Sr. Hadoop/Spark Developer Sr. Hadoop/Spark Developer Sr. Hadoop/Spark Developer - Sprint Dallas, TX   Extensive IT experience of over 9 years in Analysis, Design, Development, Implementation, Maintenance and Support with experience in developing strategic methods for deploying Big Data technologies to efficiently solve Big Data processing requirement.   Around 5 years of experience on BIG DATA using HADOOP framework and related technologies such as HDFS, Map Reduce, HIVE, PIG, YARN, APACHE SPARK, FLUME, KAFKA, OOZIE, SQOOP, ZOOKEEPER and NoSQL Databases like HBase, Cassandra.   Worked extensively on Hadoop (Gen-1 and Gen-2) and various components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node, and Resource Manager (YARN).   Experience in working with Amazon EMR, Cloudera (CDH4/CDH5) and Horton Works Hadoop Distributions.   Capable of processing large sets of structured, semi-structured and unstructured data and supporting systems application architecture.   Extensively used Apache Sqoop for efficiently importing and exporting data from HDFS to Relational Database Systems and from RDBMS to HDFS.   Worked on data load from various sources i.e., Oracle, MySQL, DB2, MS SQL Server, Cassandra, Hadoop using Sqoop and Python Script.   Experience in developing data pipeline using Sqoop, and Flume to extract the data from weblogs and store in HDFS.   Experience in managing and reviewing Hadoop Log files using FLUME and Kafka and also developed the Pig UDF's and Hive UDF's to pre-process the data for analysis. Worked on Impala for Massive parallel processing of Hive queries.   Extending HIVE and PIG core functionality by using custom User Defined Function's (UDF), User Defined Table-Generating Functions (UDTF) and User Defined Aggregating Functions (UDAF) for Hive and Pig.   Efficient in working with Hive data warehouse tool creating tables, data distributing by implementing Partitioning and Bucketing strategy, writing and optimizing the HiveQL queries. Experience in ingestion, storage, querying, processing and analysis of Big Data with hands on experience in Big Data including Apache Spark, Spark SQL and Spark Streaming.   Implemented advanced procedures like text analytics and processing using the in-memory computing capabilities like Apache Spark written in Scala.   Worked with Spark engine to process large scale data and experience to create Spark RDD and developing Spark Streaming jobs by using RDDs and leverage

Spark-Shell.   Having experience on RDD architecture and implementing Spark operations on RDD and also optimizing transformations and actions in Spark.   Hands on experience in Apache Spark jobs using Scala in test environment for faster data processing and used SparkSQL for querying.   I have been experienced with SPARK SREAMING API to ingest data into SPARK ENGINE from KAFKA.   Worked on real time data integration using Kafka - Storm data pipeline, Spark streaming and HBase.   Experienced in implementing unified data platforms using Kafka producers/ consumers, implement pre-processing using storm topologies.   Exposure to Data Lake Implementation using Apache Spark and developed Data pipe lines and applied business logics using Spark.   Good working experience on different file formats (CSV, Sequence files, XML, JSON, PARQUET, TEXTFILE, AVRO, ORC) and different compression codecs (GZIP, SNAPPY, LZO).   Hands on experience with NoSQL Databases like HBase, Cassandra and relational databases like Oracle, DB2, SQL SERVER and MySQL.   Expertise in job scheduling and monitoring tools like Oozie and ZooKeeper and experience in designing Oozie workflows for cleaning data and storing into Hive tables for quick analysis.   Strong experience in working with ELASTIC MAPREDUCE and setting up environments on Amazon AWS EC2 instances, AZURE, EMR and S3.   Installed and configured JENKINS FOR AUTOMATING Deployments and providing automation solution.   Developed build and deployment scripts using ANT and MAVEN as build tools in JENKINS to move from one environment to other environments   Extensive experience in ETL Data Ingestion, In-Stream data processing, Batch Analytics and Data Persistence Strategy. Worked extensively with Dimensional Modeling, Data Migration, Data Cleansing, Data Transformation, and ETL Processes features for Data Warehouse System.   Experience with creating the TABLEAU dashboards with relational and multi-dimensional databases including Oracle, MySQL and HIVE, gathering and manipulating data from various sources. Having experience in performance tuning, dashboards and TABLEAU reports.   Experience in understanding the security requirements for Hadoop and integrate with Kerberos authentication and authorization infrastructure.   Experience in Object Oriented Analysis Design (OOAD) and development of software using UML Methodology, good knowledge of J2EE design patterns and

Core Java design patterns.   Expertise in design and development of Web Applications involving J2EE technologies with Java, Spring, EJB, AJAX, Servlets, JSP, Struts, Web Services, XML, JMS, JSP, UNIX shell scripts, SERVLETS, MS SQL SERVER, SOAP and RESTful web services. Extensively development experience in different IDE's like Eclipse, NetBeans.   Experience in core Java, JDBC and proficient in using Java API's for application development.   Experience in Deploying web application using application servers WebLogic, Apache Tomcat, WebSphere and JBOSS   Experience in all stages of SDLC (Agile, Waterfall), writing Technical Design document, Development, Testing and Implementation of Enterprise level Data mart and Data warehouses. Authorized to work in the US for any employer Work Experience Sr. Hadoop/Spark Developer Sprint - Overland Park, KS February 2017 to Present Responsibilities: Hands on experience in Spark and Spark Streaming creating RDD & applying operations transformations and Actions.   Developed Spark applications using Scala for easy Hadoop transitions.   Used Spark and Spark-SQL to read the parquet data and create the tables in hive using the Scala API.   Performed advanced procedures like text analytics and processing, using the in-memory computing capabilities of Spark using Scala.   Developed Spark code using Scala and Spark-SQL for faster processing and testing.   Implemented Spark sample programs in python using pyspark.   Analyzed the SQL scripts and designed the solution to implement using pyspark.   Developed pyspark code to mimic the transformations performed in the on-premise environment.   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.   Responsible for loading Data pipelines from web servers and Teradata using Sqoop with Kafka and Spark Streaming API.   Developed Kafka producer and consumers, Cassandra clients and Spark along with components on HDFS, Hive. Populated HDFS and HBase with huge amounts of data using Apache Kafka.   Used Kafka to ingest data into Spark engine.   Configured deployed and maintained multi-node Dev and Test Kafka Clusters.   Managing and scheduling Spark Jobs on a Hadoop Cluster using Oozie. Experienced with different scripting language like Python and shell scripts.   Developed various Python scripts to find vulnerabilities with SQL Queries by doing SQL injection, permission checks

and performance analysis.   Tested Apache TEZ, an extensible framework for building high performance batch and interactive data processing applications, on Pig and Hive jobs.   Designed and implemented Incremental Imports into Hive tables and writing Hive queries to run on TEZ.   Experienced data pipelines using Kafka and Akka for handling large terabytes of data.   Written shell scripts that run multiple Hive jobs which helps to automate different Hive tables incrementally which are used to generate different reports using Tableau for the Business use.   Experienced in Apache Spark for implementing advanced procedures like text analytics and processing using the in-memory computing capabilities written in Scala.   Developed Solr web apps to query and visualize and Solr indexed data from HDFS.   Involved in converting Hive/SQL queries into Spark transformations using Spark RDD, Scala and Python.   Worked on Spark SQL, created Data frames by loading data from Hive tables and created prep data and stored in AWS S3.   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.   Involvement in creating custom UDFs for Pig and Hive to consolidate strategies and usefulness of Python into Pig Latin and HQL (HiveQL)   Extensively worked on Text, ORC, Avro and Parquet file formats and compression techniques like Snappy, Gzip and Zlib.   Implemented Hortonworks NiFi (HDP 2.4) and recommended solution to inject data from multiple data sources to HDFS and Hive using NiFi.   Developed various data loading strategies and performed various transformations for analyzing the datasets by using Hortonworks Distribution for Hadoop ecosystem.   Ingested data from RDBMS and performed data transformations, and then export the transformed data to Cassandra as per the business requirement and used Cassandra through Java services.   Experience in NoSQL Column-Oriented Databases like Cassandra and its Integration with Hadoop cluster.   Build servers using AWS, importing volumes, launching EC2, RDS, creating security groups, auto-scaling, load balancers (ELBs) in the defined virtual private connection and open stack to provision new machines for clients.   Implemented AWS solutions using EC2, S3, RDS, ECS, EBS, Elastic Load Balancer, and Auto scaling groups, Optimized volumes and EC2 instances.   Creating S3 buckets and managing policies for S3 buckets and utilized S3 bucket and Glacier for

storage and backup AWS.   Performed AWS Cloud administration managing EC2 instances, S3, SES and SNS services.   Wrote ETL jobs to read from web APIs using REST and HTTP calls and loaded into HDFS using java and Talend.   Along with the Infrastructure team, involved in design and developed Kafka and Storm based data pipeline.   ORM framework with spring framework for data persistence and transaction management.   Environment: Hadoop, Hive, Map reduce, Sqoop, Kafka, Spark, Yarn, Pig, Cassandra, Oozie, shell Scripting, Scala, Maven, Java, JUnit, agile methodologies, NIFI, MySQL, Tableau, AWS, EC2, S3, Hortonworks, power BI, Solr. Hadoop Developer Charter Communications - St. Louis, MO August 2015 to January 2017 Description: Its services for businesses include internet access, data networking, phone, and wireless backhaul. Charter chose Spectrum as the brand name for the combined cable, internet, voice, and business services. The combined company serves about more than 26 million residential and commercial customers across the US   Responsibilities: Worked on Hadoop cluster and data querying tools Hive to store and retrieve data.   While developing applications involved in complete Software Development Life Cycle (SDLC).   Reviewing and managing Hadoop log files by consolidating logs from multiple machines using flume.   Developed Oozie workflow for scheduling ETL process and Hive Scripts.   Started using apache NiFi to copy the data from local file system to HDFS. Involved in teams to analyze the Anomaly detection and ratings of data.   Implemented custom input format and record reader to read XML input efficiently using SAX parser.   Involved in writing queries in SparkSQL using Scala. Worked with SPLUNK to analyze and visualize data.   Analyze database and compare it with other open-source NoSQL databases to find which one of them better suites the current requirement   Integrated Cassandra as a distributed persistent metadata store to provide metadata resolution for network entities on the network   Implemented Spark using Scala and SparkSQL for faster testing and processing of data.   Having experience on RDD architecture and implementing Spark operations on RDD and also optimizing transformations and actions in Spark.   Involved in working with Impala for data retrieval process.   Exported data from Impala to Tableau reporting tool, created dashboards on live connection.   Designed multiple Python packages that were used within a large ETL process used to load 2TB of data from an existing

Oracle database into a new PostgreSQL cluster   Involved in converting Hive/SQL queries into Spark transformations using Spark RDD's.   Loaded data from Linux file system to HDFS and vice-versa   Developed UDF's using both DataFrames/SQL and RDD in Spark for data Aggregation queries and reverting back into OLTP through Sqoop.   POC for enabling member and suspect search using Solr.   Worked on ETL methods for data extraction, transformation and loading in corporate-wide ETL Solutions and Data warehouse tools for reporting and data analysis.   Used CSVExcelStorage to parse with different delimiters in PIG.   Installed and monitored Hadoop ecosystems tools on multiple operating systems like Ubuntu, CentOS.   Performed advanced procedures like text analytics and processing, using the in-memory computing capabilities of Spark using Scala.   Modified reports and Talend ETL jobs based on the feedback from QA testers and Users in development and staging environments. Involved in setting QA environment by implementing pig and Sqoop scripts.   Got chance working on Apache NiFi like executing Spark script, Sqoop scripts through NiFi, worked on creating scatter and gather pattern in NiFi, ingesting data from Postgres to HDFS, Fetching Hive metadata and storing in HDFS, created a custom NiFi processor for filtering text from Flow files etc.   Responsible for designing and implementing ETL process using Talend to load data from Worked extensively with Sqoop for importing and exporting the data from HDFS to Relational Database systems/mainframe and vice-versa.   Developed Pig Latin scripts to do operations of sorting, joining and filtering enterprise data.   Implemented test scripts to support test driven development and integration.   Developed multiple MapReduce jobs in java to clean datasets.   Involved in loading data from Linux file systems, servers, java web services using Kafka producers and consumers.   Involved in developing code to write canonical model JSON records from numerous input sources to Kafka Queues.   Performed streaming of data into Apache ignite by setting up cache for efficient data analysis.   Collected the logs data from web servers and integrated in to HDFS using Flume.   Developed UNIX shell scripts for creating the reports from Hive data.   Manipulate, serialize, model data in multiple forms like JSON, XML. Involved in setting up MapReduce 1 and MapReduce 2.   Prepared Avro schema files for generating Hive tables and Created Hive tables and loaded the data in to tables and query data

using HQL.   Installed and Configured Hadoop cluster using Amazon Web Services (AWS) for POC purposes.   Environment: Hadoop MapReduce 2 (YARN) , Nifi, HDFS, PIG, Hive, Flume, Cassandra, Eclipse, Ignite Core Java, Sqoop, Spark, Splunk, Maven, SparkSQl, Cloudera, SolrTalend, Linux shell scripting. Java/Hadoop Developer Grubhub - Chicago, IL June 2013 to July 2015 Description: Grubhub is an online and mobile food-ordering company that connects diners with local restaurants. Based in Chicago, the company has more than 9.18 million active diners, and more than 55,000 restaurant partners in over 1,100 cities across the United States and the United Kingdom. The idea of Integrated Data Ware House Project is to ingest data from different multiple sources to Hadoop Data Lake, perform transformations on it according to business requirements and loading the data to external systems   Responsibilities: Exported data from DB2 to HDFS using Sqoop and Developed MapReduce jobs using Java API.   Designed and implemented Java engine and API to perform direct calls from front-end JavaScript (ExtJS) to server-side Java methods (ExtDirect).   Used Spring AOP to implement Distributed declarative transaction throughout the application.   Designed and developed Java batch programs in Spring Batch. Worked on Data Lake architecture to build a reliable , scalable, analytics platform to meet batch, interactive and on-line analytics requirements   Concerned and well-informed on Hadoop Components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node, YARN and Map Reduce programming.   Developed Map-Reduce programs to get rid of irregularities and aggregate the data.   Implemented Hive UDF's and did performance tuning for better results   Developed Pig Latin Scripts to extract data from log files and store them to HDFS. Created User Defined Functions (UDF's) to pre-process data for analysis   Implemented optimized map joins to get data from different sources to perform cleaning operations before applying the algorithms.   Experience in using Sqoop to import and export the data from Oracle DB into HDFS and HIVE.   Implemented CRUD operations on HBase data using thrift API to get real time insights.   Developed workflow in Oozie to manage and schedule jobs on Hadoop cluster for generating reports on nightly, weekly and monthly basis.   Used various compression codecs to effectively compress the data in HDFS. Used Avro SerDe's for serialization and de-serialization and also implemented hive custom UDF's

involving date functions.   Responsible for troubleshooting issues in the execution of MapReduce jobs by inspecting and reviewing log files.   Worked in Agile development environment in sprint cycles of two weeks by dividing and organizing tasks. Participated in daily scrum and other design related meetings.   Installed and configured Pig and wrote Pig Latin scripts.   Created and maintained Technical documentation for launching Cloudera Hadoop Clusters and for executing Hive queries and Pig Scripts.   Developed workflow-using Oozie for running MapReduce jobs and Hive Queries.   Done the work in importing and exporting data into HDFS and assisted in exporting analyzed data to RDBMS using SQOOP.   Involved in loading data from UNIX file system to HDFS.   Created java operators to process data using DAG streams and load data to HDFS.   Assisted in exporting analyzed data to relational databases using Sqoop.   Involved in Develop monitoring and performance metrics for Hadoop clusters.   Continuous monitoring and managing the Hadoop cluster through Cloudera Manager.   Environment: Hadoop, HDFS, Hive, Flume, Sqoop, HBase, PIG, Eclipse, Spark, My SQL and Ubuntu, Zookeeper, Maven, Jenkins, Java (JDK 1.6), Oracle10g.

Java Developer Avineon, Inc - McLean, VA September 2011 to May 2013 Description: Avineon has been providing high quality information technology (IT) solutions to clients operating in both government and business sectors. Its aim is to meet our client's business and organizational objectives with the help of easy-to-use, feature-rich IT solutions. Avineon offers IT services in Software Solutions, Security Engineering, Cloud Computing, and Strategic Business Services and IT Service Management.   Responsibilities: Involved in the design, development and deployment of the Application using Java/J2EE Technologies.   Performed Requirements gathering and analysis and prepared Requirements Specifications document. Provided high level systems design specifying the class diagrams, sequence diagrams and activity diagrams   Involved in designing user interactive web pages as the front-end part of the web application using various web technologies like HTML, JavaScript, Angular JS, AJAX and implemented CSS for better appearance and feel.   Integrated AEM to the existing web application and created AEM components using JavaScript, CSS and HTML.   Programmed Oracle SQL, T-SQL Stored Procedures, Functions, Triggers and Packages as back- end processes to create and update staging tables, log and audit

tables, and creating primary keys.   Provided further Maintenance and support, this involves working with the Client and solving their problems which include major Bug fixing.   Deployed and tested the application using Tomcat web server.   Analysis of the specifications provided by the clients.   Developed JAVABEAN components utilizing AWT and SWING classes.   Extensively used Transformations like Aggregator, Router, Joiner, Expression, Lookup, Update Strategy, and Sequence Generator.   Used Exception handling and Multi-threading for the optimum performance of the application.   Used the Core Java concepts to implement the Business Logic.   Provided on call support based on the priority of the issues.   Designed and implemented a generic parser framework using SAX parser to parse XML documents which stores SQL.   Perform Functional testing, Performance testing, Integration testing, Regression testing, Smoke testing and User Acceptance Testing (UAT).   Environment: Core Java, Servlets, struts, JSP, XML, XSLT, JavaScript, Apache, Oracle 10g/11g. Jr. Java Developer Emorphosys - Hyderabad, Telangana July 2009 to August 2011 Description: This is a business telephone application known as Softphone which is designed to work with IP-based phone systems, and Local Area Network or Wide Area Network (WAN). The server supports the Softphone features like Call Conference, Call Transfer, and Call Forward. The Softphone provides Voice over IP (VoIP) service using a telephony server. It is used to make and take calls through computer using a headset or handset connected to Avaya Hard phone, Responsibilities: Documented in SharePoint and Confluence of projects.   Created the PTO (Permit To Operate) of the project to provide the supported documents before the development phase.   Developed the Web Application using Spring MVC.   To develop the web view of the application used CSS, HTML, JSP, and Java Script.   Used Control M to configure and schedule the Batch Jobs.   IT Service Management (ITSM) is used to plan, deliver, operate and control IT services offered   In the TDD Development process of the project, created and tested the functionality of the classes with mocking framework (Mockito).   For writing the code used Eclipse Neon IDE for the software development   For build and deploy the application to development environment used Maven   Managed the different versions of the source code with TortoiseSVN client.   For Agile Project Management used JIRA   Created Application Flow Diagram using Visio

at the time creation of PTO.   For continuous delivery, testing and deployment used Jenkins

Environment: JAVA 1.8, Spring MVC 3.2, Hibernate 4.2, JavaScript, JSON 2.2, Control M, JIRA, ITSM, Eclipse Neon, Maven 3.3.9, Jenkins, Visio, Mockito 2.2, SVN, Linux, Windows 7 Education Bachelor of Technology in Electronics and Communication Engineering University of Illinois - Urbana-Champaign, IL Masters of Science in Computer Science Governors State University - Chicago, IL Skills JAVA (9 years), SQL (6 years), ORACLE (5 years), Hadoop (5 years), HADOOP (5 years) Additional Information Operating Systems Windows, Linux, UNIX, Ubuntu, Centos Programming or Scripting Languages C, C++, Core Java/J2EE, Unix Shell Scripting, Python, SQL, Pig Latin, Hive QL, Scala Hadoop Distributions Cloudera(CDH4/CDH5) , Hortonworks (HDP2.5) IDE/GUI Eclipse3.2, IntelliJ, Scala IDE Build Tools Jenkins, Maven, ANT Database Microsoft SQL Server ,MS SQL ,Oracle 11g/10g, DB2, MySQL, MS-Access, MS-Access, NoSQL (HBase, Cassandra) Cloud Computing Tools Amazon AWS,AZURE Versioning Tools JIRA, CVS, SVN, and GitHub. SDLC Methodologies Agile, Scrum, Waterfall Model.

Name: Chelsea Zavala

Email: iwright@example.com

Phone: 785-404-3119