CloudVista: Interactive and Economical Visual Cluster Analysis for Big Data in the Cloud

Huiqi Xu, Zhen Li, Shumin Guo, Keke Chen
{xu.39, li.108, guo.18, keke.chen}@wright.edu
DIAC Lab, Kno.e.sis Center, Department of Computer Science and Engineering
Wright State University, Dayton, Ohio 45435, USA

ABSTRACT

Analysis of big data has become an important problem for many business and scientific applications, among which clustering and visualizing clusters in big data raise some unique challenges. This demonstration presents the CloudVista prototype system to address the problems with big data caused by using existing data reduction approaches. It promotes a whole-big-data visualization approach that preserves the details of clustering structure. The prototype system has several merits. (1) Its visualization model is naturally parallel, which guarantees the scalability. (2) The visual frame structure minimizes the data transferred between the cloud and the client. (3) The RandGen algorithm is used to achieve a good balance between interactivity and batch processing. (4) This approach is also designed to minimize the financial cost of interactive exploration in the cloud. The demonstration will highlight the problems with existing approaches and show the advantages of the CloudVista approach. The viewers will have the chance to play with the CloudVista prototype system and compare the visualization results generated with different approaches.

1. INTRODUCTION

With the development and deployment of ubiquitous information-sensing mobile devices, wireless sensor networks, RFID readers, simulation, and software logs, big data (e.g., terabytes to petabytes) have become normal in many business and scientific applications. Because data analysis is often iterative and exploratory, big data brings significant challenges. The current research on big data analysis has been focused on MapReduce-based processing and SQL-like interfaces such as Hive and Pig. Without intuitive understanding of the big data, it is often time consuming to form hypotheses based on interactive SQL-style queries. Previous studies have shown that visualization enables rapid perception of patterns, while interaction can quickly tune visualization to form and validate hypotheses. Thus, it is highly desired to have an interactive visual analysis tool to support existing big data analysis.

Data clustering has been widely used in the initial stages of data analysis, typically exploratory data analysis, to group data items and partition large data into smaller sections, so that the data is better understood and easier to manipulate with successive analytic operations. Traditionally, sampling and summarization [5] are the major methods used to reduce the size of large data so that existing clustering analysis algorithms can be applied. While data grows to big data, the amount of data that most existing clustering analysis and visualization algorithms can handle stays limited because of their algorithm complexity, resulting in severe loss of fidelity. In particular, when one needs to analyze rare events or small clusters in big data, data reduction methods are inappropriate to apply. Algorithms and systems working on the whole dataset or large-size samples are necessary for analyzing big data.

Interactively visualizing and analyzing data clusters would be a unique addition to the existing approaches for big data analysis. However, it also brings several unique challenges. (1) It requires visualization algorithms scalable to big data and large-scale parallel processing infrastructures. (2) As big datasets are often stored and processed remotely in the cloud, the latency caused by network and batch processing severely conflicts with the requirement of interactivity. (3) Exploring data in the cloud also needs to address the unique economics problem - how to minimize the financial cost while ensuring the quality of service not compromised.

We propose the CloudVista approach to address the above challenges. The multidimensional big data in the cloud can be reduced to "visual frames", the size of which is only subject to the resolution of visualization, independent of the original big data. In the demo system, we use the star-coordinate visualization model, e.g., the VISTA model [1] to generate visual frames, which can be implemented as MapReduce programs so that the system can scale up to big data. We address the problem of latency by promoting a hybrid batch-interaction method based on the RandGen algorithm. The RandGen algorithm uses a continuous random projection model that can produce a series of statistically meaningful visual frames without the user's intervention. We have proven that the RandGen algorithm can effectively preserve the clustering structure in a stochastic way [2]. The economics problem is addressed by an adaptive processing model and optimal resource provisioning strategies.

The rest of the proposal is organized as follows. Section 2 will present the CloudVista architecture and its components. Section 3 describes how the demonstration will be performed, including three parts: an introduction to the demonstration, the live system, and the comparison with existing data reduction approaches. Finally, we summarize our demonstration.

2. CLOUDVISTA ARCHITECTURE

The CloudVista approach has a three-tier architecture: the interactive visualization client, the application server, and the hadoop cluster (the latter two are possibly in the cloud), as shown in Figure

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Articles from this volume were invited to present their results at The xth International Conference on Very Large Data Bases.
Proceedings of the VLDB Endowment, Vol. X, No. Y
Copyright 20xy VLDB Endowment 2150-8097/11/XX... $ 10.00.

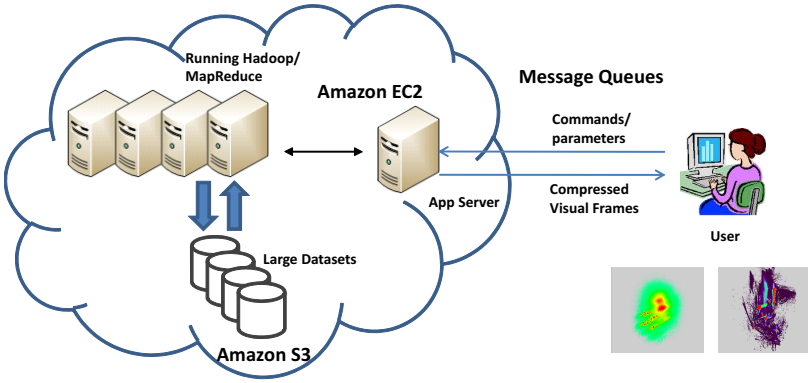

Figure 1: The CloudVista Demo Framework

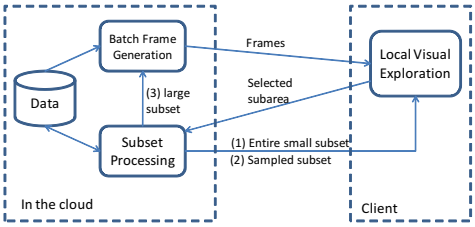

Figure 2: The exploration model.

1. It processes multidimensional numerical data for exploratory cluster analysis. The data and computing intensive tasks are finished in the hadoop cluster, which will generate the intermediate visual representations - the visual frames, or the user selected subsets. The application server manages the visual frames and subset information, issues cloud processing commands, gets the results from the cloud, and compresses data for transmission. The client renders the visual frames, takes care of user interaction, and, if the selected subsets are small, works on these small subsets directly with the local VISTA visualization system [1]. The communication between the client and the application server is via a message queue service. There are some key technical components in this architecture.

Visualization Model. We use the VISTA visualization model [1] in our prototype. It is derived from the star-coordinate model. Let $\mathbf{s}_i = (\cos(\theta_i), \sin(\theta_i)), i = 1, \ldots, k$ be unit vectors arranged in a "star shape" around the origin on the display, and a k-dimensional point $\mathbf{x} = (x_1, \ldots x_i, \ldots, x_k)$, where x_i is appropriately normalized (often using standardization or max-min normalization). $\alpha = (\alpha_1, \ldots, \alpha_k)$, $\alpha_i \in [-1, 1]$ are the dimensional weights and c is a positive scaling factor. The mapping is defined as

$$f(\mathbf{x}, \alpha, \{\mathbf{s}_i\}, c) = c \sum_{i=1}^{k} \alpha_i x_i \mathbf{s}_i. \quad (1)$$

This model is essentially a random projection model if we randomly select the parameters α_i. We have shown that close points will be surely mapped to close 2D points, while distant points might also be mapped to close 2D points, which creates visual cluster overlapping and needs the user to interactively explore and distinguish [2]. By tuning α values, the user can interactively find the possible overlapping. Compared to the popular principal component analysis (PCA) method, this model has unique advantages: it allows dimensional tuning to identify visual cluster overlapping, and its statistical properties in randomized projections also validate the RandGen algorithm [2]. Since this model is applied to individual records, it is naturally parallel and can be implemented with MapReduce [3].

Visual Frame. The VISTA model fits point-based visualization. When the number of points is too big, we need to consider density visualization instead. The key structure in CloudVista is the *visual frame* structure. With the visualization model, the density of mapped result is preserved - dense "'point clouds" are mapped to 2D dense point clouds, while the 2D point clouds might overlap each other. A visual frame is a two-dimensional density map, which encodes the 2D point clouds. Each cell is represented with (u_i, v_i, d_i), where (u_i, v_i) is the coordinate and d_i represents the density on the cell. Multiple visual frames with different α settings can be generated in parallel with a MapReduce program. The granularity of visual frame (i.e., the number of cells in the u and v directions) can be defined according to the resolution of the display. Since most of the cells will be empty in practice, this structure is space-efficient and independent of the size of the visualized dataset (about 150KB-300KB per frame). We use the widely adopted heatmap method to visualize the density information.

Batch Frame Generation Algorithm: RandGen. In local exploration of small datasets, the user can tune the α values to immediately observe the changing visualization to find the clue of possible visual overlapping. However, this is impractical for exploring big data in the cloud due to the latency of updating the visualization. We develop the RandGen algorithm to generate the continuously changing visualization in batches of visual frames. The basic idea is to allow the user to quickly capture the global patterns from randomly generated visual frames. Then, they can drill down to the subsets for detailed exploration, which can be done locally.

The algorithm is defined as follows. Starting from an initial set of α values. Let $\alpha_i^{(\phi)}$ represent the α values at the frame ϕ, the key problem is how to determine the $\alpha_i^{(\phi+1)}$ values for the frame $\phi+1$. RandGen applies small stochastic updates $\pm\delta$ to all dimensional weights α_i simultaneously, where δ is a user defined small value, and α_i are bounded to the range [-1,1]. The small stochastic updates create a randomized but smooth change of visualization. We have proved that with a sufficiently large number of continuously changing frames (e.g., 100 frames), the clustering structure can be statistically preserved and visually observed [2].

Economical Exploration. Cloud economics is one of the unique features in cloud computing. Minimizing the financial cost in computing is often an important goal of cloud-based applications. Different from the traditional computing paradigms, now the algorithm and system designers need to explicitly consider the financial factor for *each job*. We try to address this problem from three aspects.

(1) Optimizing the cost of generating the top-level visualization. The RandGen algorithm is implemented with MapReduce to generate the top-level visualizations on the whole big data. The optimization will involve a tradeoff between the financial cost and the response time. We have studied the resource optimization for running MapReduce programs in the cloud [4], which is applied to find the optimal resource provisioning strategy for RandGen, subject to the user's time and budget preferences.

(2) A hybrid exploration model. We provide drill-down operations for users to focus on interested subsets. The drill-down op-

erations will exponentially reduce the amount of focused data by levels. Once the amount of subset data is small enough for sampling, the sample set or the entire subset is sent back to the client for local exploration, which minimizes the needs of cloud operations. Figure 2 shows the hybrid processing model.

(3) Fully utilizing cloud resources. The cloud resources are often charged by time units (e.g., hours). If a task is done in a half of hour, the resources will still be charged by an hour. Thus, we should find ways to effectively utilize the charged but not used resources. For example, for the RandGen algorithm we can generate multiple batches with the provisioned resources, or try to predict the user's intention to pre-compute some frames. The studies on the concrete solutions are still ongoing.

3. DEMONSTRATION

This demonstration will show the novel features of CloudVista and also provide an intuitive way of understanding the limits of existing approaches in handling big data. Through the demonstration, users will observe the performance of RandGen algorithm to generate batches of visual frames and the utility of the batch frames. We implemented the CloudVista prototype with the VISTA visualization model [1], hadoop/MapReduce programs, and Amazon EC2/S3/SQS services. The original datasets are stored in S3. The hadoop cluster is provisioned dynamically in Amazon EC2 to generate visual frames or process subsets. Once a dataset is requested, it is loaded to the hadoop cluster from S3. The simple queue service (SQS) is used as the communication channel between the application server and the client software. The demonstration consists of three parts:

Introduction: The first part uses a poster and slides to underline the unique problems of interactive visual exploration of big data in the cloud, and introduces the viewer to the CloudVista approach.

Live System: A fully interactive demonstration of the entire CloudVista prototype will be presented. The user will be able to use the client-side system to load a dataset, run RandGen to generate batches of visual frames, interactively explore visual frames, and drill down the selected subsets. Various system statistics will be fed back from the cloud to the client to help understand the cloud side cost.

Comparison with Existing Approaches: The existing approaches depend on the data reduction approaches to obtain a manageable size of data, which will severely downgrade the visualization quality. We will compare the visualization results on the whole big dataset with those on sample sets or summary sets, to show the uniqueness of cluster exploration on big data.

3.1 Introduction to the Demonstration

A poster and a set of slides are used to highlight the inherent problems of visually exploring big data in the cloud and introduce the viewer to the CloudVista approach. First, we will show the unique problems with interactive visual cluster exploration for big data in the cloud, in particular, why the existing sampling and summarization methods cannot be used. Second, we will highlight the key features of the system, including an explanation of the visualization model, and illustrate how CloudVista handles the challenges on scalable processing, interactivity, and the economics of exploration. We will show how the three tiers collaborate to generate meaningful batches of frames in parallel and how user's interaction is fed back to the back-end system and executed. We also show how the subsets are generated and organized to efficiently support the drill-down operations. Finally, we will show some experimental results from the preliminary study [2].

3.2 Live System

The live system consists of a preliminary version of the client side user interface, the application server, and the MapReduce programs. The user will be able to use them to explore a sample dataset (e.g., the extended census dataset with 68 dimensions and 25 millions of records, and the extended KDD99Cup data with 40 dimensions and 40 millions of records [2]).

Client User Interface (UI). Figure 3 shows the client-side UI. The visual frame is visualized with the heatmap method. The UI includes frame browsing (the lower slider and control widgets), frame zooming (the left vertical slider), and view moving (the four-direction widget). The top right also shows the status and statistics for some operations.

Visual Frame Generation: In the beginning of exploring a dataset, the user needs to select the dataset in the hadoop and set the necessary parameters. Figure 4 shows the parameter configuration. The "resolution" parameter defines the number of cells in the u and v directions of the display area. The "number of frames" defines the number of frames that will be generated in each batch (e.g., 100). The "max sample size" is the maximum number of sample records that the original VISTA system can comfortably handle locally in the client side (e.g., 50,000). The "sample rate" is the user acceptable minimum sample rate that can preserve sufficient details of the clustering structure (e.g., 0.05). The top right corner of the main UI will show the progress and statistics of the RandGen algorithm, including the number seconds spent to generate the frames, the amount of data transferred to the client, the number of frames and the resolution. Once the user is not satisfied with the current batch of frames, she/he can generate another batch. The user can also observe the performance and visualization differences between two resolutions: 1000x1000 and 250x250.

Basic Interactions. After the frames are generated, transferred, and loaded into the client system. The user can "autoplay" the frames or select any frame to observe its clustering details. The zooming widget allows the user to zoom in/out of the view. The view moving widget allows the user to move the view to certain part of the frame. This is especially useful when the view is zoomed in. Note that zooming does not invoke cloud-side operations, which means the details are restricted by the resolution of current frame. For more details of a focused area, a drill-down operation is needed.

Drill-down. The drill-down operation supports the hybrid exploration model that can efficiently reduce the size of data to be processed. If the user is interested in exploring more details in some area of a visual frame, she/he can drag the mouse to select the interested rectangle area. The drill-down operation may cause three different subset operations at the backend, according to the size of the selected subset: (1)subset RandGen for subsets still too large, (2) sampling for medium subsets, if sampling with the acceptable rate will generate an acceptable sample size that the client can handle locally, and (3) directly returning the entire subset for small subsets. The subset processing strategy and statistics are shown on the top right part of the window, including the time to finish the operation and the amount of data transferred. The subset exploration is also encoded as "dataset.exp_id.layer$_1$_seq.layer$_2$_seq...", where layer$_i$ means the drill-down layer, and "layer$_i$_seq" represents the sequential number of drill-down operations at the layer i. With this organization it is easy to roll back from the exploration of low-level subsets. All the intermediate results are buffered on the client side. Once a roll-back operation happens, the system can retrieve the related dataset based on the naming convention.

3.3 Visualization of Reduced Data

In this part of demonstration, we allow the viewer to compare the

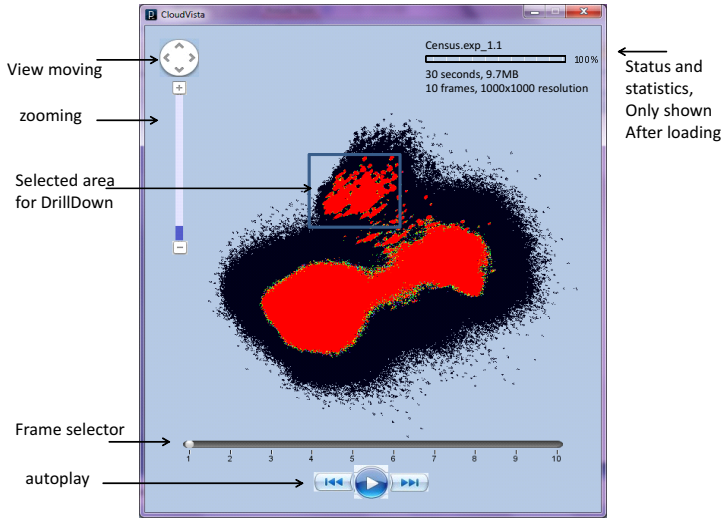

Figure 3: The client-side user interface

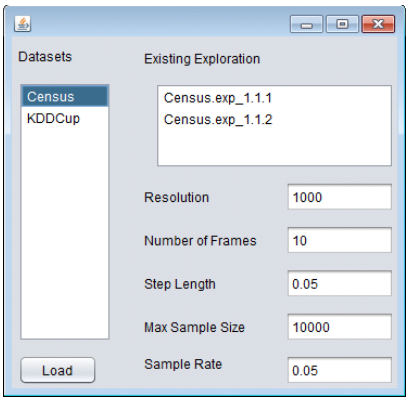

Figure 4: Parameter setting and frame loading.

visualization results of the CloudVista entire-data approach with those of reduced data. Two common methods, sampling and summarization, will be used to reduce the big data.

Sampling. The uniform sampling algorithm is applied to get acceptable sample sizes (e.g., 50,000) from the big data in the cloud. Then, we will use the existing VISTA system [1] to visualize the sample datasets.

Summarization. We implement the BIRCH approach [5] for summarization. BIRCH uses a clustering-feature (CF) tree to process the data records. Each data record is absorbed into one of the leaves of the CF tree. At last, each entry in the leaf node represents a cluster in the dataset. Depending on the height and the fanout of the tree, we can derive a number of small clusters in the end. These clusters are described by the number of points, the centroid (mean), and the variance. Each cluster can be approximated by a multidimensional normal distribution with the same mean and variance. We develop a MapReduce program proportionally draw samples from these multidimensional distributions to generate the density map, which are then visualized with the CloudVista frame viewer.

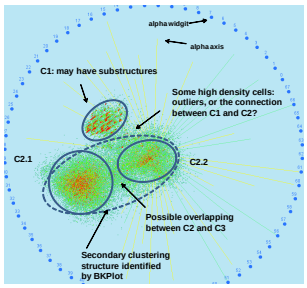

Figure 5: Visualization and analysis of 10000 samples of Census data with the VISTA system.

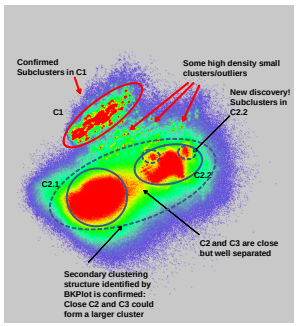

Figure 6: Visualization and analysis of 25 Million Census records with CloudVista (1000x1000 resolution).

Figure 5 and 6 show a comparison between the visualization of a sample dataset and that of the entire dataset. The viewers will find it is difficult to explore small clusters with reduced data. In addition, since visual details are not well preserved inside the clusters, it is difficult to identify interesting areas to drill down. We leave other details for the viewers to discover in the demonstration.

4. SUMMARY

This demonstration highlights the inherent research problems with visual clustering analysis of big data in the cloud. It presents a working prototype system CloudVista to address these problems. CloudVista works on the entire big dataset with the support of parallel processing framework such as hadoop/MapReduce. It uses the RandGen algorithm to generate batches of statistically meaningful visual frames for users to explore, which addresses the conflict between latency and interactivity. Cloud economics is addressed by the hybrid exploration model, optimal resource provisioning, and the price-model aware processing. In the demonstration, the viewer will be able to play with the prototype system, understand these unique problems, and observe the limitation of the data reduction methods to big data.

5. REFERENCES

[1] CHEN, K., AND LIU, L. VISTA: Validating and refining clusters via visualization. *Information Visualization 3*, 4 (2004), 257–270.

[2] CHEN, K., XU, H., TIAN, F., AND GUO, S. Cloudvista: Visual cluster exploration for extreme scale data in the cloud. In *Proceedings of International Conference on Scientific and Statistical Database Management (SSDBM)* (2011).

[3] DEAN, J., AND GHEMAWAT, S. MapReduce: Simplified data processing on large clusters. In *USENIX Symposium on Operating Systems Design and Implementation* (2004).

[4] TIAN, F., AND CHEN, K. Towards optimal resource provisioning for running mapreduce programs in public clouds. In *IEEE Conference on Cloud Computing* (2011).

[5] ZHANG, T., RAMAKRISHNAN, R., AND LIVNY., M. BIRCH: An efficient data clustering method for very large databases. In *Proceedings of ACM SIGMOD Conference* (1996), pp. 103–114.