Student LMS Data: Data Pipeline and Privacy Considerations

Dustin Tingley

Glenn Lopez

HARVARD

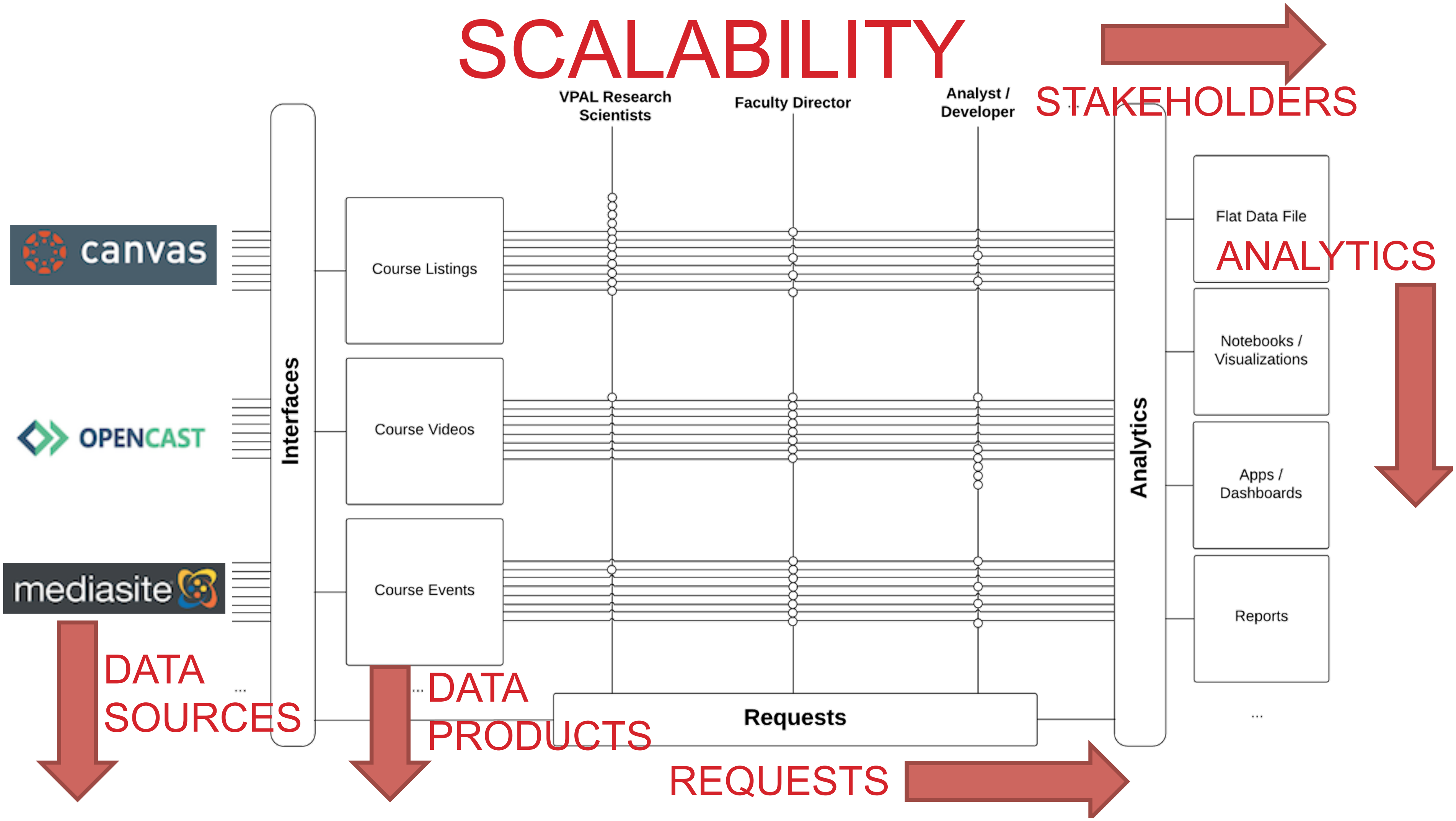
OFFICE OF THE VICE PROVOST FOR ADVANCES IN LEARNING

RESEARCH

CHALLENGES

- Canvas LMS Flat Files (Daily dumps vs. Megadumps)
- Canvas Rest API
- Video Management Platforms
- External LTI tools / Additional Platforms
- Harvard is a Decentralized Institution (3 Canvas Instances)
- Multiple Stakeholders (Administrators, Researchers, Faculty, IT, Course Teams, etc…)
- Different Data Sources with differing levels of data quality
- Security
- Answering Questions leads to more Questions. How do we keep up? (OWN YOUR DATA!)

SCALABILITY



STAKEHOLDERS

VPAL Research Scientists

Faculty Director

Analyst / Developer

...

canvas

OPENCAST

mediasite

...

Interfaces

Course Listings

Course Videos

Course Events

...

Requests

Analytics

Flat Data File

Notebooks / Visualizations

Apps / Dashboards

Reports

...

ANALYTICS

DATA SOURCES

DATA PRODUCTS

REQUESTS

QUESTIONS FOR DATA INFRASTRUCTURE

- How can we generalize Data Pipelines for different Data Sources (e.g. SIS, other instances, LTI-tools)
- How can we generalize Approaches to Processing data for different Stakeholders?
- How do you support a grow # of stakeholders who want access to the data?
- For those that are given access to the data, how do you protect the identities of students and other sensitive information that shouldn’t be used for research?
- How do we reveal identities in the event that there is an approved use case (e.g.: Faculty Dashboards)?

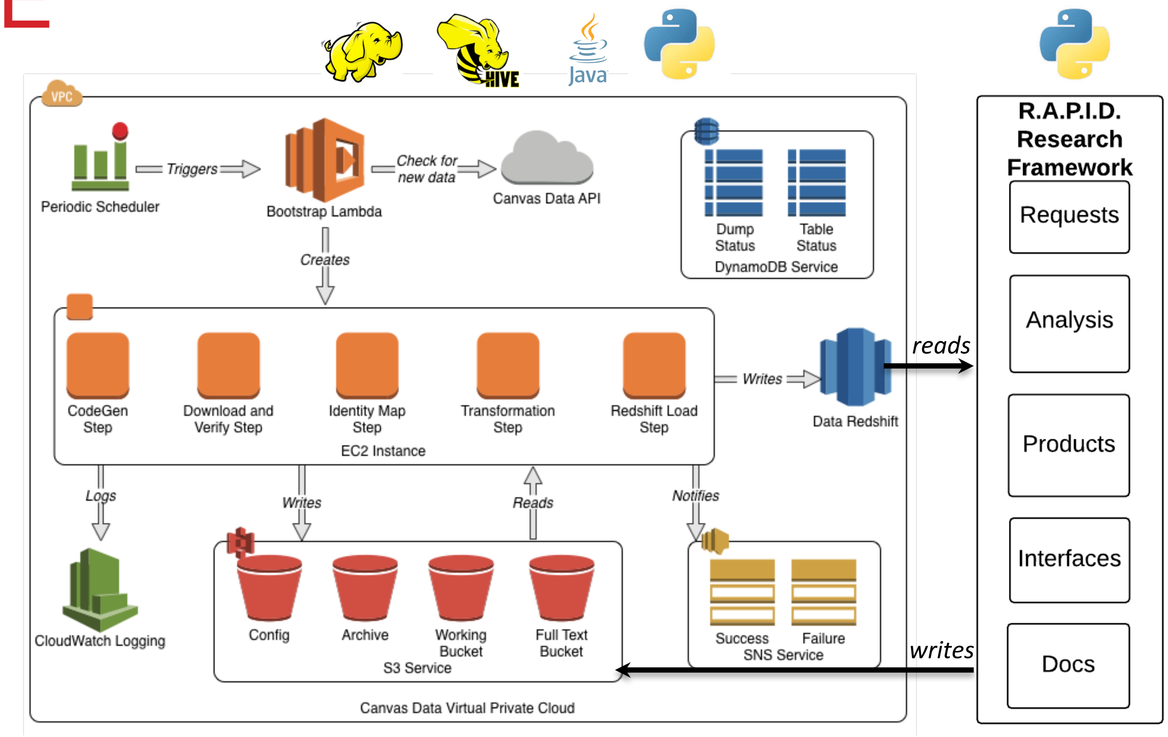
QUESTIONS FOR PRIVACY AND SENSITIVE DATA

- Sensitive Data
  - Are there Variables that could be directly, or with some degree of work, used to identify students? Identification of individuals using Sensitive variables would be more direct. Level of effort would require direct Canvas access to the course or through external Canvas resources.
- Data to Exclude
  - Are there Variables that should never provided to anyone under any circumstance? If so, should these variables be purged upon coming into the system?

HARVARD
OFFICE OF THE VICE PROVOST FOR ADVANCES IN LEARNING
RESEARCH

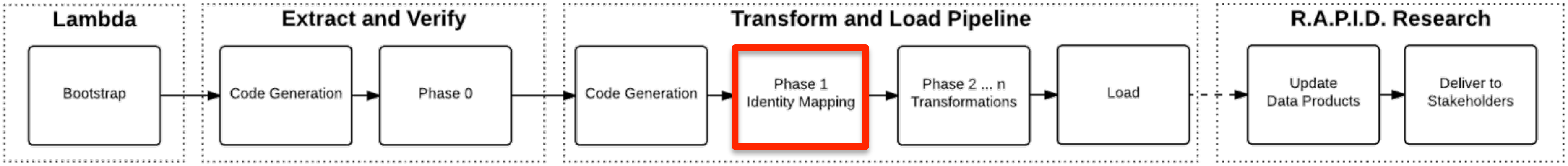
GENERIC DATA PIPELINE & ARCHITECTURE

- Hadoop, Hive, Redshift, S3, Lambda, Python, Java, SQL
- Harvard has developed a Data Pipeline to **process**, **de-identify** and **analyze** Canvas Data Flat file dumps (+ Canvas Rest API)
- Generalize for other largely used Platforms (LTI-Tools)



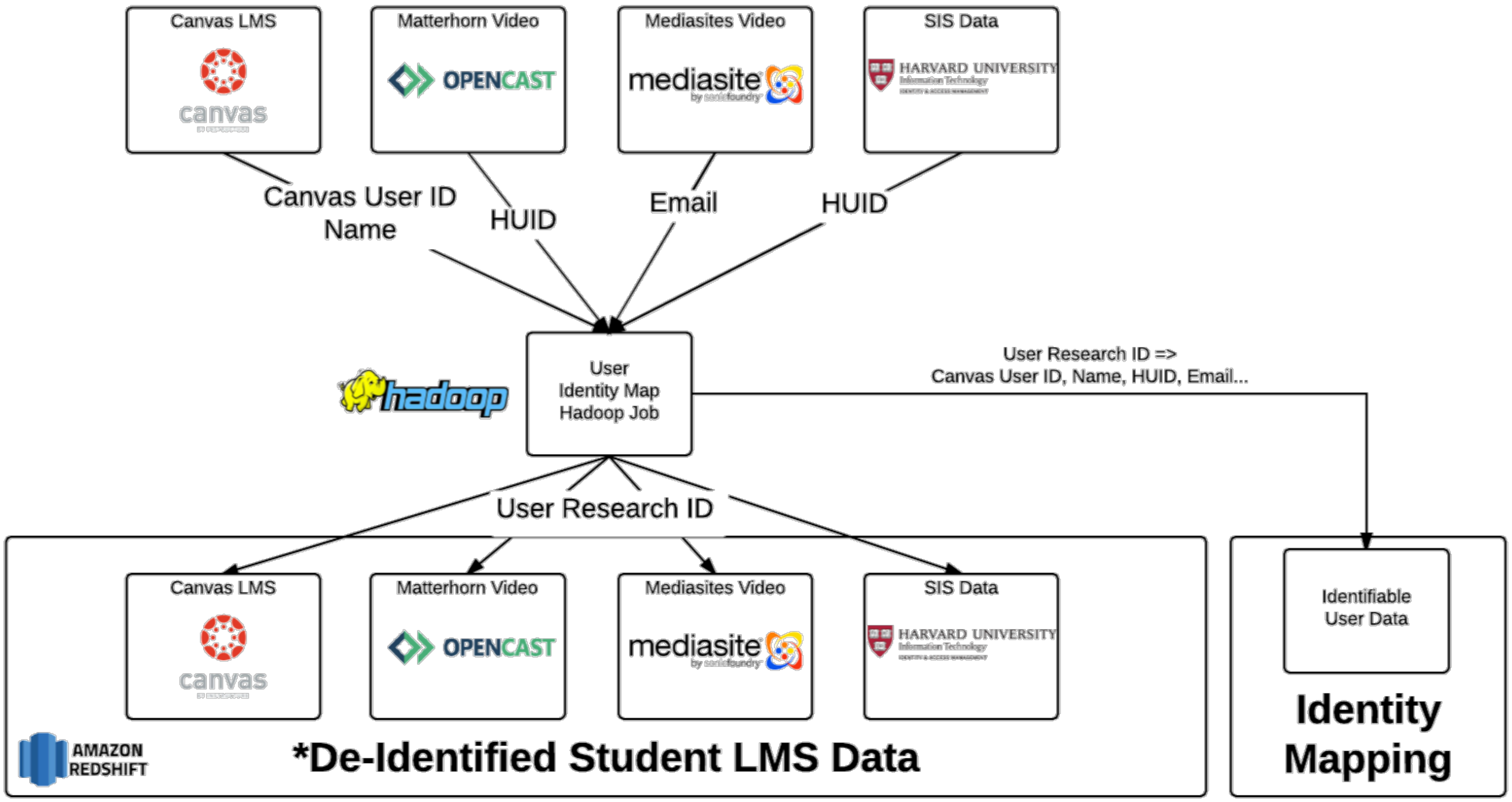
PHASE 1: IDENTITY MANAGEMENT

- Runs on EMR (Elastic Map Reduce) Instance
- Hadoop job that performs Identity Map
  - For every table that has a user id, produce a mapping from data set's main identifier to all known identifiers for that user
- Hadoop job that performs Identity Scrub
  - Replace PII with Research UUID persistent across all tables



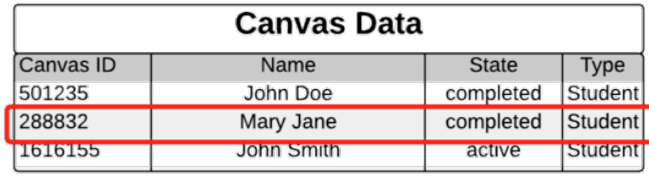
PHASE 1: IDENTITY MANAGEMENT

Identifiable Student LMS Data



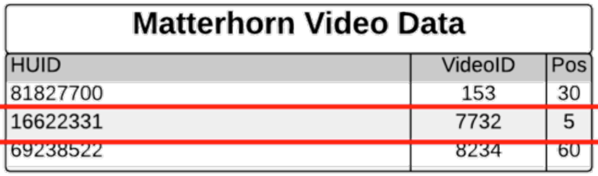
EXAMPLE: IDENTITY MANAGEMENT

Identifiable Student LMS Data



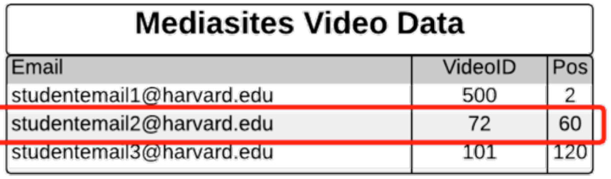
Canvas Data

Canvas ID	Name	State	Type
501235	John Doe	completed	Student
288832	Mary Jane	completed	Student
1616155	John Smith	active	Student



Matterhorn Video Data

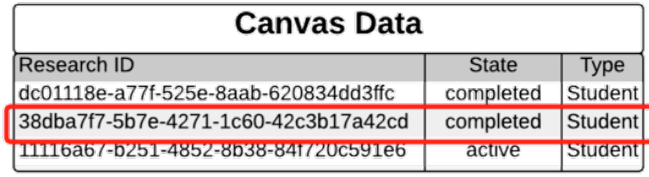
HUID	VideoID	Pos
81827700	153	30
16622331	7732	5
69238522	8234	60



Mediasites Video Data

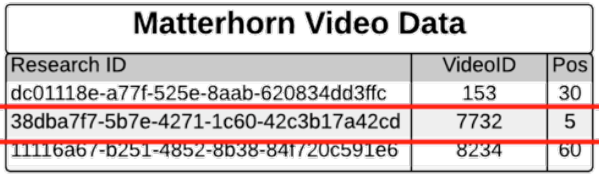
Email	VideoID	Pos
studentemail1@harvard.edu	500	2
studentemail2@harvard.edu	72	60
studentemail3@harvard.edu	101	120

*De-Identified Student LMS Data



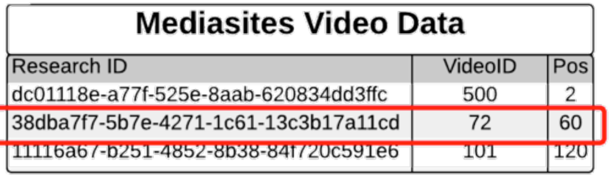
Canvas Data

Research ID	State	Type
dc01118e-a77f-525e-8aab-620834dd3ffc	completed	Student
38dba7f7-5b7e-4271-1c60-42c3b17a42cd	completed	Student
11116a67-b251-4852-8b38-84f720c591e6	active	Student



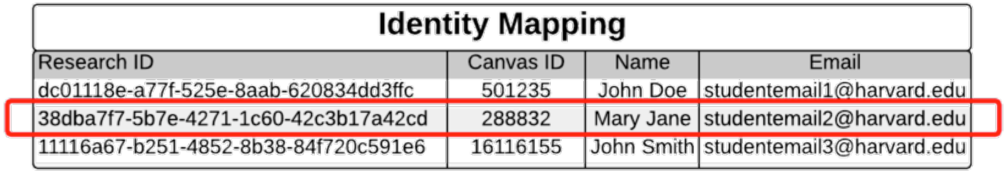
Matterhorn Video Data

Research ID	VideoID	Pos
dc01118e-a77f-525e-8aab-620834dd3ffc	153	30
38dba7f7-5b7e-4271-1c60-42c3b17a42cd	7732	5
11116a67-b251-4852-8b38-84f720c591e6	8234	60



Mediasites Video Data

Research ID	VideoID	Pos
dc01118e-a77f-525e-8aab-620834dd3ffc	500	2
38dba7f7-5b7e-4271-1c61-13c3b17a11cd	72	60
11116a67-b251-4852-8b38-84f720c591e6	101	120



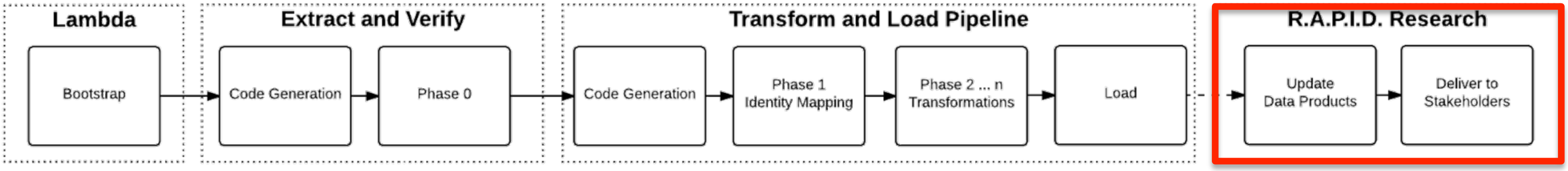
Identity Mapping

Research ID	Canvas ID	Name	Email
dc01118e-a77f-525e-8aab-620834dd3ffc	501235	John Doe	studentemail1@harvard.edu
38dba7f7-5b7e-4271-1c60-42c3b17a42cd	288832	Mary Jane	studentemail2@harvard.edu
11116a67-b251-4852-8b38-84f720c591e6	16116155	John Smith	studentemail3@harvard.edu

HARVARD
OFFICE OF THE VICE PROVOST FOR ADVANCES IN LEARNING
RESEARCH

R.A.P.I.D.
RESEARCH FRAMEWORK

- Requests, Analysis, Products, Interfaces and Docs
- Flexible Framework for R.A.P.I.D. Prototyping of Analytics Data Products
- Common Data Products supports multiple stakeholders
- Collaborative and iterative data product life cycle between researchers, data scientists and engineers



Lambda

Bootstrap

Extract and Verify

Code Generation

Phase 0

Transform and Load Pipeline

Code Generation

Phase 1
Identity Mapping

Phase 2 ... n
Transformations

Load

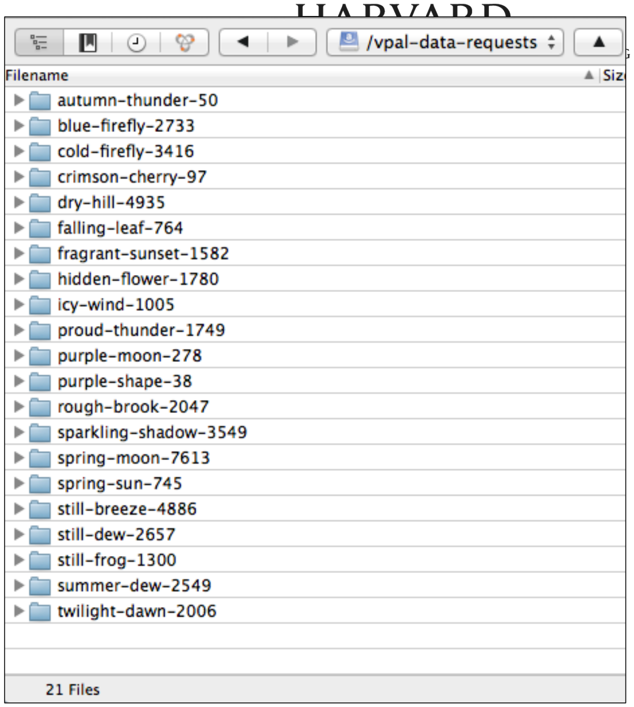
R.A.P.I.D. Research

Update
Data Products

Deliver to
Stakeholders

DELIVERY TO STAKEHOLDERS

- Secure delivery via Amazon S3
- Folder names are obscured/randomized to prevent identification of requestors and content
- Encrypted data using Stakeholders GPG public key

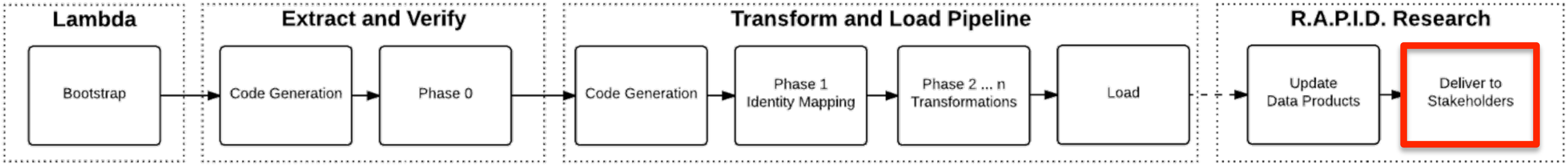


/vpal-data-requests

Filename | Siz

autumn-thunder-50
blue-firefly-2733
cold-firefly-3416
crimson-cherry-97
dry-hill-4935
falling-leaf-764
fragrant-sunset-1582
hidden-flower-1780
icy-wind-1005
proud-thunder-1749
purple-moon-278
purple-shape-38
rough-brook-2047
sparkling-shadow-3549
spring-moon-7613
spring-sun-745
still-breeze-4886
still-dew-2657
still-frog-1300
summer-dew-2549
twilight-dawn-2006

21 Files



Lambda

Bootstrap

Extract and Verify

Code Generation

Phase 0

Transform and Load Pipeline

Code Generation

Phase 1 Identity Mapping

Phase 2 ... n Transformations

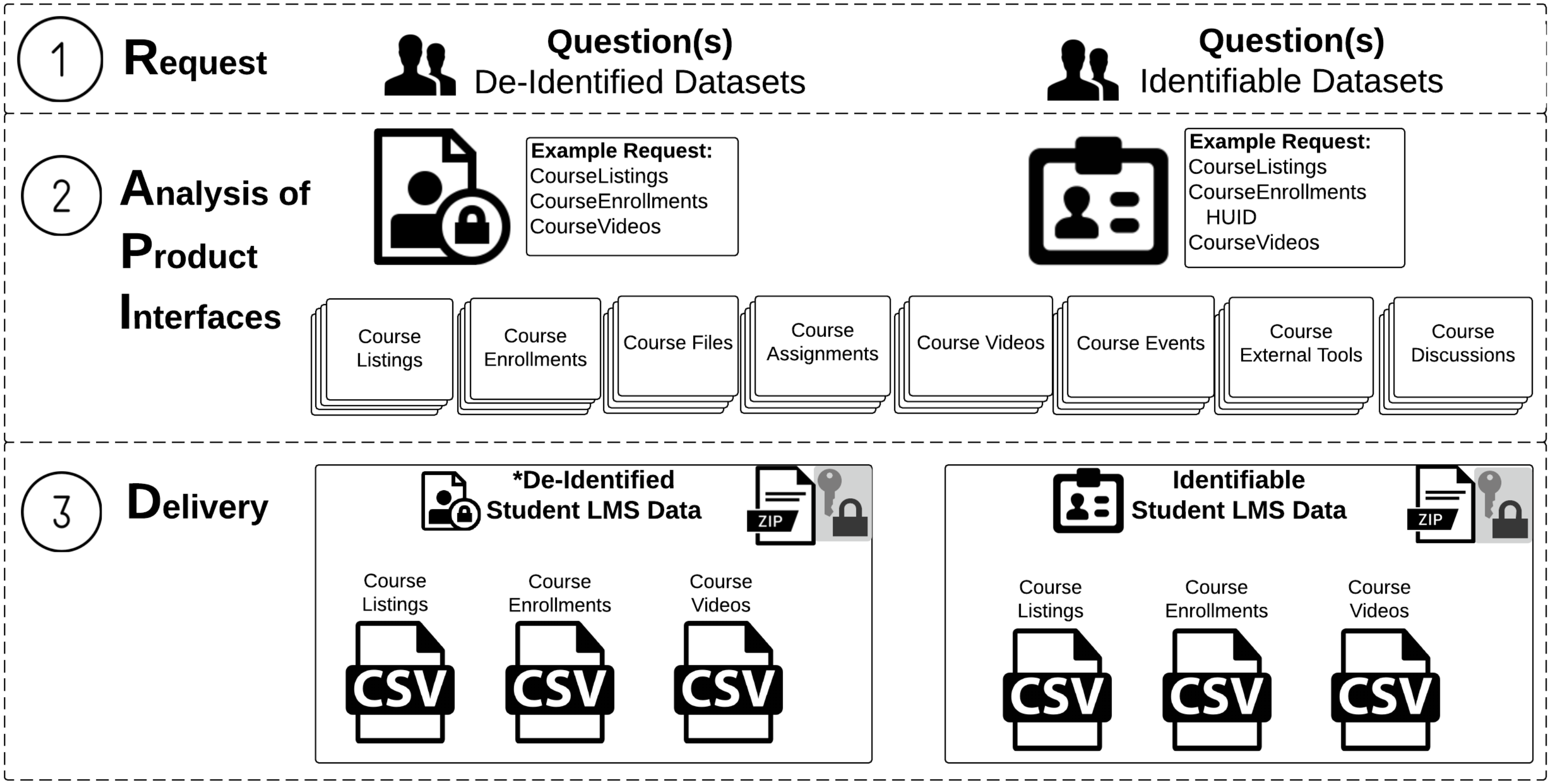
Load

R.A.P.I.D. Research

Update Data Products

Deliver to Stakeholders

EXAMPLE:
RAPID RESEARCH FRAMEWORK



1 **Request**

Question(s)
De-Identified Datasets

Question(s)
Identifiable Datasets

2 **Analysis of Product Interfaces**

Example Request:
CourseListings
CourseEnrollments
CourseVideos

Example Request:
CourseListings
CourseEnrollments
HUID
CourseVideos

- Course Listings
- Course Enrollments
- Course Files
- Course Assignments
- Course Videos
- Course Events
- Course External Tools
- Course Discussions

3 **Delivery**

***De-Identified Student LMS Data**

- Course Listings
- Course Enrollments
- Course Videos

Identifiable Student LMS Data

- Course Listings
- Course Enrollments
- Course Videos