**RAKESH** 

**Sr. Spark/Hadoop Developer**

[**Rakesh.hdpdev@gmail.com**](mailto:Rakesh.hdpdev@gmail.com)

**(+1) 913-717-9808**

**PROFESSIONAL SUMMARY:**

* + - * **Around 8+ years** of professional IT experience with **5** **years** of extensive experience in data **integration**, data **engineering** and data **analytics** using Big Data Systems like **Hadoop** and **Spark**.
      * Very strong knowledge on Hadoop ecosystem components like **HDFS, MapReduce, Spark, Hive, Pig, Sqoop, Flume, Kafka, Oozie and HBase**.
      * Strong experience utilizing programming languages like Java, Scala and Python.
      * Strong knowledge on Architecture of **Distributed** systems and **Parallel** processing frameworks.
      * In-depth understanding of **Spark** execution model and internals of **MapReduce** framework.
      * Expertise in developing production ready Spark applications utilizing **Spark**-**Core**, Data-frames, **Spark**-**SQL**, **Spark**- **ML** and **Spark**-**Streaming** **API’s**.
      * Experience working with various Hadoop distributions like **Cloudera** (Cloudera distribution **CDH3**, 4 and 5) **Horton** **Works** **Distributions** (HDP).
      * Worked extensively on AWS Cloud services like EMR, S3, Redshift, Athena, Glue etc.,
      * Good knowledge in fine tuning resources for long running **Spark** **Applications** to utilize better parallelism and executor memory for more caching.
      * Strong experience working with both **batch** and **real**-**time** processing using **Spark** **frameworks**.
      * Proficient knowledge on Apache **Spark** and programming **Scala** to analyze large datasets using **Spark** to process real time data.
      * Experience working with Cloudera &Hortonworks Distribution of Hadoop.
      * Worked extensively on **Hive** for data analytics and ETL modelling.
      * Strong knowledge on performance tuning **Hive** queries and troubleshooting various issues related to **Joins**, **memory** exceptions in **Hive**.
      * Very good understanding of **Partitions**, **bucketing** concepts in **Hive** and designed both internal and external tables in **Hive** to optimize performance.
      * Strong experience using different columnar file formats like **Avro**, **RCFile**, **ORC** and **Parquet** formats.
      * Hands on experience in **installing**, **configuring** and **deploying** Hadoop distributions both in-house and on cloud.
      * Experience in optimizing **Map**-**Reduce** algorithms by using **Combiners** and custom **practitioners**.
      * Experience in **NoSQL** Column – Oriented Databases like **HBase**, **Apache** **Cassandra**, **MongoDB** and its Integration with **Hadoop** cluster.
      * Expertise in back-end / server- side java technologies such as: Web services, **java** **persistence** **API** (JPA), **Java** **Messaging** **Service** (JMS), **Java** **Database** **Connectivity** (JDBC).
      * Experienced with different scripting language like **Python** and **Shell** **Scripts**.
      * Experienced data processing like **collecting**, **aggregating**, **moving** from various sources using Apache **Flume** and **Kafka**.
      * Extensive experience in **ETL** process consisting of **data** **transformation**, **data** **sourcing**, **mapping**, **conversion** and **loading**.
      * In-depth understanding/knowledge of Hadoop Architecture and various components such as **HDFS**, **Job** Tracker, **Task** Tracker, **Name** Node, **Data** Node, **MapReduce** programming paradigm.
      * Worked with **Sqoop** to move (import / export) data from a relational database into **Hadoop**.
      * Knowledge in **UNIX** **Shell** **Scripting** for automating deployments and other routine tasks.
      * Experienced in using agile methodologies including extreme programming, **SCRUM** and **Test**- **Driven** **Development** (TDD).
      * Used custom **SerDes** like **Regex** **SerDe**, **JSON** **SerDe**, **CSV** **SerDe** etc.., in **hive** to handle multiple formats of data. Intensive work experience in developing enterprise solutions using **Java**, **J2EE**, **Servlets**, **JSP**, **JDBC**, **Struts**, **spring**, **Hibernate**, **JavaBeans**, **JSF**, **MVC**.
      * Experience in building and deploying web applications in multiple applications servers and middleware platforms including **Web** **logic**, **Web** **Sphere**, **Apache** **Tomcat**, **JBoss**.
      * Experience in using version control tools like **Bit**-**Bucket**, **GIT**, **and SVN** etc.
      * Experience in writing build scripts using **MAVEN**, **ANT** and **Gradle**.
      * Flexible, enthusiastic and project-oriented team player with excellent communication skills with leadership abilities to develop creative solutions for challenging requirement of client.

**Technical Skills:**

|  |  |
| --- | --- |
| **Big Data Ecosystems** | HDFS, MapReduce, YARN, Hive, Storm, Sqoop, Pig, Spark HBase, Scala, Flume, Zookeeper, Oozie. |
| **NO SQL Databases** | HBase, Cassandra, MongoDB. |
| **Java & J2EE Technologies** | JDBC, JAVA, SQL, JavaScript, J2EE, C, JDBC, SQL, PL/SQL, Hibernate 3.0, Spring 3.x, Structs |
| **AWS technologies** | Data Pipeline, Redshift, EMR. |
| **Languages** | Java, Scala, Python, SQL, Pig Latin, HiveQL, Shell Scripting. |
| **Database** | Microsoft SQL Server, MySQL, Oracle, DB2. |
| **Web/Application Servers** | Web logic, Web Sphere, JBoss, Tomcat. |
| **IDE’s & Utilities** | Eclipse, JCreator, NetBeans. |
| **Operating Systems** | UNIX, Windows, Mac, LINUX. |
| **GUI Technologies** | HTML, XHTML, CSS, JavaScript, Ajax, AngularJS. |
| **Business Intelligent tools** | Tableau, Splunk, Qlik View. |
| **Development Methodologies** | Agile, V-Model, Waterfall Model, Scrum. |

**PROFESSIONAL EXPERIENCE:**

**CarMax, Richmond, VA Jul 2016 – present Hadoop/Spark Developer**

**Description:**

The objective of the project is to support digital marketing activities of the organization by providing useful insights and intelligence by processing massive amounts of click stream data, campaign and customer profile data using Big Data tools and Ecosystem. In high level, the data is processed and analyzed to find out purchase propensity, attribution of various market channels and customer grouping for effective marketing, customer reach and profits.

**Responsibilities:**

* Ingested gigabytes of click stream data from external servers such as **FTP server** and **S3** buckets on daily basis using customized home-grown Input Adapters.
* Created **Sqoop** scripts to import/export data from **RDBMS** to **S3 data store.**
* Developed various spark applications using **Scala** to perform cleansing, transformation and enrichment of these click **stream data**.
* Involved in data cleansing, event enrichment, data aggregation, de-normalization and data preparation needed for machine learning and reporting.
* Used Spark API over Hortonworks **Hadoop** YARN to perform analytics on data in Hive.
* Utilized **Spark Scala API** to implement batch processing of jobs.
* Trouble Shooting **Spark applications** for improved error tolerance and reliability.
* **Fine-tuning** **spark applications/jobs** to improve the efficiency and overall processing time for the pipelines.
* Created **Kafka** producer API to send live-stream json data into various **Kafka topics.**
* Developed **Spark-Streaming applications** to consume the data from **Kafka topics** and to insert the processed streams to **HBase.**
* Utilized **Spark in Memory capabilities**, to handle large datasets.
* Used **Broadcast variables** in Spark, effective & efficient Joins, transformations and other capabilities for data processing.
* Experienced in working with **EMR** cluster and **S3** in **AWS** cloud.
* Creating Hive tables, loading and analyzing data using hive scripts. Implemented Partitioning, Dynamic Partitions, Buckets in HIVE.
* Involved in continuous Integration of application using Jenkins.
* Interacted with the infrastructure, network, database, application and BA teams to ensure data quality and availability

**Environment: AWS EMR**, Spark, Hive, HDFS, Sqoop, Kafka, Hortonworks,Oozie, HBase, Scala, Java.

**PayPal, San Jose, CA Nov 2015 – Jun 2016**

**Spark Developer**

**Description:**

Aim of this project is to migrate provider data from all the legacy provider platform into Hadoop cluster and build a Centralized Analytical data store for the entire organization to be able to deliver all the analytical requirements from this centralized data lake.

**Responsibilities:**

* Involved in developing roadmap for migration from legacy system to **Hadoop** cluster.
* Create, validate and maintain scripts to load data using **Sqoop** manually.
* Load and transform large sets of structured, semi structured and unstructured data coming from various downstream systems.
* Migrated data between **RDBMS** and **HDFS/Hive** with Sqoop.
* Create, validate and maintain scripts to extract and transform data from **MySQL** to flat files and JSON format.
* Created Oozie workflows and coordinated to automate the workflows daily and monthly.
* Worked on reading multiple data formats on **HDFS** using **Apache spark.**
* Wrote **Scala** scripts for **spark** to perform operations like data inspection, cleaning, load and transforms the large sets of structured and semi-structured imported data.
* Experience in implementing and administering, highly available(HA) Hadoop clusters secured with Kerberos, using the Hortonworks Hadoop distribution.
* Involving in Migrating the coding from **Hive** to **Apache Spark** and Scala using Spark-SQL, RDD.
* Developed **Spar**k with Scala and Spark-SQL for testing and processing of data.
* Developed, validate and maintain **HiveQL** queries.
* Designed Hive tables to load data to and from external **HDFS** datasets.
* Created partitioned and bucketed tables in Hive and designed both Managed and External tables in Hive for optimized performance.
* Managing and scheduling Oozie jobs on a **Hadoop cluster**.

**Environment:** Hadoop, HDFS, Apache Spark 1.6, Spark-SQL, Unix, Hortonworks, Hive, Sqoop, Flume, Scala, Oozie, DB2

**Willis Towers Watson, Nashville, TN Jul 2014 - oct 2015**

**Hadoop Developer**

**Description:**

Client Proposal Integrator application validates data which comes from different Systems mainly from Teradata. This data will undergo several Cleansing and value-added processing and then finally views will be created on this data as Hadoop warehouse. This data will be consumed by downstream for analyzing and generating reports.

**Responsibilities:**

* Extensively involved in Installation and configuration of **Cloudera** distribution Hadoop, Name Node, Secondary Name Node, Job Tracker, Task Trackers, and Data Nodes.
* Developed MapReduce programs in Java and Sqoop the data from **ORACLE database**.
* Responsible for building scalable distributed data solutions using Hadoop. Written various Hive and Pig scripts.
* Moved data from HDFS to **Cassandra** using Map Reduce and Bulk Output Format class.
* Experienced with different scripting language like **Python** and **shell scripts.**
* Developed various Python scripts to find vulnerabilities with SQL Queries by doing SQL injection, permission checks and performance analysis.
* Installed **Oozie** workflow engine to run multiple Hive and Pig jobs which run independently with time and data availability.
* Experienced with handling administration activations using **Cloudera manager.**
* Expertise in understanding Partitions, Bucketing concepts in **Hive.**
* Used Oozie Scheduler system to automate the pipeline workflow and orchestrate the Map Reduces jobs that extract the data on a timely manner. Responsible for loading data from UNIX file system to HDFS.
* Analyzed the weblog data using the HiveQL, integrated Oozie with the rest of the Hadoop stack.
* Utilized cluster co-ordination services through **Zookeeper**.
* Got good experience with various NoSQL databases and Comprehensive knowledge in process improvement, normalization/de-normalization, data extraction, data cleansing, data manipulation.
* Experience with creating script for data modeling and data import and export. Extensive experience in deploying, managing and developing **MongoDB** clusters.
* Created Partitioned Hive tables and worked on them using HiveQL.
* Developed Shell scripts to automate routine DBA tasks.
* Used **Maven** extensively for building jar files of **MapReduce** programs and deployed to Cluster.
* Responsible for cluster maintenance, adding and removing cluster nodes, cluster monitoring, troubleshooting, managing and reviewing data backups and Hadoop log files.

**Environment:** HDFS, Map Reduce, Pig, Hive, Oozie, Sqoop, Flume, HBase, Java, Maven, Avro, Cloudera, Eclipse and Shell Scripting.

**Plat Fora, San Mateo, CA Feb 2012 - Jun 2014 Hadoop Developer**

**Description:**

Plat fora’s software works with open-source software framework Apache Hadoop; when a user queries a database, the product delivers answers in real time via a graphical user interface. Bloomberg Business week called it “Big Data for Dummies”. A corporate or government data analyst can use the interface to filter results or drag and drop fields to create graphs, overlays and other visualizations of the data. The analyst can then share those data visualizations and answers with others.

**Responsibilities:**

* Aggregations and analysis done on large set of log data, collection of log data done using **custom built Input Adapters and Sqoop**.
* Developed **MapReduce** programs for data extraction, transformation and aggregation.
* Monitor and troubleshoot Map Reduce Jobs those are running on the cluster.
* Implemented solutions for ingesting data from various sources and processing the data utilizing Hadoop services like **Sqoop, Hive Pig, HBase, Map Reduce etc**.
* Worked on creating Combiners, Practitioners and Distributed cache to improve the performance of Map Reduce jobs.
* Wrote Pig scripts to generate Map Reduce jobs and performed ETL procedures on the data in **HDFS.**
* Experienced in handling **Avro** data files by passing schema into HDFS using Avro tools and Map Reduce.
* Optimization of Map Reduce algorithms using combiners and practitioners to deliver the best results and worked on Application performance optimization for a HDFS cluster.
* Orchestrated many Sqoop scripts, Pig scripts, Hive queries using **Oozie** workflows and sub workflows.
* Used Flume to collect, aggregate and store the web log data from different sources like web servers and pushed to HDFS.
* Involved in creating Hive tables, loading with data and writing Hive queries which will invoke and run Map Reduce jobs in the backend.
* Created **HBase** tables to load large sets of structured, semi-structured and unstructured data coming from UNIX, **NOSQL** and a variety of portfolios.
* Involved in debugging Map Reduce jobs using **MR Unit framework** and optimizing Map Reduce jobs.
* Involved in troubleshooting errors in Shell, Hive and Map Reduce.
* Worked on debugging, performance tuning of **Hive & Pig jobs**.
* Design and implement Map Reduce jobs to support distributed processing using Map Reduce, Hive and Apache Pig.
* Created Hive external tables on the Map Reduce output before partitioning, bucketing is applied on top of it.

**Environment:** Hadoop, HDFS, MapReduce, HIVE, Pig, Sqoop, HBase, Oozie, MySQL, SVN, Putty, Zookeeper, UNIX, Shell Scripting, HiveQL, NOSQL database (HBASE), RDBMS, Eclipse, Oracle 11g.

**Airtel, India Dec 2010 - Nov 2011**

**Sr Java/ J2EE Developer**

**Description:**

Airtel is credited with pioneering the business strategy of outsourcing all its business operations except marketing, sales and finance and building the minute’s factory model of low cost and high volumes. The strategy has since been adopted by several operators. Airtel’s equipment is provided and maintained by Ericsson, Huawei, Nokia Solutions and Networks whereas IT support is provided by IBM.

**Responsibilities:**

* Created Use case, Sequence diagrams, functional specifications and User Interface diagrams using **Star UML**.
* Involved in complete requirement analysis, **design**, **coding** and **testing** phases of the project.
* Participated in **JAD** meetings to gather the requirements and understand the End Users System.
* Developed user interfaces using **JSP, HTML, XML and JavaScript**.
* Generated **XML schemas** and used **XML Beans** to parse XML files.
* Created Stored Procedures & Functions. Used JDBC to process database calls for **DB2/AS400** and **SQL Server** databases.
* Developed the code which will create **XML** files and **Flat** Files with the data retrieved from **Databases** and **XML** files.
* Created Data sources and Helper classes which will be utilized by all the interfaces to access the data and manipulate the data.
* Developed web application called **IHUB (integration Hub)** to initiate all the interface processes using **Structs Framework**, **JSP** and **HTML**.
* Developed the interfaces using **Eclipse** 3.1.1 and **JBoss** 4.1 involved in **integrated** **testing**, **Bug** **fixing** and **production** **Support**.

**Environment:** Java 1.3, Servlets, JSPs, Java Mail API, Java Script, HTML, MYSQL 2.1, Swing, Java Web Server 2.0, JBoss 2.0, RML, Rational Rose, Red Hat Linux 7.1.

**Phoenix InfoTech Pvt. Ltd, India Jan 2010 - Nov 2010**

**Java Developer**

**Description:**

Software publishing, consultancy and supply Software publishing includes production, supply and documentation of ready-made (non-customized) software, operating systems software, and business & other applications software, computer games software for all platforms. Consultancy includes providing the best solution in the form of custom software after analyzing the user’s needs and problems. Custom software also includes made-to-order software based on orders from specific users. Also, included are writing of software of any kind following directives of the users.

**Responsibilities:**

* Involved in developing the application using **Java/J2EE platform**. Implemented the **Model View Control (MVC**) structure using Structs.
* Responsible to enhance the Portal UI using **HTML**, **JavaScript, XML, JSP, Java, CSS** as per the requirements and providing the client-side Java script validations and Server-side bean Validation Framework (JSR 303).
* Used **Spring** **Core** Annotations for Dependency Injection.
* Used **Hibernate** as persistence framework mapping the **ORM** objects to table using Hibernate annotations.
* Responsible to write the different service classes and utility API which will be used across the frame work.
* Used Axis to implementing Web Services for integration of different systems.
* Developed Web services component using **XML, WSDL** and **SOAP** with **DOM** parser to transfer and transform data between applications.
* Exposed various capabilities as Web Services using **SOAP/WSDL**.
* Used **SOAP** **UI** for testing the **Restful Webservices** by sending and **SOAP** request.
* Used **AJAX** framework for server communication and seamless user experience.
* Created test framework on **Selenium** and executed Web testing in **Chrome, IE** and **Mozilla** through **Web driver**.
* Used client-side java scripting: **JQUERY** for designing **TABS** and **DIALOGBOX**.
* Created **UNIX** shell scripts to automate the build process, to perform regular jobs like file transfers between different hosts.
* Used **Log4j** for the logging the output to the files.
* Used **JUnit/ Eclipse** for the unit testing of various modules.
* Involved in **production support**, monitoring server and error logs and foreseeing the Potential issues and escalating to the higher levels.

**Environment:** Java, J2EE, JSP, Servlets, Spring, Servlets, Custom Tags, Java Beans, JMS, Hibernate, IBM MQ Series, AJAX, Junit, Log4j, JNDI, Oracle, XML, SAX, Rational Rose, UML.

**Education:**

Bachelor’s in computer science, India.