BIG DATA ENGINEER BIG DATA ENGINEER Chris Peng - Hadoop Big Data Atlanta, GA ? Good Knowledge on Spark framework on both batch and real-time streaming data processing. ? Hands-on experience processing data using Spark Streaming API and Spark SQL. ? Skilled in AWS, Redshift, DynamoDB and various cloud tools. ? Have streamed over millions of messages per day through Kafka and Spark Streaming. ? Responsible for moving and transforming big data for insightful information using Sqoop. ? Capable of building big data pipelines to optimize utilization of data and configure end-to-end systems. ? Used Kafka for data ingestion and extraction into HDFS Hortonworks system. ? Used Spark SQL to perform preprocessing using transformations and actions on data residing in HDFS. ? Created Spark Streaming jobs to divide streaming data into batches as an input to Spark engine for data processing. ? Constructed a Kafka broker with proper configurations for the needs of the organization using big data. ? Hands-on big data experience in writing spark dataframes to NoSQL databases like Cassandra. ? Responsible for building quality for big data transfer pipelines for data transformation using Kafka, Spark, Spark Streaming, and Hadoop. ? Able to design and develop new systems and tools to enable clients to optimize and track using Spark. ? Worked with highly available, scalable and fault tolerant big data systems using Amazon Web Services (AWS). ? Provide end-to-end data solutions and support using Hadoop big data systems and tools on AWS cloud services as well as on-premise nodes. ? Well versed in big data ecosystem using Hadoop, Spark, Kafka with column-oriented big data systems such as Cassandra and HBase. ? Implemented Spark in EMR for processing Big Data across our Data Lake in AWS System ? Worked with various file formats (delimited text files, click stream log files, Apache log files, Avro files, JSON files, CSV, XML Files). ? Used, Kafka, and HiveQL scripts to extract, transform, and load the data into multiple databases. ? Able to perform cluster and system performance tuning on big data systems. Work Experience BIG DATA ENGINEER TECHFIELD - Atlanta, GA January 2019 to Present The data science department at the company is building a project to analyze sentiment data. A streaming pipeline is used, and data is processed and stored in distributed file systems. As a Big Data engineer, I was in charge of Build and Prepare the data pipelines for the data science team consumption. Worked with Apache Spark which

provides fast and general engine for large data processing integrated with functional programming language Scala. Created a Kafka broker in structured streaming to get structured data by schema using case classes. Used Spark Structured Streaming to structure real time data frame and update it in real time. Integrated Kafka and Spark with JSON for serializing and deserializing data, and for Kafka producer and consumer. Fine-tuned resources for long-running Spark Applications to utilize better parallelism and executor memory for more caching. Wrote custom SQL queries and hooked dataframes into larger Spark applications. Streamed data into Spark using Kafka to compare micro-batch and structured streaming. Integrated Kafka with Spark streaming for high speed data processing. Applied the latest development approaches including applications in Spark using Scala. Integrated Spark code into the SDLC with the CI/CD pipeline using Jenkins CI with Git versioning. Implemented Spark procedures of feature engineering for data science team using the in-memory computing capabilities like Apache Spark written in Scala. Configured Kafka broker for the Kafka cluster of the project and streamed the data to Spark for structured streaming to get structured data by schema. Handled over millions of messages funneled through Kafka topics. Created and optimized multiple Kafka brokers to handle message retention and deliveries Connected multiple consumers to kafka to maximize parallelism Worked with Jenkins CI for CICD and Git version control. Spark used in optimizing ETL jobs to reduce memory and storage consumption. Created Spark SQL to create real-time processing of structured data with Spark Streaming processed through structured streaming. DISTRIBUTED SYSTEMS DEVELOPER (Big Data) Gattic, Inc - Long Island, NY October 2018 to January 2019 Consumed JSON / CSV information using C++ and sockets API to consume streaming from polygon.io Created custom consumer similar to Kafka consumer to ingest real time data from polygon API, IEX API Consumed up to 6 megabytes of data per second during peak data ingestion Created API wrapper in C++ for all endpoints for polygon and IEX data providers Used TravisCI in conjunction with github for continuous deployment (CI/CD) Helped create proprietary NoSQL data storage similar to CSV storage with schema to store financial market data. Designed NoSQL queries based on Network Request database like Cassandra and HBASE Designed execution and backend testing engine

using data storage for machine learning simulations. Serialize network packets from server dump file to CSV file using C++ and dataframes Used cmake for code linking and compilation similar to SBT and Maven Contributed in the creation of a proprietary NoSQL time-series database with similar structure like Mongo Re-wrote existing batch processing framework similar to Apache Spark that processed gigabytes of financial data in batches through the network Created Similar to Spark-API for loading data into memory for machine learning Created multi-threaded application managing data transmission through the network based on the Kafka consumer/producer principles Performed ETL process to deliver results and reports for upper management Worked PCI/PII information standards to deliver accurate data Extracted information from multiple sources to create nested data from previous transformations Used rapidJSON library to parse large structured json data into dataframes-like C++ code Used proprietary distributed system to send and retrieve data similar to Hadoop cluster with spark Used github for version control Used github for collaboration with colleagues for development Benchmarked multiple Processing tools like Spark Flink to deliver a rapid deployment platform PYTHON DATA DEVELOPER YVNT, Inc - New York, NY August 2016 to September 2018 Worked on Amazon Web Services (EC2, ELB, VPC, S3, CloudFront, IAM, RDS, Route 53, CloudWatch, SNS) Experience in supporting dozens of Amazon AWS implementations including Amazon EC2 (IaaS) and all Amazon RDS (DBaaS) offerings Designed and coded unit testing procedures and provided production application support. Computed functional requirements to define technical design using Python. Contributed to reduction of run time errors by designing and executing automation procedures. Utilized agile methodologies to support ongoing improvement of processes. Built processes and software tools to support data warehousing and third-party deployment of Python applications. Created web application backend in Python to handle millions of transactions Worked on front end application using bootstrap CSS Responsible for day to day defect resolution of YVNT issues Used Git for repository and version control for the codebase. Used Maven as a built tool Participated in daily scrum meeting with onshore and offshore developers. Had weekly meetings with stakeholders for new requirements or current requirements. Develop easy to use documentation for the frameworks

and tools developed for adaption by other teams Develops processing, archiving, and recovery procedures for systems. Provides production procedures for programs Built and design SQL database in AWS RDS Manipulation of shell scripts; python scripts and JavaScript during database connection Built NoSQL database in DynamoDB calculating the amount of data to be handled Created AWS EC2 Instances to deploy web applications, using the Web Console Configured AWS route-53 to host DNS for web applications Used AWS RDS and AWS DynamoDB to store information from web application Worked with github for development and collaboration, versioning Used PyCharm IDE for python development and testing Collaborated to convert Python Web Application to MeteorJS for multiplatform support Education Bachelor's in Computer Systems Engineering Rensselaer Polytechnic Institute - Troy, NY August 2012 to August 2016 Skills Hadoop, Big Data, Java, Python, Scala, Unix Shell Scripting, Agile, TDD, Unit Testing, Git, GitHub, SVN, Jenkins, Jira, Eclipse, Visual Studio, Atom, Cloudera, SQL and Kibana Additional Information TECHNICAL SKILLS PROGRAMMING Java, Python, Scala SCRIPTING Python, Unix Shell Scripting SOFTWARE DEVELOPMENT Agile, Continuous Integration, Test-Driven Development, Unit Testing, Functional Testing, Gradle, Git, GitHub, SVN, Jenkins, Jira DEVELOPMENT ENVIRONMENTS Eclipse, IntelliJ, PyCharm, Visual Studio, Atom AMAZON CLOUD Amazon AWS (EMR, EC2, EC3, SQL, S3, DynamoDB, Cassandra, Redshift, Cloud Formation, Lambda) DATABASE NoSQL: Cassandra, Hbase, Mongo | SQL: SQL, MySQL, PostgreSQL HADOOP DISTRIBUTIONS Cloudera, Hortonworks QUERY/SEARCH SQL, HiveQL, Apache SOLR, Kibana, Elasticsearch BIG DATA COMPUTE Apache Spark, Spark Streaming, Flink, SparkSQL MISC: Hive, Yarn, Spark, Spark Streaming, Kafka, Flink VISUALIZATION: Kibana, Tableau, PowerBI, Grafana FORMATS: Parquet, Avro, Orc, JSON Data Pipeline Tools Apache Airflow, Apache Camel, Apache Flink/Stratosphere, Nifi Admin Tools Oozie, Cloudera Manager, Ambari, Zookeeper, Active Directory, PowerShell

Name: Robert Nelson

Email: morrowpatrick@example.com

Phone: 001-417-577-1189x83453