## **Waseem**

**Big Data Engineer**

**Summary:**

* 8+ years of strong experience in software development using **Bigdata, Hadoop, Apache Spark, Java/J2EE, Python & Scala technologies.**
* In-depth knowledge of **Scala** and experienced in building the **Spark** applications using Scala.
* Very good hands-on in **Spark Core, Spark SQL, Spark Streaming** and **Spark ML (Machine Learning)**.
* Experienced in working with **Cloudera**, **Hortonworks**, &**Microsoft Azure HDInsight** Distributions.
* Installed **APACHE NIFI and MINIFI** to make data ingestion Fast, Easy and Secure.
* Good troubleshooting skills on **Hue**, which provides **GUI** for developers/business users for day to day activities.
* Highly skilled in integrating **Kafka** with **Spark streaming** for high speed data processing.
* Sound knowledge in using **Apache** **Solr** to search against structured and un-structured data.
* Involved in loading data from **UNIX** file system to HDFS And Created custom **Solr** Query components to enable optimum search matching.
* Worked on migrating **MapReduce programs** into Spark transformations using **Scala.**
* Experienced in building Computation framework in **Scala** for **Spark POC**
* Solid understanding of **RDD**operations in **Apache Spark** i.e., Transformations & Actions,Persistence (Caching), Accumulators, Broadcast Variables, Optimizing Broadcasts.
* Hands on experience on **AWS cloud** services (**VPC, EC2, S3, RDS, Redshift, Data Pipeline, EMR, DynamoDB, Workspaces, Lambda, Kinesis, RDS, SNS, SQS**).
* Good expertise in using AWS services like **EMR, EC2 and S3** to run apache spark development andproduction jobs.
* Used **Amazon Kinesis** Video Streams to capture, process, and store video streams for analytics and machine learning.
* Experienced in running query using **Impala** and used BI tools to run ad-hoc queries directly on Hadoop.
* Processed **flat** **files** in various **file formats** and stored them as in various partition models in **HDFS**.
* Experienced in **importing data** from various data sources like **Mainframes, Teradata, Oracle and Netezza** using **Sqoop,** performed transformations using **Hive, Pig and Spark** and loaded data into **HDFS.**
* Used **File System check (FSCK)** to check the health of files in **HDFS**.
* Very well-versed knowledge with workflow scheduling and monitoring tools like **Oozie** and**Zookeeper.**
* Used **Zookeeper** on a distributed **HBase** for cluster configuration and management.
* Well Experienced in working with **Flume** to load the log data from multiple sources directly into **HDFS**.
* Implemented **Ambari** to keep track of node status, job progression and running analytical jobs in Hadoop clusters.
* Having good working experience of **No SQL** database like **HBase, Cassandra and Mongo DB**.
* Developed various **Map Reduce** and spark applications to perform **ETL** workloads onterabytes of data.
* Familiar with **Data Ingestion pipeline** design and advanced data processing tools.
* Strong Experience of Data Warehousing **ETL** concepts using **Informatica Power Center, OLAP, OLTP and AutoSys.**
* Extensive experience in **PL/SQL**, developing stored procedures with optimization techniques.
* Skilled in rendering and delivering reports in desired formats by using reporting tools such as **Tableau**.
* Experienced in using IDEs and Tools like **Eclipse, Net Beans, GitHub, Jenkins, Maven and IntelliJ.**
* Excellent Java development skills using **J2EE, Servlets, Junit** and familiar with popular frameworks such as **Spring, MVC and AJAX**.
* Adept at Web Development and experience in developing front end applications using **JavaScript, CSS and HTML.**
* Expertise in **Waterfall and Agile - SCRUM** methodologies.
* Proficient with **Core Java, AWT** and with the mark-up languages like **HTML, XHTML, DHTML, XQuery, Angular JS.**
* Worked with version control systems like **Subversion** and **GIT** for providing common platform for all the developers.
* Knowledge on **IDE tools**such as **Eclipse, NetBeans,**and **Maven.**
* Highly motivated, dedicated and hardworking, with strong Analytical & Logical Development skills.

**Skills in Review**

|  |  |
| --- | --- |
| **Languages& Scripting** | Python, Scala, Java8.0,Shell Script, JavaScript |
| **Big Data Frameworks** | HDFS, Map Reduce, Hive, YARN,HBase, Pig, Impala, Apache Solr, Apache Spark, Sqoop, Zookeeper, Storm, Drill, Flume, Kafka, MongoDB, Cassandra |
| **Cloud Technologies** | Amazon EC2, S3, EMR, Dynamo DB,**Lambda, Kinesis,** ELB, RDS, SNS, SQS, EBS |
| **Development Tools** | Microsoft SQL Studio, IntelliJ, Eclipse, NetBeans,Maven |
| **Database/ RDBMS** | MYSQL, MS-SQL server, DB2, Oracle 11g/10g/9 |
| **Web Development** | HTML, XML, AJAX, SOAP, WSD |
| **Application Servers** | WebLogic, JBoss, Apache Tomcat 8.0, IBM WebSphere |
| **Operating Systems** | Unix, Linux, Windows, Mac |
| **Version Control** | GIT, SVN,**Win CVS, VSS** |
| **BI Tools** | Tableau, Splunk, Pentaho |

**Education:**

Bachelor of Engineeringin Information Technology, Osmania University, IND

**Professional Experience:**

**MetLife, Cary, NC Feb 18 – till date**

**Sr. Big Data Developer**

**Responsibilities:**

* Designed high performing batch programs and adhering to standards, best practices for development.
* Developed **Spark** with **Scala** and **Spark SQL** for testing and processing of data.
* Implemented **POC** to **migrate** map reduce jobs into **Spark** **RDD** transformations.
* Developed**Scala** scripts for spark to perform operations as data inspection, cleaning, load and transforms the large sets of structured and semi-structured imported data.
* Used **Spark-SQL** to Load **JSON** data and create Schema RDD and loaded it into Hive Tables and handled structured data using Spark SQL.
* Involved in creating **External Hive tables** and involved in data loading and writing **Hive UDFs.**
* Involved in developing **Hive DDLs** to create, alter and drop tables.
* Created Hive snapshot tables and **Hive** ORC tables from hive tables.
* Implemented **Partitioning, Dynamic Partitions, Buckets** in **HIVE**.
* Applied skills in **UNIX** commands and **shell scripting** in Oracle, PL/SQL, SQL tuning, Informatica, Sqoop, Spark SQL and data modelling.
* Prepared **Linux shell scripts** for automating the process.
* Used **Putty -SSH Client** to connect remotely to the servers.
* Experienced in **ETL** developmen**t**, enhancement and maintenance of database applications.
* Used **SSIS** as an **Extract Transform Loading (ETL)** tool of **SQL Server** to populate data from various data sources, creating packages for different data loading operations.
* Involved in End to End implementation of **ETL** logic.
* Designed dynamic processing through **PL/SQL** for projects and created **SQL database** packages, procedures.
* Worked with data transfer in **Oracle** and **Hadoop** environments.
* Stacked data from **LINUX** file system to **HDFS.**
* Involved in unit testing using **Junit.**
* Worked on **ORC**, **Avro** file formats and some compression techniques like **LZO**.
* Evaluated usage of **Oozie** for Workflow Orchestration.
* Involved in story-driven **agile development** methodology and actively participated in daily scrum meetings.

**Environment:** Sqoop,Oozie, UNIX, Shell Scripting, ETL, Spark, SSIS, HUE, AZURE, MS Visio, Microsoft SQL server, Python, Scala, Talend, IntelliJ, Tableau, Microsoft Azure HDInsight, Junit, GIT,WinSCP, Putty

**Capital One, Wilmington, DE Jan 17 – Feb18**

**Big Data Developer**

**Description:**Retail Enterprise Credit Risk application calculates Bank's retail data such as credit cards, auto, student and home loans for risk domains including Enterprise Capital Management. Data comes from different System of records mainly from Teradata. This data will undergo several cleansings and value-added processing and then finally views will be created on this data as Hadoop warehouse. This data will be consumed by downstream like ECM for analysing and generating reports.

**Responsibilities:**

* Interacting with multiple teams understanding their business requirements for designing flexible and common component.
* Validating the source file for Data Integrity and Data Quality by reading header and trailer information and column validations.
* Used **Spark SQL** for creating data frames and performed transformations on data frames like adding schema manually, casting, joining data frames before storing them.
* Implemented Spark SQL to access **Hive** tables into spark for faster processing of data.
* Worked on **Spark streaming** using **Apache Kafka** for real time data processing.
* Developing and testing **Data** **Ingestion/Preparation/Dispatch** jobs.
* Used **Hive** to do transformations, joins, filter and some pre-aggregations before storing the data onto HDFS.
* Created external **Hive tables** to store data which is loaded.
* Used Sqoop for importing and exporting data from **Netezza, Teradata** into **HDFS and Hive**.
* Worked on three layers for storing data such as raw layer, intermediate layer and publish layer.
* Optimizations techniques include partitioning&bucketing.
* Experience working with **Apache SOLR** for indexing and querying.
* Created custom **SOLR** Query segments to optimize ideal search matching.
* Using **Avro file format** compressed with Snappy in intermediate tables for faster processing of data.
* Used **Kinesis Data Analytics** to analyse data streams with SQL.
* Used **parquet file format** for published tables and created views on the tables.
* Created sentry policy files to provide access to the required databases and tables to view from **Impala** to the business users.
* Automated the jobs with **Oozie** and scheduled them with **Autosys**.
* Experience in **Amazon AWS** to spin up the **EMR** cluster to process the huge data which is stored in**Amazon S3** and push it to **HDFS**.
* Participated in evaluation and selection of new technologies to support system efficiency.
* Participated in development and execution of system and disaster recovery processes.

**Environment**:  Hadoop, Cloudera, Amazon AWS, HDFS, Hive, Impala, Apache Spark, Autosys, Kafka, DynamoDB, Lambda, s3, SQS, SNS, Sqoop, Java, Scala, Eclipse, Tableau, Teradata, UNIX, and Maven, SBT.

**Citigroup, Irving, TX July 15 – Dec 16**

**Sr. Hadoop/Spark Developer**

**Description**: Citi works tirelessly to provide consumers, corporations, governments and institutions with a broad range of financial services and products. We work with companies to optimize their daily operations, whether they need working capital, to make payroll or export their goods overseas.

**Responsibilities:**

Responsible for migrating the code base from Cloudera Platform to Amazon EMR and evaluated Amazon eco systems components like **RedShift, Dynamo DB**.

* Implemented Data loading using **Spark, Storm, Kafka, Elastic Search**.
* Stored data in **AWS S3** like **HDFS** and performed EMR programs on data stored in S3.
* Exported the analysed data to the relational databases using **Sqoop** for visualization and to generate reports for the BI team.
* Migrated Map reduce jobs to Spark Jobs to achieve better performance.
* Design and Implementation of Real time applications using **Apache Storm, Trident Storm, Kafka**.
* Used Spark Data Frame API to process **Structured** and **Semi Structured** files and load them back into S3 Bucket.
* Experienced with batch processing of data sources using **Apache Spark**, **Elastic Search**.
* Created Hive External tables and loaded the data in to tables and query data using **HQL**.
* Responsible to manage data coming from different sources.
* Good knowledge in cloud integration with **Amazon Elastic MapReduce (EMR**).
* Experience in integrating **Cassandra** with **Elastic Search** and Hadoop.
* Extensive experience in Spark Streaming through core **Spark API** running Scala, Java & Python scripts to transform raw data from several data sources into forming baseline data.
* Hands on expertise in running the **Spark**&**Spark SQL** on **Amazon (EMR).**
* Implemented SPARK batch jobs on **AWS EC2** instances through Amazon Simple Storage Service (Amazon S3).
* Performed performance tuning for **Spark Steaming** e.g. setting right Batch Interval time, correct level of Parallelism, selection of correct Serialization& memory tuning.
* Created **HBase** tables to store variable data formats of input data coming from different portfolios.
* Involved in adding huge volumes of data in rows and columns to store data in HBase.
* Used Spark API over **Hadoop YARN** to perform analytics on data in Hive.
* Developed Spark code using **Scala** and **Spark-SQL** for batch processing of data.
* Involved in requirement and design phase to implement Streaming Lambda Architecture to use real time streaming using **Spark and Kafka**.
* Hands on experience working on **NoSQL** databases like **HBase**.
* Used **Spark API** over Cloudera **Hadoop YARN** to perform analytics on data in **Hive**.
* Migrated Map reduce jobs to Spark Jobs to achieve better performance.

**Environment**: Hadoop, Map Reduce, Spark, Spark SQL, Kafka, Storm, HDFS, Hive, Sqoop, Oozie, Java, SQL, Shell script.

**Expedia, Chicago, IL Dec 13–Jun15**

**Sr. Java Developer**

**Responsibilities:**

* Development involved technologies like **Java, J2EE** (**Multithreading, JNDI, XML Parsers**), **JSP, Servlets**, and **Spring MVC**.
* Implemented business logic using **Session Beans, Servlets and stored procedures**.
* Created user-friendly GUI interface and Web pages using **HTML** and **DHTML** embedded in **JSP**.
* Deployed in different environments like **DB-DEV**, DEV and Stable. And worked with testing team for Staging Deployment.
* Developed **JSP** pages, Controller, Helper & Validator classes for the application.
* Designed customized GUI pages based on the business requirement.
* Worked with client-side and server-side validations for the web pages developed.
* Prepared all related documentation (Master Test Plan and Test Project Schedule).
* Worked closely with **Application Architects**, **Business Analysts**, and **Project Managers** to discuss business requirements and overall status of assigned projects.
* Used **JIRA** for defect tracking and raising different type of severity of the defects.
* Worked on cross browser compatibility issues for **Chrome, Firefox, Safari, IE 11**.
* Worked in Agile (SCRUM) environment and used **Git Bash** as version control tool.
* Worked on Rest Web services, to specify the business behaviours and backend calls.
* Opened **CRQs** to deploy the application in **Stable and Production environments**.
* Deployed the application in Development,Stable and Production environments and validated the application upon successful deployment.
* Performed manual testing on the changes made to the application.

**Environment:** Soap UI 5.0, Spring Framework, JIRA, Maven, Crucible, Confluence, HTML/HTML5, CSS/CSS3, XML, JavaScript, XPATH, XSLT, Web Services, SQL Server 2008, Command Editor, Hibernate ORM, Java7 & 8, JEE, Multithreading, JAX\_RS, JSON, GitHub, Junit, IntelliJ**.**

**HSBC Software Development, India Apr 2012 – Oct 2013**

**Sr. Java Developer**

**Responsibilities:**

* Involved in **Analysis, Designing, Development and Testing** phases of the application.
* Was involved in creation and maintenance of the backend services using **Spring, Hibernate, SQL Server and Oracle.**
* Developed Web pages using **JSPs with Tag libraries, HTML, and JavaScript**.
* Writing **J2EE** code using **Spring, hibernate** to upload input **CSV files** for credit risk data.
* Implemented **Dependency Injection (IOC**) feature of spring framework to inject dependency into objects and AOP is used for Logging.
* Designed and developed persistence layer build on **ORM framework** and developed it using Hibernate
* Implemented various Design patterns like Business Delegate, Data Transfer Objects DTO, Service locator, Session Facade and Data Access Objects **DAO patterns**.
* Involved in writing **SQL, Stored procedure, and PL/SQL** for back end.
* Used Views and Functions at the **Oracle Database** end.
* Developed various documents within the application using **XML** by using **Eclipse** as IDE tool.
* Developed **SOAP** requests to interact with billing schedule system.
* Used Webservices like **Soap & WSDL** to send information or remote procedure calls encoded as **XML.**
* Integrating and deploying the application on **WebLogicapplication server** using **ANT**.
* Developed user interfaces for presenting the expense reports, transaction details using **JSP, XML, HTML, and Java Script.**
* Used **Log4J** for logging the application exceptions and debugging statements.
* Proficient in doing Object Oriented Design using **UML-Rational Rose.**

**Environment:** Java, JSP, Servlets, Web Sphere Application Server, Eclipse, Java Script, Web Services (SOAP & WSDL), Microsoft VSS, Oracle, PL/SQL and JDBC.

**Assurant, Hyderabad, India May 2010–Apr 2012**

**Java /J2EE Developer**

**Responsibilities:**

* Maintained the UI screens using web technologies like **HTML, Java Script, jQuery**, and **CSS**.
* Involved in requirements like **analysis, design**, **development** and **testing**.
* Designed, deployed, and tested Multi-tier application using the **Java** technologies.
* Involved in front end development using **JSP, HTML and CSS**.
* Documented the changes for future development projects.
* Responsible to write the different service classes and utility **API** which will be used across the frame work.
* Used **Axis** to implement Web Services for integration of different systems.
* Used **MYSQL** database to store data and execute **SQL** queries on the backend.
* Exposed various capabilities as Web Services using **SOAP/WSDL**.
* Used **SOAP UI** for testing the Restful Webservices by sending an SOAP request.
* Used Hibernate as Persistence framework mapping the **ORM objects** to table using **Hibernate** annotations.
* Involved in developing **JSP** for client data presentation and data validation on the client side with in the forms.
* Used **JDBC** connections to store and retrieve the data from the database.
* Developed Web services component using **XML, WSDL,** and **SOAP** with **DOM parser** to transfer and transform data between applications.
* 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, JNDI, Oracle, XML, SAX, Rational Rose, UML.