data from the stored graphs.
* Worked with **Neo4j** and **spark** integration to get the data to and from **Neo4j** and analyse the data using **Tableau**
* Experience in writing batch processing huge Scala programs.
* Worked with Data warehouse tools to perform ETL using **Informatica** and **Talend**.
* Querying with **ANSI SQL** which works on oracle SQL.
* Load and transform large sets of structured, semi structured, and unstructured data using **Hadoop/Big Data** concepts.
* Responsible for creating Hive External tables and loaded the data in to tables and query data using HQL.
* Handled importing data from various data sources, performed transformations using **Hive, MapReduce**, and loaded data into HDFS.

**Environment:** Hadoop Cluster, HDFS, Hive, Pig, Sqoop, OLAP, Kafka, Data Modelling, Linux, Hadoop Map Reduce, HBase, Shell Scripting, MongoDB, and Cassandra, Apache Spark,Neo4J.

**Mind Tree, Hyderabad, India. Apr 2012 – May 2014**

**JAVA/J2EE Developer**

**Description:** Mind Tree specializes in application Development and Maintenance, Data Analytics, Digital Services, EAI BPM, Independent Testing, and IT Strategy & Consulting. Mind tree provides various research and Development services such as Bluetooth Solutions, Digital video surveillance, Mind Test.

**Responsibilities:**

* Involved in **Java, J2EE, struts, web services** and Hibernate in a fast-paced development environment.
* Followed agile methodology, interacted directly with the client on the features, implemented optimal solutions, and tailor application to customer needs.
* Involved in design and implementation of web tier using Servlets and JSP.
* Used Apache POI for Excel files reading.
* Developed the user interface using JSP and Java Script to view all online trading transactions.
* Used **JSP** and **JSTL** Tag Libraries for developing User Interface components.
* Performing Code Reviews.
* Developed the Apache Storm, Kafka, and HDFS integration project to do a real-time data analysis.
* Performed unit testing, system testing and integration testing.
* Designed and developed **Data Access Objects (DAO)** to access the database.
* Used DAO Factory and value object design patterns to organize and integrate the JAVA Objects
* Coded Java Server Pages for the Dynamic front end content that use Servlets and EJBs.
* Coded **HTML** pages using CSS for static content generation with JavaScript for validations.
* Used **JDBC** API to connect to the database and carry out database operations.
* Involved in building and deployment of application in Linux environment.

**Environment: Java, J2EE, JDBC, Struts, SQL. Hibernate, Eclipse, Apache POI, CSS**.