Paper 122-31

Litigation Data Analysis/Auditing/ETL for BI: Triples Separated at Birth

Nicholson Warman, Peopleclick Inc., Raleigh, NC

ABSTRACT

This paper shows how the steps in Litigation Data Analysis (hereafter Litigation Analysis), Auditing and Extract-Transform-Load for Business Intelligence (ETL for BI, or data warehousing for business data analysis) are inherently similar; it is just the emphasis and time spent on each step that varies, based on the participant's goal. This is due to the commonality of the underlying principles of "Truth," Reproducibility and Time-Invariance (also seen as Time-Invariant). If the process is flawed, or if the data or processes are "out of sync," one with the other, then any subsequent decision based on that data or process is without merit. The paper will also cover the issue of time series data, and some of the problems it poses for each function. With more and more companies needing to meet the IS Audit Standard SAS 70 (and any/all of the other 90+ general plus numerous specialty) audit standards, with ever increasing numbers of lawsuits, and continuing pressure to get the "right data" at "the right time" for decision making, you need to get to the fundamentals or you truly are "betting the firm" that you are ready for anything from a new data warehouse through an audit to presenting your corporate defense in court.

INTRODUCTION

Litigation data analysis, auditing and data warehousing have long been considered three separate and distinct functions. In reality, while the focus/intent is different when compared one to the other, the three share many similar functions and steps. This paper relates the life cycle of the three activities, showing how they are, in truth, "triplets separated at birth." Likewise, time-series data play an important part in all our triplets.

To explore our triplets in more full measure, we'll start with the characteristics common to each of our triplets then examine their specifics. The underpinnings that apply to all stages of the lifecycle for each of our triplets will then be explored, focusing on "Truth," Reproducibility and Time-Invariance. Then we'll look more carefully at the Extract, Transform, Load, Anomalies Resolution and Analysis stages, in general then for each lifecycle. We'll then explore Time Series as an issue that applies to all three of our triplets.

DATA WAREHOUSING, AUDITING AND LITIGATION ANALYSIS, OH MY!

I'd like to briefly cover the major steps in each of these three efforts, so we can start by seeing the commonality between the three. What all have in common is the need to:

- Plan for the activity to be conducted, examining standards, policies and procedures that are relevant, and any chartering materiel (such as the nature of the complaint before the courts, etc.)
- Select the resources to work on the activity (if a permanent team not already in place)
- Get data from the appropriate source(s)
- Prepare datasets/tables for subsequent analysis
- Identify data records that have missing/invalid values, either to be suppressed from the created datasets and subsequent analysis, or as the subject of further analysis
- Conduct the analyses needed, be it the User using the data warehouse in support of their daily work, the Auditor investigating variances from procedure and standards, or the Litigation Analyst looking for support for or argument(s) against the complaint being studied

THE DATA WAREHOUSING LIFECYCLE

The general lifecycle is:

- Define the scope of the data warehouse and the data sources for inclusion in the data warehouse
- Assign the appropriate resources (if any needed) to supplement the data warehousing team
- Define the frequency, processes and techniques for the data extraction and loading into the data warehouse
- State the method of identifying data that does not conform, with notification processes and actions to be taken in case of discrepancies
- Develop and implement the data warehouse, and document any future enhancements to be addressed

THE AUDIT LIFECYCLE

The general lifecycle is:

- Define the scope of the audit
- Assign the appropriate resources to the audit (if not a permanent team for this audit)
- Define the frequency, processes and techniques for the audit
- State the method of recording and reporting audit findings, particularly any nonconforming issues
- Develop and implement corrective action(s) arising from the audit findings

THE LITIGATION ANALYSIS LIFECYCLE

The general lifecycle is:

- Define the scope of the analytical studies to be performed, and the type of data to allow that analysis
- Assign the appropriate resources to the analysis team
- Build analytical datasets from the supplied data
- Discuss with counsel examples of nonconforming data to gain feedback from the data provider on the explanation/understanding of the anomalous findings
- Create the documentation of the final results, including final copies of associated programs and outputs, from initial data reading through final analytical studies, plus the final report to be supplied to the court

Let's now look at some basics, common to all our triplets, before we drill down into the respective lifecycles of our triplets.

"TRUTH" - WHATEVER THAT MEANS

The following quotes are from The American Heritage® Dictionary of the English Language:

- "*Truth* is a comprehensive term that in all of its nuances implies accuracy and honesty."
- "*Veracity* is adherence to the truth."

How then does one reconcile staff following policy and procedure, yet an audit delivering negative audit findings? Why is it that a data warehouse can show results that differ from operational reality? Why, in a legal case, where all parties are working from the same data (and perspective), do both sides reach diametrically different results?

One explanation is a bad definition of the problem or results (think two sides of the same coin, seen from just one side and without knowledge of the other side). Another is timing and completeness: if the data used are not complete for the most recent period, then any analysis, regardless of good faith, will render incorrect results. Another problem is one of scope. As an example, if the majority of a company's business is conducted outside the country, exchange rates will change results, month over month, even if the actual dollar value is unchanged.

What is required in a litigation setting, and is expected in audit and data warehousing environments, is complete disclosure of relevant material and assisting elements (think footnotes and the like) to explain variances. Otherwise, the consumer of these results, be it a court or senior management or an investor, would be misled into believing that the information provided is complete, accurate and a fair representation of reality. If, as in one of the previous explanations, the current period data are incomplete, then all consumers of that information need to be informed that current number are approximate and should be used with caution. Or better still, that period of data should not be made available at all, until it IS complete and accurate.

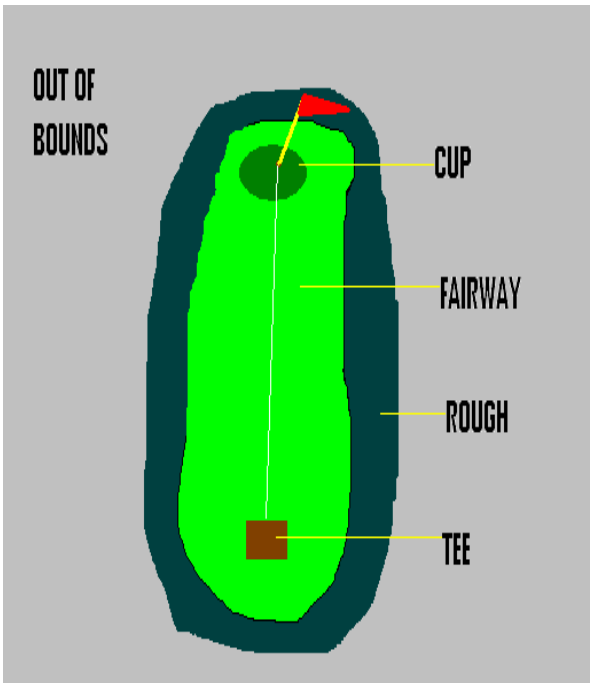
Another example of this problem is most quickly seen in financial data. If it takes six weeks to post payments to the general ledger, then only when that period is complete (that is, six weeks after the period has ended) should the data be extracted for any purpose. Until that period is complete, it would be highly misleading to report the information on that quarter. This issue of completeness is common to our triplets, as incomplete information can do more harm than good.

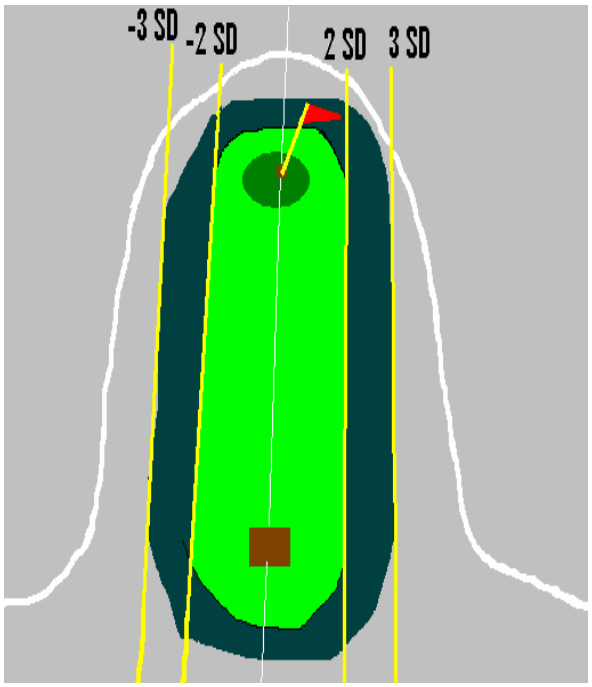
ACCURACY

This means, at its heart, addressing all possibilities that may impinge on the situation at hand. Simplifying assumptions are one such cause of breaks with the TRUTH, and typically will cause expending more time and effort than having grappled with the issue being avoided by adopting the simplifying assumption in the first place. Exploring only those factors that support your point of view does not represent an accurate analysis. Rather, an accurate review requires looking at, at a minimum, your and the opposing position on any analysis.

Accuracy means preserving the context of the data, and not varying from the definitions/intentions/meanings of the originator, legislation, or accepted practices.

Accuracy also means using the appropriate measures for the situation. To cope with this issue, courts have mandated Normal-Equivalent Standard Deviations (or NESDs) as one of the statistical measurements presented, although it is always discussed as simply Standard Deviations (SDs). While a Poisson distribution may best fit the data, and therefore be the more accurate statistical measure, the courts need a standard for comparison and the standard normal distribution is that standard. SAS® functions such as `PROBIT(your_pValue_goes_here)` allow the conversion of a non-normal distribution to a normal distribution, for the appropriate presentation of SDs. In order to show discrimination or bias, the courts have identified that statistical disparities in excess of two to three SDs are needed. If your results are within two to three SDs (< ±2-3 SDs), random chance may be the cause of the observed disparities.





Since not everyone has a Ph.D. in Statistics, let me explain the "statistical disparities in excess of two to three Standard Deviations (SDs) rule." Since golf is so significant in North Carolina, and The US Open Golf Tournament was recently held in Pinehurst (near Raleigh) in North Carolina, a golf analogy is in order. In golf, you have the fairway, with the tee at one end and the green (with the cup) at the other end. The goal is to get your ball from the tee to the cup in the shortest number of strokes (hits at the ball), the winner being the person with the lowest score (sum of each player's strokes) after playing 18 holes. Think of this shortest path as the Mean of your distribution, and assume that the fairway is equally wide on either side of this path, so you have a normal distribution. Further, the fairway is bounded by the Rough, and further out, the Rough is bounded by "Out of Bounds." If your ball lands in the Rough, you will likely expend extra effort or even incur additional strokes to achieve the cup for that hole. Think of the Rough as the two to three SDs range of your distribution; you aren't out of the game, but you are on the losing side of the argument. If your ball is Out of Bounds (that is, in excess of three SDs), you take a one stroke penalty and start with a new ball from the point where you lost the ball; your case is worsening. In the Rough or Out of Bounds does not mean you have lost the game (or your case), but you have lost ground against opposing counsel/player(s), and if this occurs frequently enough, you have lost the game/your case. Second place isn't good enough in this instance, as there are only two players, you and opposing counsel. We won't discuss the fact that my golf game is only rivaled by my artistry.

This "in excess of two to three standard deviations" rule is not as simple as it first appears, due to the nature of the definition of standard deviations. When working with small populations, no matter how egregious the appearance of discriminatory actions, it may not be possible to reach to the level of two to three SDs, as there are insufficient data points to give a strong enough indication of bias, and therefore, you cannot reach let alone exceed two standard deviations.

DEFINITIONS, INTENTIONS, MEANINGS

Lewis Carroll wrote in 1871 in *Through the Looking-Glass (and, What Alice Found There)*, "When I use a word, … it means just what I choose it to mean -- neither more nor less." However, for our purposes, it is of paramount importance that everyone agrees on the definitions, usages and meanings, and has good intentions regarding the words used.

The following example has been modified to protect the guilty. It highlights the problems caused by breaking the rules associated with "… comprehensive … accuracy …". A certain company's data warehouse was built for a single office/territory. Management liked it so much that they decreed that all branches, offices, and territories would now use it. They operated on multiple continents, and each branch/district/territory input their historical and current information into this common data warehouse, for information sharing purposes. Unfortunately, the units of measure in use varied by responding branch/district/territory, and were not recorded in the data warehouse. So you had liters and quarts from the European offices mixed with gallons and quarts from the US offices with grams and kilograms from the Asia-Pacific Rim branches; the quarts were Imperial measure in the UK but American quarts in the US, as were the gallons. The company had to launch a very expensive project to introduce the units of measure used for this ONE field in their data warehouse, with a conversion routine from any unit of measure to any other unit of measure. For the users in the originating offices, it wasn't an issue, since they were already familiar with the data. The issue

came from looking at information from outside the originating district/territory. Fortunately, other information on file allowed a computer program to determine the implicit units of measure for the records, saving the data warehouse from becoming but another failed footnote in data warehousing history. Clearly the answer " … but everyone knows …" didn't work there and won't work for you either.

One reason for so many standards in the auditing arena is to ensure common language and processes, so that all participants are working in a consistent fashion. Confusion over the definitions, intentions and meanings cannot be tolerated.

In Litigation Analysis, if there is incomplete agreement on the definitions, intentions and/or meanings of terms used in the case, there can be no effective communication. This is the reason that at every stage of the process, questions seeking clarification are asked of opposing counsel and assumptions made are clearly documented, for current or future resolution. This is also one of several reasons that expert witnesses are deposed (interviewed by opposing counsel) as one of the stages in a case moving from a complaint through a court.

SAS® is so popular in data warehousing and litigation analysis studies because of the robustness of the SAS environment's statistical functions and procedures, and the clarity or openness of the SAS language used in the various programs to reach the presented results. These programs are unambiguous (even if the results of their analyses are sometimes less revealing), and the same program used with the same data will render the same result, unless `RANUNI` or other randomness generation functions are included in the process (and even then, if a constant seed is used, you will get about the same results). The definition for the variables used, and the meaning of the various statistics are clear and consistent. Datasets are self-documenting (to a degree), and when matched to the creating program, the intention and meaning should be easily determined, if but only with the occasional comment to make the meaning clear.

REPRODUCIBILITY AND THE RUN BOOK

This is fundamentally the concept that the same program run against the same data will yield the same result; this is also called Repeatability. It also means that another person (vis. opposing counsel's expert witness) can be given the data and the programs, and produce the same result. This does not speak to the "truth" of the data or the analysis process used, but is key to ensuring confidence in any analytical study. As a scientific principle, if the same procedure is followed, given the same starting point, you should reach the same conclusion. This presumes that you have been provided the correct data and all of the program code used. If not, the results being presented are suspect, and many expert witnesses have had their credibility compromised by such process errors. If there are too many such challenges to the expert witness in a case, their "evidence" may be suppressed under a Daubert[1] motion.

For those who lived through the punch card days, in order to put a program into production, you needed to supply a Run Book, so the operators knew what conditions had to be met before running each of your programs, and in what order they were to be run. Also included would be how to recover should the process fail, who to notify, and any other special handling that would be required. In some cases, disk packs, tape, printer chains and even the paper stock required would be documented in the Run Book.

An example of what a Run Book might look like, borrowing from the litigation analysis arena:

Hierarchy of processing					*external source datasets*
Inputs	**CD #**	**Order**	**Program**	**Outputs**	**Source is**
In1	1	1	ReadMetrics.sas	location_state.sas7bdat jobtitle_jobeeo.sas7bdat	Counsel - Original data
In3	10			supp_data.sas7bdat	Counsel - Original data
In2.mdb	1	2	ReadRound1.sas	job_groupings.sas7bdat	Counsel - Original data
				paygrades.sas7bdat	
				misc_codings.sas7bdat	
In2.xls	3	3	ReadOrgChart.sas	orgchart_structure.sas7bdat	Counsel - Original data
In4	2	4	ReadREF1.sas	REF1_*.sas7bdat	Counsel - Original data
Stuff.txt	1	5	ReadREF2.sas	REF2_*.sas7bdat	Counsel - Original data
jobtitle_jobeeo.sas7bdat	5	6	JobsByOrgUnit.sas	orgStruct_jobs.sas7bdat	Step 2
stock_recoding_inclusion.inc	4				Included data steps to recode data acc. Counsel Direction 7/4/2000
orgchart_structure.sas7bdat	4				Step 3
orgStruct_jobs.sas7bdat	11	7	furtherExample.sas		Step 6
paygrades.sas7bdat	4				Step 2

The rest of the Run Book runs for another 10-15 pages. As you can see, it becomes possible to reconstruct the work done, step by step in this fashion, leaving no doubt about what is done when, where and in what order. Should time dependences become an issue, this run book arbitrates all such ambiguities (start at the beginning, and when you reach the end, stop). CD# indicates the physical CD on which the input file(s) are (to be) stored.

If you think this is too much trouble to go to, understand that the Run Book spreadsheet may:

1. be generated by a SAS program (not available for distribution – easily developed, however);
2. include narrative that may only be available in metadata stores or in correspondence with counsel (or other relevant parties);
3. be the only way to wrap one's mind around the complexity of the case, given the hundreds of programs, data files, inclusions, spreadsheets (both as data and as output) and assorted other output files;
4. be documenting a situation that takes 10-20 CDs for a single copy of all these files, with some files having to be archived/zipped in order to fit on a single CD;
5. be also used to create the set of CDs, assigning files to the next CD that has the capacity to hold the file. Folder information should be included unless folder references are sufficient clear that such labeling will not be required by the recipient. The program to do this assignment is not difficult to write, but of necessity is custom to your individual situation and the CD/DVD burning environment you use.

Given the complexity of this not-atypical example, the chance of making a mechanical error is too great to risk taking an "easier" approach. The loss of credibility inherent in the making of such a mistake is far more expensive than the time it takes to "do things right."

Experience dictates that, for large projects, realize that work may be parceled out to various analysts on a team. If each team member makes copies of some files, or maps shared directories with different drive letters, it makes matching the programs to their correct data input nearly impossible. That is why, without this Run Book, you may find multiple copies of the (same?) file in different locations with the same name. Without the context provided by the Run Book, you will have to match files together on a hit-or-miss basis until success. By keeping the same-name files on different CDs, the above process eliminates the confusion of which file belongs with which program.

TIME-INVARIANCE

This is a standard requirement and even part of the definition of dimensional data warehousing. It likewise is fundamental to auditing and litigation data analysis. Time-Invariate means that once that period's data are loaded into the data warehouse, they will NEVER be changed. This is true in litigation analysis as well, as you cannot report to the court today that the situation is X and next week, on purportedly-the-same data, report that the situation is Y without some very strong arguments to explain this change. Likewise, think of the auditor who today informs senior management that there is a problem, and tomorrow says that there really was no problem. Credibility would suffer. "Close enough" only counts in horseshoes and atom bombs, but has no place where trust is involved.

Time-Invariance means that the data do not change over time from one use to another, for the same period/extract. This means that the data have been "snapshot", meaning that the data are frozen in time and will never be modified. People are most familiar with this concept as annual reports, where the previous year's performance is documented, and compared to the previous year. While restatements and modifications do occur to the previous year's data, it is cause for concern and rarely should be used. This is true for all endeavors, where trust and dependability are required.

In a typical data warehouse, only once data have been prepared and meet all requirements are they added to the data warehouse and made available to its user community. That way, March data today are March data in six months' time, and the results of a particular query today will be identical in six months' time, based on the same data. The same concept applies in auditing and litigation analysis.

"TRUTH", REPRODUCIBILITY AND TIME –INVARIANCE IN OUR TRIPLETS

We'll now look at these topics are they related individually to our triplets.

THE DATA WAREHOUSING LIFECYCLE

Completeness is a problem very frequently encountered, given the general demand to get the data immediately, even though it is incomplete or inaccurate in some manner. We can update/fix/restate it later is a sentiment that has no place in data warehousing; that approach applies only to reporting environments. Completeness is at times a problematic concept for data warehousing, as you need to account for all records, but a data warehouse should not (normally) contain transactional-level data. Data warehouses should be the repository of aggregated data that reconciles to the transactional data for the same period, but needs not reproduce the individual data from the transactional data. Only in one-off cases should there be a one-to-one match between the data warehouse and the transactional data system (that is, a single transactional record of a particular type for that reporting period; the aggregation would also be that record).

The Run Book concept is equally important for data warehousing, as you cannot have credibility when the analyses produced by your users cannot be reproduced from the same data. And imagine using what is supposed to be time-Invariant data for a report, yet getting different results each time the report is produced! The implications are that extracts and the subsequent transformation and loading phases are executed in a particular order, either by date or some triggering event such as a successful run of the GL application. Until the pre-conditions have been met, and the data has met certain quality standards, the data must not be loaded into the data warehouse. The Run Book has been replaced in most if not all data warehousing operations by scheduling software running on the various source data systems, although there is still a human element of when the data are "done" and ready for extraction and loading into the data warehouse.

Since the time dimension is one of the most critical in data warehousing, time-Invariance is an essential characteristic. The time and cost to load and organize data in a data warehouse is non-trivial, so it neither should be done frequently nor should it be done if the data are not ready. Dependability and credibility are necessary if the investment in the data warehouse is to be warranted. Business Intelligence, based on faulty data, does more harm than good, unless the data are unchanged over time. If the data can change (that is, they can vary over time), then you have a reporting database and not a data warehouse. They are not the same thing, although they are often confused, one with the other, by people who do not know the difference. Often times, the data warehouse team will hear "I want up-to-the-minute information, but I want my numbers to not keep changing" from their consumers.

THE AUDIT LIFECYCLE

Referring back to our example under "Truth", completeness means that an examination of these financial records must also include an allowance for the six week delay in resolution of accounts. Any recommendation not taking this lag into account would be meaningless, and to the contrary, not examining the reasons and effects pertaining to this lag would be to fail to perform as an auditor should perform. Is the lag acceptable? What are the risks associated with this lag?

Reproducibility is likewise important for auditors, as you cannot have credibility when the analysis presented cannot be reproduced from the same data. For this reason, auditors will extract the needed data, typically at a "good time" in the business cycle, to best reflect the current position of the company while having minimal impact on operations.

A Run Book would rarely be required in an Audit setting, unless you are dealing with a complex audit or one that crosses multiple business units/national borders/currencies. In such cases, the Run Book would allow one to document the details of the multiple data extracts taken from multiple business units, and the exchange rate in effect for each data extract. This allows an easily understood means of communicating the dates, times, conditions, rates and exceptions associated with these extracts, and can aid in resolving discrepancies inherent in such complex situations. From personal experience, exchange rate changes from one day to the next can introduce significant variability in financial results. By making these factors clear from the beginning of the process and resurfacing them throughout the audit, one can eliminate needless conflict/contention. The Run Book can help surface these facts.

In order to be credible and to have consistency between different reports, auditors typically take a snapshot of the system being audited, and then perform their analyses. By using a snapshot approach, a time-Invariance approach is being used. This allows different questions to be asked, without the concern that the data will have changed in the intervening period. For an auditor, it is critically important to match the business cycles, and that includes taking snapshots of data at the appropriate time, or any analysis will doubtlessly be meaningless. At best, a great deal of effort will be required to explain any discrepancies arising from this off-cycle data extract.

THE LITIGATION ANALYSIS LIFECYCLE

Complete information is essential in order to analyze correctly the financial position under dispute (or you wouldn't be looking at the financial data). In any analysis conducted, referring to our example under "Truth", that six week lag would need to be taken into account in order to provide the correct value for any period; working without this information would cause a potentially very different result from the analysis. Hopefully, the claims period precedes this period sufficiently that we have complete data for our analysis.

The only way to communicate the order of programs to be run, with what data inputs are required, is to provide an equivalent document to, or some variation on, the Run Book. But don't be surprised if that courtesy is not extended to you by opposing counsel.

When a Run Book is provided, it has two effects:

1. You are declaring that you KNOW that your results are reproducible and that your work will stand up to opposing counsel's scrutiny.
2. You are being open and honest in your work, and are willing to reduce the time and cost of opposing counsel's expert witness in studying your evidence.

This has the added benefit that opposing counsel will now need to look elsewhere for something to defeat your side's position. If opposing counsel does not provide the equivalent to a run book, then it is left to you to determine what

they did, in what order, and to what effect. Don't be surprised if their evidence is not much better (in organization or in quality of documentation).

In Litigation Analysis it is possible for a judge to order one or more updated data extractions (typically when the case runs over several years), so that the defendant's progress/performance can be better examined. However, usually the data initially received is the data you will have for the life of the case. Supplementary datasets may arise, due to questions posed (by either side) or where the initial extract was deemed to be incomplete. Any such additional datasets would likewise need to be covered by the Run Book, as the analyses conducted would typically need to use this/these supplementary data source(s).

Time-invariance is inherent in the litigation data analysis process. Since you are working from extract files from transactional systems, they are by definition Time-Invariant. If new datasets are supplied, to cover a subsequent period of time, they should be only covering transactions the pre- or post-date the previously received extracts. As is the case with most business cycles, however, expect there to be some measure of overlap.

PLANNING, RESOURCING, EXTRACTING, TRANSFORMING, ANOMOLIES AND ANALYZING

Now we'll begin to explore the specifics of these topics, as they relate in general and to each of our triplets in more detail.

PLANNING FOR THE ACTIVITY

To set the context for this topic, here are two quotes from http://www.bluejacket.com/humor_usmc_wisdom.html:

- *9. No plan survives the first contact intact.*
- *13. Always have a plan. 13a. Have a back-up plan, because the first one won't work.*

There is nothing so important as proper planning, yet too little or badly done planning for the execution of any endeavor seems to the order of the day. Project Managers live and die by their project plans, so the urge to under-plan and "just get to it" needs to be avoided at all costs. It will cost you and in ways you won't realize until it's too late, if you didn't examine the risks sufficiently in planning.

While much has been said and written on this subject, a look at how it is not so different between our triplets is in order.

THE DATA WAREHOUSING LIFECYCLE

Once the initial multi-level planning has occurred for the introduction of a data warehouse, on-going operations take minimal planning until there is a change. The magnitude of the change dictates the magnitude of the planning to implement the change; which means that it will just be accomplished normally by a set of change requests so that major funding will not be required, and operations can do the work in their spare time. And no, I'm not a cynic.

If this is the planning for the creation of a data warehouse, much has been written on the design and collaboration needed to develop a successful (read useful and usable) data warehouse, and I won't repeat the writings of Ralph Kimball, Bill Inmon, *et al* here.

THE AUDIT LIFECYCLE

Planning is normally more of a scheduling or time-and-cost activity for typical audits, as the nature of an audit is generally the same corporately from one audit assignment to the next. The data may change and the issues may change, but most audits are based on a basic data analysis to determine variance from standards and practices, rather than a deeper understanding of trends, causes and effects (also known as secondary or tertiary effects). This is not a negative comment, but an operational reality, as the control of secondary and tertiary effects are management's responsibility. Auditors have to be generalists with sufficiently specialized skills to understand the bigger issues, and the sleuthing skills to "tease out" the micro-issues in any assignment; they are not a replacement for good management. In my opinion.

THE LITIGATION ANALYSIS LIFECYCLE

Depending on the stage in the case in which you are being asked to join, you may by under a court-imposed date by which you are to deliver your analysis. Ideally, you are engaged at the early stages of the case, so that you have the time to be comprehensive in your analysis. Otherwise, your analysis must be very directed to the wording of the complaint, rather than the bigger issues that may be present, and this can lead to missing the "root cause" of the complaint. Too tight a deadline also leads to shortcuts being taken to see results, rather than spending the time on the deeper understanding needed to show why the shortcuts give misleading results. We will be revisiting shortcuts again.

RESOURCING FOR THE ACTIVITY

There are two general resourcing strategies: use a permanent team, supplementing it with specialist knowledge as required, and use an ad-hoc team for each effort. This is true across most disciplines and is not restricted to our

triplets.

THE DATA WAREHOUSING LIFECYCLE

If the data warehouse is viewed as business essential, then the organization will have a permanent team responsible for the care and maintenance of the data warehouse. It is the rare organization that will have no-one tasked with support of the data warehouse if it has corporate value. The only time that additional resources would be added to the data warehouse team in the case of a permanent team is if an addition is being made to the data warehouse, and no-one on the team has experience with that subject matter. As an example, a human resources data warehouse was previously built and is in production. Management now wants to add financial analysis capabilities to the data warehouse; the team will need accounting/finance experience added to the team, at least on a temporary basis.

THE AUDIT LIFECYCLE

With the emphasis on Compliance, businesses that did not have full-time audit staff have moved to full-time staff. If they have not before been involved in a SAS 70 audit, however, expect them to need to add one or more specialists to the audit team, at least for the duration of this type of audit. If such audits will be conducted regularly and sufficiently frequently, options include adding resources permanently with this experience, or training existing auditors in the conduct of such an audit.

THE LITIGATION ANALYSIS LIFECYCLE

Unless the organization is being sued on a nearly weekly basis, there normally is no staff assigned to permanent litigation analysis. For each case, an ad-hoc team would be expected to be formed to support the stakeholders in the case, such as the VP of the area being sued and In-House Counsel. Typically, Expert Witnesses have their own staff at hand to conduct the data analyses needed to defend/prosecute the case, so corporate involvement would be limited to explaining or investigating data anomalies or the applicable corporate policies and practices. As well, expert witnesses bring credibility, impartiality and a breadth of knowledge to the question at hand not typically available otherwise to most organizations.

GETTING DATA FOR THE ACTIVITY (EXTRACT)

Here is where we see begin to see the largest differences appear between data warehousing, auditing and litigation analysis. While the triplets are all data-intensive, the differences that gave birth to each function changes how data are perceived handled and managed. What differentiates the triplets the most is the degree of ownership of the data. Auditors, as agents for the organization's management team, have near unfettered access to any corporate data holding *relevant to their current studies.* On the other hand, data warehousing staff and litigation analysts are given data to use; the litigation analyst typically has little to no say in what data they will receive, when they will be received, or how the data are prepared or delivered. The data extracts each depend upon are characterized as either a Push or Pull extract.

THE PUSH EXTRACT

A Push extract is where the data source authority "pushes" the data to the recipient. The responsibility for the completeness and appropriateness of the extract reside with the data source authority; the recipient needs take the completeness and accuracy for granted, until/unless problems arise. The recipient's timing requirements may be considered by the data source authority, but until the data source authority is satisfied with the status of processing for that data, no extract will be forthcoming. It is then up to the recipient to take the appropriate action.

THE PULL EXTRACT

A Pull extract is where the recipient goes to the data source and takes the extract that they require on their terms. While the data source authority needs to have been consulted when the extract was first taken (except possibly for some audit situations), thereafter the recipient would just take the needed extract. For Pull extracts, communications need to take place regularly so that process lag, failed updates, etc. are all communicated to the recipient, so that their extract will only cover valid data conditions and will reflect the true situation at the point in time the extract is taken. As part of this Pull extract, the rest of the processing needed by that recipient would then take place.

THE DATA WAREHOUSING LIFECYCLE

In the Push, the organization/department providing the data provides the data in a previously agreed-to format at a set interval, and loads that data into a staging area. The data warehouse team takes this information, transforms and loads it into the data warehouse, baring any data problems being encountered that preclude its loading into the data warehouse.

In the Pull, the data warehouse team awaits a signal that the appropriate processing has been completed, and the data are available and ready for extract for the applicable reporting period. The data extracted for that period are loaded into a staging area. The data warehouse team takes this information, transforms and loads it into the data warehouse, barring any data problems being encountered that preclude its loading into the data warehouse.

In some cases, the data are purchased from an external party, and updates are made available on a semi-regular basis. Examples would be postal ZIP code mapping to geographic area coding being updated quarterly, or the decennial Census of Housing and Population data released every 10 years. The more infrequent the data release, the bigger the impact on the organization of the data warehouse (and its consumers), and the more opportunities for

changes and transition strategies. This is particularly true where the definitions underpinning the data have changed from one data release to another (consider the definition changes to the 2000 Census, compared to the 1990 Census).

Where data do not conform to definitions/types/code-sets/etc., agreements need to be in place on how to address the problem, be it to replace offending data with an 'invalid data' indicator, or to reject the offending data extract in its entirety. In my experience, there are two levels to this exception handling: for the first _*n*_ exceptions, replace the offending values with the 'invalid data' indicator, but after this threshold has been exceeded, reject the data extract entirely. Suitable manual and automated notification mechanisms would likewise be needed so that the supplier of the data is made aware of a situation that may warrant further action on their part.

One of the greatest risks facing a data warehouse is in accommodating external data sources. While normal business changes affect data coding, wholesale changes to data sources and structures are infrequent and when they do occur, the cost of the data warehouse changes should be expected to be part of the cost of that system change. When external data sources are used, such as the Census data, and major changes occur in the definition and meaning of the data, one of the possible courses of action is to incorporate the new data into the data warehouse as new relationships, and plan to phase-out the old structure and definitions of previous version of this data. Planning for this kind of change rarely is possible, as you cannot know the future plans and intentions of external organizations until they are ready to communicate their plans and actions. You can only plan to react to such changes when they arise.

THE AUDIT LIFECYCLE

Most audit organizations, be it in-house or external contractor, have their own tools that are specialized for the type of audits generally being performed. In my experience, the tools the auditors use work against production databases, and extract the required information very easily. This extract is then used to analyze the data against preset and previously determined metrics to measure both compliance with standards/policies, and progress in moving towards those standards/policies compared to the last audit. This would be a Pull extract.

When large databases and complex logic are involved, a Push extract, but on the audit team's terms, may be the preferred way to proceed. This way, the knowledge associated with the extraction and preparation of subsets of data from complex data structures and systems, found in the data source authority's area alone, can be taped for audit purposes. From this structured extract, audit can extract the needed data, without the problems and risks associated with working with the transactional systems' complexities.

Risks from changes in external data sources would be comparatively limited, since most audits focus on the internal rather than external. Only where industry norms or other external measures experience change would changes in these data sources have an impact on the audit process and results.

THE LITIGATION ANALYSIS LIFECYCLE

This would clearly be a Push extract. While getting the extract is generally very easy, it can degenerate into an exercise akin to giving a root canal without anesthetic to a rampaging mastodon. It hurts and it's no fun for the mastodon either. The issue of complexity, completeness, accuracy and timeliness all remain with the data source authority. Spreadsheets for simple/small data and database extracts for complex/large data extracts, with some degree of explanation as to what is being provided, is the general rule.

The rules of evidence dictate the time limits that are permitted in cases, according to the jurisdiction in which the complaint is filed, so data would be extracted (with exceptions, of course) for a specific claim period, for example, January 1997 to December 2000, subject to these rules. Counsel (most typically for the defendant) should advise the people extracting this data as to the pertinent period of time to use for the extract. These data would be made available to Opposing Counsel (the Prosecution in that case), along with any explanatory notes, caveats and limitations. Expert witnesses for both sides would then conduct their analyses, based on this data. It is a multi-million (if not multi-billion) dollar per year industry to convert paper documents into indexed optical images available on the web for the parties to such a case. For the remainder of this paper, however, we will be restricting the discussion to machine-readable data, from text files, through EXCEL™ and SAS datasets.

PREPARING THE DATA FOR THE ACTIVITY (TRANSFORM)

No matter what form the data are in and no matter what analysis is required, almost without exception, there is a need for recoding, summarization, and/or reclassification of data to prepare them for analytical actions. Some examples of changes are:

- Adding standard coding, such as the EEO and EEO-1 codes appropriate to the position in-hand
- Adding geographic coding such as Census County-Set Area code for the physical work location
- Adding summary statistics, so that counts/sums/means/minima/maxima are calculated for the various levels of analysis expected
- Deriving segmentation dates, so that every transactional record now has a start and end date (think "good

until")

- Creating snapshot files, so that information for one snapshot period can be used in comparison to any other snapshot set
- Converting salary or account balances to coded ranges, such as "0='< $10,000'; 1='< $25,000'; etc. so that any value in that range will have the same code. This means creating a discrete variable as proxy for a continuous variable.
- Rendering anonymous identifying data, so distinguishing characteristics of any individual are removed, and only demographic aggregates remain. Think of the Census datasets, where thousands of people are represented by any cell in any record, and no individual can possibly be determined.

THE DATA WAREHOUSING LIFECYCLE

Transformation processes are well known in data warehousing, since data are typically "burst" into fact tables and dimension tables (known collectively as a star schema). See books by Dr. Ralph Kimball and Bill Inmon for details on the construction of fact and dimension tables. The mapping or transformation rules to convert the raw data sources into these fact and dimension tables are based on structured data definitions and exception processing agreements. These agreements are fundamental to the processing of the raw data to create the data warehouse data feeds and the understanding needed to create analyses based on the contents of the data warehouse. The data warehouse, even after all these transformations, must still report the "Truth" or the data warehouse will be valueless.

Typically, there is a collapsing strategy that is defined to move records coded _X_ to the various summary or aggregation levels. When a value is received that is not defined by this collapsing strategy, there are two choices: reject the batch, or recode the "bad data" to a known error code, indicating that the value was seen but did not belong to a known category. In the example to follow, think of another Category or high-level aggregation such as UNKNOWN, so that no detail is lost and the various data relationships are preserved. This also means that the data warehouse has not compromised the "truth" of the data.

While a trivial example, and incomplete for any serious analysis, think of a financial chart of accounts. There are typically three levels of data in such a chart, namely, Grand Total (Total Assets/Total Liabilities/etc.), Category Totals (Prepaid Rent, Salaries, etc.) and Category Detail (accumulated Prepaid Rent, etc.). For any one cell in the chart of accounts, there are three levels of summarization, for each period of time. For drill-down in the data warehouse, you would have: Net-Total } Category-Total } Category-Detail-Total } Category-Detail, with the three Totals being the summarization levels. The sample Chart of Accounts extract for our example would look something like:

1. Net-Total (highest level aggregation)
    1.1. Assets-Total (high-level aggregation for each category)
        1.1.1. …
    1.2. Liabilities-Total
        1.2.1. …
        1.2.2. Prepaid Rent (Category-Level aggregation)
            1.2.2.1. Prepaid Rent [accumulated]
            1.2.2.2. Prepaid Rent [expensed] ← this is a contra-account, used in Accrual Accounting
        1.2.3. …
    1.3. Equity-Total
        1.3.1. …
    1.4. Unknown-Total
        1.4.1. Unknown aggregation
            1.4.1.1. … unknown detail …

THE AUDIT LIFECYCLE

Unless the auditor needs external data as a measure against which to compare the organization being audited, virtually all the information that the auditor requires is contained within the corporate data sources being examined. Again, auditors typically have specialized tools for the extraction of data, where transformations are not required, and the loading stage is an inherent part of the tool's capabilities. During the analysis, transformations may be added or derived from the initially extracted data to facilitate the comparing and contrasting of the data for analysis.

THE LITIGATION ANALYSIS LIFECYCLE

One of the most common transforms for litigation analysis is the categorization of data. While continuous variables can be used, for some types of analysis, creating suitable categories for salary or time variables may make the analysis much simpler. For example, in a promotions study, you might look at the number of months between promotions (a continuous value), but would rather work with a discrete value, created by bucketing the time between promotions into a set of buckets. An example would be: under six months, six months but less than one year, at least one year but less than two years, and so on. You now will have a number of people with like category values, where before they would have had different (but similar) periods of time between promotions. Since you now have created cohorts, you now can measure differences between members of each cohort. A similar approach for salary data may also be useful, breaking the spectrum of compensation into, for example, five to ten buckets.

Another key tool in litigation analysis is the creation of snapshot files. These files either are created to represent quarterly, semi-annual, annual or subject-period units of time. Subject-period snapshot files are best explained by an example. Suppose a company introduces a new policy. At some time later, the company moves away from that policy. To see the impact of that policy on the company, you would create three snapshot files: Pre-Policy, Policy and Post-Policy so that the effect of the policy could be determined by comparing and contrasting the data from the three periods. Done correctly, you can now do Stock and Flow analysis, based on these snapshot files.

Time series issues often figure into litigation analysis, since the change over time of subject-matter variables figures prominently in most cases. Discrimination and bias cases are about what is expected to have occurred during the indicated period versus what actually happened. An extension to Time Series is 'Time Series In Context', where you have multiple, dependent time series, and you need to bring the data into context, one time series with another. While this will be covered in more detail later, think of the lag between the date of promotions and the date by which the corresponding pay change is reflected in your pay checks; while it would be nice, you don't get a promotion with each pay check. However, your pay check should reflect your promotions (else why would you ever want a promotion?).

LOADING BLACK HOLES, STELLAR PHENOMENA AND OTHER ANOMOLIES – WHAT TO DO

The way people cope with exceptions tells you a great deal about them. Some are paralyzed with fear and concern, and others see it as a challenge, in fulfillment of the Chinese curse "may you live in interesting time!" Working with others' data quickly teaches you that if they can muddle the data in some fashion, they will. How these anomalies are handled during the data loading stage varies between data warehousing, auditing and litigation analysis only in whether the anomaly is noteworthy/a stopping point, or just something to be "handled."

THE DATA WAREHOUSING LIFECYCLE

There are two courses of action when unexpected data values are received. Depending on the severity of the anomaly (e.g., receiving the text string "n/a" instead of the number for year-to-date expenditures), the data warehouse data loader would either:

- Replace the anomalous value with a known value for unacceptable values (think the infamous Missing), or
- Stop the loading process, reject the dataset entirely, and notify the appropriate parties of this event being due to this data error.

For minor values, setting to a known value is the accepted approach, but for critical values such as date, organization code, part number, or other primary or essential key value, the only option is to reject the dataset until either the offending record is replaced with acceptable values, or the acceptability criteria are amended to accept this new value (for instance, a new division is reporting for the first time in the company).

THE AUDIT LIFECYCLE

When faced with invalid data, the auditor has two choices, either:

- The anomaly leads to an audit finding, and management needs to address the cause and resolution of the anomaly; or
- Ignore the anomaly until/unless it points out a more substantial problem, at which time the more substantial problem leads to an audit finding and the anomaly is just background/substantiating detail.

THE LITIGATION ANALYSIS LIFECYCLE

When an anomaly is detected, if it is in a variable needed for analysis, then all such anomalies are identified to counsel for resolution with opposing counsel. In some cases, the analysis would continue, but records of this nature would be removed from consideration in the analysis. An example of this type of anomaly would be in a gender pay discrimination case, where Gender or Salary is missing. Since these are the variables of primary importance in such an analysis, the record missing any of these data must be excluded from consideration, and counsel would be advised of data problems should there be a large number of records that are "defective" for some analytical reason.

And since record counts can become a point of dispute, explanations/listings at every point where records are excluded (with the reason for that exclusion) is standard procedure.

ANALYZING AND REPORTING FOR THE ACTIVITY

Here the bulk of the work falls on the auditing and litigation analysis functions, as by this point, data warehousing has completed the work associated with populating the data warehouse, and the Business Intelligence users (or consumers) now being working with the data.

THE DATA WAREHOUSING LIFECYCLE

For data warehousing, this activity is primarily in the realm of business intelligence and belongs to their consumers.

Once the data warehouse has been updated, the only concern of the data warehouse staff is tuning, performance and training. If designed correctly, and with the proper software (think SAS), tuning and performance is more a hardware issue. Training is one of the data warehouse key success factors, though, as if the consumers who use the data

warehouse do not understand the data and structure, they will produce results which are not credible, and then who gets the blame? While outside the scope of this paper, I would suggest books by Ralph Kimball or Bill Inmon to address the issue of designing for analysis and reporting using the data warehouse. One consideration is to revisit the use made of the data warehouse on a regular basis, so that measures needed by the consumer can be implemented in the data warehouse. If you look at the sample problem at the end of the paper, you will see summary values that, made available through the data warehouse, would allow more sophisticated answers to be derived from the data warehouse.

The Business Intelligence analysis may well be similar to audit or litigation analysis but is a user not data warehousing responsibility. Unlike audit or litigation analysis, the data warehouse team is a facilitator and not the party taking action based on the data being processed.

THE AUDIT LIFECYCLE

Audit reports will typically address the tests performed, results discovered, and any "findings" of value to note. Typically these constitute records of procedural violations and hopefully not financial violations, as recently seen on the news. The intent is both to inform management of potential problems and to comfort one and all that there is no risk inherent in the operations just reviewed. Depending on the degree of openness in the communications before, during, and after the audit, the auditor may be seen as being as welcome as an IRS auditor or OFCCP inspector.

The report will typically include quotes from the data extracted to support the finding(s) presented, and should be available for review should the interpretation of the facts be challenged by the manager of the area/operations under audit.

THE LITIGATION ANALYSIS LIFECYCLE

There are multiple pieces to this activity, and the complexity/scope of the complaint and the data available will dictate the volume of information analyzed and presented. It will also indicate the number of stages between the first final report, deposition, response to opposing counsel's expert witness' final report, response to opposing counsel's expert witness' deposition, etc., until the case ends.

Often, a final report is produced (not in all cases though), giving a précis of the complaint, the methodology used to analyze the data supplied, and the findings of this analysis. It also would include a *curriculum vitae* of the expert witness, to establish their expertise in analyzing and reporting on such matters.

In other cases, there may be a deposition, where opposing counsel interviews the expert witness under oath, resulting in a transcription of this deposition. This is a fact-finding stage to determine the robustness of the evidence and the ability of the expert witness' findings to stand up to close scrutiny. Should the case proceed to trial, the deposition would be submitted in evidence, and the expert witness may be called to give further testimony.

While not done in many cases, it is reasonable to expect that opposing counsel will request a copy of all programs used to extract, transform and analyze the data, leading to the conclusions and facts stated in the expert witness' final report. To make the most sense of the material, and to ensure that the integrity, accuracy and veracity of the conclusions are supported and not subject to mechanical challenges, the Run Book is an invaluable tool to document the process of recreating the conclusions, from reading the initial data to presenting the Rank Sum or Regression or study quoted in the final report to support the conclusion(s). One source of challenges to an expert witness' credibility is to challenge them on mechanical errors such as:

- Programs run out of order (how do you show "cause and effect" when the effect precedes the cause?)
- Programs that won't run or produce different results than those reported (we're to believe these results really were created by the program somehow/at some time? Under what conditions?? Only when Aquarius is rising???) This is often seen during interactive development, when a file is being reused, often to define the variables, labels and/or formats associated with the data. In batch, the program fails as the file does not yet exist. This is a significant fault.
- Log (.LOG) or List (.LST) files last modified **before** the program's last modified date (think *"Let's do the time warp again"* from *Rocky Horror Picture Show,* 1975).
- Datasets with more/fewer columns and/or observations than can be reproduced. This too is typified by developing and executing programs interactively but not saving the final version of the program which included the code to filter observations, or add/remove/transform columns.

Think "if the data or programs don't fit, you must acquit!"

See the previous section on Repeatability for the Run Book example. That example shows the number of possible points of failure that you are facing, even with such a simple example.

TIME SERIES (WAS THAT A FENCE POST OR A FENCE RAIL?)

Regardless of which one of the triplets one examines, the correct handling and analysis of Time Series data can be a significant issue. There are two types of events over time: the milestone (also known as a point-in-time, transactional data or historical data), and the duration (typically start date and the number of units of time the situation had life, OR start and end dates) expressed through a snapshot. The issue comes with the type of analysis being conducted, as the two types of date information are not interchangeable; you however can (and often do) convert transactional data into snapshots for various periods.

A way to "bucket" transactional records into snapshot files is to determine and use the most appropriate granularity specified by the nature of the complaint, type of audit or other data-related differentiation. When working with multi-year data, you would typically produce quarterly, semi-annual or annual snapshots, while in a different case, you would produce pre-, during- and post-situation snapshots. To include someone in a snapshot is defined, in human resource examples, most typically as anyone who was:

- Hired on/before the snapshot date, and
- Never terminated or terminated some time after the snapshot date.

Since this will exclude those persons who were hired and terminated during the snapshot period (that is, they were hired after the snapshot date and were terminated before the next snapshot date), a further refinement is to flag such persons and include them in the snapshot, so that head count numbers will better reconcile. This refinement to the type of structure for the data allows "stock and flow" counts to be created, so that the on-going population can be differentiated from the new hires and the in-period terminations, and the on-going population count (or stock) can be determined, along with the additions/subtraction to the study population (or flow). As an example, you will see this expressed explicitly or implicitly in Annual Reports, be it financial or other data being displayed.

The remainder of this section will be an example of a study on possible promotion discriminations. As is (hopefully) a familiar consideration, think about a person's promotion date versus the date that the pay change due to the promotion first appears in their pay check. Both can be thought of as points-in-time (respectively, the start date of the appointment, and the date of the first pay check to reflect the promotion). While frequent promotions would be appreciated by most people, the reality is that one receives many pay checks between each promotion. Also, there is no relationship (or at least no immediate relationship) between the promotion event and a pay check amount. Therefore, at first blush, one would assume that they are independent events, when they are actually loosely coupled in time. Part of the problem is that pay checks are typically paid in arrears, and corrections that occur during one pay period will be posted to your check for the next pay check. On the other hand, some organizations are able to make the update during the pay period (with a few last-minute exceptions) so that the pay check is kept current with events in that pay period. The third possible approach is that at some point in time after the promotion has occurred, a retroactive check is issued, bringing the pay check(s) in the interim "up to date," and all subsequent pay checks reflect this promotion. The latter approach becomes problematic for data warehousing, as Time-Invariance rules cannot be correctly met. What will be often done, from a pragmatic approach, is to simply report the pay check in the reporting period in which it was issued, and to not attempt to apply the proportion of the pay check to each retroactive reporting period as appropriate. In a settlement in the past year or two, the Government of Canada paid staff retroactively for a three-plus year period for an Affirmative Action equalization complaint; staff in certain professions that were female-dominated were deemed to have been paid less for equivalent work compared to other, similar occupations that were male-dominated.

When in doubt on pay issues, look to when the IRS or other appropriate national income taxation agency deems the amount subject to taxation and when. If the settlement reached back 10 years, you have not been in arrears for the taxes for those 10 years, just from the point in time where the settlement payment date was established; this is not normally the date when the settlement was reached, but rather an agree-to date some weeks/months in the future by which time the payee (in this case, the employer) needed to issue/begin issuing the arrears checks!

The fence rail in this example is the time between pay checks, or the time between promotions. The fence posts are the effective dates (i.e., the promotion dates or the pay days for the various pay checks). In a litigation analysis context, what may well be important is the time between the promotion and the payment. If minorities or females are only paid for their promotions three months after the event, while white males are paid in at most ten days, then there is discrimination. As indicated in the section on Accuracy, whether this can be demonstrated to the court's satisfaction (that is, a statistical disparity in excess of two-three standard deviations) is in part a function of the number of people in this case.

THE DATA WAREHOUSING LIFECYCLE

Time dimensions are typically Calendar Date or Reporting Periods (such as Week, Month, Quarter and/or Year). With this example, the promotion would be shown as a Calendar Date, while the pay check would be a Reporting Period (typically pay period 1, 2, etc.) with an additional fact or dimension attribute of the pay date. For most considerations, the reporting period in which the pay check was issued is taken as the reporting period of the complete check. What can become tricky is attempting to span data periods, where for instance the pay check is issued in reporting period 2, but belongs in part to reporting period 1 (e.g., 4 days in period 1 and 6 days in period 2; 10 work days to the pay period). The resolution is most often to leave to the user to resolve the situation, by determining the number of days of the total applicable to each reporting period and scaling the pay check amount

accordingly (i.e., 4/10 * $X in period 1 and 6/10 * $X in period 2). Alternatively, the rules agreed-to may say that the period in which the pay check is issued is the reporting period. What is most important is that the rules that have previously been defined for the handling of this data when the scope of the data warehouse was defined are followed. And even more important, all consumers of results from this data warehouse must be aware of this agreement.

When working with human resources data, it has been my experience that stock and flow is the most common approach to analyze populations, as it is the easiest to explain to a general audience.

THE AUDIT LIFECYCLE

Generally, an audit focuses on a point-in-time or snapshot, and is concerned with then determining the frequency or count of occurrences, or the duration (or interval) between events. For instance, if the rule that is being audited is that the promotion will be reflected in the pay check for the first complete pay period following the promotion, then the auditor will examine the difference between the date of the promotion and the date of the pay check. This would be, for a biweekly pay period, from 15 to 28 days (think within two pay periods); biweekly pay is every 14 calendar days. If monthly paid, then the audit would look to see if the month of the pay change is the month following the month of the promotion; this is a milestone comparison. A factor that should be expected to be reviewed by an audit is the time taken to complete these back-pay transactions, as the lag needs to be noted in accounting statements; the amount pending completion of this lag is an expense or liability against the company's accounts and needs to be addressed by the company books.

THE LITIGATION ANALYSIS LIFECYCLE

Here is an area where data are often in a snapshot form for analytical purposes. Transactional data are often used, as both milestones and durations are important. For some analytical questions, you need point-in-time data (to answer "how many?"), while for other analyses, you need the duration expressed by start and end date information (to answer "how long?"). These data allow you to determine inclusion/exclusion rules for each record's participation in a particular study. If you are studying the number of promotions or terminations by minority/non-minority or male/female due to allegations of discrimination, you are examining milestone data, and pay questions are not included (at least not initially). If however the claim of discrimination is due to the time BETWEEN the promotion and the first payment of the revised wages, due to gender or racial bias, then you are looking at duration and not events. This is because the promotion is not in question, just how long (i.e., duration) it took for pay to be adjusted to account for the promotion.

CONCLUSION

By now, it should be clear that the lifecycle of data warehousing, auditing and litigation analysis are clearly similar, but with substantive differences. Issues faced are largely the same, but focus and actions permitted in context vary from one triplet to the other. "Truth," Reproducibility and Time-Invariance are underpinnings for every stage in the life cycle of all these endeavors. The Extract, Transform, Loading, Anomaly Processing and Analysis stages are common to our triplets, with the emphasis/responsibilities varying from one triplet to another.

Data warehousing staff are the people who can help an organization extract and present data to opposing counsel, as they have the experience in working with and knowledge of the data (and the more important metadata), together with the appropriate handling of the various exceptions inherent in the data/system being used. They also can speak to the resolution of missing/invalid data, given that they deal with this whenever they load the data warehouse with this information. This assumes however that the organization has a data warehouse, and it pertains in whole or in part to the files that are included in the litigation.

Given that auditors are concerned with the legality of operations, above all else, it makes sense, when engaged in litigation analysis, to consider having Audit assist with the process. Be aware, however, that auditors are typically not trained in the specifics that would lead an organization to engage an expert witness, so their assistance would of necessity be limited to the Extraction and Transform stages of the process.

Expert witnesses have the knowledge and experience to speak to the methods of litigation data analysis and interpretation that are relevant to the case at hand, and bring both substantive opinion and credibility to the analysis. The art of selecting an expert witness is best left to counsel, but know that their expertise is not inexpensive. Engaging an expert witness is a potentially large expense, unless you win, in which case they save your organization far larger dollar outlays.

WEB REFERENCES

1. The American Heritage® Dictionary of the English Language, Fourth Edition
   Copyright © 2000 by Houghton Mifflin Company.
   From the website: http://dictionary.reference.com/search?q=truth
2. The rules of Golf for the USA
   http://www.usga.org
3. For more information on Daubert motions and decisions
   http://www.DaubertOnTheWeb.com

4. Rocky Horror Picture Show CD
   http://www.amazon.com/exec/obidos/ASIN/B0000032LS/qid=1124376342/sr=2-1/ref=pd_bbs_b_2_1/002-6554896-0584017
5. For Marine Corp humor, see
   http://www.bluejacket.com/humor_usmc_wisdom.html
6. The best source I've found for ZIP Code information in a non-commercial environment:
   http://mcdc2.missouri.edu/webrepts/geography/ZIP.resources.html
7. Books by Dr. Ralph Kimball on data warehousing
   http://www.ralphkimball.com/html/books.html
8. Information on the Information Systems Professional of Canada (I.S.P.) certification
   http://www.cips.ca/standards/ispcert/

REFERENCES

1. Kimball, Ralph
   The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses
   John Wiley & Sons, 1996
   ISBN: 0471153370
2. U.S. Census Bureau; Census 2000 EEO Special Tabulations File [table1]; <http://www.census.gov/Press-Release/www/releases/archives/census_2000/001633.html> and take the EEO Data Tool link for online extracts or CD-ROM purchase options; (February 17, 2004).

ACKNOWLEDGMENTS

My thanks to the staff and colleagues at the Peopleclick Research Institute for sharing their experience and knowledge with me; the mistakes however are mine, all mine!

RECOMMENDED READING

1. *see* Kimball, Ralph in References and Web References
2. some representative audit standards and procedures:
   a. SAS 70 – Statement of Auditing Standards No. 70
      http://www.sas70.com/
   b. Office of the Federal Environmental Executive (random Googling)
      http://www.ofee.gov/ems/resources/Generic_Audit_Procedure.pdf
3. *Litigation and SAS: some DOs and DON'Ts (or, you call that Evidence?)* [SUGI paper 144-30]
   http://www2.sas.com/proceedings/sugi30/144-30.pdf

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Nicholson Warman, MBA, I.S.P.[2]
Peopleclick Research Institute
Peopleclick, Inc.
7th floor, Two Hannover Square
Raleigh NC 27601
Work Phone: 919-645-3674
Email: Nick.Warman@peopleclick.com
Web: www.peopleclick.com

[1] A Daubert motion is a motion to the court to have an expert witness' testimony and evidence set aside or suppressed by the court. It is a challenge to the scientific integrity of the evidence, and causes include:

- Sloppy or incomplete/incorrect work
- Misrepresentation of fact, experience or detail
- Methods not peer-reviewed and not (yet) accepted by the appropriate scientific community

See www.DaubertOnTheWeb.com for further insight and details.

[2] See www.cips.ca/standards/ispcert for information on the Information Systems Professional (I.S.P.) certification program in Canada.