Sr. Spark Developer Sr. Spark Developer Sr. Spark Developer - CenterPoint Energy 8 years of extensive hands-on experience in IT industry including 5+ years experience in deployment of Hadoop Ecosystems like MapReduce, Yarn, Sqoop, Flume, Pig, Hive, HBase, Cassandra, Zoo Keeper, Oozie, and Ambari, BigQuery, Big Table and 5+ years experience on Spark, Storm, Scala, Python. Experience in OLTP and OLAP design, development, testing, implementation and support of enterprise Data warehouses. Strong Knowledge in Hadoop Cluster Capacity Planning, Performance Tuning, Cluster Monitoring Extensive experience in business data science project life cycle including Data Acquisition, Data Cleaning, Data Manipulation, Data Validation, Data Mining, Machine Learning Algorithms, and Visualization Good Hands on experience in working with Ecosystems like Hive, Pig, Sqoop, Map Reduce, Flume, Oozie. Strong knowledge in HIVE and PIG core functionality by using custom User Defined Function's (UDF), User Defined Table-Generating Functions (UDTF) and User Defined Aggregating Functions (UDAF) for Hive. Experience on Productionizing Apache Nifi. for dataflows with significant processing requirements and controlling security of data flow. Designed and developed RDD Seeds using Scala and Cascading. Streaming data to Sparkstreaming using Kafka Exposure to Spark, Spark Streaming, Spark MLlib, Scala and Creating the Data Frames handled in Spark with Scala. Good Exposure on Map Reduce programming using Java, PIG Latin Scripting and Distributed Application and HDFS. Experienced Good understanding of NoSQL databases and hands on work experience in writing applications No SQL Databases HBase, Cassandra and MongoDB. Very good implementation experience of Object Oriented concepts, Multithreading and Java/Scala Experienced with the Scala, Spark improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark -SQL, Pair RDD's, Spark YARN. Experienced in installation, configuration, supporting and managing Hadoop Clusters using Apache Cloudera distributions, Horton works, Cloud Storage and Amazon web services (AWS) and related technologies DynamoDB, EMR, S3, ML. Experience in deploying NiFi Data flow in Production team and Integrating data from multiple sources like Cassandra, MongoDB. Deploying templates to environments can be done via NiFi RestAPI integrated with other automation tools Complete end to end design and development of

Apache NiFi flow which acts as the agent between middleware team and EBI team and executes all the actions mentioned above Experienced in Python programming, wrote Web Crawlers using Python. Experience in bench marking Hadoop cluster for analysis of queue usage Experienced in working with Mahout for applying machine learning techniques in the Hadoop Ecosystem. Good Experience on Amazon Web Services like Redshift, Data Pipeline, ML. Good experienced on moving the data in and out of Hadoop RDBMS, No-SQL and UNIX from various systems using SQOOP and other traditional data movement technologies. Experience on Integration of Quartz scheduler with Oozie work flows to get data from multiple data sources in parallel using fork. Experience in installation, configuration, support and management of a Hadoop Cluster using Cloudera Distributions. Experienced Spark scripts by using Scala shell as per requirements. Good knowledge on tuning the Spark jobs by changing the configuration properties and using broadcast variables. Developed REST APIs using Java, Play framework and Akka. Expertise in search technology's like SOLR, Informatica & Lucene. Experience in converting SQL queries into Spark Transformations using Spark RDDs and Scala and Performed map-side joins on RDD's. Experienced in writing Hadoop Jobs for analyzing data using Hive Query Language (HQL), Pig Latin (Data flow language), and custom MapReduce programs in Java. Good understanding of NoSQL databases like MongoDB, Cassandra, and HBase. Strong analytical skills with ability to quickly understand client's business needs. Involved in business meetings for requirements gathering form business clients. Experienced in Storm builder topologies to perform cleansing operations before moving data into HBase. Hands on experience in configuring and working with Flume to load the data from multiple sources directly into Hdfs. Experience on configuring fully the Flume agent, suitable for all type of logger data and store them in Avro Sink in Parquet file format and developing 2-tier architecture connecting channels between Avro sinks and Source. Experience creating Visual report, Graphical analysis and Dashboard reports using Tableau, Informatica of historical data saved in Hdfs and data analysis using Splunk enterprise edition. Good experience in utilizing Cloud Storage Services like Git. Extensive knowledge in using GitHub and Bit Bucket. Experienced in job scheduling and monitoring using Oozie, Zookeeper. Work Experience Sr. Spark

Developer CenterPoint Energy - Houston, TX June 2018 to Present Responsibilities: Hands on experience in installation, configuration, supporting and managing Hadoop Clusters. Knowledge of Cassandra security, maintenance and tuning both database and server. Chipped away at outlining and building up the Real Time Analysis module for Analytic Dashboard utilizing Cassandra, Kafka, Spark Streaming. Installed and configured Confluent Kafka in R&D line. Validated the installation with HDFS connector and Hive connectors. Deployed high availability on the Hadoop cluster quorum journal nodes. Experience on implementing SAX (Symbolic Aggregate approXimation) in Java to use with Apache Spark for normalizing time series data. Involved in defining job flows, managing and reviewing log file. Set-up configured and optimized the Cassandra cluster. Developed real-time Spark based application to work along with the Cassandra database. Responsible to manage data coming from different sources through Kafka. Installed Kafka Producer on different severs and Scheduled to produce data for every 10 seconds Integrated Kafka with Spark Streaming to listen onto multiple Kafka Brokers with different Kafka topics for every 5 Seconds. Enhanced and optimized product Spark code to aggregate, group and run data mining tasks using the Spark framework and handled Json Data. Handled Json Data comes from Kafka Direct Stream on each partitions and transformed them into required Data Frame Formats. Upgraded Spark 1.6 to latest Version Spark 2.2 and configure Kafka Version 0.10. Managing Kafka Offsets, Saving Offsets in external data base like HBase and to its own Kafka. Worked on Import & Export of data using ETL tool Sqoop from MySQL to HDFS. Worked on Lambda Architecture for both Batch processing and Real Streaming purposes. Used Oozie to Schedule Spark and Kafka Producer Jobs to run in parallel. Appended the Data Frames into Cassandra Key Space Tables using DataStax Spark-Cassandra Connector. Experience with Cassandra YAML, Configuration files, RACK DC properties file, Cassandra-env file for JMX configurations etc. Installed and configured Datastax OpsCenter and Nagios for Cassandra cluster maintenance and alert. Configured Authentication and security in Apache kafka pub-sub system. Good experience with Century Link Cloud for provisioning virtual machines, creating resource groups, configuring key vaults for storing encryption keys, Monitoring etc. Great Hands on Experience in seat stamping

Hadoop bunch for investigation of line utilization Performing OS level setups and Kernel level tuning Implement and test integration of BI (Business Intelligence) tools with Hadoop stack. Installed/Configured/Maintained Apache Hadoop clusters for application development and Hadoop tools like Hive, Pig, HBase, Zookeeper, Sqoop, Yarn, Spark2, Kafka and Oozie. Formulated procedures for installation of Hadoop, Spark2 patches, updates and version upgrades. Environment: Cloudera, HDFS, Spark, Hive, Pig, Map Reduce, Hue, Sqoop, Putt, Apache Kafka, Apache Drill, Century Link Cloud, AWS, Java Netezza, Cassandra, Oozie, Spark , SPARK SQL, Maven, SBT, Java, Scala, SQL and Linux, YARN, Agile Methodology, Solr, PHP Admin, XAMPP, DataStax Cassandra. Sr. Hadoop/Spark Developer CPS Energy - San Antonio, TX January 2017 to June 2018 Responsibilities: Involved in deploying systems on Amazon Web Services (AWS) Infrastructure services EC2. Experience in configuring, deploying the web applications on AWS servers using SBT and Play. Migrated Map Reduce jobs into Spark RDD transformations using Scala. Used Spark API over Cloudera Hadoop YARN to perform analytics on data in Hive. Developed Spark code using Spark RDD and Spark-SQL/Streaming for faster processing of data. Performed configuration, deployment and support of cloud services including Amazon Web Services (AWS). Working knowledge of various AWS technologies like SQS Queuing, SNS Notification, S3 storage, Redshift, Data Pipeline, EMR. Responsible for all Public (AWS) and Private (Openstack/VMWare/DCOS/Mesos/Marathon) cloud infrastructure Developed Flume ETL job for handling data from HTTP Source and Sink as HDFS and configuring Data Pipelining. Used Hive data warehouse tool to analyze the unified historic data in HDFS to identify issues and behavioral patterns. Involved in Developing a Restful service using Python Flask framework. Expertise in working with Python GUI frameworks - PyJamas, Jython. Experienced in using Apache Drill data-intensive distributed applications for interactive analysis of large-scale datasets. Developed end to end ETL batch and streaming data integration into Hadoop(MapR), transforming data. Used Python modules such as requests, urllib, urllib2 for web crawling. Tools developed extensively include Spark, Drill, Hive, HBase, Kafka & MapR Streams, PostgreSQL, Stream Sets. Used Hive Queries in Spark-SQL for analysis and processing the data. Worked as a key role in a

team of developing an initial prototype of a NiFi big data pipeline. This pipeline demonstrated an end to end scenario of data ingestion, processing. Used HUE for running Hive queries. Created Partitions according to day using Hive to improve performance. Wrote Python routines to log into the websites and fetch data for selected options. Worked on custom Pig Loaders and storage classes to work with variety of data formats such as JSON and XML file formats. Loaded some of the data into Cassandra for fast retrieval of data. Worked in provisioning and managing multi-tenant Hadoop clusters on public cloud environment - Amazon Web Services (AWS) and on private cloud infrastructure - Open stack cloud platform and worked on DynamoDB, Ml. Worked on large-scale Hadoop YARN cluster for distributed data processing and analysis using Data Bricks Connectors, Spark core, Spark SQL, Sqoop, Pig, Hive, Impala and NoSQL databases. Created HBase tables to load large sets of structured, semi-structured and unstructured data coming from UNIX, NoSQL and a variety of portfolios. Worked on a POC to compare processing time of Impala with Apache Hive for batch applications to implement the former in project. Worked with various HDFS file formats like Avro, Sequence File and various compression formats like Snappy, bzip2. Used the RegEx, JSON and Avro for serialization and de-serialization packaged with Hive to parse the contents of streamed log data. Converted all the vap processing from Netezza and implemented by using Spark data frames and RDD's. Worked in writing Spark Sql scripts for optimizing the query performance. Responsible for handling different data formats like Avro, Parquet and ORC formats. Implemented Spark Scripts using Scala, Spark SQL to access hive tables into Spark for faster processing of data. Environment: Cloudera, Horton Works distribution, HDFS, Spark, Hive, Pig, Map Reduce, Hue, Sqoop, Putty, HaaS (Hadoop as a Service), Apache Kafka, Apache Mesos and the AWS, Java Netezza, Cassandra, Oozie, Spark, SPARK SQL, Maven, Java, Scala, SQL and Linux, Toad, YARN, Agile Methodology. Hadoop Developer BANK of America - Dallas, TX May 2016 to December 2016 Responsibilities: Concerned and well-informed on Hadoop Components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node, YARN and Map Reduce programming. Developed Map-Reduce programs to get rid of irregularities and aggregate the data. Developed Cluster coordination services through Zookeeper. Implemented

Hive UDF's and did performance tuning for better results Developed Pig Latin Scripts to extract data from log files and store them to HDFS. Created User Defined Functions (UDFs) to pre-process data for analysis Implemented Optimized Map Joins to get data from different sources to perform cleaning operations before applying the algorithms. Created highly optimized SQL queries for MapReduce jobs, seamlessly matching the query to the appropriate Hive table configuration to generate efficient report. Used other packages such as Beautifulsoup for data parsing in Python. Tuned, and developed SQL on HiveQL, Drill and SparkSQL Experience in using Sqoop to import and export the data from Oracle DB into HDFS and HIVE, HBase. Implemented CRUD operations on HBase data using thrift API to get real time insights. Developed workflow in Oozie to manage and schedule jobs on Hadoop cluster for generating reports on nightly, weekly and monthly basis. Worked on integration independent microservices for real-time bidding (scala/akka, firebase, cassandra, Elasticsearch) Used slick to query and storing in database in a Scala fashion using the powerful Scala collection framework Using HIVE processed extensively ETL loadings on a Structured Data. Defined job flows and developed simple to complex Map Reduce jobs as per the requirement. Optimized Map/Reduce Jobs to use HDFS efficiently by using various compression mechanisms. Developed PIG UDFs for manipulating the data according to Business Requirements and also worked on developing custom PIG Loaders. Created various Parser programs to extract data from Autosys, Tibco Business Objects, XML, Informatica, Java, and database views using Scala PIG UDF was required to extract the information of the area from the huge data which we get from the sensors. Responsible for creating Hive tables based on business requirements. Implemented Partitioning, Dynamic Partitions and Buckets in HIVE for efficient data access. Involved in NoSQL database design, integration and implementation. Loaded data into NoSQL database HBase. Worked on debugging, performance tuning PIG and HIVE scripts by understanding the joins, group and aggregation between them. Used Flume to collect, aggregate and store the web log data from different sources like web servers and pushed to HDFS. Connected the hive tables to Data analyzing tools like Tableau for Graphical representation of the trends. Experienced in managing and reviewing Hadoop log files. Involved in loading data from

UNIX file system to HDFS. Responsible for design & development of Spark SQL Scripts based on Functional Specifications. Used Apache HUE interface to monitor and manage the HDFS storage.Concerned and well-informed on Hadoop Components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node, YARN and Map Reduce programming. Developed Map-Reduce programs to get rid of irregularities and aggregate the data. Developed Cluster coordination services through Zookeeper. Implemented Hive UDF's and did performance tuning for better results Developed Pig Latin Scripts to extract data from log files and store them to HDFS. Created User Defined Functions (UDFs) to pre-process data for analysis Implemented Optimized Map Joins to get data from different sources to perform cleaning operations before applying the algorithms. Created highly optimized SQL queries for MapReduce jobs, seamlessly matching the query to the appropriate Hive table configuration to generate efficient report. Used other packages such as Beautifulsoup for data parsing in Python. Tuned, and developed SQL on HiveQL, Drill and SparkSQL Experience in using Sqoop to import and export the data from Oracle DB into HDFS and HIVE, HBase. Implemented CRUD operations on HBase data using thrift API to get real time insights. Developed workflow in Oozie to manage and schedule jobs on Hadoop cluster for generating reports on nightly, weekly and monthly basis. Worked on integration independent microservices for real-time bidding (scala/akka, firebase, cassandra, Elasticsearch) Used slick to query and storing in database in a Scala fashion using the powerful Scala collection framework Using HIVE processed extensively ETL loadings on a Structured Data. Defined job flows and developed simple to complex Map Reduce jobs as per the requirement. Optimized Map/Reduce Jobs to use HDFS efficiently by using various compression mechanisms. Developed PIG UDFs for manipulating the data according to Business Requirements and also worked on developing custom PIG Loaders. Created various Parser programs to extract data from Autosys, Tibco Business Objects, XML, Informatica, Java, and database views using Scala PIG UDF was required to extract the information of the area from the huge data which we get from the sensors. Responsible for creating Hive tables based on business requirements. Implemented Partitioning, Dynamic Partitions and Buckets in HIVE for efficient data access. Involved in NoSQL database design,

integration and implementation. Loaded data into NoSQL database HBase. Worked on debugging, performance tuning PIG and HIVE scripts by understanding the joins, group and aggregation between them. Used Flume to collect, aggregate and store the web log data from different sources like web servers and pushed to HDFS. Connected the hive tables to Data analyzing tools like Tableau for Graphical representation of the trends. Experienced in managing and reviewing Hadoop log files. Involved in loading data from UNIX file system to HDFS. Responsible for design & development of Spark SQL Scripts based on Functional Specifications. Used Apache HUE interface to monitor and manage the HDFS storage. Environment: HDFS, Map Reduce, Pig, Mesos, AWS Hive, Sqoop, Scala, Flume, Mahout, HBase, Spark, SPARK SQL, Yarn, Java, Maven, Git, Cloudera, MongoDB, Eclipse and Shell Scripting. Hadoop Developer DELL - Bengaluru, Karnataka June 2013 to August 2015 Responsibilities: Designed and developed data movement framework for multiple sources like SQL Server, Oracle, and MySQL. Created Sqoop import and export jobs for multiple sources. Developed scripts to automate the creation Sqoop jobs for various workflows. Developed Hive scripts to alter the tables and perform required transformations. Developed a java MapReduce and PIG cleansers for data cleansing. Worked on Hive UDFS to mask confidential information in the data. Designed and developed MapReduce programs for data lineage. Designed and developed the framework to log information for auditing and failure recovery. Closed worked with the web application development team to develop the user interface for data movement framework. Designed Oozie workflows for Job Automation. Created Map Reduce programs to handle semi/unstructured data like xml, Json, Avro data files and sequence files for log files. A RESTful web service, built with python and cherrypy, retrieves data from an accumulo data warehouse Maintaining the MySQL server and Authentication to required users for Databases. Appropriately documented various Administrative technical issues. Developed MapReduce programs to extract and transform the data sets and results were exported back to RDBMS using Sqoop. Built a RESTful web service for storing and retrieving documents in an apache accumulo data store Involved in collecting and aggregating large amounts of log data using Apache Flume and staging data in HDFS for further analysis. Optimized our Hadoop

infrastructure at both the Software and Hardware level. Experience in troubleshooting in MapReduce jobs by reviewing log files. Developed end-to-end search solution using web crawler, Apache Nutch & Search Platform, Apache SOLR. Environment:Hadoop, Cloudera Manager, Linux, RedHat, CentOs, Ubuntu Operating System, Scala, Map Reduce, HBase, SQL, Sqoop, HDFS, Kafka, UML, Apache SOLR, Hive, Oozie, Cassandra, maven, Pig, UNIX, Python, MR Unit, Git. Java Developer STATE FARM - Bengaluru, Karnataka July 2011 to June 2013 Responsibilities: Developed the J2EE application based on the Service Oriented Architecture by employing SOAP and other tools for data exchanges and updates. Developed the functionalities using Agile Methodology. Used Apache Maven for project management and building the application. Worked in all the modules of the application which involved front-end presentation logic developed using Spring MVC, JSP, JSTL and JavaScript, Business objects developed using POJOs and data access layer using Hibernate framework. Used JAX-RS (REST) for producing web services and involved in writing programs to consume the web services with Apache CXF framework. Used Restful API and SOAP web services for internal and external consumption. Used Spring ORM module for integration with Hibernate for persistence layer. Involved in writing Hibernate Query Language (HQL) for persistence layer. Used Spring MVC, Spring AOP, Spring IOC, Spring Transaction and Oracle to create Club Systems Component. Wrote backend jobs based on Core Java & Oracle Data Base to be run daily/weekly. Coding the core modules of the application compliant with the Java/J2EE coding standards and Design Patterns. Written Java Script, HTML, CSS, Servlets, and JSP for designing GUI of the application. Worked on Service-side and Middle-tier technologies, extracting catching strategies/solutions. Design data access layer using Data Access Layer J2EE patterns. Implementing the MVC architecture Struts Framework for handling databases across multiple locations and display information in presentation layer. Used XPath for parsing the XML elements as part of business logic processing. Environment: Java, Struts 1.2, Hibernate 3.0, JSP, JavaScript, HTML, XML, Oracle, Eclipse, JBoss Application Server, ANT, CVS, and SQL. Education Masters Michigan State University - Ann Arbor, MI Skills Hdfs, Impala, Oozie, Sqoop, Apache kafka, Kafka, Db2, Flume, Jboss, Jms, Map reduce, Mongodb, Ms

visual studio, Visual studio, Apache spark, Api, Hive, Html, Javascript, Node.js Additional Information TECHNICAL SKILLS: Big Data Ecosystems HDFS and Map Reduce, Pig, Hive, Pig Latin, Impala, YARN, Oozie, Zookeeper, Apache Spark, Apache Crunch, Apache NiFi, Apace STORM, Apache Kappa, Apache Kafka, Sqoop, Flume. Streaming Technologies Spark Streaming, Storm Scripting Languages Python, Perl, Shell, Sheme, Tcl, Unix Shell Scripts, Windows Power Shell Programming Languages Java, J2EE, JDK1.4/1.5/1.6/1.7/1.8, JDBC, Hibernate, XML Parsers, JSP 1.2/2, Servlets, EJB, JMS, Struts, Spring Framework, Java Beans, AJAX, JNDI. Databases MongoDB, Netezza, SQL Server, MySQL, ORACLE, DB2 IDEs / Tools Eclipse, JUnit, Maven, Ant, MS Visual Studio, Net Beans Methodologies Agile, Waterfall Virtualization Technologies VMware ESXi, Windows Hyper-V, Power VM, Virtual box, Citrix Xen, KVM. Web Technologies HTML, JavaScript, JQuery, Ajax, Boot Strap, Angular JS, Node.js, Express.js Web Servers Web Logic, Web Sphere, Apache Tomcat, JBOSS. Web Services SOAP, RESTful API, WSDL

Name: Michelle Mcdaniel

Email: darren93@example.net

Phone: 997-717-6528x832