**Sneha B**

**Hadoop-Spark Developer**

**Email:** [**sneta999@gmail.com**](mailto:sneta999@gmail.com)

**Ph #: (929)-399-9396**

**Professional Summary:**

* Overall **8** years of extensive IT experience in all phases of Software Development Life Cycle (SDLC) with skills in data analysis, design, development, testing and deployment of software systems.
* **4+** years of strong working experience with **Big Data** and Apache Hadoop ecosystem components like **Map-Reduce, HDFS, Sqoop, Flume, Spark, Spark Streaming Pig, Hive, HBase, Oozie, Kafka, and Zookeeper.**
* Worked with different flavors of Hadoop distributions, which includes **Cloudera (CDH4&5 Distributions) and Hortonworks and AWS.**
* Excellent understanding/ knowledge of Hadoop architecture and various components such as **HDFS, Job Tracker, Task Tracker, Name Node, Data Node, Resource Manager** and **Map Reduce**.
* Experienced in installation, configuration, support and managing of Big Data and underlying infrastructure **of Hadoop Cluster**
* Worked on Cloud computing infrastructure such as **EC2 Cloud Computing with AWS.**
* Experience in importing and exporting data using **Sqoop** from HDFS to Relational database systems and Vice-versa.
* Involved in importing Streaming data using **Flume** to HDFS and good experience in analyzing and cleansing raw data using **HiveQL, Pig Latin**.
* Experience in **Partitioning, Bucketing, Join Optimizations** and **Query Optimizations** in **Hive** and automating the Hive Queries with the Dynamic Partitioning.
* Good understanding of **NoSQL** databases and hands on experience in writing applications on **NoSQL databases like HBase** and worked on **HBase** to load and retrieve data for real time processing using Rest API.
* Working experience and Good understanding of **NoSQL** databases like Cassandra and Mongo DB
* Hands On experience on **SPARK and SPARK Sql, Spark Streaming, SCALA.** Creating the Data Frames handle in **SPARK** with **Scala**
* Developed **Scala** scripts, **UDF's** using both Data frames/SQL and **RDD’s in Spark** for Data Streaming, Aggregation and Testing Purposes
* Experience in installation, configuration, Management, supporting and monitoring Hadoop cluster using various distributions such as **Apache SPARK**, Cloudera and AWS Service console.
* Implemented POC's to migrate map reduce programs into **Spark** transformations using spark and Scala.
* Worked with both **Scala** and Java, Created frameworks for processing data pipelines through **Spark.**
* Implemented batch-processing solution to certain unstructured and large volume of data by using Hadoop Map Reduce framework.
* Experience in optimization of **MapReduce** algorithm using Combiners and Partitioners to deliver efficient results.
* Extending **HIVE** and **PIG** core functionality by using custom User Defined Function’s (UDF), User Defined Table-Generating Functions (UDTF) and User Defined Aggregating Functions (UDAF).
* Assisted in Cluster maintenance, Cluster Monitoring and Troubleshooting, Managing and Reviewing data backups and log files.
* Configured **Zookeeper** to coordinate the servers in clusters to maintain the data consistency.
* Proficient in **ETL tools** for Designing Data warehouse, Business Intelligence, Analytics, Data Mining, Data Mapping, Data conversion, Data Migrations and Transformations from Source to Target Systems
* Used **Oozie, ControlM** and **Autosys** workflow engine for managing and scheduling Hadoop Jobs.
* Experienced in using **Kafka** as a distributed publisher-subscriber messaging system.
* Continuous integration and automated deployment and management using **Jenkins and Udeploy.**
* Diverse experience in working with variety of Database like **Oracle, MySQL, IBM DB2** and **Netezza.**
* Experienced in using IDEs and Tools like **Eclipse, NetBeans, GitHub, Jenkins, Maven and IntelliJ.**
* Extensive Experience in creating **Tableau Dashboards** using Stack Bars, Bar Graphs, and geographical maps.
* Good knowledge on various scripting languages like **Linux/Unix shell scripting and Python.**
* Experience in developing Client-Side Web applications using **Core Java** and **J2EE** technologies such as **HTML, JSP, jQuery, JDBC, Hibernate** and Custom Tags while implementing the client-side validations using **JavaScript** and Server-side validations using **Struts** and **Spring** Validations Framework
* Strong team player, ability to work independently and in a team as well, ability to adapt to a rapidly changing environment, commitment towards learning, Possess excellent communication, project management, documentation, interpersonal skills
* Experienced in developing web applications in various domains Telecommunications, Retail, Insurance and Health Care

**Technical Skills:**

|  |  |
| --- | --- |
| **Big Data** | Hadoop, HDFS, MapReduce, Hive, Pig, Sqoop, Flume, Oozie, Avro, Spark, Spark Streaming, Storm, Kafka, YARN, Zookeeper, HBase, Impala, Cassandra. |
| **Hadoop Distributions** | Cloudera, Hortonworks and MapR |
| **Databases** | SQL Server, MySQL, Oracle, Netezza. |
| **Languages** | Java, C, HTML Scala, SQL, PL/SQL, Unix Shell Script, Python. |
| **JEE Technologies** | JSP, JDBC |
| **FRAME WORKS** | MVC, Struts, Spring, Hibernate. |
| **Build Tools** | SBT, Maven and Gradle |
| **IDE’s** | Eclipse, Intellij |
| **CI Tools** | Hudson/Jenkins, NetBeans |
| **Cloud Solutions** | AWS EMR, S3 |
| **Version Control / Configurations** | GIT, SVN, CVS |
| **Defects Triage** | JIRA and Bugzilla |
| **Operating Systems** | Windows, UNIX, LINUX, Ubuntu, Cent OS |
| **Packages** | MS Office Suite, MS Visio, MS Project Professional |
| **File Formats** | Avro, JSON, Parquet, Sequence, XML, CSV |
| **Reporting Tools** | Tableau |

**Work Experience:**

**Change healthcare – San Mateo, CA** Sep’18- Present

**AWS Big Data Developer**

* Developed Spark applications by using Scala
* Deployment of Spark steaming applications with optimized no. of executors, write ahead logs & check point configurations.
* Developed Spark SQL queries for the analysts via Zeppelin.
* Strong experience in working with ELASTIC MAPREDUCE(EMR)and setting up environments on Amazon AWS EC2 instances.
* Implemented Spark using Scala and Spark SQL for faster testing and processing of data.
* Experience in extracting appropriate features from data sets in order to handle bad, null, partial records using Spark SQL.
* Collected data using Spark Streaming from AWS S3 bucket in near-real-time and performs necessary Transformations and Aggregations to build the data model and persists the data in HDFS
* Imported data from AWS S3 into Spark RDD, Performed transformations and actions on RDD's.
* Experienced in working with spark eco system using Spark SQL and Scala queries on different formats like Text file, CSV file. Extensively work with Parquet file formats.
* Implemented a mechanism for triggering the spark applications on EMR on file arrival from the client.
* Implemented Sqoop for large data transfer from RDBMS to S3.
* Administering the cluster and tuning the memory based on the RDD usage.
* Developed a process to identify the personal health information(PHI) and mask them with the predefined reference keys.
* Implemented an application to do the address normalization for all the clients datasets
* Experienced in working with Amazon Web Services (AWS) using EC2 for computing and S3 as storage mechanism.
* Good Knowledge on continuous Integration of application using Jenkins, Rundeck and CICD Pipelines.
* Implemented a mechanism to start a EMR based on the input file size using Lambda function and s3 events.
* Worked with the team on many design decisions.
* Responsible in creating mappings and workflows to extract and load datafrom relational databases, flat file sources and legacy systems using Azure.
* Experienced in using Zeppelin tool for data analytics and data visualization.

**Environment**: Amazon Web Services, Spark3 HDFS, Hive, Pig, Python Impala, Scala, Java, SQL, CDH 4.6-5.3, Sqoop, Flume, Oozie, Java, IntelliJ IDEA, Jenkins, Artifactory, Rundeck, Parquet.

**Global atlantic financial group – Boston, MA** Aug’17-Aug’18

**Hadoop Developer**

**Responsibilities:**

* Coordinated with business customers to gather business requirements. And, interact with other technical peers to derive Technical requirements and delivered the BRD and TDD documents
* Worked on analyzing Hadoop 2.6.2 cluster and different Big Data analytic tools including Hive 2.0 HBase 1.1.2 database and SQOOP 1.4.6
* Implemented Spark 2.0 using Scala 2.11.8 and Spark SQL for faster processing of data.
* Implemented algorithms for real time analysis in Spark.
* Used Spark for interactive queries, processing of streaming data and integration
* Installed and Configured Apache Hadoop clusters for application development and Hadoop tools like Hive, Pig, Oozie, Zookeeper, HBase, Flume and Sqoop.
* Implemented multiple Map Reduce Jobs in java for data cleaning and pre-processing.
* Worked in a team with 30 node cluster and increased cluster by adding Nodes, the configuration for additional data nodes were done by Commissioning process in Hadoop.
* Importing the data from the MySQL and Oracle into the HDFS using Sqoop.
* Loading files to HDFS and writing Hive queries to process required data.
* Loading data to Hive tables and writing queries to process.
* Involved in loading data from LINUX file system to HDFS.
* Load and transform large sets of structured, semi structured and unstructured data.
* Experience in managing and reviewing Hadoop log files.
* Worked on Hive for exposing data for further analysis and for generating transforming files from different analytical formats to text files.
* Used Reporting tools like Tableau to connect with Hive for generating daily reports of data.
* Developed Hive queries to process the data and generate the data cubes for visualizing.
* Used Hive QL to analyze the partitioned and bucketed data, Executed Hive queries on Parquet tables stored in Hive to perform data analysis to meet the business specification logic
* Importing and exporting data into HDFS and Hive using Sqoop.
* Involved in creating Hive tables, loading with data and writing hive queries that will run internally in MapReduce way.
* Worked on configuring multiple MapReduce Pipelines, for the new Hadoop Cluster.
* Performance tuned and optimized Hadoop clusters to achieve high performance.
* Written Hive queries for data analysis to meet the business requirements.
* Monitored System health and logs and respond accordingly to any warning or failure conditions
* Responsible to manage the test data coming from different sources.
* Involved in scheduling Oozie workflow engine to run multiple Hive jobs
* Weekly meetings with technical collaborators and active participation in code review sessions with senior and junior developers.
* Created and maintained Technical documentation for launching Hadoop Clusters and for executing Hive queries and Pig Scripts
* Implemented schedulers on the Job tracker to share the resources of the cluster for the MapReduce jobs given by the users.
* Extensive hands on experience in Hadoop file system commands for file handling operations.

**Environment:** Hadoop, Map Reduce, HDFS, Hive, Java, Hortonworks,Oozie, Linux, XML, Java 6, Eclipse, Oracle 10g, PL/SQL, YARN, Spark, Pig, Sqoop, , DB2, java, XML, Pig, PVCS, UNIX, DB2, XML, PVCS, UNIX, HCatalog.

**Kohls – Menomonee, WI** Sep’16- July’17

Hadoop-Spark Developer

**Responsibilities:**

* Developed Pig Latin scripts to extract the data from the web server output files to load into HDFS.
* Used IMPALA for querying the HDFS data.
* Developed and implemented two Service Endpoints (end to end) in Java using Play framework, Akka server Hazelcast.
* Services like EC2 and S3 for small data sets.
* Ingested data from RDBMS and performed data transformations, and then export to Cassandra.
* Developed the Pig UDF'S to pre-process the data for analysis.
* Used Apache Kafka to get the data from Kafka producer which in turn pushes data to broker.
* Written robust/reusable Hive Scripts and UDF's in Hive using Java.
* Experience with Test Driven Development (TDD) and acceptance- test using Behave.
* Implemented partitioning, bucketing in Hive for better organization of the data.
* Designed and built unit tests and executed operational queries on HBase.
* Built Apache Avro schemas for publishing messages to topics and enabled relevant serializing formats for message publishing and consumption.
* Connected Tableau from client end with AWS IP addresses and view the end results.
* Used Spark API over Hadoop YARN to perform analytics on data in Hive.
* Implemented a script to transmit information from Oracle to HBase using Sqoop.
* Installed Hadoop, Map Reduce, and HDFS and developed multiple MapReduce jobs in PIG and Hive for data cleaning and pre-processing.
* Writing MapReduce programs to convert text files into AVRO and loading into Hive (Hadoop) tables
* Assisted in upgrading, configuration and maintenance of various Hadoop infrastructures like Pig, Hive, and HBase.
* Performed real time analysis on the incoming data using Pig, Hive and Map Reduce.
* Performed transformations, cleaning and filtering on imported data using Hive, Map Reduce and loaded final data into HDFS.
* Loaded data into HBase using Bulk Load and Non-bulk load.
* Developed Spark scripts by using Scala shell commands as per the requirement.
* Used Spark API over Cloudera Hadoop YARN to perform analytics on data in Hive.
* Developed Scala scripts, UDFFs using both Data frames/SQL and RDD/MapReduce in Spark for Data Aggregation, queries and writing data back into OLTP system directly or through Sqoop.
* Developed Spark code using Scala and Spark-SQL/Streaming for faster testing and processing of data.
* Imported the data from different sources like HDFS/HBase into Spark RDD.
* Developed a data pipeline using Kafka to store data into HDFS.
* Connected Tableau from client end with AWS IP addresses and view the end results.
* Used Spark API over Hadoop YARN to perform analytics on data in Hive.
* Implemented a script to transmit information from Oracle to HBase using Sqoop.
* Worked on migrating MapReduce programs into Spark transformations using Spark
* Loaded the data into Spark RDD and do in memory data Computation to generate the Output response
* Created and maintained Technical documentation for launching HADOOP Clusters and for executing Hive queries and Pig Scripts
* Involved in the identifying, analyzing defects, questionable function error and inconsistencies in output.

**Environment:** Hadoop, HDFS, MapReduce, YARN, Spark, Pig, Hive, Sqoop, Flume, Kafka, HBase, Oozie, Scala, Java, SQL scripting, Linux shell scripting, Eclipse, AWS, HBase, AVRO, Oracle, Unix, Tableau.

**Ericsson – Atlanta, GA** Nov’14– Aug’16

Sr. Hadoop Developer

**Responsibilities:**

* Worked extensively on Hadoop Components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node, YARN and Map Reduce programming.
* Developed Map-Reduce programs to clean and aggregate the data.
* Responsible for building scalable distributed data solutions using Hadoop and Spark
* Worked hands on with ETL process using Java
* Implemented Hive Ad-hoc queries to handle Member data from different data sources such as Epic and Centricity.
* Implemented Hive UDF's and did performance tuning for better results.
* Analyzed the data by performing Hive queries and running Pig Scripts.
* Involved in loading data from UNIX file system to HDFS
* Implemented optimized map joins to get data from different sources to perform cleaning operations before applying the algorithms.
* Experience in using Sqoop to import and export the data from Netezza and Oracle DB into HDFS and HIVE.
* Implemented POC to introduce Spark Transformations.
* Worked with NoSQL database HBase, MongoDB and Cassandra to create tables and store data
* Handled importing data from various data sources, performed transformations using Hive and Map Reduce, streamed using Flume and loaded data into HDFS
* Worked in transforming data from map reduce into HBase as bulk operations.
* Implemented CRUD operations on HBase data using thrift API to get real time insights.
* Installed Ozzie workflow engine to run multiple MapReduce, Hive, Impala, Zookeeper and Pig jobs which run independently with time and data availability
* Developed workflow in Oozie to manage and schedule jobs on Hadoop cluster for generating reports on nightly, weekly and monthly basis.
* Used Zookeeper to manage Hadoop clusters and Oozie to schedule job workflows.
* Implemented test scripts to support test driven development and continuous integration.
* Involved in data ingestion into HDFS using Apache Sqoop from a variety of sources using connectors like JDBC and import parameters
* Coordination with Hadoop Admin's during deployment to production
* Developed Pig Latin Scripts to extract data from log files and store them to HDFS. Created User Defined Functions (UDFs) to pre- process data for analysis.
* Developing Scripts and Batch Job to schedule various Hadoop Program.
* Continuously monitoring and managing the Hadoop cluster through Cloudera Manager
* Participated in design and implementation discussion for the developing Cloudera 5 Hadoop eco system.
* Used JIRA and Confluence to update tasks and maintain documentation.
* Worked in Agile development environment in sprint cycles of two weeks by dividing and organizing tasks. Participated in daily scrum and other design related meetings.
* Created final reports of analyzed data using Apache Hue and Hive Browser and generated graphs for studying by the data analytics team.
* Used SQOOP to export the analyzed data to relational database for analysis by data analytics team.

**Environment:** Hadoop, Cloudera Hadoop, Map Reduce, Hive, Pig, Sqoop, Flume, HBase, Java, JSON, Spark, HDFS, YARN, Oozie Scheduler, Zookeeper, Mahout, Linux, UNIX, ETL, My SQL.

**Blue Cross Blue Shield**­**– Birmingham, AL** Sep’13– Oct’14

Hadoop Developer

**Responsibilities:**

* Design, Installation and Configuration of Flume, Hive, Pig and Oozie on the Hadoop Cluster.
* Designed workflow by scheduling Hive processes for Log file data, which is streamed into HDFS using Flume.
* Effectively used Sqoop to transfer data between databases and HDFS.
* Import data from open data sources into Amazon S3 and other private clusters.
* Developed scripts to automate the creation of Sqoop Jobs for various workflows.
* Used Hive data warehouse tool to analyze the unified historic data in HDFS to identify issues and behavioralpatterns.
* Using HiveQL developed many queries and extracted the business required information.
* Developed scripts to automate the creation of hive tables and partitions.
* Developed MapReduce programs to extract and transform the data sets and results were exported backto RDBMS using Sqoop.
* Developed MR jobs for analyzing the data stored in the HDFS by performing map-side joins, reduce-sidejoins.
* Involved in implementing High Availability and automatic failover infrastructure to overcome single point offailure for Name node.
* Developed Pig Latin scripts to extract the data from the web server output files to load into HDFS.
* Designed and developed the framework to log information for auditing and failure recovery.
* Design & Develop ETL workflow using Oozie for business requirements, which includes automating theextraction of data from MySQL database into HDFS using Sqoop.

**Environment:** HDFS, MapReduce, Hive, Pig, Oozie, Sqoop, Flume

**Nationwide – Columbus, OH** Apr’12– Aug’13

ETL Data Stage Developer

**Responsibilities:**

* Designed jobs involving various cross reference lookups and joins, shared containers which can be used in multiple jobs.
* Sequencers are created at job level to include multiple jobs and a layer level sequence which include all job level sequences.
* Involved in the designing of marts and dimensional and fact tables.
* Extensively used Parallel Stages like Row Generator, Column Generator, Head, and Peek for development and de-bugging purposes.
* Worked hands on with ETL process using Python and Java
* Knowledge of configuration files for Parallel jobs.
* Migrated the jobs from 7.5 to 8.1 and developed new DataStage jobs using data stage/quality stage designer Imported and exported repositories across projects.
* Extensive experience in working with DataStage Designer for developing jobs and DataStage Director to view the log file for execution errors.
* Created DataStage Parallel Jobs to Fact and Dimension Tables.
* Wrote Shell Scripts to run data stage jobs, PL/SQL blocks.
* Wrote SQL queries for checking the data from Source system as well as Staging.
* Used Parallel Extender for splitting the data into subsets, utilized Lookup, Sort, Merge and other stages to achieve job performance.
* Extensive experience in working with DataStage tools like DataStage Designer and DataStage Director for developing the jobs and view the log for errors.

**Environment:**IBM WebSphere Data stage 8.1, Datastage 7.5.2, Python, UNIX Shell Scripting (Korn /KSH), SQL, Oracle 9i/10g, UNIX and Windows XP.

**Stryker – Kalamazoo, MI** Jul’10– Mar’12

Jr. Java Developer

**Responsibilities:**

* The application was developed in J2EE using an MVC based architecture.
* Implemented MVC design using Struts1.3 frameworks, JSP custom tag Libraries and various in-house custom tag libraries for the presentation layer.
* Created tile definitions, Struts-config files, validation files and resource bundles for all modules using Struts framework.
* Wrote prepared statements and called stored Procedures using callable statements in MySQL.
* Executed SQL queries to perform crud operations on customer records.
* Gathered requirements and then developed complex workflows which involved Templates. Open Deploy.
* Used Eclipse 6.0 as IDE for application development Configured Struts framework to implement MVC design patterns.
* Validated all forms using Struts validation framework and implemented Tiles framework in the presentation layer.
* Designed and developed GUI using JSP, HTML, DHTML and CSS. Worked with JMS for messaging interface.
* Used Hibernate for handling database transactions and persisting objects deployed the entire project on WebLogic application server.
* Part of the team involved in the design and coding of the Data capture templates, presentation & component templates.
* Developed and configured templates to capture and generate multi-lingual content. With this approach North US branch content is encoded in BIG5.
* Used Apache web sphere as the application server for deployment.
* Used Web services for transmission of large blocks of XML data over HTTP.

**Environment:**Java/J2EE, Oracle 10g, SQL, PL/SQL, JSP, EJB, Struts, Hibernate, WebLogic 8.0, HTML, AJAX, Java Script, JDBC, XML, JMS, XSLT, UML, JUnit, Log4j, Eclipse 6.0.