Sr. Spark/ Python Developer Sr. Spark/Python Developer Sr. Spark/ Python Developer - Twitch, CA

Hadoop Developer with 5+ years of overall IT experience in a variety of industries, which includes hands on experience in Big Data tools and technologies. Have 4+ years of comprehensive experience in Big Data processing using Hadoop and its ecosystem (MapReduce, Pig, Hive, Sqoop, Flume, Spark, Kafka and HBase). Good working experience on Spark (spark streaming, spark SQL) with python and Kafka. Hands on experience with Spark Core, Spark SQL and Data Frames/Data Sets/RDD API. Developed applications using Spark and python for data processing. Strong experience and knowledge of real time data analytics using Flume and Spark. Replaced existing map-reduce jobs and Hive scripts with Spark Data-Frame transformation and actions. Good knowledge on Spark architecture and real-time streaming using Spark. Expert in working with Hive data warehouse tool-creating tables, data distribution by implementing partitioning and bucketing, writing and optimizing the HiveQL queries. Performed maintenance, monitoring, deployments, and upgrades across infrastructure. Involved in Debugging Pig and Hive scripts and used various optimization techniques in MapReduce jobs. Wrote custom UDFs and UDAF for Hive and Pig core functionality. Expertise in writing Hadoop Jobs for analyzing data using Hive QL (Queries), Pig Latin (Data flow language), and custom MapReduce programs in Java. Experience in importing and exporting data using Sqoop from Relational Database Systems to HDFS. Good knowledge about YARN configuration. Hands on experience in configuring and working with Flume to load the data from multiple web sources directly into HDFS. Hands on experience working with NoSQL databases such as HBase, MongoDB and Cassandra. Used HBase in accordance with PIG/Hive as and when required for real time low latency queries. Scheduled job workflow for FTP, Sqoop and hive scripts using Oozie and Oozie coordinators. Experience in developing solutions to analyze large data sets efficiently. Good experience in creating and designing data ingest pipelines using technologies such as Apache Storm- Kafka. Maintained list of source systems and data copies, tools used in data ingestion, and landing location in Hadoop. Developed various shell scripts and python scripts to automate Spark jobs and hive scripts. Generated Java APIs for retrieval and analysis on No-SQL database such as HBase and Cassandra. Imported the data from different

sources like AWS S3, Local file system into Spark RDD and worked on cloud Amazon Web Services (EMR, S3, EC2, RedShift, Lambda). Integrated clusters with Active Directory for Kerberos and User Authentication / Authorization. Experience with developing and maintaining Applications written for Amazon Simple Storage, AWS Elastic Beanstalk, and AWS Cloud Formation. Dealt with huge transaction volumes while interfacing the frontend application written in Java, JSP, Struts, Hibernate, SOAP Web service and with Tomcat Web server. Experience in Job scheduling using Control-M. Experience with all stages of the SDLC and Agile Development model right from the requirement gathering to Deployment and production support. Involved in daily SCRUM meetings to discuss the development/progress and was active in making scrum meetings more productive. Also have experience in understanding of existing systems, maintenance and production support, on technologies such as Java, J2EE and various databases (Oracle, SQL Server). Work Experience Sr. Spark/ Python Developer Twitch, CA January 2017 to Present Description: Twitch is a global community that comes together each day to create multiplayer entertainment: unique, live, unpredictable experiences created by the interactions of millions. Responsibilities: Working on Big Data infrastructure for batch processing as well as real-time processing. Responsible for building scalable distributed data solutions using Hadoop. Used Spark-Streaming APIs to perform necessary 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. Experience in designing and developing applications in Spark using Python to compare the performance of Spark with Hive and SQL/Oracle.

Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Python. Experience in both SQLContext and SparkSession and implementing Log Error Alarmer in Spark. Optimized Hive QL/ pig scripts by using execution engine like TEZ, Spark. Tested Apache TEZ, an extensible framework for building high performance batch and interactive data processing applications on Pig and Hive jobs. Experience in managing nodes on Hadoop cluster and monitor Hadoop cluster job performance using Cloudera manager. Managing and scheduling Jobs on a Hadoop Cloudera cluster using Oozie work flows and java schedulers. Used Spark API over

Cloudera Hadoop YARN to perform analytics on data in Hive and also involved in creating Hive Tables, loading with data and writing Hive queries which will invoke and run Map Reduce jobs in the backend. Designed and implemented Incremental Imports into Hive tables. Created Partitions and Bucketing concepts in Hive and designed both Managed and External tables in Hive to optimize performance. Experience in importing and exporting tera bytes of data using Sqoop from Relational Database Systems to HDFS. Moved Relational Database data using Sqoop into Hive Dynamic partition tables using staging tables. Involved in collecting, aggregating and moving data from servers to HDFS using Apache Flume. Involved in developing Pig Scripts for change data capture and delta record processing between newly arrived data and already existing data in HDFS.

Migrated ETL jobs to Pig scripts do transformations, even joins and some pre-aggregations before storing the data onto HDFS. Implemented the workflows using Apache Oozie framework to automate tasks. Used Zookeeper to co-ordinate cluster services. Worked on different file formats like Sequence files, XML files and Map files using Map Reduce Programs. Used Impala where ever possible to achieve faster results compared to Hive during data Analysis. Configured deployed and maintained multi-node Dev and Test Kafka Clusters and implemented data ingestion and handling clusters in real time processing using Kafka. Worked on writing transformer/mapping Map-Reduce pipelines using Java. Transform the logs data into data model using apache pig and written UDF's functions to format the logs data. Setting up Confluent Kafka for ingesting large volumes of events/logs into Hadoop. Developed and Configured Kafka brokers to pipeline server logs data into spark streaming. Used Teradata Data Mover to copy data and objects such as tables and statistics from one system to another. involved in Analyzing / building Teradata EDW using Teradata ETL utilities and Informatica. Implemented usage of Amazon EMR for processing Big Data across a Hadoop Cluster of virtual servers on Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3). Implemented installation and configuration of multi-node cluster on the cloud using Amazon Web Services (AWS) on EC2. Used Amazon Web Services S3 to store large amount of data in identical/similar repository. Environment: Hadoop, HDFS, Pig, Hive, Sqoop, Flume, Kafka, Spark, TEZ, Storm, Shell Scripting, HBase, Python scripting, Kerberos, Agile,

Zoo Keeper, Maven, AWS, AWS EMR, MySQL. Hadoop/Bigdata Developer TDS Telecom, IL December 2015 to December 2016 Description: The purpose of the project is to analyze the data coming from the various sources into the Hadoop data center unit. Created programs to process large volumes of data through a lot of prepay concepts, which analyze, produce suspect claims and it helps to generate Datasets for visualization. These suspect claims verified again and it saves millions of dollars to the company every year. Responsibilities: Involved in implementation of Hadoop Cluster and Hive for Development and Test Environment. Worked on analyzing Hadoop Cluster and different big data analytic tools including Pig, Hive and MongoDB. Extracted files from MongoDB through Sqoop and placed in HDFS for processed. Designed and developed functionality to get JSON document from MongoDB document store and send it to client using RESTful web service. Successfully loaded files to Hive and HDFS from MongoDB. Exploring with the Spark for improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frame, Pair RDD's, Spark YARN. Experienced with batch processing of data sources using Apache Spark and developing predictive analytic using Apache Spark Scala APIs. Imported millions of structured data from relational databases using Sqoop import to process using Spark and stored the data into HDFS in CSV format and developed Spark streaming application to pull data from cloud to hive table. Involved in importing data from Oracle into HDFS and Hive using Sqoop. Created tables using Impala, and involved in creating Queries which are stored in HBase. Implemented complex scripts to support test driven development and continuous integration. Improving the performance and optimization of existing algorithms in Hadoop using Spark context, Spark-SQL and Spark YARN using Scala Analyzed the data as per the business requirements using Hive queries. Installed and configured Hadoop stack and different big data analytic tools, export and imports from data preprocessing. Collected and aggregated large amounts of log data using Apache Flume and staging data in HDFS for further analysis. Hands on experience in writing custom UDF's and also custom input and output formats and created Hive Tables, loaded values and generated adhoc-reports using the table data. Showcased strong understanding on Hadoop architecture including HDFS, MapReduce, Hive, Pig,

Sqoop and Oozie. Used spark with Yarn and got performance results compared with MapReduce. Loaded existing data warehouse data from Oracle database to Hadoop Distributed File System (HDFS). Developed MapReduce programs in Java to search production logs and web analytics logs for use cases like application issues, measure page download performance respectively. Loaded data into BIG SQL Tables from Hadoop Distributed File System (HDFS) to provide access through JDBC/ODBC connections. Evaluated multiple tools available in IBM infosphere Biginsights like IBM BIG SQL , Big Sheets and developing a data warehousing platform. Good understanding of core java concepts and implementation in MapReduce Programs. Involved and actively interacted with cross-functional teams like Web Team. Unix and DBA Team for successful Hadoop implementation. Involved in User Training of Hadoop system for cross-functional teams. Environment: Hadoop, Sqoop, Hive, Pig, Oracle, Java, Oozie, Spark, Scala, Mongo DB, Eclipse IDE, HortonWorks. Hadoop Developer Game Stop, TX December 2014 to November 2015 Description: GameStop's analysis is about which games are being bought, which ones are being traded, which ones move at a certain price or not. Ideally, they learn the trends in the game industry before anyone else. The value of analysis is determining who is likely to buy which series of games. Rapid analytics helps us target individual customers. Responsibilities: Responsible for analyzing large data sets and derive customer usage patterns by developing new MapReduce programs. Written MapReduce code to parse the data from various sources and storing parsed data into HBase and Hive. Created HBase tables to store different formats of data as a backend for user portals. Developed Kafka producer and consumers, HBase clients, Apache Spark and Hadoop MapReduce jobs along with components on HDFS, Hive. Worked on creating combiners, partitions, and distributed cache to improve the performance of MapReduce jobs. Developed Shell Script to perform data profiling on the ingested data with the help of HIVE Bucketing. Developed Hive UDF for performing Hashing mechanism on the Hive Column. Involved in creating Hive tables, loading with data and writing Hive queries, which will run internally in map reduce way. Used Hive to analyze the partitioned and bucketed data and compute various metrics for reporting. Written Hive jobs to parse the logs and structure them in tabular format to facilitate effective querying on the log

data. Writing Hive join query to fetch info from multiple tables, writing multiple Map Reduce jobs to collect output from Hive. Ingest data into Hadoop /Hive/HDFS from different data resources. Involved in loading and transforming large sets of Structured, Semi-Structured and Unstructured data and analyzed them by running Hive queries and Pig scripts. Experienced in writing Hive validation scripts that are used in validation framework (for daily analysis through graphs and presented to business users). Developed workflow in Oozie to automate the tasks of loading data into HDFS and pre-processing with Pig and Hive. Developed code in Python to use MapReduce framework by Hadoop streaming. Used Pig as ETL tool to do transformations, joins and some pre-aggregations before storing the data into HDFS. Imported all the customer specific personal data to Hadoop using Sqoop component from various relational databases like Netezza and Oracle.

Develop testing scripts in Python and prepare test procedures, analyze test results data and suggest improvements of the system and software. Experience in streaming log data using Flume and data analytics using Hive. Extracted the data from RDBMS (Oracle, MySQL & Teradata) to HDFS using Sqoop. Environment: Hadoop, MapReduce, HDFS, Pig, HiveQL, Oozie, Flume, Impala, Cloudera, MySQL, Shell Scripting, HBase. Java/J2EE Developer Talent Nid Pvt Ltd January 2014 to November 2014 Description: This Project aims to implement the infrastructure of Java Message Service (JMS). This project developed in J2EE package using JMS API's provides services for Exchange for message between components in a distributed environment. It supports both Synchronous and Asynchronous messaging and the receiver receives the message according to selection of the message format. The message will be stored in Database and it will be retrieved whenever sender or receiver requires. Responsibilities: Involved in Full Life Cycle Development in Distributed Environment using Java and J2EE framework. Designed the application by implementing Struts Framework based on MVC Architecture. Designed and developed the front end using JSP, HTML and JavaScript and jQuery. Implemented the Web Service client for the login authentication, credit reports and applicant information Apache Axis 2 Web Service. Extensively worked on User Interface for few modules using JSPs, JavaScript and Ajax. Developed framework for data processing using Design patterns, Java, XML. Used the lightweight

container of the Spring Framework to provide architectural flexibility for Inversion of Controller (IOC).

Used Hibernate ORM framework with spring framework for data persistence and transaction management. Designed and developed Session beans to implement the Business logic. Developed EJB components that are deployed on Web logic Application Server. Written unit tests using JUnit Framework and Logging is done using Log4J Framework. Designed and developed various configuration files for Hibernate mappings. Designed and documented REST/HTTP APIs, including JSON data formats and API versioning strategy. Developed Web Services for sending and getting data from different applications using SOAP messages. Actively involved in code reviews and bug fixing. Applied CSS (Cascading style Sheets) for entire site for standardization of the site. Assisted QA Team in defining and implementing a defect resolution process including defect priority, and severity. Environment: Java 5.0, Struts, Spring 2.0, Hibernate 3.2, Web Logic 7.0, Eclipse 3.3, Oracle , JUnit 4.2, Maven, Windows XP, HTML, CSS, JavaScript, and XML. Education Bachelor's Additional Information Technical Skills: Big Data Cloudera Distribution, HDFS, Zookeeper, Yarn, Data Node, Name Node, MapReduce, PIG, SQOOP, HBase, Hive, Flume, Cassandra, MongoDB, Oozie, Kafka, Spark, Storm, Scala, Impala, TEZ, AWS, S3, EC2, RedShift, Elastic Beanstalk, EMR, Lambda, FTP. Operating System Windows, Linux, Unix. Languages Java, J2EE, SQL, PL/SQL Databases Oracle, SQL Server, MySQL, MS Access. Web Technologies JSP, Servlets, HTML, CSS, Java Script, JDBC, SOAP, XSLT. XML Technologies XML, XSLT. IDE Eclipse, STS, IntelliJ. Web/App Server UNIX server, Apache Tomcat. Cloud Platform Amazon Web services (AWS), Azure

Name: April Jennings

Email: ibyrd@example.net

Phone: 540.700.8575x1633