Azure Databricks

https://z.umn.edu/databricks_asr

Getting Data

Query ➡

- PeopleSoft
- uAchieve
- DORS
- Etc.

Transform ➡

- Filter
- Flatten
- Convert
- Etc.

Save

- Oracle
- Tableau
- BI
- Etc.

https://z.umn.edu/databricks_asr

Current Approaches

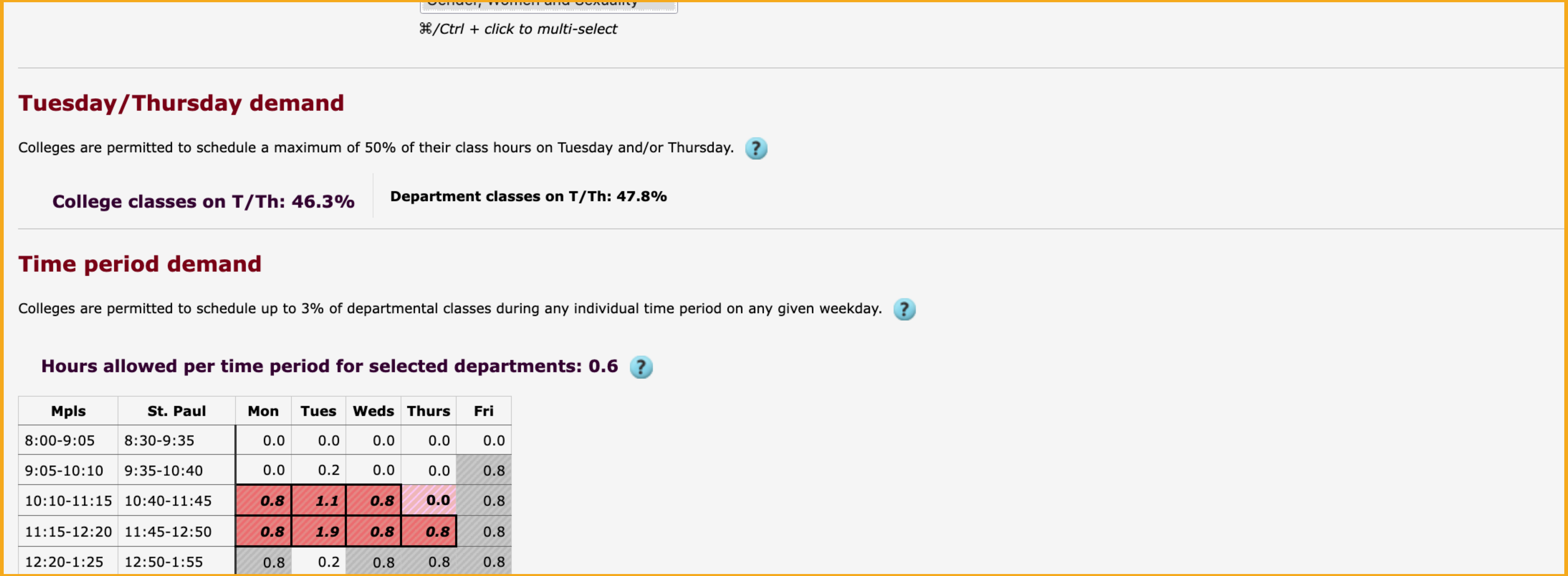
1. Application Development Team
2. OIT/EDMR

Downsides

- Changes need to go back through App Dev or OIT
- The solution is maintained by people unfamiliar with the data
- Solutions are usually single use

Example

CSDS



Gender, Women and Sexuality

⌘/Ctrl + click to multi-select

Tuesday/Thursday demand

Colleges are permitted to schedule a maximum of 50% of their class hours on Tuesday and/or Thursday.

College classes on T/Th: 46.3%

Department classes on T/Th: 47.8%

Time period demand

Colleges are permitted to schedule up to 3% of departmental classes during any individual time period on any given weekday.

Hours allowed per time period for selected departments: 0.6

Mpls	St. Paul	Mon	Tues	Weds	Thurs	Fri
8:00-9:05	8:30-9:35	0.0	0.0	0.0	0.0	0.0
9:05-10:10	9:35-10:40	0.0	0.2	0.0	0.0	0.8
10:10-11:15	10:40-11:45	***0.8***	***1.1***	***0.8***	**0.0**	0.8
11:15-12:20	11:45-12:50	***0.8***	***1.9***	***0.8***	***0.8***	0.8
12:20-1:25	12:50-1:55	0.8	0.2	0.8	0.8	0.8

https://z.umn.edu/databricks_asr

Example

CSDS

- Changes to data must be done by App Dev
- App Dev are not experts in Class Scheduling data
- No one else can use this data

https://z.umn.edu/databricks_asr

Solutions

- Works with a variety of data
- Lets people manage data they know and need
- Cloud based

https://z.umn.edu/databricks_asr

Data Lake(houses)

- Lots of options in this space
- Modern way of managing and analyzing data
- Similar to a Data Warehouse, but more flexible

https://z.umn.edu/databricks_asr

Databricks

- Cloud hosted data lakehouse
- Well integrated with Azure, OIT's preferred cloud

https://z.umn.edu/databricks_asr

Databricks

- Store data from many sources
- Send data to many targets
- Easy transformation
- Lots of automation options
- Analytics

https://z.umn.edu/databricks_asr

CSDS in Databricks

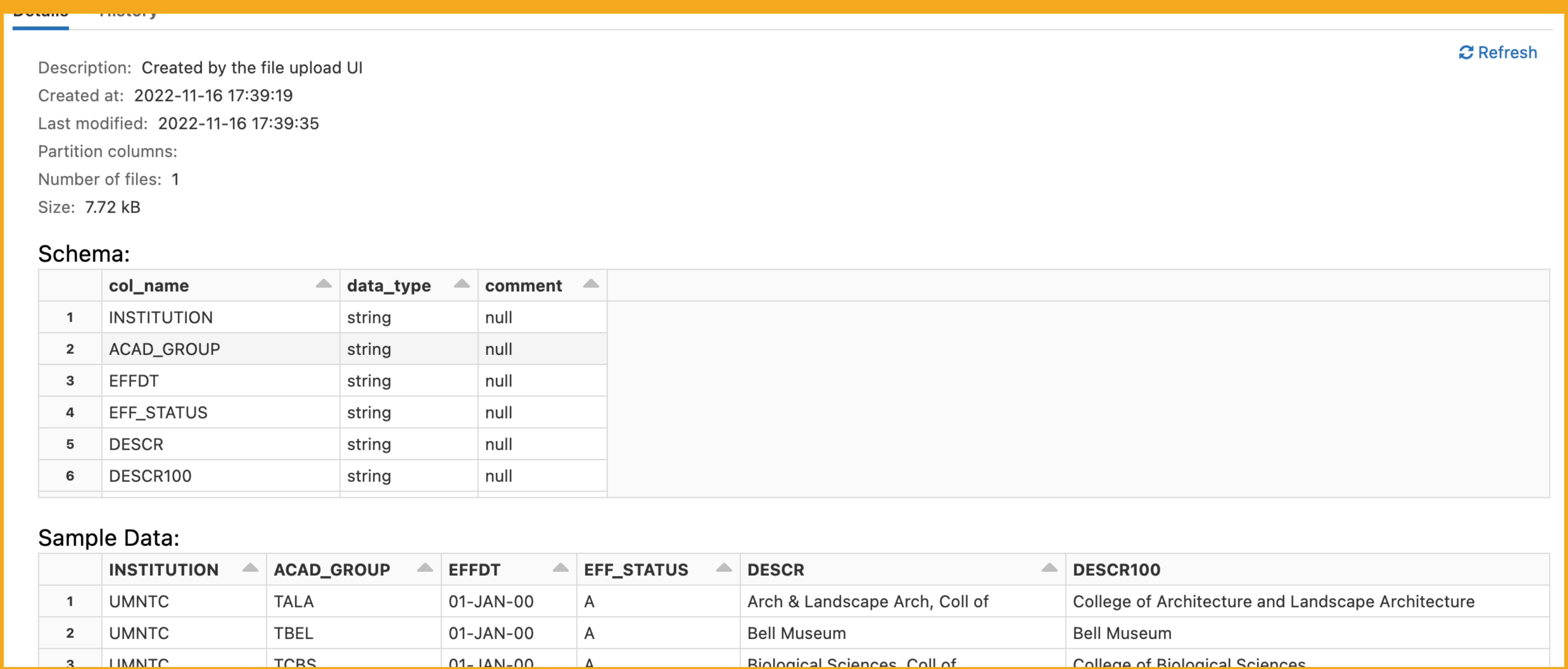
https://z.umn.edu/databricks_asr

Goals

- Easy to load data in from a variety of sources
- Data and Business Analysts can manage data they use
- Resulting data can be used by multiple systems and people

https://z.umn.edu/databricks_asr

Query



Description: Created by the file upload UI

Created at: 2022-11-16 17:39:19

Last modified: 2022-11-16 17:39:35

Partition columns:

Number of files: 1

Size: 7.72 kB

Schema:

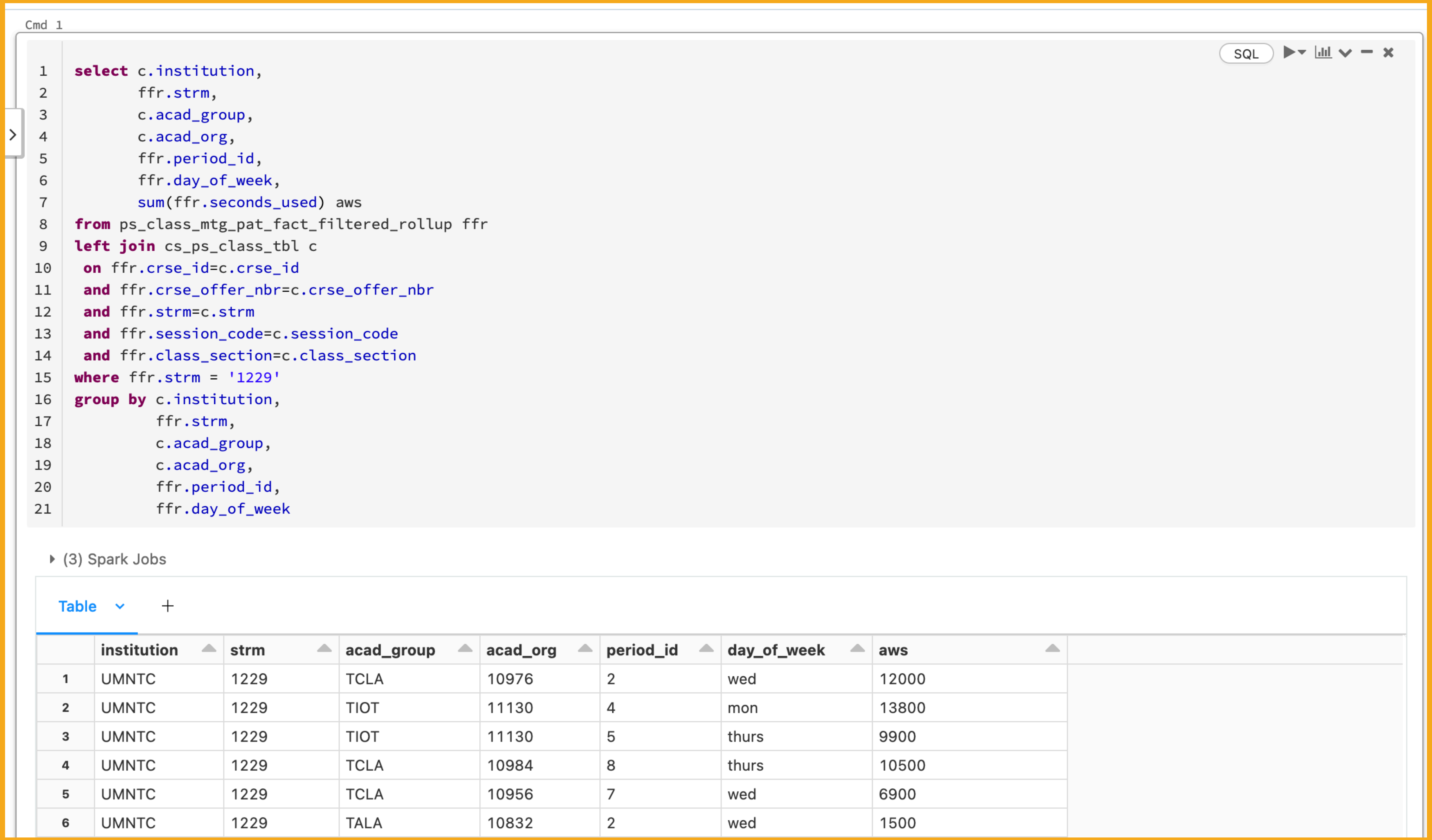
	col_name	data_type	comment
1	INSTITUTION	string	null
2	ACAD_GROUP	string	null
3	EFFDT	string	null
4	EFF_STATUS	string	null
5	DESCR	string	null
6	DESCR100	string	null

Sample Data:

	INSTITUTION	ACAD_GROUP	EFFDT	EFF_STATUS	DESCR	DESCR100
1	UMNTC	TALA	01-JAN-00	A	Arch & Landscape Arch, Coll of	College of Architecture and Landscape Architecture
2	UMNTC	TBEL	01-JAN-00	A	Bell Museum	Bell Museum
3	UMNTC	TCBS	01-JAN-00	A	Biological Sciences, Coll of	College of Biological Sciences

https://z.umn.edu/databricks_asr

Transformation Process

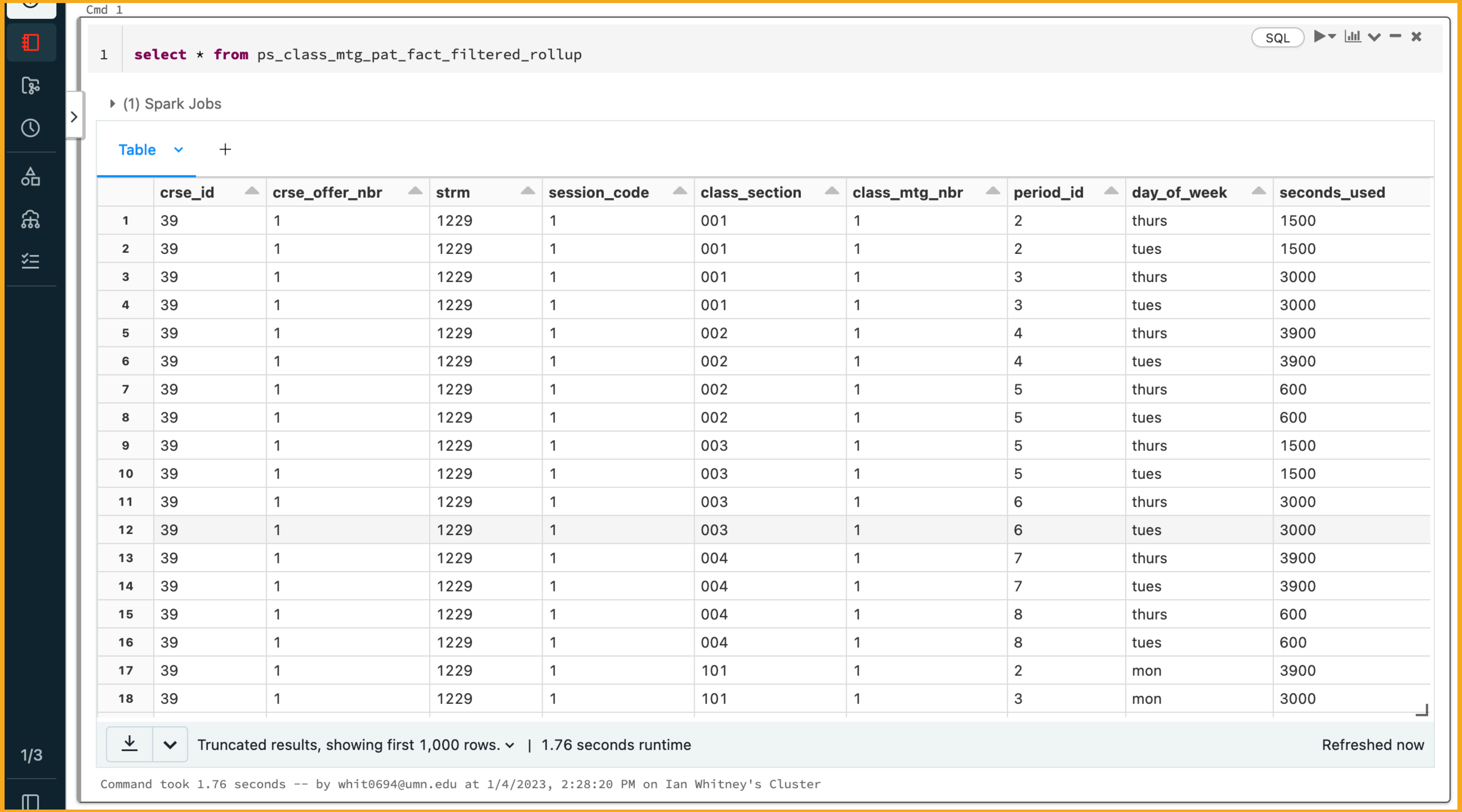


```sql
select c.institution,
       ffr.strm,
       c.acad_group,
       c.acad_org,
       ffr.period_id,
       ffr.day_of_week,
       sum(ffr.seconds_used) aws
from ps_class_mtg_pat_fact_filtered_rollup ffr
left join cs_ps_class_tbl c
 on ffr.crse_id=c.crse_id
 and ffr.crse_offer_nbr=c.crse_offer_nbr
 and ffr.strm=c.strm
 and ffr.session_code=c.session_code
 and ffr.class_section=c.class_section
where ffr.strm = '1229'
group by c.institution,
         ffr.strm,
         c.acad_group,
         c.acad_org,
         ffr.period_id,
         ffr.day_of_week
```

	institution	strm	acad_group	acad_org	period_id	day_of_week	aws
1	UMNTC	1229	TCLA	10976	2	wed	12000
2	UMNTC	1229	TIOT	11130	4	mon	13800
3	UMNTC	1229	TIOT	11130	5	thurs	9900
4	UMNTC	1229	TCLA	10984	8	thurs	10500
5	UMNTC	1229	TCLA	10956	7	wed	6900
6	UMNTC	1229	TALA	10832	2	wed	1500

https://z.umn.edu/databricks_asr

Transformation Results



```sql
select * from ps_class_mtg_pat_fact_filtered_rollup
```

	crse_id	crse_offer_nbr	strm	session_code	class_section	class_mtg_nbr	period_id	day_of_week	seconds_used
1	39	1	1229	1	001	1	2	thurs	1500
2	39	1	1229	1	001	1	2	tues	1500
3	39	1	1229	1	001	1	3	thurs	3000
4	39	1	1229	1	001	1	3	tues	3000
5	39	1	1229	1	002	1	4	thurs	3900
6	39	1	1229	1	002	1	4	tues	3900
7	39	1	1229	1	002	1	5	thurs	600
8	39	1	1229	1	002	1	5	tues	600
9	39	1	1229	1	003	1	5	thurs	1500
10	39	1	1229	1	003	1	5	tues	1500
11	39	1	1229	1	003	1	6	thurs	3000
12	39	1	1229	1	003	1	6	tues	3000
13	39	1	1229	1	004	1	7	thurs	3900
14	39	1	1229	1	004	1	7	tues	3900
15	39	1	1229	1	004	1	8	thurs	600
16	39	1	1229	1	004	1	8	tues	600
17	39	1	1229	1	101	1	2	mon	3900
18	39	1	1229	1	101	1	3	mon	3000

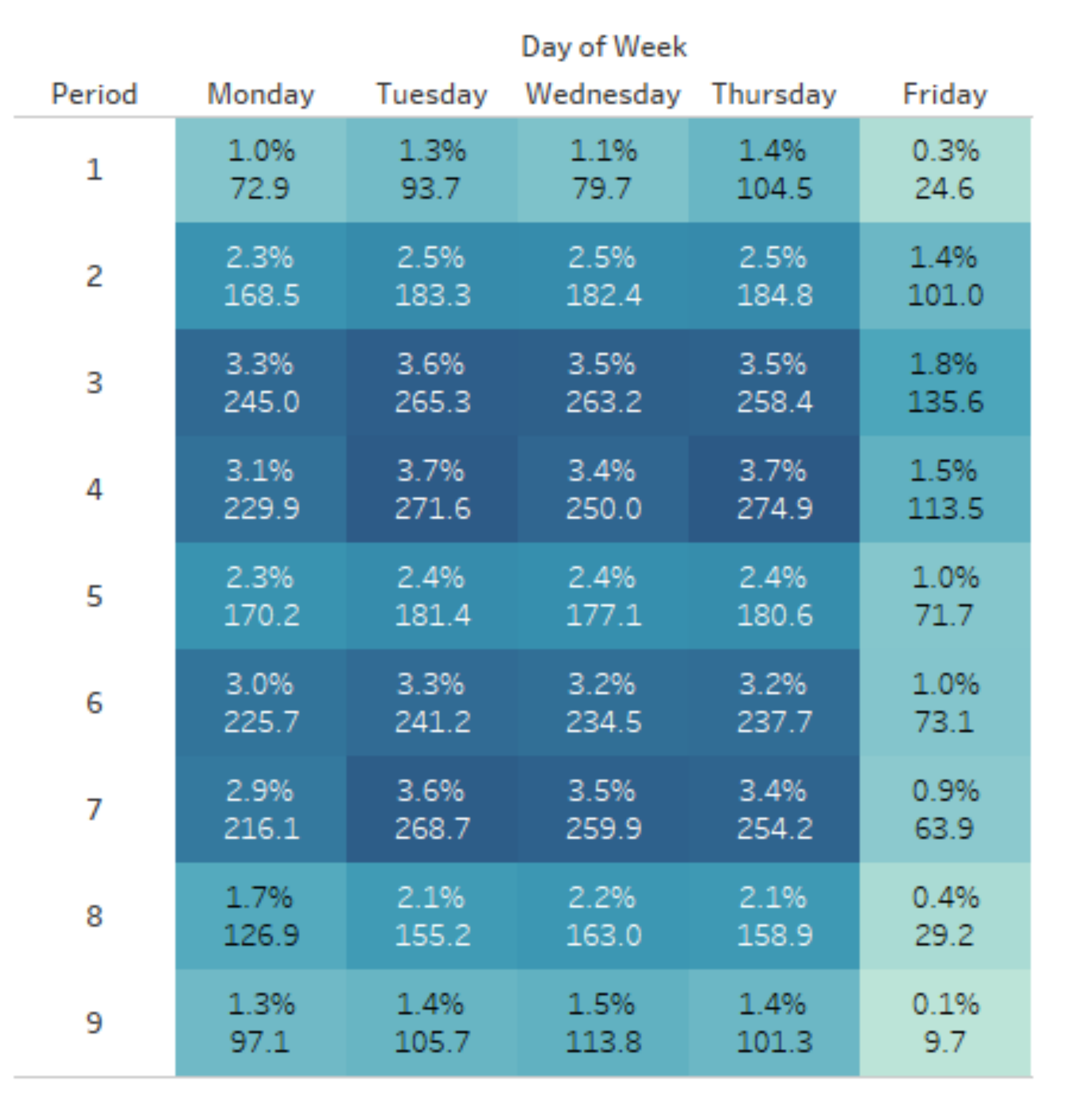
Truncated results, showing first 1,000 rows. | 1.76 seconds runtime

Refreshed now

```
Command took 1.76 seconds -- by whit0694@umn.edu at 1/4/2023, 2:28:20 PM on Ian Whitney's Cluster
```

https://z.umn.edu/databricks_asr

Load



	Day of Week				
Period	Monday	Tuesday	Wednesday	Thursday	Friday
1	1.0% 72.9	1.3% 93.7	1.1% 79.7	1.4% 104.5	0.3% 24.6
2	2.3% 168.5	2.5% 183.3	2.5% 182.4	2.5% 184.8	1.4% 101.0
3	3.3% 245.0	3.6% 265.3	3.5% 263.2	3.5% 258.4	1.8% 135.6
4	3.1% 229.9	3.7% 271.6	3.4% 250.0	3.7% 274.9	1.5% 113.5
5	2.3% 170.2	2.4% 181.4	2.4% 177.1	2.4% 180.6	1.0% 71.7
6	3.0% 225.7	3.3% 241.2	3.2% 234.5	3.2% 237.7	1.0% 73.1
7	2.9% 216.1	3.6% 268.7	3.5% 259.9	3.4% 254.2	0.9% 63.9
8	1.7% 126.9	2.1% 155.2	2.2% 163.0	2.1% 158.9	0.4% 29.2
9	1.3% 97.1	1.4% 105.7	1.5% 113.8	1.4% 101.3	0.1% 9.7

https://z.umn.edu/databricks_asr

Databricks

Benefits

- We can store data from multiple sources in many formats
- We can let data experts transform or analyze that data
- The results of their work can be used by others

https://z.umn.edu/databricks_asr

Databricks

Differences from other systems

- Boomi API
- Data Warehouse

Databricks

Next Steps

- Work with OIT to improve access to Oracle
- Investigate Automation and Operation
- More proof of concepts

https://z.umn.edu/databricks_asr

Questions

https://z.umn.edu/databricks_asr