

Sr. Hadoop/Spark Developer Sr. Hadoop/Spark Developer Sr. Hadoop/Spark Developer - JPMC
Jersey City, NJ Extensive IT experience of over 9 years with multinational clients which includes of Big Data related architecture experience developing Spark/Hadoop applications. Excellent understanding / knowledge of Hadoop architecture and various components such as HDFS, JobTracker, TaskTracker, NameNode, DataNode and MapReduce programming paradigm. Experience in all major components of Hadoop Eco-components such as HDFS, HIVE, PIG, Oozie, Sqoop, Map Reduce and YARN on Cloudera, MapR and Hortonworks distributions. Experience in tuning and troubleshooting performance issues in Hadoop cluster. Designing and creating HIVE external tables using shared meta-store instead of the derby with partitioning, dynamic partitioning and buckets. Experience in Oozie and workflow scheduler to manage Hadoop jobs by Direct Acyclic Graph (DAG) of actions with control flows. Experience in integrating Hive and HBase for effective operations. Developed the Pig UDF'S to pre-process the data for analysis. Experience working on different file formats like Avro, Parquet, ORC, Sequence and Compression techniques like Gzip, Lzo, and Snappy in Hadoop. Strong understanding of NoSQL databases and hands on work experience in writing applications on NoSQL databases like HBase, Cassandra and MongoDB, Redis, Neo4j. Experience in working on CQL (Cassandra Query Language), for retrieving the data present in Cassandra cluster by running queries in CQL. Proficient with Cluster management and configuring Cassandra Database. Experienced in working with Spark ecosystem using Spark-SQL and Scala queries on different data file formats like .txt, .csv etc. Implemented POC to migrate Map Reduce jobs into Spark RDD transformations using SCALA. Have good experience in creating real time data streaming solutions using Apache Spark/Spark Streaming/Apache Storm, Kafka and Flume. Working knowledge on major Hadoop ecosystems PIG, HIVE, Sqoop, and Flume. Good experience in Cloudera, Hortonworks & Apache Hadoop distributions. Knowledge on AWS (Amazon EC2) Hadoop distribution. Developed high-throughput streaming apps reading from Kafka queues and writing enriched data back to outbound Kafka queues. Wrote and worked on complex performance improvements on PL/SQL queries, stored procedures, triggers, indexes with databases like MySQL and Oracle. Also, working towards improvement of knowledge on

No-SQL databases like MongoDB. Experience on NoSQL databases including HBase, Cassandra.

Hands-on experience in scripting skills in Python, Linux and UNIX Shell. Good working experience using Sqoop to import data into HDFS from RDBMS and vice-versa. Knowledge on creating Solr collection configuration to scale up the infrastructure. Experience in developing web-based applications using Python. Experience in application development using Java, J2EE, EJB, Hibernate, JDBC, Jakarta Struts, JSP and Servlets. Experience in using various IDEs Eclipse, My Eclipse and repositories SVN and CVS. Experience of using build tools Ant and Maven. Working with relative ease with different working strategies like Agile, Waterfall and Scrum methodologies. Excellent communication and analytical skills and flexible to adapt to evolving technology. Authorized to work in the US for any employer

Work Experience

Sr. Hadoop/Spark Developer JPMC - Chicago, IL April 2018 to Present

Roles & Responsibilities:

- Responsible for building scalable distributed data solutions using Hadoop.
- Responsible for managing and scheduling Jobs on a Hadoop cluster.
- Loading data from UNIX file system to HDFS and vice versa.
- Improving the performance and optimization of existing algorithms in Hadoop using Spark context, Spark-SQL and Spark YARN.
- Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Python and Scala.
- Worked with Apache Spark for large data processing integrated with functional programming language Scala.
- Developed POC using Scala, Spark SQL and MLlib libraries along with Kafka and other tools as per requirement then deployed on the Yarn cluster.
- 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.
- Implemented Data Ingestion in real time processing using Kafka.
- Developed Spark code using Scala and Spark-SQL/Streaming for faster processing of data
- Configured Spark Streaming to receive real time data and store the stream data to HDFS.
- Developed Spark scripts by using Scala shell commands as per the requirement
- Used Spark API over Cloudera Hadoop YARN to perform analytics on data in Hive.
- Documented the requirements including the available code which should be implemented using Spark, Hive, HDFS and SOLR.
- Tested Apache TEZ, an extensible framework for building high performance batch and interactive data processing applications, on Pig

and Hive jobs. Used Kafka Streams to Configure Spark streaming to get information and then store it in HDFS. Developed multiple Kafka Producers and Consumers as per the software requirement specifications. 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.

Real time streaming the data using Spark with Kafka. Responsible for creating Hive tables and working on them using Hive QL. Implementing various Hive UDF's as per business requirements.

Exported the analyzed data to the databases using Sqoop for visualization and to generate reports for the BI team. Involved in Data Visualization using Tableau for Reporting from Hive Tables.

Developed Python Mapper and Reducer scripts and implemented them using Hadoop Streaming.

Developed multiple Map Reduce jobs in java for data cleaning and preprocessing. Optimized Map Reduce Jobs to use HDFS efficiently by using various compression mechanisms. Responsible for writing Hive queries for data analysis to meet the business requirements. Customized Apache Solr to handle fallback searching and provide custom functions. Responsible for setup and benchmarking of Hadoop/HBase clusters.

Environment: Hadoop, HDFS, HBase, Sqoop, Hive, Map Reduce, Spark- Streaming/SQL, Scala, Kafka, Solr, Sbt, Java, Python, Ubuntu/Cent OS, MySQL, Linux, GitHub, Maven, Jenkins.

Scala Developer T-Mobile - Atlanta, GA May 2016 to March 2018

Responsibilities: Creating end to end Spark-Solr applications using Scala to perform various data cleansing. Involved in converting Hive/SQL queries into Spark transformations using Spark RDD'S and Scala. Developed Spark scripts by using Scala shell commands as per the requirement. Used Akka as a framework to create reactive, distributed, parallel and resilient concurrent applications in Scala. Used Slick to query and storing in database in a Scala fashion using the powerful Scala collection framework. Developed POC using Scala & deployed on Yarn cluster, compared the performance of Spark, with Hive and SQL. Deployed and maintained multi-node Dev and Test Kafka Clusters. Using Spark API over Cloudera Hadoop YARN to perform analytics on data in Hive. Developed Scala scripts, UDFFs using both Data frames/SQL and RDD/MapReduce in Spark 1.6 for Data Aggregation, queries and writing data back into OLTP system through Sqoop and Developed enterprise application using Scala as well Performed

advanced procedures like text analytics and processing, using the in-memory computing capabilities of Spark using Scala. Developed an equivalent Spark Scala code for existing SAS code to extract summary insights on the hive tables. Implemented Spark using Scala and utilizing Data frames and Spark SQL API for faster processing of data. Implemented the Data Bricks API in Scala program to push the processed data to Redshift DB. Redshift is columnar and compressed storage, scale linearly and seamlessly. Worked on the performance tuning of spark data frames for aggregation using dynamic partition, creating the temp views needed. Developed Spark applications for the entire batch processing by using Scala. Developed Spark scripts by using Scala shell commands as per the requirement. Environment: Hive, Flume, Java, Maven, Impala, Spark, Oozie, Oracle, Yarn, GitHub, Junit, Tableau, Unix, Cloudera, Flume, Sqoop, HDFS, Tomcat, Java, Scala, Hbase. BigData Developer GE - Waukesha, WI October 2014 to April 2016 Roles & Responsibilities: Involved in Automation of clickstream data collection and store into HDFS using Flume. Involved in creating Data Lake by extracting customer's data from various data sources into HDFS. Used Sqoop to load data from Oracle Database into Hive. Developed MapReduce programs to cleanse the data in HDFS obtained from multiple data sources. Implemented various Pig UDF's for converting unstructured data into structured data. Developed Pig Latin scripts for data processing. Optimizing of existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frames and Pair RDD's. Load the data into Spark RDD and performed in-memory data computation to generate the output response. Developed the Apache Spark, Flume, and HDFS integration project to do a real-time data analysis. Developed data pipeline using Flume, Spark and Hive to ingest, transform and analyzing data. Wrote Flume configuration files for importing streaming log data into MongoDB with Flume. Performed masking on customer sensitive data using Flume interceptors. Used IMPALA to analyze data ingested into Hive tables and compute various metrics for reporting on the dashboard. Implemented Spark using Scala and utilizing Data frames and Spark SQL API for faster processing of data. 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 WEB HDFS REST API to make the HTTP GET, PUT, POST and DELETE requests from the webserver to perform analytics on the data lake. Involved in creating Hive tables as per requirement defined with appropriate static and dynamic partitions. Used Hive to analyze the data in HDFS to identify issues and behavioral patterns. Involved in production Hadoop cluster set up, administration, maintenance, monitoring and support. Logical implementation and interaction with HBase. Assisted in creation of large HBase tables using large set of data from various portfolios. Cluster coordination services through Zookeeper. Efficiently put and fetched data to/from HBase by writing MapReduce job. Developed MapReduce jobs to automate transfer of data from/to HBase. Assisted with the addition of Hadoop processing to the IT infrastructure. Used flume to collect the entire web log from the online ad-servers and push into HDFS. Implemented custom business logic by writing UDF's in Java and used various UDF's from Piggybank and other sources. Implemented MapReduce job and execute the MapReduce job to process the log data from the ad-servers. Load and transform large sets of structured, semi structured and unstructured data. Performing analysis using high level languages like Python. Launching Amazon EC2 cloud instances using Amazon images and configuring launched instances with respect to specific applications. Back-end Java developer for Data Management Platform (DMP) and building RESTful APIs to build and let other groups build dashboards. Environment: Hadoop, Pig, Sqoop, Oozie, MapReduce, HDFS, Hive, Java, Python, Eclipse, HBase, Flume, AWS, Oracle 10g, UNIX Shell Scripting, GitHub, Maven. Java/J2EE Developer Capital Coast Inc - Columbia, MD January 2012 to September 2014

Roles & Responsibilities: Involved in writing programs for XA transaction management on multiple databases of the application. Developed java programs, JSP pages and servlets using Struts framework. Involved in creating database tables, writing complex TSQL queries and stored procedures in the SQL server. Worked with AJAX framework to get the asynchronous response for the user request and used JavaScript for the validation. Used EJBs in the application and developed Session beans to implement business logic at the middle tier level. Actively involved in writing SQL using SQL Query Builder. Involved in coordinating the on-shore/Off-shore

development and mentoring the new team members. Extensively Used Ant tool to build and configure J2EE applications and used Log4J for logging in the application Used JAXB to read and manipulate the xml properties. Used JNI for calling the libraries and other implemented functions in C language. Used prototype MooTools and script.aculo.us for fluid User Interface. Involved in fixing defects and unit testing with test cases using JUnit. Environment: Java, EJB, Servlets, XSLT, CVS, J2EE, AJAX, Struts, Hibernate, ANT, Tomcat, JMS, UML, Log4J, Oracle 10g, Eclipse, Solaris, JUnit and Windows 7/XP, Maven. Java Developer Choice Perficient - St. Louis, MO April 2010 to December 2012 Roles & Responsibilities: Played an active role in the team by interacting with business and program specialists and converted business requirements into system requirements. Conducted Design reviews and Technical reviews with other project stakeholders. Implemented Services using Core Java. Involved in development of classes using java. Good proficiency in developing algorithms for serial interfaces. Involved in testing of CAN protocols. Developed the flow of algorithm in UML. Used Servlets to implement Business components. Designed and Developed required Manager Classes for database operations Developed various Servlets for monitoring the application. Designed and developed the front end using HTML and JSP Developed XML files, DTDs, Schema's and parsing XML by using both SAX and DOM parser. Wrote deployment descriptors using XML and Test java classes for a direct testing of the Session and Entity beans. Did Packaging and Deployment of builds through ANT script. Wrote stored procedure and used JAVA APIs to call these procedures. Database designing that includes defining tables, views, constraints, triggers, sequences, index, and stored procedures. Developed verification and validation scripts in java. Followed verification and validation cycle for development of algorithms. Developed Test cases for Unit Test cases and as well as System and User test scenarios. Involved in Unit Testing, User Acceptance Testing and Bug Fixing. Environment: Java, JSP, Servlets, JDBC, JavaScript, MySQL, JUnit, Eclipse IDE, Windows 7/XP/Vista, UNIX, LINUX. Education Bachelor's Skills Cassandra, Hdfs, Mapreduce, Oozie, Sqoop Certifications/Licenses Driver's License

Name: George Lucero

Email: anthonygarrett@example.net

Phone: 3422688379