Extending BlinkDB With Fundamental Analytical Operations: Linear Regression

Nathan Harada, Julian Katz-Samuels

University of Michigan
{nharada, jkatzsam}@umich.edu

I. Evaluation

I. Evaluation Datasets

We evaluated our extensions to BlinkDB via a generated synthetic benchmark as well as the bike sharing dataset from the University of California Irvine Machine Learning Repository (UCI-ML), a collection of standard and well studied datasets available to the research community. The synthetic dataset was generated in Matlab via the equation $y = X\beta + \varepsilon$, where $X \sim \mathcal{N}(0, 1)$ and ε represents additive white Gaussian noise with a signal to noise ratio of 0dB. We fixed the values fo β to $[1, -2, 3]$.

II. Evaluation Setting

Experiments were perfomed on a single-node system, configured with 8 GB of RAM, 6 CPU cores (2.67 GHz), 1TB of disk running Ubuntu 14.04. BlinkDB alpha release 0.2 was used as the base software, along with Scala 2.10 and Spark SQL 1.1.0. To evaluate the correctness of the implemented algorithms, we used IBM SPSS Statistics 21.

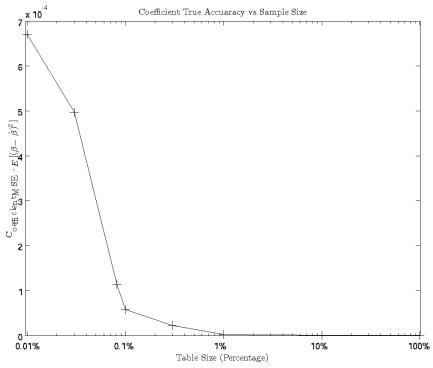
III. Synthetic Dataset

The full synthetic dataset consists of 10 million rows of data, each consisting of independent variables x, y, z and dependent variable u. The dataset was sampled with various sampling probabilities, with the true sample sizes listed below:

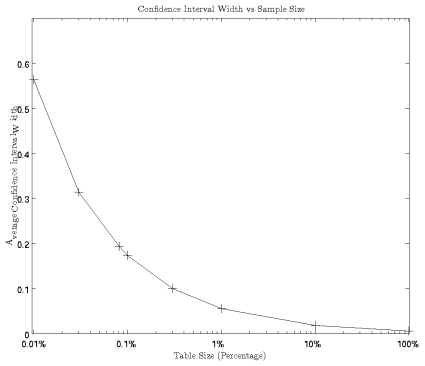
Sampling Probability	Resulting Sample Size
0.0001	999
0.0003	3000
0.0008	8112
0.001	9950
0.003	29967
0.01	100390
0.1	999682
Full	10000000

To test the accuracy of our system, we first compared the true coefficient values with the estimated coefficients. The mean squared error between β and $\hat{\beta}$ was calculated for each of the sample sizes. We show that as the sample size increases, the values of $\hat{\beta}$ asymptotically approach the true values of β. Error rate drops quickly as sample size increases, with a fraction of a

percentage of the original dataset presenting nearly accurate results. We repeated the experiment for the calculated confidence intervals on each sample. Again, as sample size increases, the confidence interval width shinks, approaching zero as the sample size grows large.

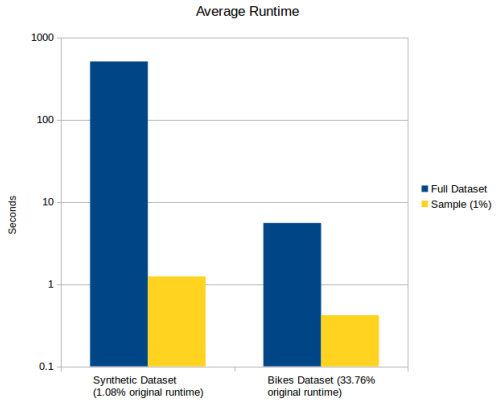


(a) *Coefficient accuracy vs Sample Size (log scale)*

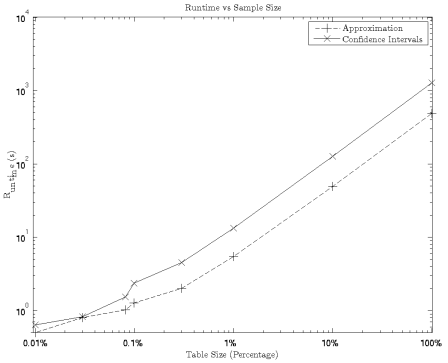


(b) *Confidence Interval Width vs Sample Size (log scale)*

We additionally recorded and plotted the runtime for each query. Runtime appears to scale linearly with table size, meaning that a user who wishes to perform an approximate linear regression can easily trade off error for time to achieve reasonable confidence intervals within an interactive time frame.



(a) *Query Runtime on a synthetic and real world dataset*



(b) *Query Runtime (log scale) vs Sample Size (log scale)*

IV. Bike Sharing Dataset

The bike sharing dataset contains the hourly and daily counts of bike rentals for Washington DC's bikeshare program, along with relevant weather data such as humidity and windspeed. We evaluated the runtime performance of our system on this dataset, and compared it to the performance of our larger synthetic dataset. While both datasets achieved a significant increase in speed, the real world dataset saw less of an improvement compared to our synthetic data. We suspect this is because, while we only chose a few independent variables to regress, the full

dataset contains many more columns. This additional overhead likely resulted in additional disk overhead that limited our performance gains.

References

[1] Agarwal, Sameer, et al. "BlinkDB: queries with bounded errors and bounded response times on very large data." Proceedings of the 8th ACM European Conference on Computer Systems. ACM, 2013.

[2] Brown, Paul G. "Overview of SciDB: large scale array storage, processing and analysis." Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. ACM, 2010.