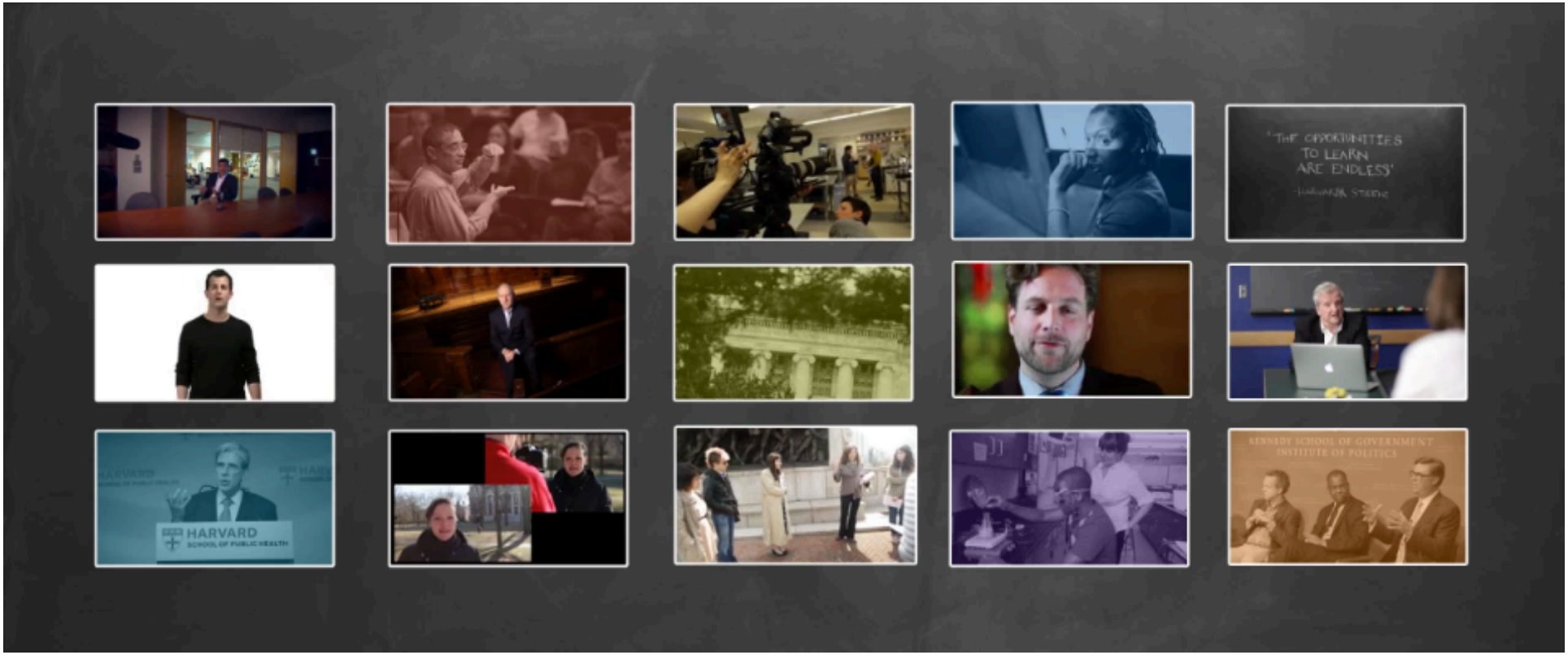


A Common Research Data Infrastructure Between HarvardX and MITx Using Google BigQuery

HarvardX

Glenn Lopez

2015-04-02

HarvardX Goals

HarvardX is a University-wide strategic initiative, overseen by the Office of the Vice Provost for Advances in Learning, to enable faculty to build and create open online learning experiences for residential and online use, and to enable groundbreaking research in online pedagogies.

- Expand access to education worldwide
- Improve teaching and learning on campus
- Advance our understanding of teaching and learning through research

http://harvardx.harvard.edu/

HarvardX

Research Data

How “Big Data” is our Data?

- **Volume**
- **Variety**
- **Velocity**

HarvardX

Research Data

How “Big Data” is our Data?

- **Volume**
  - 60+ Courses w/ over 1.7 Million participants
  - ~1 Billion+ activity log events total, ~1 million activity log events / day
  - ~30 GB of course enrollment data exported / week

HarvardX

Research Data

How “Big Data” is our Data?

- **Variety**
  - 3 major types of data are managed including activity logs, course enrollment data & survey data
  - Course enrollment data format changes slowly (once every ~4-6 months)
  - Activity logs format change frequently (once every 1-2 months)
  - Survey Data format varies the most and changes most frequently, with different questions asked for different courses (although some are standardized)

HarvardX

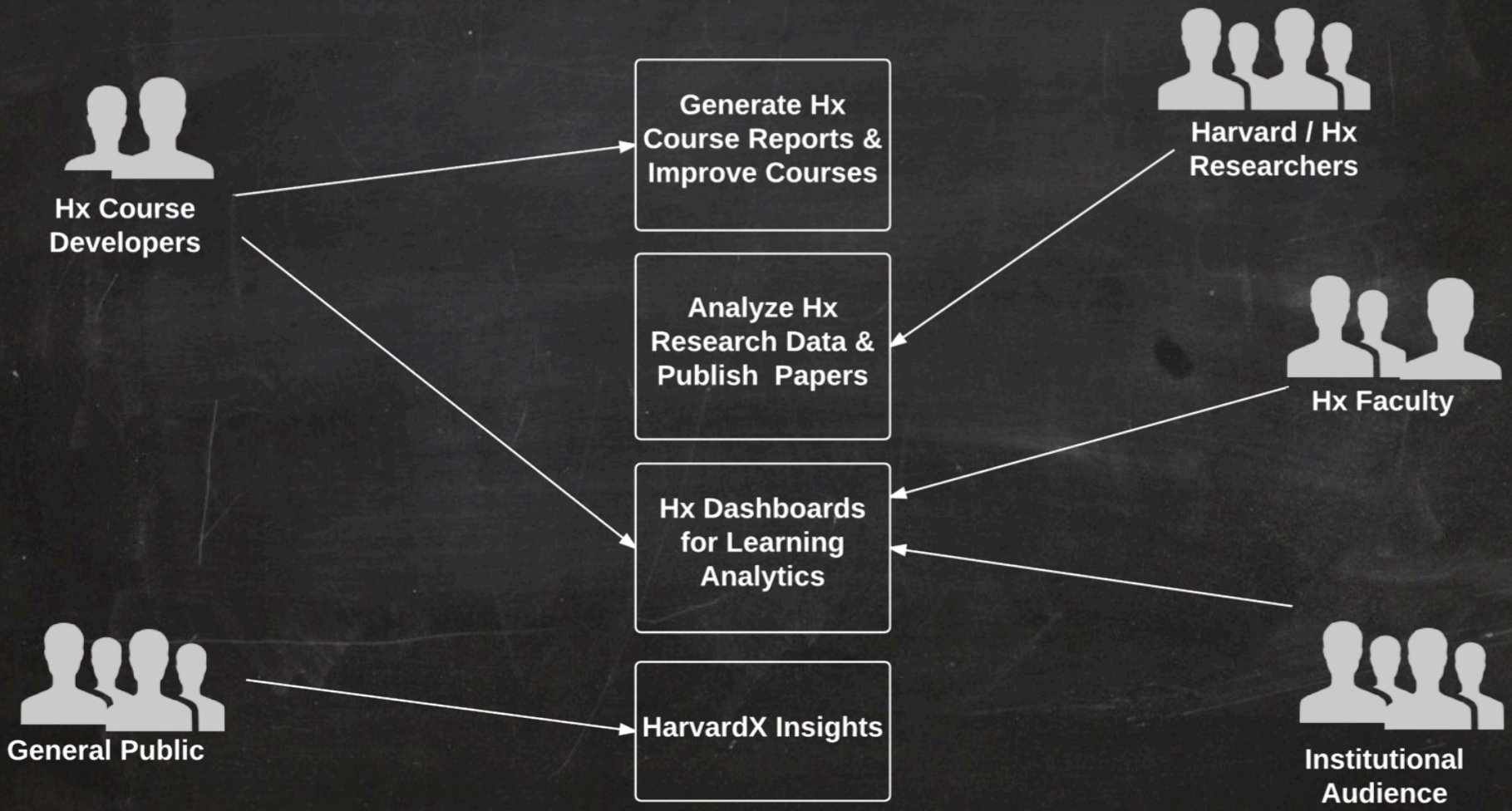
Research Data

How “Big Data” is our Data?

- **Velocity**
  - 1 day delay for activity log event extraction
  - 1 week delay for course enrollment data exports

HarvardX

Who Uses the Data and How?



HarvardX

Raw Learner Data

- **EdX Database exports**
  - Enrollment Data [CSV]
  - Profile Data [CSV]
  - Certificate Data [CSV]
  - Discussion Forum Data [JSON]
  - Courseware Data [CSV]
- **EdX Transactional Data**
  - Activity Log Data [JSON]
- **Survey Data**
  - Pre-course Survey [CSV]
  - End-of-course Survey [CSV]
  - Dynamic Survey [CSV]
  - In-Course Survey Data [CSV]

HarvardX

EdX Database Export Schemas

The `auth_user` table has the following columns.

Column	Type	Null	Key	Comment
id	int(11)	NO	PRI	
username	varchar(30)	NO	UNI	
first_name	varchar(30)	NO		# Never used
last_name	varchar(30)	NO		# Never used
email	varchar(75)	NO	UNI	
password	varchar(128)	NO		
is_staff	tinyint(1)	NO		
is_active	tinyint(1)	NO		
is_superuser	tinyint(1)	NO		
last_login	datetime	NO		
date_joined	datetime	NO		

The `student_courseenrollment` table has the following columns.

Field	Type	Null	Key	Default	Extra
id	int(11)	NO	PRI	NULL	auto_increme
user_id	int(11)	NO	MUL	NULL	
course_id	varchar(255)	NO	MUL	NULL	
created	datetime	YES	MUL	NULL	
is_active	tinyint(1)	NO		NULL	
mode	varchar(100)	NO		NULL	

The `certificates_generatedcertificate` table has the following columns.

Field	Type	Null	Key	Default	Extra
id	int(11)	NO	PRI	NULL	auto_increment
user_id	int(11)	NO	MUL	NULL	
download_url	varchar(128)	NO		NULL	
grade	varchar(5)	NO		NULL	
course_id	varchar(255)	NO	MUL	NULL	
key	varchar(32)	NO		NULL	
distinction	tinyint(1)	NO		NULL	
status	varchar(32)	NO		NULL	
verify_uuid	varchar(32)	NO		NULL	
download_uuid	varchar(32)	NO		NULL	
name	varchar(255)	NO		NULL	
created_date	datetime	NO		NULL	
modified_date	datetime	NO		NULL	
error_reason	varchar(512)	NO		NULL	
mode	varchar(32)	NO		NULL	

The `courseware_studentmodule` table has the following columns.

Field	Type	Null	Key	Default	Extra
id	int(11)	NO	PRI	NULL	auto_increment
module_type	varchar(32)	NO	MUL	problem	
module_id	varchar(255)	NO	MUL	NULL	
student_id	int(11)	NO	MUL	NULL	
state	longtext	YES		NULL	
grade	double	YES	MUL	NULL	
created	datetime	NO	MUL	NULL	
modified	datetime	NO	MUL	NULL	
max_grade	double	YES		NULL	
done	varchar(8)	NO	MUL	NULL	
course_id	varchar(255)	NO	MUL	NULL	

Source: EdX Research Data, https://edx.readthedocs.org/

HarvardX

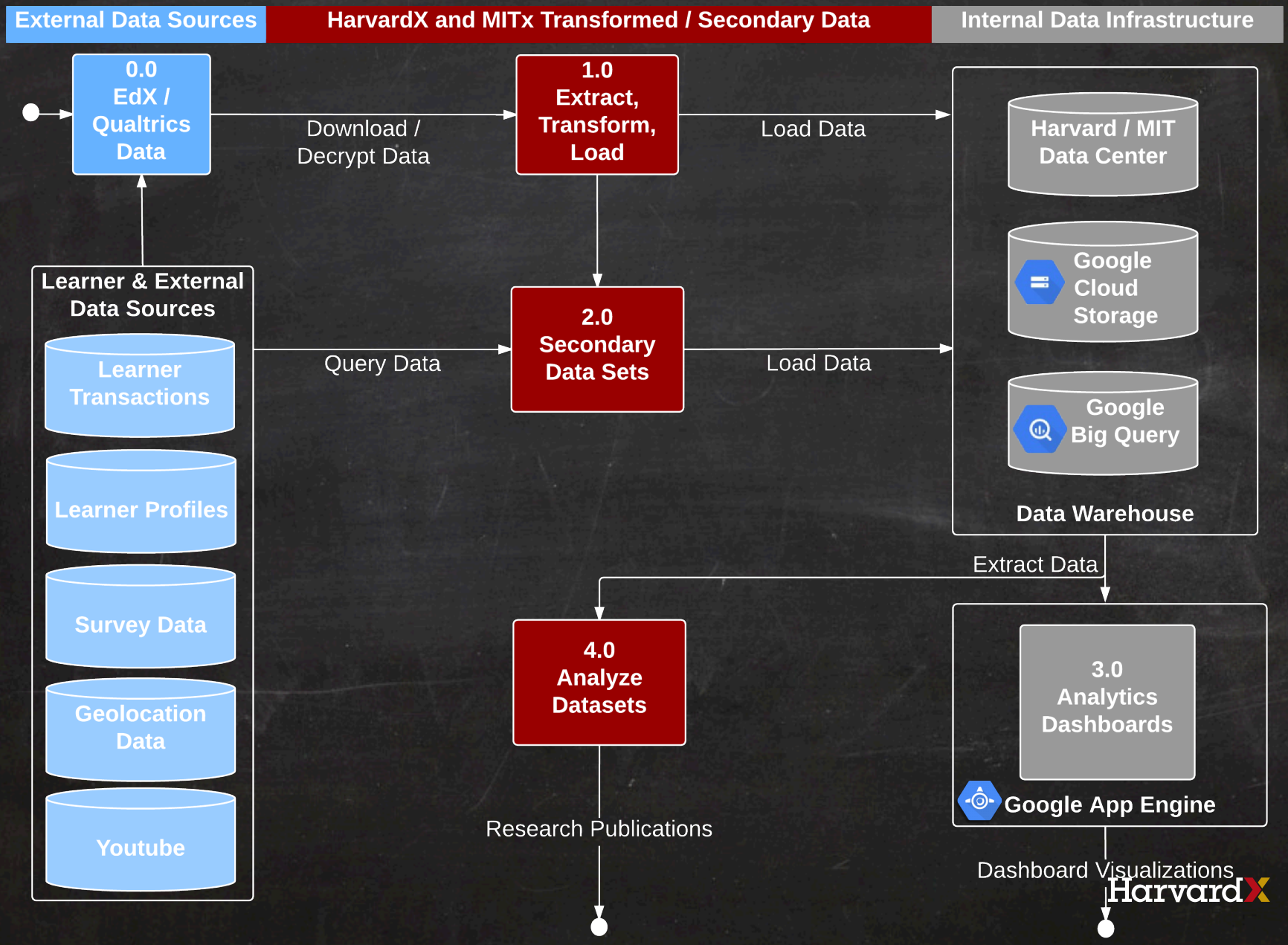
Sample EdX Tracking Log


```
{"agent": "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/30.0.1599.101 Safari/537.36", "context": {"course_id": "edx/AN101/2014_T1",
"module": {"display_name": "Multiple Choice Questions"}, "org_id": "edx", "user_id":
9999999}, "event": {"answers": {"i4x-edx-AN101-problem-a0effb954cca4759994f1ac9e9434bf4_2_1":
"yellow", "i4x-edx-AN101-problem-a0effb954cca4759994f1ac9e9434bf4_4_1": ["choice_0", "choice_2"]
"attempts": 1, "correct_map": {"i4x-edx-AN101-problem-a0effb954cca4759994f1ac9e9434bf4_2_1":
{"correctness": "incorrect", "hint": "", "hintmode": null, "msg": "", "npoints": null,
"queuestate": null}, "i4x-edx-AN101-problem-a0effb954cca4759994f1ac9e9434bf4_4_1":
{"correctness": "correct", "hint": "", "hintmode": null, "msg": "", "npoints": null,
"queuestate": null}}, "grade": 2, "max_grade": 3, "problem_id": "i4x://edx/AN101/problem/
a0effb954cca4759994f1ac9e9434bf4", "state": {"correct_map": {}, "done": null, "input_state":
{"i4x-edx-AN101-problem-a0effb954cca4759994f1ac9e9434bf4_2_1": {}, "i4x-edx-AN101-problem-
a0effb954cca4759994f1ac9e9434bf4_4_1": {}}, "seed": 1, "student_answers": {}}, "submission":
{"i4x-edx-AN101-problem-a0effb954cca4759994f1ac9e9434bf4_2_1": {"answer": "yellow", "correct":
false, "input_type": "optioninput", "question": "What color is the open ocean on a sunny day?",
"response_type": "optionresponse", "variant": ""}, "i4x-edx-AN101-problem-
a0effb954cca4759994f1ac9e9434bf4_4_1": {"answer": ["a piano", "a guitar"], "correct": true,
"input_type": "checkboxgroup", "question": "Which of the following are musical instruments?",
"response_type": "choiceresponse", "variant": ""}}, "success": "incorrect"}, "event_source":
"server", "event_type": "problem_check", "host": "precise64", "referer": "http:\/\/localhost:80
container\/i4x:\/\/edX\/DemoX\/vertical\/69dedd38233a46fc89e4d7b5e8da1bf4?action=new",
"accept_language": "en-US,en;q=0.8","ip": "NN.N.N.N", "page": "x_module",
"time": 2014-03-03T16:19:05.584523+00:00", "username": "AAAAAAAAAA"}
```

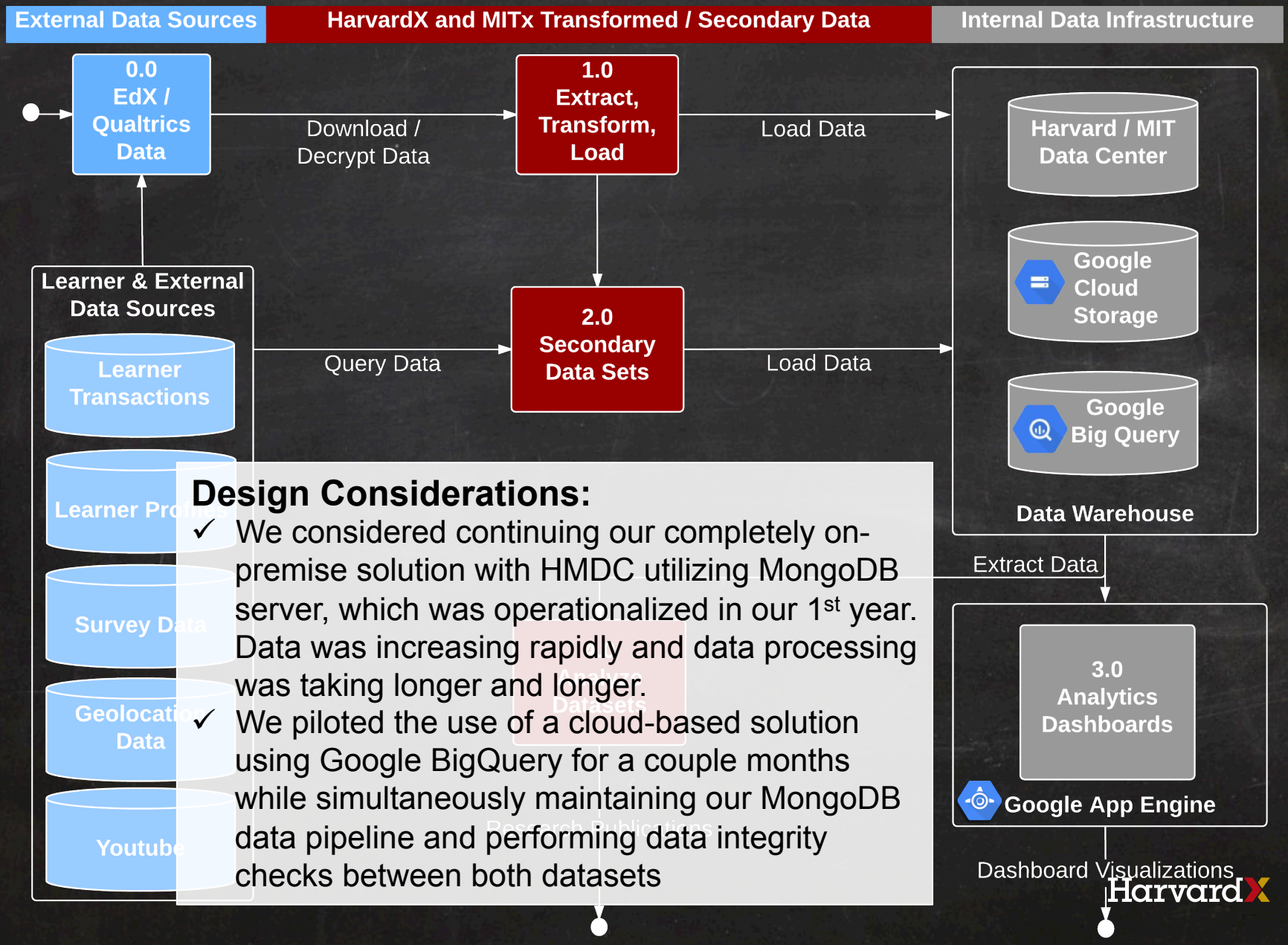

Source: EdX Research Data, https://edx.readthedocs.org/

HarvardX

HarvardX and MITx Data Flow



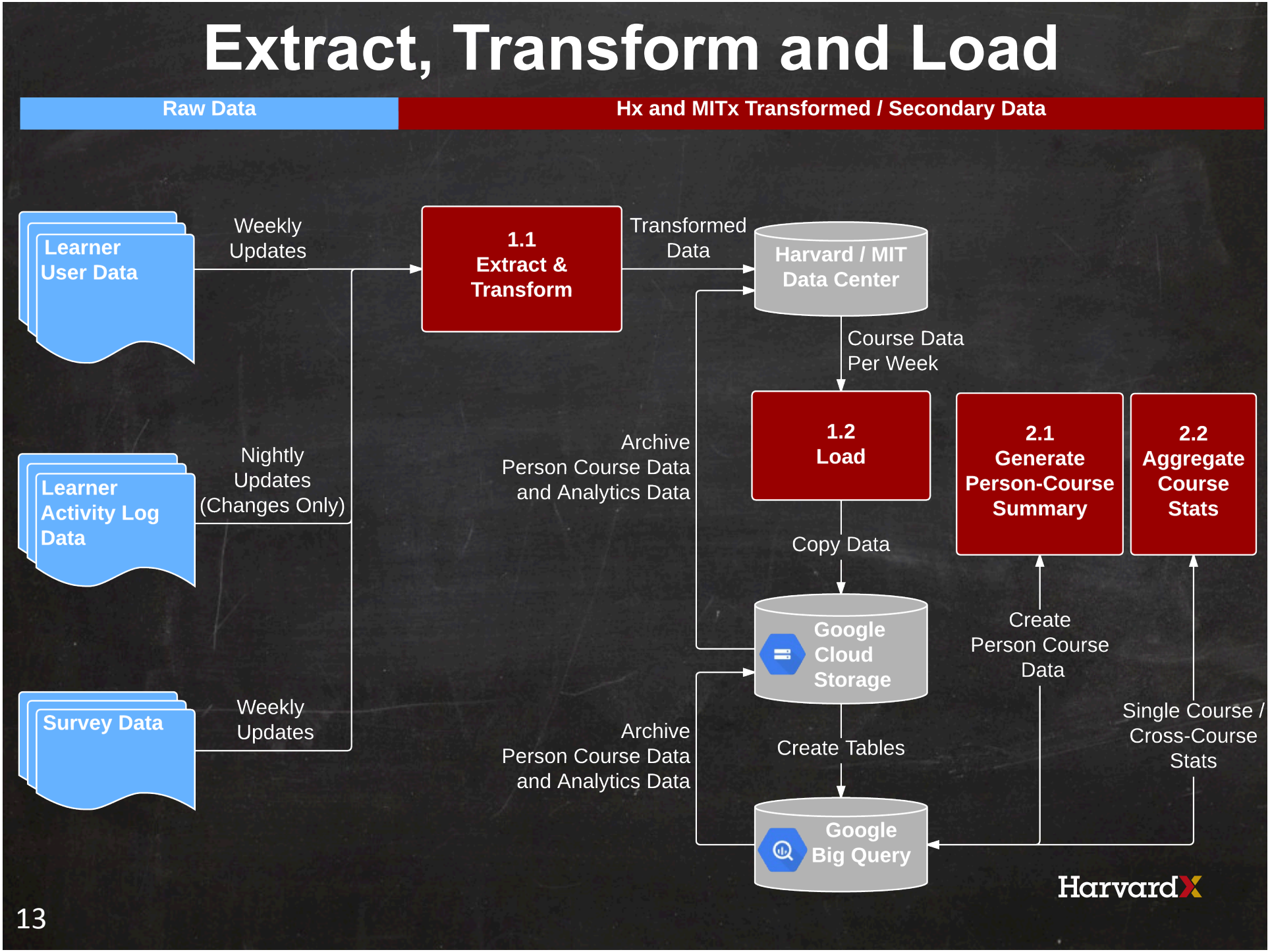
HarvardX and MITx Data Flow



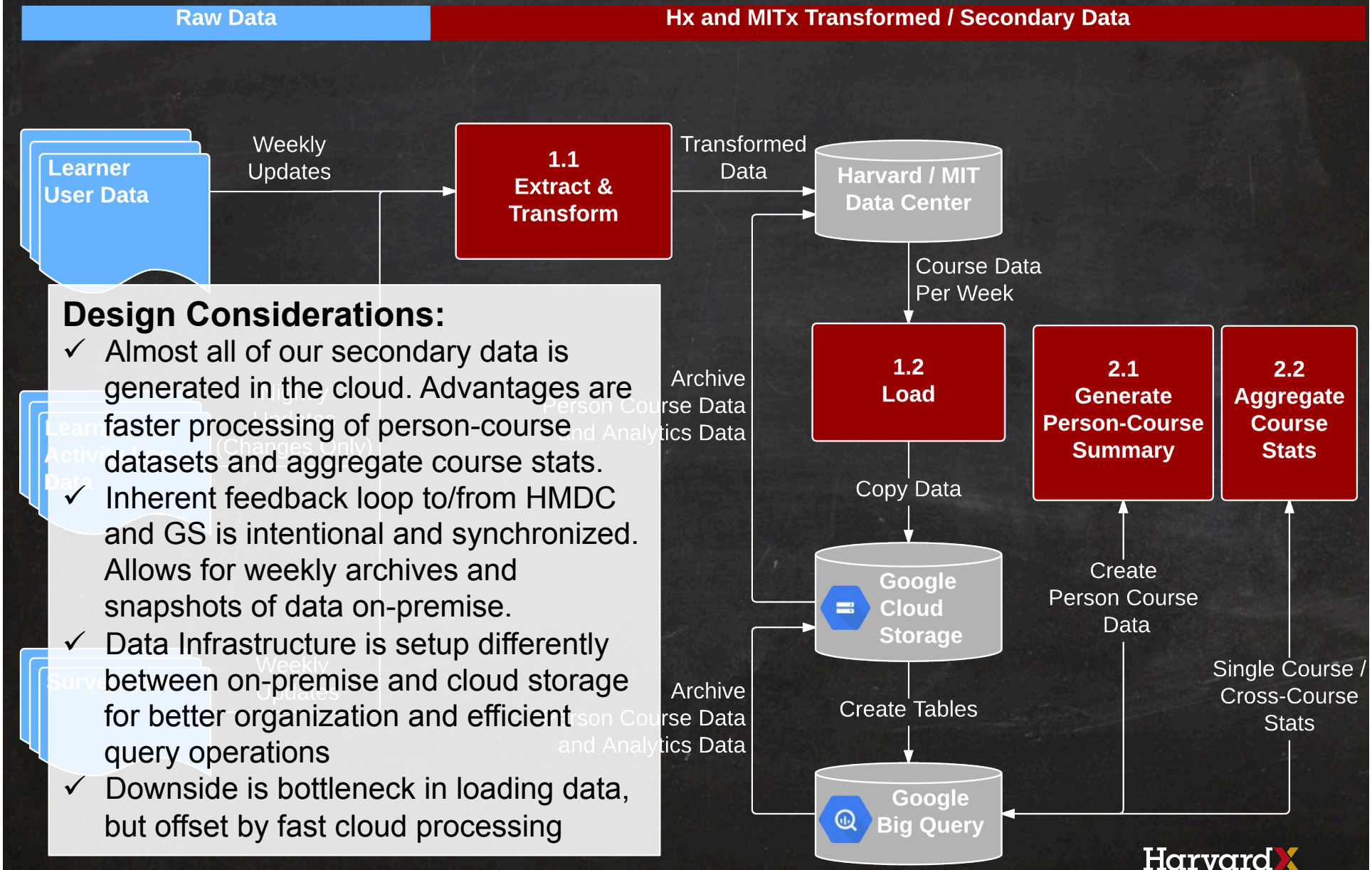
Design Considerations:

- ✓ We considered continuing our completely on-premise solution with HMDC utilizing MongoDB server, which was operationalized in our 1st year. Data was increasing rapidly and data processing was taking longer and longer.
- ✓ We piloted the use of a cloud-based solution using Google BigQuery for a couple months while simultaneously maintaining our MongoDB data pipeline and performing data integrity checks between both datasets

Extract, Transform and Load



Extract, Transform and Load



Design Considerations:

- ✓ Almost all of our secondary data is generated in the cloud. Advantages are faster processing of person-course datasets and aggregate course stats.
- ✓ Inherent feedback loop to/from HMDC and GS is intentional and synchronized. Allows for weekly archives and snapshots of data on-premise.
- ✓ Data Infrastructure is setup differently between on-premise and cloud storage for better organization and efficient query operations
- ✓ Downside is bottleneck in loading data, but offset by fast cloud processing

Data Infrastructure

Harvard-MIT Data Center (HMDC)

- HMDC retains archive of Weekly snapshots, including
  Raw, Transformed and Secondary Data sets from Google Storage

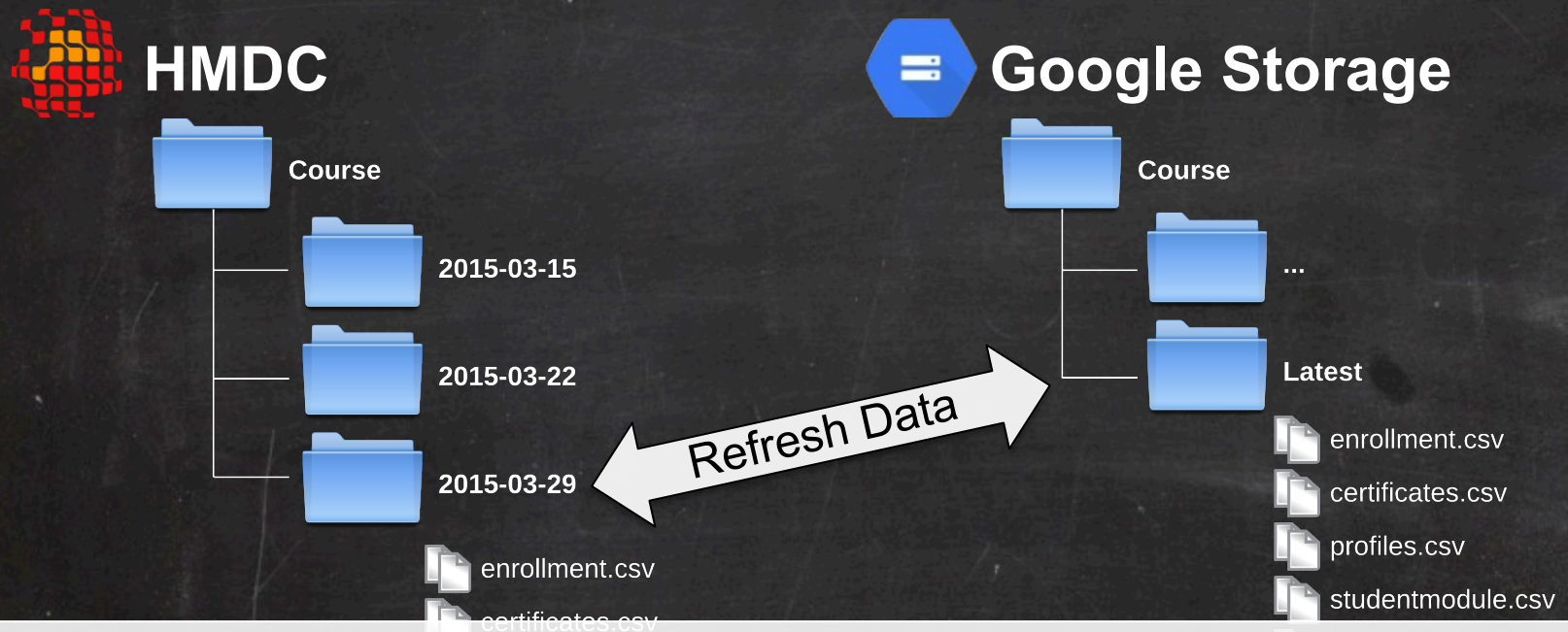
Google Storage

- Google Storage mirror the latest and most current Weekly Snapshot

HarvardX

Data Infrastructure

Learner Data Directory Structure

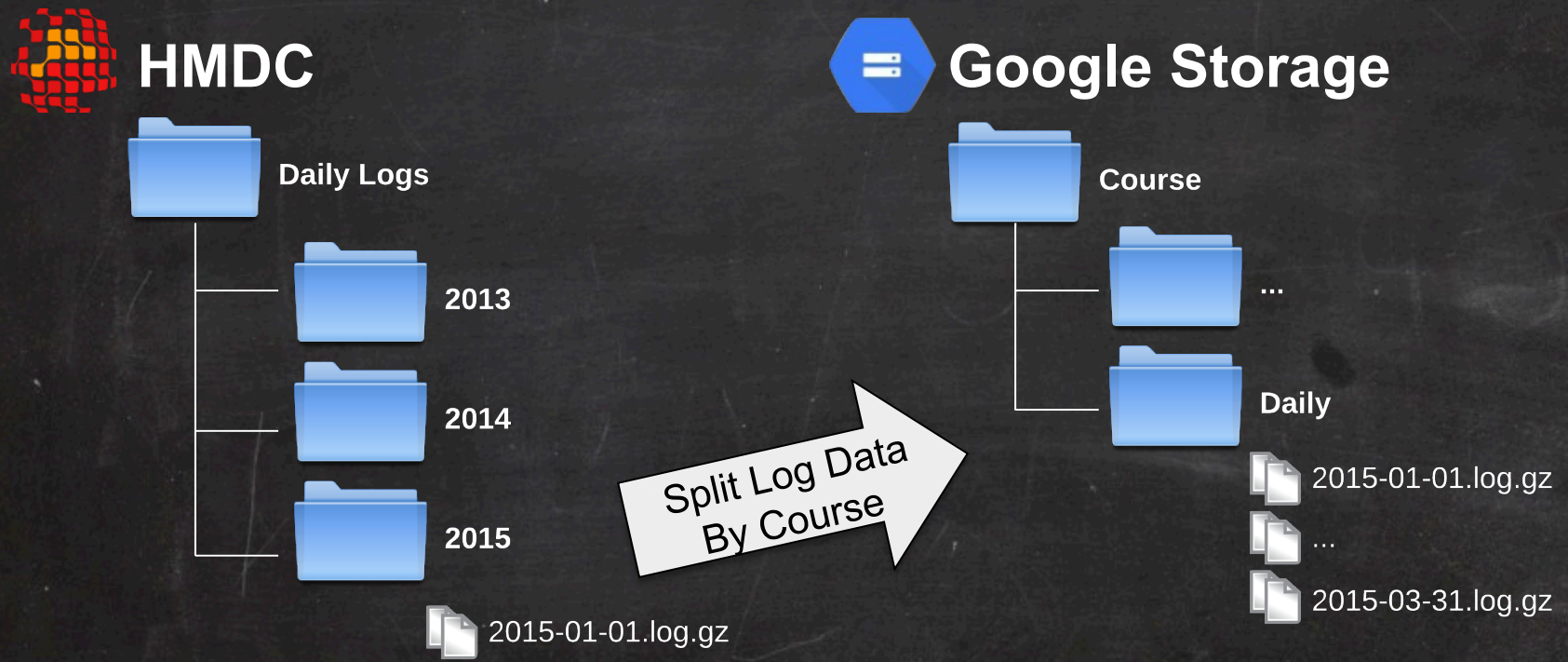


Design Considerations:

- ✓ "Latest" bucket in Google Storage is refreshed and updated weekly with raw files from HMDC with checks on file dates to ensure latest data
- ✓ Similarly, Google Storage will push changes for secondary data to HMDC for archival purposes
- ✓ HMDC is meant to archive weekly snapshots of learner data, while Google Storage is meant to reflect aggregate snapshots of desired time windows (i.e.: Year 1 from 2012-2013 or Year 2 from 2012-2014)
- ✓ Segregating subdirectories as time windows in Google Storage allows for data to be readily accessible for learning analytics dashboards

Data Infrastructure

Learner Activity Data Directory Structure



Design Considerations:

- ✓ The output data resulting from the command for splitting daily log data by course are idempotent, which means only new files are processed and not older files which have already been processed and remain unchanged.

HarvardX

Data Infrastructure [cont.]

Google BigQuery

- HarvardX / MITx have separate Project ID's
- Each Course has a "latest" dataset for current weeks data
- Each Course has Log Dataset, containing 1 table / day

"Project ID"

- **Course ID latest "Dataset"**
  - course axis
  - person course
  - stats overall
  - user info combo
- **Course ID Log "Dataset"**
  - Tracklog 20130101
  - ....
  - Tracklog 20150331

Datasets contain Tables from Transformed or Secondary Data created via BigQuery "Jobs"

HarvardX

Data Infrastructure [cont.]

Google BigQuery

- HarvardX / MITx have separate Project ID's
- Each Course has a "latest" dataset for current weeks data
- Each Course has Log Dataset, containing 1 table / day

"Project ID"

- **Course ID latest "Dataset"**
  - course axis
  - person course
  - stats overall
  - user info combo
- **Course ID Log "Dataset"**
  - Tracklog 20130101
  - ….
  - Tracklog 20150331

…**Google Biquery**, a fully-managed and cloudbased interactive query service for massive datasets… a cloud-enabled massively parallel query engine…

[Source: https://cloud.google.com/files/BigQueryTechnicalWP.pdf]

Datasets contain Tables from Transformed or Secondary Data created via BigQuery "Jobs"

HarvardX

Data Infrastructure [cont.]

Google BigQuery

- HarvardX / MITx have separate Project ID's
- Each Course has a "latest" dataset for current weeks data
- Each Course has Log Dataset, containing 1 table / day

"Project ID"

- **Course ID latest "Dataset"**
  - course axis
  - person course
  - stats overall
  - user info combo
- **Course ID Log "Dataset"**
  - Tracklog 20130101
  - ....
  - Tracklog 20150331

Projects are top-level containers in Google Cloud Platform. They store information about billing and authorized users, and they contain BigQuery data.

[Source: https://cloud.google.com/bigquery/what-is-bigquery]

Datasets contain Tables from Transformed or Secondary Data created via BigQuery "Jobs"

Data Infrastructure [cont.]

Google BigQuery

- HarvardX / MITx have separate Project ID's
- Each Course has a "latest" dataset for current weeks data
- Each Course has Log Dataset, containing 1 table / day

"Project ID"

- **Course ID latest "Dataset"**
  - course axis
  - person course
  - stats overall
  - user info combo
- **Course ID Log "Dataset"**
  - Tracklog 20130101
  - ….
  - Tracklog 20150331

Datasets allow you to organize and control access to your tables.
[Source: https://cloud.google.com/bigquery/what-is-bigquery]

Datasets contain Tables from Transformed or Secondary Data created via BigQuery "Jobs"

HarvardX

Data Infrastructure [cont.]

Google BigQuery

- HarvardX / MITx have separate Project ID's
- Each Course has a "latest" dataset for current weeks data
- Each Course has Log Dataset, containing 1 table / day

"Project ID"

- **Course ID latest "Dataset"**
  - course axis
  - person course
  - stats overall
  - user info combo
- **Course ID Log "Dataset"**
  - Tracklog 20130101
  - ....
  - Tracklog 20150331

Tables contain your data in BigQuery, along with a corresponding table schema that describes field names, types, and other information. BigQuery also supports views, virtual tables defined by a SQL query.

[Source: https://cloud.google.com/bigquery/what-is-bigquery]

Datasets contain Tables from Transformed or Secondary Data created via BigQuery "Jobs"

HarvardX

Data Infrastructure [cont.]

Google BigQuery

- HarvardX / MITx have separate Project ID's
- Each Course has a "latest" dataset for current weeks data
- Each Course has Log Dataset, containing 1 table / day

"Project ID"

- **Course ID latest "Dataset"**
  - course axis
  - person course
  - stats overall
  - user info combo
- **Course ID Log "Dataset"**
  - Tracklog 20130101
  - ….
  - Tracklog 20150331

Jobs are actions you construct and BigQuery executes on your behalf to load data, export data, query data, or copy data

[Source: https://cloud.google.com/bigquery/what-is-bigquery]

Datasets contain Tables from Transformed or Secondary Data created via BigQuery "Jobs"

HarvardX

Data Infrastructure [cont.]

Google BigQuery

- HarvardX / MITx have separate Project ID's
- Each Course has a "latest" dataset for current weeks data
- Each Course has Log Dataset, containing 1 table/day

"Project ID"

- **Course ID latest "Dataset"**
  - course axis
  - person course
  - stats overall
  - user info combo
- **Course ID Log "Dataset"**
  - Tracklog 20130101
  - ....
  - Tracklog 20150331

"Dataset" Design Considerations:

- ✓ "Latest" dataset contains aggregate statistics from ALL datasets until now (default for analytics dashboards)
- ✓ Similarly, other time specific datasets can be created, representing different snapshots in time (i.e.: 2012-2013, or 2012-2014 for example). Benefits include readily selectable datasets for our analytics dashboard
- ✓ "Log" dataset grows in size daily, with only new daily log tables being added during the batch process

Datasets contain Tables from Tracking Logs, Courseware Data created via BigQuery Jobs

HarvardX

Data Infrastructure [cont.]

Google BigQuery

“Project ID”

- **Course ID latest “Dataset”**
  - course axis
  - person course
  - stats overall
  - info combo
  - Course ID Log “Dataset”

Schema

Metadata

Table Details: person_course

Schema

course_id	STRING	NULLABLE	course ID in the st
user_id	INTEGER	NULLABLE	user ID number
username	STRING	NULLABLE	username

Table Details: [redacted]

Description

Computed by [redacted] at 2014-12-12 16:57:21.716025 processing 895338970 bytes in 26.17 sec with this SQL:

```
select username,
    [redacted] as course_id,
    date(time) as date,
    sum(bevent) as nevents,
    sum(bprogress) as nprogcheck,
    sum(bshow_answer) as nshow_answer,
    sum(bvideo) as nvideo,
    sum(bproblem_check) as nproblem_check,
```

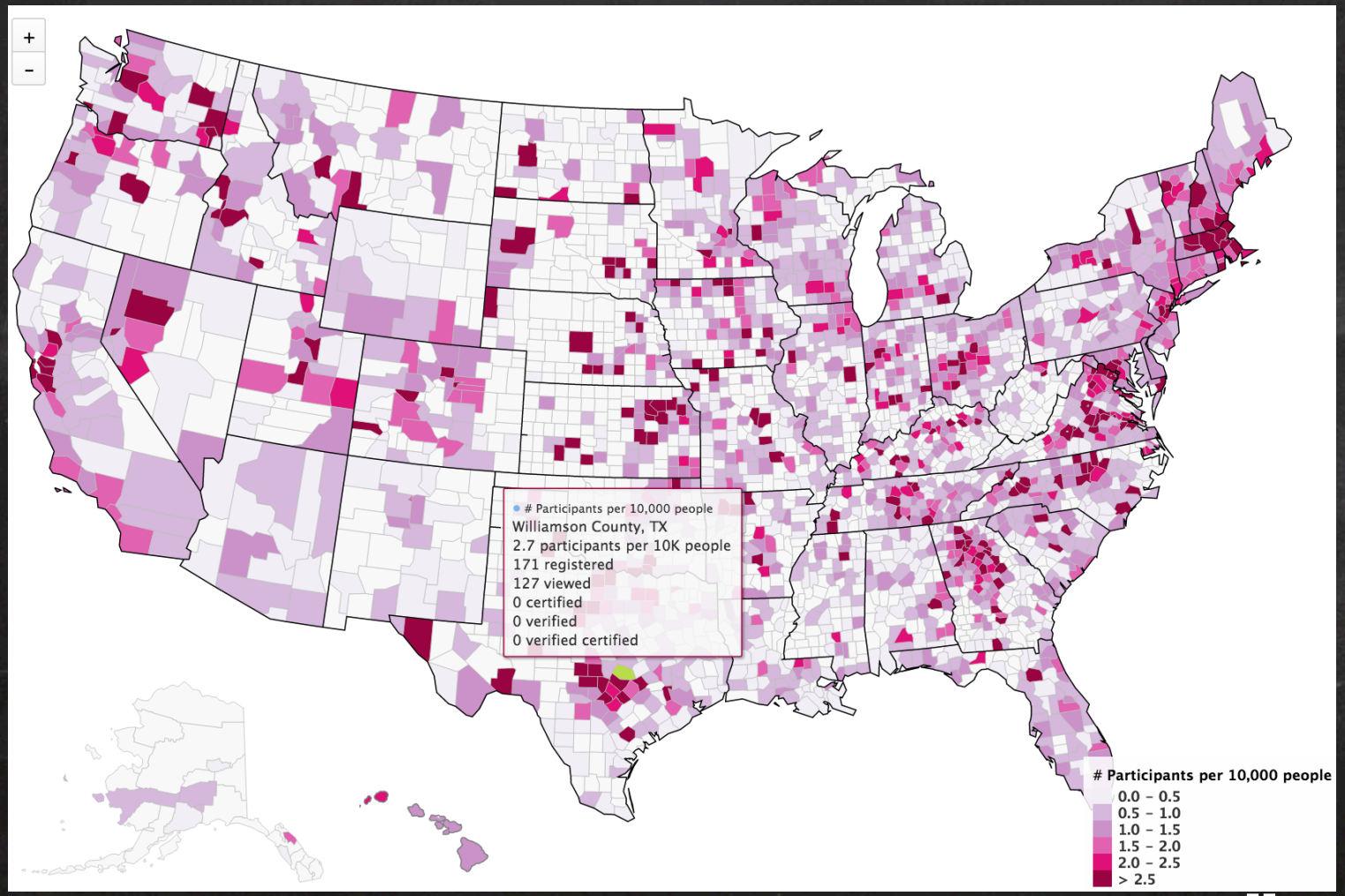
Design Considerations:

- ✓ Schema details can be stored for each table with a type and field description.
- ✓ Additionally, upon import of primary “secondary” datasets, the load scripts perform a schema check to ensure the ingested data is correct
- ✓ Custom Metadata about each table is generated with details including what user computed the data, time, how much data was processed, how long the job took, and what query was used to generate the table.
- ✓ Additional data such as process status variables and debugging can be added as well

HarvardX

Analytics Dashboard

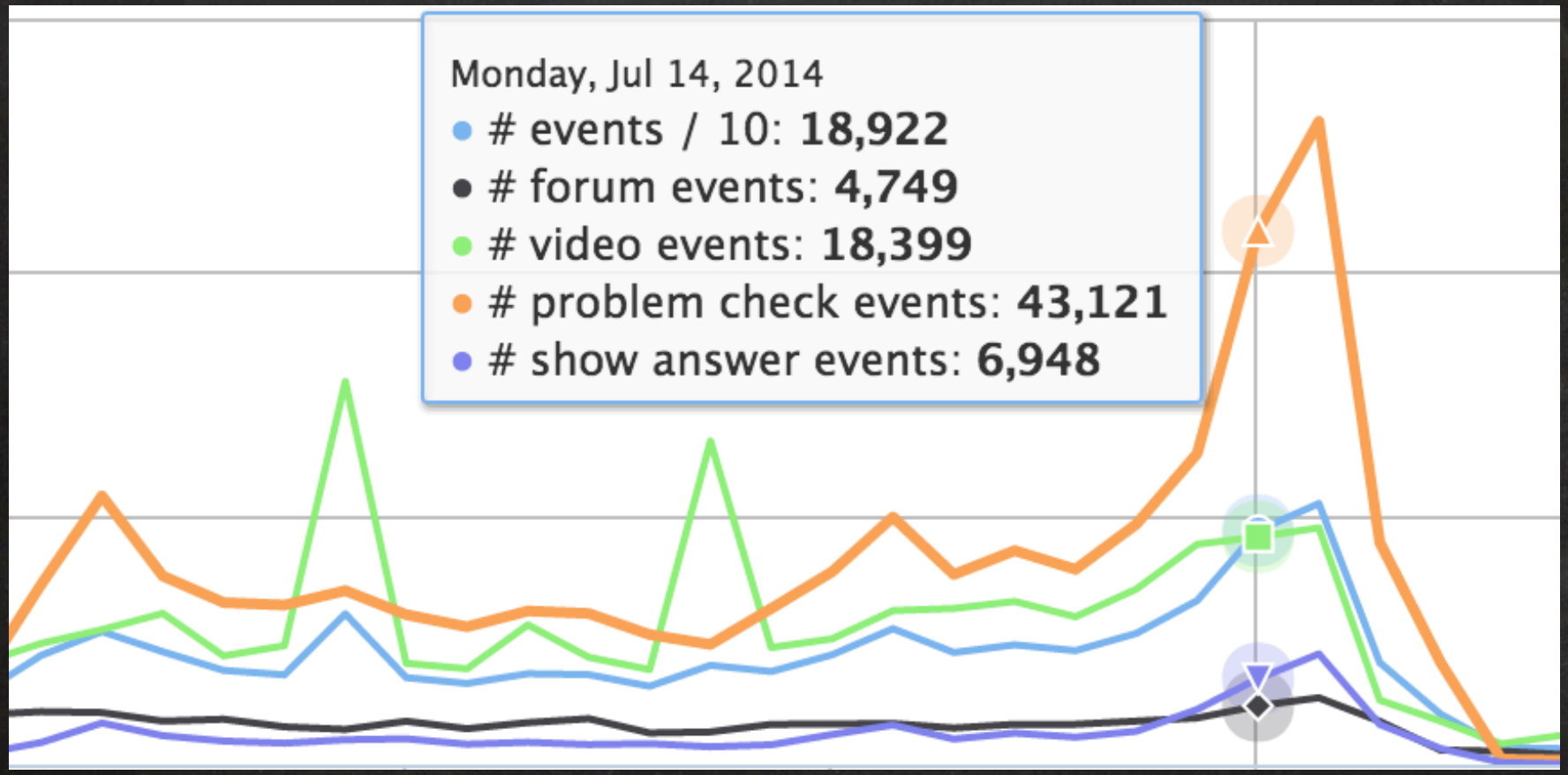
Sample Visualizations



Description

Analytics Dashboard

Sample Visualizations [cont.]



HarvardX

Description

Analytics Dashboard
Framework

Back-end powered by Google App Engine, Python and WebApp2

- Runs on Google's scalable infrastructure
- WebApp2, web application environment, is installed w/ App Engine environment
- Google App Engine features include:
  - MemCache: distributed, in-memory cache
  - OAuth: Use google app accounts for authentication
  - Python Runtime: Built-in web application environment, WebApp2, interacts with App Engine web server using WSGI protocol
  - Logs: Maintains two types; application logs and access request logs (time, username, URL), viewable from Admin console

Front-end technologies/libraries/plug-ins:

- Javascript + jQuery + Datatables
- Ajax approach

HarvardX

Analytics Dashboard
Visualizations

Provides Analytics and Visualizations in multiple ways:

1] Single Course Analytics

- Determined by Course ID
- Used by Course Developers

2] Across-Course Analytics

- Across all of HarvardX / MITx
- Used by institutional administration

3] Multi-Course Analytics

- Specify groups of courses, categorized anyway, such as by sequence of modules, department/school or curricular area
- Used by researchers

Daily / Cum. Enrollments
Demographic Distribution
Problem Analysis
Time-on-Task
Log Event Activity

Design Considerations:

✓ The analytics dashboard views should function to satisfy the different stakeholders

HarvardX

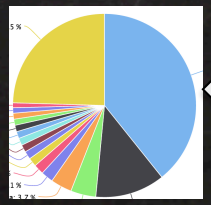
Analytics Dashboard
Visualizations [Cont.]

Provides Analytics and Visualizations in multiple ways [cont]:

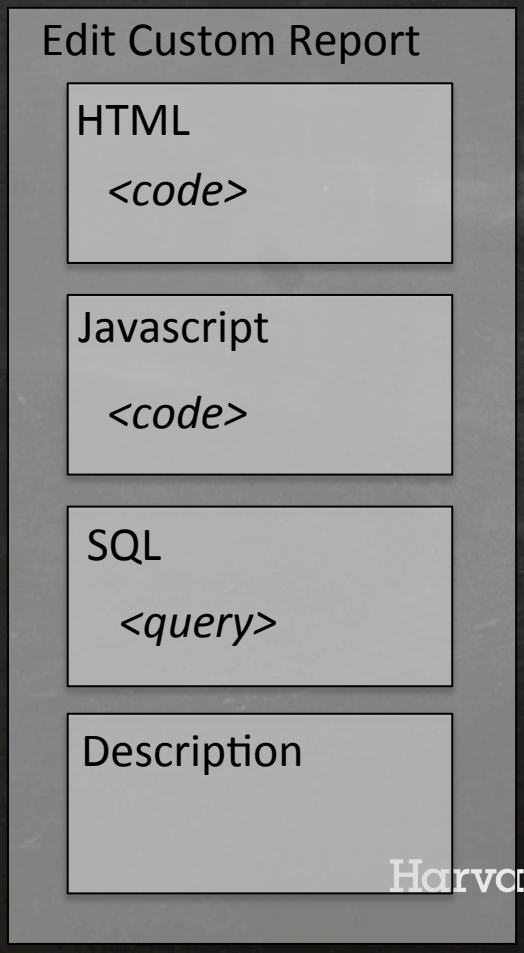
4] Custom Reports

- YAML Format
- Reports can be imported or exported
- Allows for experimentation of graphs using custom HTML, CSS, Javascript and BigQuery jobs
- If generalizable, custom Report Graphs can then be embedded in desired analytics page(s)

[i.e: Provide age histogram for Single Course View, across all courses]



View



Analytics Dashboard
Visualizations [Cont.]

Optionally Provide Data Tables for Visualizations

- Dynamically Sort and Search graph data

Appendix A: Course Listings

Show 10 entries Search:

Institution	Course Number	Short Title[1]	Course Title	Instructors[2]	Course Launch
HarvardX	PH525x	Genomics	Data Analysis for Genomics	Rafael Irizarry	4/7/2014
HarvardX	PH201x	HealthSoc	Health and Society	Ichiro Kawachi	11/15/2013
HarvardX	GSE1x	ImmunChange	Unlocking the Immunity to Change	Robert Kegan, Lisa Lahey	3/11/2014
HarvardX	AI12.1x	PoetNewEng	Poetry in America: The Poetry of Early New England	Elisa New	10/31/2013
HarvardX	PH278x	HealthEnv	Human Health & Global Environmental Change	Aaron Bernstein and Jack Spengler	5/15/2013
HarvardX	HSPH210x	USHealth	United States Health Policy	John McDonough	4/7/2014
HarvardX	HSPH-HMS214x	HealthTrials	Fundamentals of Clinical Trials	James Ware, Elliott Antman, Julie Buring, Graham McMahon, Marcia Testa, Robert Truog	10/14/2013
HarvardX	SW12.5x	ChinaX-e	From Global Empire to Global Economy (ChinaX)	Peter Bol, Bill Kirby	4/24/2014
HarvardX	ER22x	Justice-1	Justice	Michael Sandel	3/2/2013
HarvardX	ER22.1x	Justice-2	Justice, v2	Michael Sandel	4/8/2014

Showing 1 to 10 of 68 entries Previous 1 2 3 4 5 6 7 Next

Hx/MITx Year 2 Online Appendix A: Course Listings
http://harvardx-mitx-year2.odl.mit.edu/appendix.html

HarvardX

Description

Benefits thus far…

On-premise vs. Hybrid-Cloud solution

Immediate benefits

Reduced Cost

Instead of spending thousands $$$ to upgrade hard disk space to expand our MongoDB server for both Harvard / MIT, we now spend under $40 / month for redundant storage and queries (charged at $5 / TB). Google BigQuery charges by the query and stores data in columnar storage structure, meaning queries can specify only relevant columns which is more efficient and cost effective. Significantly reduced total cost of ownership, and requires no overhead to manage servers, capacity planning, patch updates, etc…

Faster Processing of Research Datasets

Scaling up with more courses and registrants is more manageable now, due to faster query times. For instance, a BigQuery job to process a year's worth of tracking log data [26.9 million events] to produce a person-course dataset for Hx largest course, CS50, takes 23 seconds! Previously, nightly jobs that created our secondary data sets usually started at 1AM and then finished at ~12PM (even after code improvements reducing time by ~50%).

Quicker Interactive Data Analysis

Since Google BigQuery is incredibly fast, being able to query billions of rows, write and return results in seconds, analysis that took hours using on-premise MongoDB can be done much quicker. Researchers can find out answers much more quickly and test different theories.

And Lastly…

A special thanks goes to I. Chuang @ MIT and J. Waldo @ Harvard for implementing a lot of the ground work that HarvardX and MITx uses jointly for our Research Data Infrastructure

HarvardX

Questions?

HarvardX

Backup

HarvardX

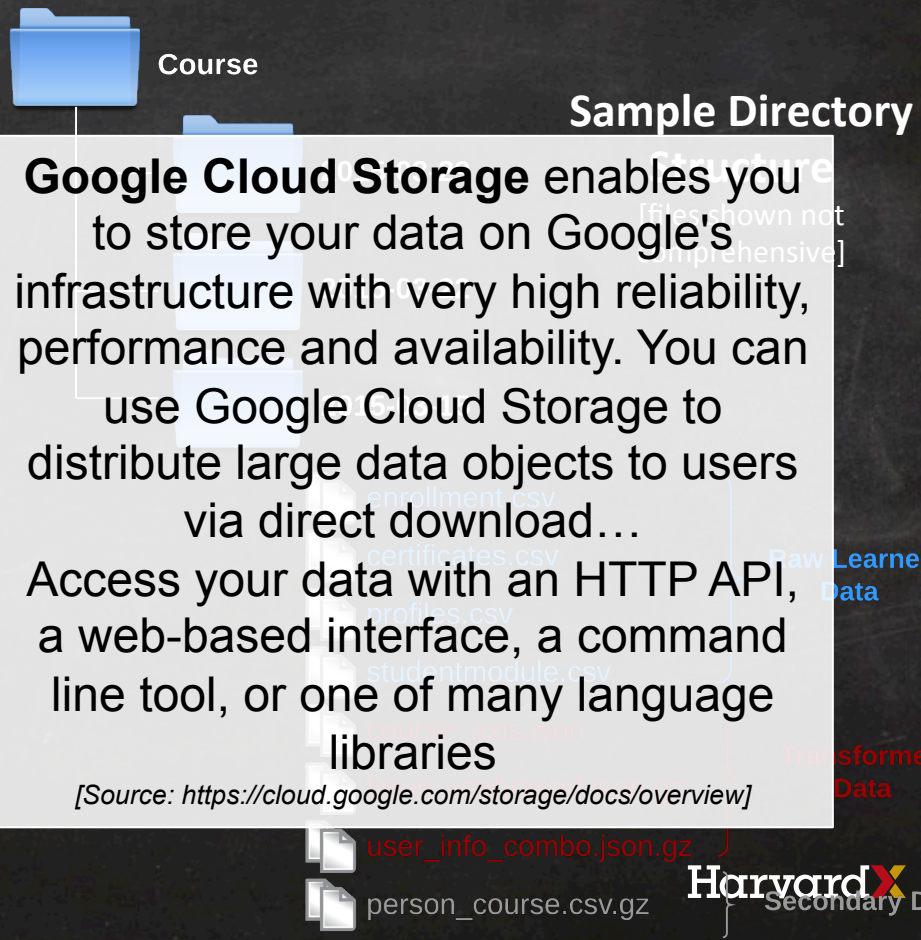
Data Infrastructure

Harvard-MIT Data Center (HMDC)

- HMDC retains archive of Weekly snapshots, including Raw, Transformed and Secondary Data sets from Google Storage

Google Storage

- Google Storage mirrors the latest and most current Weekly Snapshot



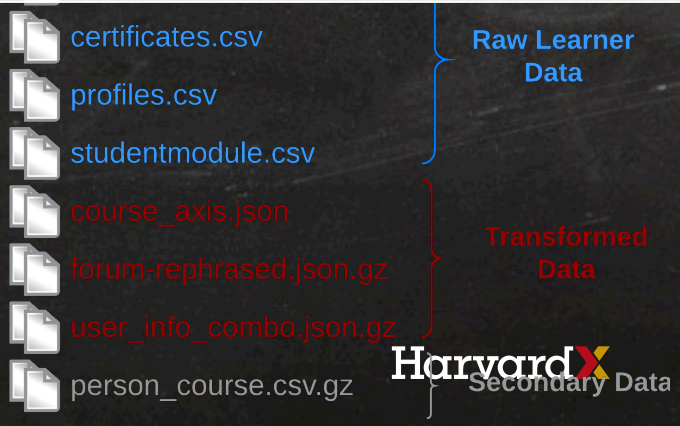
Data Infrastructure

Harvard-MIT Data Center (HMDC)

- HMDC retains archive of Weekly snapshots, including Raw, Transformed and Secondary Data sets from Google Storage

HMDC offers advanced computing facilities and infrastructure, including powerful and yet easy-to-use research computing tools, cluster computing, application and server hosting, and on-site computer labs... [HMDC] continues to serve as the principal distributor of social science data for Harvard and MIT.

[Source: http://projects.iq.harvard.edu/hmdc]



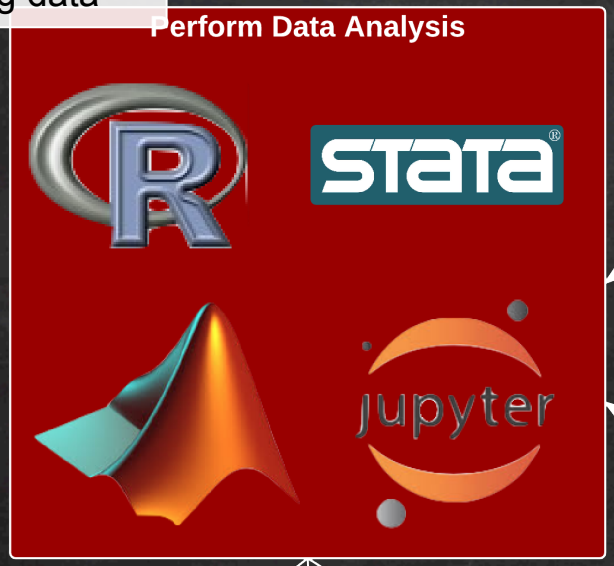
Google Storage

- Google Storage mirrors the latest and most current Weekly Snapshot

Analyzing Research Data

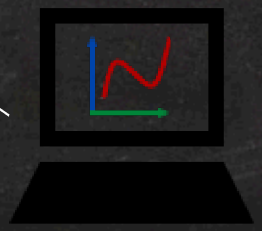
Design Considerations:

- ✓ Google BigQuery was built to be Interactive ad-hoc and trial-and-error data analysis tool for big data



Extract Data

Download Dashboard Tables



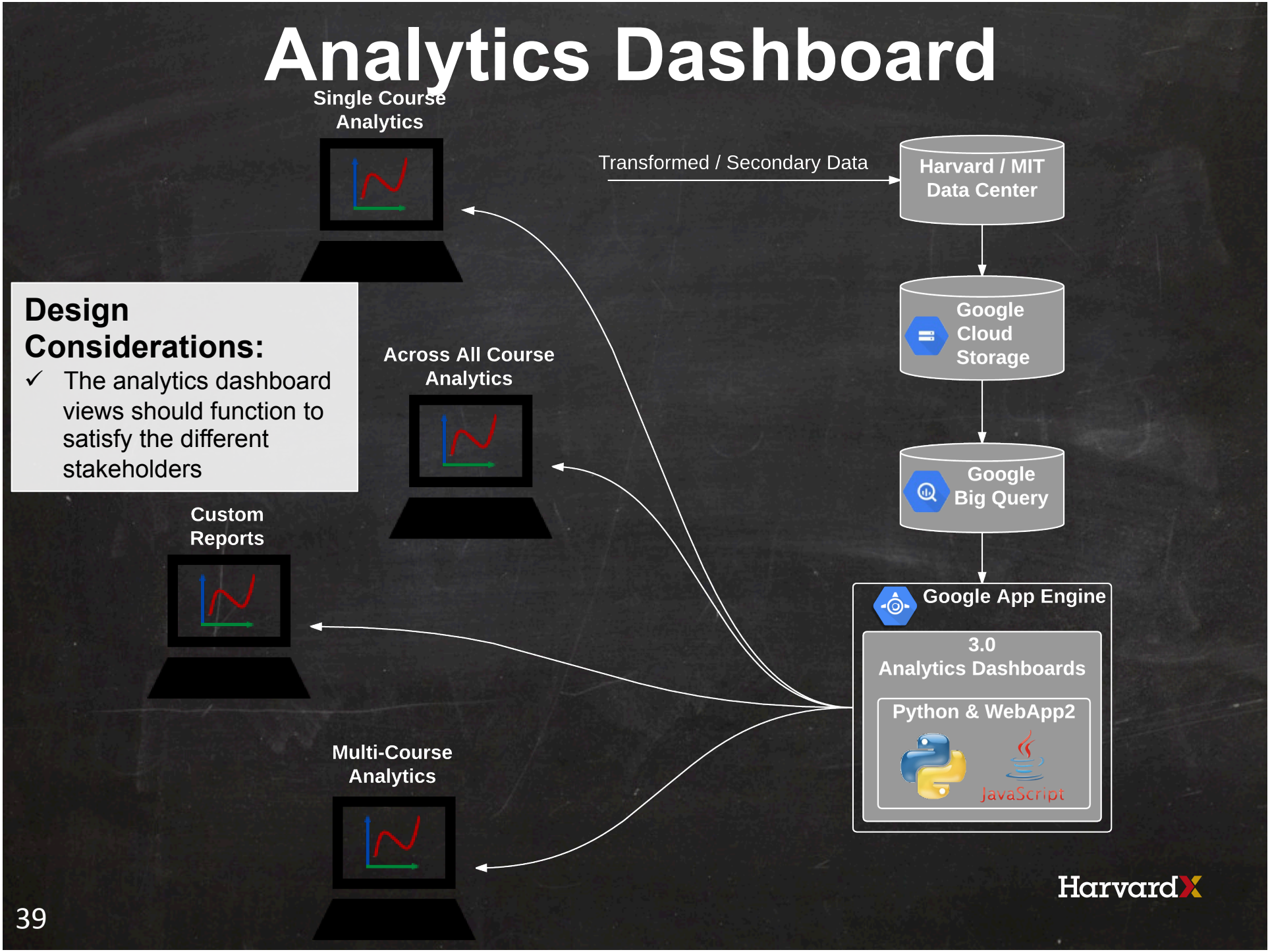
Year End Report

Research Publications

Business Analytics

HarvardX

Analytics Dashboard



Design Considerations:

- ✓ The analytics dashboard views should function to satisfy the different stakeholders

HarvardX