**KAVYA SREE MUSUNURU**

|  |  |
| --- | --- |
| Contact Number | **732-631-4349** |
| Alternate Contact Number | **N/A** |
| Current location of the candidate (City/State) | **West Des Moines, IA** |
| Email Address | **hadoopksm@gmail.com** |
| Skype ID | **sreemusunuru531** |

**Sr. Big Data Engineer**

**PROFESSIONAL EXPERIENCE**

* Around 8 years of IT experience in a variety of industries including 4 years of **Big Data** experience in data ingestion, storage, querying, processing and analysis.
* In depth understanding and knowledge of Hadoop Architecture and its components such as HDFS, Map Reduce, Job Tracker, Task Tracker, Name Node, Data Node, Resource Manager, Node Manager.
* Well versed in installing, configuring, administrating and tuning Hadoop cluster of major Hadoop distributions **Cloudera CDH 3/4/5**, **Hortonworks HDP 2.3/2.4** and **Amazon Web Services (AWS).**
* Hands on experience using **YARN**, and tools like **Pig** and **Hive** for data analysis, **Sqoop** for data ingestion, **Oozie** for scheduling, **Zookeeper** for coordinating cluster resources.
* Experienced in **Amazon Web Services (AWS)** Cloud services like EMR, EC2, S3, EBS and IAM roles for users.
* Solid knowledge in importing and exporting data from different databases like **Oracle**, **MySql** into **HDFS** and vice versa using **Sqoop**.
* Experience in developing Spark applications using the **Core**, **Spark SQL**
* Experienced in handling large datasets using **Partitions**, Spark in Memory capabilities, **Broadcasts** in Spark, Effective & efficient **Joins**, **Transformations** and other during ingestion process itself.
* Integrate **Kafka** and **Storm** by using **Avro** for serializing and deserializing the data and Kafka producer and consumer.
* Experience in writing custom **Map Reduce** jobs and writing **User Defined Functions** (UDF) for evaluation, filtering, loading and storing data in both **Hive** and **Pig**.
* Experience in Collecting log data from various sources and integrating into HDFS using **Flume** and staging data in HDFS for further analysis.
* Knowledge of using **Teradata** connector for Hadoop by loading files to **HDFS** and further to **Hive**.
* Experience in developing **Oozie** workflow for scheduling and orchestrating the ETL process.
* Hands on experience in working with NoSQL databases like **HBase**, **Cassandra**.
* Developed workflows using **Talend** for data pre-processing
* Experience in analyzing data using **Hive**, **Pig**, **HBase** and custom **Map** **Reduce** programs in **Java**.
* Developed **Pig** **Latin** scripts to extract the data from the web server output files to load into HDFS.
* Familiar with almost all the SDLCycles like agile, waterfall, spiral methodologies.
* Dealt with huge transaction volumes while interfacing the front end application written in **Java, JSP, Struts, Hibernate, SOAP Web service** an**d** with **Tomcat Web server.**
* Familiar using building tools like **ANT** and **MAVEN**.
* Good Knowledge of data compression formats like **Snappy**, **Avro**.
* Experienced in Database development, ETL, OLAP, OLTP.
* Experience in writing SQL queries, Stored Procedures for accessing and managing databases such as Oracle 11g/10g/9ix, MS SQL Server 2012/2008/2005, MS Access.
* High level on hands on experience using version control tools like GIT, CVS, SVN, and Clear Case, using IDE’s like Eclipse, IntelliJ, NetBeans and worked with Linux Red Hat, Ubuntu and Windows Operating Systems.
* Experience with all stages of the SDLC and **Agile Development** model right from the requirement gathering to Deployment and production support.
* Involved in daily **SCRUM** meetings to discuss the development/progress and was active in making scrum meetings more productive.
* Collaborated with technical team members to integrate back/front end issues.

**AREAS OF EXPERTISE**

Languages C, Java, J2EE, SQL, PL/SQL, Python, Scala

Big Data Ecosystem Datalake, HDFS, MapReduce, YARN, Hive, Pig, HBase, Impala, Tez, Zookeeper, Sqoop, Pentaho, Oozie, Apache Cassandra, Flume, Spark, Scala, AWS, Hcatolog, Kafka, Avro, Teradata

Scripting language Pig, Python, UNIX, LINUX

Hadoop Clusters Cloudera CDH 4/5, Hortonworks HDP 2.3/2.4, Amazon Web Services (AWS)

J2EE Technologies JSP, Servlets, HTML, CSS, JDBC, SOAP, XSLT.

XML Technologies XML, XSLT

Databases MySQL, Oracle 11g/10g/9i, SQL Server, MS Access

IDE’s Eclipse, Net Beans, IntelliJ.

Development Tools Maven, TOAD, SQL Workbench, Ant, Talend 6.0

Operating Systems Linux (Redhat, CentOS, Ubuntu), UNIX, Mac OS, Sun Solaris and Windows

Version Control SVN, CVS, Clear case, GIT

**EDUCATION**

Bachelor of Technology in Computer Science, JNTUK-2009.

**WORK EXPERIENCE**

**Client: CDS Global Inc. Location: Des Moines, IA**

**Role: Big Data Engineer Duration: Oct 2016 – May 2017**

**Description:** CDS Global Inc. is a leading provider of third party application services to various sector clients such as Hearst Publishing, Experian, Taunton. Moving the data from Hortonworks environment to AWS EMR as the companies other divisions use AWS to handle the data and storage.

**Responsibilities**

* Due to storage issue we have moved the data from **Hortonworks** cluster to **AWS** **EMR** cluster.
* Involved in running Hadoop jobs for processing millions of records and data gets updated on daily and weekly basis and documenting each and every task is done.
* Worked directly with the Big Data Architecture Team that is creating the foundation of this Enterprise Analytics initiative in a Hadoop-based Data Lake.
* Continuous data integration from **Mainframe** systems to Amazon **S3** which is connected via **Attunity** an ETL tool.
* POC involved in loading data from **LINUX** file system to AWS **S3** and **HDFS**.
* Worked on **AWS** to create, manage **EC2** instances and Hadoop Clusters.
* Developed **shell** **scripting** and **Python** **programs** to automate the data flow on day to day tasks.
* Created both internal and external tables in **Hive** and developed **Pig scripts** to preprocess the data for analysis.
* Implemented optimization and performance testing and tuning of **Hive and Pig**
* Designed appropriate **partitioning**/**bucketing** schema to allow faster data retrieval during analysis using **HIVE**.
* Experienced in performance tuning of **Spark** jobs for setting right Batch Interval time, correct level of Parallelism and memory tuning, changing the configuration properties and using broadcast variables.
* Involved in migrating **MapReduce** jobs into **Spark** jobs and used **Spark** **SQL** and **Data** **frames** **API** to load structured data into Spark clusters
* Involved in converting **HiveQL/SQL** queries into **Spark** transformations using Spark **RDDs**, **Python** and **Scala**.
* Implemented **Spark** using **Scala** and utilizing **Data** **frames** and **Spark** **SQL** **API** for faster processing of data.
* Implemented **Spark** **RDD** **transformations** to Map business analysis and apply **actions** on top of transformations.
* Experience in Job management using Fair scheduler and Developed job processing scripts using **Oozie** workflow to run multiple Spark Jobs in sequence for processing data
* Used **TOAD** to view the hive tables, views and the data.
* Involved in connecting **Pentaho 7.0** to target database to get data.
* Used ETL to transfer the data from the target database to **Pentaho** to send it to reporting tool MicroStrategy.
* Used **Zookeeper** for various types of centralized configurations, GIT for version control, Maven as a build tool for deploying the code.

**Environment:** Hadoop 2.7.3-amzn-0, Hive 1.0.0, Pig 0.14.0, Hue 3.7.1, Spark 1.6.2, Python 2.7.12, Oozie-Sandbox 4.2.0, Tez 0.8.4, ZooKeeper-Sandbox 3.4.8, Sqoop-Sandbox 1.4.6, Scala 2.11.8, HCatalog 1.0.0, Linux 4.4.19-29, Bedrock 4.2.1, Cassandra, Pentaho 7.0, MariaDB, TOAD 1.3.0 Beta, Eclipse, Hortonworks, Java 1.8, Maven 3.3.9, GIT.

**Client: BB & T Location: Winston- Salem, NC**

**Role: Hadoop Developer Duration: Aug 2015 – July 2016**

**Description:** BB & T is a multi- national financial organization with millions of customers. My role is to create MapReduce programs for analyzing data. Loading RDBMS of large datasets to big data by using sqoop. Migrating the mainframe to Hadoop platform was involved in the entire ETL flow.

**Responsibilities**

* Interaction with the business users from the client side to discuss and understand ongoing enhancements and changes at the upstream business data and performing data analysis.
* Experienced on loading and transforming of large sets of structured and semi structured data from **HDFS** through **Sqoop** and placed in **HDFS** for further processing.
* Worked on various file formats AVRO, ORC, Text, CSV, Parquet using Snappy compression.
* Involved in transforming data from legacy tables to **HDFS** and **HBase** tables using **Sqoop**.
* Proficient in writing **Flume** and **HiveQL** scripts to extract, transform and load the data into Database.
* Experience in writing **Map** **Reduce** programs and using Apache Hadoop API for analyzing the data.
* Analyzed the data by performing **HiveQL** queries and running **Pig** scripts to study data patterns.
* Created custom **UDFs** for **Hive** and **Pig**.
* Developed **PIG** scripts to arrange incoming data into suitable and structured data before piping it out for analysis.
* Created extensive SQL queries for data extraction to test the data in **Hive** on the top of S3
* Load and transform data into HDFS from Oracle/SQL server using **Talend** Open Studio.
* Used **Cassandra** to work on JSON documented data.
* Worked on **Cassandra** Querying language to load bulk of data and execute queries.
* Integrating bulk data into **Cassandra** file system using **MapReduce** programs.
* Involved in creating data-models for customer data using **Cassandra** Query Language.
* Used Subversion for maintaining the component and for release and version management and **JIRA** for defect tracking.
* Worked on Spark SQL and Data frames for faster execution of Hive queries using **Spark SqlContext**.
* Used Spark API over Cloudera Hadoop YARN to perform analytics on data in Hive.
* Developed Python scripts, **UDFs** using **Data** **frames** and RDD in Spark 1.6 for **Data** Aggregation, queries and writing data back into OLTP system through Sqoop.
* Used **Spark** **DataFrame** API over Cloudera platform to perform analytics on hive data.
* Implemented many **Impala** **scripts** and **Shell** **scripts** for data validation and data analytics
* Responsible for developing multiple **Kafka** Producers and Consumers from scratch as per the software requirement specifications.
* Involved in scheduling **Oozie** workflow engine to run multiple **HiveQL** and **Pig** jobs.
* Experienced in importing real time logs to **HDFS** using **Flume**.
* Followed agile software development with Scrum methodology.

**Environment:** CDH 4, 5 Hadoop 2.6, MapReduce, YARN, HDFS, Hive 1.1.0, Hue 3.9.0, Java, Impala Spark 1.5.0, Hbase, Kafka, Cassandra, PIG, Zookeeper, Talend 6.0,Ganglia, Nagios, Kerberos, Sqoop, Flume, Oracle, Oozie, JIRA, Eclipse, and CVS, Maven.

**Client: Health Care Service Corp. Location: Chicago, IL**

**Role: Hadoop Developer Duration: Sept 2014– June 2015**

**Description:** Health Care Service Corp. is an insurance company with millions of customers. My role is to use Hadoop and its related tools to analyze millions of customers and their records. Writing software to clean messy and large datasets. Performing web analytics for the site and storing data.

**Responsibilities**

* Collected the business requirement from the subject matter experts like data scientists and business partners.
* Load and transform large sets of structured, semi structured and unstructured data.
* Written multiple **MapReduce** programs in java for data extraction, transformation and aggregation from multiple file formats.
* Used different file formats like Text files, Sequence Files, Avro.
* Solved performance issues in **Hive** and **Pig** scripts with understanding of **Joins**, **Group** and aggregation and how does it translate to **MapReduce** jobs.
* Experience in configuring **Zookeeper** to coordinate the servers in clusters to maintain the data consistency and Monitored services.
* Extensively used **Spark**-**SQL** & **Hive** **Query** **Language** (HQL) for getting customer insights which will be used for critical decision making by business users.
* Used **Spark** **SQL** to perform transformations and actions on data residing in Hive.
* Designed batch processing jobs using Apache **Spark** to increase speed compared to that of **MR** jobs.
* Load data from various data sources into **HDFS** using **Kafka**.
* Developed a data pipeline using **Kafka** to store data into HDFS
* Shell scripts to dump the data from **MySQL** to **HDFS**.
* Involved in **NOSQL** databases like **HBase** in implementing and integration.
* Worked on streaming the analyzed data to the **HBase** using **SQOOP** for making it available for visualization and report generation by the BI team.
* Involved in scheduling **Oozie** workflow engine to run multiple **HiveQL**, **Sqoop** and **pig** jobs.
* Consumed the data from **Kafka** queue using **Storm**

**Environment:** HDP 2.4 Hadoop Cluster, MapReduce, HDFS, Hive, Java, Hbase, Spark 1.2.0, AVRO, Kafka, PIG, Zookeeper, Sqoop, Flume, Kafka, Storm, Oozie, MySQL, Ambari, Eclipse, and GIT.

**Client: Charter Communications Location: St. Louis, MO**

**Role: Hadoop Developer Duration: August 2013 – July 2014**

**Responsibilities:**

* Involved in full life-cycle of the project from Design, Analysis, logical and physical architecture modeling, development, Implementation, testing.
* Analyzing business requirements and translating requirements into functional and technical design specifications.
* Used **Oozie** to automate/schedule business workflows which invoke **Sqoop**, **MapReduce** and **Pig** jobs as per the requirements.
* Wrote Scripts to generate **Map** **Reduce** jobs on the data in **HDFS**.
* Responsible for designing and managing the **Sqoop** jobs to imports data from Data warehouse platform to **HDFS**.
* Created **HBase** tables to load large sets of structured, semi-structured and unstructured data coming from **UNIX**, **NoSQL** and a variety of portfolios.
* Developed the **Pig UDF's** to pre-process the data for analysis.
* Analyzed the data by performing **Hive** queries and running **Pig** scripts to know user behavior.
* Optimized **Map**/**Reduce** jobs to use **HDFS** efficiently by using various compression mechanisms.
* Developed **Hive** queries for data sampling and analysis to the analysts.
* Participated in end-to-end agile SDLC.
* Interfaced with business analysts, architects, developers, testers and clients in completing the project.

**Environment:** CDH 4, MapReduce, HDFS, Hive, Java, Flat files, Hbase, PIG, HCatolog, Zookeeper, Sqoop, Flume, Oozie, Oracle, Eclipse and SVN.

**Client: Progressive Inc. Location: Columbus, OH**

**Role: Java Developer Duration: Aug 2011 – July 2013**

**Description:** Progressive Inc. is ranked one of the best insurance companies. My role is to maintain all the user information, application provisioning and de-provisioning access request for an individual user. Coding the application in a way to make it user- friendly.

**Responsibilities:**

* Analyzing the design specs, business requirements and implemented the design using UML.
* Followed Waterfall development life cycle strategy for developing of the application.
* Generating Java classes from the respective APIs so that they can be incorporated in the overall application.
* Writing entities in Java along with named queries to interact with database.
* Responsible for client side UI Validation and implementing Business logic based on user selection using JQuery, JQuery UI and Angular JS.
* Used JBoss application server to deploy and JDBC, Hibernate for accessing Oracle database.
* Developed the build script using Maven and server side code using Spring Dependency Injection.
* Navigation through various screens of application is implemented using Spring Web Flow.
* Used Hibernate to access the database using entity POJO classes.
* Created shell scripts and PL/SQL scripts that were executed daily to refresh data feeds from multiple systems.
* Involved in developing the REST based web services for various data formats including JSON.
* Used PL/SQL stored procedures, triggers, cursors extensively to achieve effective results while dealing with huge transaction.
* Extensive experience in working with Struts and Spring MVC (Model View Controller) architecture for developing applications using various Java/J2EE technologies like Servlets, JSP, JDBC, JSTL.
* Created conditional logics in pages using JSF tags and JSTL.

**Environment:** Java/J2EE, JSP, Spring MVC, Spring Web Flow, JQuery, EJB, JBoss, Maven, Hibernate 3.0, SOAP UI, HTML, XML, PL/SQL, JDBC, GIT, Rest, Oracle, AGILE Methodology, NetBeans, dependency injection.

**Client: App Soft Solutions Location: Hyderabad, India**

**Role: Java Developer Duration: Jun 2009 – Jun 2011**

**Responsibilities:**

* Worked in Spiral development environment.
* Implement JUnit test cases for Struts/Spring components and built SOAP Web services.
* Developed server side presentation layer using Struts MVC Framework.
* Developed ActionForm classes, Form beans, Action Classes using Struts.
* Participated in Server upgrades, code migrations, and also worked on important enhancements in the business requirements.
* Performed Inheritance based OR mappings in tables to simplify the data in Hibernate.
* Used Hibernate for connecting to database with Oracle as its back- end.
* Interact with offshore team to coordinate and guide for the development happening in offshore
* Used Oracle database for SQL, PL/SQL scripts, stored procedures, functions, triggers, Oracle forms/Reports.
* Performance Tuning and Optimization with Java Performance Analysis Tool.
* Made use of CVS for checkout and check in operations.

**Environment:** Java, J2EE, Struts, WebSphere, Hibernate, JSP, JavaScript, SOAP web services, Oracle 10g, CVS, Eclipse.