

Hadoop Developer Hadoop Developer Hadoop Developer - UBS AG Orlando, FL Over 8 years of professional IT experience with over 3 Years of Big Data experience in ingestion, storage, querying, processing and analysis. Excellent understanding of HDFS, Spark, MapReduce, YARN and tools including Hive, Impala and Pig for data analysis, Sqoop and Flume for data ingestion, Oozie for scheduling and ZooKeeper for coordinating cluster resources. Good knowledge on Spark components like SparkSQL, MLLib, Spark.ML, Spark Streaming and GraphX. Experience in writing HiveQL queries to store processed data into Hive tables for analysis. Experience in building Pig scripts to extract, transform and load data onto HDFS for processing. Experience working with NoSQL databases: Cassandra, HBase and MongoDB. Experience in using different file formats: Parquet, Avro, ORC, RCFile, etc., Experience in working with BI team and transform requirements into Hadoop/NoSQL centric technologies. Experience in Administering, Installation, Configuration, Troubleshooting, Security, Backup, Performance Monitoring and Fine-tuning of Linux Red Hat. Strong experience in all the phases of SDLC including requirements gathering, analysis, design, implementation, deployment and support. Experience in using Maven and ANT for build automation. Experience in using version control and configuration management tools like SVN, CVS and Tortoise. Experience working in environments using Agile (SCRUM) and Waterfall methodologies. Involved in design and development of various web and enterprise applications using various technologies like JSP, Servlets, Struts, Hibernate, Spring, JDBC, JSF, XML, AJAX, SOAP and Amazon Web Services. Experience in using of web/application servers Apache Tomcat, Web Logic and WebSphere. Developed databases using SQL and PL/SQL and experience working on databases like Oracle, SQL Server, PostgreSQL and MySQL. Good experience in database design, creating Tables, Views, Stored Procedures, Functions, Triggers and Indexes. Excellent interpersonal skills, good experience in interacting with clients with good team player and problem solving skills. Strong team player, ability to work independently and in a team as well, ability to adapt to a rapidly changing environment and commitment towards learning. Ability to blend technical expertise with strong Conceptual, Business and Analytical skills to provide quality solutions. Work Experience Hadoop Developer UBS AG - Orlando, FL December 2018 to

Present Responsibilities: Upgraded the Cloudera distribution from CDH 4 to CDH 5, configured high availability for both the NameNode, Impala and other services. Worked on a 30 node Hadoop cluster with highly unstructured and semi structured data of 90 TB in size (270 TB with replication factor of 3) Developed Puppet modules to automate the installation, configuration and deployment of ecosystem tools, OS's and network infrastructure at a cluster level. Performed cluster co-ordination and assisted with data capacity planning and node forecasting using ZooKeeper. Executed custom interceptors for Flume to filter data and defined channel selectors to multiplex the data into different sinks. Extracted transactional data from Netezza and MySQL databases to HDFS using Sqoop. Wrote and executed various MySQL database queries from python using Python-MySQL connector and MySQL DB package. Optimized MapReduce jobs to use HDFS efficiently by using Gzip, LZO, Snappy and Bzip2 compression techniques. Experience in writing Pig scripts to transform raw data from several data sources into forming baseline data. Created Hive tables to store the processed results in a tabular format and written Hive scripts to transform and aggregate the disparate data. Automated the process for extraction of data from warehouses and weblogs into HIVE tables by developing workflows and coordinator jobs in Oozie. Transferred data from Hive tables to HBase via stage tables using Pig and used Impala for interactive querying of HBase tables. Written Python scripts to automate the jobs and improve the performance. Utilize PyUnit, the Python unit test framework, for all Python applications. Exported the aggregated data to SQL Server using Sqoop for creating dashboards in the Tableau and helped to Responsible for cluster maintenance, rebalancing blocks, commissioning and decommissioning of nodes, monitoring and troubleshooting, manage and review data backups and log files. Integrated Hadoop Security with Active Directory by implementing Kerberos for authentication and Sentry for authorization. Implemented POC to migrate iterative map reduce programs into Spark transformations using Spark. Cross Verification of the XML logs for integration of the different web services, REST Services and involved in Functionality, User Interface, System, Integration Testing. Scheduled snapshots of volumes for backup to find root cause analysis of failures and document bugs and fixes for downtimes and maintenance of cluster. Automated processes for

troubleshooting, resolution and tuning of Hadoop clusters. Utilized Agile Scrum Methodology to manage and organize the team with regular code review sessions. Environment: Cloudera (CDH 5), HDFS, MapReduce, YARN, Spark, Hive, Pig, Flume, Sqoop, Puppet, Oozie, ZooKeeper, Clouder Manager, Oracle SQL server, MySQL, HBase, Impala, SparkSQL, Cassandra, Avro, Parquet, RCFile, JSON, UDF, Java (jdk1.7), Kerberos, Sentry, Tableau, CentOS Hadoop Developer Cisco Systems - San Jose, CA January 2017 to November 2018 Project Description: Cisco spending on test equipment continues to increase year over year and outpaces revenue growth. The Supply Chain Operations Test Transformation program is established to enhance tools, processes, policies and governance that will decrease cost and product lead time. The Test Strategy project is established within the program to identify tests where the fault detection rate vs. modeled escape rate does not justify the capital investment. This will prevent the purchase of unnecessary test equipment and allow existing equipment to be repurposed, thereby reducing capital expenses for test equipment. Module: Manufacturing test process of Cisco products, build and shipping status, returns and service shipment. Responsibilities: Involved in architecture design, development and implementation of Hadoop application deployment, backup and recovery systems. Migrated Supply Chain use cases from Cloudera distribution to MapR distribution as part of Cisco's Datalake Initiative. Worked on 135 node MapR Hadoop Production Cluster with structured and semi-structured Supply Chain data. Involved in creation and modeling of Hive tables and automated the process of ingestion and transformation by building workflows in the scheduler. Created custom Sqoop jobs to incrementally pull hourly data from SQL Server to HDFS into Hive table partitions. Assisted in optimization of queries used for applications by using Hive partitions, bucketing and different file formats. Experience in using Parquet, Avro and RCFile file formats for efficient compression and query performance improvements. Initiated and implemented the conversion of Hive MapReduce jobs into Spark In-Memory execution model using PySpark and SparkSQL. Developed UDF/UDAFs using SparkSQL to create custom transformations and aggregations from datasets. Implemented multiple POCs using PySpark and SparkSQL for ETL operations and integrate transactional data with DataStax Enterprise Cassandra and Hadoop

cluster. Involved in data modelling of Cassandra tables to enable real time reporting applications using Spark components. Created dashboards in the Spotfire and Tableau using generated datasets and aggregations by connecting to the corresponding Hive tables using Impala/Hive ODBC connector. Utilized CA Agile Central/Rally quality module to maintain and execute test cases, defect tracking, test case - user story mapping and metrics. Participated in weekly meetings with technical collaborators and involved in code review sessions with developers using Agile methodology. Experience in working with Apache Solr in setting up of the collections and querying the schema. Developed text analytics workflow for Cisco's Customer Assurance Program Initiative, including sentiment scoring and theme categorization of Service Request Notes using PySpark, SparkSQL and Python Natural Language Processing packages. Implemented various machine learning models like Linear Regression, Logistic regression, Dimensionality Reduction via feature hashing using Spark MLLib, Spark ML and Python NumPy packages.

Hadoop Developer
Charter Communications - St. Louis, MO April 2015 to December 2016 Responsibilities:

Responsible for building scalable distributed data solutions on a 40-node cluster using Cloudera Distribution (CDH 4) Developed data pipeline using Flume, Sqoop, Pig and Java MapReduce to ingest customer behavioral data and financial histories onto HDFS. Involved in moving all log files generated from various sources to HDFS for further processing through Flume and process the files by using PiggyBank functions. Developed Sqoop scripts to import and export data from MySQL and handled incremental and updated changes into HDFS layer. Developed workflow in Oozie to orchestrate a series of Pig scripts to cleanse data, such as removing unnecessary information or merging many small files into large, compressed files using pig pipelines in the data preparation stage. Created Hive tables and loaded the data into tables to query using HiveQL. Implemented partitioning and bucketing in HIVE tables and executed the scripts in parallel to improve the performance. Created HBase tables to store various data formats as input coming from different sources. Developed different kind of custom filters and handled pre-defined filters on HBase data using Java API. Connected Hadoop cluster to MongoDB for implementation and executed programs on servers. Used python scripts to update content in the database and manipulate files.

Involved in building the REST API using Spring for fetching the data from MongoDB. Built the shell scripts to monitor the health of hadoop daemon services and responded accordingly to any warning or failure conditions. Implemented Fair Schedulers on the Job tracker to share the resources of the cluster for the MapReduce jobs given by the users. Experience in monitoring and managing of Hadoop cluster using Cloudera Manager for optimum performance and utilization of the resources. Created the users and groups in LDAP and configured mappings for Hadoop services.

Centralized service for maintaining configuration information and provided distributed synchronization and group services using ZooKeeper. Environment: Cloudera (CDH 4), MapReduce, HDFS, Pig, Hive, Flume, Sqoop, HBase, MongoDB, ZooKeeper, Oozie, Fair Schedulers, LDAP, MySQL, Cloudera Manager, Linux Big Data Engineer Eli Lilly - Indianapolis, IN January 2013 to March 2015 Responsibilities: Deployed Hadoop cluster on Amazon Web Services (AWS) with Elastic MapReduce (EMR) as EC2 instances. Installed and configured Hive and Pig environment on Amazon EC2. Ingested and integrated the unstructured log data from the web servers onto cloud using Flume. Configured Sqoop and developed scripts to extract structured data from PostgreSQL onto Amazon S3 cloud. Used Pig as ETL tool to do transformations, event joins, filters both traffic and some pre-aggregations before storing the data onto cloud. Developed multiple MapReduce jobs for data cleaning and preprocessing. Created Hive tables to store the processed results in a tabular format and automated the jobs for extracting data from FTP server into Hive tables using Oozie workflows. Performed queries using HiveQL and exported the analyzed data for visualization to the reporting team. Implemented algorithms and built profiles using Hive and stored the results in HBase and performed CRUD operations using HBase Java Client API and Rest API. Wrote script for Location Analytic project deployment on a Linux cluster/farm & AWS Cloud deployment using Python. Used Aspera Client on Amazon EC2 instance to connect and store data in the Amazon S3 cloud. Managed and monitored the infrastructure of the Hadoop cluster and the Amazon AWS using Amazon CloudWatch. Worked with administration team and created the Load Balancer on AWS EC2 for unstable cluster. Secured the encrypted data on Amazon AWS using the CloudHSM and authenticated node

communication using Kerberos. Utilized ZooKeeper to implement high availability for Namenode and automatic failover infrastructure to overcome single point of failure. Used SVN for version control and Maven to build the application and implemented unit testing using MRUnit.

Environment: Amazon S3, AWS, Elastic MapReduce, EC2, Hive, Pig, Flume, Sqoop, HBase, ZooKeeper, Oozie, CloudHSM, Kerberos, PostgreSQL, Aspera, CloudWatch Software Engineer

Aditri Technologies Pvt. Ltd July 2011 to December 2012 Responsibilities: Designed the system with object-oriented methodology. Participate in the whole SDLC lifecycle from the re-architecture stage to maintenance stage for this product. Gathered, analyzed and coded Business Requirements. Developed presentation layer components comprising of JSP, Servlets and JavaBeans using the struts framework. Designed the presentation layer using JSP, XML & XSLT.

Implemented the complex style-sheet using XSLT to present XML data in the presentation layer. Developed and deployed EJB components on IBM WebSphere Application Server. Developed XML and Action classes to implement framework. Participated in development and validation of XML, XSD. Designed and developed a highly convenient front end user interface using HTML and Java Server Pages (JSP), for customer profile setup. Extensively worked on SQL Queries, Stored procedures and Triggers. Used Struts validation framework for validations. Created the database tables with indexes and views in the database-using Oracle. Responsible for Analysis, Coding and Unit Testing and Support. Environment: Java, MQ Series, Struts, Servlets, JSP, EJB, IBM

WebSphere application server, WSAD, SQL, XML, XSLT, XHTML, SQL Server, Windows.

Education Bachelors in Computer Science Osmania University Skills APACHE HADOOP MAPREDUCE (6 years), APACHE HADOOP SQOOP (6 years), Hadoop (6 years), HADOOP (6 years), Hive (6 years) Additional Information TECHNICAL SKILLS

Big Data Ecosystem: Hadoop, Spark, MapReduce, YARN, Flink, Hive, SparkSQL, Impala, Drill, Pig, Sqoop, HBase, Flume, Oozie, Zookeeper, Avro, Parquet, Maven, Snappy, Bzip2. Hadoop Distributions: Cloudera, MapR, and Hortonworks NoSQL Databases: Cassandra, Mongo DB, HBase Languages: Java, Python, Scala, SQL, HTML, JavaScript, XML, and C/C++ Java Technologies: JSP, Servlets, JavaBeans, JDBC, JNDI, EJB Web Design Tools: HTML, DHTML, AJAX, JavaScript, JQuery and CSS DB Languages:

SQL Server, MySQL, PL/SQL, PostgreSQL, Oracle Frameworks: Struts, spring, Hibernate

Operating systems: UNIX, Linux, and Windows Variants

Name: Jessica Decker

Email: laura30@example.com

Phone: 593-847-3151x503