Hadoop/Spark Developer Hadoop/Spark Developer Hadoop/Spark Developer - BECU Seattle, WA Around 5 years of experience in Big data application development through Hadoop ecosystem components like Hadoop, Spark, Hive, Kafka, Sqoop, Yarn, Oozie. Hands-on experience in working with Cloudera, Hortonworks Hadoop Distributions. Having knowledge and understanding of Distributed Computing and Parallel processing frameworks. Experienced at performing read and write operations on HDFS filesystem. 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 applications in Spark using Scala. Worked with Spark RDD for parallel processing of datasets in HDFS, SQL Server and other data sources. Used job scheduling tool Oozie to manage and schedule Spark Jobs on a Hadoop cluster. Having good knowledge and experience on Apache Spark, Spark Streaming, Spark SQL. Hands on experience in using Spark Streaming programming model for real time processing of data stored in HDFS. Skilled in integrating Kafka with Spark streaming for faster data processing. Knowledge of using Producer and Consumer API's of Apache Kafka. Experience in creating Hive Tables and loading the data from different file formats. Good Experience in Data importing and Exporting to Hive and HDFS with Sqoop, and processing data using Hive QL. Implemented Partitioning, Dynamic Partition, Buckets in HIVE. Extending Hive Core functionality by writing UDF's for Data Analysis. Experience in processing the data using HiveQL for data Analytics. Experience in converting Hive/SQL queries into Spark transformations using Spark RDD and Dataframe API in Scala and Python and performing map-side joins on RDD's. Good exposure to Python programming. Good knowledge on Python Collections and Python Scripting. Experience working with large data sets and making performance improvements. Experience dealing with file formats like Sequence files, Avro, JSON, Parquet, ORC. Sufficient knowledge on NOSQL databases HBASE. Experience in creating and driving large scale ETL pipelines. Knowledge of working with Amazon's Elastic Cloud Compute (EC2) cluster instances for computational tasks, Simple Storage Service (S3) as Storage mechanism and setting up EMR (Elastic MapReduce). Experience in working with Tableau visualization tool. Good with version control systems like Git. Experience in using different build

tools like SBT and Maven. Strong Knowledge on UNIX/LINUX commands and shell scripting. Adequate knowledge of Scrum, Agile methodologies. Good communication and presentation skills, willing to learn and adapt to emerging new technologies. Highly motivated with the ability to work independently or as an integral part of a team and committed to highest levels of profession. Work Experience Hadoop/Spark Developer BECU - Tukwila, WA August 2018 to Present Responsibilities: Having good Knowledge on Spark Architecture and core components of Hadoop frameworks. Experience in designing and deployment of Hadoop cluster and different Big Data analytic tools including Spark, Hive, Sqoop, Oozie, with Cloudera distribution. Worked under the Cloudera distribution CDH 5.6 version. Imported and transformed large scale volumes of data from various data sources to HDFS. Loading data from Linux Filesystems to HDFS and vice-versa. Experience in importing and exporting data using Sqoop from Relational Database Systems to HDFS and vice-versa. Processed different data sets files from HDFS into Spark code using Scala and Spark-SQL for faster testing and processing of data. Load the data into Spark RDD's & Spark Data Frame API's and performed in-memory data computation to generate the output response. Worked on writing various spark transformations using Scala for Data Validation, Cleansing and Filtering in Hadoop HDFS. Developed Scala scripts using both Data frames and RDD's in Spark for Data aggregation queries. Responsible in performing a sort, join, filter, and other transformations on the datasets. Involved in creating Hive tables to load the transformed data and stored it in HDFS. Did various performance optimizations like using distributed cache for small datasets, partitioning, bucketing of the tables in hive and Map Side joins. Written customized Hive UDF's in Scala where the functionality is too complex, to extend Hive functionality. Performed analysis on the hive tables based on the business logic, by writing queries using HiveQL for faster data processing. Appended the Dataframes to pre-existing data in hive. Performed data cleansing to meet business requirements and stored the output data to Hive and HDFS. Implemented the workflows using Apache Oozie framework to automate tasks. Used Git extensively as versioning tool. Worked with Jenkins for continuous integration. Environments: Cloudera 5.6, Hadoop 3.0, HDFS, Spark 2.4, Hive 3.0, Spark SQL, Scala, Sqoop, Oozie, Linux

shell, GIT, Jenkins, Agile. Spark/Hadoop Developer T-Mobile - Seattle, WA April 2017 to July 2018 Responsibilities: Worked under the Hortonworks HDP Enterprise. Worked on large sets of structured and semi-structured data. Identifying data sources and create appropriate data ingestion procedures. Worked on importing and exporting data from Oracle into HDFS and HIVE using Sqoop. Involved in creating Hive tables, loading data into them. Analyzed the data by performing Hive queries (Hive QL) that will run internally in Map reduce way. Developed UDF's to analyse/transform the data. Implemented Partitioning, Dynamic Partitions, Buckets in HIVE and designed both Managed and External tables in Hive to optimize performance. Written PySpark scripts for data extraction, transformation and aggregation. Experience converting HiveQL/SQL queries into Spark transformations through Spark RDD and Dataframe API in Python. Wrote various spark transformations in Python to perform data cleansing, validation and summarization activities on the data. Implemented the persistence of frequently used transformed data from data frames for faster processing. Used Spark SQL to perform sort, join, and filter the data. Performed data aggregation operations using Spark SQL queries. Implemented data ingestion and handling clusters in real time processing using Kafka. Configured Spark streaming to receive data from Kafka and store the streamed data to HDFS using Scala for real time processing. Transformed the Dstreams into Dataframes using spark SQL. Build hive tables on the transformed data and used different SERDE's to store the data in HDFS in ORC format. Copied the ORC files to Amazon S3 buckets using Sqoop for further processing in amazon EMR. Exported the analyzed data to the relational databases using Sqoop, to further visualize and generate reports for the BI team. Used Oozie and Oozie coordinators to deploy end to end data processing pipelines and scheduling the work flows. Used Git as Version Control System. Automated the code deployment process using Jenkins. Environments: HDP 2.5, Hadoop 2.6, HDFS, Spark 2.4, Hive 2.1, Scala, Sqoop, Kafka, Oozie, Amazon S3, Linux shell, Git, Jenkins, Agile. Software Engineer Dhanush IT Solutions May 2014 to July 2016 Responsibilities: Was involved in Big Data project implementation and support. Worked on big data platform and ecosystem, created complex data processing pipelines for data management functionalities. Was involved in design and

implementation of highly scalable system that meet the architectural requirements for system scalability, availability, and performance. Implemented CDH4 Hadoop cluster on Linux. Assisted with performance tuning and monitoring. Worked extensively with Sqoop for importing and exporting the data from HDFS to Relational Database systems and vice-versa loading data into HDFS. Developed and maintained application that run on custom architecture, using diverse technologies including Core Java, J2EE, XML, JMS. Used Oozie to automate data loading into the Hadoop Distributed File System. Shared responsibility for administration of Hadoop, Hive. Organized tasks and resources to complete work and meet deadlines according to established departmental procedures. Conferred with systems analysts, data engineers, programmers and others to design system and to obtain information on project performance requirements and interfaces. Was involved in building distributed, scalable, and reliable data pipelines that ingest and process data at large scale. Experienced working in a Linux environment. Detected data quality issues, identified their root causes, developed fixes. Design, build and maintain data pipelines to support data and analytical needs. Utilized programming tools to bring together a diverse set of data sources and making them easily accessible and useful for analysis. Developed user interfaces using JSP, HTML, XML and JavaScript for interactive cross browser functionality and complex user interface. Provided Technical support for production environments resolving the issues, analyzing the defects, providing and implementing the solution defects. Environments: Hadoop, HDFS, Sqoop, Hive, Java, JSP, Cloudera, Linux, Agile Skills Hdfs, Mapreduce, Oozie, Sqoop, Hbase, Kafka, Hadoop, Nosql, C++, Git, Hadoop, Hbase, Hive, Mapreduce, Python, Zookeeper, Sql server, Oracle, Sql, Linux Additional Information Technical Skills: Big-Data Technologies Spark, Hadoop, MapReduce, HDFS, Hive, Yarn, Oozie, Sqoop, Kafka, Zookeeper Hadoop Distributions Horton Works, Cloudera Cloud platforms Aws Databases Oracle 11g/10g, SQL Server NOSQL Databases HBase Version Control Tools Git Build Tools Maven, sbt Languages Scala, Python, Java, C, C++, SQL, HiveQL, Unix shell scripts Operating System Linux (Various Versions), Windows 7/10 Development Tools IntelliJ, NetBeans, Scala IDE for Eclipse

Name: Amanda Perez

Email: joshua92@example.net

Phone: (420)279-4892x006