Spark/Kafka Developer Spark/Kafka Developer Spark/Kafka Developer - Citizens Property Insurance Corporation Florida City, FL ? 10+ years of experience in IT, which includes experience in Bigdata Technologies, Hadoop ecosystem, Data Warehousing, SQL related technologies in Retail, Manufacturing, Financial and Communication sectors ? 5 Years of experience in Big Data Analytics using Various Hadoop eco-systems tools and Spark Framework and Currently working on Spark and Spark Streaming frameworks extensively using Scala as the main programming dialect ? Experience installing/configuring/maintaining Apache Hadoop clusters for application development and Hadoop tools like Sqoop, Hive, PIG, Flume, HBase, Kafka, Hue, Storm, Zoo Keeper, Oozie, Cassandra, Sqoop, Python ? Worked with major distributions like Cloudera (CDH 3&4) & Horton works Distributions and AWS. Also worked on Unix and DWH in support for various Distributions ? Hands on experience in developing and deploying enterprise-based applications using major components in Hadoop ecosystem like Hadoop 2.X, YARN, Hive, Pig, MapReduce, Spark, Kafka, Storm, Oozie, HBase, Flume, Sqoop and Zookeeper ? Experience in handling large datasets using Partitions, Spark in memory capabilities, Broadcasts in Spark with Scala and python, Effective and efficient Joins, Transformations and other during ingestion process itself ? Experience in developing data pipeline using Pig, Sqoop, and Flume to extract the data from weblogs and store in HDFS and accomplished developing Pig Latin Scripts and using HiveQL for data analytics ? Extensively dealt with Spark Streaming and Apache Kafka to fetch live stream data. ? Experience in converting Hive/SQL queries into Spark transformations using Java and experience in ETL development using Kafka, Flume and Sqoop ? Good experience in writing Spark applications using Scala and Java and used Scala set to develop Scala projects and executed using Spark-Submit ? Experience working on NoSQL databases including HBase, Cassandra and MongoDB and experience using Sqoop to import data into HDFS from RDBMS and vice-versa ? Developed Spark scripts by using Scala shell commands as per the requirement ? Good experience in writing Sqoop queries for transferring bulk data between Apache Hadoop and structured data stores ? Substantial experience in writing Map Reduce jobs in Java, PIG, Flume, Zookeeper, Hive and Storm ? Created multiple Map Reduce Jobs using Java API, Pig and Hive for data extraction ? Strong expertise in

troubleshooting and performance fine-tuning Spark, Map Reduce and Hive applications ? Good experience on working with Amazon EMR framework for processing data on EMR and EC2 instances ? Created AWS VPC network for the installed Instances and configured security groups and Elastic IP's Accordingly ? Developed AWS Cloud formation templates to create custom sized VPC, subnets, EC2 instances, ELB and security groups ? Extensive experience in developing applications that perform Data Processing tasks using Teradata, Oracle, SQL Server and MySQL database ? Worked on data warehousing and ETL tools like Informatica, Tableau, and Pentaho ? Experience in understanding the security requirements for Hadoop and integrate with Kerberos authentication and authorization infrastructure ? Acquaintance with Agile and Waterfall methodologies. Responsible for handling several clients facing meetings with great communication skills Work Experience Spark/Kafka Developer Citizens Property Insurance Corporation February 2018 to Present Responsibilities: ? Involved in complete Big Data flow of the application starting from data ingestion from upstream to HDFS, processing and analyzing the data in HDFS. ? Developed Spark API to import data into HDFS from Teradata and created Hive tables. ? Developed Sqoop jobs to import data in Avro file format from Oracle database and created hive tables on top of it. ? Created Partitioned and Bucketed Hive tables in Parquet File Formats with Snappy compression and then loaded data into Parquet hive tables from Avro hive tables. ? Involved in running all the hive scripts through hive, Impala, Hive on Spark and some through Spark SQL. ? Development and Review of spark code containing Airflow DAG's, Databricks Notebooks, Delta Tables in DDL's and Metadata SQL's, other SQL scripts. ? Deploying the Code to Dev, QA, PreProd and Prod Environments by adhering to GIT process flow and following the standards mentioned by the release management process. ? Creating Technical Design Documentation and Support/OPS Turnover documentation by following the OPS checklist. ? Raising Change Request once the code is PreProd. ? Airflow Orchestration especially configuring the DAG start date and scheduled time and other parameters. ? Worked on mainly developing Pyspark code in Databricks code using existing load patterns(Full, Incremental and Backfill) for forecasting(Region and Country) rawCustomerSales and pubCustomerSales. ? Wrote Spark Dataframes that uses mainly CSV files,

Parquet, Delta file formats. Used Spark SQL, Joins, views, partitioning extensively. ? Validating the source data and generating the output data in the required format using Pyspark transformations ? Submitting Jobs for cluster administered by other Linux teams. ? Environment: Used Databricks, Azure Data Lake storage(Gen1), Oracle EDW, PySpark mainly & Spark SQL, Scala Spark occasionally , Jenkins , PyCharm, Git, Spark BDA server, Putty for Tunneling into Airflow environments etc., ? Involved in performance tuning of Hive from design, storage and query perspectives. ? Developed Flume ETL job for handling data from HTTP Source and Sink as HDFS. ? Collected the Json data from HTTP Source and developed Spark APIs that helps to do inserts and updates in Hive tables. ? Developed Spark scripts to import large files from Amazon S3 buckets. ? Developed Spark core and Spark SQL scripts using Scala for faster data processing. ? Developed Kafka consumer's API in Scala for consuming data from Kafka topics. ? Involved in designing and developing tables in HBase and storing aggregated data from Hive Table. ? Integrated Hive and Tableau Desktop reports and published to Tableau Server. ? Developed shell scripts for running Hive scripts in Hive and Impala. ? Orchestrated number of Sqoop and Hive scripts using Oozie workflow and scheduled using Oozie coordinator. ? Used Jira for bug tracking and Bit Bucket to check-in and checkout code changes. ? Continuous monitoring and managing the Hadoop cluster through Cloudera Manager. ? Environment: HDFS, Yarn, Map Reduce, Hive, Sqoop, Flume, Oozie, HBase, Kafka, Impala, Spark SQL, Spark Streaming, Eclipse, Oracle, Teradata, PL/SQL UNIX Shell Scripting, Cloudera. Environment: HDFS, Yarn, Map Reduce, Hive, Sqoop, Flume, Oozie, HBase, Kafka, Impala, SparkSQL, Spark Streaming, Eclipse, Oracle, Teradata, PL/SQL UNIX Shell Scripting, Cloudera Hadoop/ Java Developer Experian - Chicago, IL March 2016 to January 2018 Responsibilities: ? Responsible for architecting Hadoop clusters with CDH3 and involved in installation of CDH3 and up gradation to CDH4 from CDH3 ? Worked on creating Key space in Cassandra for saving the Spark Batch output ? Worked on Spark application to compact the small files present into hive ecosystem to make it equivalent to block size of HDFS ? Manage migration of on-perm servers to AWS by creating golden images for upload and deployment ? Manage multiple AWS accounts with multiple VPC's for both production and non-production where primary

objectives are automation, build out, integration and cost control ? Implemented the real time streaming ingestion using Kafka and Spark Streaming ? Loaded data using Spark-streaming with Scala and Python ? 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 SparkScala 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). ? Involved in requirement and design phase to implement Streaming Lambda Architecture to use real time streaming using Spark and Kafka and Scala ? Experience in loading the data into Spark RDD and performing in-memory data computation to generate the output responses ? Migrated complex map reduce programs into In-memory Spark processing using Transformations and actions ? Developed full text search platform using NoSQL and Logstash

Elastic Search engine, allowing for much faster, more scalable and more intuitive user searches ? Developed the Sqoop scripts to make the interaction between Pig and MySQL Database ? Worked on Performance Enhancement in Pig, Hive and HBase on multiple nodes ? Worked with Distributed n-tier architecture and Client/Server architecture ? Supported Map Reduce Programs those are running on the cluster and developed multiple Map Reduce jobs in Java for data cleaning and pre-processing ? Developed MapReduce application using Hadoop, MapReduce programming and HBase ? Evaluated usage of Oozie for Workflow Orchestration and experienced in cluster coordination using Zookeeper ? Developing ETL jobs with organization and project defined standards and processes ? Experienced in enabling Kerberos authentication in ETL process ? Implemented data access using Hibernate persistence framework ? Design of GUI using Model View Controller Architecture (STRUTS Framework) ? Integrated Spring DAO for data access using Hibernate and involved in the Development of Spring Framework Controllers Environment: Hadoop 2.X, HDFS, MapReduce, Hive, Pig, Sqoop, Oozie, HBase, Java, J2EE, Eclipse, HQL. Hadoop Developer BMW - Woodcliff Lake October 2013 to February 2016 Responsibilities: ? Involved in the Complete Software development life cycle (SDLC) to develop the application. ? Worked on analyzing Hadoop cluster using different big data analytic tools including Pig, Hive and Map Reduce on EC2. ? Worked with the Data Science team to gather requirements for various data mining projects. ? Worked with different source data file formats like JSON, CSV, and TSV etc. ? Experience in importing data from various data sources like MySQL and Netezza using Sqoop, SFTP, performed transformations using Hive, Pig and loaded data back into HDFS. ? Performed transformations, cleaning and filtering on imported data using Hive, Map Reduce. ? Import and export data between the environments like MySQL, HDFS and deploying into productions. ? Used Pig as ETL tool to do transformations, event joins and some pre-aggregations before storing the data onto HDFS. ? Worked on partitioning and used bucketing in HIVE tables and setting tuning parameters to improve the performance. ? Involved in developing Impala scripts to do Adhoc queries. ? Experience in Oozie workflow scheduler template to manage various jobs like Sqoop, MR, Pig, Hive, Shell scripts, etc. ? Involved in importing and exporting data from HBase using

Spark. ? Involved in POC for migrating ETLS from Hive to Spark in Spark on Yarn Environment. ? Actively participating in the code reviews, meetings and solving any technical issues. Environment: Apache Hadoop, AWS, EMR, EC2, S3, Horton works, Map Reduce, Hive, Pig, Sqoop, Apache Spark, Zookeeper, HBase, Java, Oozie, Oracle, MySQL, Netezza and UNIX Shell Scripting. Hadoop Developer WNS Global Services - Pune, Maharashtra June 2011 to September 2013 Responsibilities: ? Installed and configured Hadoop MapReduce, HDFS, Developed multiple MapReduce jobs in java for data cleaning and preprocessing ? Installed and configured Apache Hadoop to test the maintenance of log files in Hadoop cluster ? Importing and exporting data into HDFS and Hive using Sqoop ? Experienced in defining job flows and managing and reviewing Hadoop log files ? Load and transform large sets of structured, semi structured and unstructured data ? Responsible to manage data coming from different sources and for implementing MongoDB to store and analyze unstructured data ? Supported Map Reduce Programs those are running on the cluster and involved in loading data from UNIX file system to HDFS ? Installed and configured Hive and written Hive UDFs ? Involved in creating Hive tables, loading with data and writing hive queries that will run internally in map reduce way ? Involved in Hadoop cluster task like Adding and Removing Nodes without any effect to running jobs and data ? Created HBase tables to store variable data formats of PII data coming from different portfolios ? Implemented best income logic using Pig scripts ? Load and transform large sets of structured, semi structured and unstructured data ? Cluster coordination services through Zookeeper ? Exported the analyzed data to the relational databases using Sqoop for visualization and to generate reports for the BI team ? Supported in setting up QA environment and updating configurations for implementing scripts with Pig and Sqoop ? Continuous monitoring and managing the Hadoop cluster using Cloudera Manager ? Used Hibernate ORM framework with Spring framework for data persistence and transaction management and involved in templates and screens in HTML and JavaScript Environment: Hadoop, HDFS, MapReduce, Pig, Sqoop, UNIX, HBase, Java, JavaScript, HTML SQL/ Java Developer ILogix IT Services Pvt Ltd - Hyderabad, Telangana July 2009 to May 2011 Responsibilities: ? Importing the data from the MySQL and Oracle into the HDFS using Sqoop. ?

Implemented CDH3 Hadoop cluster on Centos. Worked on installing clusters, commissioning & decommissioning of data node, name node recovery, capacity planning, and slots configuration. ? Developing parser and loader map reduce application to retrieve data from HDFS and store to HBase and Hive. ? Importing the unstructured data into the HDFS using Flume. ? Written Map Reduce java programs to analyze the log data for large-scale data sets. ? Involved in creating Hive tables, loading and analyzing data using hive queries. ? Involved in using HBase Java API on Java application. ? Automated all the jobs for extracting the data from different Data Sources like MySQL to pushing the result set data to Hadoop Distributed File System. ? Vodafone is a British Multinational Telecommunications Services Company headquartered in Swindon (UK). ? Initially it has started with an infrastructure company which supplies infrastructure and cables to other companies. After that it has started Telecommunications Company. So they want to communicate with different components. ? Project Description ? In this project all the Network elements for activating the network service will be stored in InventoryMangementSystem(IMS).Using the procedure IMS will allocate, deallocate, modify, cease the Service based on order request send by other components and request will be in the form of xml.Response also will be sent as xml to other components ? InventoryMangementSystem GUI is used to add any new Network elements and more features are added to display the details of network elements.InventoryManagementGUI is designed with MVC (Spring & Hibernate framework) architerature. ? All the code changes are deployed in Weblogic server. ? Responsibilities I was working as a Team Member ? Developed Pig Latin scripts to extract the data from the output files to load into HDFS. ? Responsible for managing data from multiple sources. ? Created and maintained Technical documentation for launching HADOOP Clusters and for executing Hive queries and Pig Scripts. ? Designed and built many applications to deal with vast amounts of data flowing through multiple Hadoop clusters, using Java-based map-reduce. ? Worked with application teams to install operating system, Hadoop updates, patches, version upgrades as required. Environment: Hadoop 1.0.0, Map Reduce, Hive, HBase, Flume, Sqoop, Pig, Zookeeper, Java, ETL, SQL, Centos. Skills Cassandra, Hdfs, Impala, Mapreduce, Oozie, Sqoop, Hbase, Kafka, Flume, Hadoop, Jboss, Mongodb, Nosql, Teradata,

Visual studio, Apache spark, Application server, Git, Hadoop, Hbase Additional Information Skills: Big Data Technologies HDFS, MapReduce, Hive, Pig, Sqoop, Flume, Oozie, Zookeeper, Kafka, Cassandra, Apache Spark, Spark Streaming, HBase, Flume, Impala Hadoop Distribution Cloudera, Horton Works, Apache, AWS Languages Java, SQL, PL/SQL, Python, Pig Latin, HiveQL, Scala, Regular Expressions Web Technologies HTML, CSS, JavaScript, XML, JSP, Restful, SOAP Operating Systems Windows (XP/7/8/10), UNIX, LINUX, UBUNTU, CENTOS. Portals/Application servers WebLogic, WebSphere Application server, WebSphere Portal server, JBOSS Build Automation tools SBT, Ant, Maven Version Control GIT IDE & Build Tools, Design Eclipse, Visual Studio, Net Beans, Rational Application Developer, Junit Databases Oracle, SQL Server, MySQL, MS Access, NoSQL Database (HBase, Cassandra, MongoDB), Teradata.

Name: David Perry

Email: dennisjamie@example.net

Phone: 347.685.5379x197