**Udisha Agarwal**

**Big Data/ Scala Developer/Hadoop Developer**

**PROFESSIONAL SUMMARY:**

* A dynamic professional with over **7**+ years of diversified experience in the field of Information Technology with an emphasis on **Big Data/Hadoop Eco System, SQL/NO-SQL databases, Java / J2EE technologies** and tools using industry accepted methodologies and procedures.
* **Hadoop Development:** Extensively worked on Hadoop tools which include **Hive, Oozie, Sqoop, and Spark, Data frames, Spark Streaming, HBase and MapReduce programming, Apache Flink**. Created Partitions and Bucketing concepts in **Hive** and designed both Managed and External tables in **Hive** to optimize performance.
* Experience in installation, configuring, supporting and managing Hadoop Clusters using **Apache, Cloudera (CDH 5.X) distributions** and on **Amazon web services (AWS**).
* Experience in Amazon AWS services such as **EMR, EC2, S3, cloud Formation, Red shift** which provides fast and efficient processing of Big Data.
* Developed **Spark** applications using **Scala** for easy Hadoop transitions. Used **Spark API** over **Cloudera Hadoop YARN** to perform analytics on data in **Hive**. Developed **Spark** code and **Spark- SQL/Streaming** for faster testing and processing of data.
* **Hadoop Distributions:** Worked with Apache Hadoop along enterprise version of **Cloudera and Hortonworks** and good Knowledge on MAPR distribution.
* **Data Ingestion in to Hadoop (HDFS):** Ingested data into Hadoop from various data sources like **Oracle, MySQL using Sqoop tool**. Created **Sqoop** job with incremental load to populate **Hive** External tables. Involved in importing the real-time data to Hadoop using **Kafka** and also worked on **Flume**. Exported the analyzed data to the relational databases using **SQOOP** for visualization and to generate reports for the BI team.
* **File Formats:** Involved in running Hadoop streaming jobs to process terabytes of text data and worked with different file formats such as **Text, Sequence files, Avro, ORC and Parquet.**
* **Scripting and Reporting**: Created scripts for performing data-analysis with **HIVE and IMPALA** and used the **ANT script** for creating and deploying.**jar, ear and .war files**. Generated reports, extracts and statistics on the distributed data on Hadoop cluster. Generated **Java APIs** for retrieval and analysis on **No-SQL** database such as **HBase and Cassandra.**
* **Java Experience:** Created applications in **core Java**, built application that satisfy use of database and constant connectivity such as a client-server model using **JDBC, JSP, spring and Hibernate** and implemented web-services for network related applications in java.
* In depth understanding/knowledge of Hadoop Architecture and various components such as **HDFS, MR, Hadoop GEN2 Federation, High Availability and YARN** architecture and good understanding of workload management, scalability and distributed platform architectures.
* Strong Knowledge and experience on architecture and components of **Spark**, and efficient in working with **Spark Core, Spark SQL, Spark streaming** and implemented Spark Streaming jobs by developing **RDD's (Resilient Distributed Datasets)** and used **PySpark** and **spark-shell** accordingly.
* Good knowledge in using **Apache NiFi** to automate the data movement between different Hadoop systems.

**EDUCATION:**

* **PGDCA, Kumaun university, India – 2008.**
* **Master in English, Kumaun university, India – 2011.**

**TECHNICAL SKILLS:**

* **Bigdata Technologies:** HDFS, MapReduce, Hive, HBase, Scala, Spark, Apache, Sqoop, Oozie, Kafka, MongoDB, Apache NiFi.
* **Databases:** MongoDB, HBase, Cassandra, Oracle 10g/11g/12c, PL/SQL, MySQL, MS SQL Server 2016/2012. MySQL
* **Cloud:** AWS S3, AWS EMR, AWS EC2, Redshift and AWS Glue.
* **Programming Languages:** Scala, Java, J2EE, PL/SQL and SQL
* **Java/J2EE Technologies:** JDK 1.8, Microservices, JDBC, JNDI, JSON, JSTL, RMI, JMS, JSP, Servlets
* **Development Methodologies:** Agile, Waterfall
* **ETL Tools:** Import Export Data, Talend.
* **IDE Tools:** Eclipse, IntelliJ
* **Operating System:** Windows 7/8/10, Vista, UNIX, Linux, Ubuntu, Mac OS X.

**PROFFESIONAL EXPERIENCE:**

**BNY Mellon, NY (Remote)**

**Scala/Hadoop Developer Dec ’22– Present**

**Description:**

The Bank of New York Mellon Corporation, commonly known as BNY Mellon, is an American investment banking services holding company headquartered in New York City. BNY Mellon was formed from the merger of The Bank of New York and the Mellon Financial Corporation in 2007. It is the world's largest custodian bank and securities services company. it is one of the three oldest banking corporations in the United States and among the oldest banks in the world, having been established in June 1784 by a group that included American Founding Father Alexander Hamilton.

**Responsibilities:**

* Handled importing of data from various data sources, performed transformations using **Hive, MapReduce, loaded data** into **HDFS** and Extracted the data from **MySQL** into **HDFS** using **Sqoop**.
* Generated **Java APIs** for retrieval and analysis on **No-SQL** database such as **HBase** and **Cassandra.**
* Built pipeline using **Kafka streaming** and provided the data for analytics team.
* Developed data quality engine using the spark **MVEL** framework, which can be used to provide quality analysis of data on any type of data.
* Used Scala sbt to develop scala coded spark project and executed using spark-submit.
* Develop new modules in scala async and scala futures.
* Strong core scala with collections and type variance.
* Implemented **Partitioning, Bucketing** in **Hive** for better organization of the data.
* Developed spark application using **Scala** to develop the connected components using sparkling **graph-x algorithm**.
* Developed spark application to create row number for billions of records within 3 mins and provides hive table on top of data. Provided this as reusable component for the team.
* Worked on performance tuning of spark application, to resolve memory issues on huge data.
* Worked with the **Spark** for improving performance and optimization of the existing algorithms in **Hadoop** using **Spark Context, Spark-SQL, Spark MLlib, Data Frame, Pair RDD's, Spark YARN.**
* Configurating and deploying using **AWS apps** technologies such as **RDS, DynamoDB, RedShift** and **AWS IaaS** Technologies such as **EC2, S3** etc.
* Developed many hive **UDFs** as reusable components for business team for analysis and data transformation.
* Created hive tables on huge data using hive orc and compression techniques.
* Have worked on large volumes of structured, unstructured and semi structured data and have developed many **MapReduce** for data cleansing. i.e., filtering and projection of data.
* Hands on experience on Hue to import data on to the **Graphical User Interface.**
* Executed different performance tuning techniques on **HBase schema** creation, region splitting and row key design.
* Implemented the index for fast and effective search for Business and address matching in **Elastic search**.
* Developed **Apache Lucene** indexing to perform fast searching on huge flat files.
* Extracted data from **HBase** and stored as external tables in **Hive using** serve, to perform **Data analytics**.
* Written complex **map reduce** programs to perform data transformation.

**Deutsche Bank, Chicago, IL**

**Big Data Developer / Hadoop Developer Nov ’19– Nov’21**

**Description:**

Deutsche bank is a leading global financial services corporation that, through its subsidiaries and affiliates, advises, and originates, trades, manages and distributes capital for, governments, institutions and individuals. The company operates in three business segments: Institutional Securities, Wealth Management, and Investment Management.

**Responsibilities:**

* Analyze and define researcher's strategy and determine system architecture and requirement to achieve goals and developed multiple **Kafka** Producers and Consumers from as per the software requirement specifications.
* Used **Kafka** for log accumulation like gathering physical log documents off servers and places them in a focal spot like **HDFS** for handling.
* Configured **Spark** Streaming to get ongoing information from the **Kafka** and store the stream information to **HDFS**.
* Implemented usage of **Amazon EMR** for processing Big Data across Hadoop Cluster of virtual servers on **Amazon Elastic Compute Cloud (EC2)** and **Amazon Simple Storage Service (S3).**
* Involved in development of Hadoop System and improving multi-node Hadoop Cluster performance and responsible for developing data pipeline with **Amazon AWS** to extract the data from weblogs and store in **MongoDB.**
* Developed Real time data processing applications by using **Scala and Python** and implemented **Apache Spark** Streaming from various streaming sources like **Kafka, Flume and JMS.**
* Storing and loading the data from **HDFS to Amazon S3** and backing up the Namespace data into NFS and integrated Hive server 2 with Tableau using Horton Works **Hive ODBC driver,** for auto generation of Hive queries for non-technical business user.
* Responsible for building scalable distributed data solutions using **Hadoop** cluster environment with Horton works distribution and ingested streaming data with **Apache NiFi into Kafka.**
* Wrote **MapReduce** jobs using **Java API**, Optimized **Hive queries** by using execution engine like **Spark.**
* Worked in **AWS** environment for development and deployment of custom Hadoop applications and involved in working with **Elastic Map Reduce (EMR)** and setting up environments on **Amazon AWS EC2** instances.
* Load and transform large sets of structured, semi structured data using Hive and extract real time feed using **Kafka** and **Spark** Streaming and convert it to RDD and process data in the form of Data Frame and save the data as Parquet format in **HDFS.**
* Used **Spark and Spark-SQL** to read the Parquet data and create the tables in **Hive** using the **Scala API.**
* Used **Java API** and **Sqoop** to export data into **DataStax Cassandra cluster from RDBS** worked on **Cassandra** for retrieving data from **Cassandra** clusters to run queries.
* Implemented **Spark** using **Scala and Spark SQL** for faster testing and processing of data files and involved in making code changes for a module in turbine simulation for processing across the cluster using spark-submit.
* Involved in performing the analytics and visualization for the data from the logs and estimate the error rate and study the probability of future errors using regressing models.
* Used **Kafka** to patch up a customer activity taking after pipeline as a course of action of steady appropriate subscribe supports.

**Amex, New York**

**Big Data Engineer / Hadoop Developer Jan’18 – Oct’19**

**Description:**

Amex is a leading Organization in financial services. Amex used the Bigdata technology to load the customer data into Hadoop clusters for the linking of vast data transactions between merchants and clients and to provide better services.

**Responsibilities:**

* Developed Simple and complex Map Reduce streaming jobs using **Java** language.
* Ingested data into **HDFS** from Oracle and vice-versa using **Sqoop**.
* Handled structured and unstructured data and applying **ETL** processes.
* Written multiple **UDF** programs in Java for data extraction, transformation and aggregation from multiple file formats (**XML**, **JSON**, and **CSV**).
* Develop and maintained complex outbound notification applications that run on custom architectures, using languages **Core Java**, **J2EE**, **SOAP**, **XML**, **JMS**, **JBoss** and Web Services.
* Experienced in monitoring and debug performance issues on Linux (**RHEL& Centos**).
* Involved in Production Rollout Support which includes monitoring the solution post go-live and resolving any issues that are discovered prior to Rollout.
* Integrating Hadoop with **Kafka**. Expertise in uploading **Click stream** data from **Kafka** to HDFS.
* Designed and documented operational issues by following standards and procedures in a software reporting tool **JIRA**.

**Bank of Ireland, India**

**Big Data/ Hadoop Developer June 2015 – Feb 2017**

**Description:**

Bank of Ireland is one for the largest financial services group in Ireland and provide a broad range of banking and other financial services. Data cleansing and business transformations are implemented in Hadoop ecosystem. The data is provisioned to downstream systems for reporting and dash boarding purposes.

**Responsibilities:**

* Evaluated business requirements and prepared detailed specifications that follow project guidelines required to develop written programs.
* Exploring with the **Spark** improving the performance and optimization of the existing algorithms in **Hadoop** using **Spark Context, Spark-SQL, Data Frame, Pair RDD's, Spark YARN.**
* Managed and reviewed **Hadoop log files** to identify issues when job fails and used **HUE** for UI based **Oozie scheduling**.
* Involved in creating **data-lake** by extracting customer's data from various data sources to **HDFS** which include data from **Excel, databases, and log data from servers.**
* Automated workflows using shell scripts pull data from various databases into **Hadoop** and developed scripts to automate the process and generate reports.
* Created detailed **AWS Security groups** which behaved as virtual firewalls that controlled the traffic allowed reaching one or more **AWS EC2** instances.
* Designed number of partitions and replication factor for **Kafka** topics based on business requirements and worked on migrating **MapReduce** programs into **Spark** transformations using **Spark** and **Scala,** initially done using **python (PySpark).**
* Used various **Spark** Transformations and Actions for cleansing the input data and involved in using the **Spark** application master to monitor the **Spark jobs** and capture the logs for the spark jobs.
* Worked with **Amazon EMR** to process data directly in **S3** when we want to copy data from **S3** to the **Hadoop Distributed File System (HDFS)** on **Amazon EMR** cluster by setting up the **Spark** Core for analysis work.
* Implemented **Spark** using **Scala** and utilizing **Data frames** and **Spark SQLAPI** for faster processing of data and worked on extensible framework for building high performance batch and interactive data processing application on hive.
* Extracted Real time feed using **Spark streaming** and convert it to **RDD** and process data into **Data Frame** and load the data into **Cassandra.**
* Integrated **Oozie** with **Map-Reduce, Hive, and Sqoop** and developed **Oozie workflow** for scheduling and orchestrating the **ETL** process within the **Cloudera Hadoop** system.
* Solved performance issues in **Hive** with understanding of Joins, and group and aggregation and how does it translate to MapReduce jobs.