Spark Developer Spark Developer Spark Developer - State Farm Work Experience Spark Developer State Farm - Bloomington, IL August 2018 to Present Responsibilities Worked under the Cloudera distribution CDH 5.13 version. Worked on Ingesting weblog data into HDFS using Kafka. Used Spark SQL to process JSON data. Performed Cleansing the data to get a desired format. Wrote Spark Sql Data frames into Parquet Files. Worked on Tuning Spark Jobs for optimal Efficiency. Wrote the Python functions, procedures, Constructors and Traits. Created Hive tables to load the transformed Data. Performed partitions and bucketing in hive for easy data classification. Involved in Analyzing data by writing queries using HiveQL for faster data processing. Involved in working with Sqoop for loading the data into RDBMS. Created a data pipeline using oozie which runs on daily basis. Involved in Persisting Metadata into HDFS for further data processing. Loading data from Linux Filesystems to HDFS and vice-versa. Involved in creating tables, partitioning, bucketing of table and creating UDF's along with fine tuning in Hive. Loaded the Cleaned Data into the hive tables and performed some analysis based on the requirements. Responsible in performing sort, join, aggregations, filter, and other transformations on the datasets. Utilized Agile and Scrum Methodology to help manage and organize a team of developers with regular code review sessions. Environment: HDFS, Apache Spark, Apache Hive, Python, Oozie, Flume, Kafka, Agile, Methodology Cloudera, Cassandra. Hadoop/Spark Developer Caterpillar - Chicago, IL February 2017 to July 2018 Responsibilities: Worked on Cluster size of 150-200 nodes. Responsible for building scalable distributed data solutions using Hadoop Worked on migrating Map Reduce programs into Spark transformations using Spark and Python. Using Spark-Streaming APIs to perform transformations and actions on the fly for building the common learner data model which gets the data from Kafka in near real time and Persists into Cassandra. Developed Spark scripts by using Python shell commands as per the requirement. Used Spark API over Cloudera Hadoop YARN to perform analytics on data in Hive. Developed Python 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 Python as well Expertise in performance tuning of Spark Applications for setting

right Batch Interval time, correct level of Parallelism and memory tuning. Loaded the data into Spark RDD and do in memory data Computation to generate the Output response. Experience and hands-on knowledge in Akka and LIFT Framework. Used PostgreSQL and No-SQL database and integrated with Hadoop to develop datasets on HDFS Involved in creating partitioned Hive tables, and loading and analysing data using hive queries, Implemented Partitioning and bucketing in Hive. Worked on a POC to compare processing time of Impala with Apache Hive for batch applications to implement the former in project. Developed Hive queries to process the data and generate the data cubes for visualizing Implemented schema extraction for Parquet and Avro file Formats in Hive. Good experience with Talend open studio for designing ETL Jobs for Processing of data. Experience designing, reviewing, implementing and optimizing data transformation processes in the Hadoop and Talend /Informatica ecosystems. Implemented Partitioning, Dynamic Partitions, Buckets in HIVE. Coordinated with admins and Sr. Technical staff for migrating Teradata to Hadoop and Ab Initio to Hadoop as well Configured Hadoop clusters and coordinated with Big Data Admins for cluster maintenance. Environment: Hadoop YARN, Spark-Core, Spark-Streaming, Spark-SQL, Python, Kafka, Hive, Sqoop, Amazon AWS, Elastic Search, Impala, Cassandra, Tableau, Informatica, Cloudera, Oracle 10g, Linux. Hadoop Developer Tachyon Technologies LLC October 2015 to November 2016 Responsibilities Used Flume as a data pipeline system to ingest the unstructured events from various web servers to HDFS. We altered the unstructured events from web servers on the fly using various flume interceptors. Wrote various spark transformations using Python to perform data cleansing, validation and summarization activities on user behavioral data. Parsed the unstructured data into semi-structured format by writing complex algorithms in spark using Python. Developed generic parser to transform any format of unstructured data into a consisted data model. Configured Flume using Python with the Spark Streaming to transfer the data into HDFS at regular intervals of time from web servers to process the data. Implemented the persistence of frequently used transformed data from data frames for faster processing. Build hive tables on the transformed data and used different SERDE's to store the data in HDFS in different formats. Loaded the transformed Data into the hive

tables and perform some analysis based on the requirements. Implemented portioning on the Hive data to increase the performance of the processing of data. Analyzed the data by performing Hive queries (Hive QL) to study customer behavior. Created Pig Latin scripts to sort, group, join and filter to transform the data. Worked on various performance optimizations like using distributed cache for small datasets, Partitioning, Bucketing in Hive and Map Side joins. Exported the analyzed data to the relational databases using Sqoop, to further visualize and generate reports for the BI team. Implemented custom workflow to automate the jobs on daily basis. Created custom workflows to automate Sqoop jobs weekly and monthly. Environment: HDFS, Python, Hive, Sqoop, Flume, Spark, MapReduce, Oracle 11g, YARN, UNIX Shell Scripting, Agile Methodology, Cloudera. Python Developer Tachyon Technologies LLC May 2014 to September 2015 Responsibilities Worked with Open stack Command-line client. Created backend database T-SQL stored procedures and Jasper Reports. Created a Git repository and added the project to GitHub. Used Python modules such as requests, urllib, urllib2 for web crawling. Used other packages such as Beautiful soup for data parsing. Worked on writing and as well as read data from csv and excel file formats. Worked on resulting reports of the application and Tableau reports. Performed QA testing on the application. Held meetings with client and worked all alone for the entire project with limited help from the client. Utilize PyUnit, the Python unit test framework, for all Python applications. Exported/Imported data between different data sources using SQL Server Management Studio. Designed and developed the UI of the website using HTML, XHTML, AJAX, CSS and JavaScript. Developed views and templates with Python and Django's view controller and templating language to create a user-friendly website interface. Working with Database procedures, Triggers, PL/SQL statements for data retrievingand also for migration purpose. Environment: Python 2.7/3.0, PL/SQL C++, Redshift, XML, Agile (SCRUM), PyUnit, MYSQL, Apache, CSS, MySQL, DHTML, HTML, JavaScript, Shell Scripts, Git, Linux, Unix and Windows. Education Master's Skills Apache (3 years), Apache hadoop hdfs (3 years), Apache hadoop oozie (Less than 1 year), Apache hadoop sqoop (3 years), Apache kafka. (2 years), databases (1 year), Flume (1 year), Hadoop (2 years), Hadoop distributed file system (3 years), Hdfs (3 years), Hive (3

years), Kafka. (2 years), Linux (3 years), map reduce (1 year), Mysql (1 year), Oozie. (Less than 1 year), Oracle (2 years), Pig (1 year), Python (4 years), Sqoop (3 years) Additional Information TECHNICAL SKILLS: Big Data Technologies: Apache Spark, Apache Hadoop, Map Reduce, Apache Hive, Apache Pig, Apache Sqoop, Apache Kafka, Apache Flume, Apache oozie, Hue, Apache Zookeeper, HDFS, amazon S3, EC2, EMR. Languages: Scala, Python, Java. Databases: MySQL, Oracle 11g. Operating Systems: Mac OS, Windows 7/10, Linux (Cent OS, Redhat, Ubuntu). Development Tools: IntelliJ, Maven, Scala Test, GitHub, Jenkins.

Name: John Stephens

Email: thompsonmegan@example.net

Phone: 475.351.7141