From Data Sets to Databricks

Faron Kincheloe

Baylor University

What is Databricks?

What is Databricks?

- **Unified Data Warehouse/Analytics Platform**: Combines data engineering, data science, machine learning, and analytics in one collaborative workspace.
- **Integrated Programming Tools**: Enables teams to work together using notebooks with support for Python, SQL, R, and Scala.
- **Enterprise-Grade Security & Governance**: Offers robust tools for data access control, auditing, and compliance.
- **Optimized for Cloud**: Available on AWS, Azure, and Google Cloud, with seamless scalability and integration.

Databricks Migration Project

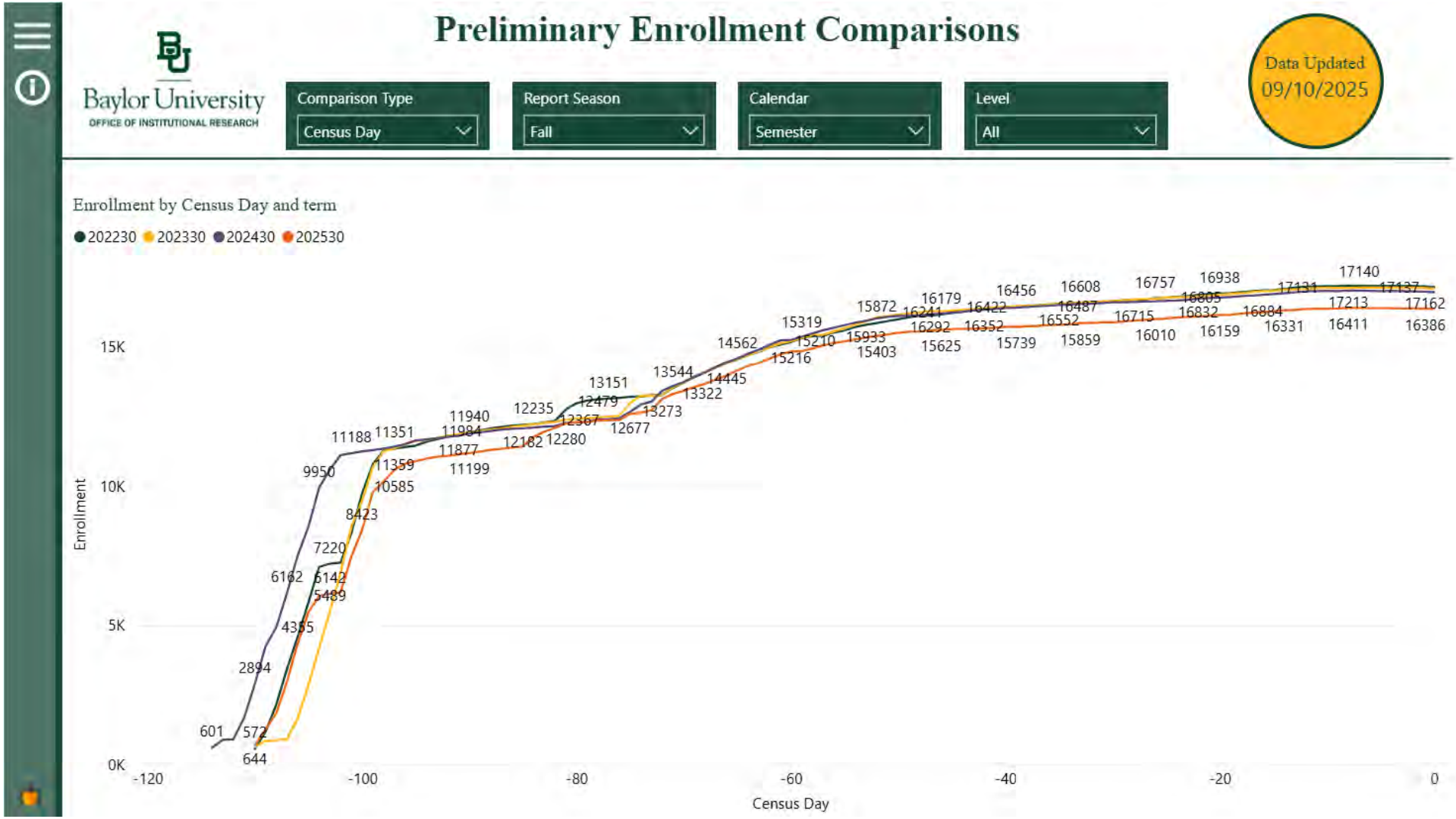
Major Goals/Requirements

- ✓Read/Write between SAS and Databricks
- ✓Load new and historical data into Databricks
- ✓Point in time comparisons
- ✓Manual drops
- ✓Manual edits (e.g.: retroactive major code changes)
- ✓Error recovery
- ✓Apply user formats
- ✓Connect from Power BI
- ✓Manage users and security (Power BI does not inherit permissions from Databricks)

Pilot Milestones

- Start with the greatest challenge
- By Summer II Census date – Have automatic data uploads in place to observe handling of records on a census date
- Start of Fall Semester – Deliver a Preliminary Enrollment Power BI dashboard using only Databricks data as a source
  - Required loading historical data for comparison
  - Initially load two years for proof of concept

Pilot Milestones



Writing to Databricks with SAS

- SAS SQL Passthrough cannot read local data sets
- Use PROC HTTP to write SAS data set to Databricks landing zone

```
proc http
url="https://&databricks_host./api/2.0/fs/files&volume_path."
     method="PUT"
     in=infile
     out=response;
     headers
          "Authorization"="Bearer &access_token."
          "Content-Type"="application/octet-stream";
run;
```

Writing to Databricks with SAS

- Specially crafted filename tells Databricks what to do

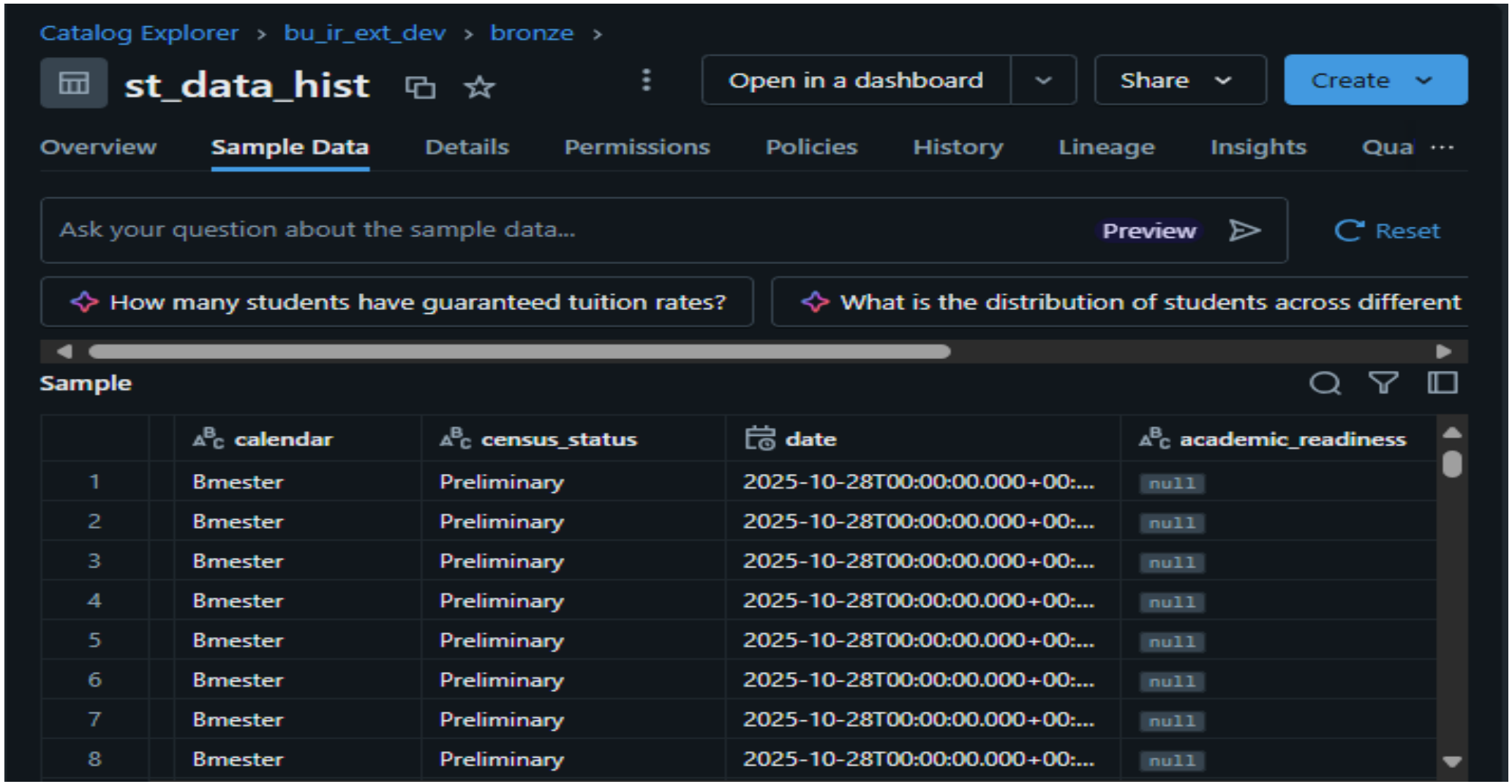
```
%let long_filename = &timestamp.-&table.-&pkeys..sas7bdat;
```

- “File Arrival” trigger launches script to ingest data set
- Python function converts data set to Databricks table

```
encodings_to_try = ['latin-1', 'utf-8', 'cp1252']
    spark_df = None
    for encoding in encodings_to_try:
      try:
        pandas_df = pd.read_sas(volume_path, format='sas7bdat', encoding=encoding)
```

- Changed Data Capture & flags determine which rows to process/load to end table

Changed Data Capture(CDC)
Slowly Changing Dimensions(SCD)



Sample Daily Load Data

```sql
select term, calendar, pidm, last_name, first_name, middle_initial, census_status, effective_start_date, effective_end_date, is_current
  from bu_ir_ext_dev.silver.st_data_hist
  where term = '202533'
    and calendar='Bmester'
 order by pidm, effective_start_date
```

Add parameter

Table

	term	calendar	pidm	last_name	first_name	middle_initial	census_status	effective_start_date
1	202533	Bmester	[redacted]8171	Al[redacted]	Yvonne	A	Preliminary	2025-08-12T00:00:00.000+00:...
2	202533	Bmester	[redacted]8171	Ac[redacted]	Yvonne	null	Preliminary	2025-08-29T00:00:00.000+00:...
3	202533	Bmester	[redacted]8171	Ac[redacted]	Yvonne	null	Preliminary	2025-09-01T00:00:00.000+00:...
4	202533	Bmester	[redacted]8171	Ac[redacted]	Yvonne	null	Preliminary	2025-09-23T00:00:00.000+00:...
5	202533	Bmester	[redacted]8171	Ac[redacted]	Yvonne	null	Preliminary	2025-10-28T00:00:00.000+00:...
6	202533	Bmester	[redacted]8436	Bl[redacted]	Pamela	M	Preliminary	2025-07-28T00:00:00.000+00:...
7	202533	Bmester	[redacted]8436	Bl[redacted]	Pamela	M	Preliminary	2025-08-11T00:00:00.000+00:...
8	202533	Bmester	[redacted]8436	Bl[redacted]	Pamela	M	Preliminary	2025-08-18T00:00:00.000+00:...
9	202533	Bmester	[redacted]8436	Bl[redacted]	Pamela	M	Preliminary	2025-09-03T00:00:00.000+00:...
10	202533	Bmester	[redacted]8436	Bl[redacted]	Pamela	M	Preliminary	2025-09-23T00:00:00.000+00:...

Daily Student Data with SCD

```sql
select term, calendar, pidm, last_name, first_name, middle_initial, census_status, effective_start_date, effective_end_date, is_current
  from bu_ir_ext_dev.silver.st_data_hist
  where term = '202533'
    and calendar='Bmester'
 order by pidm, effective_start_date
```

Add parameter

Table +

	last_name	first_name	middle_initial	census_status	effective_start_date	effective_end_date	is_current
1	Al[redacted]	Yvonne	A	Preliminary	2025-08-12T00:00:00.000+00:...	2025-08-29T00:00:00.000+00:...	false
2	Ac[redacted]	Yvonne	null	Preliminary	2025-08-29T00:00:00.000+00:...	2025-09-01T00:00:00.000+00:...	false
3	Ac[redacted]	Yvonne	null	Preliminary	2025-09-01T00:00:00.000+00:...	2025-09-23T00:00:00.000+00:...	false
4	Ac[redacted]	Yvonne	null	Preliminary	2025-09-23T00:00:00.000+00:...	2025-10-28T00:00:00.000+00:...	false
5	Ac[redacted]	Yvonne	null	Preliminary	2025-10-28T00:00:00.000+00:...	null	true
6	Bl[redacted]	Pamela	M	Preliminary	2025-07-28T00:00:00.000+00:...	2025-08-11T00:00:00.000+00:...	false
7	Bl[redacted]	Pamela	M	Preliminary	2025-08-11T00:00:00.000+00:...	2025-08-18T00:00:00.000+00:...	false
8	Bl[redacted]	Pamela	M	Preliminary	2025-08-18T00:00:00.000+00:...	2025-09-03T00:00:00.000+00:...	false
9	Bl[redacted]	Pamela	M	Preliminary	2025-09-03T00:00:00.000+00:...	2025-09-23T00:00:00.000+00:...	false
10	Bl[redacted]	Pamela	M	Preliminary	2025-09-23T00:00:00.000+00:...	null	true

Daily Student Data with SCD (cont.)

```
11  from bu_ir_ext_dev.silver.daily_term_data
12  where term = '202533'
13  order by date, calendar desc
```

Add parameter

Table

	term	calendar	is_term_reported	date	start_date	start_day_key	
1	202533	TRIMESTER	Y	2025-07-21T00:00:00.000+00:...	2025-09-02T00:00:00.000+00:...	-31	2
2	202533	BMESTER	N	2025-07-21T00:00:00.000+00:...	2025-10-27T00:00:00.000+00:...	-70	2
3	202533	TRIMESTER	Y	2025-07-22T00:00:00.000+00:...	2025-09-02T00:00:00.000+00:...	-30	2
4	202533	BMESTER	N	2025-07-22T00:00:00.000+00:...	2025-10-27T00:00:00.000+00:...	-69	2
5	202533	TRIMESTER	Y	2025-07-23T00:00:00.000+00:...	2025-09-02T00:00:00.000+00:...	-29	2
6	202533	BMESTER	N	2025-07-23T00:00:00.000+00:...	2025-10-27T00:00:00.000+00:...	-68	2
7	202533	TRIMESTER	Y	2025-07-24T00:00:00.000+00:...	2025-09-02T00:00:00.000+00:...	-28	2
8	202533	BMESTER	N	2025-07-24T00:00:00.000+00:...	2025-10-27T00:00:00.000+00:...	-67	2
9	202533	TRIMESTER	Y	2025-07-27T00:00:00.000+00:...	2025-09-02T00:00:00.000+00:...	-27	2
10	202533	BMESTER	N	2025-07-27T00:00:00.000+00:...	2025-10-27T00:00:00.000+00:...	-66	2

Time Intelligence Table

```
11  from bu_ir_ext_dev.silver.daily_term_data
12  where term = '202533'
13  order by date, calendar desc
```

Add parameter

Table +

		start_date	1.2 start_day_key	registration_date	1.2 reg_day_key	census_date	1.2
1	+00:...	2025-09-02T00:00:00.000+00:...	-31	2025-07-21T00:00:00.000+00:...	0	2025-09-16T00:00:00.000+00:...	
2	+00:...	2025-10-27T00:00:00.000+00:...	-70	2025-07-21T00:00:00.000+00:...	0	2025-11-03T00:00:00.000+00:...	
3	+00:...	2025-09-02T00:00:00.000+00:...	-30	2025-07-21T00:00:00.000+00:...	1	2025-09-16T00:00:00.000+00:...	
4	+00:...	2025-10-27T00:00:00.000+00:...	-69	2025-07-21T00:00:00.000+00:...	1	2025-11-03T00:00:00.000+00:...	
5	+00:...	2025-09-02T00:00:00.000+00:...	-29	2025-07-21T00:00:00.000+00:...	2	2025-09-16T00:00:00.000+00:...	
6	+00:...	2025-10-27T00:00:00.000+00:...	-68	2025-07-21T00:00:00.000+00:...	2	2025-11-03T00:00:00.000+00:...	
7	+00:...	2025-09-02T00:00:00.000+00:...	-28	2025-07-21T00:00:00.000+00:...	3	2025-09-16T00:00:00.000+00:...	
8	+00:...	2025-10-27T00:00:00.000+00:...	-67	2025-07-21T00:00:00.000+00:...	3	2025-11-03T00:00:00.000+00:...	
9	+00:...	2025-09-02T00:00:00.000+00:...	-27	2025-07-21T00:00:00.000+00:...	4	2025-09-16T00:00:00.000+00:...	
10	+00:...	2025-10-27T00:00:00.000+00:...	-66	2025-07-21T00:00:00.000+00:...	4	2025-11-03T00:00:00.000+00:...	

Time Intelligence Table (cont.)

```
select
  a.term,
  a.calendar,
  a.pidm ... a.other_columns,
  a.census_status,
  b.date,
  b.reg_day_key,
  b.start_day_key,
  b.census_day_key
from
  bu_ir_ext_dev.silver.st_data_hist as a,
  bu_ir_ext_dev.silver.daily_term_data as b
where
  a.term = b.term
  and a.calendar = initcap(b.calendar)
  and date(b.date) >= date(a.effective_start_date)
  and (date(b.date) < date(a.effective_end_date)
    or effective_end_date is null  )
```

Daily Enrollment View

	term	calendar	pidm	name	census_status	date	reg_day_key	start_day_key	census_day_key
1	202533	Bmester	[redacted]8171	Al-[redacted]Yvonne	Preliminary	2025-08-11T00:00:00.000+00:...	15	-55	-60
2	202533	Bmester	[redacted]8171	Al-[redacted]Yvonne	Preliminary	2025-08-12T00:00:00.000+00:...	16	-54	-59
3	202533	Bmester	[redacted]8171	Al-[redacted]Yvonne	Preliminary	2025-08-13T00:00:00.000+00:...	17	-53	-58
4	202533	Bmester	[redacted]8171	Al-[redacted]Yvonne	Preliminary	2025-08-14T00:00:00.000+00:...	18	-52	-57
5	202533	Bmester	[redacted]8171	Al-[redacted]Yvonne	Preliminary	2025-08-17T00:00:00.000+00:...	19	-51	-56
6	202533	Bmester	[redacted]8171	Al-[redacted]Yvonne	Preliminary	2025-08-18T00:00:00.000+00:...	20	-50	-55
7	202533	Bmester	[redacted]8171	Al-[redacted]Yvonne	Preliminary	2025-08-19T00:00:00.000+00:...	21	-49	-54
8	202533	Bmester	[redacted]8171	Al-[redacted]Yvonne	Preliminary	2025-08-20T00:00:00.000+00:...	22	-48	-53
9	202533	Bmester	[redacted]8171	Al-[redacted]Yvonne	Preliminary	2025-08-21T00:00:00.000+00:...	23	-47	-52
10	202533	Bmester	[redacted]8171	Al-[redacted]Yvonne	Preliminary	2025-08-24T00:00:00.000+00:...	24	-46	-51
11	202533	Bmester	[redacted]8171	Al-[redacted]Yvonne	Preliminary	2025-08-25T00:00:00.000+00:...	25	-45	-50
12	202533	Bmester	[redacted]8171	Al-[redacted]Yvonne	Preliminary	2025-08-26T00:00:00.000+00:...	26	-44	-49
13	202533	Bmester	[redacted]8171	Al-[redacted]Yvonne	Preliminary	2025-08-27T00:00:00.000+00:...	27	-43	-48
14	202533	Bmester	[redacted]8171	Ac[redacted]onne	Preliminary	2025-08-28T00:00:00.000+00:...	28	-42	-47

Daily Enrollment View Sample Results

```sql
select term, calendar, pidm, last_name, first_name, middle_initial, census_status, effective_start_date, effective_end_date, is_current
  from bu_ir_ext_dev.silver.st_data_hist
  where term = '202533'
    and calendar='Trimester'
  order by pidm, effective_start_date
```

Add parameter

Table

	term	calendar	pidm	last_name	first_name	middle_initial	census_status	effective_start_date
1	202533	Trimester	[redacted]8171	Al[redacted]	Yvonne	A	Preliminary	2025-08-12T00:00:00.000+00:...
2	202533	Trimester	[redacted]8171	Ac[redacted]	Yvonne	null	Preliminary	2025-08-29T00:00:00.000+00:...
3	202533	Trimester	[redacted]8171	Ac[redacted]	Yvonne	null	Preliminary	2025-09-01T00:00:00.000+00:...
4	202533	Trimester	[redacted]8171	Ac[redacted]	Yvonne	null	Census	2025-09-17T00:00:00.000+00:...
5	202533	Trimester	[redacted]8175	Sc[redacted]	Juli	J	Preliminary	2025-07-23T00:00:00.000+00:...
6	202533	Trimester	[redacted]8175	Sc[redacted]	Juli	J	Preliminary	2025-07-30T00:00:00.000+00:...
7	202533	Trimester	[redacted]8175	Sc[redacted]	Juli	J	Preliminary	2025-08-11T00:00:00.000+00:...
8	202533	Trimester	[redacted]8175	Sc[redacted]	Juli	J	Preliminary	2025-08-13T00:00:00.000+00:...
9	202533	Trimester	[redacted]8175	Sc[redacted]	Juli	J	Preliminary	2025-08-14T00:00:00.000+00:...
10	202533	Trimester	[redacted]8175	Sc[redacted]	Juli	J	Preliminary	2025-08-18T00:00:00.000+00:...

Daily Student Data after Census

```sql
select term, calendar, pidm, last_name, first_name, middle_initial, census_status, effective_start_date, effective_end_date, is_current
  from bu_ir_ext_dev.silver.st_data_hist
  where term = '202533'
    and calendar='Trimester'
 order by pidm, effective_start_date
```

Add parameter

Table

	last_name	first_name	middle_initial	census_status	effective_start_date	effective_end_date	is_current
1	Al[redacted]	Yvonne	A	Preliminary	2025-08-12T00:00:00.000+00:...	2025-08-29T00:00:00.000+00:...	false
2	Ac[redacted]	Yvonne	null	Preliminary	2025-08-29T00:00:00.000+00:...	2025-09-01T00:00:00.000+00:...	false
3	Ac[redacted]	Yvonne	null	Preliminary	2025-09-01T00:00:00.000+00:...	2025-09-17T00:00:00.000+00:...	false
4	Ac[redacted]	Yvonne	null	Census	2025-09-17T00:00:00.000+00:...	2025-09-18T00:00:00.000+00:...	false
5	Sc[redacted]	Juli	J	Preliminary	2025-07-23T00:00:00.000+00:...	2025-07-30T00:00:00.000+00:...	false
6	Sc[redacted]	Juli	J	Preliminary	2025-07-30T00:00:00.000+00:...	2025-08-11T00:00:00.000+00:...	false
7	Sc[redacted]	Juli	J	Preliminary	2025-08-11T00:00:00.000+00:...	2025-08-13T00:00:00.000+00:...	false
8	Sc[redacted]	Juli	J	Preliminary	2025-08-13T00:00:00.000+00:...	2025-08-14T00:00:00.000+00:...	false
9	Sc[redacted]	Juli	J	Preliminary	2025-08-14T00:00:00.000+00:...	2025-08-18T00:00:00.000+00:...	false
10	Sc[redacted]	Juli	J	Preliminary	2025-08-18T00:00:00.000+00:...	2025-09-17T00:00:00.000+00:...	false

Daily Student Data after Census

Handling Manual Drops

	calendar	term	1.2 pidm	app_term	class_adjusted	incentivized_program	name	new_student_t
1	Semester	202520	[redacted]5296.0	202520	GM	null	Naig[redacted]	G
2	Semester	202520	[redacted]7739.0	202410	SR	null	Spec[redacted]	C
3	Semester	202520	[redacted]7346.0	202520	RS	null	Gom[redacted]	H
4	Semester	202520	[redacted]7638.0	202520	SM	null	Bang[redacted]	S
5	Semester	202520	[redacted]4210.0	202520	SM	null	Wrig[redacted]	S
6	Semester	202520	[redacted]3390.0	202230	GD	null	Pope[redacted]	C
7	Semester	202520	[redacted]5696.0	202130	SR	null	Angu[redacted]	C

Manual Drops Table

```
select
  *
from
  bu_ir_ext_dev.silver.st_data_hist
where
  census_status = 'Census'
  and calendar || term || pidm not in (
    select
      calendar || term || pidm
    from
      bu_ir_ext_dev.silver.manual_student_drops
  )
```

Census Enrollment View (st_data_census)

Reading Databricks with SAS

```
proc sql;  /* Requires SAS version with Unicode Support */
/* Connection string to enable direct query with databricks  (could be libref)*/
connect to jdbc as db (  < JDBC connection parameters >  );

/*      SQL Passthrough Query:
     - The outer SELECT pulls data into SAS.
     - The inner SQL (inside the parentheses) is sent directly to Databricks via JDBC.    */

create table st_data_census_&term as
select * from connection to db (
    /* This inner SQL is executed by databricks, not SAS */
       select *
       from bu_ir_ext_dev.silver.st_data_census   /* [Catalog].[Schema].[Table] */
       where term = "&term"
 );
```

SAS Access to Census Enrollment View

Questions?

Faron_Kincheloe@baylor.edu

Baylor University

Scan the QR code to complete the session survey.

2026