Hadoop/Spark Developer Hadoop/Spark Developer Hadoop/Spark Developer - Edward Jones Boston, MA Around 8+ years of professional experience in Information Technology which includes 5 years in Big Data and Hadoop Ecosystem. Experience in working with BI team and transform big data requirements into Hadoop centric technologies. Expert in creating indexes, views, complex stored procedures, user-defined functions, cursors, derived tables, common table expressions (CTEs) and triggers to facilitate efficient data manipulation and data consistency. Excellent understanding/knowledge of Hadoop architecture and various components of Big Data. Hands on experience in installing, configuring, and using Hadoop ecosystem components like MapReduce, HDFS, HBase, Oozie, Hive, Sqoop, Pig, and Zookeeper. Experience in analyzing data using HQL, Pig Latin and custom MapReduce programs in Java. Knowledge on various file formats such as Avro, Parquet, ORC, etc. and on various compression codecs such as GZip, Snappy, LZO etc. Strong competency in HIVE schema design, partitions and bucketing. Experience in ingestion, storage, querying, processing and analysis of Big Data with hands-on experience in Apache Spark and Spark Streaming. Owned the design, development and maintenance of ongoing metrics, reports, analyses, dashboards, etc., to drive key business decisions and communicate key concepts to readers. Expertise in designing and developing a distributed processing system running into a Data Warehousing platform for reporting. Data cleaning, pre-processing and modelling using Spark and Python. Experience in performing ETL on top of streaming log data from various web servers into HDFS using Flume. Performed data analytics and insights using Impala and Hive. Expert in developing and scheduling jobs using Oozie and Crontab. Hands-on Git, Agile (Scrum), JIRA and Confluence. Authorized to work in the US for any employer Work Experience Hadoop/Spark Developer Edward Jones - Boston, MA March 2019 to Present Responsibilities: Developed Spark jobs written in Scala to perform operations like data aggregation, data processing and data analysis. Load the data into Spark RDD and performed in-memory data computation to generate the output response. Developed Hive UDF's for extended use and wrote HiveQL for sorting, joining, filtering and grouping the structure data. Used Spark for series of dependent jobs and for iterative algorithms. Developed a data pipeline using Kafka and Spark Streaming to store

data into HDFS. Developed workflow in Oozie to automate the tasks of loading the data into HDFS. Developed data pipeline using Flume, Sqoop, Pig and Java MapReduce to ingest customer behavioral data and financial histories into HDFS for analysis. Developed ETL Applications using Hive, Spark, and Impala & Sqoop for Automation using Oozie. Used Pig as ETL tool to do transformations, event joins and some pre-aggregations before storing the data onto HDFS. Used Spark for series of dependent jobs and for iterative algorithms. Developed a data pipeline using Kafka and Spark Streaming to store data into HDFS. Hadoop Developer/Engineer Navy Federal Credit Union - Vienna, VA September 2018 to March 2019 Responsibilities: Automated the process for extraction of data from warehouses and weblogs by developing work-flows and coordinator jobs in Oozie. Assisted in upgrading, configuration and maintenance of various Hadoop infrastructures like Pig, Hive, and HBase. Developed Spark jobs written in Scala to perform operations like data aggregation, data processing and data analysis. Load the data into Spark RDD and performed in-memory data computation to generate the output response. Implemented various MapReduce Jobs in custom environments and updating them to HBase tables by generating hive queries. Responsible for Cluster maintenance, adding and removing cluster nodes, Cluster Monitoring and Troubleshooting, manage and review data backups and log files. Developed workflow in Oozie to manage and schedule jobs on Hadoop cluster to trigger daily, weekly and monthly batch cycles. Used Spark for series of dependent jobs and for iterative algorithms. Developed a data pipeline using Kafka and Spark Streaming to store data into HDFS. Performance Tuning for Hive and Pig Job's performance parameters along with native MapReduce parameters to avoid excessive disk spills, enabled temp file compression between jobs in the data pipeline to handle production size data in a multi-tenant cluster environment. Developed data pipeline using Flume, Sqoop, Pig and Java MapReduce to ingest customer behavioral data and financial histories into HDFS for analysis. Environment: Hadoop HDFS, Flume, CDH, Pig, Hive, Oozie, Zookeeper, HBase, Spark, Storm, Spark SQL, NoSQL, Scala, Kafka, MongoDB, Linux, Sqoop Hadoop Developer Bank of America - Charlotte, NC September 2017 to September 2018 Responsibilities: Worked on implementation and data integration in developing large-scale system

software experiencing with Hadoop ecosystem components like HBase, Sqoop, Zookeeper, Oozie, Hive and Pig. Developed Hive UDF's for extended use and wrote HiveQL for sorting, joining, filtering and grouping the structure data. Developed ETL Applications using Hive, Spark, and Impala & Sqoop for Automation using Oozie. Used Pig as ETL tool to do transformations, event joins and some pre-aggregations before storing the data onto HDFS. Automated the process for extraction of data from warehouses and weblogs by developing workflows and coordinator jobs in Oozie. Developed workflow in Oozie to automate the tasks of loading the data into HDFS. Creating Hive tables, dynamic partitions, buckets for sampling, and working on them using HiveQL. Used Sqoop for importing the data into HBase and Hive, exporting result set from Hive to MySQL using Sqoop export tool for further processing. Enumerated Hive queries to do analysis of the data and to generate the end reports to be used by business users. Worked on scalable distributed computing systems, software architecture, data structures and algorithms using Hadoop, Apache Spark and Apache Storm etc. and ingested streaming data into Hadoop using Spark, Storm Framework and Scala. Implemented POCs with Spark SQL to interpret complex JSON records. Experience in transferring Streaming data, data from different data sources into HDFS and NoSQL databases using Apache Flume. Developed Spark jobs written in Scala to perform operations like data aggregation, data processing and data analysis. Used Kafka, Flume for building robust and fault tolerant data Ingestion pipeline between JMS and Spark Streaming Applications for transporting streaming web log data into HDFS. Used Spark for series of dependent jobs and for iterative algorithms. Developed a data pipeline using Kafka and Spark Streaming to store data into HDFS. Environment: Hadoop HDFS, Flume, CDH, Pig, Hive, Oozie, Zookeeper, HBase, Spark, Storm, Spark SQL, NoSQL, Scala, Kafka, MongoDB Big Data/Hadoop Developer Capital One - Dallas, TX August 2016 to September 2017 Responsibilities: Developed Scala, UDF's using both Data frames/SQL and RDD/MapReduce in Spark for Data Aggregation, queries and writing data back into RDBMS through Sqoop. Exported analyzed data to relational databases using Sqoop in deploying data from various sources into HDFS and building reports using Tableau. Exported analyzed data to relational databases using Sqoop for visualization to generate reports for the BI

team. Used the JSON and Avro for serialization and deserialization packaged with Hive to parse the contents of streamed log data and implemented Hive custom UDF's. Configured Spark streaming to receive real time data from Kafka and store the stream data to HDFS using Scala. Involved in developing ETL data pipelines for performing real-time streaming by ingesting data into HDFS and HBase using Kafka and Storm. Involved in moving log files generated from varied sources to HDFS, further processing through Flume. Involved in creating Hive tables by using Impala and working on them using HiveQL and perform data analysis using Hive and Pig. Developed workflow in Oozie to manage and schedule jobs on Hadoop cluster to trigger daily, weekly and monthly batch cycles. Assisted in upgrading, configuration and maintenance of various Hadoop infrastructures like Pig, Hive, and HBase. Worked on Apache Flume for collecting and aggregating huge amount of log data and stored it on HDFS for doing further analysis. Load the data into Spark RDD and performed in-memory data computation to generate the output response. Efficiently put and fetched data to/from HBase by writing MapReduce job. Environment: Hadoop, Flume, Kafka, Spark, Sqoop, Spark SQL, Spark-Streaming, Hive, Scala, pig, NoSQL, Impala, Oozie, HBase, Zookeeper. Hadoop Consultant AT&T - Dallas, TX June 2015 to August 2016 Responsibilities: Worked on loading the customer's data and event logs from Kafka into HBase using REST API. Responsible for Cluster maintenance, adding and removing cluster nodes, Cluster Monitoring and Troubleshooting, manage and review data backups and log files. Implemented various MapReduce Jobs in custom environments and updating them to HBase tables by generating hive queries. Developed data pipeline using Flume, Sqoop, Pig and Java MapReduce to ingest customer behavioral data and financial histories into HDFS for analysis. Collecting and aggregating large amounts of log data using Apache Flume and staging data in HDFS for further analysis. Created Hive tables from JSON data using data serialization framework using Avro, Parquet File formats and Snappy compression. Implemented generic export framework for moving data from HDFS to RDBMS and vice-versa. Worked on analyzing Hadoop cluster using different big data analytic tools including Kafka, Pig Hive and Map Reduce (MR1 and MR2). Automated all the jobs for pulling data from FTP server to load data into Hive tables using

Oozie workflows. Involved using HCATALOG to access Hive table metadata from Pig code. Implemented SQL, PL/SQL Stored Procedures. Actively involved in code review and bug fixing for improving the performance. Worked on tuning the performance Pig queries and involved in loading data from LINUX file system to HDFS. Importing and exporting data into HDFS using Sqoop and Kafka. Created HBase tables to store various data formats of PII data coming from different portfolios, Implemented Map-reduce for loading data from Oracle database. Used NoSQL database with HBase and MongoDB. Exported the result set from Hive to MySQL using Shell scripts. Gained experience in managing and reviewing Hadoop log files. Involved in scheduling Oozie workflow engine to run multiple Pig jobs. Environment: Hadoop, HDFS, HBase, Pig, Hive, Spark, Hortonworks, Oozie, MapReduce, Sqoop, MongoDB, Kafka, LINUX, Java APIs, Java collection. Hadoop Developer Sprint - Overland Park, KS August 2013 to April 2015 Responsibilities:

Worked extensively on Hadoop Components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node, YARN and MapReduce programming. Involved in loading data from UNIX file system to HDFS. Imported and exported data into HDFS and Hive using Sqoop. Experience in developing batch processing framework to ingest data into HDFS, Hive and Cassandra. Worked on Hive and Pig extensively to analyze network data in collecting metrics for Hadoop clusters. Automation of data pulls from SQL Server to Hadoop for analyzing large amounts of data sets to determine optimal way to aggregate and report on it. Provided quick response for client requests and experienced in creating ad hoc reports. Performance Tuning for Hive and Pig Job's performance parameters along with native MapReduce parameters to avoid excessive disk spills, enabled temp file compression between jobs in the data pipeline to handle production size data in a multi-tenant cluster environment. Designed workflows and coordinators in Oozie to automate and parallelize Hive and Pig jobs on Apache Hadoop environment by Hortonworks. Developed a process for the Batch ingestion of CSV Files, Sqoop from different sources and also generating views on the data source using Shell Scripting and Python. Delivered Hadoop migration strategy, roadmap and technology fitment for importing real time network log data into HDFS. POCs on moving existing Hive / Pig Latin jobs to Spark for Deploying and configuring agents to stream log

events into HDFS for analysis. Load the data into Hive tables using HiveQL along with Deduplication and Windowing to generate ad-hoc reports using Hive to validate customer viewing history and debug issues in production. Experienced with multiple Input Formats such as Text File, Key Value, Sequence File input format load to HDFS. Worked on business specific custom UDF's in Hive and Pig for data extraction, transformation and aggregation from multiple file formats including XML, JSON, CSV and other compressed file formats. Environment: HDFS, MapReduce, Pig, Hive, Oozie, Sqoop, Cassandra, Hortonworks. Java Developer Blue Cross Blue Shield - Detroit, MI February 2012 to July 2013 Responsibilities: Experience using middleware architecture using Java technologies like J2EE, Servlets, and application servers like Web Sphere and Web logic. Worked on loading data from Linux file system to HDFS. Understanding and analyzing the requirements. Designed, developed and validated User Interface using HTML, Java Script, and XML. Involved with writing SQL queries using Joins and Stored Procedures using Maven to build and deploy the applications in JBoss application Server in Software Development Lifecycle Model. Worked on Eclipse IDE for front end development environment for insertions, updating and retrieval operations of data from oracle database by writing stored procedures. Developed MapReduce jobs to convert data files into Parquet file format and included MRUnit to test the correctness of MapReduce programs. Experienced in working with various kinds of datasets for structured, semi structured and unstructured data with Teradata and Oracle for successfully loading files to HDFS from Teradata and loaded from HDFS to Hive. Installed Oozie workflow engine to run multiple Hive. Developed Hive queries to process the data and generate the data cubes for visualizing Concatenated ETL logics from RDBMS to Hive. Implemented partitioning, bucketing and worked on Hive, using file formats and compressions techniques with optimizations. Computed various metrics using Map Reduce to calculate metrics that define user experience, revenue etc. Environment: Hadoop, HDFS, Pig, Oozie, Hive, Python, MapReduce, Java, SQL Scripting and Linux Shell Scripting, Cloudera, Cloudera Manager. Software Developer Amazon - Hyderabad, Telangana September 2010 to November 2011 Responsibilities: Worked on Business logic for web service using spring annotations which enables dependency injection. Developed Spring Application

Framework for Dependency Injection, support for the Data Access Object (DAO) pattern and integrated with Hibernate ORM. Developed user interface for designing and developing the application using Java, JEE and spring core using JSP, JavaScript, Ajax, jQuery, HTML, CSS and JSTL. Outlining agile methodology with daily scrums using TDD and continuous integration in the SDLC process and used JIRA for bug tracking and task management. Developed Talend jobs to populate the claims data to data warehouse - star schema. Used Jenkins for continuous integration purpose in using SVN, Junit and Mockito as version control and Unit testing by Creating design documents and test cases for development work. Environment: Java, Servlets, JSP, HTML, CSS, Talend, Ajax, JavaScript, Hibernate, Spring, WebLogic, JMS, REST, SVN Education Bachelor of Technology in Computer Science and Engineering in Computer Science and Engineering Gokaraju Rangaraju Institute of Engineering and Technology - Hyderabad, Telangana Skills Cassandra, Hdfs, Impala, Mapreduce, Oozie Military Service Branch: United States Navy Rank: second Additional Information Technical Skills: Languages: Java, Python, Scala, HiveQL. Hadoop Ecosystem: HDFS, Hive, MapReduce, HBase, YARN, Sqoop, Flume, Oozie, Zookeeper, Impala. Databases: Oracle, RDBMS, DB2, SQL Server, MySQL. NoSQL Databases: HBase, MongoDB, Cassandra. Scripting Languages: JavaScript, CSS, Python, Perl, Shell Script.

Name: Samuel Perez

Email: taylormiller@example.org

Phone: 946.913.2012x9272