Job Seeker Plano, TX Cloudera Certified Hadoop and Spark developer with 4 years of experience in IT industry which includes design, development, testing and application support of various Web Based applications using Java, J2EE Technologies and Big Data Ecosystem. Around 2 years of comprehensive experience in Big Data processing and its ecosystem like HDFS, Hive, Sqoop, Oozie, Falcon, HBase, Yarn, Zookeeper, Attunity Replicate , Hue and Spark components. Excellent knowledge on Spark, Scala, Spark SQL, Spark Streaming, Flume, Pig ,Hadoop MapReduce concepts and NiFi. Around 1.5 years of experience in Software development using Java, J2EE(JSF and Chordiant Framework) and Oracle Database in various areas like Requirements gathering, Analysis, Design, Development, Testing, Implementation and Maintenance. Good hands-on knowledge in Hortonworks and Cloudera Hadoop Distributions. Experience in importing and exporting data using Sqoop from HDFS to Relational Database systems (RDBMS) and vice-versa. Expertise in loading different file based sources into Hadoop Landing zone. Hands on experience in developing SPARK applications using Spark API's like Spark transformations and actions (RDD) ,Spark MLlib and Spark SQL(Dataframes) using Scala and Python. Scheduled job workflow for FTP, Sqoop and hive scripts using Oozie coordinators. Designed and Developed Data Ingestion Framework using Java to ingest data from various RDBMS and file based sources and housed them in Hive tables from which reports are generated. Worked with various file formats such as delimited text files, click stream log files, Sequence Files, Avro files, ORC files, JSON files, Parquet and XML File formats. Excellent knowledge of Hadoop Architecture and various components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node and MapReduce Programming paradigm. Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Scala and have a good experience in using Spark-Shell and Spark Streaming. Extensive experience in working with structured data using Hive QL, join operations, Partitioning , Bucketing and experienced in optimizing Hive Queries. Experience in implementing Kerberos authentication protocol in Hadoop for data security. Used Scala SBT to develop Scala coded spark projects and executed using spark-submit. Having experience in analyzing large amounts of data sets to determine optimal way to aggregate and report on it using

Qlikview. Experienced with code versioning and dependency management systems such as Accurev and Maven. Monitoring Job failures and performing the root cause analysis and corrective action. Developed workflows for the batch jobs and scheduled using Oozie. Experience in data workflow management tools such as Falcon, Oozie and Autosys which is job scheduling software. Experience with cloud computing platforms like Amazon Web Services(AWS). Experience in developing applications using Java technologies include Core Java, J2EE, Java Server Pages (JSP), Servlets and Java Script. Well versed in using Software development methodologies like Agile and DevOps Methodology. Experience in preparing and executing unit test plan and unit test cases after software development using JUNIT. Highly motivated and passionate to learn and explore emerging technologies. Authorized to work in the US for any employer Work Experience Udemy(Self-Learning) November 2018 to Present Project: Similar Movies Prediction based on rating in Apache Spark GroupLens Research has collected and made available rating data sets from the MovieLens website. The data sets were collected over various periods of time, depending on the size of the set. For my project, I took a 20 million ratings and 465,000 tag applications applied to 27,000 movies by 138,000 users and analyzed them using Spark ,Spark SQL , Python, Scala on Amazon EMR to find the similar movies to each other based on the user ratings. Project Details: Worked on Cloudera distribution and deployed on AWS EC2 Instances. Analyzed movie rating dataset by building data ETL pipeline using PySpark, Scala and Spark SQL. Designed movies recommendation using Spark MLlib's collaborative filtering algorithm by parallel processing million of records on Amazon s3. Implemented Alternative Least Square model for customized movie recommendation. Migrated the historical data from movieLens site to Amazon s3 and cached the data in memory for repetitive use. Used Rating class for parsing the .dat file and applied transformation to create RDD objects. Used Matrix Factorization Model to make product predictions for users. Evaluated the performance of models and got recommendations successfully, proven boosted running speed by parallelism. 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). Good Understanding on Extracted Real time feed using Spark streaming and convert it to RDD and process data into Data Frame and load the data into HDFS. Worked on improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frame, Pair RDD's, Spark YARN. Environment: AWS, S3, Spark, Spark SQL, Hive, LINUX, Scala, UNIX Shell Scripting, Python, YARN, Cloudera Hadoop Developer Ford Motor Company - Dearborn, MI July 2016 to November 2017 Project Details: Data Supply Chain(DSC) is a Data lake on Hadoop platform to receive, consolidate data. DSC is the Ford information platform for analytics that balances the demands of data management and information access. It helps integrate structured, semi-structured and unstructured information into single logical information. The DSC Application supports the Global Data Insight and Analytic Skill team(GDIA) and their mission to drive evidence-based decision making, provide timely, actionable ,forward-looking insights to their business partners. Responsibilities: Developed a Custom Sqoop project based on Data Ingestion Framework using Java to ingest Incremental/Full load data from Teradata source into Hadoop Landing Zone based on timestamp partitions. Analyze and understanding the requirement given by downstream users. Hands on experience on extracting data from different databases ,file based sources and scheduled Oozie, falcon workflows to execute this job on daily and monthly basis. Responsible for loading various sources to HDFS using SQOOP and command line bash scripts. Prepared and executed HBase entries to run Oozie jobs in Adhoc. Created Hive tables to load the Data and stored as ORC files for processing. Implemented Hive Partitioning and bucketing for further classification of data. Involved in creating Hive Tables, loading with data and writing Hive queries which will invoke and run Map Reduce jobs in the backend. Setting up and worked on Kerberos authentication principals to establish secure network communication on cluster and testing of HDFS, Hive, Pig and MapReduce to access cluster for new users. Autosys scheduler to automate the jobs and time scheduling. Worked with No SQL databases like HBase and created HBase tables to load large sets of semi structured data coming from various sources. Used Attunity Replicate web based Interface tool to load data efficiently and quickly into HDFS from different sources. Performed data masking and special character removal tasks in the data transformation using SPARK. Been part of Design Reviews & Daily Project Scrums and sprint planning based on Agile methodology. Used Maven to build .jar files and Accurev as software configuration management tool. Used Rally as a work tracking tool and BMC for incident management. Worked with different file formats such as Text, Sequence files, Avro, ORC and Parquet. Involved in Qlikview development of business analytic and visualization reports for DSC Management Dashboard. Support/Troubleshoot hive programs running on the cluster and Involved in fixing issues arising out of duration testing. Environment: Hortonworks HDP 2.5.3 ,Hadoop(YARN), HDFS, Hive, Sqoop, Oozie, Falcon , Accurev ,LINUX, Hue, HBase, Zookeeper, Java , Maven, Autosys , Mainframe , Teradata , Shell scripting , Qlikview, Rally Edureka November 2015 to April 2016 Project Details: In Internet Telephony, a call detail record is a data record that contains information related to a telephone call, such as the origination and destination addresses of the call, the time the call started and ended, the duration of the call, the time of day the call was made and any toll charges that were added through the network or charges for operator services, among other details of the call. This project is to find out the customers who are facing frequent call drops in Roaming. This is a very important report which telecom companies uses to prevent customer churn out, by calling them back and at the same time contacting their roaming partners to improve the connectivity issues in specific areas. Responsibilities: Developed Spark code using Scala and Spark-SQL for faster testing and data processing as per the requirement. Installed and configured Spark, Hive, Pig, Sqoop and Oozie on the Hadoop cluster. Imported millions of structured data from relational databases using Sqoop import to process using Spark and stored the data into HDFS in CSV format. Used Spark SQL to process the huge amount of structured data. Developed traits and case classes etc in Scala implemented business logic using Scala 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. Used DataFrame API in Scala for converting the distributed collection of data organized into named columns. Registered the datasets as Hive Table. Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs and Scala. Used various spark Transformations and Actions for cleansing the input data. Experienced in handling large datasets using Partitions, Spark in Memory capabilities, Broadcasts in Spark, Effective & efficient Joins, Transformations and other during ingestion process itself. Responsible in performing sort, join, aggregations, filter, and other transformations on the datasets using Spark. Developed solutions to pre-process large sets of structured, with different file formats (Text file, Avro data files, Sequence files, Xml and JSON files, ORC and Parquet). Experienced with batch processing of data sources using Apache Spark. Environment: HDFS, YARN, Sqoop, Apache Spark, Spark-SQL, Cloudera CDH 5.X, Spark-shell, Hive, Hue, MYSQL Java/J2EE Tata Consultancy Services February 2013 to May 2014 Developer Project Details: Lloyds Banking Group is a major British financial institution formed through the acquisition of HBOS by Lloyds with millions of UK customers with a presence in nearly every community. Its business is focused on retail and commercial financial services. The project Savings Re-Engineering project is one of the big milestone deliveries which went on live in three different group releases working in agile methodology. The main requirements of this project were around Savings swim-lanes. It is the fundamental rethinking and radical redesign of business process to achieve dramatic improvements in critical measures of performance such as cost, service, and speed. Responsibilities: Involved in complete Software Development Life cycle (SDLC) of the project from Analysis, Design, Programming, Testing and Deploying the application. Used Java Server Faces (JSF) and Java Server Pages (JSP) for developing UI pages. Developed HTML+ Java script prototypes for the Savings re-engineering screens. Developed web application using Chordiant MVC Framework. Generated Web Service Client from a Web Services Description Language(WSDL). Used JUNIT for unit testing the code changes. Involved in testing of application on various levels like Unit and Integration testing. Provided support for SIT and UAT. Used Rational Clear Case / Clear Quest for source control and defect management. Submitted the unit tested code for Review and rectified the concerns raised. Interact with business people, SMEs, onshore partners and other supporting teams through mails, voice calls, video conference calls, online communicator chats to ensure the delivery of an optimized solution to the customer. Fixed the defects in SIT phase immediately and provided build with minimal turnaround time. Used Quality Center (QC) for requirements and bug tracking. Environment: Core Java, Chordiant MVC, JSF,JSP, Servlets, Oracle SQL, Windows XP/Vista, RAD, TOAD, WebSphereApplication Server Web Developer GEE Communication - IN June 2012 to January 2013 Project Details: Gee communication is part of a telecom company where they created a portal that helps to create and send E-Mail campaigns, track leads, automate sales-force activities and help in effective Customer Relationship Management (CRM). Responsibilities: Involved in developing easy-to-use user interface and step-by-step process that helps to create eye-catching E-Mail campaign. Design and developed web-based software using Java Server Faces(JSF) framework. Data has been handled using the JDBC connection. Java Beans were used to handle business logic as a Model and Servlets to control the flow of application as Controller. Proficient in developing responsive Front End components using HTML, CSS, JSP tags, JavaScript. Used IBM's WebSphere Application Server to deploy code. Involved in Code Reviews, Defect Fixing and UAT support. Familiarity with all aspects of Software Development Life Cycle. Make code modifications for the assigned projects according to business specifications and application standards Perform unit and system testing for all coding changes. Environment: HTML, CSS, Servlets, JSF, JSP, JUNIT, Oracle 11g, Eclipse, JavaScript, Core Java Education Bachelor's Skills Apache hadoop hdfs (2 years), Hadoop (2 years), Hadoop distributed file system (2 years), Java (3 years), Sql (3 years), Spark (Less than 1 year) Additional Information TECHNICAL SKILLS Hadoop Ecosystem HDFS, MapReduce, Hive, Sqoop, Oozie, Zookeeper, Spark, Falcon, Kafka, Attunity Replicate, Pig Spark Ecosystem Spark Core, Spark SQL, Spark MLlib Databases HBase(NoSQL),MYSQL, Teradata Programming Languages Java, Scala, Hive QL , Shell Scripting Framework Hadoop API, JSF Web Technologies HTML, Java script, CSS, JSP, Servlets, XML IDE/Interfaces Eclipse, IntelliJ , Spark-Shell, PySpark, JUNIT, Maven Methodologies Agile ,DevOps Operating Systems Windows, UNIX, LINUX Cloud Platform AWS EMR

Name: Shelby Rodriguez

Email: sarahcobb@example.org

Phone: +1-626-987-2325x668