Technical Paper

Scalability of the SAS/STAT® HPGENSELECT High-Performance Analytical Procedure: A Comparison with RevoScaleR

Effectively implementing high-performance analytics software solutions in the insurance industry

Table of Contents

Executive Summary

At the Strata Conference on October 25, 2012, the research and planning division of a large insurance corporation (hereafter "insurer") presented various methods that they used to model 150 million observations of insurance data. A summary of their presentation is available at: http://blog.revolutionanalytics.com/2012/10/allstate-big-data-glm.html.

The presentation compared the following software:

- The GENMOD procedure in SAS/STAT® software
- Custom MapReduce code on a Hadoop cluster
- Open-source R
- Revolution R Enterprise using RevoScaleR

The insurer reported that PROC GENMOD took more than five hours to fit a Poisson regression model, whereas Revolution Analytics used the RevoScaleR package to fit the model in 5.7 minutes. However, this is not an "apples to apples" comparison because RevoScaleR was run on a cluster of computers, whereas the GENMOD procedure executes only on a single server.

A more informative comparison can be made by using the HPGENSELECT procedure, which had not yet been released at the time of the Strata comparison. Introduced in SAS/STAT 12.3 in June 2013, PROC HPGENSELECT runs in either single-machine mode (multiple threads on a single machine) or distributed mode (multiple threads on multiple machines). Distributed mode requires SAS® High-Performance Statistics.

Purpose

This paper compares the performance of the HPGENSELECT procedure with results cited for the RevoScaleR package by using data that are similar to the insurer's data. The paper also demonstrates the scalability of the HPGENSELECT procedure by using two sizes of data sets and three different computing environments.

Results

On a small grid with two nodes, the HPGENSELECT procedure fits a Poisson regression model with 150 million observations in 159 seconds, which is less than half the time that RevoScaleR required on a somewhat larger grid. On a grid with 140 nodes, the HPGENSELECT procedure solves the problem in 22 seconds.

The scalability of the HPGENSELECT procedure is demonstrated by increasing the size of the data set. For a data set that has the same variables and one billion observations, the procedure executes in less than one minute.

These results, which are summarized graphically in Figure 1, show that the HPGENSELECT procedure provides a faster alternative.

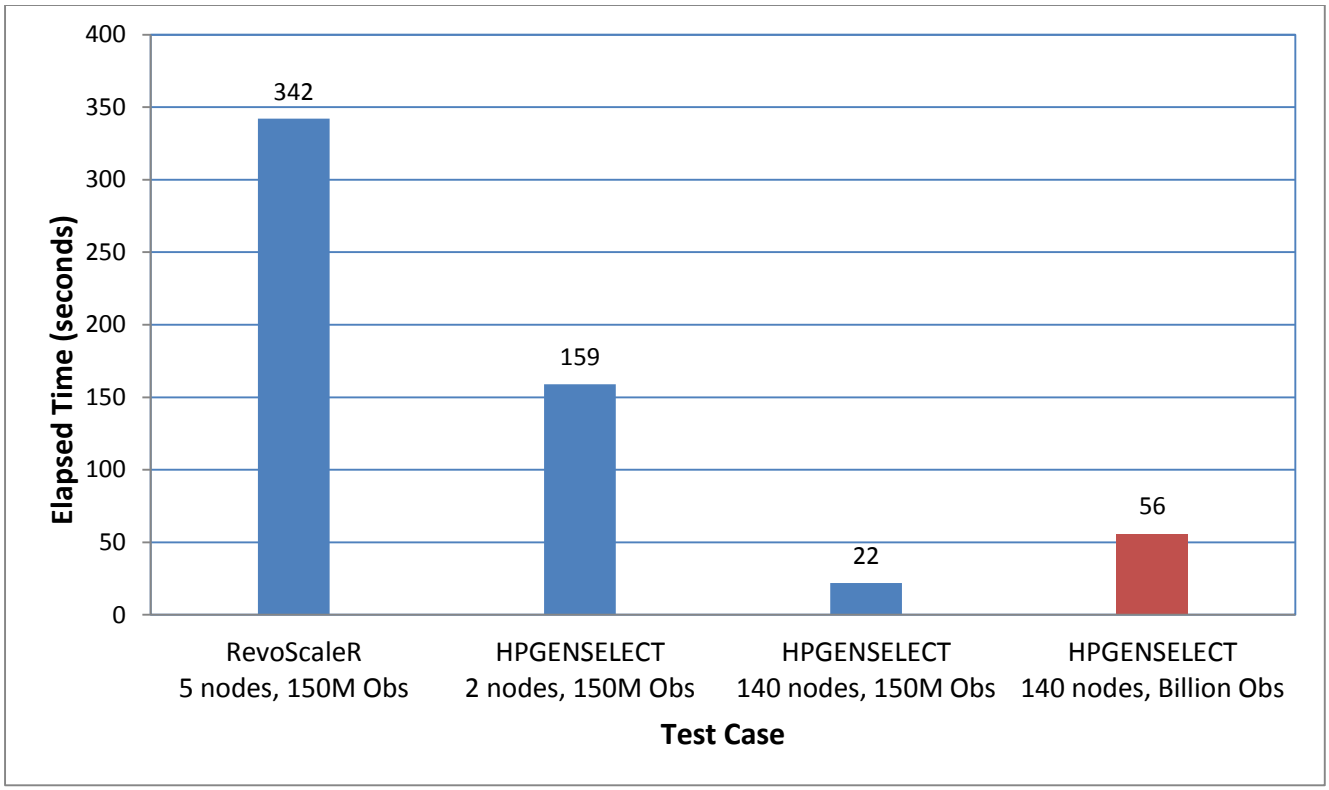

Figure 1. Times Required to Fit Poisson Model

Introduction

At the Strata conference, the insurer described four ways in which they attempted to fit a Poisson regression model to insurance data that contain 150 million observations:

- Running the GENMOD procedure in SAS/STAT® software on a 16-core Sun Server (potentially with 256 gigabytes of memory).
- Writing MapReduce code to run on a Hadoop cluster.
- Using open-source R, both on the full data set and by using a sampling strategy. The full data set could not fit in the memory, and this strategy was abandoned. The sampling strategy was time-consuming, and therefore not favored by the insurer.
- Running RevoScaleR on a five-node cluster. This was done by a Revolution Analytics employee.

The insurer found that PROC GENMOD took more than five hours to return results for a model that has approximately 70 degrees of freedom. With the RevoScaleR package in Revolution R Enterprise, the same model was fit in 5.7 minutes. However, a direct comparison of these benchmarks is misleading because the RevoScaleR package implements algorithms that execute in parallel when run on a cluster of computers, whereas the GENMOD procedure runs only on a single server.

An informative comparison can be made by using the HPGENSELECT procedure, which was introduced in SAS/STAT 12.3 in June 2013. Like all high-performance analytical procedures in SAS, PROC HPGENSELECT is designed to run either on a single workstation (single-machine mode) or in a distributed mode that uses a cluster of computers running high-performance analytical software, such as SAS® High-Performance Statistics. High-performance analytical procedures are multithreaded and can take advantage of all the cores available in either single-machine mode or distributed mode. For an introduction to high-performance procedures for statistical modeling, see Cohen and Rodriguez (2013).

The HPGENSELECT procedure provides model fitting and model building for the family of generalized linear models. It fits standard models in this family, including the Poisson regression model and the Tweedie model, which are used often in the insurance industry. PROC HPGENSELECT also fits multinomial models for ordinal and nominal responses and zero-inflated models for count data. For all these models, the HPGENSELECT procedure provides forward, backward, and stepwise variable selection. The HPGENSELECT procedure is primarily designed for large-data tasks such as predictive model building. In contrast, the GENMOD procedure excels at a broad range of inferential tasks for samples of moderate size; it also excels at postfitting analysis.

The main purpose of this paper is to highlight the scalability of the HPGENSELECT procedure by using proxy data that are similar to the data used by the insurer. The authors of this paper ran PROC HPGENSELECT in several computing environments, one of which is a small cluster that approximately compares to the one used by Revolution Analytics. The remainder of this paper describes how the data were constructed, the environments in which the HPGENSELECT procedure was run, and the results.

Construction of Proxy Data

The data modeled by the insurer and Revolution Analytics had 145.8 million observations and 139 variables. Their presentation did not discuss the details of the variables, such as their types and levels. However, the Poisson regression model is known to have 70 degrees of freedom, which indicates the number of effects in the model. For example, these degrees of freedom could have come from one categorical variable that had 71 levels, from multiple continuous variables, or from a combination of categorical and continuous variables.

Because the exact data modeled by the insurer could not be accessed, a proxy data set of representative insurance claims data was created by sampling 150 million observations with 44 variables from a data set that contained 10 billion observations. Tables 1 and 2 provide a brief description of the variables.

Table 1 *Categorical Variables*

NAME	# levels
City	71623
Country	203
Rating_Category	5
Region	17
Occupation	9
Mstatus	2
Car_Use	2
Education	5
State	50
Gender	2
Urbanicity	2
Policyno	10
Car_Type	6
DUI	2
Parent1	2
Claim_Flag	2

Table 2 *Continuous Variables*

NAME	MISS	N	MIN	MAX	MEAN	STD
Age	0	150000000	16	81	42.12	11.61
Bluebook	0	150000000	5000	120000	40151.31	25533.38
Car_Age	0	150000000	1	19	4.97	2.57
Clm_Freq **	0	150000000	0	5	1.46	1.24
Home_Val	3	149999997	0	1000000	160793.25	193828.56
Income	0	150000000	0	500000	67423.80	83843.97
Kidsdriv	0	150000000	0	3	0.62	0.93
Prem_Adjustment	0	150000000	-150	536	24.90	99.25
Premium	0	150000000	250	6000	2007.57	1276.67
Proj_Premium	0	150000000	100	6448	2032.47	1280.21
Risk_Score	0	150000000	1	1371	349.81	198.49
Travtime	0	150000000	0	40	23.44	12.03

The target variable is claim frequency, which is marked with ** in Table 2. This variable counts the number of claims made by a policy holder, with results ranging from 0 to 5 claims.

The Poisson regression model for these data has exactly 70 degrees of freedom. The model includes the categorical variables listed in Table 1 and the continuous variables listed in Table 2.

This paper compares the performance of the HPGENSELECT procedure with results cited for the RevoScaleR package by using data similar to those in the insurer's exercise. The paper also demonstrates the scalability of the HPGENSELECT procedure by using two sizes of data sets and three different computing environments.

The HPGENSELECT Procedure

SAS/STAT® high-performance procedures such as PROC HPGENSELECT provide predictive modeling tools that have been specially developed to take advantage of parallel processing in both single-machine mode and distributed mode. Furthermore, PROC HPGENSELECT uses all the memory available on the server nodes. Therefore, for large data sets, using PROC HPGENSELECT is more appropriate than using PROC GENMOD.

You can use high-performance statistical procedures in single-machine mode without any additional license required. However, to run these procedures in distributed mode, you must purchase SAS High-Performance Statistics. Performance results from both distributed mode and single-machine mode are shown in the next section.

The following statements illustrate the basic syntax of the HPGENSELECT procedure for fitting a Poisson regression model:

```
proc hpgenselect data=<dataset-name>;
            class <classification-variables>;
            model <target-variable> = <classification-variables> <other-variables> /
           dist=Poisson;
       run;
```

The syntax of the HPGENSELECT procedure is similar to the syntax of the GENMOD procedure. The HPGENSELECT procedure also enables you to specify performance-related parameters, such as the number of threads on which to run the procedure. For details, see Cohen and Rodriguez (2013).

Results

Performance in Distributed Mode

The analysis of 150 million observations was first conducted on a cluster of 144 nodes. Each node has two sockets of Xeon E5-2680 CPUs running at 2.7 GHz. Each CPU has eight cores and supports Intel Hyper-Threading technology. Each node is equipped with 256 gigabytes of main memory and a 640-gigabyte hard drive. Each computational node in this configuration provides slightly less computational power and more memory than the entire five-node cluster on which RevoScaleR was run.

The data were loaded in SASHDAT format onto a Cloudera CDH 4.2 Hadoop server on the cluster. The data have a footprint of approximately 70 gigabytes in the Hadoop file system, with a block size of 512 megabytes. The analysis was also run with a much larger sample of one billion observations, which uses more than 500 gigabytes in the Hadoop file system and is also stored with a block size of 512 megabytes.

The results for these analyses are provided in Table 3 and are illustrated in Figure 1.

Table 3 *Scalability and Timing of HPGENSELECT procedure*

Timing (Seconds)	Nodes	Observations
22	140 nodes	150 million
159	2 nodes	150 million
56	140 nodes	1 billion

Performance in Single-Machine Mode

The analysis for 150 million observations was also performed by running PROC HPGENSELECT on a single machine. This machine has two sockets of Xeon E5-2667 CPUs running at 2.90 GHz, each with six cores and 128 gigabytes of

main memory. Because the CPUs support Intel Hyper-Threading technology, the machine can run a maximum of 24 threads concurrently.

Table 4 provides the elapsed times for different numbers of threads. Even on one machine, the HPGENSELECT procedure can solve the problem in five minutes.

Table 4 *Timing of HPGENSELECT Procedure on Varying Threads*

Thread Count	Average Elapsed Time (Minutes)
4	11.9
6	8.5
12	6.9
18	5.4
24	5.1

Conclusion

The SAS/STAT HPGENSELECT procedure provides a performance benchmark that compares well with that of Revolution Enterprise's RevoScaleR package. For a data set that contains 150 million observations, the HPGENSELECT procedure fits the Poisson model considerably faster than the RevoScaleR package. The scalability of the procedure is demonstrated by increasing the size of the data set. For a data set that contains one billion observations, the HPGENSELECT procedure executes in under one minute. Although these comparisons were not done in an identical environment and were not done with identical data, these results indicate that the HPGENSELECT procedure is a faster alternative than the RevoScaleR method on a comparable data set.

References

Cohen, R. and R. Rodriguez. 2013. “High-Performance Statistical Modeling.” *Proceedings of the SAS Global Forum 2013 Conference*. Cary, NC: SAS Institute Inc. *Available at: **http://support.sas.com/resources/papers/proceedings13/401-2013.pdf**.*