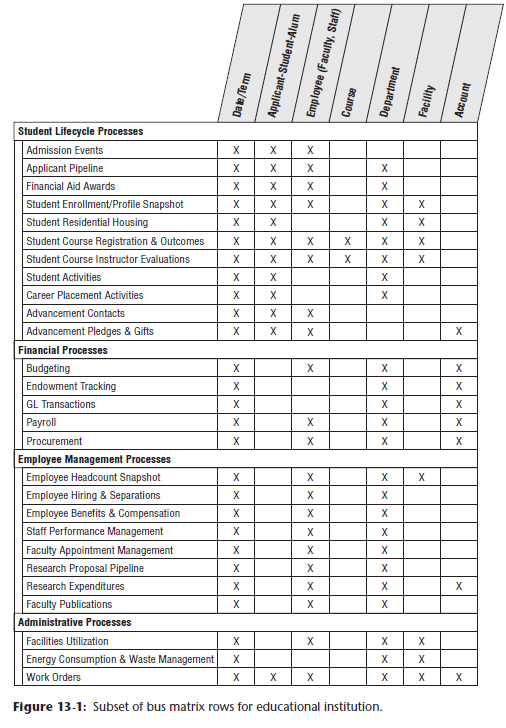
# Kimball Data Warehouse Toolkit

## Ch 13 – Education

* We step into the world of an educational institution in this chapter, looking first at the applicant pipeline as an **accumulating snapshot**
* When **accumulating snapshot fact tables** were introduced in Chapter 4: Inventory, a product movement pipeline illustrated the concept, + order fulfillment workflows were captured in an accumulating snapshot in Chapter 6: Order Management
* Now, rather than watching products/orders move through various states, an accumulating snapshot is used to monitor prospective student applicants as they progress through admissions milestones
* The other primary concept discussed in this chapter is the **fact-less fact table**
* Explore several case study illustrations drawn from higher education to further elaborate on these **special fact tables** and discuss **the analysis of events that didn’t occur**
* **Concepts:**
* **Bus matrix** snippet for a university or college
* Applicant tracking and research grant proposals as **accumulating snapshot fact tables**
* **Fact-less fact table** for admission events, course registration facilities management, + student attendance
* Handling of **nonexistent events**

### University Case Study and Bus Matrix

* Ex: Working for a university, college, or other type of educational institution
* Someone once remarked that running a university is akin to operating all the businesses needed to support a small village
* **Universities** = **simultaneously** a **real estate property management company** (residential student housing), **restaurant** w/ **multiple outlets** (dining halls), **retailer** (bookstore), **events management** and **ticketing agency** (athletics + speaker events), **police department** (campus security), **professional fundraiser** (alumni development), **consumer financial services company** (financial aid), **investment firm** (endowment management), **venture capitalist** (research + development), **job placement firm** (career planning), **construction company** (buildings + facilities maintenance), + **medical services provider** (health clinic)
* In addition to these varied functions, higher education institutions are obviously also focused on attracting high caliber students + talented faculty to create a robust educational environment.
* **Traditionally, there has been less focus on revenue and profit in higher education, but with ever-escalating costs and competition, universities and colleges cannot ignore these financial metrics**
* They want to attract + retain students who align w/ their academic + other institutional objectives
* There’s a strong interest in analyzing what students are “buying” in terms of courses each term + the associated academic outcomes
* Colleges + universities want to understand many aspects of the student’s experience, along with maintaining an ongoing relationship well beyond graduation
* The bus matrix snippet seen below covers several core processes within an educational institution

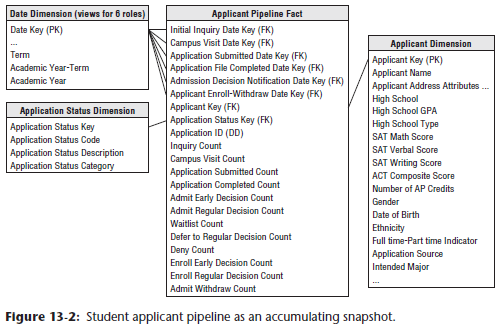


### Accumulating Snapshot Fact Tables

* Chapter 4 used an **accumulating snapshot fact table** to **track** products identified by serial or lot numbers as they move through various inventory stages in a warehouse.
* Recall the **distinguishing characteristics of an accumulating snapshot fact table**:
* A **single row represents the *complete* history of a workflow/pipeline instance**
* ***Multiple* dates represent the standard pipeline milestone events**
* The **accumulating snapshot facts often included metrics corresponding to *each* milestone**, + **status counts and elapsed durations**
* **Each row is *revisited* + *updated* whenever the pipeline instance changes**
* **Both FKs + measured facts may be changed during the fact row updates**

#### Applicant Pipeline

* Now envision these *same* accumulating snapshot characteristics as applied to the prospective student admissions pipeline
* For other industries, there are obvious similarities to tracking job applicants through the hiring process or sales prospects as they are qualified and become customers.
* In the case of applicant **tracking**, prospective students progress through a standard set of admissions hurdles or **milestones**
* Perhaps you’re interested in tracking activities around key dates, such as initial inquiry, campus visit, application submitted, application file completed, admissions decision notification, + enrolled or withdrawn
* ***At any point in time*, admissions + enrollment management analysts are interested in how many applicants are at each stage in the pipeline**
* **The process is much like a funnel, where many inquiries enter the pipeline, but far less progress through to the final stage**
* Admission personnel also would like to **analyze the applicant pool by a variety of characteristics**
* The **grain of the applicant pipeline accumulating snapshot is one row per prospective student**, which **represents the lowest level of detail captured when the prospect enters the pipeline**
* **As more information is collected while the prospective student progresses** toward application, acceptance, + enrollment, you **continue to revisit and update the fact table row**, as below:



* **Like earlier accumulating snapshots, there are *multiple* dates in the fact table corresponding to the standard milestone events**
* Want to **analyze a prospect’s progress by these dates to determine the pace of movement through the pipeline + spot bottlenecks**
* Especially important if you see a significant lag involving a candidate whom you’re interested in recruiting
* **Each of these dates is treated as a role-playing dimension, with a default surrogate key to handle the unknown dates for new and in-process rows**
* The **applicant dimension** contains many interesting **attributes about prospective students**
* **Analysts are interested in slicing + dicing by applicant characteristics** such as geography, incoming credentials (GPA, admissions test scores, AP credits, + high school), gender, DOB, ethnicity, preliminary major, application source, + a multitude of others
* **Analyzing these characteristics *at various stages of the pipeline* can help admissions personnel adjust their strategies to encourage more (or fewer) students to proceed to the next mile marker**
* The **facts in the applicant pipeline fact table include a variety of counts closely monitored by admissions personnel**
* *If available*, this table could include estimated probabilities that the prospect will apply + subsequently enroll if accepted to predict admission yields

##### Alternative Applicant Pipeline Schemas

* **Accumulating snapshots are appropriate for *short*-lived processes that *have a defined beginning and end*, with *standard intermediate milestones***
* This type of fact table **enables you to see an updated status + ultimately final disposition of each applicant**
* *However*, **because accumulating snapshot rows are *updated*, they do NOT preserve applicant counts + statuses at critical points in the admissions calendar**, such as the early decision notification date
* Given the close scrutiny of these numbers, **analysts might also want to retain snapshots at several important cut-off dates.**
* ***Alternatively*, could build an admission transaction fact table with 1 row per transaction per applicant for counting + period-to-period comparisons**

#### Research Grant Proposal Pipeline

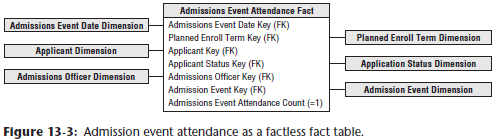
* The research proposal pipeline is another education-based example of an accumulating snapshot
* Faculty + administration are interested in **viewing the lifecycle of a grant proposal as it progresses through the pipeline** from preliminary proposal to grant approval and award receipt
* This **would support analysis of the number of outstanding proposals in each stage of the pipeline by faculty, department, research topic area, or research funding source**
* Likewise, you **could see success rates by various attributes**
* Having this information in a common repository would allow it to be leveraged by a broader university population

### Fact-less Fact Tables

* So far we’ve **largely designed fact tables with very similar structures** 🡪 Each one **typically has 5 to approximately 20 FK columns, followed by one to potentially several dozen numeric, continuously-valued, *preferably additive* facts**
* The **facts can be regarded as measurements taken at the intersection of the dimension key values**
* *From this perspective,* the **facts are the justification for the fact table, + the key values are simply administrative structure to identify the facts.**
* **There *are*, however, a number of business processes whose fact tables are similar to those we’ve been designing *with one major distinction*** 🡪 **NO measured facts!**
* Introduced fact-less fact tables while discussing promotion events in Chapter 3: Retail Sales, as well as in Chapter 6 to describe sales rep/customer assignments
* There are numerous examples of fact-less events in higher education.

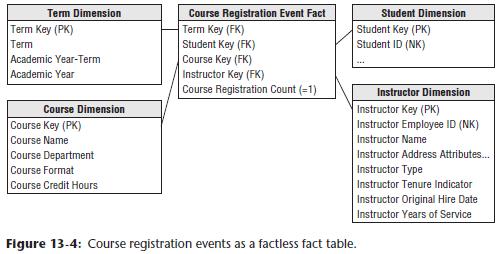
#### Admissions Events

* Can envision a fact-less fact table to **track each prospective student’s attendance at an admission event,** such as a high school visit, college fair, alumni interview or campus overnight, as below:



#### Course Registrations

* Similarly, you can **track student course registrations by term using a fact-less fact table**
* The **grain = 1 row for each registered course *by student and term***, as illustrated below



##### Term Dimension

* In this fact table, the **data is at the *term* level** rather than at the more typical calendar day, week, or month granularity
* The **term dimension** **still should conform to the calendar date dimension**
* i.e., Each date in the daily calendar dimension should identify the term (ex: Fall), term + academic year (ex: Fall 2013), *and* academic year (ex: 2013-2014)
* **Column labels + values *must* be identical for the attributes common to both the calendar date and term dimensions**

##### Student Dimension and Change Tracking

* The **student dimension is an expanded version of the applicant dimension** discussed earlier
* **Still want to retain some information garnered from the application process** (ex: geography, credentials, intended major) but **supplement it with on-campus info**, such as part-time or full-time status, residence, athletic involvement indicator, declared major, + class level status (ex: Senior)
* As discussed in Chapter 5: Procurement, you **could imagine placing some of these attributes in a type 4 mini-dimension because factions throughout the university are interested in tracking changes to them, especially for declared major, class level, + graduation attainment**
* People in administration and academia are keenly interested in academic progress + retention rates by class, school, department, + major
* Alternatively, **if there’s a strong demand to preserve *student profiles at the time of course registration*, plus filter + group by the students’ *current* characteristics,** **consider handling the student information as a type 7 SCD with dual student dimension keys in the fact table**, also described in Chapter 5
* The **surrogate student key would link to a dimension table with type 2 attributes, + the student’s durable identifier would link to a view of the complete student dimension containing only the current row for each student**

##### Artificial Count Metric

* **A fact table represents the robust set of many-to-many relationships among dimensions + records the *collision* of dimensions at a point in time and space**
* **The course registration fact table could be queried to answer a number of interesting questions regarding registration** for the college’s academic offerings, such as: Which students registered for which courses? How many declared engineering majors are taking an out-of-major finance course? How many students have registered for a given faculty member’s courses during the last 3 years? How many students have registered for more than 1 course from a given faculty member?
* The only **peculiarity** in these examples = **you don’t have a numeric fact tied to this registration data**
* As such, **analyses of this data will be based largely on counts**.
* **NOTE**: ***Events* are modeled as fact tables containing a series of keys, each representing a participating dimension in the event**
* **Event tables sometimes have *no* variable measurement facts associated with them + hence are called fact-less fact tables**
* **SQL for performing counts in this fact-less fact is *asymmetric* because of the absence of any facts**
* When counting the number of registrations for a faculty member, *any* key can be used as the argument to the COUNT function
* Ex: SELECT faculty, COUNT(term\_key)... GROUP BY faculty gives the simple count of the number of student registrations by faculty, subject to any constraints that may exist in a WHERE clause
* **An oddity of SQL is that you can count *any* key and still get the same answer because you are counting the number of keys that fly by the query, *not* their distinct values**
* You’d need to use a COUNT DISTINCT if you want to count the unique instances of a key rather than the number of keys encountered.
* **The inevitable confusion surrounding the SQL statement, although not a serious semantic problem, causes some designers to create an artificial implied fact, perhaps called course registration count (as opposed to “dummy”), that is *always* populated by the value 1**.
* **Although this fact does not add any information to the fact table, it makes SQL more readable**, such as:SELECT faculty, SUM(registration\_count)... GROUP BY faculty
* **At this point the table is no longer *strictly* fact-less, but “1” is nothing more than an artifact**
* The **SQL will be a bit cleaner + more expressive with the registration count**, + some BI query tools have an easier time constructing this query w/ a few simple user gestures
* **More important, if you build a summarized aggregate table above this fact table, you need a real column to roll up to meaningful aggregate registration counts**
* Finally, **if deploying to an OLAP cube, you typically include an explicit Count column (always equal to 1) for complex counts because dimension JOIN keys are *not* explicitly revealed in a cube**
* **If a *measurable* fact *does* surface during the design, it can be added to the schema, *assuming it is consistent with the grain of student course registrations by term***
* Ex: Could add tuition revenue, earned credit hours, + grade scores to this fact table, but then it’s no longer a fact-less fact table.

##### Multiple Course Instructors

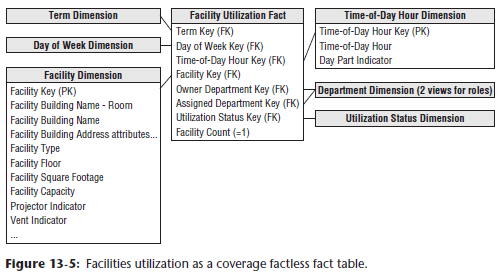
* If **courses are taught by a *single* instructor, you can associate an instructor key to the course registration events**, as shown in the prior fact-less schema
* *However*, **if some courses are co-taught, then it is a *dimension* attribute that takes on *multiple* values for the fact table’s declared grain**
* You have several options:
* **1) Alter the grain of the fact table** to be 1 row per instructor per course registration per student per term
* Although this would address the multiple instructors associated with a course, it’s an **unnatural granularity** that’d be **extremely prone to overstated registration count errors**
* **2) Add a bridge table** with an **instructor *group* key** in either the fact table or as an outrigger on the course dimension, as introduced in Chapter 8: CRM
* There would be 1 row in this table for each instructor who teaches courses on his own
* In addition, there would be 2 rows for each instructor team, + these rows would associate the same group key with individual instructor keys
* The **concatenation of the group key + instructor key would uniquely identify each bridge table row**
* As described in Chapter 10: Financial Services, you **could assign a weighting factor to each row in the bridge if the teaching workload allocation is clearly defined**
* Though **this approach would be susceptible to the potential overstatement issues surrounding the bridge table usage** described in Chapter 10
* **3) Concatenate the instructor names into a single, delimited attribute** on the course dimension, as discussed in Chapter 9: HR Management.
* Enables users to easily label reports with a single dimension attribute but **would NOT support analysis of registration events by instructor characteristics.**
* 4) If one of the instructors is identified as the **primary instructor**, then their instructor key could be handled as **a single FK in the fact table, joined to a dimension where attributes were prefaced with “primary” for differentiation**

#### Course Registration Periodic Snapshots

* The **grain of the fact table** **in the earlier fact-less schema is 1 row for each registered course by student and term**
* Some users at the college or university might be interested in **periodic snapshots** **of the course registration events at key academic calendar dates**, such as preregistration, start of term, course drop/add deadline, + end of term
* ***In this case,* the fact table’s grain would be 1 row for each student’s registered courses for a term *per snapshot date***

#### Facility Utilization

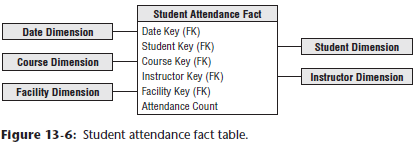
* **2nd type of fact-less fact table deals with coverage**, which can be illustrated with a facilities management scenario
* Universities invest a tremendous amount of capital in their physical plant + facilities, so it’d be helpful to understand which facilities were being used for what purpose during every hour of the day during each term
* Ex: Which facilities were used most heavily? What was the average occupancy rate of the facilities as a function of time of day? Does utilization drop off significantly on Fridays when no one wants to attend (or teach) classes?
* Again, the fact-less fact table comes to the rescue, + in this case you’d insert **1 row in the fact table for each facility for standard hourly time blocks during each day of the week during a term regardless of whether the facility is being used**



* The facility dimension would include all types of descriptive attributes about the facility, such as the building, facility type (ex: classroom, lab, office), square footage, capacity, + amenities (ex: built-in projector or whiteboard)
* The utilization status dimension would include a text descriptor with values of Available or Utilized.
* Meanwhile, multiple organizations may be involved in facilities utilization
* Ex: One organization might own the facility during a time block, but the same or a different organization might be assigned as the facility user

#### Student Attendance

* Can visualize a similar schema as above to **track student attendance** in a course
* In this case, **grain = 1 row for each student who walks through a course’s classroom door each day**
* This fact-less fact table would share a number of the same dimensions present in registration events
* **Primary difference = the granularity is by calendar date in this schema rather than merely term**
* This dimensional model (below) allows business users to answer questions concerning: Which courses were the most heavily attended? Which courses suffered least attendance attrition over the term? Which students attended which courses? Which faculty member taught the most students?



##### Explicit Rows for What Didn’t Happen

* Perhaps people are interested in monitoring students who were registered for a course but didn’t show up
* In this example you can **envision adding explicit rows to the fact table for attendance events that *didn’t occur***
* **The fact table would no longer be fact-less, as there is an attendance metric equal to either 1 or 0**
* **Adding rows is viable in this scenario because non-attendance events have the *same exact dimensionality* as the attendance events**
* Likewise, the fact table won’t grow at an alarming rate, presuming (or hoping) the no-shows are a small % of the total students registered for a course
* Although this approach is **reasonable in *this* scenario, creating rows for events that didn’t happen is ridiculous in many other situations**, such as adding rows to a customer’s sales transaction for
* promoted products that weren’t purchased by the customer.

##### What Didn’t Happen with Multidimensional OLAP

* **Multidimensional OLAP databases do an excellent job of helping users understand what didn’t happen**
* **When the cube is constructed, multidimensional databases handle the sparsity of the transaction data while minimizing the overhead burden of storing explicit 0’s**
* **As such, at least for fact cubes that are not too sparse, the event and nonevent data is available for user analysis while reducing some of the complexities just discussed in the relational star schema world**

### More Educational Analytic Opportunities

* Many business processes described earlier, such as procurement and HR, are obviously applicable to the university environment, given the desire to better monitor and manage costs
* Research grants + alumni contributions are key sources of **revenue**, in addition to the tuition revenue
* Research grant analysis is often a variation of **financial analysis** (Chapter 7: Accounting), but at a lower level of detail, much like a **subledger**
* The grain would include additional dimensions to further describe the research grant, such as the corporate or governmental funding source, research topic, grant duration, + faculty investigator
* There is a strong need to better understand + manage the budgeted + actual spending associated with each research project
* The **objective is to optimize the spending so a surplus or deficit situation is avoided**, **+ funds are deployed where they will be most productive**
* Likewise, understanding research spending rolled up by various dimensions is necessary to ensure proper institutional control of such monies
* Better understanding the university’s alumni is much like better understanding a customer base, as described in Chapter 8
* Obviously, there’re many interesting characteristics that would be helpful in maintaining a relationship with your alumni, such as geographic, demographic, employment, interests, + behavioral information, in addition to the data you collected about them as students (ex: affiliations, residential housing, school, major, length of time to graduate, + honors designations)
* Improved access to a broad range of attributes about the alumni population would allow the institution to better target messages and allocate resources
* In addition to alumni contributions, alumni relationships can be leveraged for potential recruiting, job placement, and research opportunities
* **To this end, a robust CRM operational system should track all the touch points with alumni to capture meaningful data for the DW/BI analytic platform**