

Data Visualization in the Cloud — Software as a Service Application for D3.js

0456024 Pei-Shan Tsai
National Chiao Tung University, pstsai@nclab.tw

Abstract - Data visualization, as data-intensive and computing-intensive as other tasks of big data, is a perfect cloud computing application. This project presents a software as a service for a web-based interactive data visualization JavaScript library D3.js, and we make efforts to two aspects: client-side offload and easy-to-use interface.

Index Terms - Big data, Cloud computing, Data visualization

INTRODUCTION

Big data is one of the most important cloud service since it requires clusters of servers to efficiently process the large volumes, high velocity, and varied formats of data. The term often refers simply to data analysis, while other challenges include capture, curation, search, sharing, storage, transfer, visualization, interaction, and information privacy [1].

Data visualization is the techniques used to is to communicate information clearly and efficiently to users by encoding it as visual objects contained in graphics. Effective visualization can help users in analyzing and reasoning about data and evidence. It makes complex data more accessible, understandable and usable [2].

D3.js, referred to as Data-Driven Documents, is an open source JavaScript library for web-based interactive data visualization applications. It helps users bring data to life by using HTML, SVG, and CSS [3].

Although its emphasis on web standards provides full capabilities of browsers without tying users to a proprietary framework, it is almost unusable for a nonprogrammer. For programmers, development becomes painful when the size of dataset is larger than browsers can handle over. In this case, there are two solutions: (1) reducing dataset, such as deleting unnecessary data, precalculating interested statistics as new input. (2) embedded in server-side scripting. The first solution may make analysis procedures inflexible and they all increase users' workload.

Hence, we want to build a software as a service application for D3.js which has higher scalability by offloading browsers' tasks to server and provides easy-to-use interface for both programmers and nonprogrammers by specifying D3.js visualization modules.

RELATED WORK

Table. 1 provides a side-by-side comparison of all major on-premise and cloud big data vendors [4]. Compared to these platform, cloud D3.js is an open source for data

visualization. It is more lightweight and more portable since the only required installation is a web browser.

PROPOSED APPLICATION

The application is built on Google App Engine and it runs on Google's powerful infrastructure. On client-side, the main task of a browser is drawing a representation. On server-side, Google App Engine takes over the computing. Here we use PHP and Google Cloud SQL, a relational database for data storage and manipulation. Fig. 1 depicts the process of cloud D3.js data visualization and Fig. 2 shows a series of screenshots of one use case.

- (1) User selects a D3.js data visualization method.
- (2) User uploads dataset.
- (3) User selects a set of interested data and customizes setting for the specific D3.js data visualization method.
- (4) Server computes statistics according to user's requests at step 1, 2 and 3.
- (5) Server returns the results of step 4.
- (6) Client renders the representation.

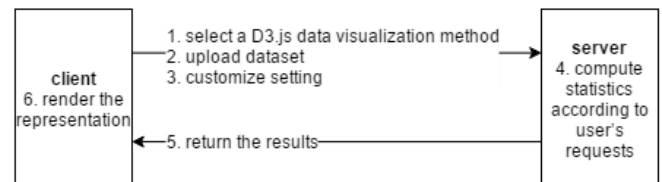


Fig. 1. The process of cloud D3.js data visualization.

EVALUATION AND DISCUSSION

A key element of cloud service is customization and the creation of a user experience. The following are some evaluation criteria for such data visualization applications. The topmost is more data-visualization-related and the bottommost is more cloud-computing-related.

















(1) *Accuracy of results*. The basic requirement of data analysis, are there any errors during computation and whether the final representation is exact the one we desire or not.

(2) *A balance between generality and usefulness*. Since data visualization is a highly customized techniques, how to module D3.js visualization methods becomes an important issue. Generality means if a module is suitable for most users' cases, and usefulness means how well can a module fit the specific case.

(3) *Integrity of data*. Dataset is uploaded to the cloud, can Google Cloud SQL assure the accuracy and the consistency

December 29, 2015

Table 1. Big Data Vendors/Platform Comparison.

	Hortonworks	Cloudera	HDInsight	Altscale	TreasureData	Databricks	AmazonEMR	Qubole	Cloud D3.js
Product Summary	100% open Source Hadoop	OpenSource Hadoop with proprietary management	Big Data Infrastructure as a Service in the Azure Cloud	Big Data in Dedicated Cloud Service	Cloud-based data warehousing	Standalone Spark Service	Big Data Infrastructure as a Service in the AWS Cloud	Cross-platform Big Data Service with Unified Metadata	Open Source for Data Visualization
Must Migrate Data To Platform	YES	YES	NO*	YES	YES	NO**	NO**	NO***	YES
Out-of-the-box Data Processing Engines	Installation required	Installation required	MapReduce, Hive, Pig, Spark, HBase, Storm	MapReduce, Hive, Pig, Spark	Hive, Presto	Spark	MapReduce, Hive, Pig, HBase, Cascading, Impala, Spark, Presto	MapReduce, Hive, Pig, Cascading, Spark, Presto	Web browser required
Deployment Model	 On-Premises/Hosted	 On-Premises/Hosted	 Cloud	 Cloud	 Cloud	 Cloud	 Cloud	 Cloud	Cloud
Data Store	On-Premises	On-Premises	Azure	Altscale Data Cloud	TreasureData Cloud	AWS	AWS	AWS, GCP, Azure	Google Cloud Platform
Setup	 Manual	 Manual	 Automatic	 Automatic	 Automatic	 Automatic	 Automatic	 Automatic	Automatic
Management	Support and 3rd Party Consulting	Support and 3rd Party Consulting	No Big Data-specific Support	Full Management and Support	Full Management and Support	Full Management and Support	No Big Data-specific Support	Full Management and Support	Full Management and Support
Economic Structure	Software License and Support, Infrastructure Purchase and Personnel	Software License and Support, Infrastructure Purchase and Personnel	Elastic compute pricing	Fixed Rate	Pay-per-use	Elastic compute pricing	Elastic compute pricing	Elastic compute pricing	Elastic computing pricing
Scalability	Fixed Cluster	Fixed Cluster	Manual scaling, elastic, on-demand, no graceful downscaling	Manual scaling, elastic, on-demand	Manual scaling, elastic, on-demand	Manual scaling, elastic, on-demand, no graceful downscaling	Manual scaling, elastic, on-demand	Automatic, elastic, on-demand	Automatic, elastic, on-demand

of data over its entire life cycle, and prevent unintentional change.

(4) *Real-time response*. When we use ordinary client-side D3.js, we usually use precalculated interested statistics rather than whole dataset as input, but it makes analysis procedures inflexible because we have to preprocess data for each visualization scene. Now we put the computing task on server-side, is the response time of rendering and interaction acceptable.

(5) *Scalability*. One of the most significant feature of cloud service, in contrast to ordinary client-side D3.js, how much more data can we deal with in the cloud. Can system provide enough storage to store data and computing power to handle requests on demand.

(6) *Cost-efficiency*. One of the most significant feature of cloud service, how effectively clusters of servers can be used to compute statistics.

This application takes advantage of Google App Engine, that most of the above aspects are took good care of by platform itself and we can focus on modulating D3.js.

CONCLUSIONS AND FUTURE WORK

Big data and cloud computing are two popular topics of information technology in both academia and industry across the globe. How they work together to offer a better model has become critical issue and new terminologies such as storage as a service, data as a service, analysis as a service come out. Such a plenty of big data cloud services provides more flexible and more cost-efficient solutions than traditional relational database management systems for both personal and enterprise use with data-on-demand model.

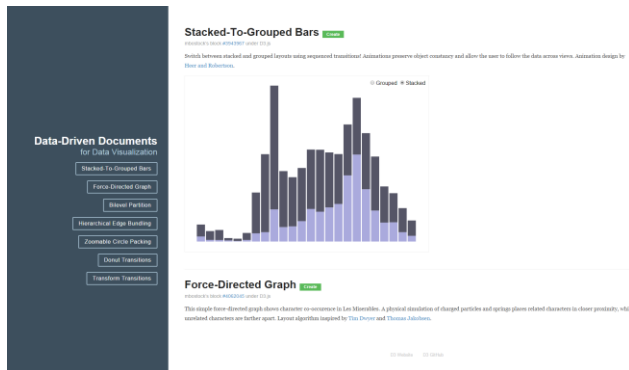
Beside mainstreams of storage and analysis, more and more attention is paid to other subfields of big data such as visualization and interaction. Services of different aspects can be combined and provides a total solution [5].

REFERENCES

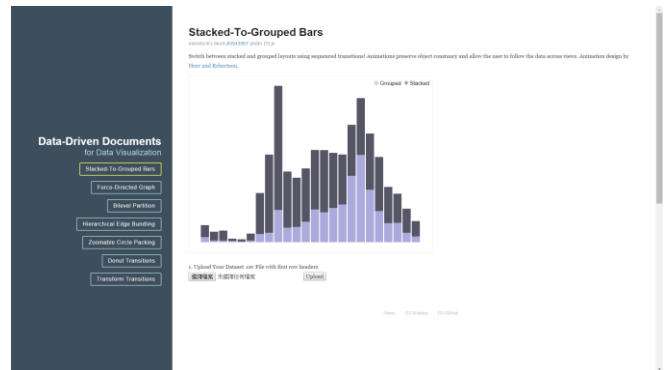
- [1] Big data. (n.d.). In *Wikipedia*. Retrieved November 3, 2015, from https://en.wikipedia.org/wiki/Big_data
- [2] Data visualization. (n.d.). In *Wikipedia*. Retrieved November 3, 2015, from https://en.wikipedia.org/wiki/Data_visualization
- [3] Mike Bostock. (n.d.). *Data-Driven Documents*. Retrieved from <http://d3js.org>

- [4] Qubole, Inc. (n.d.). *Big Data Vendors/Platform Comparison*. Retrieved from <http://www.qubole.com/resources/solution/big-data-vendors-comparison>

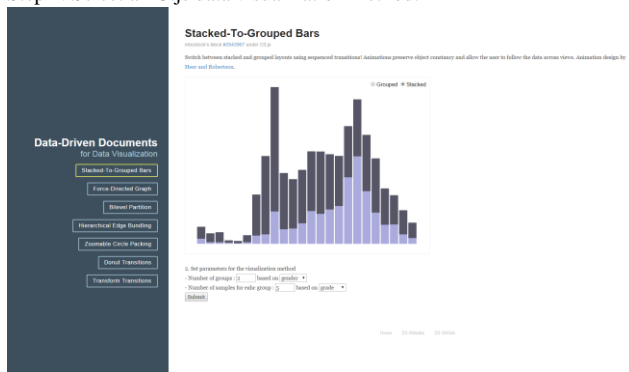
Marcos D. Assunção, Rodrigo N. Calheiros, Silvia Bianchi, Marco A.S. Netto, Rajkumar Buyya. *Big Data computing and clouds: Trends and future directions*. Journal of Parallel and Distributed Computing, volumes 79–80, pages 3-15, May 2015, ISSN 0743-7315 (KEY REFERENCE)



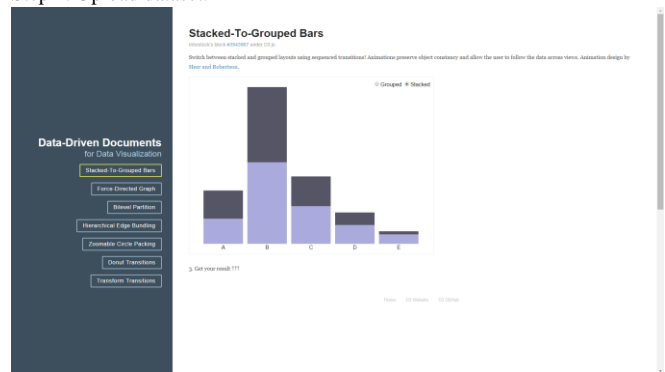
Step 1. Select a D3.js data visualization method.



Step 2. Upload dataset.



Step 3. Customize setting.



Step 4. Get a result.

Fig. 2. A series of screenshots of one use case.