Data Warehouse Quality Testing

Afshin Karimi
Sunny Moon

Institutional Research & Analytical Studies
CSU Fullerton

2016 CAIR Conference - Los Angeles, CA
11/17/2016

Why Do We Need Data Warehouses?

- **Definition**: Large store of time-variant, non-volatile data accumulated from different sources used for reporting/analysis

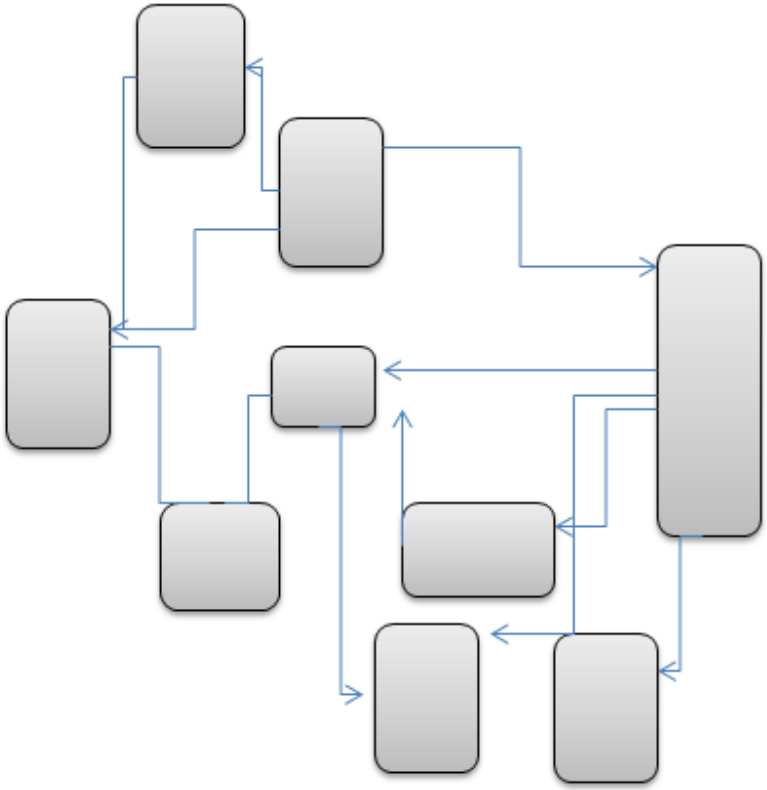
- **Data Warehouses are needed because:**
  - Live operational systems are not easily accessible; not designed for end-user analysis
  - Separate analysis/decision support from the operational systems
  - Querying operational databases causes performance issues
  - Needed data may reside in different databases on different servers in different formats
  - DW supports ad-hoc, unplanned exploration of the data

Differences between Live Operational Systems and DWs

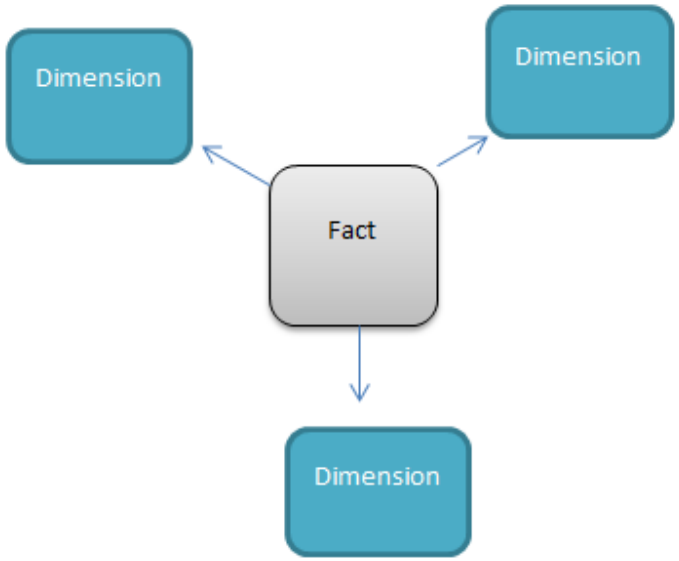
- Operational vs. Informational
- Transactional vs. Analytical
- Relational Data Model vs. Multi-Dimensional Data Model (star schema)
- Ease of Access

Relational vs. Dimensional Data Model

Relational Model (normalized data)



Multi-Dimensional Model (star schema)



What's Needed to Test a DW

1. Data warehouse & ETL business rules (mapping document, transformation rules)

2. Environment other than production (test and/or development)

1. Read/Write access to test instances of the source databases (data sandboxes)

2. Ability to launch the ETL process and have visibility into DW

What's Needed… #1 DW& ETL business rules

- Example: CSUF Student Success Dashboard – need to know the rules behind the three Key Performance Indicator flags

Cohort Description	Enrollment Type Description	Initial Cohort Size	Graduated Count	Enrolled Count	Not Graduated not Enrolled Count
fa04	First-time Full-Time Freshman	3,542	2,180	20	1,342
fa05	First-time Full-Time Freshman	3,820	2,252	21	1,547
fa06	First-time Full-Time Freshman	3,737	2,245	23	1,469
fa07	First-time Full-Time Freshman	4,042	2,449	69	1,524
fa08	First-time Full-Time Freshman	4,519	2,750	104	1,665
fa09	First-time Full-Time Freshman	3,845	2,395	223	1,227
fa10	First-time Full-Time Freshman	3,749	1,843	581	1,325
fa11	First-time Full-Time Freshman	4,091	900	1,738	1,453
fa12	First-time Full-Time Freshman	4,419	24	3,353	1,042
fa13	First-time Full-Time Freshman	4,512	1	3,603	908
fa14	First-time Full-Time Freshman	4,243	0	3,616	627
fa15	First-time Full-Time Freshman	4,287	0	4,097	190

What's Needed… #1DW& ETL business rules continued

Current Enrolled Flag:

Set for currently enrolled in an Undergraduate Academic program (state-support degree program only,). Include enrollment in one or more credit units (excluding courses with 'W' grade). Self-support degree program enrollments, other extended education program enrollments, post-bacc., graduate level program as well as certificate, credential only enrollments are excluded.

```
select fc.cohort_sid, fc.person_sid, per.person_id from ps_f_cohort_csu fc, ps_d_person per
   where fc.person_sid = per.person_sid  and exists (
      select 1 from ps_stdnt_enrl
         where stdnt_enrl_status = 'E' and crse_grade_off <> 'W' and acad_career = 'UGRD' and acad_prog = 'UGD' and unt_taken > 0
         and strm = (select term_cd from ps_cohort_term_csu
            where sysdate between term_begin_dt and term_end_dt) and emplid = per.person_id
```

Graduated flag:

Set for undergraduate baccalaureate degree recipients (exclude certificates, minors, 2nd bacc. degrees, graduate degrees, teaching credentials)

```
select 1 from ps_f_degrees fd, ps_d_acad_car ac, ps_d_deg dg
   where fd.person_sid = ps_f_cohort_csu.person_sid and fd.acad_car_sid = ac.acad_car_sid  and fd.deg_sid = dg.deg_sid
    and ac.acad_car_cd = 'UGRD' and dg.deg_cd like 'B%'
     --02/14/2013  added filter only look for B% degrees
     and fd.degree_count_awd = 1
```

Not graduated, not enrolled Flag:

set to true if both above flags are false

```
update ps_f_cohort_csu a   set a.not_grad_enr_cnt = 1 where a.enrolled_cnt = 0 and a.graduated_cnt = 0
```

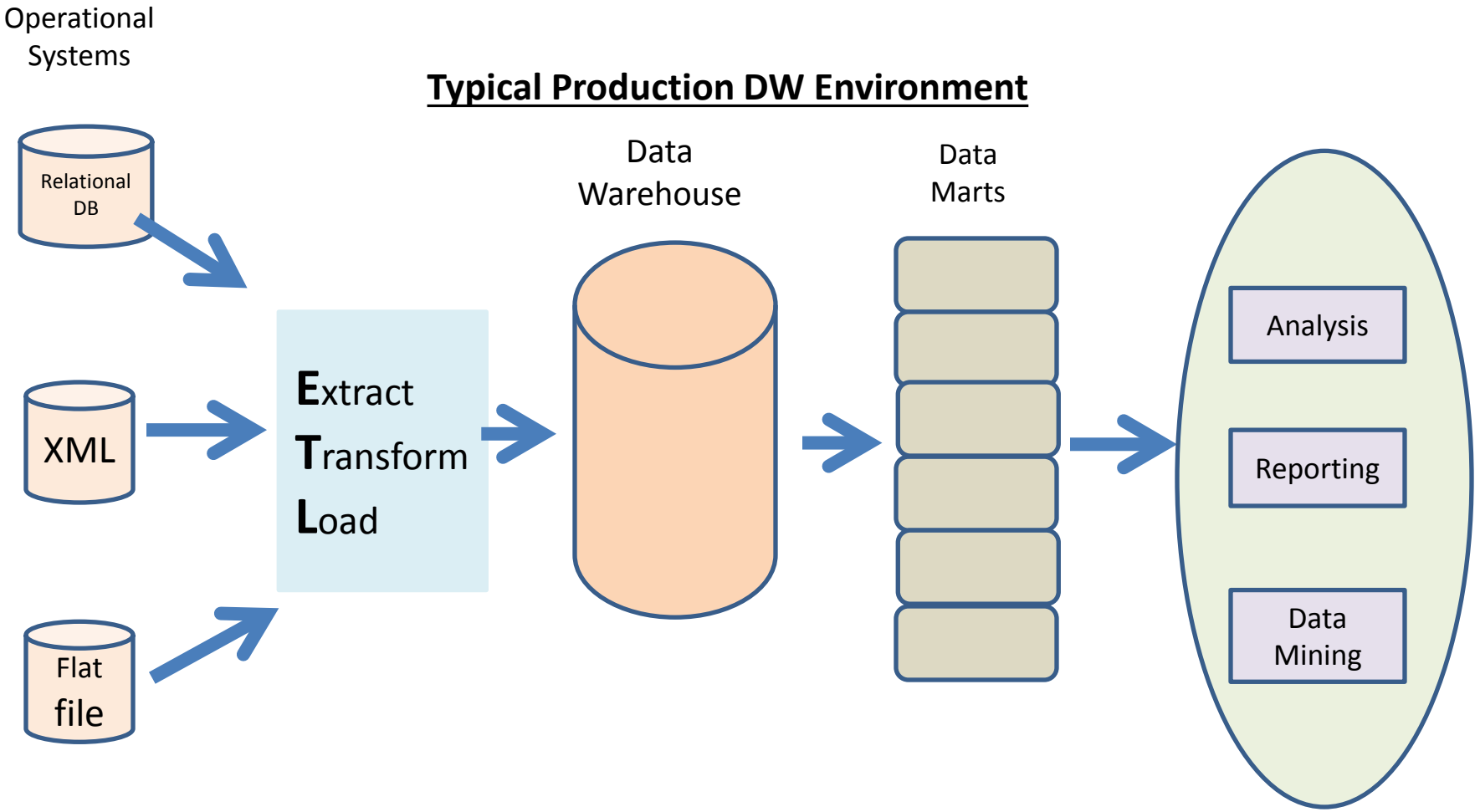
- Three flags are mutually exclusive (no overlap), sum of three flags=cohort count
- If 'Graduated' flag is set, overwrite Enrolled flag to zero

What's Needed… #1 DW& ETL business rules continued

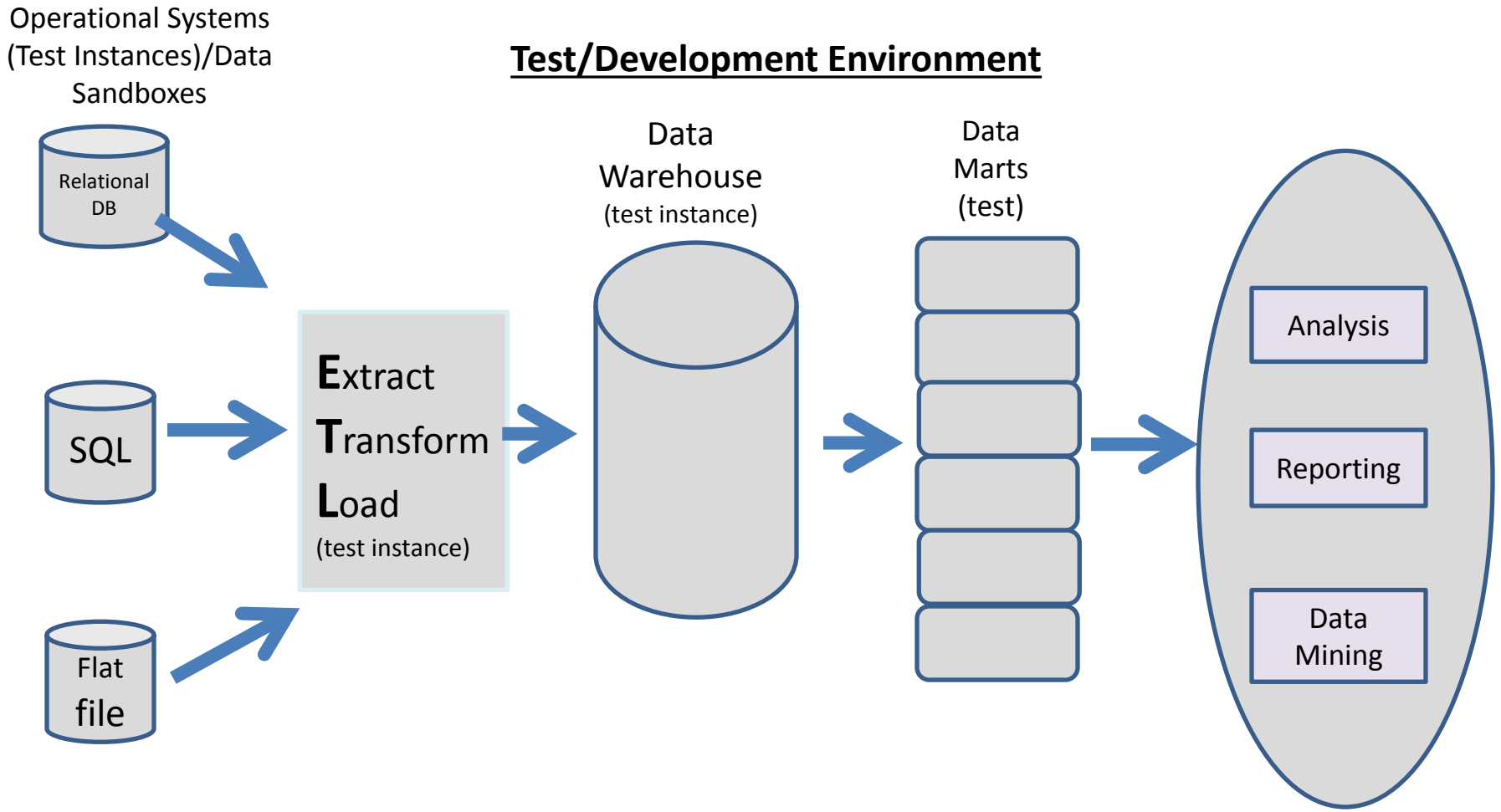
Test Cases

Test #	Area	Test Steps	Expected results
1	3 KPI flags	for a given cohort, add up the three values (enrolled, graduated, neither graduated nor enrolled)	sum should equal cohort size
2	3 KPI flags	download student lists under the three flags	no student should be cross-listed in more than 1 group
2	Enrolled flag	find an student with degree objective of Master's degree	student should not be included in the cohort
3	Enrolled flag	locate a student (in SIS) who had withdrawn from all classes this current semester	student should have enroll flag=0, neither graduated, nor enrolled flag=1
4	Enrolled flag	locate an undergraduate degree-seeking student who is also a Credential student enrolled in current term	student's enroll flag should be 1
5	Enrolled flag	locate a 2nd bacc. enrolled degree seeking student	student should not be included in cohort
6	Enrolled flag	locate a student who graduated last term	student should have degree flag=1, enrolled flag=0
7	Enrolled flag	locate a recent CSUF bachelor degree graduate who is currently enrolled in a graduate program	student's graduated flag should be 1, enrolled flag should be 0

What's Needed… #2 Test Environment



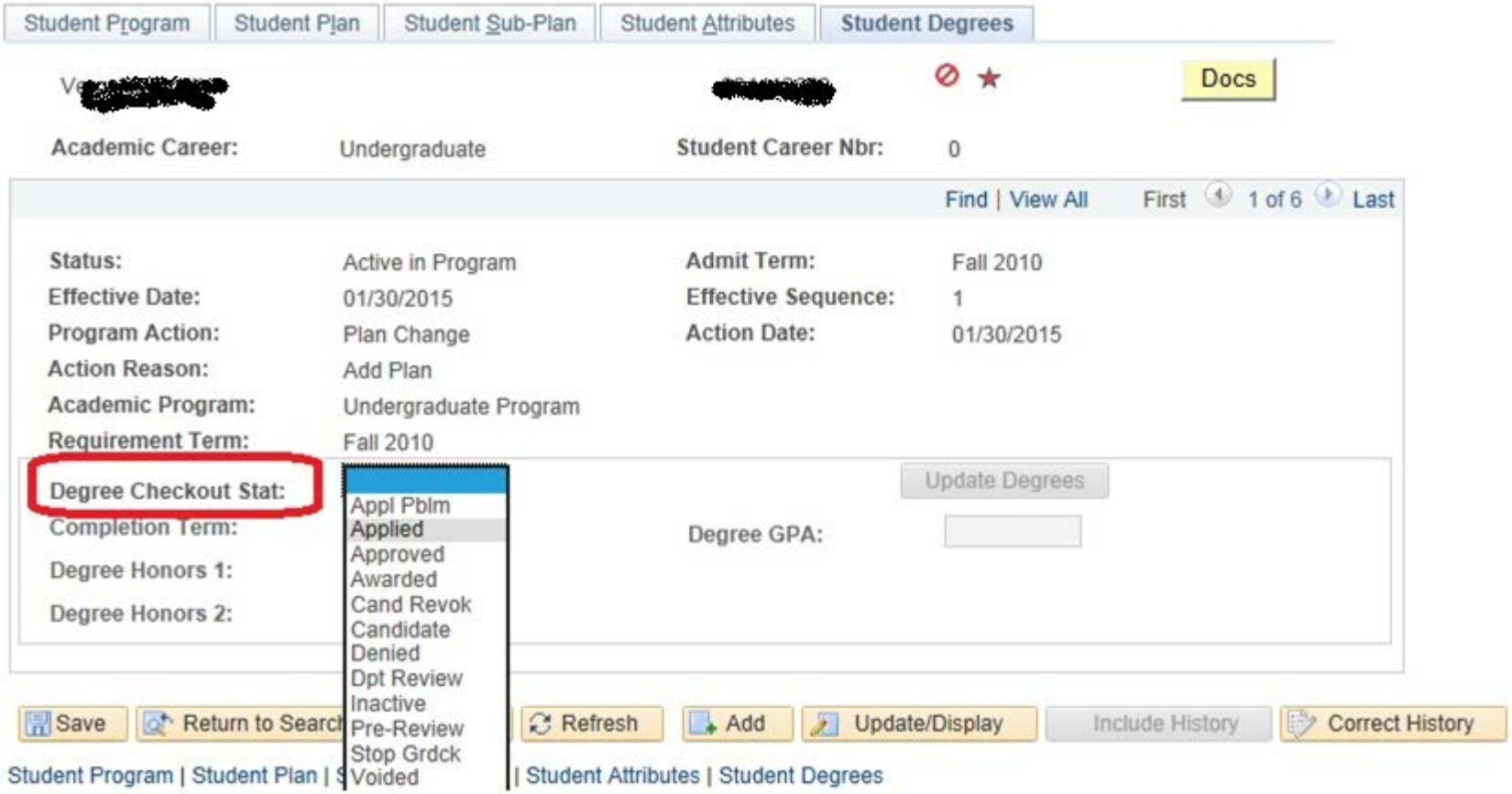
What's Needed… #2 Test Environment



Different Types of DW Testing

- Data Transformation Testing
- Data Completeness Testing
- Data Accuracy Testing
- Database Constraint Testing (including ‘NotNull’, ‘Unique’, ‘Primary Key’, ‘Foreign Key’ constraints)
- Regression Testing

Data Transformation Testing - Example



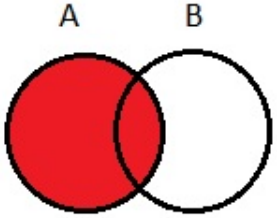
Data Transformation Testing – Example Contd.

Major (Latest)	Sex	Ethnicity	Underrepresented?	Units Earned	Units Attempted (current)	Current Term	Units Attempted (Future)	Future Term	Group Qual	Cumulative GPA	Degree Candidate Term	Degree Checkout Consideration Term	Degree Checkout Consideration Status	Academic Standing	EOP Participation	Freshman Programs Participation	Degree Audit Program
ENGL	WOMEN	Hispanic	Yes	167.0	11.0	Fall 2016	0.00	Spring 2017	EI 3300-3399	2.78	-	-	Pre-Review	Good Standing	No	No	BA ENGL
AMST	MEN	Hispanic	Yes	83.0	9.0	Fall 2016	0.00	Spring 2017	EI 3200-3299	1.92	-	-		Continued Probation	No	No	BA CMPH
ANTH	WOMEN	White	No	102.0	12.0	Fall 2016	0.00	Spring 2017	EI 3500-3599	2.77	-	-		Academic Disqual	No	No	BA ANTH
HIST	MEN	Hispanic	Yes	128.0	9.0	Fall 2016	0.00	Spring 2017	EI 3600-3699	2.34	-	-		Good Standing	No	No	BA HIST
PSYC	WOMEN	Black	Yes	126.0	5.0	Fall 2016	0.00	Spring 2017	EI 3300-3399	2.25	Fall 2016	Spring 2016	Approved	Good Standing	No	No	BA PSYC
CRJU	WOMEN	Hispanic	Yes	79.0	12.0	Fall 2016	0.00	Spring 2017	EI 3900 or more	3.18	-	-		Good Standing	No	No	BA CRJU
GEOG	WOMEN	White	No	185.0	9.0	Fall 2016	0.00	Spring 2017	EI 3500-3599	2.90	Fall 2016	Spring 2016	Approved	Good Standing	No	No	BA GEOG
POSC	WOMEN	Hispanic	Yes	72.0	15.0	Fall 2016	0.00	Spring 2017	EI 3000-3099	2.52	-	-		Academic Disqual	No	No	BA POSC
PSYC	MEN	Race/Ethnicity Unknown	No	112.0	5.0	Fall 2016	0.00	Spring 2017	EI 2900-	2.59	-	Fall 2016	Dpt Review	Good Standing	No	No	BA PSYC

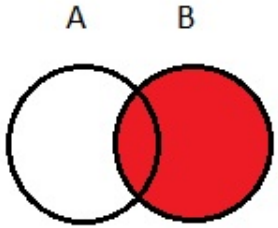
Data Completeness Testing

- Verify that all projected data is loaded without any data loss or termination
- Break down data by different variables and compare record counts
- Erroneous data join operation a common cause

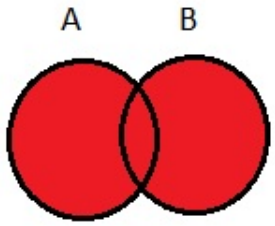
Data Completeness Testing – Example Database Join Operation



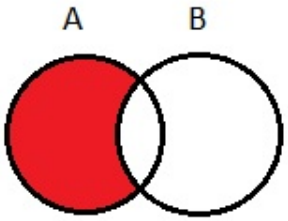
SELECT xxxx FROM A
LEFT JOIN B ON
A.key=B.key



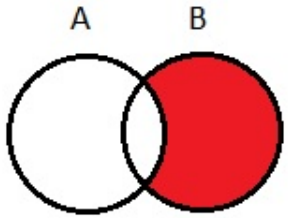
SELECT xxxx FROM A
RIGHT JOIN B ON
A.key=B.key



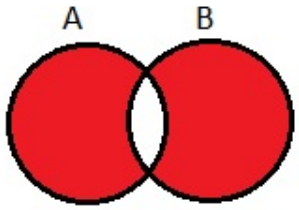
SELECT xxxx FROM A
FULL OUTER JOIN B ON
A.key=B.key



SELECT xxxx FROM A
LEFT JOIN B ON
A.key=B.key WHERE
B.key=NULL

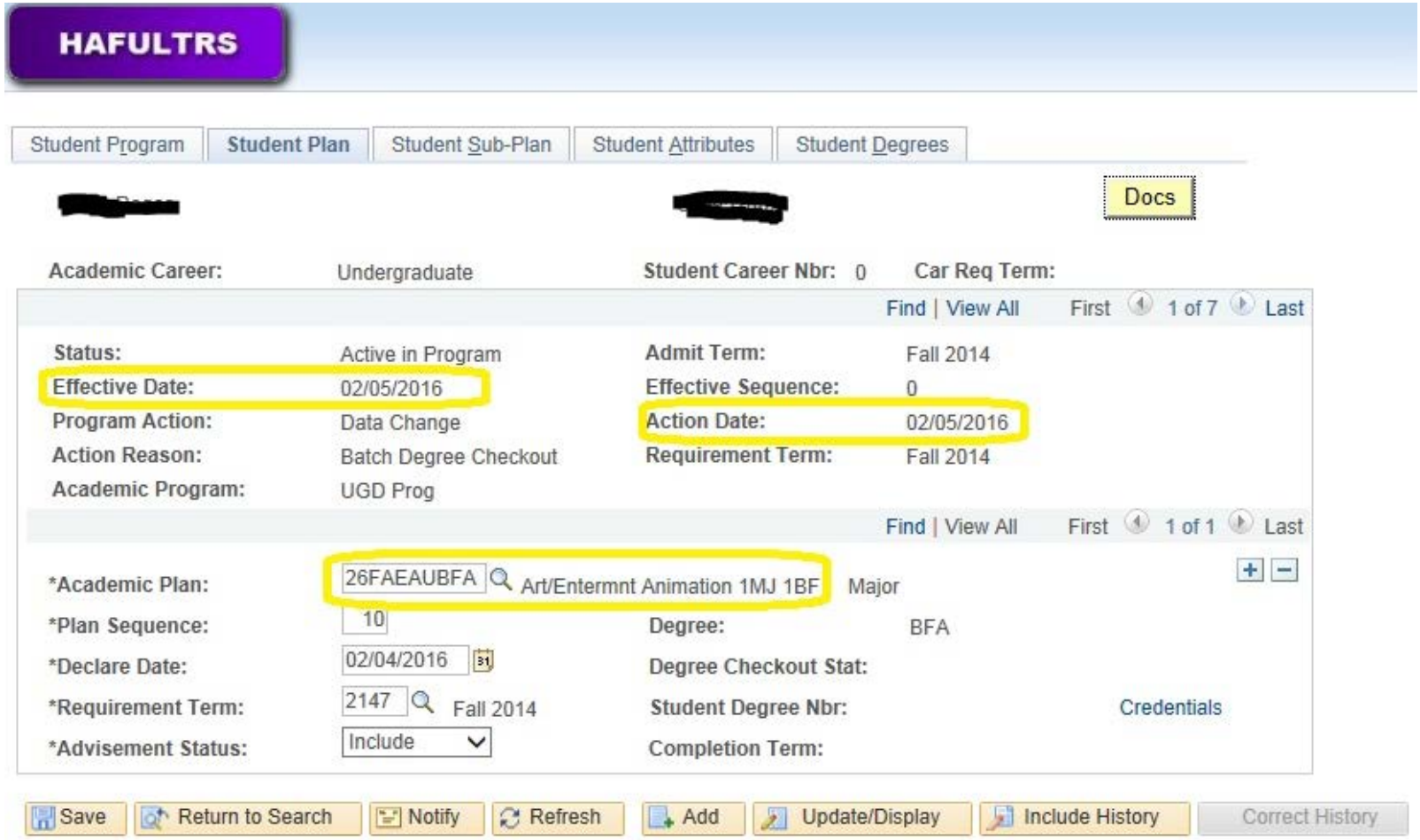


SELECT xxxx FROM A
RIGHT JOIN B on
A.key=B.key WHERE
A.key=NULL

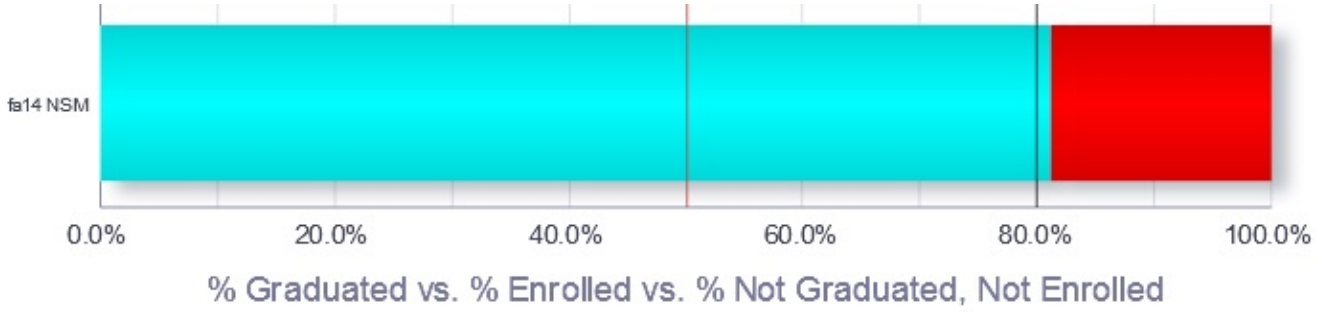


SELECT xxxx FROM A
FULL OUTER JOIN B ON
A.key=B.key WHERE
A.key=NULL or
B.key=NULL

Data Accuracy Testing
Example: effective dating



Data Accuracy Testing
Example: effective dating



Cohort Description	College (latest)	Size	Degree Count	% Graduated	Enrolled Count	% Enrolled	Not Graduated & Not Enrolled Count	% No
fa14	Arts	253	0	0.0%	222	87.7%	31	
	CBE	945	0	0.0%	811	85.8%	134	
	COM	365	0	0.0%	333	91.2%	32	
	ECS	646	0	0.0%	536	83.0%	110	
	HHD	540	0	0.0%	493	91.3%	47	
	HSS	745	0	0.0%	629	84.4%	116	
	MISC	362	0	0.0%	278	76.8%	84	
	NSM	387	0	0.0%	314	81.1%	73	
fa14 Total		**4,243**	**0**	**0.0%**	**3,616**	**85.2%**	**627**	

Save | Return to Search | Notify | Refresh | Add | Update/Display | Include History | Correct History

Regression Testing

- Verifies that software previously developed still functions correctly after changes were made to the product
- Goal is to catch unintended defects introduced when source code was updated
- Start with a number of test cases that verify basic functionalities of ETL
- After defects are fixed or enhancements are made, add corresponding test cases to test suite
- Tester needs to execute regression test cases prior to *every* release of the product

Final Thoughts

- ETL does more than changing data structure
- We covered data warehouse testing approaches. DW Quality Assurance, however, covers more than testing
- Whose job is it to test DW's data quality?
- Testing activities should start early
- Questions/Comments?