Data Science Workloads for Big data Benchmarking

Milind Bhandarkar
Greenplum, A division of EMC

Big Data has been characterized by three V's: Volume, Variety, and Velocity. This has implications about the scalability, ability to process different data genres, with high performance. Also, the availability of new cost-effective system infrastructure has given rise to new types of workloads.

Enterprises are interested in deriving value from the variety of data sets both public and private, by extracting business insights from them. Analyzing these datasets using statistical modeling and machine learning has been described as a new discipline, called Data Science.

According to Hilary Mason, Chief Scientist at Bitly, the Data Science workloads can be described as five stages of data processing:

1. Obtain: Collect raw data from various sources. For the data corpus to be usable and sufficient, it needs to be obtained for multiple semi-independent sources. These sources could be structured databases, server logs, document corpuses etc. Efficient ingestion, and format translation is a desired quality of a data analytics workbench. In addition, some data sources need periodic or streaming ingestion.
2. Scrub: Most of the raw data sources, such as social media streams, server logs etc are by nature unstructured, and thus very "messy". For example, in twitter streams, the location indication, such as Ney York City, mugh be described as NY, NYC, Big Apple etc. A mature data science practice, typically has dictionaries that are used to "normalize" such messy data. Also, a small percent of the data is often dropped since it does not conform to the parsing algorithms.
3. Explore: Before data modeling can begin, a data scientist often explores various dimensions of the data, to discover relationship among various fields in the data set. This typically involves calculating basic statistical parameters, such as Min, Max, Median, Mean, Quartiles etc; as well as plotting histograms for single features and scatter plots among various features. Often scrubbing and exploring data is an iterative process, where outliers are dropped, and correlated features are discarded, resulting in a very structured data set on which statistical modeling can be conducted.
4. Model: Depending on the business problems at hand, a variety of statistical modeling techniques, such as linear or logical regressions, decision trees, naïve bayes classification, various clustering techniques are used. Many of these algorithms are iterative in nature. Some of the modeling techniques involve generating "ensemble models", by generating multiple models, and in the evaluation phase, attaching weights to each of these models to form a

"model of models". This is typically a very easily parallelizable computation, since each of these models could be generated from the same data independently of others.

5. Interpret: A generated model is applied to the "hold-out" data, and results are compared against expected results to determine precision-recall ratios and confidence intervals.

Most data practitioners spend a majority of time in the first three phases of the data science workflow, i.e. obtaining raw data, scrubbing them, and exploring data sets. In addition, most of them tend to work with multiple infrastructure components, depending on ready availability. For large data sets, massively parallel databases, and Hadoop are increasingly common. A diverse sets of tools, ranging from simple shell scripts, to SQL queries, to small user-defined functions written in languages such as Python, Perl, and R are commonly used to perform many of the tasks.

Any benchmark that intends to be representative of these real-world use-cases, cannot restrict itself to a single framework, query language, or toolset. Instead, it needs to provide a "paper-and-pencil" specification of the task at hand, and the implementation could be in a variety of languages, in different frameworks.