INSTRUCTURECON 2019

BIG DATA IDEAS & USAGE FOR TINY BUDGETS & SMALL TEAMS

Learnin' Safari
INSTRUCTURECON 2019
July 9-11, Long Beach Convention Center

PRESENTATION BY:
Robert Carroll
robert.c@nv.ccsd.net
twitter@rbtcar

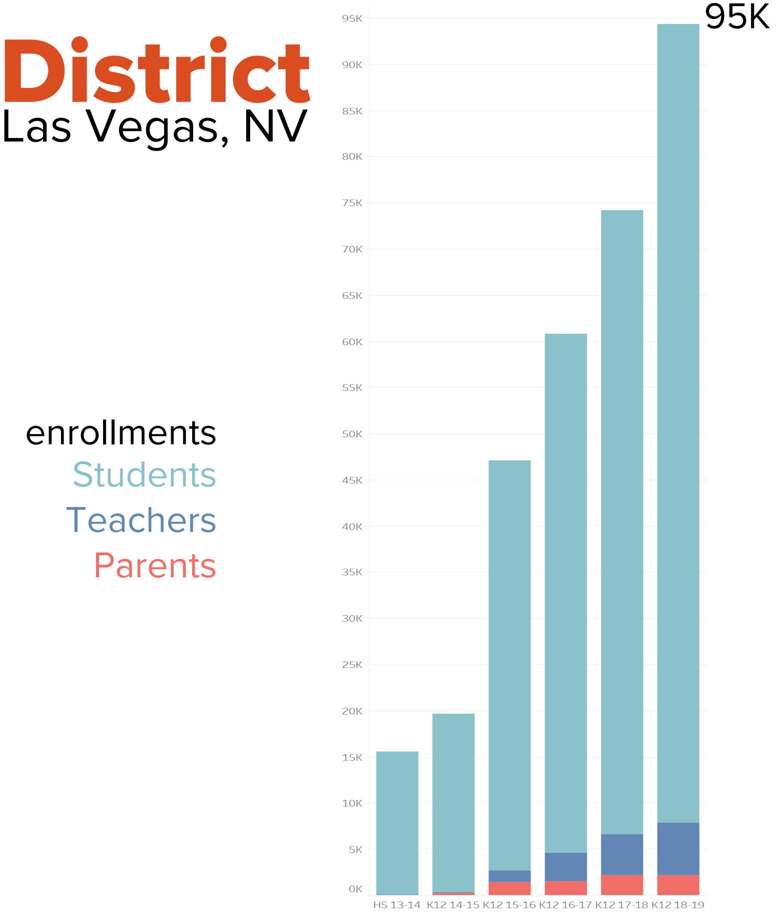
Clark County School District

Las Vegas, NV

Using Canvas since 2013

362 Schools (SY19-20)

320,000 Students



Clark County School District

NVLA Fall & Spring

Students	8,369
Observers	1,282
Teachers	843

NVLA Summer

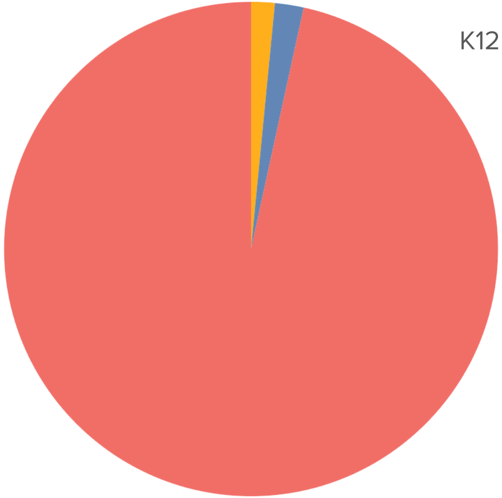
Students	10,436
Observers	1,128
Teachers	243

Blended Fall & Spring

Students	67,193
Teachers	4,029
Observers	714

Professional Development

Students	4,023
Teachers	260



DISTINCT USERS

Students	40,321
Employees	786
Parents	639

K12 fall and spring

*last_activity_at IS NOT NULL

a *disclaimer* about the "costs"

not *realized* by our department

...you have a small tech team
...you are the tech team
...you were 'given a Linux box'
...local infrastructure is sufficient

DB Managers, Server Admins, Security, Networking

sometimes I use my Mac Mini too

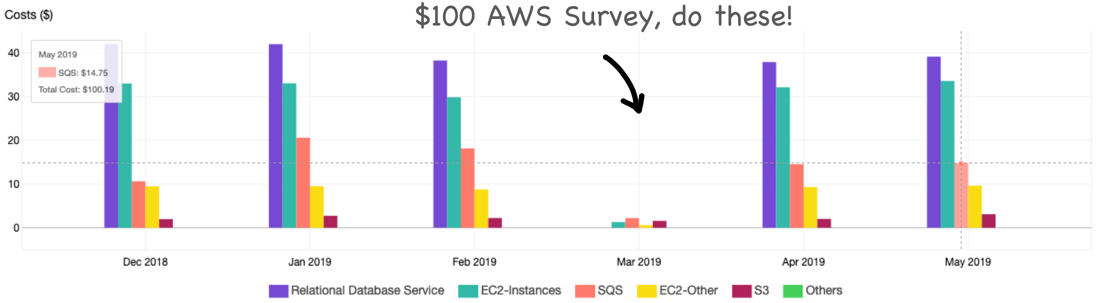
this is what we use for our canvas data pipeline

RHEL7 Virtual Machine (onsite)
6 CPUs, 12 Cores
16GB RAM
120GB HD

$475/month Azure estimate

Some solutions often charge by GB ingested
... not very Canvas Data friendly without deltas

$100 AWS Survey, do these!



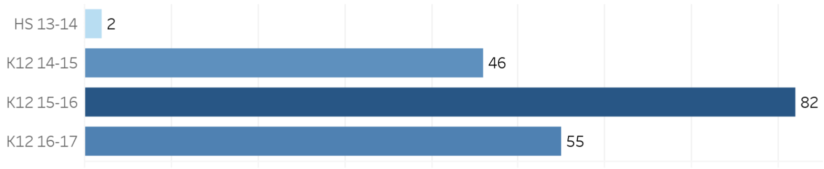
Clark County School District

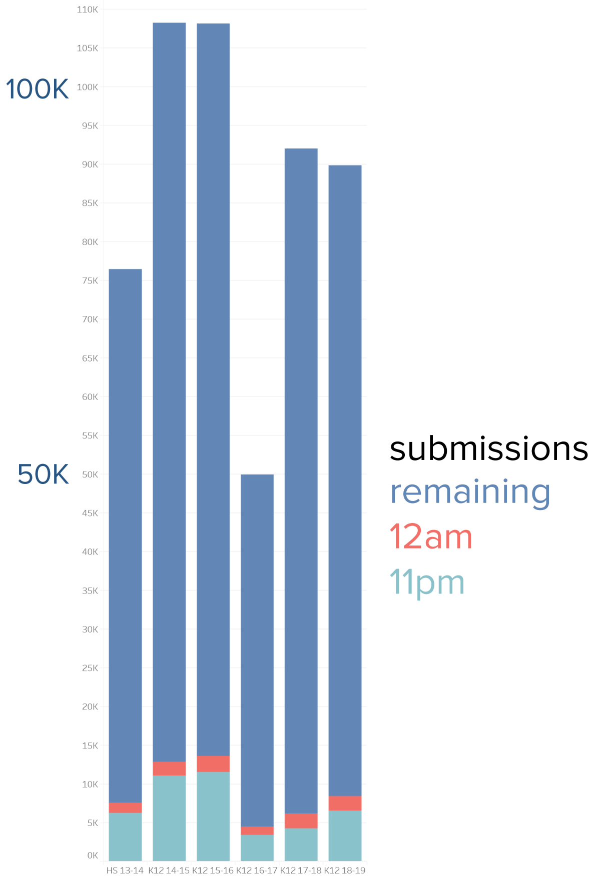
First few submission queries revealed a bad default

Most assignments had a Sunday 11:59:59 due date

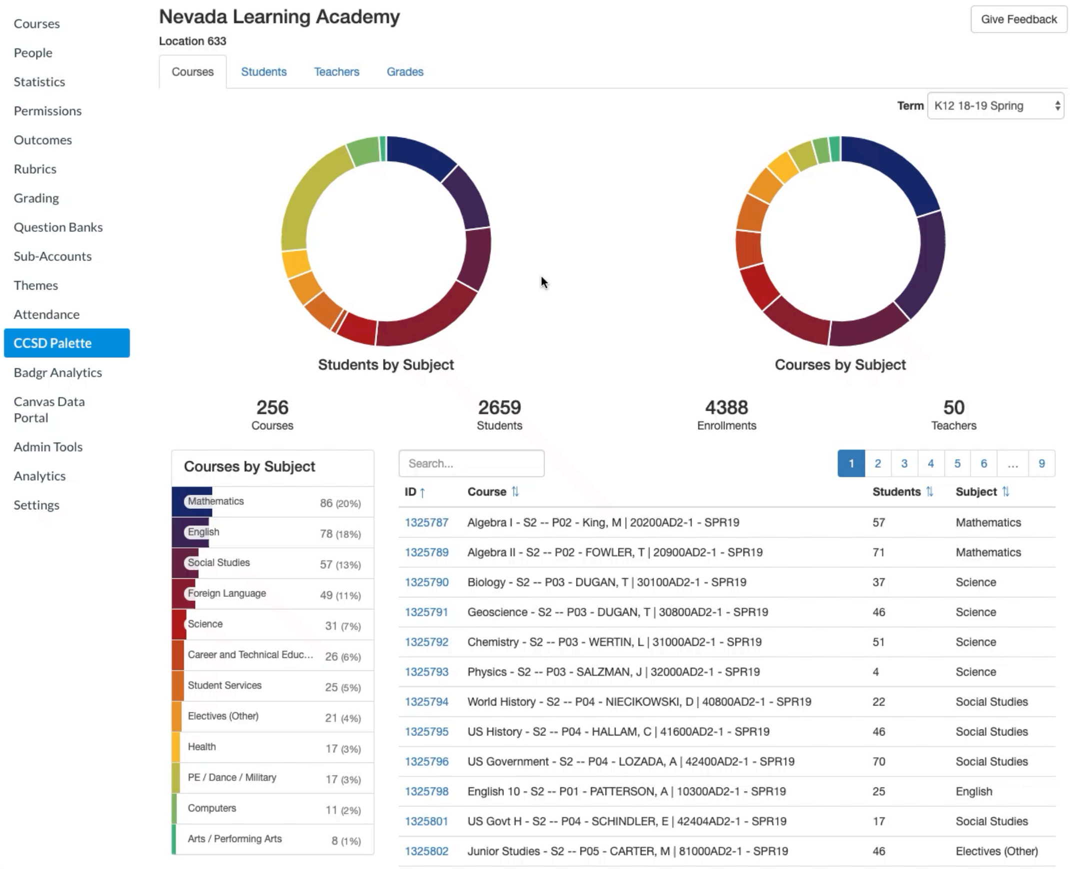
Students staying up until midnight

Assignments due Sunday at Midnight





Clark County School District



- Courses
- People
- Statistics
- Permissions
- Outcomes
- Rubrics
- Grading
- Question Banks
- Sub-Accounts
- Themes
- Attendance
- CCSD Palette
- Badgr Analytics
- Canvas Data Portal
- Admin Tools
- Analytics
- Settings

Nevada Learning Academy

Location 633

Give Feedback

Courses | Students | Teachers | Grades

Term: K12 18-19 Spring

256	2659	4388	50
Courses	Students	Enrollments	Teachers

Courses by Subject

Subject	Courses
Mathematics	86 (20%)
English	78 (18%)
Social Studies	57 (13%)
Foreign Language	49 (11%)
Science	31 (7%)
Career and Technical Educ...	26 (6%)
Student Services	25 (5%)
Electives (Other)	21 (4%)
Health	17 (3%)
PE / Dance / Military	17 (3%)
Computers	11 (2%)
Arts / Performing Arts	8 (1%)

Search...

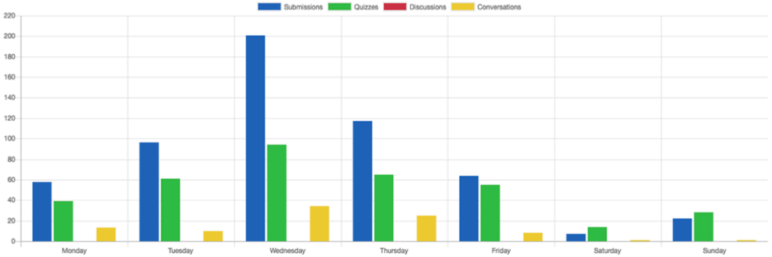
1 2 3 4 5 6 ... 9

ID ↑	Course ⇅	Students ⇅	Subject ⇅
1325787	Algebra I - S2 -- P02 - King, M | 20200AD2-1 - SPR19	57	Mathematics
1325789	Algebra II - S2 -- P02 - FOWLER, T | 20900AD2-1 - SPR19	71	Mathematics
1325790	Biology - S2 -- P03 - DUGAN, T | 30100AD2-1 - SPR19	37	Science
1325791	Geoscience - S2 -- P03 - DUGAN, T | 30800AD2-1 - SPR19	46	Science
1325792	Chemistry - S2 -- P03 - WERTIN, L | 31000AD2-1 - SPR19	51	Science
1325793	Physics - S2 -- P03 - SALZMAN, J | 32000AD2-1 - SPR19	4	Science
1325794	World History - S2 -- P04 - NIECIKOWSKI, D | 40800AD2-1 - SPR19	22	Social Studies
1325795	US History - S2 -- P04 - HALLAM, C | 41600AD2-1 - SPR19	46	Social Studies
1325796	US Government - S2 -- P04 - LOZADA, A | 42400AD2-1 - SPR19	70	Social Studies
1325798	English 10 - S2 -- P01 - PATTERSON, A | 10300AD2-1 - SPR19	25	English
1325801	US Govt H - S2 -- P04 - SCHINDLER, E | 42404AD2-1 - SPR19	17	Social Studies
1325802	Junior Studies - S2 -- P05 - CARTER, M | 81000AD2-1 - SPR19	46	Electives (Other)

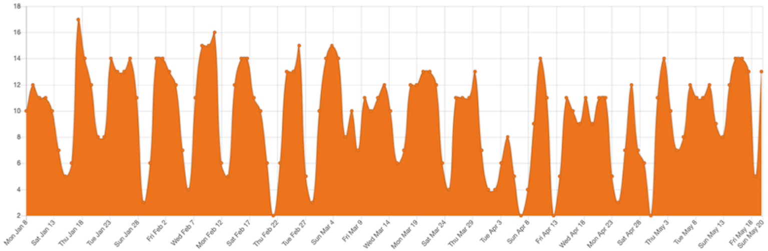
Nevada Learning Academy at CCSD

Full Time Students

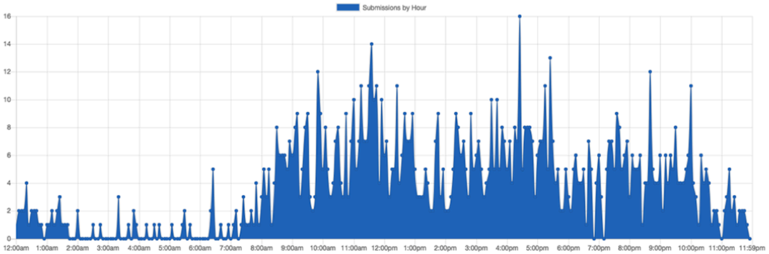
Activity



Attendance By Day

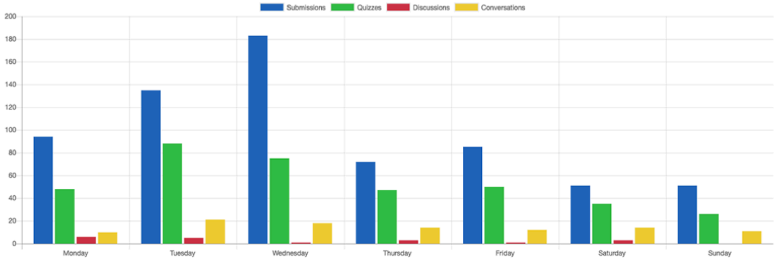


Submissions By Time of Day

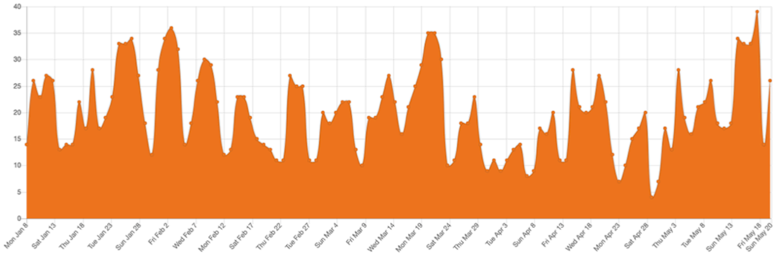


Part Time Students

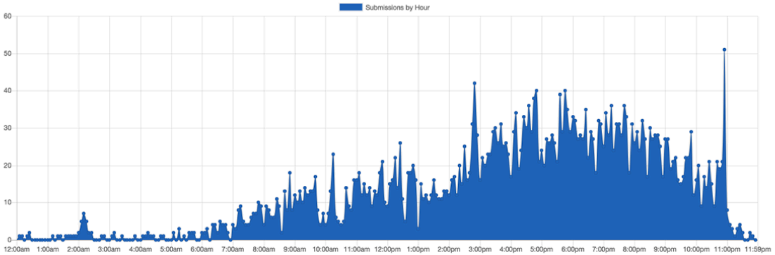
Activity

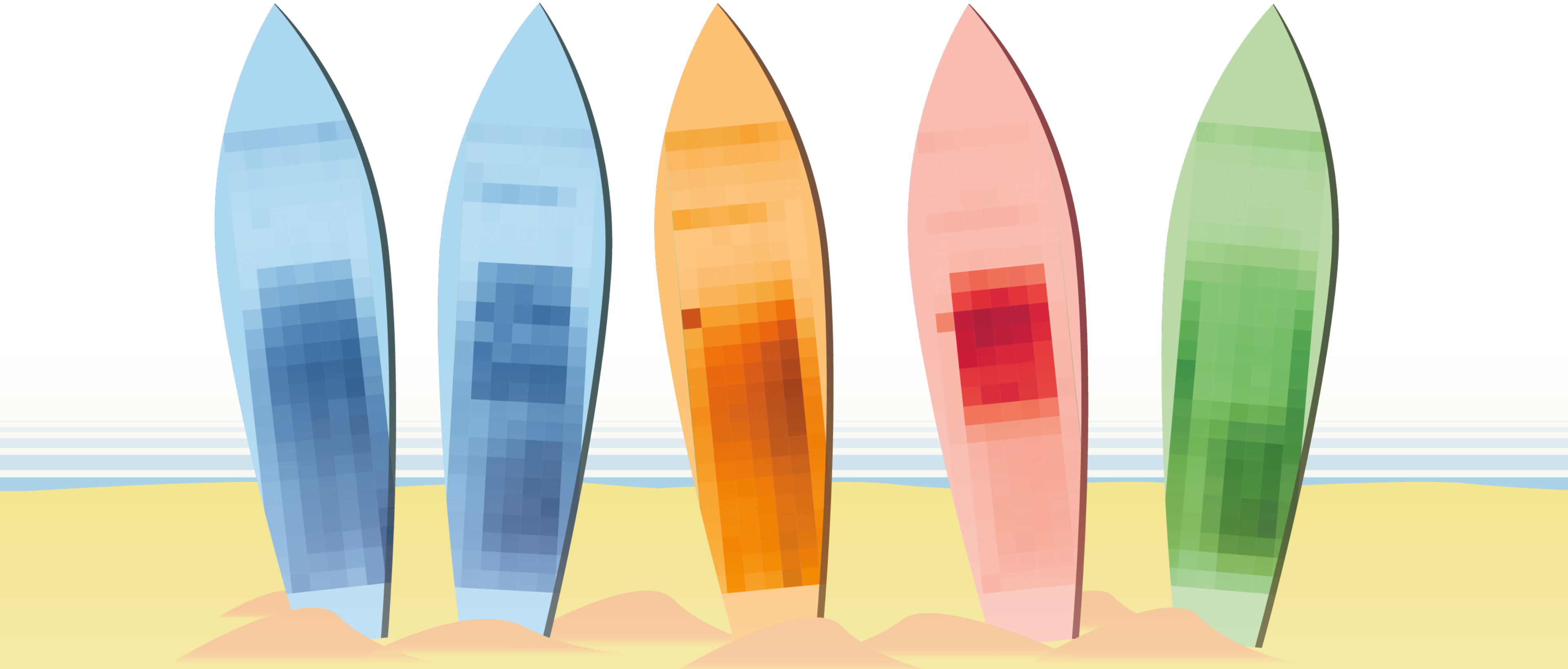


Attendance By Day



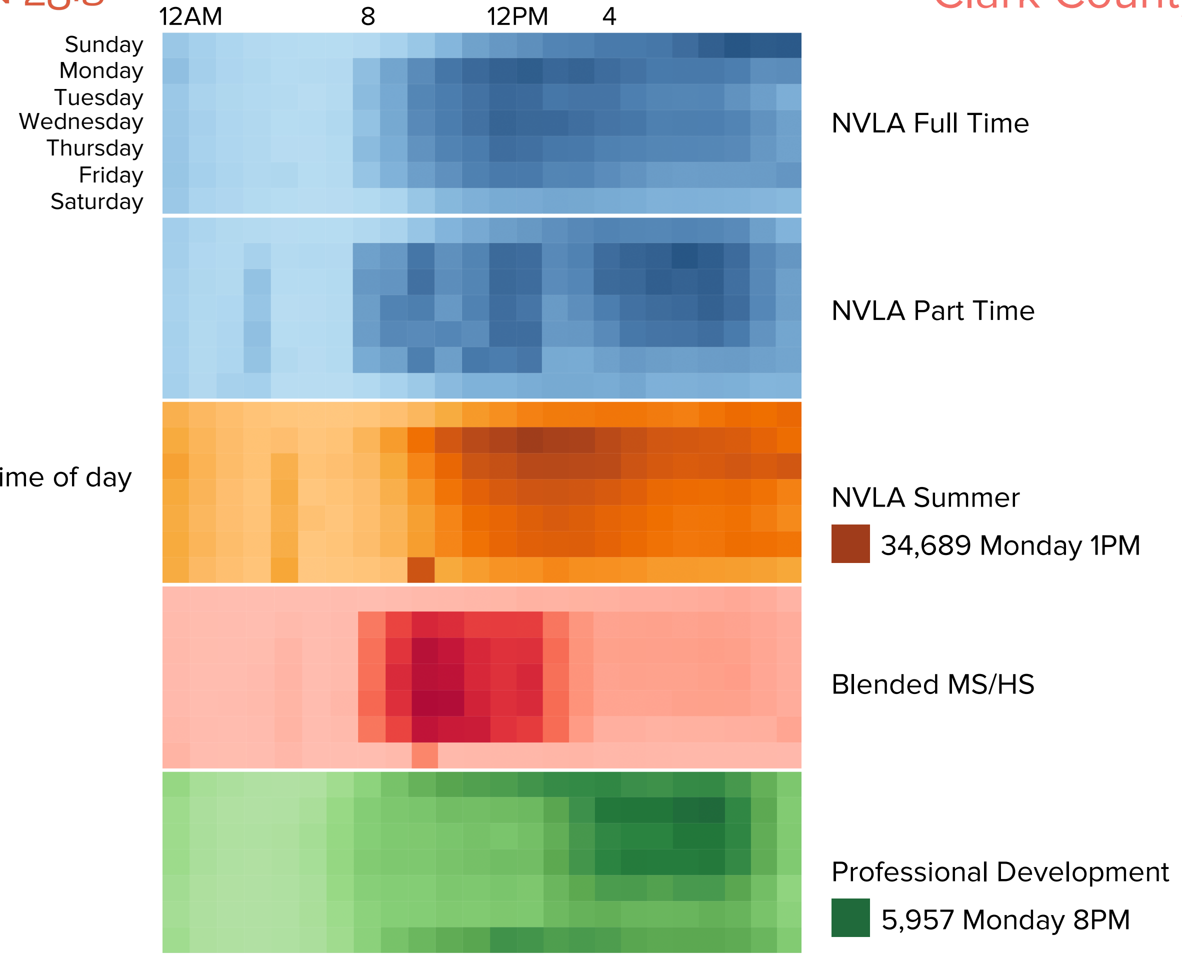
Submissions By Time of Day



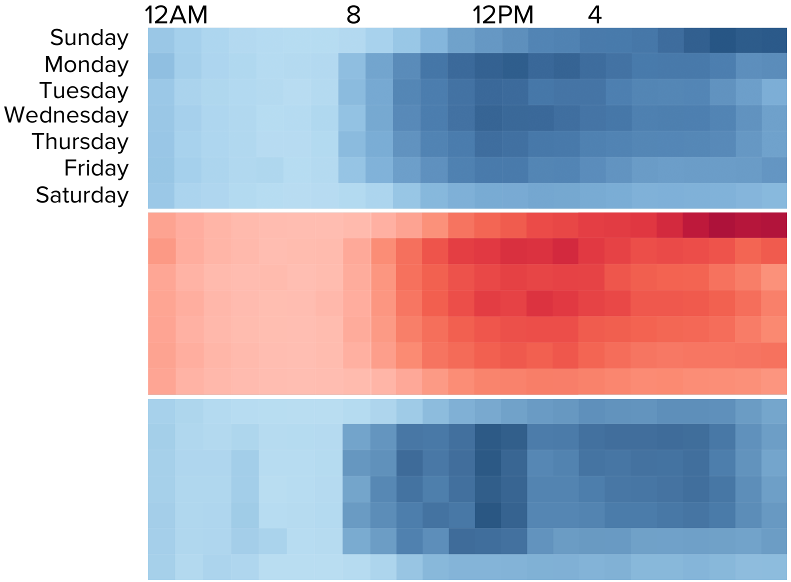


Clark County School District

submissions by time of day



effects of assignment due dates



NVLA All Submissions

🟦 27,101 Sunday 9PM

NVLA 13-14, 14-15, 15-16

🟥 22,676 Sunday 9PM

🟧 5,126 Monday 12AM ☹

NVLA 16-17, 17-18, 18-19

🟦 14,831 Thursday Noon

🟦 6,091 Midnight

submission_dim

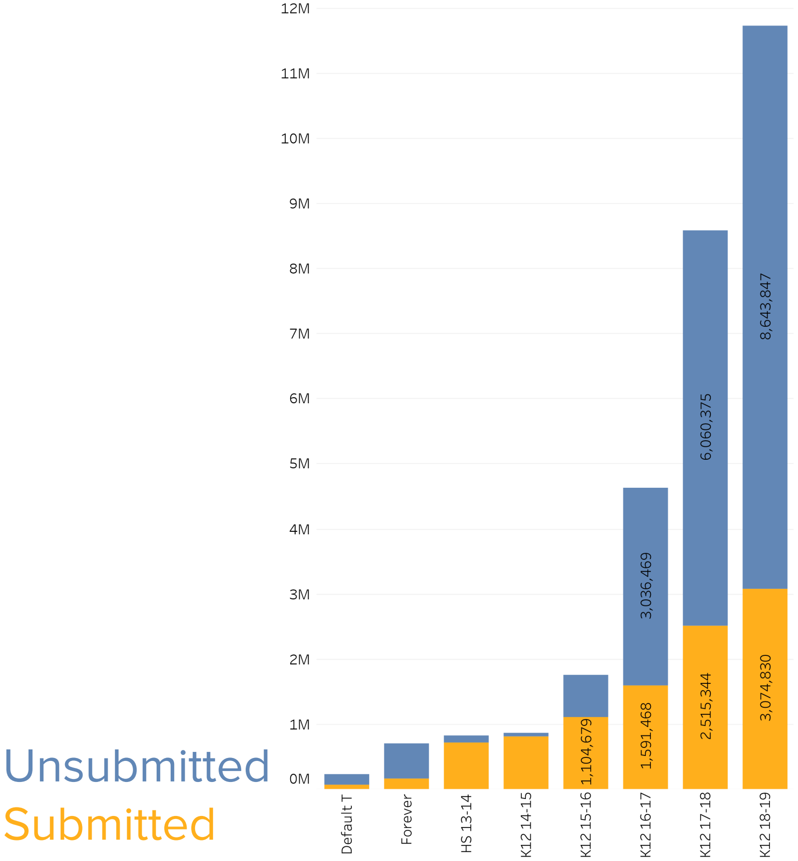
31 million rows

7.4GB unpacked submission_dim.txt

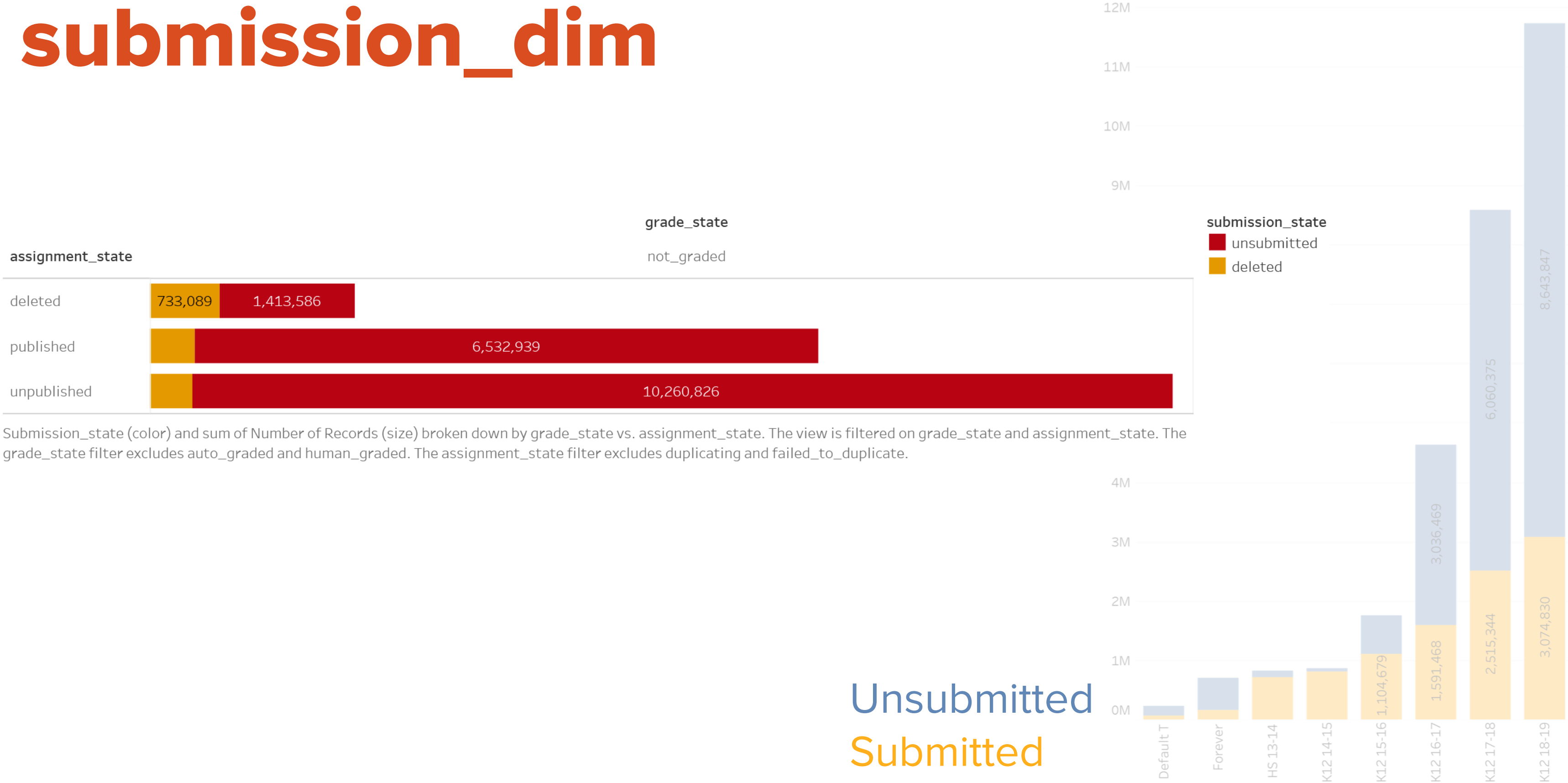
13.4GB in SQL Server

contains placeholders for submissions when students get assignment visibility

courses * assignments * students



submission_dim



Submission_state (color) and sum of Number of Records (size) broken down by grade_state vs. assignment_state. The view is filtered on grade_state and assignment_state. The grade_state filter excludes auto_graded and human_graded. The assignment_state filter excludes duplicating and failed_to_duplicate.

Unsubmitted

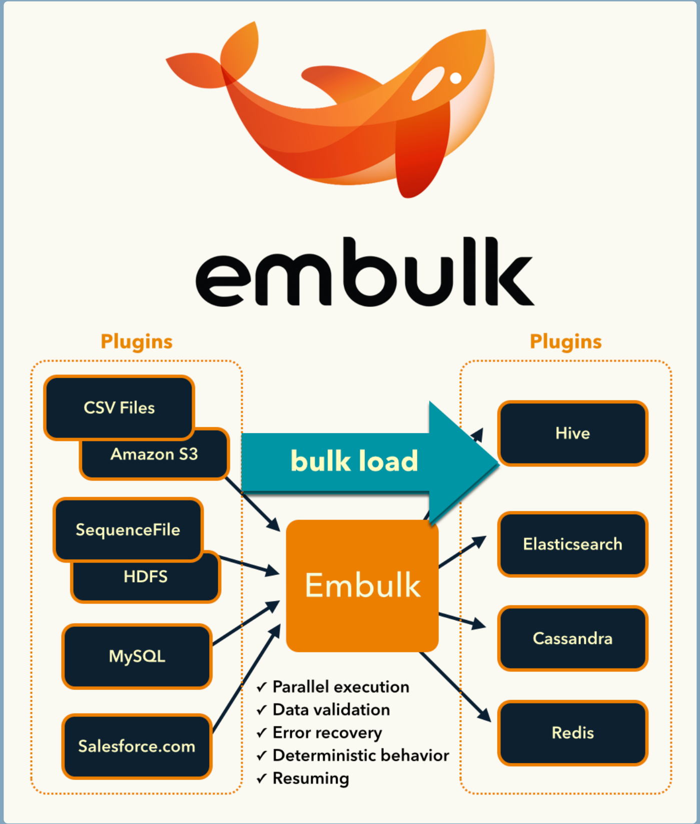
Submitted

Canvas Data **Sting Rays**

- Flat files from the beginning of time, no delta
- GB of downloads become TB of SQL
- >24 hours old when batch is published and installed
- Import time takes longer every day
- Changes to schema, tables, fields, new and deprecated
- Correcting TimeZones in SQL is annoying
- Generating a DDL from schema.json brings bad vibes
- Documentation is not complete… enumerables?

Managing Canvas Data **with Embulk**

- Bulk Data Loader for multiple storage types and storage destinations
- Parallel Tasks each CPU Core
- Intelligent guess for CSV input
- Plugin Based
    1. Input Plugin - CSV
    2. Filter Plugins
    3. Output Plugin - SQL
- YAML Configs
    - Define input/output settings
    - Define DDL



embulk

Plugins

CSV Files

Amazon S3

SequenceFile

HDFS

MySQL

Salesforce.com

bulk load

Embulk

✓ Parallel execution
✓ Data validation
✓ Error recovery
✓ Deterministic behavior
✓ Resuming

Plugins

Hive

Elasticsearch

Cassandra

Redis

embulk.org

github.com/embulk/embulk

Managing Canvas Data **with Embulk**

embulk

Input CSV

default, reads gzip

Filter Plugins

Column Filter
Add PKs
Remove deprecated fields

Row Filter
skip rows with SQL like syntax

UNIQUE and DISTINCT

JOIN files

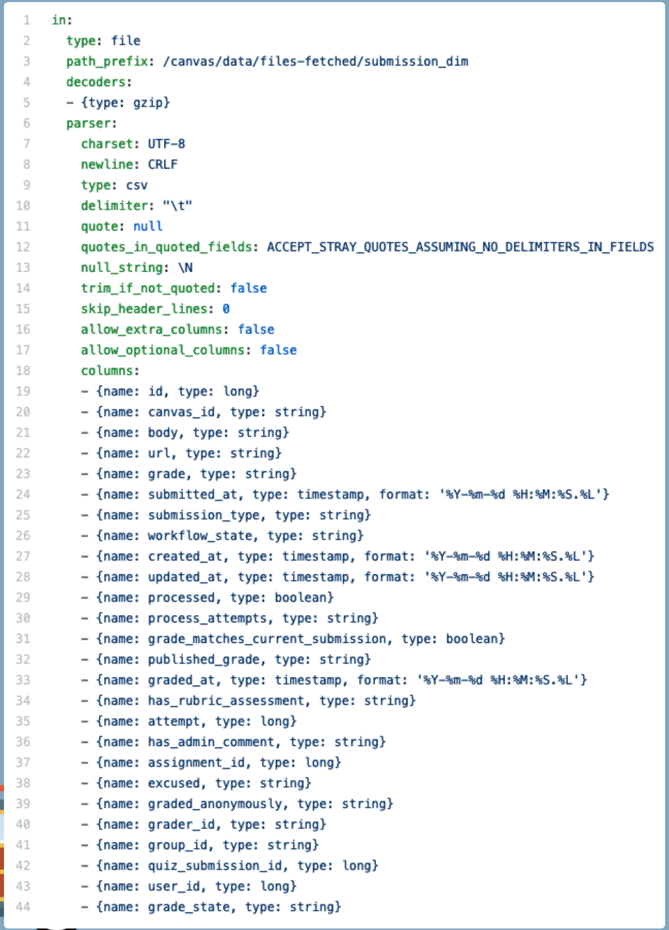
Output SQL

Uses JDBC and Native Drivers

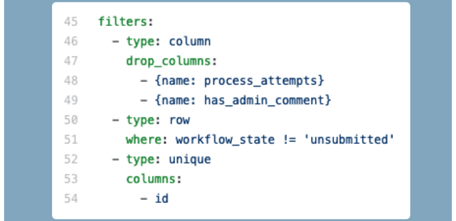
Modes
insert, insert_direct, truncate_insert, replace, merge

Overwrite TimeZones for each timestamp column

INSTRUCTURECON 2019

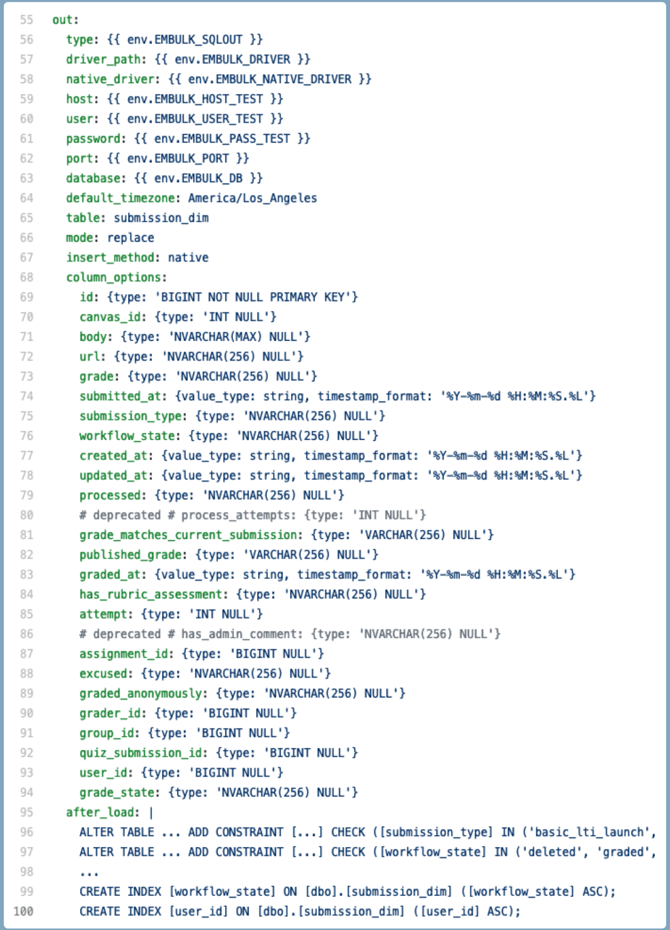


```yaml
in:
  type: file
  path_prefix: /canvas/data/files-fetched/submission_dim
  decoders:
  - {type: gzip}
  parser:
    charset: UTF-8
    newline: CRLF
    type: csv
    delimiter: "\t"
    quote: null
    quotes_in_quoted_fields: ACCEPT_STRAY_QUOTES_ASSUMING_NO_DELIMITERS_IN_FIELDS
    null_string: \N
    trim_if_not_quoted: false
    skip_header_lines: 0
    allow_extra_columns: false
    allow_optional_columns: false
    columns:
    - {name: id, type: long}
    - {name: canvas_id, type: string}
    - {name: body, type: string}
    - {name: url, type: string}
    - {name: grade, type: string}
    - {name: submitted_at, type: timestamp, format: '%Y-%m-%d %H:%M:%S.%L'}
    - {name: submission_type, type: string}
    - {name: workflow_state, type: string}
    - {name: created_at, type: timestamp, format: '%Y-%m-%d %H:%M:%S.%L'}
    - {name: updated_at, type: timestamp, format: '%Y-%m-%d %H:%M:%S.%L'}
    - {name: processed, type: boolean}
    - {name: process_attempts, type: string}
    - {name: grade_matches_current_submission, type: boolean}
    - {name: published_grade, type: string}
    - {name: graded_at, type: timestamp, format: '%Y-%m-%d %H:%M:%S.%L'}
    - {name: has_rubric_assessment, type: string}
    - {name: attempt, type: long}
    - {name: has_admin_comment, type: string}
    - {name: assignment_id, type: long}
    - {name: excused, type: string}
    - {name: graded_anonymously, type: string}
    - {name: grader_id, type: string}
    - {name: group_id, type: string}
    - {name: quiz_submission_id, type: long}
    - {name: user_id, type: long}
    - {name: grade_state, type: string}
```



```yaml
filters:
  - type: column
    drop_columns:
      - {name: process_attempts}
      - {name: has_admin_comment}
  - type: row
    where: workflow_state != 'unsubmitted'
  - type: unique
    columns:
      - id
```

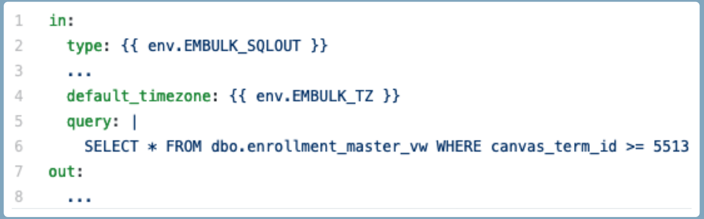
embulk



```yaml
out:
  type: {{ env.EMBULK_SQLOUT }}
  driver_path: {{ env.EMBULK_DRIVER }}
  native_driver: {{ env.EMBULK_NATIVE_DRIVER }}
  host: {{ env.EMBULK_HOST_TEST }}
  user: {{ env.EMBULK_USER_TEST }}
  password: {{ env.EMBULK_PASS_TEST }}
  port: {{ env.EMBULK_PORT }}
  database: {{ env.EMBULK_DB }}
  default_timezone: America/Los_Angeles
  table: submission_dim
  mode: replace
  insert_method: native
  column_options:
    id: {type: 'BIGINT NOT NULL PRIMARY KEY'}
    canvas_id: {type: 'INT NULL'}
    body: {type: 'NVARCHAR(MAX) NULL'}
    url: {type: 'NVARCHAR(256) NULL'}
    grade: {type: 'NVARCHAR(256) NULL'}
    submitted_at: {value_type: string, timestamp_format: '%Y-%m-%d %H:%M:%S.%L'}
    submission_type: {type: 'NVARCHAR(256) NULL'}
    workflow_state: {type: 'NVARCHAR(256) NULL'}
    created_at: {value_type: string, timestamp_format: '%Y-%m-%d %H:%M:%S.%L'}
    updated_at: {value_type: string, timestamp_format: '%Y-%m-%d %H:%M:%S.%L'}
    processed: {type: 'NVARCHAR(256) NULL'}
    # deprecated # process_attempts: {type: 'INT NULL'}
    grade_matches_current_submission: {type: 'VARCHAR(256) NULL'}
    published_grade: {type: 'VARCHAR(256) NULL'}
    graded_at: {value_type: string, timestamp_format: '%Y-%m-%d %H:%M:%S.%L'}
    has_rubric_assessment: {type: 'NVARCHAR(256) NULL'}
    attempt: {type: 'INT NULL'}
    # deprecated # has_admin_comment: {type: 'NVARCHAR(256) NULL'}
    assignment_id: {type: 'BIGINT NULL'}
    excused: {type: 'NVARCHAR(256) NULL'}
    graded_anonymously: {type: 'NVARCHAR(256) NULL'}
    grader_id: {type: 'BIGINT NULL'}
    group_id: {type: 'BIGINT NULL'}
    quiz_submission_id: {type: 'BIGINT NULL'}
    user_id: {type: 'BIGINT NULL'}
    grade_state: {type: 'NVARCHAR(256) NULL'}
  after_load: |
    ALTER TABLE ... ADD CONSTRAINT [...] CHECK ([submission_type] IN ('basic_lti_launch',
    ALTER TABLE ... ADD CONSTRAINT [...] CHECK ([workflow_state] IN ('deleted', 'graded',
    ...
    CREATE INDEX [workflow_state] ON [dbo].[submission_dim] ([workflow_state] ASC);
    CREATE INDEX [user_id] ON [dbo].[submission_dim] ([user_id] ASC);
```

Embulk **Hang Ten** with *SQL in SQL Out*

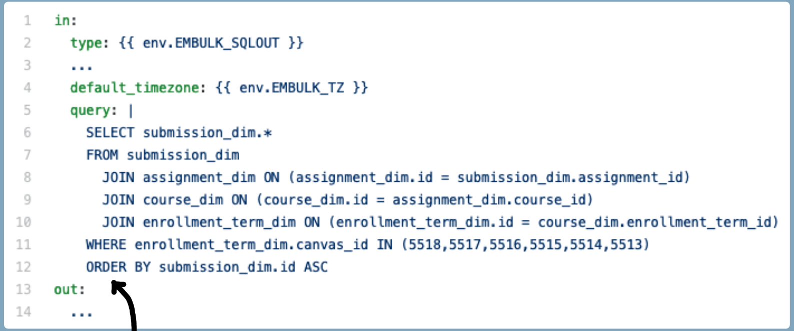
Materialized enrollment view to production



```
in:
  type: {{ env.EMBULK_SQLOUT }}
  ...
  default_timezone: {{ env.EMBULK_TZ }}
  query: |
    SELECT * FROM dbo.enrollment_master_vw WHERE canvas_term_id >= 5513
out:
  ...
```

JOINS are expensive
create new tables from JOINS and views

School Year Term submissions to production



```
in:
  type: {{ env.EMBULK_SQLOUT }}
  ...
  default_timezone: {{ env.EMBULK_TZ }}
  query: |
    SELECT submission_dim.*
    FROM submission_dim
      JOIN assignment_dim ON (assignment_dim.id = submission_dim.assignment_id)
      JOIN course_dim ON (course_dim.id = assignment_dim.course_id)
      JOIN enrollment_term_dim ON (enrollment_term_dim.id = course_dim.enrollment_term_id)
    WHERE enrollment_term_dim.canvas_id IN (5518,5517,5516,5515,5514,5513)
    ORDER BY submission_dim.id ASC
out:
  ...
```

final sort

Managing Canvas Data **with Embulk**

7.7GB downloaded

47GB unpacked

>1TB in SQL, with 1 day of requests

Imported for Staging

4 hours 44 minutes

*with indexes

Pro Tip

account_dim 1 file 8,600 rows
max_threads=12
output tasks 6 = input tasks 1 * 6

submission_dim 31 files 1M rows each
max_threads=12
tasks=31

Don’t unpack to single .txt
...more files more tasks

Don’t sort before import
...parallel tasks, imported out of order

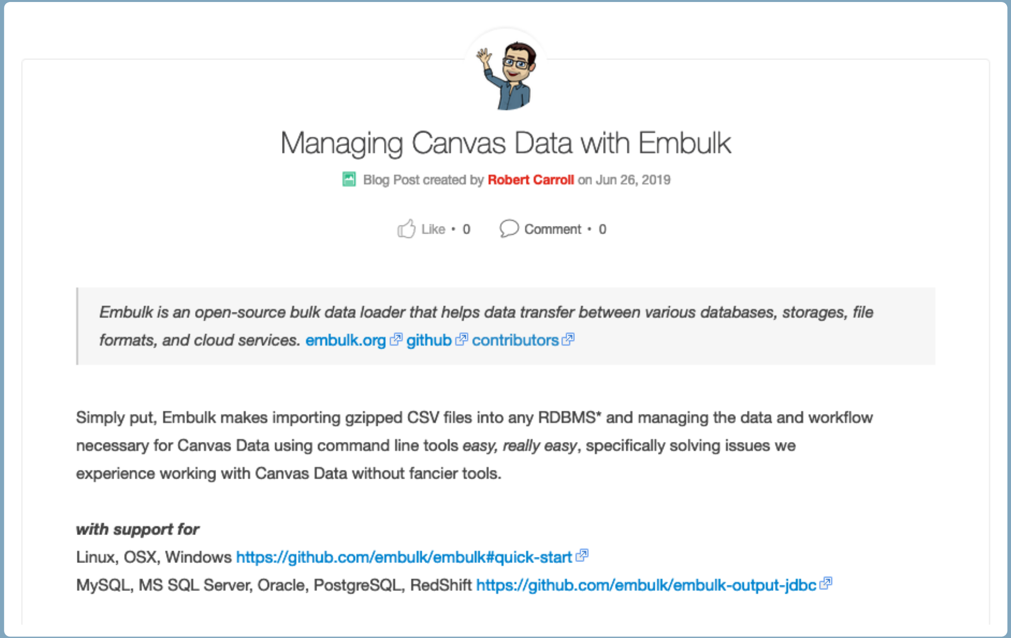
Embulk Canvas Community Compatibility

Linux
OSX
Windows
github.com/embulk/embulk#linux--mac--bsd

SQL Input & Output Plugins

MySQL
MSSQL
Oracle
PostgreSQL
RedShift
DB2

github.com/embulk/embulk-input-jdbc
github.com/embulk/embulk-output-jdbc



Managing Canvas Data with Embulk

Blog Post created by Robert Carroll on Jun 26, 2019

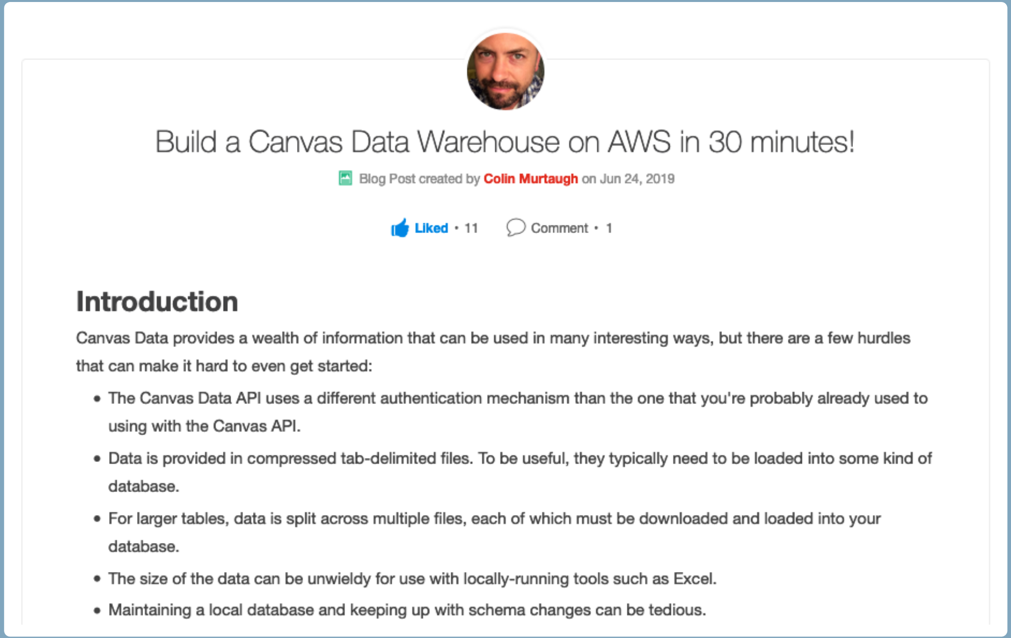
Like • 0 Comment • 0

> *Embulk is an open-source bulk data loader that helps data transfer between various databases, storages, file formats, and cloud services.* embulk.org github contributors

Simply put, Embulk makes importing gzipped CSV files into any RDBMS* and managing the data and workflow necessary for Canvas Data using command line tools *easy, really easy*, specifically solving issues we experience working with Canvas Data without fancier tools.

with support for
Linux, OSX, Windows https://github.com/embulk/embulk#quick-start
MySQL, MS SQL Server, Oracle, PostgreSQL, RedShift https://github.com/embulk/embulk-output-jdbc

github.com/ccsd/canvas-data-embulk-configs



Build a Canvas Data Warehouse on AWS in 30 minutes!

Blog Post created by **Colin Murtaugh** on Jun 24, 2019

Liked • 11 Comment • 1

Introduction

Canvas Data provides a wealth of information that can be used in many interesting ways, but there are a few hurdles that can make it hard to even get started:

- The Canvas Data API uses a different authentication mechanism than the one that you're probably already used to using with the Canvas API.
- Data is provided in compressed tab-delimited files. To be useful, they typically need to be loaded into some kind of database.
- For larger tables, data is split across multiple files, each of which must be downloaded and loaded into your database.
- The size of the data can be unwieldy for use with locally-running tools such as Excel.
- Maintaining a local database and keeping up with schema changes can be tedious.

...the total cost to run this data warehouse should be **under about $10/month**. There's also a cost associated with the queries that you run against this data, but typical queries will only cost pennies.

github.com/Harvard-University-iCommons/canvas-data-aws

...I want it all, I want it now...

***Queen**, 1988*

Consuming Live Events

You could batch them in hourly, nightly…
or in real time, with long polling on SQS queue…
wait for new events

In real time… it's a tsunami
hundreds to thousands of events per minute

Canvas Live Events Services
working on granular subscription to events by type, and format
limit what you need

AWS
Glue, Lambda,
S3, Redshift

Fluentd
Makers of Embulk
SQS in SQL out

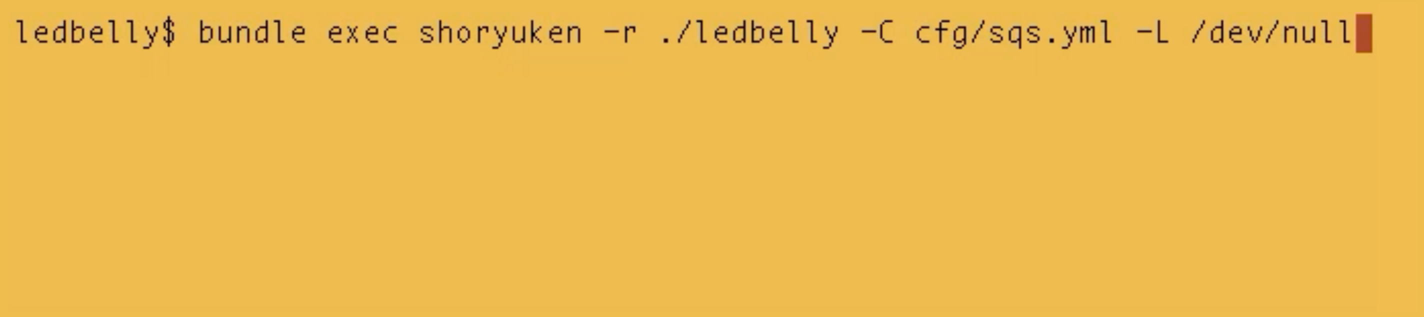
↰ 1000 Community Points

LEDbelly
Maker of this presentation
SQS to SQL
Canvas Raw and IMS Caliper

LEDbelly, Live Events Daemon

Lightweight middleware for consuming Canvas Live Events from SQS to SQL

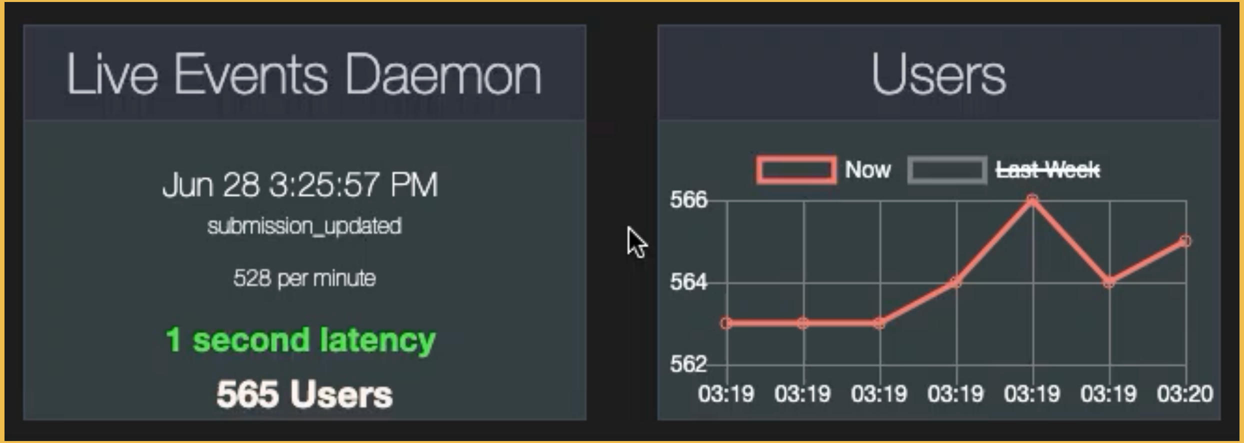
- **Shoryuken** provides fast multi-threaded message processing
- **Sequel** supports MSSQL, MySQL, PostgreSQL, and Oracle
- Canvas Raw and IMS Caliper formats, simultaneously, use both… why not



```
ledbelly$ bundle exec shoryuken -r ./ledbelly -C cfg/sqs.yml -L /dev/null
```

github.com/ccsd/ledbelly

LEDbelly, Live Events Daemon



Live Events Daemon

Jun 28 3:25:57 PM

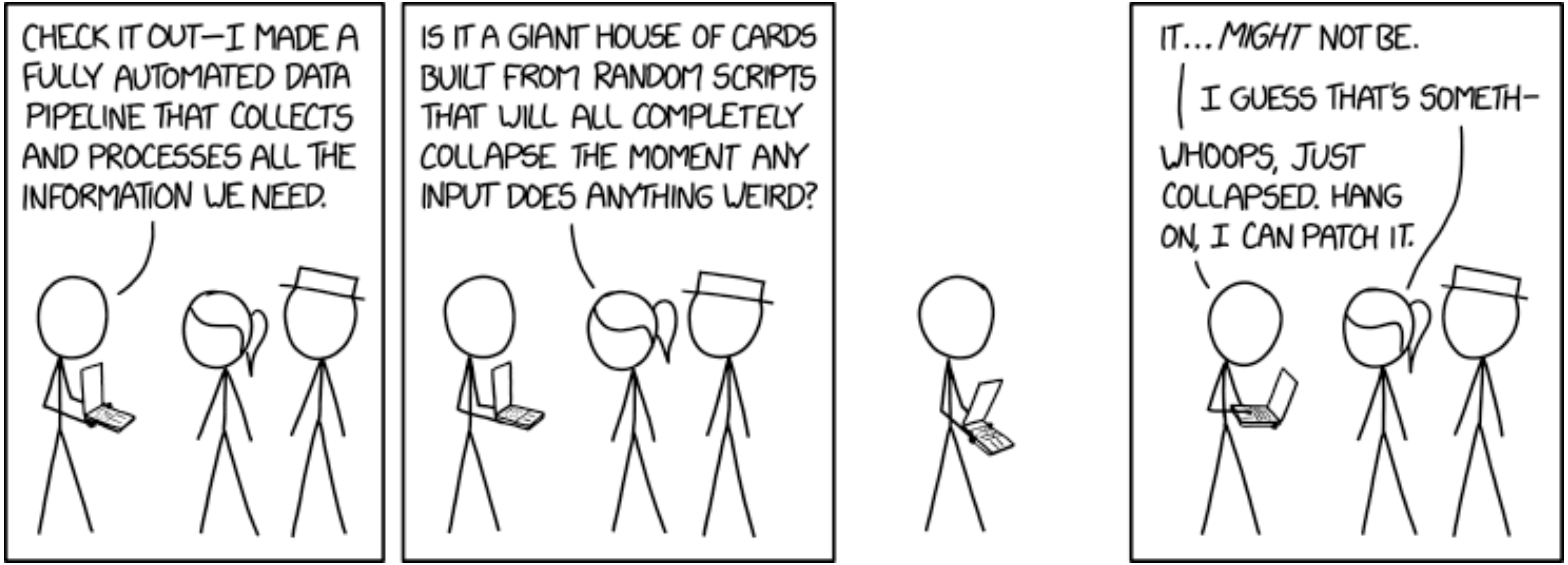
submission_updated

528 per minute

1 second latency

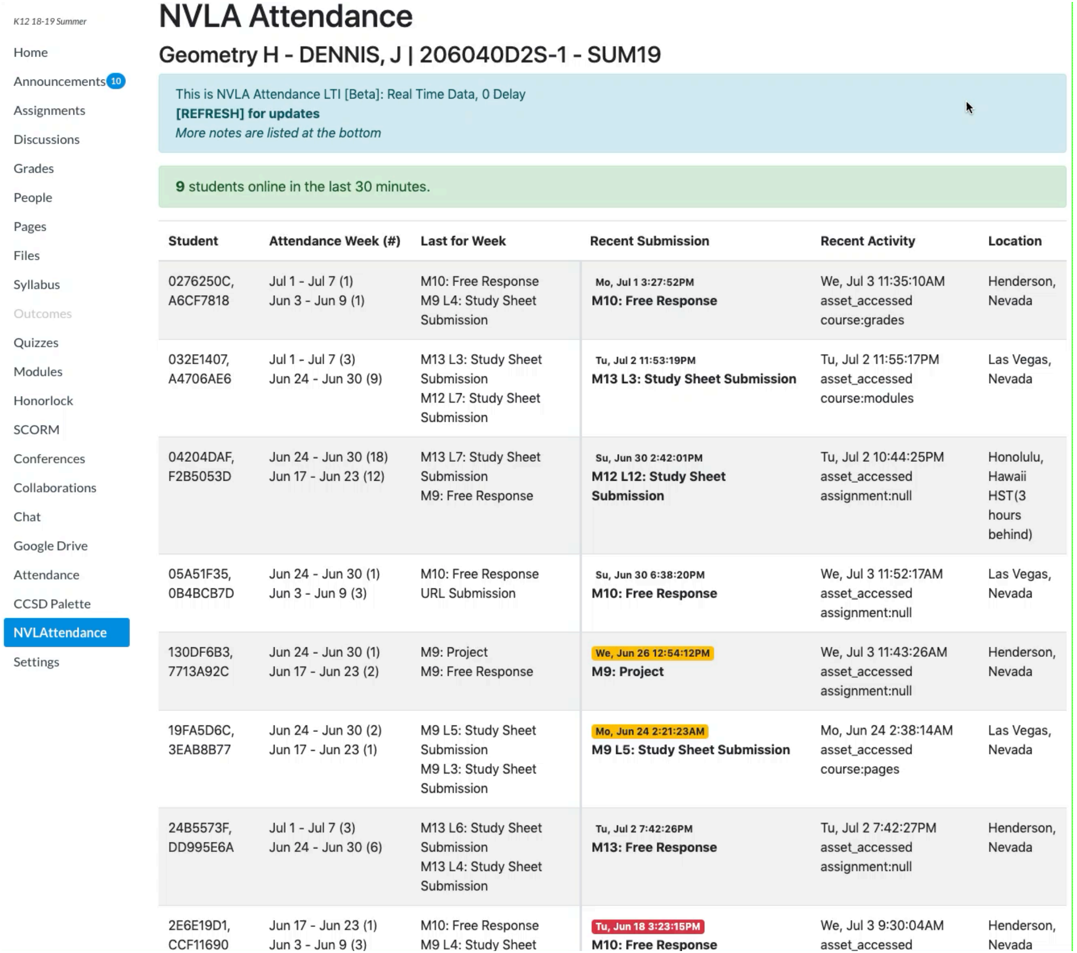
565 Users

Users


CHECK IT OUT—I MADE A FULLY AUTOMATED DATA PIPELINE THAT COLLECTS AND PROCESSES ALL THE INFORMATION WE NEED.
IS IT A GIANT HOUSE OF CARDS BUILT FROM RANDOM SCRIPTS THAT WILL ALL COMPLETELY COLLAPSE THE MOMENT ANY INPUT DOES ANYTHING WEIRD?
IT... *MIGHT* NOT BE.
I GUESS THAT'S SOMETH–
WHOOPS, JUST COLLAPSED. HANG ON, I CAN PATCH IT.

https://xkcd.com/2054/

alt
Is the pipeline literally running from your laptop?
Don't be silly, my laptop disconnects far too often to host a service we rely on.
It's running on my phone.



NVLA Attendance

Geometry H - DENNIS, J | 206040D2S-1 - SUM19

This is NVLA Attendance LTI [Beta]: Real Time Data, 0 Delay
[REFRESH] for updates
More notes are listed at the bottom

9 students online in the last 30 minutes.

Student	Attendance Week (#)	Last for Week	Recent Submission	Recent Activity	Location
0276250C, A6CF7818	Jul 1 - Jul 7 (1) Jun 3 - Jun 9 (1)	M10: Free Response M9 L4: Study Sheet Submission	Mo, Jul 1 3:27:52PM **M10: Free Response**	We, Jul 3 11:35:10AM asset_accessed course:grades	Henderson, Nevada
032E1407, A4706AE6	Jul 1 - Jul 7 (3) Jun 24 - Jun 30 (9)	M13 L3: Study Sheet Submission M12 L7: Study Sheet Submission	Tu, Jul 2 11:53:19PM **M13 L3: Study Sheet Submission**	Tu, Jul 2 11:55:17PM asset_accessed course:modules	Las Vegas, Nevada
04204DAF, F2B5053D	Jun 24 - Jun 30 (18) Jun 17 - Jun 23 (12)	M13 L7: Study Sheet Submission M9: Free Response	Su, Jun 30 2:42:01PM **M12 L12: Study Sheet Submission**	Tu, Jul 2 10:44:25PM asset_accessed assignment:null	Honolulu, Hawaii HST(3 hours behind)
05A51F35, 0B4BCB7D	Jun 24 - Jun 30 (1) Jun 3 - Jun 9 (3)	M10: Free Response URL Submission	Su, Jun 30 6:38:20PM **M10: Free Response**	We, Jul 3 11:52:17AM asset_accessed assignment:null	Las Vegas, Nevada
130DF6B3, 7713A92C	Jun 24 - Jun 30 (1) Jun 17 - Jun 23 (2)	M9: Project M9: Free Response	We, Jun 26 12:54:12PM **M9: Project**	We, Jul 3 11:43:26AM asset_accessed assignment:null	Henderson, Nevada
19FA5D6C, 3EAB8B77	Jun 24 - Jun 30 (2) Jun 17 - Jun 23 (1)	M9 L5: Study Sheet Submission M9 L3: Study Sheet Submission	Mo, Jun 24 2:21:23AM **M9 L5: Study Sheet Submission**	Mo, Jun 24 2:38:14AM asset_accessed course:pages	Las Vegas, Nevada
24B5573F, DD995E6A	Jul 1 - Jul 7 (3) Jun 24 - Jun 30 (6)	M13 L6: Study Sheet Submission M13 L4: Study Sheet Submission	Tu, Jul 2 7:42:26PM **M13: Free Response**	Tu, Jul 2 7:42:27PM asset_accessed assignment:null	Henderson, Nevada
2E6E19D1, CCF11690	Jun 17 - Jun 23 (1) Jun 3 - Jun 9 (3)	M10: Free Response M9 L4: Study Sheet	Tu, Jun 18 3:23:15PM **M10: Free Response**	We, Jul 3 9:30:04AM asset_accessed	Henderson, Nevada

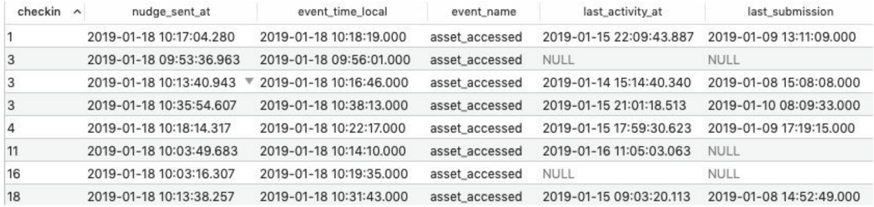
NVLA Nudges

12,696 sent*

Students	8,369
Enrollments	4,388

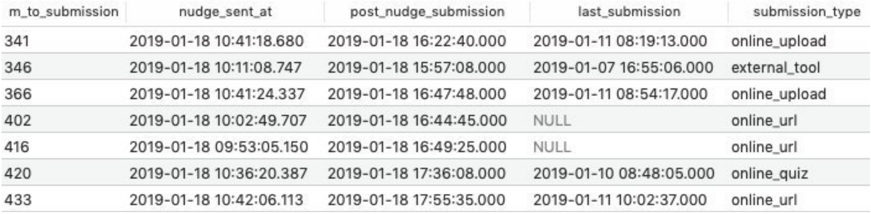
FT and PT students

Checking in

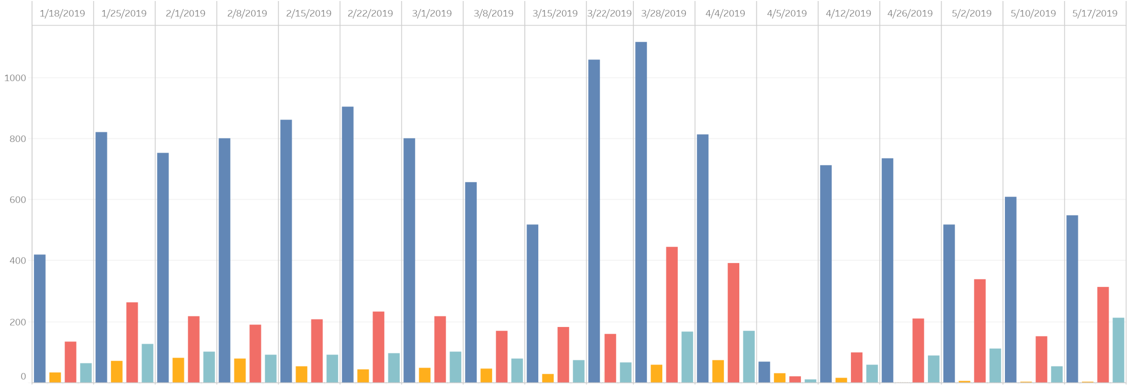


checkin	nudge_sent_at	event_time_local	event_name	last_activity_at	last_submission
1	2019-01-18 10:17:04.280	2019-01-18 10:18:19.000	asset_accessed	2019-01-15 22:09:43.887	2019-01-09 13:11:09.000
3	2019-01-18 09:53:36.963	2019-01-18 09:56:01.000	asset_accessed	NULL	NULL
3	2019-01-18 10:13:40.943	2019-01-18 10:16:46.000	asset_accessed	2019-01-14 15:14:40.340	2019-01-08 15:08:08.000
3	2019-01-18 10:35:54.607	2019-01-18 10:38:13.000	asset_accessed	2019-01-15 21:01:18.513	2019-01-10 08:09:33.000
4	2019-01-18 10:18:14.317	2019-01-18 10:22:17.000	asset_accessed	2019-01-15 17:59:30.623	2019-01-09 17:19:15.000
11	2019-01-18 10:03:49.683	2019-01-18 10:14:10.000	asset_accessed	2019-01-16 11:05:03.063	NULL
16	2019-01-18 10:03:16.307	2019-01-18 10:19:35.000	asset_accessed	NULL	NULL
18	2019-01-18 10:13:38.257	2019-01-18 10:31:43.000	asset_accessed	2019-01-15 09:03:20.113	2019-01-08 14:52:49.000

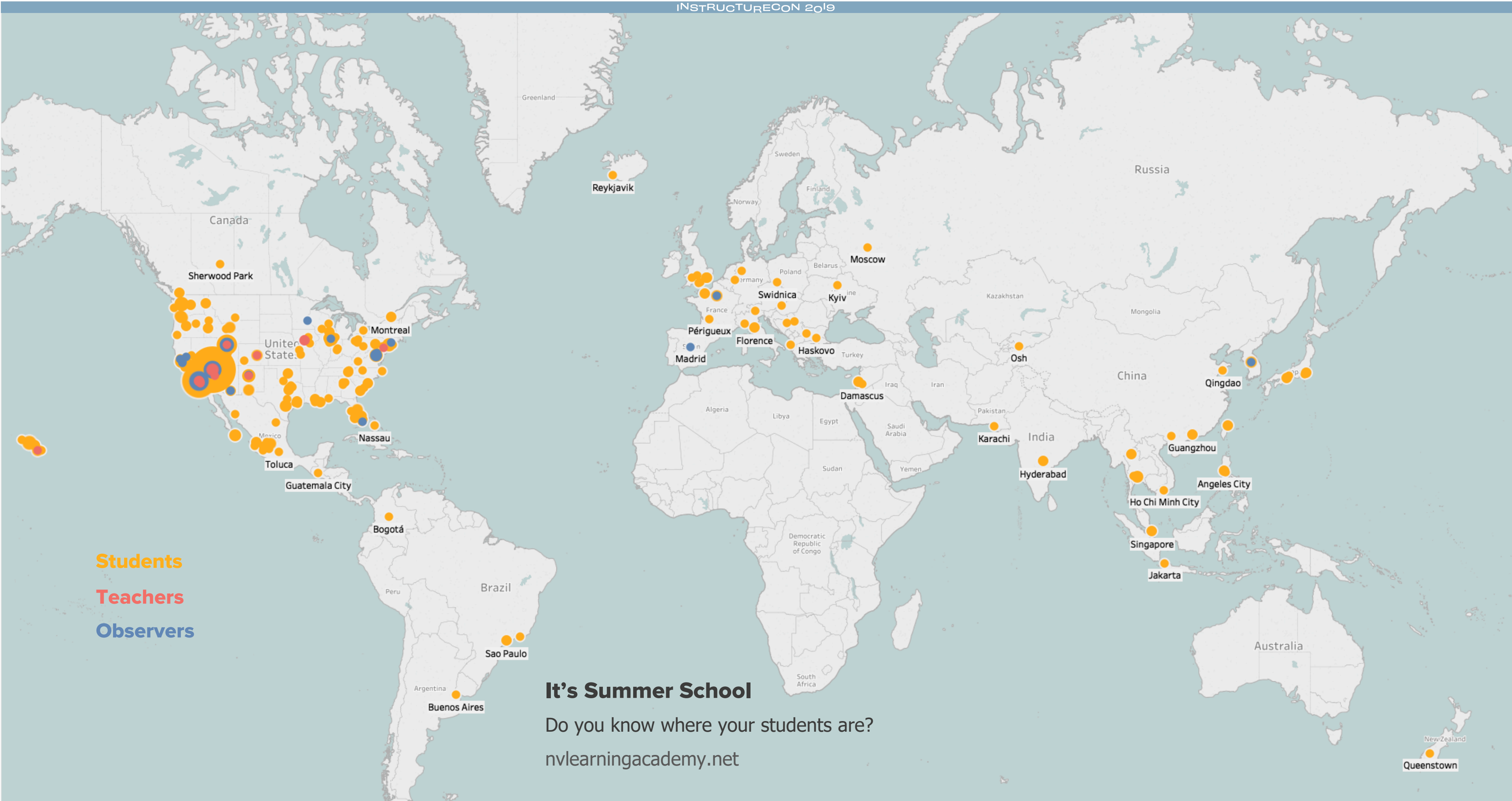
Submission



m_to_submission	nudge_sent_at	post_nudge_submission	last_submission	submission_type
341	2019-01-18 10:41:18.680	2019-01-18 16:22:40.000	2019-01-11 08:19:13.000	online_upload
346	2019-01-18 10:11:08.747	2019-01-18 15:57:08.000	2019-01-07 16:55:06.000	external_tool
366	2019-01-18 10:41:24.337	2019-01-18 16:47:48.000	2019-01-11 08:54:17.000	online_upload
402	2019-01-18 10:02:49.707	2019-01-18 16:44:45.000	NULL	online_url
416	2019-01-18 09:53:05.150	2019-01-18 16:49:25.000	NULL	online_url
420	2019-01-18 10:36:20.387	2019-01-18 17:36:08.000	2019-01-10 08:48:05.000	online_quiz
433	2019-01-18 10:42:06.113	2019-01-18 17:55:35.000	2019-01-11 10:02:37.000	online_url



704	Responses
3,932	EOW Submissions
1,752	Met Attendance



It's Summer School

Do you know where your students are?

nvlearningacademy.net

INSTRUCTURECON 2019

PRESENTED BY:

Robert Carroll
robert.c@nv.ccsd.net
twitter@rbtcar

Neal Shebeck
neal@nv.ccsd.net
twitter@CanvasNinja

Palette LTI
Aaron Leonard

#ElvisPanda
CCSD Canvas Team

Learnin' Safari
INSTRUCTURECON 2019
July 9-11, Long Beach Convention Center

obl.ccsd.net
twitter@CCSDBlendED

nvlearningacademy.net
twitter@ NVLearningAcad

Heatboards on a Beach
Stingray Shuffle Remix
Elvis Panda

Zac Clark

THANK YOU

THANK YOU

LONG BEACH InstructureCon CALIFORNIA

THANK YOU

7. 9-11 Learning Safari 2019

THANK YOU

THANK YOU