Spark/Hadoop Developer Spark/Hadoop Developer Spark/Hadoop Developer - DXC Technology ? Spark developer with 5+ years of experience in Big data application development through frameworks Hadoop, Spark, Hive, Sqoop, Flume, Oozie, Kafka. ? Hands on experience with Hadoop/Spark Distribution - Cloudera, Hortonworks. ? Experience in implementing Spark with the integration of Hadoop Ecosystem. ? Experience in data cleansing using Spark map and Filter Functions. ? Experience in designing and developing application in Spark using Scala. ? Experience migrating map reduce programs into Spark RDD transformations or actions to improve performance. ? Experience in creating Hive Tables and loading the data from different file formats. ? Implemented Partitioning, Dynamic Partition, Buckets in HIVE. ? Experience developing and Debugging Hive queries. ? Experience in processing the data using HiveQL and Pig Latin scripts for data Analytics. ? Extending Hive Core functionality by writing UDF's for Data Analysis. ? Experience converting HiveQL/SQL queries into Spark transformations through Spark RDD and Data frames API in Scala. ? Used Oozie to Manage and schedule Spark Jobs on a Hadoop Cluster. ? Used HUE GUI to implement Oozie scheduler and workflows. ? Good Experience in Data importing and exporting to Hive and HDFS with Sqoop. ? Experience in using Producer and Consumer API's of Apache Kafka. ? Skilled in integrating Kafka with Spark streaming for faster data processing. ? Experience in using Spark Streaming programming model for Real-time data processing. ? Experience dealing with the file formats like text files, Sequence files, JSON, Parquet, ORC. ? Extensively used Apache Kafka to collect the logs and error messages across the cluster. ? Excellent knowledge and understanding of Distributed Computing and Parallel processing frameworks. ? Experienced at performing read and write operations on HDFS file system. ? Experience working with large data sets and making performance improvements. ? Experience working with EC2 (Elastic Compute Cloud) cluster instances, setup data buckets on S3 (Simple Storage Service), setting up EMR (Elastic MapReduce). ? Extensive programming knowledge in developing Java application using Java ,J2EE and JDBC. ? Good experience working on Tableau and enabled the JDBC/ODBC data connectivity from those to Hive tables. ? Experience creating and driving large scale ETL pipelines. ? Good with version control systems like GIT. ? Strong knowledge on UNIX/LINUX commands. ? Adequate Knowledge on Python scripting Language. ? Adequate knowledge of Scrum, Agile and Waterfall methodologies. ? Highly motivated and committed to the highest levels of professionalism. ? Exhibited strong written and oral communication skills. Rapidly learn and adapt quickly to emerging new technologies and paradigms. Work Experience Spark/Hadoop Developer DXC Technology - Plano, TX January 2019 to Present The main goal of this project (Product Recommendation) was get data from different databases in real-time into HDFS and redesigning the imported data into useable format by cleansing and performing transformations on the data to provide an aggregated overview of the trials to be accessed by the clients. And store the data in hive for further processing by data scientists. Responsibilities: ? Worked under the Cloudera distribution CDH 5.13 version. ? Involved in working with Sqoop for fetching the data from RDBMS. ? Transformed and stored the ingested data into Data frames using spark SQL. ? Created Hive tables to load the transformed Data. ? Performed partitions and bucketing in hive for easy data classification. ? Worked on Performance and Tuning optimization of Hive. ? Involved in exporting Spark SQL Data frames into hive tables stored as Parquet Files. ? Involved in Ingesting real-time log data from various producers using Kafka. ? Used spark streaming to subscribe to desired topics for real time processing. ? Transformed the DStreams into Data frames using spark engine. ? Experienced in performance tuning of Spark Application for setting right Batch Interval time, level of Parallelism and memory tuning for optimal Efficiency. ? Responsible for performing sort, join, aggregations, filter, and other transformations on the data. ? Appended the Data frames to pre-existing data in hive. ? Performed analysis on the hive tables based on the business logic. ? Created a data pipeline using Oozie workflows which performs jobs on a daily basis. ? Involved in Analyzing data by writing queries in HiveQL for faster data processing. ? Involved in Persisting Metadata into HDFS for further data processing. ? Loading data from Linux File systems to HDFS and vice-versa using shell commands. ? Used GIT as Version Control System. ? Worked with Jenkins for continuous integration. ? Build hive tables on the transformed data and used different SERDE'S to store data in HSFS in different formats. ? Used different API's to perform necessary transformation and actions on the data which gets from kafka in real time. ? Involved in collecting and transferring the data from various web servers to HDFS using Apache Kafka. Environment: CDH 5.1, HDFS, Hadoop 3.0, Spark 2.4, Scala, Hive 3.0, Pig, Hue, Oozie, Sqoop, Kafka, Linux shell, Git, Jenkins, Agile. Spark/Hadoop Developer Vanguard - Charlotte, NC February 2018 to December 2018 The main goal of this project (Predict Stock Prices) was to move market data and News data from RDBMS to Hive and perform data analysis using spark SQL and store output to hive for use by the R&D team and setup Kafka for incremental loading for new data from sensors to be appended directly to the respective Hive tables. Responsibilities: ? Worked under the Hortonworks Enterprise. ? Worked on large sets of structured and semi-structured historical data. ? Involved in working with Sqoop to import the data from RDBMS to Hive. ? Created Hive tables to load the Data and stored as ORC files for processing. ? Implemented Hive Partitioning and bucketing for further classification of data. ? Worked on Performance and Tuning optimization of Hive. ? Involved in cleansing and transforming the data. ? Used spark SQL to perform sort, join and filter the data. ? Copied the ORC files to amazon s3 buckets using Sqoop for further processing in amazon EMR. ? Wrote custom UDF's in Spark SQL using Scala. ? Performed data Aggregation operations using Spark SQL queries. ? Copied output data back to Hive from Amazon S3 buckets using Sqoop after getting the output desired by the business. ? Setup Kafka to subscribe to topics(sensors) and load data directly to Hive table. ? Automated filter and join operations to join new data with the respective Hive tables using Oozie workflows daily. ? Used Oozie and Oozie coordinators to deploy end to end data processing pipelines and scheduling workflows. ? Compared the sensor data to a persisted table on a 24hr period to check if the machine is operating at optimal conditions and Used Kafka as a messaging system to notify the producer of that data and the maintenance department in case a maintenance is required. ? Used Git as Version Control System. ? Worked with Jenkins for continuous integration. Environment: HDP 2.5, HDFS, Hadoop 2.7, Spark 2.1, Kafka, Amazon S3, EMR, Sqoop, Oozie, Hive 2.1, Pig, Hue, Linux shell, Git, Jenkins, Agile. Hadoop Developer MITS - Hyderabad, Telangana July 2014 to June 2017 The primary objective of this project was focused on creating a Hive metastore and moving data from RDBMS to the Hadoop environment and writes Map Reduce jobs to process the data previously cleaned, transformed and loaded into Hive to output results requested as input by the BI teams for further ETL processes. Responsibilities: ? Worked under the Cloudera distribution. ? Responsible for building scalable distributed data solutions using Hadoop. Developed Simple to complex Map Reduce jobs. ? Created and populated bucketed tables in Hive to allow for faster map side joins and for more efficient jobs and more efficient sampling. ? Also performed partitioning of data to optimize Hive queries. ? Handled importing of data from Oracle 11g to Hive tables using Sqoop on a regular basis, later performed join operations on the data in the Hive. ? Develop User defined functions in Hive to work on multiple input rows and provide an aggregated result based on the business requirement. ? Wrote user defined custom counters to add to the Map Reduce job to gain further insight and for debugging purposes. ? Developed a Map Reduce job to perform lookups of all entries based on a given key from a collection of Map files that were created from the data. ? Performed side data distribution using the distributed cache to make read only data available to the job to process the main dataset. ? Used Combine File Input Format to make sure maps had sufficient data to process when there is a large number of small files. Also packaged a collection of small files into a Sequence File which was used as input to the Map Reduce job. ? Implemented LZO compression of Map output to reduce I/O between mapper and reducer nodes. ? Continuous monitoring and managing the Hadoop cluster using Web console. ? Developed Pig Latin scripts to extract the data from the output files to load into HDFS. ? Installed Oozie workflow engine to run multiple Hive and Pig jobs. Environment: CDH 5.0, HDFS, Hadoop 2.7, Map Reduce, spark 1.6, Hive 1.2, Pig, Hue, Oozie, Sqoop, Scala, Oracle 12c, YARN, Linux shell, GIT, Jenkins, Agile. Python Developer MITS - Hyderabad, Telangana June 2013 to June 2014 Responsibilities: ? Experienced with Python frameworks like WPebapp2 and, Flask. ? Experienced in WAMP (Windows, Apache, MYSQL, and Python PHP) and MVC Struts ? Developed mobile cross-browser web application Angular JS, JavaScript API. ? Successfully migrated the Django database from SQLite to MySQL to PostgreSQL with complete data integrity. ? Used Celery with Rabbit MQ and Flask to create a distributed worker framework. ? Created Automation test framework using Selenium. ? Responsible for design and development of Web

Pages using PHP, HTML, JOOMLA, CSS including Ajax controls and XML. ? Developed intranet portal for managing Amazon EC2 servers using Tornado and MongoDB. ? Expertise in developing different web applications implementing the Model-View-Controller (MVC) architectures using Full stack frameworks such as Turbo Gears. ? Implemented monitoring and established best practices around using Elastic search. ? Strong experience in building large, responsive based REST web application experienced in Cherrypy framework, Python. ? Used Test driven approach (TDD) for developing services required for the application. ? Developed mobile cross-browser web application Angular JS, JavaScript API.   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. Skills Hdfs, Oozie, Sqoop, Apache kafka, Kafka, Flume, Hadoop, Map reduce, Apache spark, Hadoop, Hive, Pig, Zookeeper, Apache, Linux, Mysql, Oracle, Scala, Tomcat, Eclipse Additional Information TECHNICAL SKILLS: Big Data Technologies: Apache Hadoop, Apache Spark, Map Reduce, Apache Hive, Apache Pig, Apache Sqoop, Apache Kafka, Apache Flume, Apache Oozie, Apache Zookeeper, HDFS Databases: MySQL, Oracle 11g. Languages: Scala, JAVA Operating Systems: Mac OS, Windows 7/10, Linux (Cent OS, Redhat, Ubuntu). Development Tools: Apache Tomcat, Eclipse, NetBeans, IntelliJ.

Name: Maria Bush

Email: larry06@example.org

Phone: +1-396-644-5204x6622