Before the meeting, please review the following, a set of readings compiled by Berkeley's EDW team, D-Lab staff, and others:

- Big data vendors should stop dising data warehouse systems by Wayne Eckerson
- ETL for America by Dave Guarino
- ETL vs ELT: We Posit, You Judge by David Friedland

Optionally, have a look at (browse) the CalAnswers and EDW sites, to get a sense of what the EDW currently offers to the campus.

Participating:

- Steve Masover (RIT)
- Patrick Schmitz (RIT)
- David Greenbaum (RIT)
- Aaron Culich (RIT)
- Aron Roberts (RIT)
- Richard Millet (RIT)
- Chris Hoffman (RIT)
- Mark Chiang (EDW)
- Jonathan Lye -- Research administration and compliance (RAC)
- Dan Percival (RAC)
- Neil Maxwell (RAC)
- Ken Geis (RAC)
- Nancy Han (RAC)
- Dav Clark (DLab)

Mark’s team: EDW / CalPlanning / BAIRS. Hyperion. Mark passed around the diagram on the home page of CalAnswers (click thumbnail below to embiggen):

Data sources that feed EDW: HCM, CARS, student systems. Transform with Informatica. Load into Oracle db w/ star schema. Oracle BI (CalAnswers) is end-user portal.

Walking through how to deal with integrating new source system or other major new work:

- a long process
- ingest looks at business use/scope/need
- goes onto a long list of things that the requestors want, generally ‘yesterday’

Analysis / engineering that goes into something that reaches the top of the list:

- first analysis, SMEs
- data analysis, metadata layer, Oracle BI - "RPD" - semantic layer. maps between db and what end user will see in the report -- hides the db layer / field names / star schema joins / etc. from end-user, who is generally not a database person, doesn't know anything about joins
- business process / definition / governance are the ‘real’ problems, hardest to solve. things users argue about, generate distrust of db -- lots of "my way is the right way"
- the "headcount" example -- in a university where many people hold multiple jobs (e.g., Chancellor and Prof. of History), how are they counted in reports that want to count a person only once? (after the meeting addition from Steve: the definition of how "primary appointment" among multiple jobs held is defined from original rollout of HR-BAIRS; this definition is still in use, or something close to it, as far as SJM knows)

Mark: typically not a lot of published history or rational about why decisions are made -- there are definitions, but not so much the rationales -- wiki and JIRA is where that stuff might be surfaced

Dav: there's deep and careful thought that goes into these definitions, how to answer questions correctly. It would be helpful for people to be able to access that: why is this the correct definition/answer.

Patrick: Consider difference of high end tools used by EDW team vs. open-source ETL tools? Open-source worked for CollectionSpace, but what does Mark know about comparisons between.
Mark: Familiar with Informatica, DataStage (IBM), Talend. Haven't worked with/looked at open-source. In contexts I've worked, need is to run every night, need bulletproof tools and processes.

Patrick: Our development (e.g., of CollectionSpace) predicated on funder requirement that what we use be open-source. So tools with high license cost were immediately off the table. This becomes an issue too when sharing across campuses in research contexts come into play.

Mark: Worth noting that Informatica in use at multiple campuses, including multiple UCs.

Patrick: Question of loading data produced by technical teams to EDW specs, modeled by folks other than the EDW team -- streamlined w/o a lot of the analysis workflow that comes into play for administrative data sets.?

Neil: RAC wanted to do this too. Didn't make it into EDW queue at a high enough level, so we're looking elsewhere.

Mark: Good news is that we're currently wrapping up a number of projects. Can start conversations soon. We know that Tableau is in use across the campus.

Neil: RAC does a lot of reporting, maybe that's why Mark and team aren't hearing about it: requests come to RAC not EDW.

David: Is it realistic to add other front ends to the EDW?

Mark: Definitely an option. OBIE great for standard reports. Tableau better / faster for ad hoc.

Ken: Tableau is faster because it engages people on a different level. There's a desktop product that you can build a scaled down version of a report on a portion of data; then you can scale up to the complete data set and publication to world.

Patrick: Possible to decompose EDW services, to offer pieces of expertise that are not full-service, one-stop-shop?

Mark: Possible, sure. Reporting on ODS, APIs drawing from ODS is where this is happening. (ODS = Operational Data Store)

Ken: Tableau can't sit atop APIs, unfortunately. And what you don't get when using a tool that does not integrate well with EDW or APIs is benefit of all the work done defining metadata, work that can't easily be pulled into Tableau.

Patrick: So you get benefit of full stack if you use the full stack. That's not necessarily a bad thing.

Discussion of why put research into a campus EDW at all, as opposed to national repository infrastructures maintained by domain-specific organizations ...

o e.g., is there an onramping role for campus infrastructure ...

o and what of the many gaps in domain-centric repositories

o is there too great a gap between domains of administration and requirements around enterprise data; and the kinds of issues and processes that surround research data sets? ... and even if so, is there benefit to leveraging the technical infrastructure and expertise to support multiple different processes/requirements

DAG: What about EDW work at systemwide level?

Mark: UC Path. Date for go-live pushed out on this project, was supposed to happen this year, now looking at 2017.

DAG: Research? Student systems?

Mark: Hard across campuses for student systems: unique models/processes. Working first to try procurement, chart of accounts, other financial data, across campuses, likely based on a model that has been under development at UCLA for a couple of years.

Chris: Does EDW accommodate different user needs or always drive to common definitions

Mark: We make accommodations. Similar but slightly different fields given different names and plenty of documentation -- where needed.

Patrick: Excess capacity in technical (as opposed to organizational) infrastructure?

Mark: Again, ODS. Example: Bill Allison / Tamer Sakr API needs around HR data a current area where we're looking at exposing data through ODS + APIs.

[Winding down]

Patrick: How do we continue discussion -- if it makes sense to pursue -- about whether and where there is capacity, opportunity? Fit/gap?

Chris: Might lessons learned in EDW's history on this campus be described to researchers in a way that helps researchers as we move into a world of better-managed research data.

David: Good suggestion, and bring data owners who may not have seen value in EDW process 5-7 years ago, but now are calling for greater investment because the value has become apparent.

Patrick: How to inject this into research workflow, instruction, methods?

Dav: Researchers don't want to learn about technology until it's (often) later than would have been ideal, no different from administrators' history with EDW. Methodology classes (for grad students) are a place where discussion of these questions can be injected into the early phases of research projects.

Aaron: what governments are doing ....

[various examples: Paul Wadell in SF; Raymond Yee re: opendata.gov; etc.]

Dav: Researchers going to DMP are interested in EDW's lessons-learned.....
Patrick: Let's continue to explore through individual discussions......