**Archana**

**Bigdata Developer || Email: archana.y0929@gmail.com|| Phone no: (469) 846 8345**

**PROFESSIONAL SUMMARY:**

* 8+ years of Full Software Development Life Cycle experience in Software, System analysis, design, development, testing, deployment, maintenance, enhancements, re-engineering, migration, troubleshooting and support of multi-tiered web applications in high performing environments.
* In depth knowledge in HDFS, Job Tracker, Task Tracker, Name Node, Data Node and Map Reduce programming.
* Expertise in converting Map Reduce programs into Spark transformations using Spark RDD's.
* Expertise in Spark Architecture including Spark Core, Spark SQL, Data Frames, Spark Streaming and Spark MLlib.
* Configured Spark streaming to receive real time data from the Kafka and store the stream data to HDFS using Scala.
* Experience in implementing Real-Time event processing and analytics using messaging systems like Spark Streaming.
* Experience in using Kafka and Kafka brokers to initiate spark context and processing live streaming information with the help of RDD.
* Good knowledge on Amazon AWS concepts like EMR and EC2 web services which provides fast and efficient processing of Big Data.
* Experience with all flavor of Hadoop distributions, including Cloudera, Hortonworks, Mapr and Apache.
* Experience in installation, configuration, supporting and managing Hadoop Clusters using Apache, Cloudera (5.X) distributions and on Amazon web services (AWS).
* Expertise in implementing Spark Scala application using higher order functions for both batch and interactive analysis requirement.
* Extensive experienced working with Spark tools like RDD transformations, spark MLlib and spark QL.
* Hands on experience in writing Hadoop Jobs for analyzing data using Hive QL (Queries), Pig Latin (Data flow language), and custom MapReduce programs in Java.
* Experienced in working with structured data using HiveQL, join operations, Hive UDFs, partitions, bucketing and internal/external tables.
* Extensive experience in collecting and storing stream data like log data in HDFS using Apache Flume.
* Experienced in using Pig scripts to do transformations, event joins, filters and some pre-aggregations before storing the data onto HDFS.
* Involvement in creating custom UDFs for Pig and Hive to consolidate strategies and usefulness of Python/Java into Pig Latin and HQL (HiveQL).
* Good Experience with NoSQL Databases like HBase, MongoDB and Cassandra.
* Experience on using Cassandra CQL with Java APIs to retrieve data from Cassandra tables.
* Hands on experience in querying and analyzing data from Cassandra for quick searching, sorting and grouping through CQL.
* Experience working with MongoDB for distributed storage and processing.
* Good knowledge and experienced in Extracting files from MongoDB through Sqoop and placed in HDFS and processed.
* Worked on importing data into HBase using HBase Shell and HBase Client API.
* Experience in designing and developing tables in HBase and storing aggregated data from Hive Table.
* Good knowledge in working with scheduling jobs in Hadoop using FIFO, Fair scheduler and Capacity scheduler.
* Experienced in designing both time driven and data driven automated workflows using Oozie and Zookeeper.
* Experience working on Solr for developing search engine on unstructured data in HDFS.
* Extensively used Solr to enable indexing for enabling searching on Non-primary key columns from Cassandra key spaces.
* Experience in writing stored procedures and complex SQL queries using relational databases like Oracle, SQL Server, and MySQL.
* Experience in Extraction, Transformation and Loading (ETL) of data from multiple sources like Flat files, XML files, and Databases.
* Supported various reporting teams and experience with data visualization tool Tableau.
* Implemented Data Quality in ETL Tool Talend and having good knowledge in Data Warehousing and ETL Tools like IBM DataStage, Informatica and Talend.
* Experienced and in-depth knowledge of cloud integration with AWS using Elastic Map Reduce (EMR), Simple Storage Service (S3), EC2, Redshift and Microsoft Azure.
* Designed end to end Azure cloud-based analytics dashboard for a state government for showing real time updates for the recently their state assembly elections 2016. Solution utilized PowerBI, Enterprise Gateway and Azure SQL Server.
* Software Development Life Cycle (SDLC) and strong knowledge in project implementation methodologies like Waterfall and Agile.

## **EDUCATION:**

Bachelor in Information Technology at JNTUH - 2013

Masters in Information Assurance at Willmington University - 2016

## **TOOLS AND TECHNOLOGIES:**

|  |  |
| --- | --- |
| **BigData/Hadoop Technologies** | Hadoop, HDFS, YARN, MapReduce, Hive, Pig, Impala, Sqoop, Flume, Spark, Kafka, Storm, Drill, Zookeeper, and Oozie. |
| **Language** | C, C++, Java, Python, Scala |
| **Application Servers** | Web Logic, Web Sphere, JBoss, Tomcat. |
| **Cloud Computing Tools** | Amazon AWS,S3, EMR, EC2, Lambda, VPC, Route 53, Cloud Watch |
| **Databases** | Microsoft SQL Server 2008 […] MySQL 4.x/5.x, Oracle 10g, 11g, 12c, DB2, Teradata, Netezza |
| **NO SQL Databases** | HBase, Cassandra, MongoDB, MariaDB. |
| **Build Tools** | Jenkins, Maven, Ant, Toad, SQL Loader, RTC, RSA, Control-M, Oozie, Hue, SOAP UI |
| **Modeling** | Rational Rose, Star UML, Visual paradigm for UML |
| **Reporting Tools** | MS Office (Word/Excel/PowerPoint/ Visio/Outlook), Crystal Reports XI, SSRS, Cognos7.0/6.0. |
| **Operating Systems** | All versions of UNIX, Windows, LINUX, Macintosh HD, Sun Solaris |

**PROFESSIONAL EXPERIENCE:**

|  |  |
| --- | --- |
| **Client: Fannie Mae** | **Mar 2020 - Till Date** |
| **Role: Big Data/Hadoop Developer** |  |

## **Responsibilities:**

* Writing Spark SQL in Scala for extracting data from AWS S3, transform it and load the data into AWS S3(Onelake).
* Enhancing External Web Application Server (Capital One Partnership Portal) using Java, JSP and servlets with underlying tomcat server.
* Developed Python application in AWS EMR to perform two necessary steps (quality check and tokenization) before loading Bluetarp business data into our centralized S3 buckets.
* Writing Spark jobs for end-to-end Settlement, Reward Files data cleansing, performing transformations and actions on data in Scala to convert from ASCII to PARQUET format to land the data in S3(Onelake) and Snowflake.
* Developed a spark framework for AWS cross account Data Movement in S3 using IAM Role Authentication.
* Responsible for data extraction and direct data ingestion framework with exception handling, row count validation for data quality checks and Secure Socket Layer Authentication to AWS Account.
* Developed Spark code to Attach Schema to the processed files using Scala based direct-Ingestion framework by making API call to NEBULA (Meta Data Management Tool) and passing the Technical Dataset ID for S3 location to store the processed data.
* Configured AWS SNS notification on TSYS input files, as soon as input files arrive in S3(Onelake) as part of Inbound Data-Ingestion framework, it triggers the AWS Lambda to kick start the Airflow DAG (Directed Acyclic Graph).
* Developed a config driven core agnostic Spark Scala framework which comprises of Reader, Dynamic Joiner, Dynamic Rule Resolver and Writer Modules, for producing Enrichment Layer (Account, Customer and Transaction) files in S3(Onelake).
* Migrating Single Sign-On Page from on-cloud instance to AWS Cloud Instance in Production.
* Built a resilient and automated CICD pipeline to make continuous development and continuous integration into production a much easier and secure process.
* Orchestration of all the workflows using Apache Airflow.
* Delivering Report functionality to partner users through COPP portal with AWS S3 Download/View Report Functionality in Backend.
* Developed secured connectivity from COPP-Capital One Partnership Portal to our AWS S3 staging buckets where reports are stored, using two-way authentication factor (Login credentials and SSO–Single Sign On).
* Writing Test Cases for the applications that are been built by us and perform various kinds of Testing like Integration, Functional, JUnit, Regression and Performance testing.
* Enhancing spark-based Tokenization application to change the default values in config file to dynamically assign values at the command line while running the job.

**Environment:** Hadoop 2.7, HDFS, Hive, Scala 2.11.12, Spark Core, Spark SQL, Python 3, Amazon Web Services (AWS), EMR, EC2, S3, SNS, Lambda, Snowflake, Java 8, JSP, Java Script, CICD, Jenkins, JUnit, GitHub, Apache Airflow.

|  |  |
| --- | --- |
| **Client: USAA** | **April 2018 – Feb 2020** |
| **Role: Big Data Engineer** |  |

**Responsibilities:**

* Evaluated business requirements and prepared detailed specifications that follow project guidelines required to develop written programs.
* Implemented Web Interfacing with Hive and stored the data in Hive tables.
* Loaded data from MySQL, a relational database to HDFS regularly using Sqoop Import/Export.
* Responsible for implementing Map Reduce programs into Spark transformations using Spark and Scala.
* Used Spark Streaming APIs to perform transformations and actions on the fly for building a common learner data model which gets the data from Kafka in near real-time and persists it to Cassandra.
* Responsible for the ingestion of data from various APIs and writing modules to store data in S3 buckets.
* Transformation of batch and stream data to encrypt fields and store in the data warehouse for ad-hoc query and analysis.
* Experience in developing ETL scripts for data acquisition and transformation using Informatica and Talend.
* Validating data fields from downstream sources to ensure uniformity of data.
* Converting ingested data (csv, XML, Json) to parquet file format in compressed form.
* Wrote Spark applications for data validation, cleansing, transformation, and custom aggregation.
* Strong Experience in Application/Platform Consolidation and Re-hosting, Legacy Conversion/Retirement, ETL&ELT data pipeline development for operational data stores and analytical warehouses.
* Imported real-time weblogs using Kafka as a messaging system and ingested the data to Spark Streaming.
* Expertise in using Flume in Collecting, aggregating, and loading log data from multiple sources into HDFS.
* Involved in Data Querying and Summarization using Pig and Hive and created UDF's, UDAF's and UDTF.
* Experienced in using Spark Core for joining the data to deliver the reports and for detecting fraudulent activities.
* Developed business logic using Kafka Direct Stream of Spark Streaming and implemented business transformations.
* Implemented a distributed messaging queue to integrate with Cassandra using Apache Kafka and Zookeeper.
* Involved in loading the real-time data to the NoSQL database like Cassandra.
* Involved in NoSQL (Datastax Cassandra) database design, integration and implementation and written scripts and invoked them using CQLSH.
* Designed columnar families in Cassandra and Ingested data from RDBMS, performed data transformations, and then exported the transformed data to Cassandra as per the business requirement.
* Implemented usage of Amazon EMR for processing Big Data across a Hadoop cluster of virtual servers on Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3).

**Environments**: Hadoop YARN, Spark-Core, AWS S3, AWS EMR, Spark-SQL, GraphX, Scala, Python, Kafka, Zeppelin, Jenkins, Docker, Microservices, Hive, Pig, Sqoop, Cassandra, Informatica, Cloudera, Oracle 12c, Linux.

|  |  |
| --- | --- |
| **Client: Tayota** | **Dec 2016 to March 2018** |
| **Role: Hadoop Developer** |  |

**Responsibilities:**

* Performed Data Profiling to learn about behavior with various features such as traffic pattern, location, and time, Date and Time etc. Integrating with external data sources and APIs to discover interesting trends.
* Evaluated business requirements and prepared detailed specifications that follow project guidelines required to develop written programs.
* Worked on Cloudera distribution for Hadoop ecosystem and installed and configured Flume, Hive, Pig, Sqoop and Oozie, Automatic on the Hadoop cluster.
* 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.
* Good working knowledge on Snowflake and Teradata databases.
* Extensively worked on Spark using Scala on cluster for computational (analytics), installed it on top of Hadoop performed advanced analytical application by making use of Spark with Hive and SQL/Oracle/Snowflake.
* Work with IT security auditors to resolve security vulnerabilities in Linux, UNIX, Apache.
* Delivered zero defect code for three large projects which involved changes to both front end (web services) and back-end (Oracle, snowflake, Teradata).
* Experience in Developing Spark applications using Spark - SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Created data pipelines migrating data from on premises servers to S3 to Glue to Athena and utilized by AWS Quick sight and Tableau.
* Strong Experience in implementing Data warehouse solutions in Confidential Redshift; Worked on various projects to migrate data from on premise databases to Confidential Redshift, RDS and S3.
* Implemented Continuous Integration using GIT and GitHub from scratch.
* Involved in all the stages of Software Development Life Cycle Primarily in Database Architecture, Logical and Physical modeling, Data Warehouse/ETL development using MS SQL Server 2012/2008R2/2008, Oracle 11g/10g, and ETL Solutions/Analytics Applications development.
* Experience with Unix/Linux systems with scripting experience and building data pipelines
* 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.
* Hands on experience in writing Python and Bash Scripts.
* 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.
* Experience in refactoring the existing spark batch process for different logs written in Scala.
* Implemented Big Data tools like Spark using Scala and utilizing Data frames and Spark SQL API for faster processing of data and worked on extensible framework for building high performance batch and interactive data processing application on hive.
* Debugging and maintenance of automaton test scripts in batch mode and implemented a plan on automation scripts on based on Sprint.
* Develop Oozie workflows to schedule the Scripts on daily basis.

# **Environments**: Hadoop/Big Data Technologies: Spark-Scala, Kafka, Spark Streaming, Mlib, Map Reduce, Pig, Hive, Cassandra, HBase, HDFS, MapReduce, Hive, Sqoop, Flume, Oozie

|  |  |
| --- | --- |
| **Client:** **COMCAST** | **April 2014 to Dec 2016** |
| **Role: Hadoop Developer** |  |

**Responsibilities:**

* Exported the analyzed data to the relational databases using Sqoop for visualization and to generate reports for the BI team.
* Developed complex Hive queries to process the data for visualizing.
* Used impala to pull the data from Hive tables.
* Worked with the Data Science team to gather requirements for various data mining projects
* Written Hive queries for data analysis to meet the business requirements and writing Hive scripts to extract, transform and load the data into Database.
* Developed Pig Scripts, Pig UDFs and Hive Scripts, Hive UDFs to process the data.
* Involved in migration of Map Reduce programs into Spark transformations.
* Integrated Oozie with the rest of the Hadoop stack supporting several types of Hadoop jobs out of the box (like, Pig, Hive, and Sqoop) as well as system specific jobs (such as Perl and shell script).
* Developed POC on real time streaming data received by Kafka and processed the data using Spark and this data was further stored into HDFS cluster using Java and Scala.
* Experienced in transferring data from different data sources into HDFS systems using Kafka producers, consumers and Kafka brokers.
* Experienced in performance tuning of Spark Application for setting right Batch Interval time, correct level of parallelism and memory tuning which persists into Cassandra.
* Developed Spark scripts by using Scala shell commands as per the requirement.
* Responsible for all the activities from requirement analysis, Quality control and coordination between functional and development team.
* Wrote shell scripts and cron jobs for rolling day-to-day processes and it is automated.
* Developed scripts and Auto sys Jobs to schedule a bundle (group of coordinators) which consists of various Hadoop programs using Oozie.
* Involved in migration of MapReduce programs into Spark transformations using Spark and Scala.

Proficient in using Cloudera Manager, an end to end tool to manage Hadoop services.

**Environment**: JDK1.6, HDFS, Map Reduce, Cassandra, Hive, Pig, HBase, Sqoop, Spark, Oozie, Impala, Cloudera, Scala, Oracle 11g, My SQL, AWS, Linux, SVN, GitHub