Preview of Award 1320357 - Annual Project Report

Cover

Federal Agency and Organization Element to Which Report is Submitted:	4900
Federal Grant or Other Identifying Number Assigned by Agency:	1320357
Project Title:	III: Small: Cumulon: Easy and Efficient Statistical Big-Data Analysis in the Cloud
PD/PI Name:	Jun Yang, Principal Investigator Shivnath Babu, Co-Principal Investigator Sayan Mukherjee, Co-Principal Investigator Michael D Ward, Co-Principal Investigator
Recipient Organization:	Duke University
Project/Grant Period:	09/15/2013 - 08/31/2016
Reporting Period:	09/15/2013 - 08/31/2014
Submitting Official (if other than PD\PI):	N/A
Submission Date:	N/A
Signature of Submitting Official (signature shall be submitted in accordance with agency specific instructions)	N/A

Accomplishments

*** What are the major goals of the project?**

"Big data" have been growing in volume and diversity at an explosive rate, bringing enormous potential for transforming science and society. Driven by the desire to convert messy data into insights, analysis has become increasingly statistical, and there are more people than ever interested in analyzing big data. The rise of cloud computing in recent years, exemplified by the popularity of services such as Amazon EC2, offers a promising possibility for supporting big-data analytics. Its "pay-as-you-go" business model is especially attractive: users gain on-demand access to computing resources while avoiding hardware acquisition and maintenance costs.

However, it remains frustratingly difficult for many scientists and statisticians to use the cloud for any nontrivial statistical analysis of big data. First, developing efficient statistical computing programs requires a great deal of expertise and effort. Popular cloud programming platforms, such as Hadoop, require users to code and think in low-level, platform-specific ways, and, in many cases, resort to extensive manual tuning to achieve acceptable performance. Second, deploying such programs in the cloud is hard. Users are faced with a maddening array of choices, ranging from resource provisioning (e.g., type and number of machines to request on Amazon EC2), software configuration (e.g., number of parallel execution slots per machine for Hadoop), to execution parameters and implementation alternatives. Some of these choices can be critical to meeting deadlines and staying within budget, but current systems offer little help to users in making such choices.

This project aims to build Cumulon, an end-to-end solution for making statistical computing over big data easier and more efficient in the cloud. When developing data analysis programs, users will be able to think and code in a declarative fashion, without being concerned with how to map data and computation onto specific hardware and software platforms. When deploying such programs, Cumulon will present users with best "plans" meeting their requirements, along with information that is actually helpful in making decisions---in terms of completion time and monetary cost. For example, given a target completion time, Cumulon can suggest the best plan on Amazon EC2 that minimizes the expected total cost. A plan encodes choices of not only implementation alternatives and execution parameters, but also cluster provisioning and configuration choices. This project will develop effective cost modeling and efficient optimization techniques for the vast search space of possible plans. Once a plan is chosen, Cumulon automatically takes care of all details, including reserving hardware, configuring software, and executing the program.

*** What was accomplished under these goals (you must provide information for at least one of the 4 categories below)?**

Major Activities:

To date, we have completed two iterations of system development for Cumulon. The first version demonstrates efficient implementation and intelligent cost-based optimization of matrix-based data analysis workloads on Amazon EC2, using Hadoop/HDFS for underlying execution/storage. The optimization problem is formulated using two criteria clearly understandable to end users: expected execution time and expected monetary cost.

The second version of Cumulon further demonstrates the ability to leverage auction-based markets (in our case, Amazon spot instances) of cloud resources to lower execution costs. Cumulon can make intelligent choices of bidding strategies, and recover gracefully from sudden, massive departure of transient machines acquired through bidding. The optimization problem is extended to additionally include a user-specified risk tolerance requirement, because the expected cost alone is unable to capture the risk due to the volatility of auction-based markets.

To support cost-based optimization, we have also been working on statistical methods for performance prediction and uncertainty quantification. Besides performance, we also model the market price of transient resources using the price history of Amazon spot instances. These problems are interesting and useful in their own right.

Finally, we have been working on applying Cumulon to concrete data analysis problems of interest to statisticians and political scientists.

Specific Objectives:

At the end of Year 1, we have met our main objectives so far of building the Cumulon prototypes, evaluating their effectiveness, and learning from this experience.

Significant Results:

We have shown that the popular MapReduce programming model is a poor fit for matrix workloads. We have developed a simplified parallel execution model that is more efficient, flexible, and easier for automatic optimization. Instead of reinventing the wheel, we have demonstrated how to implement this new model on top of MapReduced-based Hadoop and HDFS, but in a way that avoids the limitations of MapReduce. This novel strategy offers large savings over more "traditional" uses of Hadoop for matrix workloads.

Additionally, we have developed storage and execution techniques capable of leveraging transient machines acquired through bidding on a market. Our techniques handle heterogeneous clusters that result from a combination of upfront provisioning and runtime bidding, and gracefully cope with sudden and massive departure of transient machines to reduce their adverse effects on execution costs.

We have learned how to obtain good cost estimates for matrix workloads in a cloud by benchmarking, simulation, and modeling. Various sources of uncertainty in the cloud make cost estimation particularly challenging. We have devised a better method for predicting job completion time, by accounting for the variance among the speeds of individual tasks. We have also developed a stochastic market price model based on historical prices of Amazon spot instances, which allows us to predict departures and their effects on overall execution costs, and to quantify the uncertainty in the predictions.

Through experiments on Amazon EC2, we have demonstrated the benefit of considering a bigger "plan" space whose dimensions include the choices of not only implementation alternatives and execution parameters, but also hardware provisioning and bidding strategies as well as software configuration settings. We have developed optimization techniques that consider time, monetary cost, and user-specified risk tolerance.

Key outcomes or Other achievements:

A paper on the first iteration of Cumulon has been published in SIGMOD 2014. Our paper on the second iteration, focusing on leveraging auction-based markets, is currently under submission. We were also invited to submit a paper on Cumulon to an upcoming special issue of IEEE Data Engineering Bulletin on the topics of databases and large-scale machine learning; the submission is currently under review. A paper providing details of modeling runtime distributions in Cumulon, intended for the statistics community, is current under preparation.

Jun Yang co-organized a panel about big data at VLDB 2014 titled "Big and Useful: What's in the Data for Me?" Jun is also co-chairing the First International Workshop on Bringing the Value of "Big Data" to Users, held in conjunction with VLDB 2014.

We have been working on testing Cumulon in application domains, and hope to be able to report on the outcomes soon.

*** What opportunities for training and professional development has the project provided?**

Botong Huang, the lead computer science PhD student on this project, and Nick Jarrett, the lead statistics PhD student on this project, have been working closely with each other; they have learned much from each other's discipline as well as about practical cloud computing skills.

Botong is interning at IBM Almaden Research Lab in the summer of 2014, working on SystemML, a system being developed at IBM with similar goals as Cumulon (albeit from a provider's perspective instead of users'). He is learning a lot from this industry perspective, and is helping with the exchange of ideas between academia and industry.

We have redesigned the undergraduate database curriculum ("Introduction to Database Systems" and "Everything Data") at Duke with heavy use of cloud-based virtual machines. The students have been introduced to cloud computing concepts and skills.

*** How have the results been disseminated to communities of interest?**

A paper on the first iteration of Cumulon has been published in SIGMOD 2014; Botong Huang delivered a well-received talk at the conference, and was invited to visit the SystemML group at IBM Almaden. Botong is now interning with the group in the summer.

Our paper on the second iteration, focusing on leveraging auction-based markets, is currently under submission. We have also been invited to submit a paper on Cumulon to an upcoming special issue of IEEE Data Engineering Bulletin on the topic of databases and large-scale machine learning; that submission is currently under review.

Jun Yang co-organized a panel about big data at VLDB 2014 titled "Big and Useful: What's in the Data for Me?" In the

panel, Jun argued that "democratizing data analysis" is as important as "democratizing data," and that we need more systems and research like Cumulon that are more user-facing and user-friendly. To further promote such research, Jun is co-chairing the First International Workshop on Bringing the Value of "Big Data" to Users, held in conjunction with VLDB 2014.

Cumulon's poster has been used in Duke Computer Science Department's graduate recruiting events.

*** What do you plan to do during the next reporting period to accomplish the goals?**

We have identified many directions and are currently exploring them before deciding which one to tackle next. Besides continuing to work with statisticians and scientists to explore more workloads and new applications, the possible next steps include better stochastic modeling of performance and market, a repository of execution traces and market prices, online optimization, and "deep extensibility" (extending not only Cumulon's functionality with new operators but also their optimizability).

Products

Books

Book Chapters

Conference Papers and Presentations

Botong Huang and Shivnath Babu and Jun Yang (2013). *{Cumulon}: Optimizing Statistical Data Analysis in the Cloud*. Proceedings of the 2013 {ACM} {SIGMOD} International Conference o n Management of Data. New York City, New York, USA. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Inventions

Journals

Licenses

Other Products

Other Publications

Botong Huang, Nicholas W.D. Jarrett, Shivnath Babu, Sayan Mukherjee, and Jun Yang (2014). *Cumulon: Cloud-Based Statistical Analysis from Users' Perspective*. Submitted to IEEE Data Engineering Bulletin (invited).. Status = UNDER_REVIEW; Acknowledgement of Federal Support = Yes

Botong Huang, Nicholas W.D. Jarrett, Shivnath Babu, Sayan Mukherjee, and Jun Yang (2014). *Leveraging spot instances for cloud-based statistical data analysis*. Technical report, Duke University, July 2014.. Status = OTHER; Acknowledgement of Federal Support = Yes

Patents

Technologies or Techniques

Thesis/Dissertations

Websites

Participants/Organizations

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked

Yang, Jun	PD/PI	4
Babu, Shivnath	Co PD/PI	1
Mukherjee, Sayan	Co PD/PI	1
Ward, Michael	Co PD/PI	0
Huang, Botong	Graduate Student (research assistant)	9
Jarrett, Nicholas	Graduate Student (research assistant)	4

Full details of individuals who have worked on the project:

Jun Yang
Email: junyang@cs.duke.edu
Most Senior Project Role: PD/PI
Nearest Person Month Worked: 4

Contribution to the Project: As the PI, Jun Yang leads the team in building Cumulon, and is responsible for data management for the project.

Funding Support: NSF-IIS-09-16027 Research grant from Amazon Web Services

International Collaboration: No
International Travel: Yes, Italy - 0 years, 0 months, 7 days; Mexico - 0 years, 0 months, 7 days

Shivnath Babu
Email: shivnath@cs.duke.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: As co-PI, Shivnath Babu leads the design and implementation of the Cumulon system together with Jun Yang, and co-advises the lead computer science PhD student on the project.

Funding Support: None other

International Collaboration: No
International Travel: No

Sayan Mukherjee
Email: sayan@stat.duke.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: As co-PI, Sayan Mukherjee oversees the statistical aspects of the project. He supervised work on statistical models of runtime as well as those of spot auction pricing.

Funding Support: None other.

International Collaboration: No

International Travel: Yes, United Kingdom - 0 years, 0 months, 7 days; Germany - 0 years, 0 months, 8 days; Canada - 0 years, 0 months, 6 days; Canada - 0 years, 0 months, 8 days

---

Michael D Ward
Email: michael.d.ward@duke.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 0

Contribution to the Project: As co-PI, Mike Ward works on statistical methods in political and social sciences, and will apply the proposed research and evaluate the system in problems from these domains. (Because of his sabbatical, his effort in Year 1 is 0 month.)

Funding Support: None other

International Collaboration: No
International Travel: No

Botong Huang
Email: bhuang@cs.duke.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 9

Contribution to the Project: Botong is the lead computer science student on the project, and the main developer of the Cumulon system.

Funding Support: NSF-IIS-09-16027

International Collaboration: No
International Travel: Yes, China - 0 years, 1 months, 0 days

Nicholas W.D. Jarrett
Email: nwj2@stat.duke.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 4

Contribution to the Project: Nick Jarrett is the lead student on the statistical aspects of the projec. He worked on developing the various statistical models used in Cumulon.

Funding Support: NSF DMS-12-09155: Collaborative: Numerical Algebra and Statistical Inference

International Collaboration: No
International Travel: No

What other organizations have been involved as partners?
Nothing to report.

Have other collaborators or contacts been involved? No

Impacts

What is the impact on the development of the principal discipline(s) of the project?

Cumulon has helped to demonstrate the power of declarative languages and automatic, cost-based optimization---stables of the database systems---in the new setting of matrix-based data analysis workloads in the cloud. Although there has been work on automatic optimization of data-parallel workloads in the cloud, what distinguishes Cumulon from others is its focus on the users' perspective, and its ability to provide an end-to-end solution. Cumulon's search space includes not only the traditional dimensions of an execution plan (e.g., alternative implementations and execution parameters) but also novel ones (e.g., software configuration parameters, cluster provisioning and bidding strategies). Cumulon's optimization criteria are also user-centric (time, monetary cost, and risk tolerance) as opposed to provider-centric (e.g., overall throughput). We are hopeful that this new perspective will open up new research directions.

What is the impact on other disciplines?

Cumulon makes principled use of statistics towards modeling and handling of uncertainty. As a result, it has motivated new applications and research questions of interest to the statistics community.

We have been working on testing Cumulon in application domains of other disciplines, and hope to report on its impact soon.

What is the impact on the development of human resources?

So far, the project has supported training and development of two PhD students from computer science and statistics. We have also revamped the undergraduate curriculum at Duke to incorporate elements of cloud computing.

What is the impact on physical resources that form infrastructure?

In conjunction with this project, we have obtained research grants from Amazon in the form of Amazon Web Services credits, which can be used to rent hardware resources as needed from Amazon EC2.

What is the impact on institutional resources that form infrastructure?

None.

What is the impact on information resources that form infrastructure?

None so far. However, we plan to develop a public repository of execution traces, historical market prices, as well as performance and price models that are derived from them. This information resource will be valuable to a cloud-based computing infrastructure.

What is the impact on technology transfer?

Botong Huang, the lead computer science PhD student on the project, is interning with the SystemML group at IBM Almaden Research Center in the summer of 2014. SystemML is an IBM project with similar aims as Cumulon. Through this connection, we hope that some of Cumulon's technology can be incorporated into commercial products.

What is the impact on society beyond science and technology?

None so far. We plan to pursue applications of Cumulon to problems in quantitative political science, with impact to the society beyond science and technology.

Changes/Problems

Changes in approach and reason for change

N/A

Actual or Anticipated problems or delays and actions or plans to resolve them

N/A

Changes that have a significant impact on expenditures

N/A

Significant changes in use or care of human subjects

N/A

Significant changes in use or care of vertebrate animals

N/A

Significant changes in use or care of biohazards

N/A