Big Data / Spark Engineer Big Data / Spark Engineer Big Data / Spark Engineer - The Hartford South Portland, ME 7 years of professional experience as a software professional industry comprising of Big Data/ Hadoop development, design, deployment. Expertise in designing scalable Big Data solutions, data warehouse models on large-scale distributed data, performing wide range of analytics. Experience in building data pipelines using Bigdata tools like HDFS, Sqoop, Flume, Kafka, Spark, Scala, Hive, Impala, Oozie, YARN. Experience in building data ingestion, data processing pipelines using On Prem and Cloud services Experience in working with Spark API (Core, Sql, Streaming) using Scala. Experience in building streaming pipelines using Spark and Kafka. Experience in working with different file formats and compressions like Parquet, Avro, ORC, Snappy, LZO etc., Experience working with AWS services like S3, EMR, Data Pipeline, Step Functions, Athena and Redshift. Experience in working with NoSQL data stores like HBase. Experience in manipulating/analyzing large datasets within structured and semi structured (JSON, XML) data. Experience in writing testcases, static code analysis and CICD process using Git, Jenkins. Experience working in agile environment with tools like Rally, Clear Case and Jira. Experience in Object Oriented Analysis Design (OOAD) and development. Experience in working with Onshore/ Offshore model, code reviews and solving the defects. Strong team player with good communication, analytical, presentation and inter-personal skills. Work Experience Big Data / Spark Engineer The Hartford - South Portland, ME December 2018 to Present Responsibilities: Design and Document the new architecture and development process to convert existing ETL pipeline in to Hadoop based systems. Extensively worked on Spark Scala to prepare data for building Prediction model which will be consumed by Data Science team. Developed Streaming data pipelines using Spark Scala and Kafka. Developed a generic framework using Spark for processing/Flatten JSON data that is re-used by various applications within the enterprise. Performance tuning of Spark Applications from code, resource and data point views. Expertise in performance tuning of Spark Streaming Applications for setting right Batch Interval time, correct level of Parallelism and memory tuning. Developed a common framework to prepare the data to feed for the machine learning models. Developed Oozie Workflows for daily incremental loads, which gets data from Teradata

and then imported into hive tables. Design and performance tuning hive tables and queries from storage, file formats and query levels Used Hive to analyses the partitioned and bucketed data and compute various metrics for reporting on the dashboard. Automated the deployment process using Git, Jenkins and IBM UDeploy. Environment: Spark Scala, Cloudera, HDFS, Hive, Sqoop, Python, Agile, YARN, Teradata, Shell Scripting, Autosys, Bit Bucket and JIRA. Hadoop Developer Paypal - San Jose, CA March 2017 to November 2018 Responsibilities: Responsible for designing, implementing and testing data pipelines on the cloud using AWS Services. Extensively used Spark to read data from S3 and preprocess it and to store in back S3 again for creating tables using Athena Designed and developed Spark application to read data json data from REST API's. Extensively used EMR, S3, Data pipeline and Step Functions for building data pipelines Created partitioned tables in Athena, also designed a data warehouse using Athena external tables and also created Athena queries for analysis. Responsible for designing and implementing the data pipeline using Big Data tools including Spark and Sqoop. Worked with different source file formats and destination source file formats like Parquet and ORC. Experience in performance tuning of long running spark applications by looking into Spark UI. Implemented the Spark Best practices to efficiently process data to meet ETAs by utilizing features like partitioning, resource tuning, memory management and Check pointing features. Used versions controls tools such as GitHub to pull data from Upstream to local branch, check conflict, cleaning also reviewing the codes of other developers. Worked on POC for exploring cutting-edge technologies in Big Data open source tools to make existing process in efficient manner. Environment: Athena, EMR, S3, Data pipeline, Step Functions, Sqoop, Spark, Scala, Linux, SQL Server, Data Warehouse and Tableau. Hadoop Developer Signa - Bloomfield, CT April 2016 to February 2017 Responsibilities: Responsible for developing solutions by working closely with Solution Architects and Business teams. Document technical and business requirements and develop architectural diagrams. Developed the code for Importing and exporting data into HDFS and Hive and Impala using Sqoop. Extensive experience in working with Hive and Impala for designing tables. Worked on data science project life cycle and actively involved in phases, data acquisition, data cleansing and data preparation. Developed an

ingestion module to ingest data into HDFS from heterogeneous data sources. Built distributed in-memory applications using Spark and Spark SQL to do analytics efficiently on huge data sets. Developed HIVE and Impala queries for the Data Transformation and Data analysis. Developed Oozie workflows and sub workflows to orchestrate the Sqoop scripts, hive queries and the Oozie workflows are scheduled through Autosys. Environment: Hive, Sqoop, Spark, Python, Scala, Linux, Impala, SQL Server Research Graduate Assistant North Western University - Fremont, CA August 2014 to December 2015 Responsibilities: Handle the installation and configuration of a Hadoop cluster. Responsible for analyzing and cleansing raw data by performing Hive queries. Created Hive tables, loaded data and wrote hive queries that run within the map. Extracted the data from RDBMS into HDFS using Sqoop and vice versa. Implemented Partitioning, Dynamic Partitioning and Bucketing in Hive. Software Engineer / Python Developer Infosys - Hyderabad, Telangana May 2012 to July 2014 Responsibilities: Involved in Requirements gathering, Requirement analysis, Design, Development, Integration and Deployment. Developed entire frontend and backend modules using Python on Django Web Framework. Developed Python batch processors to consume and produce various feeds. Wrote and executed various MYSQL database queries from python using Python-MySQL connector and MySQL dB package. Utilized PyUnit, the Python unit test framework for testing the functionality of the application. Environment: Python, Django, MySQL, PyUnit, Git and Linux. Education Master's in Computer Science in Computer Science Northwestern University December 2015 Bachelor of Engineering in Computer Science in Computer Science JNTU - Hyderabad, Telangana May 2012 Skills Impala, Mapreduce, Oozie, Sqoop, Hbase

Name: Brendan Bailey

Email: steven00@example.com

Phone: 489-427-5151x8692