**Abhinay**

**Hadoop/Spark Developer**

**E-mail: abhinayb.hadoop@gmail.com**

**Ph: +1 408-780-8394**

**Professional Summary:**

* Overall 8+ years of IT experience in analysis, design, development and implementation of business applications with thorough knowledge in Java, J2EE, Big Data, Hadoop Eco System and RDBMS related technologies with domain exposure in Retail, Healthcare, Banking, E-commerce websites, Insurance, Logistics and Financial (Mortgage) systems.
* Expertise with the tools in **Hadoop** Ecosystem including **Pig, Hive, HDFS, MapReduce, Sqoop, Storm, Spark, Kafka, Yarn, Oozie, and Zookeeper.**
* Excellent knowledge on **Hadoop** Architecture such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node and MapReduce programming paradigm.
* Strong experience on **Hadoop** distributions like Cloudera, MapR and HortonWorks.
* Good Knowledge on Hadoop Cluster architecture and monitoring the cluster.
* Hands on experience in installing, configuring Cloudera Apache Hadoop ecosystem components like Flume, Hbase, ZooKeeper, Oozie, Hive, Sqoop and Pig.
* Highly capable of processing large sets of Structured, Semi-structured and Unstructured datasets supporting Big Data applications.
* Good experience in Hive partitioning, bucketing and perform different types of joins on Hive tables and implementing Hive Sere like **JSON** and **ORC**.
* Worked on different file formats (ORCFILE, TEXTFILE) and different Compression Codecs **(GZIP, SNAPPY, LZO)**.
* Proficiency in Hadoop data formats like **AVRO** & **Parquet.**
* Have good knowledge on NoSQL databases like **HBase**, **Cassandra** and **MongoDB**.
* Proficient in implementing **HBase**.
* Used **Zookeeper** to provide coordination services to the cluster.
* Experience in **Oozie** and workflow scheduler to manage Hadoop jobs by **Direct Acyclic Graph (DAG)** of actions with control flows.
* Experience using **Sqoop** to import data into HDFS from RDBMS and vice-versa.
* Extensive Experience on importing and exporting data using stream processing platforms like **Flume** and **Kafka**.
* Implemented indexing for logs from **Oozie** to **Elastic Search**.
* Analysis on integrating **Kibana** with Elastic Search.
* Implemented POC to migrate Map Reduce jobs into **Spark RDD** transformations using **Scala**.
* Experience in creating **Spark Contexts, Spark SQL Contexts, and Spark Streaming** Context to process huge sets of data.
* Exploring with the Spark for improving the performance and optimization of the existing algorithms in Hadoop using Spark **Context, Spark-SQL, Data Frame, Pair RDD's, Spark YARN**.
* Experienced in **Spark Core, Spark RDD, Pair RDD**, Spark Deployment Architectures.
* Extensive experience using **MAVEN** and **ANT** as a Build tool for the building of deployable artifacts from source code.
* Worked with Big Data distributions like **Cloudera (CDH 3, 4 and 5)** with Cloudera Manager.
* Knowledge on Cloud technologies like **AWS Cloud** and **Amazon Elastic Map Reduce (EMR)**.
* Experience working with **Amazon** **Web Services S3, EC2, Redshift**.
* Proficient in using **OOPs** **Concepts** (Polymorphism, Inheritance, Encapsulation) etc.
* Data Warehouse experience using **Informatica Power Center** as ETL tool.
* Experience in using data integration tool like **Talend.**
* Experience in performance tuning and fine tuning of a hadoop cluster.
* Experience with **cloudera Manager** and **Ambari** on **cloudera** and **Hortonworks** clusters.
* Experience in **Unix/Linux** shell scripting.
* Worked with Admin teams to setup cluster securities like **Kerberos, Ranger, and Knox**.
* Written unit test cases using **JUnit** and **MRUnit** for Map Reduce jobs.
* Proficiency in Hadoop data formats like **AVRO & Parquet.**
* Experience with code development frameworks – **GitHub, Jenkins.**
* Deployed and monitored scalable infrastructure on cloud environment **Amazon web services (AWS)**.
* Knowledge in **Machine Learning** (Linear Regression, logistic regression, Clustering, Classification, and Decision Tree, support vector machines and dimensionality reduction).
* Comprehensive knowledge of **Software Development Life Cycle (SDLC)**, having thorough understanding of various phases like Requirements Analysis, Design, Development and Testing.
* Involved in the Software Life Cycle phases like **Agile and Waterfall** estimating the timelines for projects.
* Ability to quickly master new concepts and applications.

**Technical skills:**

* **Big Data Technologies:** Hadoop (HDFS & MapReduce), PIG, HIVE, HBASE, ZOOKEEPER, Sqoop, Apache Storm, Flume, Kafka, Spark, Spark Streaming, Mlib, Spark SQL and Data Frames, Graph X, Scala, Solr, Lucene, Elastic Search and AWS
* **Programming & Scripting Languages:** Java, C, SQL, R, Python, Impala, Scala, C++
* **J2EE Technologies:** JSP, SERVLETS, EJB, Angular JS
* **Web Technologies:** HTML, JavaScript
* **Frameworks:** Spring 3.5 – Spring MVC, Spring ORM, Spring Security, Spring ROO, Hibernate, Struts.
* **Application Servers:** IBM Web Sphere, JBoss WebLogic
* **Web Servers:** Apache Tomcat
* **Databases**: MS SQL Server & SQL Server Integration Services (SSIS), My SQL, MongoDB, Cassandra, Oracle DB, Teradata
* **Designing Tools:** UML, Visio
* **IDEs**: Eclipse, Net Beans
* **Operating System:** Unix, Windows, Linux, Cent OS
* **Others:** Putty, WinScp, DataLake, Talend, Tableau, GitHub, SVN, CVS.

**Hadoop/spark Developer**

**Project:  Hilton World Wide, TN Dec 2015 – Current**

**Description**:Hilton is one of the largest and fastest growing hospitality companies in the world. We are

Designing an analytical platform with a distributed cloud ingestor for data intake weblogs, reviews, complaint logs on cloudera CDH cluster with spark as the distributing computing engine.

**Responsibilities:**

* Working on Big Data infrastructure for batch processing as well as real-time processing. Responsible for building scalable distributed data solutions using **Hadoop**.
* Involved in creating **Hive Tables**, loading with data and writing Hive queries which will invoke and run Map Reduce jobs in the backend.
* Designed and implemented **Incremental Imports** into Hive tables.
* Very good understanding of Partitions, Bucketing concepts in Hive and designed both **Managed and External tables** in Hive to optimize performance.
* Experience in importing and exporting tera bytes of data using **Sqoop** from HDFS to Relational Database Systems and vice-versa.
* Moved Relational Database data using Sqoop into Hive **Dynamic partition** tables using **staging tables**.
* Written Hive jobs to parse the logs and structure them in tabular format to facilitate effective querying on the log data.
* Experience in manipulating/analyzing large datasets and finding patterns and insights within structured and unstructured data.
* Involved in collecting, aggregating and moving data from servers to HDFS using **Apache Flume**.
* Experienced in managing and reviewing the Hadoop **log files**.
* Involved in developing Pig Scripts for change data capture and **delta record** processing between newly arrived data and already existing data in HDFS.
* Migrated **ETL** jobs to Pig scripts do transformations, even joins and some pre-aggregations before storing the data onto HDFS.
* Implemented the workflows using Apache **Oozie** framework to automate tasks.
* Used Zookeeper to co-ordinate cluster services.
* Worked on different file formats like Sequence files, **XML** **files** and **Map files** using Map Reduce Programs.
* Used **Impala** where ever possible to achieve faster results compared to Hive during data Analysis.
* Implemented data ingestion and handling clusters in real time processing using **Kafka**.
* Worked on writing transformer/mapping Map-Reduce pipelines using **Java**.
* Developed scripts and automated data management from end to end and sync up between all the clusters.
* Transform the logs data into data model using apache pig and written **UDF’s** functions to format the logs data.
* Experience in designing and developing applications in **Spark using Scala** to compare the performance of Spark with Hive and SQL/Oracle.
* Involved in converting Hive/SQL queries into Spark transformations using **Spark RDDs, Python** and **Scala**.
* Experience in both **SQLContext** and **SparkSession**
* Experienced in working with sparkeco system using Spark SQL and Scalaqueries on different formats like **Text file, CSV file**.
* Used **Spark API** over Hadoop YARN to perform analytics on data in Hive.
* Developed and Configured **Kafka brokers** to pipeline server logs data into spark streaming.
* Developed Spark scripts by using scala shell commands as per the requirement.
* Developed spark code and **spark-SQL/streaming** for faster testing and processing of data.
* Experience in implementing **Log Error Alarmer** in Spark
* Exported the analyzed data to relational databases using sqoop for visualization and to generate reports.
* Experienced in Monitoring Cluster using Cloudera manager.
* Collaborated with the infrastructure, network, database, application and BI teams to ensure data quality and availability.

Environment: Hadoop, HDFS, Pig, Apache Hive, Sqoop, Flume, Kafka, Apache Spark, Storm, Solr, Shell Scripting, HBase, Scala, Python, Kerberos, Agile, Zoo Keeper, Maven, AWS, MySQL.

**Hadoop Developer**

**Client: Macy’s Inc., OH July 2014 – Nov 2015**

**Description**: As the fastest growing part of Macy’s Inc. business, macys.com is achieving record sales. We build data products and data systems and work on collecting, storing, processing and analyzing of huge sets of sales data. And managing and organizing the core domains such as Product, Price, Sales, Fulfillment, Inventory, Customer, Offers and Associates.

**Responsibilities:**

* Developing and running Map-Reduce jobs on YARN and Hadoop clusters to produce daily and monthly reports as per user's need.
* Worked on analyzing, writing Hadoop MapReduce jobs using Java API, Pig and Hive.
* Imported data using Sqoop to load data from **MySQL** to HDFS on regular basis.
* Implemented data access jobs through Pig, Hive, **Tez**, **Solr, Accumulo**, **Hbase**, and **Storm**.
* Worked on Developing custom MapReduce programs and User Defined Functions (UDFs) in Hive to transform the large volumes of data with respect to business requirement.
* 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) for Hive and Pig using python.
* Developed HIVE scripts for analyst requirements for analysis.
* Involved in loading and transforming large sets of Structured, Semi-Structured and Unstructured data and analyzed them by running Hive queries and Pig scripts.
* Analyzed the data by performing Hive queries (HiveQL), **Impala** and running Pig Latin scripts to study customer behavior.
* Processed HDFS data and created external tables using Hive and developed scripts to ingest and repair tables that can be reused across the project.
* Implemented data ingestion and handling clusters in real time processing using **Kafka**.
* Filter the dataset with PIG UDF, PIG scripts in HDFS and **Storm/Bolt** in Apache Storm.
* Involved in writing Pig Scripts for Cleansing the data and implemented Hive tables for the processed data in tabular format.
* Involved in collecting, aggregating and moving data from servers to HDFS using Apache Flume.
* Written Map Reduce code to process and parsing the data from various sources and storing parsed data into HBase and Hive using **HBase-Hive Integration**.
* Experienced with performing **CURD** operations in HBase.
* Used JSON, Parquet and Avro **SerDe's** for serialization and de-serialization.
* Setting up **CRON** job to delete hadoop logs/local old job files/cluster temp files.
* Using HBase to store majority of data which needs to be divided based on region.
* Used **Maven** extensively for building jar files of MapReduce programs and deployed to cluster.
* Used **Zookeeper** to provide coordination services to the cluster. Experienced in managing and reviewing Hadoop log files.
* Hands-on expertise with various architectures in **MongoDB** & **CassandraDB.**
* Very good experience in monitoring and managing the Hadoop cluster using **Hortonworks**.
* Used Amazon **Redshift** for data warehouse and to generate backend reports.
* Exported the business required information to RDBMS using Sqoop to make the data available for BI team to generate reports based on data.

Environment: Hadoop, HDFS, Pig, Hive, HBase, Map Reduce, Sqoop, Flume, Impala, Oozie, Zookeeper, LINUX, BigData, Java, Eclipse, Maven, SQL, Ambari, NoSql.

**Hadoop Developer**

**Client: Citibank, NJ Dec 2013 – June 2014**

**Description:** Citibank is an American international banking and financial services holding company with "hub quarters" throughout the country. Citibank delineates three different business segments when reporting results: Community Banking, Wholesale Banking, and Wealth, Brokerage and Retirement. Gathered requirements for big data project using technologies like Hadoop stack (Hadoop, HDFS, HIVE, PIG, and SQOOP). Data transformed to HDFS architecture and analyzed as per the requirements.

**Responsibilities:**

* Involved in all phases of the Big Data Implementation including requirement analysis, design, development, building, testing, and deployment of **Hadoop** cluster in fully distributed mode Mapping the DB2 V9.7, V10.x Data Types to Hive Data Types and validations.
* Identifying the various data sources and understanding the data schema in source environment.
* Design, build and support pipelines of data ingestion, transformation, conversion and validation.
* Provided quick response to ad hoc internal and external client requests for data and experienced in creating ad hoc reports.
* Experience in ingesting data into Cassandra and consuming the ingested data from Cassandra to Hadoop.
* Implemented Partitioning, **Dynamic Partitions** and **bucketing** in HIVE for efficient data access.
* Created final tables in **Parquet** format. Use of **Impala** to create and manage Parquet tables.
* Enhanced HIVE queries performance using **TEZ** for Customer Attribution datasets.
* Implemented data Ingestion and handling clusters in real time processing using Apache Kafka.
* Involved in running **Ad-Hoc** query through **PIG** Latin language, Hive or Java MapReduce.
* Implemented Partitioning, Dynamic Partitions, Buckets in HIVE.
* Worked on NoSQL databases including **HBase** and **Cassandra**.
* Participated in development/implementation of **Cloudera impala** Hadoop environment.
* Installed Oozie workflow engine to run multiple MapReduce jobs.
* Developed NoSQL database by using **CRUD**, Indexing, Replication and Sharing in MongoDB.
* Developed the data model to manage the summarized data.

Environment: Hadoop, HDFS, Pig, Hive, MapReduce, Java, Flume, Talend, Oozie, Linux/Unix Shell scripting, Avro, Parquet, Cassandra, MongoDB, Python, Perl, Java, Git, Maven, Jenkins.

**Hadoop Developer**

**Client: Alten calsoft labs, CA Nov 2012 – Sep 2013**

**Description:** Alten Group is a leading provider of comprehensive health informatics solutions for more than 200+ hospitals in the US. The customer wanted AltenCalsoft Labs to identify a patient's chance of getting re-admitted upon discharge within 30 days. So, Clinicians can be prepared to provide better post-discharge care for patients who are likely to get re-admitted.

**Responsibilities:**

* Worked with Big Data team responsible for building Hadoop stack and different big data analytic tools, migration from RDBMS to Hadoop using Sqoop.
* Used Bash shell scripting to perform Hadoop operations.
* Designed the **sequence diagrams** to depict the data flow into Hadoop.
* Involved in importing and exporting data between HDFS and Relational Systems like Oracle, Mysql, DB2 and Teradata using Sqoop.
* As a POC, extensively worked with Oozie workflow engine to run multiple Hive Jobs.
* Working on Hive to analyze the data and to extract report.
* Involve in creating **Hive tables**, loading with data and writing **Hive queries** which will run internally in map reduce way.
* Developed Simple to complex MapReduce Jobs using Hive and Pig. Developed Shell and Pythonscripts to automate and provide Control flow to Pig scripts.
* Responsible for managing data from multiple sources.
* Managing and scheduling Jobs on a Hadoop cluster using Oozie.
* Developed Simple to complex Map/reduce Jobs using Java programming language that are implemented using Hive and Pig.
* Supported in setting up QA environment and updating configurations for implementing scripts with Pig, Hive and Sqoop

Environment: Hadoop, Hive, Pig, Sqoop, Map Reduce, Linux, HDFS, Java.

**Java Developer**

**Client: iTwine Technologies, India Jan 2011 – Aug 2012** **Description:** The main objective of the project is to develop a fully functional Ezee Mail System that enhances communications among the members of the organization in a reliable, cost-effective and secure way. Thus it helps the organization to improve the performance of its teams in the corporate sector.

**Responsibilities:**

* Designed and developed the application using agile methodology.
* Implementation of new module development, new change requirement, fixes the code. Defect fixing for defects identified in pre-production environments and production environment.
* Wrote technical design document with class, sequence, and activity diagrams in each use case.
* Created Wiki pages using Confluence Documentation.
* Developed various reusable helper and utility classes which were used across all modules of application.
* Used Hibernate for persistence framework, involved in creating DAO's and used Hibernate for ORM mapping.
* Written Java classes to test UI and Web services through JUnit.
* Performed functional and integration testing, extensively involved in release/deployment related critical activities. Responsible for designing Rich user Interface Applications using JSP, JSP Tag libraries, Spring Tag libraries, JavaScript, CSS, HTML.
* Used SVN for version control. Log4J was used to log both User Interface and Domain Level Messages.
* Involved in system wide enhancements supporting the entire system and fixing reported bugs
* Explored Spring MVC, Spring IOC, Spring AOP, and Hibernate in creating the POC.
* Done data manipulation on front end using JavaScript and JSON.

Environment: Java, J2EE, JSP, Spring, Hibernate, CSS, JavaScript, Oracle, JBoss, Maven, Eclipse, JUnit, Log4J, AJAX, Web services, JNDI, JMS, HTML, XML, XSD, XML Schema, SVN, Git.

**Java Developer**

**Client: Asm Technologies, Bangalore, India Mar 2009 – Oct 2010**

**Description:** E-Check Payment System involves the development of ‘software modules’ for the generation and verification of ‘Electronic Check.’ The Payee System is used to verify payer's signatures. The Payee Bank Server processes and verifies the ‘deposited endorsed E-Checks’ and ‘Clears and Settles’ the transaction.

**Responsibilities:**

* Designed Use Case Diagrams, Class Diagrams and Sequence Diagrams and Object.
* Involved in designing user screens using **HTML** as per user requirements.
* Used Spring-Hibernate integration in the back end to fetch data from **Oracle** and **MYSQL** databases.
* Used **Spring** Dependency Injection properties to provide loose-coupling between layers.
* Implemented the Web Service client for the login authentication, credit reports and applicant information.
* Used Web services (SOAP) for transmission of large blocks of **XML** data over **HTTP.**
* Implemented the logging mechanism using Log4j framework.
* Wrote test cases in **JUnit** for unit testing of classes.
* Developed application to be implemented on **Windows XP.**
* Created application using **Eclipse IDE.**
* Installed Web Logic Server for handling **HTTP** Request/Response.
* Used Subversion for version control and created automated build scripts.

Environment:CSS, HTML, JavaScript, AJAX, JUNIT, Struts, Spring, and Hibernate, Oracle, and Eclipse.

**Jr. Java Developer**

**Client: Exide Life Insurance, India May 2008 – Dec 2008**

**Description:** The web application system has the policy proposals, claims, accounts and reports. This Project was developed by Java Servlets, Struts, JDBC & Oracle Database Server under Web logic tool.

**Responsibilities:**

* Developed the system by following the agile methodology.
* Involved in the implementation of design using vital phases of the Software development life cycle (SDLC) that includes Development, Testing, Implementation and Maintenance Support.
* Applied **OOAD** principles for the analysis and design of the system.
* Created real time web applications using **Node.js**
* Used **Web sphere Application Server** to deploy the build.
* Developed front-end screens using **JSP, HTML, JQuery, JavaScript and CSS.**
* Used Spring Framework for developing business objects.
* Performed data validation in **Struts Form beans and Action Classes.**
* Used Eclipse for the Development, Testing and Debugging of the application.
* Used DOM Parser to parse the xml files.
* Log4j framework has been used for logging debug, info & error data.
* Used Oracle 10g Database for data persistence.
* Used **WinSCP** to transfer file from local system to other system.
* Performed **Test Driven Development (TDD) using JUnit.**
* Used Ant script for build automation.
* Used Rational Clear Quest for defect logging and issue tracking.

Environment: HTML, JSON, JavaScript, CSS, Struts, Spring, Hibernate, Eclipse, Oracle 10g, SQL Developer