Getting Data Into Splunk

The What and The How

Ian Whitney

ASR Data Engineering

he/his

https://z.umn.edu/splunk_data_presentation

Splunk

What Is It

software for searching, monitoring, and analyzing machine-generated data

Splunk

What Do We Use It For

- Ensuring our applications are healthy
- Alerts
- Dashboards
- Research when things go awry

Splunk

Today's Topics

- Generation of data
- Ingestion of data

Another Time

- Working with data in Splunk

Generation

Options

- Application logs
- Instrumentation
- API responses
- Database queries
- Server metrics

Generation

General Advice

More structured = better
More identifiers = better

https://z.umn.edu/splunk_data_presentation

Generation

Application Logs

Generation

Application Logs

- A sensible place to start
- You probably already have these
- Write them in JSON if you can

Generation

Application Logs

```
Mar 05 08:23:01 asr-kafka-qat-04.oit.umn.edu docker-compose[130842]: distributed-connect    | \
[2024-03-05 14:23:01,561] INFO 192.168.96.1 - - [05/Mar/2024:14:23:01 +0000] \
"GET /connectors/non_credit_registration.sub_offerings/status HTTP/1.1" 200 229 "-" "Ruby" 0 \
(org.apache.kafka.connect.runtime.rest.RestServer)
```

Generation

Instrumentation

Generation

Instrumentation

- Offered by many application frameworks
- Write code to to observe code

Generation

Instrumentation

```
ActiveSupport::Notifications.instrument "your_application.magic" do |instrumentation|
  # magic happens
end
```

Generation

Instrumentation

```
{
    "severity": "INFO",
    "timestamp": 1707750727997393,
    "program": "your_application",
    "environment": "production",
    "version": "0.0.40",
    "entry": {
        "string": null,
        "instrumentation": {
            "name": "your_application.magic",
            "started": 1707750727992772,
            "finished": 1707750727997335,
            "elapsed": 4563
        }
    }
}
```

https://z.umn.edu/splunk_data_presentation

Generation

API Responses

Generation

API Responses

- Health checks, status, etc
- These usually return JSON

Generation

API Responses

```
curl -H "Content-Type: application/json" \
-X GET \
"http://127.0.0.1:8088/healthcheck"
```

Generation

API Responses

```
{
    "isHealthy": true,
    "details": {
        "metastore": {
            "isHealthy": true
        },
        "kafka": {
            "isHealthy": true
        },
        "commandRunner": {
            "isHealthy": true
        }
    },
    "serverState": "READY"
}
```

https://z.umn.edu/splunk_data_presentation

Generation

Database Queries

Generation

Database Queries

- Databases contain data and metadata that you can track
- SQL queries can return JSON

Generation

Database Queries

```
SELECT /*json*/
    max(updated_at) as most_recent_update
FROM
    important_application_table
;
```

Generation

Database Queries

```
{
    "most_recent_update": 1707750727992772
}
```

https://z.umn.edu/splunk_data_presentation

Generation

Server Metrics

Generation

Server Metrics

- Ask the server how it's doing
- A lot of data from hosts is already in Splunk
  - `top`
  - `cpu`
  - `vmstat`
  - `syslog`

Ingestion

Ingestion

Options

- Have the Splunk team do it
- Use methods that already exist, `syslog`
- Send it yourself, `HEC`
- Other (secretly also `HEC`)
- Other Other (OpenTelemetry, JMX, etc.)

Ingestion

Have the Splunk team do it

Ingestion

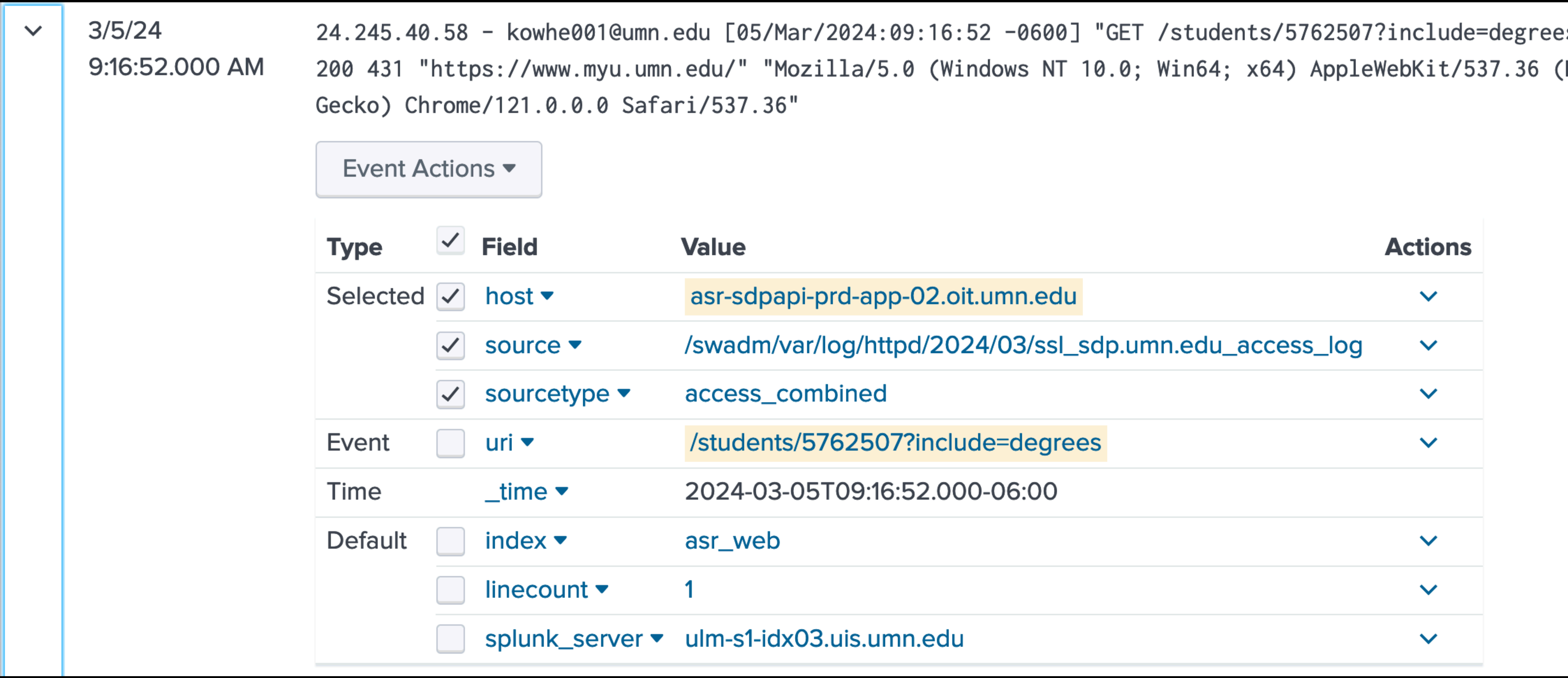
Have the Splunk team do it

- Work with the Splunk team
- Format your files a consistent way
- Store your files in a consistent location

Ingestion

Have the Splunk team do it

- Requires fewer changes on your side
- Slower. Splunk team is super busy!
- Can require you to remember to add new hosts to Splunk ingestion



3/5/24
9:16:52.000 AM

```
24.245.40.58 - kowhe001@umn.edu [05/Mar/2024:09:16:52 -0600] "GET /students/5762507?include=degree
200 431 "https://www.myu.umn.edu/" "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (
Gecko) Chrome/121.0.0.0 Safari/537.36"
```

Type	☑	Field	Value	Actions
Selected	☑	host ▾	asr-sdpapi-prd-app-02.oit.umn.edu	⌄
	☑	source ▾	/swadm/var/log/httpd/2024/03/ssl_sdp.umn.edu_access_log	⌄
	☑	sourcetype ▾	access_combined	⌄
Event	☐	uri ▾	/students/5762507?include=degrees	⌄
Time		_time ▾	2024-03-05T09:16:52.000-06:00	
Default	☐	index ▾	asr_web	⌄
	☐	linecount ▾	1	⌄
	☐	splunk_server ▾	ulm-s1-idx03.uis.umn.edu	⌄

https://z.umn.edu/splunk_data_presentation

Ingestion

Use methods that already exist, `syslog`

Ingestion

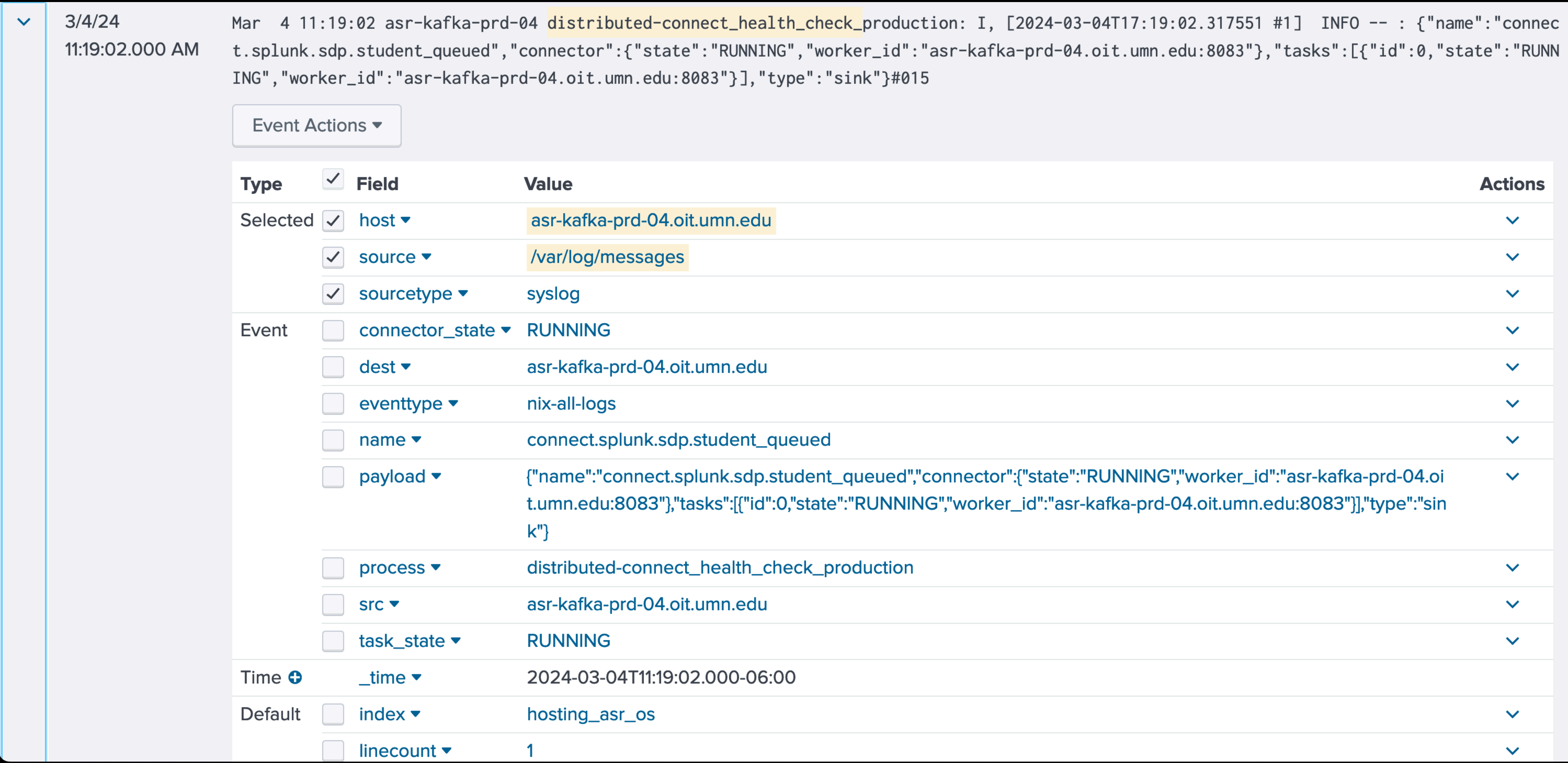
Use methods that already exist, `syslog`

- Splunk already ingests host log data
- Easy to set up if you run your process in
  - `systemd`
  - Docker
- Be sure to add unique and helpful tags

Ingestion

Use methods that already exist, `syslog`

- Ingestion is out of your control
- Written alongside a *ton* of other messages, can be noisy



3/4/24
11:19:02.000 AM

```
Mar  4 11:19:02 asr-kafka-prd-04 distributed-connect_health_check_production: I, [2024-03-04T17:19:02.317551 #1]  INFO -- : {"name":"connect.splunk.sdp.student_queued","connector":{"state":"RUNNING","worker_id":"asr-kafka-prd-04.oit.umn.edu:8083"},"tasks":[{"id":0,"state":"RUNNING","worker_id":"asr-kafka-prd-04.oit.umn.edu:8083"}],"type":"sink"}#015
```

Type	☑	Field	Value	Actions
Selected	☑	host ▾	asr-kafka-prd-04.oit.umn.edu	⌄
	☑	source ▾	/var/log/messages	⌄
	☑	sourcetype ▾	syslog	⌄
Event	☐	connector_state ▾	RUNNING	⌄
	☐	dest ▾	asr-kafka-prd-04.oit.umn.edu	⌄
	☐	eventtype ▾	nix-all-logs	⌄
	☐	name ▾	connect.splunk.sdp.student_queued	⌄
	☐	payload ▾	{"name":"connect.splunk.sdp.student_queued","connector":{"state":"RUNNING","worker_id":"asr-kafka-prd-04.oit.umn.edu:8083"},"tasks":[{"id":0,"state":"RUNNING","worker_id":"asr-kafka-prd-04.oit.umn.edu:8083"}],"type":"sink"}	⌄
	☐	process ▾	distributed-connect_health_check_production	⌄
	☐	src ▾	asr-kafka-prd-04.oit.umn.edu	⌄
	☐	task_state ▾	RUNNING	⌄
Time ⊕		_time ▾	2024-03-04T11:19:02.000-06:00	
Default	☐	index ▾	hosting_asr_os	⌄
	☐	linecount ▾	1	⌄

https://z.umn.edu/splunk_data_presentation

Ingestion

Send it yourself, `HEC`

Ingestion

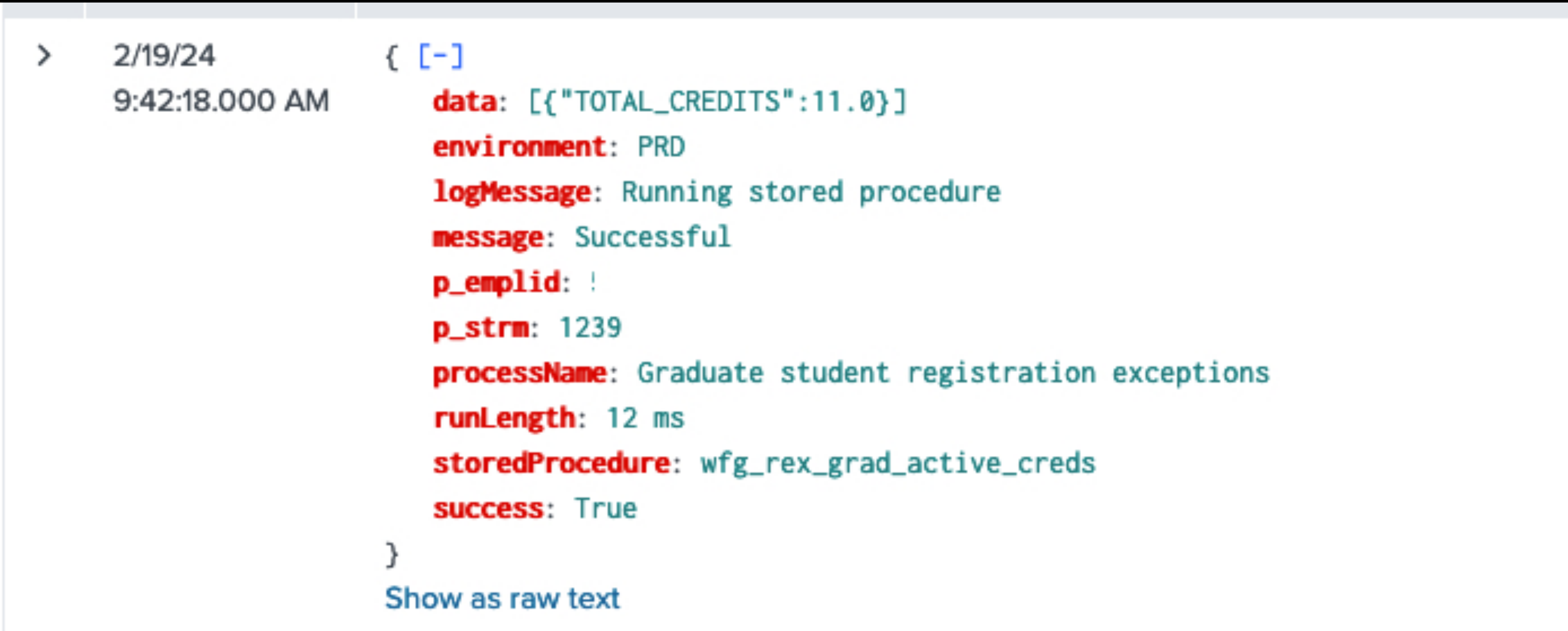
Send it yourself, `HEC`

- Get a token from the Splunk team
- Send your data over http

Ingestion

Send it yourself, `HEC`

- Code for you to manage
- Sometimes HTTP fails



2/19/24
9:42:18.000 AM

```
{ [-]
   data: [{"TOTAL_CREDITS":11.0}]
   environment: PRD
   logMessage: Running stored procedure
   message: Successful
   p_emplid:
   p_strm: 1239
   processName: Graduate student registration exceptions
   runLength: 12 ms
   storedProcedure: wfg_rex_grad_active_creds
   success: True
}
```

https://z.umn.edu/splunk_data_presentation

Ingestion

Other (secretly also `HEC`)

Ingestion

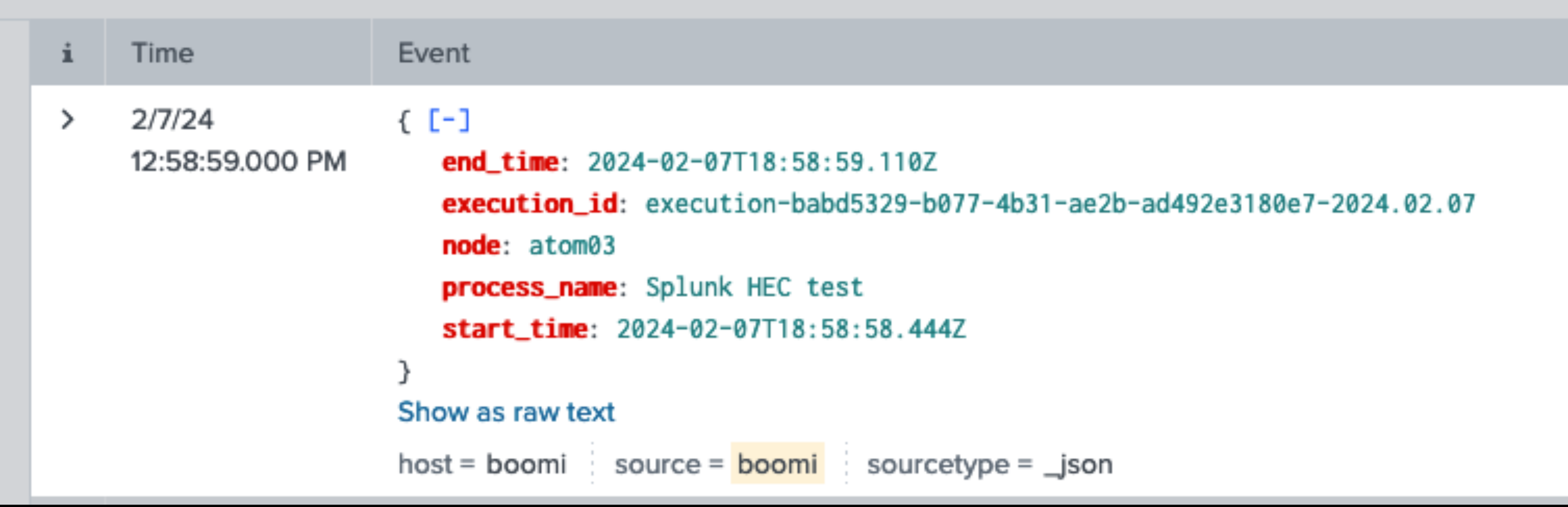
Other (secretly also `HEC`)

- Boomi
- Kafka
- Etc

Ingestion

Other (secretly also `HEC`)

- Allows you to integrate services in to Splunk
- Without managing all the integration code



```
2/7/24
12:58:59.000 PM

{ [-]
   end_time: 2024-02-07T18:58:59.110Z
   execution_id: execution-babd5329-b077-4b31-ae2b-ad492e3180e7-2024.02.07
   node: atom03
   process_name: Splunk HEC test
   start_time: 2024-02-07T18:58:58.444Z
}

host = boomi   source = boomi   sourcetype = _json
```

https://z.umn.edu/splunk_data_presentation

Ingestion

Other Other (OpenTelemetry, JMX, etc)

- Splunk offers connectors that ingests these things
- Ask the Splunk team!

Mix And Match

Generation

- Application logs
- Instrumentation
- API responses
- Database queries
- Server metrics

Ingestion

- Have the Splunk team do it
- Use methods that already exist, `syslog`
- Send it yourself, `HEC`
- Other (secretly also `HEC`)

#splunk-users

Questions

https://z.umn.edu/splunk_data_presentation