**SUMMARY OF QUALIFICATIONS**

* Hadoop Certified Professional with experience in Analysis, Design, Development, Integration and Implementation of Database concepts.
* Experiencedin SQL, Unix, Java, Python, MySQL Server, Hadoop, MapReduce, Hive, Pig, Sqoop, Flume, Oozie, Impala, PySpark, Spark SQL, REST.
* In-depth understanding of the Hadoop Architecture and Spark Architecture.
* Have written bespoke MapReduce programs and used UDFs for Pig and Hive.
* Excellent experience in Microsoft Azure, Amazon Web Services, Digital Ocean, and other web services.
* Utilized several development environments such as Eclipse, PyCharm, RStudio.
* Store and load data in a variety of formats such as Avro, Parquet, ORC with Java programs, Pig, and Hive.
* Experience working with variousdistributions such as Apache BigTop, Hortonworks, Cloudera, Amazon Elastic MapReduce.
* Effective working independently and collaboratively in teams.

**TECHNICAL SKILLS**

**Languages:** MySQL, Unix, Java, Python

**Big Data Ecosystem:** HDFS, MapReduce, Hive, Sqoop, Flume, Oozie, Impala,HBase,Kafka, PySpark, Spark Streaming, Spark SQL, Pig

**Source Control:** Github

**Relational Databases:** SQL Server, Oracle

**Operating Systems:** Red Hat Linux, Windows 7, 10

**Statistical Software:** Excel, Clear Analytics, R, Pandas Library, Stata, PASW Statistics

**Certifications**

* Hortonworks Data Platform Certified Developer

**PROFESSIONAL EXPERIENCE**

**Hudson Data, New York, NY July 2016 – Aug 2016**

Hadoop Developer

**TPC-H Benchmark Testing**

Following the Transaction Processing Performance Council’s guidelines, queries were ran on Hive, Hawq, and Spark SQL to compare database performance against VectorH for Progressive Insurance. The data was randomly generated by a database generator. Multiple runs were performed on database sizes ranging from 1 gigabyte to 500 gigabytes.

**Responsibilities**

* Alter SQL queries that are not supported by Hive 1.2 to run and return the desired results
* Change Hive and Spark configurations for optimal performance
* Automated submission of queries using a combination of shell scripts via Python programming
* Export results to a Tableau dashboard and have dashboard display information in semi-real time to a website

**Environment:** Hortonworks HDP 2.4, Cloudera Hadoop, Hive 1.2, Spark SQL 2.1, Spark SQL 1.6, Centos 6.5, Python 2.6, Tableau,Actian Vector-H, Hawq

**Ana-Data Consulting, Inc., Jersey City, NJ April 2016 – July 2016**

Hadoop Developer

**Clear Analytics Cloud**

Enhancing the Clear Analytics cloud implementation to support much broader range of cloud databases including Salesforce, CRM Dynamics, QuickBooks, Google/Yahoo/Bing analytics and Azure.

**Responsibilities:**

* Utilized UDFs and custom SerDes to expand Hive capabilities
* Employed PySpark to analyze data in a fast manner by performing actions and transformations on RDDs
* Used Spark SQL and its DataFrame API to quickly analyze large sets of data
* Optimized query performance by using Static and Dynamic Partitioning, and Bucketing for sampling with Hive
* Configured Flume to import log data from servers in real-time into HDFS
* Imported data from RBDMS into HDFS and exported data from HDFS to RDBMS with Sqoop
* Constructed Oozie coordinator jobs based on data availability and time to automate ETL
* Performed Hadoop Streaming with Python Scripts

**Environment:** Cloudera Hadoop, Linux, HDFS, Hive, Oracle, Putty, Java, MapReduce, Pig, Sqoop, WinSCP, Eclipse

**American Express, New York, NY Sept 2015– April 2016**

Hadoop Developer

**Out-Bound Telemarketing Application (OBTM)**

Implemented the data analytics solution to integrate feeds from different vendors to support the Commercial Telemarketing Cross Channel Acquisition Systems.  Analytics were performed for pre-screening the prospective customers to increase the marketing success and eliminating the high risk prospects. Developed end to end automation to support telemarketing acquisition strategies via leveraging the Big Data production platform.

**Responsibilities:**

* Migrated the business logic originally developed in SAS to a Hadoop MapReduce/Hive Framework.
* Implemented the scheduler module for feeds integration.
* Implemented the missing value imputation, pre and post-processing logic for the scoring module.
* Used and combined different languages (Java, Python) and technologies (Hadoop, MapReduce, Hive) to reduce the runtime.
* Develop scripts using Spark in Python to perform Data Quality checks and transform data for OBTM operational reporting Tables
* Convert existing ETL in PIG scripts for Call Center reporting to Spark in Python.
* Design and load data into Hive Tables using HiveQL and Parquet format
* Design and develop Hive views to be accessed by reporting team.
* Performance tuning of hive tables and used Oozie for scheduling jobs.
* Reduced the telemarketing lead time from 5 days to 4 hours.
* Provided support for the production platform.

**Environment:** Java, Python**,** Hadoop MapReduce/Hive Framework, SAS

**Ana-Data Consulting, Inc., Jersey City, NJ July 2014 – Sept2015**

Developer

**Clear Analytics**

Clear Analytics is a data consolidation tool, Business Intelligence and Reporting tool. Clear Analytics consolidates data from a variety of sources and brings that into Excel and enables users to utilize all ofExcel’s functions.

**Responsibilities**

* Designed and developed production using C#, WebAPI. Also working on enhancement of powerful self-service Business Intelligence tool.
* Worked on middle tier to extra data from database and creating wrapper classes to communicate between UI and business layer.
* Designed customizable dashboard using AngularJS,HTML5
* Retrieved data from SQL Server database and displaying that on Excel.
* Supported requirements and maintenance in Web application.
* Handled calculations and manipulations of reports at backend process.

**Environment:** Visual Studio .NET 2015, WCF Service, SQL Server 2010, Angular JS, HTML5, WebAPI, LINQ, TFS