Google BigQuery for Education: Framework for Parsing and Analyzing edX MOOC Data

Conference Paper · April 2017

DOI: 10.1145/3051457.3053980

5 authors, including:

Daniel T. Seaton
Harvard University
58 PUBLICATIONS **1,855** CITATIONS
SEE PROFILE

Isaak Chuang
Massachusetts Institute of Technology
241 PUBLICATIONS **22,203** CITATIONS
SEE PROFILE

Google BigQuery for Education: Framework for Parsing and Analyzing edX MOOC Data

Glenn Lopez
Harvard
Cambridge, MA
glenn_lopez@harvard.edu

Daniel T. Seaton
Harvard
Cambridge, MA
daniel_seaton@harvard.edu

Andrew Ang
Harvard
Cambridge, MA
andrew_ang@harvard.edu

Dustin Tingley
Harvard
Cambridge, MA
dtingley@gov.harvard.edu

Isaac Chuang
MIT
Cambridge, MA
ichuang@MIT.EDU

ABSTRACT

The size and complexity of MOOC data present overwhelming challenges to many institutions. This paper details the functionality of edx2bigquery – an open source Python package developed by Harvard and MIT to ingest and report on hundreds of MITx and HarvardX course datasets from edX, making use of Google BigQuery to handle multiple terabytes of learner data. For this application, we find that Google BigQuery provides ease of use in loading the multi-faceted MOOC datasets and near real-time interactive querying of data, including large clickstream datasets; moreover, we are able to provide flexible research and reporting dashboards, visualizing and aggregating data, by interfacing services associated with BigQuery. This framework makes it feasible for edx2bigquery to be open source, following standards which emphasize the importance of data products that transcend a particular data science platform and allow teams with diverse backgrounds to interact with data. edx2bigquery is being adopted by other institutions with an aim toward future collaboration.

ACM Classification Keywords

K.3.1. Computers and Education: Computer Use in Education.

Author Keywords

MOOC; learning analytics; big data; BigQuery; educational data mining.

INTRODUCTION

Applications found in the Learning at Scale community require efficient and pliable data infrastructure capable of processing millions of user interactions in order to study the learning process. Massive Open Online Courses (MOOCs) are one such example where data sizes and the complexity of learning environments present challenges to institutional analyses. To date, HarvardX and MIT have well over 300 total courses launched on edX.org, and the overall number of courses and learners, and dataset sizes, continue to grow.

The data generated from MOOCs are easily some of the largest and most complex educational course data institutions have encountered. The majority of course user interactions have a unique record indicating when and with which resource an interaction occurred. Each record takes up storage space, often leading to data sizes of tens of gigabytes per course. In addition to individual interaction data, there are often contextual data that must be merged with interaction data in order to interpret an action. Processing these data has been the responsibility of diverse teams, often extending to some combination of institutional analysts and academic researchers.

Harvard and MIT recently turned to the cloud – Google's BigQuery – for processing MOOC data. Google BigQuery combined with custom Python libraries offers an efficient and pliable analytics pipeline whose features satisfy stakeholders hoping to glean insights from MOOC data. And although HarvardX and MITx data are continuously growing in size and complexity, the data pipeline used to parse and aggregate MOOC data is stable to share openly with other institutions.

In this paper, we describe the primary open source package responsible for the Harvard and MIT MOOC data pipeline: edx2bigquery [7]. The edx2bigquery framework has facilitated the majority of external reporting on HarvardX and MITx MOOCs. Institutions working with edX data are encouraged to clone edx2bigquery for standing up their own data pipeline.

EDX DATA

edX is a provider of Massive Open Online Courses (MOOCs) and the unifying body for roughly 90 partner organizations from higher education, government, and industrial settings (e.g., HarvardX and MITx are both edX consortium members). edX partners create MOOC content while edX develops the open-source platform and handles distribution of courses. edX as an organization provides regular data exports to partners

L@S 2017, April 20-21, 2017, Cambridge, MA, USA

 ISBN 978-1-4503-4450-0/17/04…$15.00

DOI: http://dx.doi.org/10.1145/3051457.3053980

whose courses are running on edX.org, allowing partner organizations to prioritize their own research interests around MOOC learners and their behavior.

edX provides partners with regular research data exports [3] from courses run on edX.org and the piloting site edge.edx.org. These exports include the majority of user-interactions and arrive in two formats: 1) *Database Data* - snapshots of the platform's backend tables delivered weekly and 2) *Event Data* - all user click interactions across a partner's courses delivered nightly. Database data (also called SQL) contain all user information required to maintain a current state of a user's learning experience in a given course, such as, self-reported demographics, course enrollments, problem submissions for grading, etc. Event data (also called clickstream) contains a complete history of user interactions in the courseware. Clickstream data are generally quite complex containing a semi-structured format to allow data to evolve over time as new tools and features are developed. Events are generally the largest source of data recorded by the edX platform – often gigabytes per course – and by far the most challenging to work with pragmatically.

External data sources may also be required to better understand edX data. edx2bigquery has integrated data from the YouTube API, and geolocation via external libraries like Maxmind GeoIP2 City Database [11]. Each of these external sources has played a role in reporting from Harvard and MIT.

Finally, there are a number of stakeholder groups that have specific needs with regard to extracting insights from MOOC data. Although not discussed in detail here, data requests from Harvard and MIT come from one of four stakeholders: faculty, course teams, researchers, and administrators.

GOOGLE BIGQUERY: A POSITIVE HARVARDX AND MITX EXPERIENCE

Prior to the development of edx2bigquery, teams at Harvard and MIT developed a processing pipeline using a MongoDB database, hosted on a small computing cluster at the Office of Digital Learning at MIT. After roughly 2 years of MOOC data, processing times soon exceeded the limits of practical wait time for aggregating data, with some aggregate data taking longer than 48 hours to generate. Harvard / MIT decided to try Google BigQuery - a fully managed cloud service that enables storage and fast querying of large and multi-faceted datasets.

BigQuery is fast, returning results in seconds even when running queries on terabyte scale data; behind the scenes, the vital key making this speed uniform and easily obtained is that BigQuery does not require creation (or even specification) of indexes – any field can be queried rapidly. This is a huge advantage, compared with MongoDB, and traditional SQL systems like MySQL and Postgres, all for which queries are quick only on fields with existing indexes.

At the time of this writing, BigQuery's storage charge is $0.01-0.02 / GB stored per month, and query charges are $5 per TB processed per month. For MOOC data analysis, this per-query pricing model turns out to be remarkably affordable. As an example, Harvard spends in the range of $20-$50 / month to process 130+ courses, containing over 700 million click events generated by almost 3 million unique learners. These costs also include custom queries from an active research team. Davidson College – which processes data from a small handful of MOOCs and has fewer researchers interacting with BigQuery – spends typically under $1 / month.

Summarizing the BigQuery Choice

Summarizing the advantages provided by Google BigQuery over our previous on-premise MongoDB solution, we find:

1. No technical overhead costs for maintaining infrastructure;
2. Scalability of processing research data products across a growing number of courses and users;
3. Quick interactive data analysis, no database indexes needed;
4. Perform ad-hoc query across any and all courses, allowing joins across thousands of tables;
5. Affordable relative to many on-premise solutions.

EDX2BIGQUERY FRAMEWORK

The edx2bigquery framework provides an efficient, optimized solution for processing massive amounts of learner event data (billions of rows) per course efficiently using Google BigQuery while minimizing querying costs given a daily update frequency of event log data from edX. The framework can be used generically to process all events for a single course, or group(s) of courses, and a time interval can be specified with a custom start and end time (for the purposes of generating landmark annual reports). Figure 1 provides an illustration of how the edx2bigquery framework handles processing of edX database data and daily event data.

The dataflow for canonical daily datasets (green), begins with the stacked tables in the top left corner, representing multiple event tables in BigQuery, one table for each day. The table in the top right corner represents a generic Canonical Daily table (e.g., Person-Course-Day) produced in BigQuery as the result of the aggregate SQL query passed. Processing time in Google BigQuery takes the longest the first time the query is executed since millions of rows of event data are queried from the beginning of the course. Daily updates thereafter require performing a query only on new daily events data for each course and consequently provide quicker incremental updates by appending data to the end of the BigQuery table. As a real world example, a query to generate the Person Course Day for our largest and most popular course CS50X over a period of 73 days containing 13.4 million events took BigQuery 21.17 seconds.The results of that query created a Person Course Day table containing 1.15 million rows. New events for the next day took 7.57 seconds to process 165K events.

The dataflow for canonical datasets are highlighted in blue in Fig. 1. New database tables are delivered to each institution once a week representing a snapshot in time from edX's database. In the bottom right of the figure, canonical dataset (e.g., Person Course) in BigQuery is replaced when this occurs instead of updating and modifying each row. For efficiency, the SQL that generates a new canonical dataset (e.g., Person Course table) should simply aggregate the results of the latest

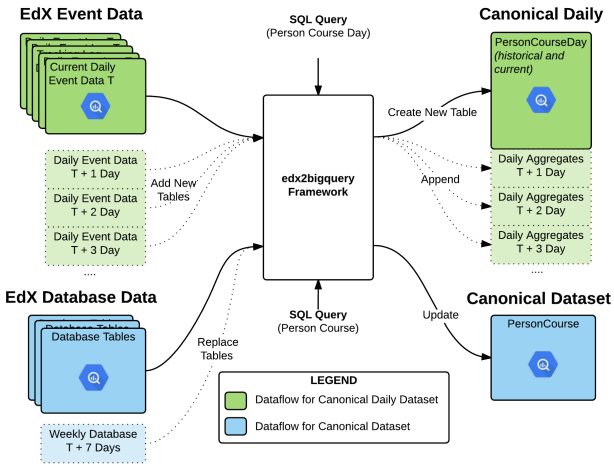

Figure 1. The edx2bigquery framework for processing event and database data to generate Person-Course-Day and Person-Course datasets. Solid lines represent up-to-date BigQuery tables. Dotted lines represent future edX data refreshed daily/weekly. Events are added as new BigQuery tables, while database tables are replaced weekly.

canonical daily dataset(s) (e.g., Person Course Day). As an example, loading for one our biggest courses took a total of 13.73 minutes to update 312K rows of data, 60+ columns wide.

Person-Course Dataset

The Person-Course dataset summarizes learning activity of a particular learner enrolled in an individual course. The key identifiers are a unique user ID and course ID. The dataset comprises enrollment time and mode, demographics, resources accessed, progression through a course, and time spent in the course. Currently, the Person-Course dataset contains 60+ variables centered on learner backgrounds and their behavior in a specific course. The Person-Course dataset is created for each MOOC run on edX and the dataset can be appended for multiple courses without duplicating data. A machine-readable Person-Course Schema in JSON format is defined, along with a Harvard VPAL Private Datasets Documentation record, maintained on Dataverse with version control by the Harvard Vice Provost for Advances in Learning (VPAL) Research Team [4]. We note that the majority of visualizations and analyses presented in the yearly reporting from Harvard and MIT are generated using the Person-Course dataset.

XANALYTICS

We have taken advantage of how Google BigQuery makes it straightforward to create dashboards. Reporting is largely handled by XAnalytics [8] – an open-source dashboard application hosted on Google's App-Engine [6] – which provides stakeholders with interactive graphics and tables based on aggregate datasets created using edx2bigquery. Dashboards can be created for every course at Harvard and MIT, as well as, for groups of many courses. XAnalytics was built around visualizing the various data products generated by edx2bigquery, hence, there are numerous interactive visualizations available. Figure 2 also contains visualizations directly available in XAnalytics (A1, B1, B2). The HarvardX and MITx administrative offices have also made significant use of data connectors offered by BigQuery, to flexibly extract, analyze, and visualize HarvardX and MITx data using external applications, in particular with Tableau, Python Jupyter Notebooks, and R Shiny.

DISCUSSION AND CONCLUSION

In summer 2016, Harvard and MIT hosted an edX data workshop that brought a number of universities together to explore using edx2bigquery [5]. Colgate University, Davidson College, Hamilton College, University of British Columbia, and Wellesley College now all have a working instance of edx2bigquery and others are working to deploy the framework. Cross-institutional use opens numerous opportunities to explore MOOC metrics across institution. Our hope is that collaborating with a diverse group of organizations will lead to richer insights into the MOOC movement.

Although we greatly value the edx2bigquery framework, we are also aware that data standards and canonical data products are the most important aspect of the Harvard and MIT data pipeline. Shared standards for data aggregation and reporting mean that institutions can readily compare metrics in a common format. Such sharing takes place in the realm of institutional research through organizations like the Association of American Universities Data Exchange (http://aaude.org/).

Cross-institutional sharing also brings up serious legal concerns related to IRB and regulations such as FERPA [13]. Open communication with policy makers is essential before committing to research and data sharing of any kind. In the very early days of the MOOC movement, Harvard and MIT setup a data use agreement to allow cross-institutional analysis of MOOC data. The Liberal Arts Collaborative – Colgate University, Davidson College, Hamilton College, and Wellesley University – modeled a similar agreement and are in the process of comparing their respective MOOC data [2].

Lastly, we note the existence of other MOOC data frameworks, e.g. MOOCdb and MOOC Czar. MOOC DB was an early framework whose focus was similar to our own in creating standards that can be shared between institutions [12]. MOOC Czar focused more on getting from raw data to dashboards with interactive visualizations [14]. We emphasize that data products (e.g., Person-Course) are the most meaningful basis for comparison. Institutions could remain platform agnostic and still share insights given common canonical datasets.

Institutionally, edx2bigquery, and its set of canonical datasets, continue to be immensely valuable for the multi-level reporting framework provided, embodying the research methodologies developed by Harvard and MIT [10] [9] [1].

ACKNOWLEDGMENTS

We are grateful for the support from Harvard University and MIT, specifically, the Office of the Vice Provost for Advances in Learning at Harvard, the Office of Digital Learning at MIT, and the VPAL-Research group. Special thanks to Professor Jim Waldo for early contributions and the many others contributing to the open source code base. Above all, we are thankful for the tireless efforts of HarvardX and MITx course designers whose MOOCs make this project possible.

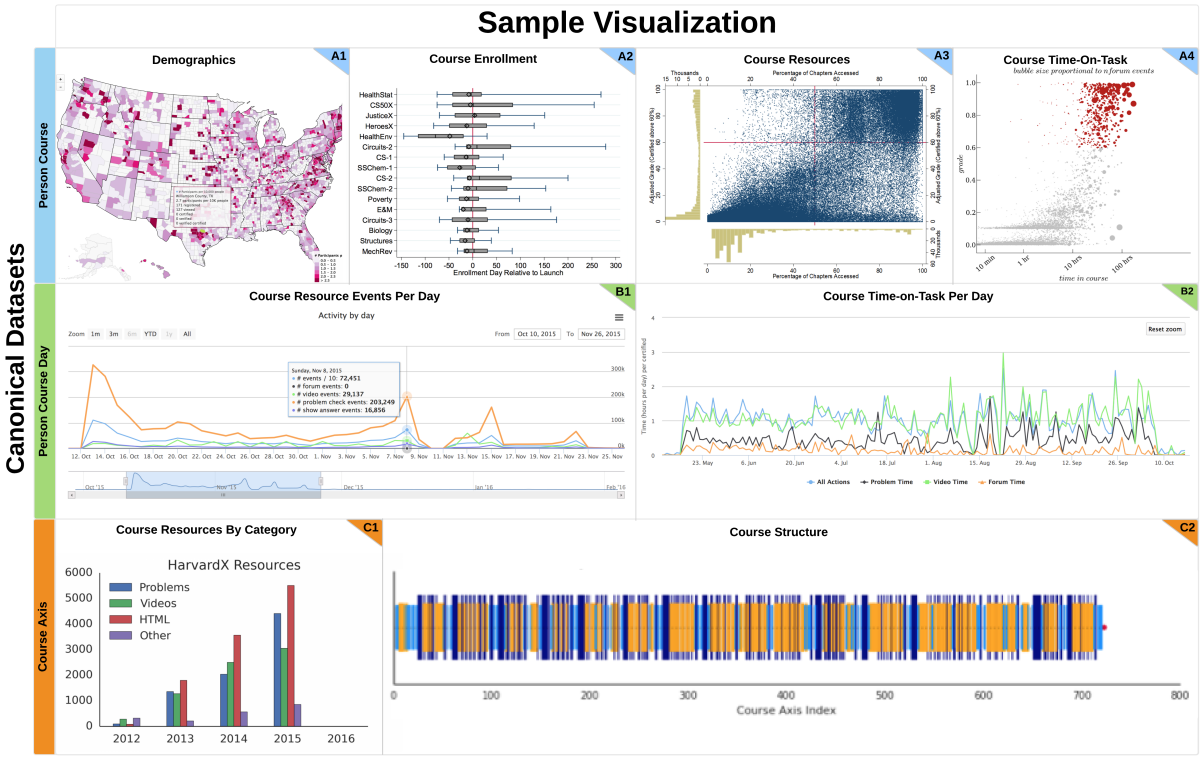

Figure 2. Visualization mosaic produced from canonical datasets generated by edx2bigquery; many are directly available with the dashboard application XAnalytics. Top row provides visualizations based on the Person-Course data product: left to right, geolocation as an example of demographic data, course enrollments across multiple HarvardX courses, performance versus amount of course accessed, and final grade versus time-on-task. Middle row provides visualizations for Person-Course-Day: left to right, number of daily events by event type and daily time on task by behavior (problem time, video time, forum time). Bottom row visualizes the course-axis data product: left to right, counts of total amount of HarvardX content created by year and a course structure visualization where bars and colors indicate course ordering for videos (orange), problems (navy), and text pages (light blue).

REFERENCES

1. Isaac Chuang and Andrew Dean Ho. 2016. Harvardx and MITx: Four Years of Open Online Courses – Fall 2012-Summer 2016. *Available at SSRN 2889436* (2016).
2. Davidson College. 2015. A Liberal Arts Take on Tech. `https://www.davidson.edu/news/news-stories/150512-davidson-co-founds-online-learning-consortium-`. (12 May 2015).
3. edX Documentation. 2016. Research Guide. (2016). `http://edx.readthedocs.io/projects/devdata/en/latest/`.
4. Harvard Vice Provost for Advances in Learning Research. 2016. Harvard VPAL Private Datasets Documentation. (2016). `http://dx.doi.org/10.7910/DVN/RTVIEM`.
5. Harvard Gazette. 2016. MOOCS Ahead. (2016). `http://news.harvard.edu/gazette/story/2016/07/moocs-ahead/`.
6. Google. 2016. Google App Engine. (25 October 2016). `https://cloud.google.com/appengine/docs`.
7. Harvard and MIT. 2016a. edx2bigquery. (25 October 2016). `https://github.com/mitodl/edx2bigquery`.
8. Harvard and MIT. 2016b. xanalytics. (25 October 2016). `https://github.com/mitodl/xanalytics`.
9. Andrew Dean Ho, Isaac Chuang, Justin Reich, Cody Austun Coleman, Jacob Whitehill, Curtis G Northcutt, Joseph Jay Williams, John D Hansen, Glenn Lopez, and Rebecca Petersen. 2015. Harvardx and MITx: Two years of Open Online Courses, Fall 2012 - Summer 2014. *Available at SSRN 2586847* (2015).
10. Andrew Dean Ho, Justin Reich, Sergiy O Nesterko, Daniel Thomas Seaton, Tommy Mullaney, Jim Waldo, and Isaac Chuang. 2014. HarvardX and MITx: The First Year of Open Online Courses, Fall 2012 - Summer 2013. *Available at SSRN 2381263* (2014).
11. Maxmind. 2016. GeoIP2 City Database. (2016). `https://www.maxmind.com/en/geoip2-city`.
12. Kalyan Veeramachaneni, Franck Dernoncourt, Colin Taylor, Zachary Pardos, and Una-May O'Reilly. 2013. Moocdb: Developing data standards for mooc data science. In *AIED 2013 Workshops Proceedings Volume*. Citeseer, 17.
13. Elise Young. 2015. Educational privacy in the online classroom: FERPA, MOOCs, and the big data conundrum. *Harv. J. Law & Tec* 28 (2015), 549–593.
14. John Zornig. 2016. MOOCczar. (25 October 2016). `https://github.com/UQ-UQx/MOOCczar`.