



ACCOUNTING

CONTINUING EDUCATION

Data Analytics for CPAs

(DATAC4)

Data Analytics for CPAs

(DATAC4)

Brad Caruso, CPA, CFE

Thomas Gancarski, CPA, CIPP/E, CISA



DATA ANALYTICS FOR CPAS (DATAC4)
©2021 Kaplan, Inc.
Published in 2021 by Kaplan Financial Education.

Printed in the United States of America.

All rights reserved. The text of this publication, or any part thereof, may not be translated, reprinted or reproduced in any manner whatsoever, including photocopying and recording, or in any information storage and retrieval system without written permission from the publisher.

ISBN: 978-1-0788-1478-2

TABLE OF CONTENTS

- 1..... 1**
- Data Analytics Basics..... 1
 - Learning Objectives1
 - Introduction1
 - Phases of a Data Analytic Project or Process7
 - Summary..... 15
- 2..... 17**
- Firm Technology 17
 - Learning Objectives 17
 - Data Analytics in the Accounting Field 31
 - Summary..... 33
- 3..... 35**
- Data Analytic Techniques 35
 - Learning Objectives 35
 - Introduction 35
 - Acquiring Data 35
 - Sampling..... 39
 - Journal Entry Testing..... 48
 - Join or Match Functions..... 49
 - Append Function..... 52
 - Benford’s Law 53
 - Specific Audit Areas 54
 - Data Visualization 59
 - Summary..... 63
- Case Studies 65**
 - Case Study #1 - Newark Watershed 65
 - Case Study #2 - DOV-Q, Customer Attrition, and Predictive Analytics 69
 - Case Study #3 – Not-for-Profit Audit..... 74

Unit

1

Data Analytics Basics

LEARNING OBJECTIVES

After completing this unit, participants will be able to:

- › Describe the phases of the data analytic process
- › Explain data strategy for a business
- › Develop best practices and tips and tricks related to data analytics

INTRODUCTION

Performing data analytics takes a significant amount of thought and commitment. Data comes from many sources, both structured and unstructured. But the most fundamental issue is that every project needs an objective. You need to have the end goal in mind in order to be effective with your quest. **The most common mistake made in performing analytics using technology is the lack of a specific and clear objective.** Technology works great and is very easy to use; however, without knowing what you want to do, you cannot truly answer the call. The first concept we need to learn is how to identify the objective and define it with clarity.

Data analytics is a multiple-faceted field and requires significant thought process before, during and after. When we look at a data analysis project, we need to be mindful of our strategy. Many of us want to just dive in, click fancy buttons, and produce some great analysis. But before even thinking about that, you must develop a plan with clear and concise objectives, goals, and overall expectations. The point of using data analytics is to solve a problem or answer a question. To do so, you must drill down into the specifics.

If you don't know what you are looking for, how can you find it? You probably can't. Would you ever build a house without a definitive objective and plan? Same goes for data analytics. Because you are accessing a different part of your brain that isn't always utilized, you need to plan your moves. What are you trying to accomplish? The author oftentimes receives requests regarding journal entry testing. The problem is the author receives no instruction on what someone is looking for. To make

effective and efficient use of data analytic software and other tools, you must provide specific directive on what you are looking for and what your expectation is for the result.

Additionally, the size of a data set has a significant impact on whether you can rely upon the results. When drawing conclusions on data, there are standard principles that can be applied which are discussed in this manual. However, to identify anomalies in data sets may require multiple years of data from the same company. Many of the technologies today are starting to incorporate machine learning for this reason. The technology can learn patterns and trends to better identify anomalies.

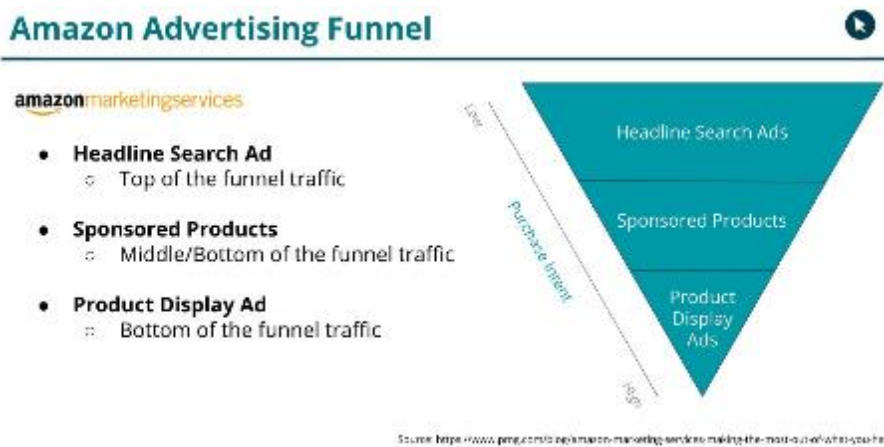
Below are the strategies and techniques in data analytic projects.

DATA STRATEGY

Use Cases by Industry

Retail

Amazon uses consumers' queries and incorporates them with advertising algorithms in order to improve customer relations. This is an area that many Big Data companies overlook.¹



[<https://www.searchenginejournal.com/amazon-search-engine-ranking-algorithm-explained/265173/>]

Financial

American Express Company uses Big Data to analyze and predict consumer behavior.²

¹ "10 companies that are using big data", Eleanor O'Neill (2016), <https://www.icas.com/thought-leadership/technology/10-companies-using-big-data>

² "10 companies that are using big data", Eleanor O'Neill (2016), <https://www.icas.com/thought-leadership/technology/10-companies-using-big-data>

Capital One analyzes the demographics and spending habits of its customers. Capital One uses the information to present offers (marketing) to customers in order to increase conversion rates.³

Media

Netflix uses customer view history (data) coupled with algorithms to predict and suggest programming to customers.⁴



[<https://www.muvi.com/blogs/deciphering-the-unstoppable-netflix-and-the-role-of-big-data.html>]

Technology

Oracle is a Big Data provider offering the following services: Cloud platform, autonomous data warehouses, over 2,000 SaaS applications, analytics models, dashboards, and machine-learning capabilities.⁵

Amazon also provides six equally Big Data storage and databases: NoSQL, Object Storage, Graph Databases, Amazon Aurora, Amazon EMR, and Relational Databases.⁶

Google offers a service called BigQuery, a serverless data warehouse used for real-time analytics that provides advanced security standards and speed.⁷







³ "10 companies that are using big data", Eleanor O'Neill (2016), <https://www.icas.com/thought-leadership/technology/10-companies-using-big-data>

⁴ "10 companies that are using big data", Eleanor O'Neill (2016), <https://www.icas.com/thought-leadership/technology/10-companies-using-big-data>

⁵ Top-15 Big Data companies, Jane Todavchych, <https://thinkmobiles.com/blog/best-big-data-companies/>

⁶ Top-15 Big Data companies, Jane Todavchych, <https://thinkmobiles.com/blog/best-big-data-companies/>

⁷ Top-15 Big Data companies, Jane Todavchych, <https://thinkmobiles.com/blog/best-big-data-companies/>

Industry Analytics Use Cases					
					
Banking	Insurance	Healthcare & Pharma	Retail & Consumer	Telecom	Energy & Utilities
<ul style="list-style-type: none"> ■ Marketing ■ Risk, AML/KYC ■ Product optimization ■ Funds & portfolio ■ Attrition ■ Eligibility 	<ul style="list-style-type: none"> ■ Claims ■ Underwriting ■ Pricing ■ Loss forecasting ■ Loyalty benchmarking ■ Reward partner and channel analysis 	<ul style="list-style-type: none"> ■ Population management ■ Disease management ■ Patient satisfaction ■ Clinical effectiveness ■ Market research trial outcome 	<ul style="list-style-type: none"> ■ Product optimization ■ Sales ■ Campaigns ■ Demand ■ Supply chain 	<ul style="list-style-type: none"> ■ Capacity ■ Maintenance & repair ■ Campaign ■ Cross-sell ■ Customer service ■ Loyalty programs 	<ul style="list-style-type: none"> ■ Account activation ■ Campaigns ■ Smart meters ■ Fraud detection ■ Forecasting ■ Bill and account maintenance ■ Energy management

[<https://www.slideshare.net/SaurabhBanerjee/evolution-of-analytics-timeline-view>]

All businesses need a strategy. Capturing data for the sake of capturing data does not produce a meaningful result and it is especially expensive. Each bit of data stored carries a cost and you need to be mindful of that as you embark on a plan. When you start formulating your plan, start with the basics: What am I trying to achieve? Is it reduction in cost? Is it maximizing revenue stream? Or are you not sure where the most logical place is to start?

According to Bernard Marr in his book, *Data Strategy*,⁸ focusing on key questions helps you to hone in on the data you really need in order to execute a sound plan. The best starting place is to look at the four key areas of your business, which are:

1. Customers, markets and competition
2. Finance
3. Internal operations
4. Personnel

It is extremely important to understand each specific area and try to focus on one area at a time. Let us put these areas into a public accounting firm practice to see how they may relate.

⁸ Data Strategy, Bernard Marr, ©2017, Kogan Page Limited

Customers, Markets and Competition

Accounting firms capture a myriad of data, with the most significant of that data being the almighty chargeable hour (unless you are on the value billing bandwagon)! From a high level, firms may segregate their customer data internally by geographic area, by niche market, by customer size, by partner, etc. In order to make that data meaningful, you must capture the data at inception in such a way to produce those results. If your client acceptance practice does not categorize things appropriately, the results that are produced will not tell you a story except that a mistake was made.

There is a myriad of data out there on customers. Benchmarking of data is an extraordinarily valuable component a CPA firm can provide to its customers for the simple reason that the data isn't publicly available. Public accountants are in an unusual spot because they have access to resources. Obviously, we need to maintain a level of confidentiality to maintain trust. Always use caution here to not reveal client specific information. Therefore, we can capture data about our customers from their financial information to draw trends. And not just high level trends such as profitability metrics, but more granular trends. Some of these trends could include Benford's Law type analysis to identify unique and non-unique data. Identify profit per transaction per day and how weather data affects those metrics.

Finance

How do accounting firms make money? Utilization and realization! When thinking about the finances of an accounting firm, we need to capture data at such a level to understand our utilization and our realization. Utilization refers to the time staff spends doing chargeable vs. non-chargeable activities. Realization refers to the rate or percentage of time incurred to billings. When we have a fixed fee engagement whereby we charge \$20,000 and incur payroll time at market rates of \$25,000, we have 80% realization. This realization percentage factors into profitability because each charge hour is also driven by a payroll dollar. But how can we benefit from this information? Easy – we need to manage our business to maximize both of these metrics. When both metrics slide, so too does our cash in pocket.

Are you spending too much time servicing one customer versus servicing another that is more profitable? How do you define your profitability? A client could be at 35% realization but could be great to work with and fills a time of the year that is not busy. That improves your utilization. But if that same client is a pain in the *** and is in the middle of your busiest time, then it may be affecting other customers in the same way. Does your data tell you this information in real time to be able to know when to cut loose a client versus when to cultivate that client?

Internal Operations

Scheduling is a dynamic and integral part of running an efficient accounting practice. If you cannot schedule time appropriately, the profitability metrics will be affected. Take a look at the key issues you are facing with operations. Do you have enough data to schedule appropriately? How much MANUAL WORK goes into scheduling? IF you capture data at the most granular of levels, there should be minimal to no manual work. How about budgeting for a job: are you able to leverage prior year data to understand where over and under runs occurred where you can make a positive change?

Leverage is an integral operational metric in public accounting. By nature, public accounting firms are built such that the chargeable hour scale goes from high to low depending on your level with staff level 1's (first year's) expected to have the most chargeable hours down to partners who are expected to have the least chargeable hours. Therefore, we need to capture data by level. When analyzing the data you captured, can you tell not only what amount of time is spent on a particular client, but also, what type of leverage is spent? You may have bad realization and utilization and when you look at the ending results of the job, you realize the partner had >50% of the hours. Regardless of how fast a partner can do a task, if staff are sitting around doing nothing and a partner is churning and burning, the firm will not be as profitable. Therefore, in an accounting firm, capture hourly data down to job, task, person and level in order to analyze this data.

How about niche – can you identify if there are trends whereby a partner is profitable in a niche versus not profitable? How about the people on those jobs? Consider capturing the data on niche verticals to identify if specialization has an impact on the metrics mentioned above.

Personnel

Public accounting is a service business. People are the greatest assets and either make or break the company. But how do you measure the important stuff? You need to define the metrics that cause your people to be your best asset and then capture data to produce meaningful results. Some of these factors are easy to see in the trends that occur from capturing your basic productivity metrics. But there is a myriad of other data that can be correlated with those utilization and realization metrics to predict if a person will or will not be successful to the company.

Specifically, take your most successful (and profitable) people and understand what makes them do well. Survey everyone and focus on specific points of data. In going through this exercise, you may realize that a manager has bad realization, but that bad realization is not because that manager is bad at their job. You may come to find out that a manager is in an industry with extreme price competition, and therefore, the fees being charged are not commensurate for the work being done. That manager's traits did not produce bad profitability, but rather, the market did and then it is a decision to stay in that market or leave.

Consider going through the HR data of your employees and understanding what characteristics on their resume and on personality profiles made them successful in your business. Not everyone is cut out for the business, but there are definitely traits that make some more successful than others. Countering this, by identifying these traits, you can also identify warning signs and areas where people need further training. A staff person may be your best staff, but once they get promoted to supervisor, they struggle. But why? You should capture data to answer this question for you. Maybe the person needs leadership training more than another person who is a natural leader. Tailoring your training approach based on data BEFORE AN ISSUE OCCURS will have a positive impact on your operating metrics.

PHASES OF A DATA ANALYTIC PROJECT OR PROCESS

We will start by looking at a data analytic process for an engagement performed by a public accounting firm. These types of projects can be as simple as a rote task or as complex as a full scale data analytics project. Big or small, some common principles can be applied to create a sound approach.

The Association of Certified Fraud Examiners identifies 4 distinct phases when approaching a data analytics project.⁹ Specifically, those phases are:

1. Planning phase
2. Preparation phase
3. Testing and interpretation phase
4. Reporting phase

Each project/task will have different phases but almost any approach will have a phase where you plan your attack, a phase where you execute on that plan, and a phase where you report the results. Let's go through an example to understand how to conceptualize a project.

Let's assume you just picked up a new client and that client told you they had a fraud uncovered during fiscal year 2020 that their previous auditor did not catch. You have been hired to perform the audit for fiscal year 2020. A condition of you being hired outside of performing the audit is to identify the magnitude of the identified fraud and conclude whether an opinion can ultimately be issued. Where do we begin? Let's talk through the phases of a data analytic task/project. For purposes of this manual, the following phases were utilized: planning, preparation and information gathering, execution, and reporting.

Planning Phase

In order to create an effective and efficient plan to address the needs identified, we need to spend time planning. First and foremost, we need to refine our goals and objective. Typically, one should answer the following questions:

- Is our objective to find fraud?
- What facts do we currently have?
- What is our budget?
- What deliverables are we issuing?

⁹ Association of Certified Fraud Examiners Manual (2015).

- What resources are available to us?
- What data is available to us?

After you answer those questions, you need to identify the limitations, if any. This typically evolves as more information is gathered; however, we want to be as focused as possible. Instead of saying “we are going to identify all fraud”, we need to focus our testing in certain areas. A focused objective could be “we are going to verify all cash related transactions that occurred in the bank that were posted in the general record keeping source” (i.e., general ledger). Carrying forward, we will use this as our example and build around this objective.

Next, the core of data analytics is... **Data**. We must understand the data that is available. In audits, fraud investigations, and other work common in public accounting, this data is typically structured data which is cross checked against other structured data. The most common sources of data are:

- General ledger data
- Banking data
- Billing software data
- Payroll data
- EXCEL reconciliation schedules
- Management reports

If we are dealing with a financial reporting system, it is important to create a map of the systems and how they talk to each other. It is recommended that the client fills in this information, namely, an individual with suitable knowledge on the IT side (IT accountant or IT department). As a simple example, see exhibit below for an application map that can be created within EXCEL and sent to your client to complete in the planning phase.

		Software Program 1	Software Program 2	Software Program 3	Software Program 4
		Insert dates software used	Insert dates software used	Insert dates software used	Insert dates software used
		Document Who has Access	Document Who has Access	Document Who has Access	Document Who has Access
		Internally create or 3rd party software or service org	Internally create or 3rd party software or service org	Internally create or 3rd party software or service org	Internally create or 3rd party software or service org
	Finance/Business Cycle				
	1 Revenue/Accounts Receivable				
	2 Accounts Payable/Cash Disbursements				
	3 Investments and Investment Income				
	4 Fixed Assets				
	5 Financial Reporting				
	6 Financial Close				
	7 Payroll and Human resources				
	8 Debt Management				

Exhibit – Application Map Example

The next step is to request the data and assess the tools required to complete the test. Data comes in all sizes and file types. The procedures and tools will vary widely depending on the availability of data. Data availability is one of the most significant struggles in the industry at large because of the wide variety of software platforms utilized and the evolution of “Big Data”.

Once the data sources are identified, you must understand the “structure” of the data. One must understand the fields that a system can produce, and the reliability of the query. This often takes several attempts to obtain the appropriate data.

Lastly, it is important to have a clear understanding of any pitfalls on the data entry side. This typically involves doing a walkthrough of controls to identify whether data is being appropriately entered. If there is an error on the data entry side, the data export will not be reliable.

Based on our objective to verify “existence” of transactions in the general ledger back to the bank, we first developed the following application map, as a starting point to requesting data.

System	Microsoft Dynamics GP	ADP	TD Bank - Checking	PNC Bank - Checking
Data in use	1/1/08-present	1/1/08 – present	1/1/08 – 6/30/19	7/1/19 – 12/31/20
Access	Everyone	Payroll Clerk, Chief Accountant	CFO	CFO
Platform Type	3 rd Party Software, Physical copy	Service Organization, Cloud Based	Cloud Based	Cloud Based
Finance/Business Cycle				
1 Revenue/Accounts Receivable	X			
2 Accounts Payable/Cash Disbursements	X			
3 Investments/Investments Income	X			
4 Fixed Assets	X			
5 Financial Reporting	X	X	X	X
6 Financial Close	X	X	X	X
7 Payroll and Human Resources	X	X		
8 Debt Management	X			

Considerations in this Stage

Area	Considerations
IT	What is the organization's IT structure (simple, complex)?
Computer applications	What systems are involved in financial reporting? Does the organization input everything in the G/L or does each process/practice area have a separate application?
Data integration	Automated or manual? For example, does the sales journal upload to G/L or is there a manual step in between? Internal review?
File types	What data types are available (PDF, EXCEL, text, data file) and how compatible are they with our tools?
Testing requirements	What are our testing objectives and where should our population come from? Completeness considerations?
Brainstorming	Risks of specific entity that need to be considered (e.g., overstatement of allocations).
Monitoring	What does the client look at? Is the data management is reviewing reliable?

Preparation Phase

Once data is requested and obtained, you must evaluate the **completeness of that data** as well as the **reliability and integrity of that data** before **organizing the data** and **developing the test plan** and doing any testing on the data. Data is only as good as the inputs used to create it.

Evaluating Completeness

In an audit environment, control totals are a logical starting point to assess or evaluate completeness. Summarizing data and arriving at the same control totals typically provides assurance as to the completeness of that data, especially if you are performing audit procedures over the control totals. In a fraud investigation environment, this becomes a more comprehensive task. The fraud examiner may need to rely upon computer controls and technology to assess the completeness of data.

Data Reliability and Integrity

In order to determine the reliability and integrity of data consider the source. Is it a reliable 3rd party or internal data? Scrub the data to remove duplicate, inaccurate or incomplete records in order to ensure that the data is clean without inaccuracies that could disrupt the workflow. Analysis of the anomalies should be completed to determine the cause and whether further analysis of the information should be completed or considered. The resulting data should also be normalized, or formatted in way to allow users to build queries and analysis of the data. Normalizing data includes grouping like data together as well as establishing relationships between groups or tables of data.

Using the example above, let's assume we requested and obtained the following:

- Microsoft Dynamics GP general ledger data file for period 1/1/2014-12/31/2020 in PDF format, each year was required to be separated based on system limitations
- Trial balance for each period end date from 2014 to 2020
- Download from TD Bank for the period 1/1/2014-6/30/2019 in .CSV format
- Download from PNC Bank Checking for the period 7/1/2019-12/31/2020 in .CSV format
- Scan of bank statements for all months included in above periods on hand at the client Organization

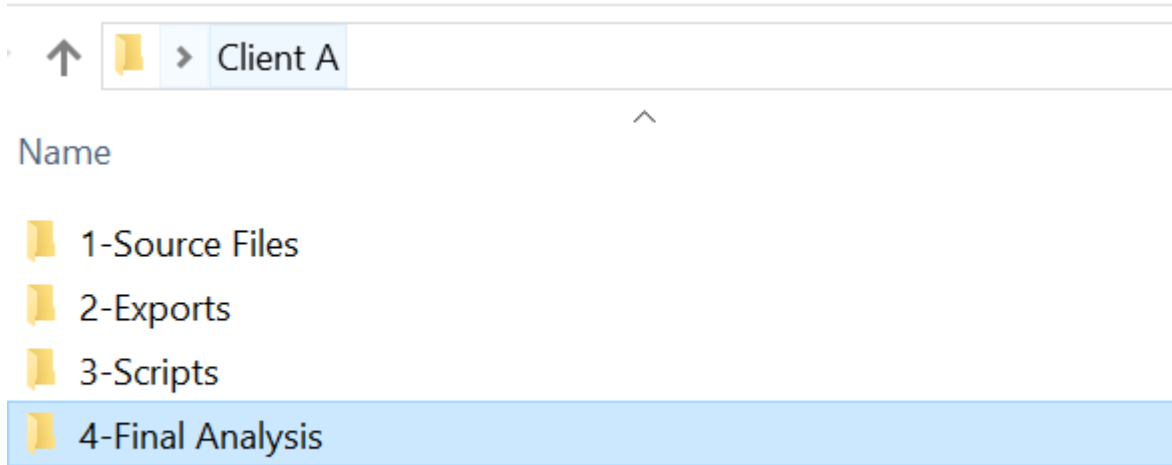
Organization

One of the most important aspects of performing procedures is being organized. Most software utilize project-based or task-based organization systems and are generally user friendly. As a best practice, it is recommended that you set up a folder for each client you work on, and within that folder, set up a specific data analytic folder (generally labeled the name of the data analytics software vendor).

The purpose of this step is to organize your files for ease of locating them when you need to utilize them and to have a central location for all files associated with a particular task. Additionally, you can save all your templates/scripts, which will allow you to save time on importing and other regular tasks. Secondly, it is important to label or organize original or native files from the calculated or result files. This can be accomplished by using a letter (such as an "r" for results) before the file name. As an example, let's assume the PDF file name is "general ledger". If you import this file and clean it up into a spreadsheet format and export, you should label the export file "r_general ledger" to clearly identify that it is an export file as opposed to an original client file.

Therefore, when you receive the client files requested, you must clearly and concisely organize the files such that you know where to find them, what was a client-provided document versus a fabricated or altered document, and what was done to the files (LOG file).

Each file would be obtained from the client, brought into a folder, and subsequently brought into our data analytic software. This then brings us to the execution phase – this is where we do stuff with the data!



Testing and Interpretation Phase

In the testing and interpretation (execution) phase, we are going to interrogate the data we have. We are going to test assumptions and draw conclusions from our data. We have the files we need, now we need to import the data into our specialized software and get moving.

Since we received the banking data as well as the general ledger data, we want to start analyzing each file independently before looking at comparing the data for anomalies. As mentioned above, we need to assess the completeness and draw conclusions on that completeness before starting our analysis.

Banking Data

Banking data was provided to us in a .CSV format. Is this banking data legitimate? Did the sender delete any rows in the data before providing it to us or alter the data? We may not know unless we received it from a quality source. Therefore, maybe we need the client to email the bank to send us a file directly (and securely of course) so that we don't need to question the integrity of the data. Sometimes this can be accommodated (typically in an audit environment dealing with 3rd party financial institutions), however, sometimes in fraud investigations this cannot be accommodated because the data that is available is all that exists. Let's assume we received it from a qualified and trusted source. Next, we want to import into our software and test control totals. In the case of banking data, there are debits (decreases to the bank balance) and credits (increases to the bank balance).

Summarization by Month. One easy analytic requires the insertion of a "month" or "period" as a field in a database. By summarizing the banking data by month, we can now compare to monthly statements with ease and quickly/efficiently test the integrity of our data. We can also perform a proof of cash if desired to tie the debits/credits by month back to the bank statement itself. If a fraudster was defalcating the bank statements, this will show up through the proof.

Gap Detection. When checks clear the bank, they leave a trail. That trail shows up on the bank statement. By performing a gap detection on the checks over the whole period, we can identify if any

checks had gone missing or if there are any unusual streams of checks (outside the normal check run). These outliers can be extracted and we can then focus specific procedures to address the risk identified.

Visualization. We can create a map or graph of the data over each year, or provide a multi-dimensional graph to show each year and look at cash activity on both the deposit side and the disbursement side. This graph can serve many purposes including identification of trends, outlier months, and unusual activity. Looking at this data visually will allow the reader a much better understanding rather than combing through lines of details.

Banking Data – Example Visual Presentations



General Ledger Data

The general ledger data was provided to us in PDF format. This means we need to convert the file into a structured data (table) using specialized software.... Or, as discussed above, since Microsoft Great Plains works with the Validis software and all the data is extracted throughout the history, we may need to change our plan from an efficiency standpoint. If that will not work, we can write a script on the first PDF file and ultimately import the file.

Assuming we imported the file, we want to verify the technical accuracy of the file. The technical accuracy of the file can be verified in several ways. Specifically, we want to first check control totals to make certain we captured all the required data in our import process. This does not only mean that

debits = credits, however, that the total debits and credits per the system or per the PDF report provided agree back to the table created through the import process. There are two reasons the reports would not agree: 1) the report was altered before it was provided to you, or 2) during the import process, the proper fields were not mapped appropriately. Typically, #2 is the reason and there is an easy way to check.

Since the trial balance and general ledger are from the same source, the expectation would be that they would agree. A common procedure performed is to “roll” the trial balance. This very simple procedure involves taking the opening period trial balance, closing out P&L accounts to retained earnings such that the opening balances are balance sheet only. This can be done as follows:

- Summarize the general ledger file by account number and perform a join with the balance sheet only trial balance.
- Perform math to add opening balance to the summarized activity.
- Perform another join with the current period end trial balance and then compare the two balances.
- If they agree, then the import performed was accurate and the data is complete.
- If they are not in agreement, a recheck of the import should be done prior to going back to the client.
- If the import was successful, then you may have identified your first finding – altered data.

The above procedures tested the completeness of the data. That completeness check is essential because it ensures that your data set is complete. Specific testing of the data will verify accuracy and sometimes specific testing (sampling procedures and substantive testing) are necessary to be able to conclude the data is captured accurately. Assuming the data is complete in this example, you can move forward to performing testing and procedures to meet your goals and objectives.

Testing is open ended when it comes to this type of engagement; therefore, it is important to keep your goals and objectives in mind and constantly revisit them. It is easy to get carried away and run down a path that wastes time and not count towards the objective.

Documentation

Equally as important as the testing is the documentation trail we leave. This trail serves the purposes of documenting the procedures that we performed as well as leaving a trail for another qualified professional to follow (whether that be a reviewer of the testing, a regulator or a staff person in a subsequent year). The data analytic software tools discussed previously typically keep a log file. This log file supports the trail left by the software and can be exported to text or to an EXCEL or another file. This trail should be kept as documentation.

Reporting Phase

In the reporting phase, your goal is to present your results to the decision makers and/or a designated body of individuals. The mechanism for the reporting is extremely important. Typically for audits, there are certain required communications and items that must be presented. For other types of engagements there are generally accepted methods of communication. Issuing a written report may be a requirement of your engagement. That being the case, the written report should contain all the technical facts, analysis and other facts that lead you to draw your conclusions and should comply with all external bodies that regulate such reporting.

You can also prepare a presentation on that information to display the data in such a way that the audience can understand it, digest it, comprehend it, and leave the presentation with a specific action item or a specific conclusion. This is where data visualization comes into play significantly. When you are trying to convince an external stakeholder, and communicate the results, it's all in the presentation. Whether you are doing a PowerPoint presentation, sending handouts, or drawing on a 3M pad, the presentation will make or break the analysis performed. Therefore, it's one thing to perform the analysis and another to report on it.

SUMMARY

In this unit we discussed the stages and steps taken in the data analytics process.

NOTES

Unit 2

Firm Technology

LEARNING OBJECTIVES

After completing this unit, participants will be able to:

- › Review the technology that can be utilized to perform data analytics
- › Discuss how technology is instrumental in the analysis process
- › Review the Computer Aided Audit Tools (CAAT)

TECHNOLOGIES THAT AID IN THE AUDIT PROCESS

When selecting a tool to utilize to perform data analytics, it is important to consider the users of the tools, the frequency of expected use and the training time that is available and afforded to those users. From our experience, if you do not spend time working with the tools frequently, the concepts of the analysis are not necessarily natural to a core accountant unless they have a computer science background or a teacher that dedicates time to help troubleshoot issues.

One of the most common challenges in the profession is access to data. And when you have access, how can you tell what is important from what is not important. Data is meaningless until we provide meaning or value to that data. At times, siphoning through mountains of data can be time consuming and often highly inefficient. To add to that complication, auditors are limited to the data available to them as part of the audit process. In order to combat these issues, specialized software exists which can aid in the audit process (computer aided audit tools or CAAT tools).

Programming Languages

Data analytic tools provide some assistance in the analytics process, nonetheless data scientists and engineers need to know how to code in at least two to three languages.¹⁰

¹⁰ “Which languages should you learn for data science?” Peter Gleeson (2017), <https://www.freecodecamp.org/news/which-languages-should-you-learn-for-data-science-e806ba55a81f/>



Python was invented in 1991 and is one of the most used programming languages used in data science. Programming languages, specifically Python, has been making waves in the accounting industry for its ease of use and versatility. Much like the subject of accounting, many believe that programming requires high level math in order to learn it, but to their surprise it actually takes very little math. When it comes to programming languages in accounting, many are familiar with VBA. VBA is used to automate processes in Excel and other Microsoft software; and although there are many benefits to learning VBA, it has its limitations. The biggest limitation is that VBA can only be used within certain software, hence why learning a more universal programming language like Python can be beneficial.

There are many resources online for Python, both for learning and practicing. One recommended resource would be the Python community's favorite, "Automate the Boring Stuff with Python". The book teaches many processes that can be automated with the use of Python and is a very good introduction to the programming language. One of the biggest advantages to learning Python is the extensive history and the large software libraries. A software library is an archive of pre-written code that is used to assist the programmers in development. In terms of analytics, Pandas is a software library that is used for data manipulation and analysis. Below is an example (www.python4cpas.com) of Pandas being used to extract data from a QuickBooks General ledger.

```
In [1]: import pandas as pd
        pd.options.display.float_format = '{:,.2f}'.format

In [2]: gl = pd.read_excel(r'npgl.xlsx')

In [3]: gl.head()
```

The first line of text is opening the software library pandas, followed by opening the general ledger EXCEL spreadsheet. The `.head()` command is used to show the first 5 lines of a table as can be seen in the figure below:

Out[30]:

	Type	Unnamed: 1	Date	Unnamed: 3	Num	Unnamed: 5	Name	Unnamed: 7	Memo	Unnamed: 9	Split	Unnamed: 11	Debit
NaN	Cash in bank - operating	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
		NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
	Check	nan	2021-01-12	nan	1888	nan	McCarthy Charities	nan	NaN	nan	Custodial funds	nan	nan
	Bill Pmt - Check	nan	2021-01-15	nan	1889	nan	East Bayshore HMO	nan	NaN	nan	Accounts payable	nan	nan
	Bill Pmt - Check	nan	2021-01-15	nan	1890	nan	Garrett Printing	nan	R 594	nan	Accounts payable	nan	nan
	Bill Pmt - Check	nan	2021-01-15	nan	1891	nan	King, Vicki - Petty Cash Custodian	nan	NaN	nan	Accounts payable	nan	nan

The script has returned a spreadsheet with cells filled with 'NaN', which stands for not a number, and several blank cells that need to be cleaned up. In the world of programming, many have run into the same issues as you and are willing to share their solutions. Below is a script found online that can be used to clean up the mess above.

```
In [4]: def acct_append(row, new_accts):
        if pd.isnull(row[1]):
            new_accts.append(row[0])
        else:
            new_accts.append('{} | {}'.format(*row))

def fix_qb_gl(gl):
    gl = gl.dropna(axis=1, how='all')
    main_acct = list(gl.index.get_level_values(1))
    sub_acct = list(gl.index.get_level_values(2))
    acct = list(zip(main_acct, sub_acct))
    new_accts = []
    acct_append(acct[0], new_accts)

    for idx, (m, s) in enumerate(acct[1:]):
        if str(m).startswith('Total'):
            m = 'DELETE'
        if str(s).startswith('Total'):
            s = 'DELETE'
        idx += 1
        acct[idx] = m, s

        if pd.isnull(m): # Fill NA if main is NA
            acct[idx] = acct[idx - 1][0], acct[idx][1]

        if pd.isnull(s): # If main is NA, then fill NA if sub is NA
            acct[idx] = acct[idx][0], acct[idx-1][1]

        acct_append(acct[idx], new_accts) # Create the new acct
    gl = gl.reset_index(drop=True)
    gl['Acct'] = pd.Series(new_accts)
    gl[['Debit', 'Credit']] = gl[['Debit', 'Credit']].fillna(0)
    gl['Net'] = gl.apply(lambda x: (x['Debit'] - x['Credit'])
                        if 'DELETE' not in x['Acct']
                        else 0), axis=1)

    gl = gl.fillna('NA')
    gl = gl.where(gl['Net'] != 0).dropna()
    columns = ['Acct', 'Type', 'Date', 'Num', 'Name', 'Memo',
              'Split', 'Debit', 'Credit', 'Balance']
    gl = gl[columns]
    gl['Date'] = gl['Date'].apply(pd.datetime.date)
    return gl
```

```
In [5]: gl = fix_qb_gl(gl)
```

```
In [6]: gl.head()
```

A common function used in EXCEL and in programming are “if statements”. The script used above has several ‘if statements’ in place that delete any empty cells and remove any cells with NaN as can be seen in the figure below:

Out[6]:

	Acct	Type	Date	Num	Name	Memo	Split	Debit	Credit	Balance
1	Cash in bank - operating	Check	2021-01-12	1888	McCarthy Charities	NA	Custodial funds	0.00	6,000.00	47,750.00
2	Cash in bank - operating	Bill Pmt -Check	2021-01-15	1889	East Bayshore HMO	NA	Accounts payable	0.00	1,250.00	46,500.00
3	Cash in bank - operating	Bill Pmt -Check	2021-01-15	1890	Garrett Printing	R 594	Accounts payable	0.00	306.00	46,194.00
4	Cash in bank - operating	Bill Pmt -Check	2021-01-15	1891	King, Vicki - Petty Cash Custodian	NA	Accounts payable	0.00	100.00	46,094.00
5	Cash in bank - operating	Bill Pmt -Check	2021-01-15	1892	Rand Properties	NA	Accounts payable	0.00	3,000.00	43,094.00

Now that we have our data in place, we can set some criteria to take a deeper dive into the general ledger accounts. Let’s say I only want to examine any entries made under the type “general journal” and only entries over \$45,000 in that category, I would proceed with following code:

```
In [13]: balance = gl.where(abs(gl.Balance) > 45000).dropna()
```

```
In [14]: balance[balance.Type == 'General Journal']
```

Out[14]:

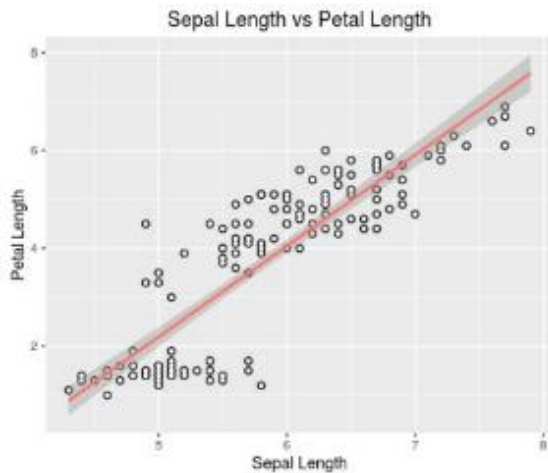
	Acct	Type	Date	Num	Name	Memo	Split	Debit	Credit	Balance
798	Temporarily restrict net asset Use restricte...	General Journal	2021-03-31	0308	HHS:Research This Year	1st qtr revenue release	Satisfaction of use restric	150,254.83	0.00	150,254.83
822	Unrestricted net assets Transfers to/from un...	General Journal	2021-03-31	0308	HHS:Research This Year	1st qtr revenue release	Satisfaction of use restric	0.00	150,254.83	-150,254.83
843	Assets released fr restrictions Satisfaction...	General Journal	2021-03-31	0308	HHS:StudentEd This Year	1st qtr revenue release	Satisfaction of use restric	121,751.09	0.00	121,751.09
845	Assets released fr restrictions Satisfaction...	General Journal	2021-03-31	0308	HHS:StudentEd This Year	1st qtr revenue release	Satisfaction of use restric	0.00	121,751.09	-121,751.09

The first line of code is creating a variable for any balance that is over \$45,000 and the code that follows is requesting that the type of entry be “General Journal”.

As you can see, the benefit with code writing is not the initial time spent; it is the benefits derived the second, third, and nth time that you use the tools. As you automate scripts and codes, you can dedicate more time to performing auditing and analysis type functions and less time doing rote tasks such as converting files.



R was invented in 1995 and is mainly used for statistical data modeling in data science.¹¹



Java virtual machine (Java) is helpful in the ETL process and offers computation with machine learning algorithms. Java also supports portability between platforms.¹²

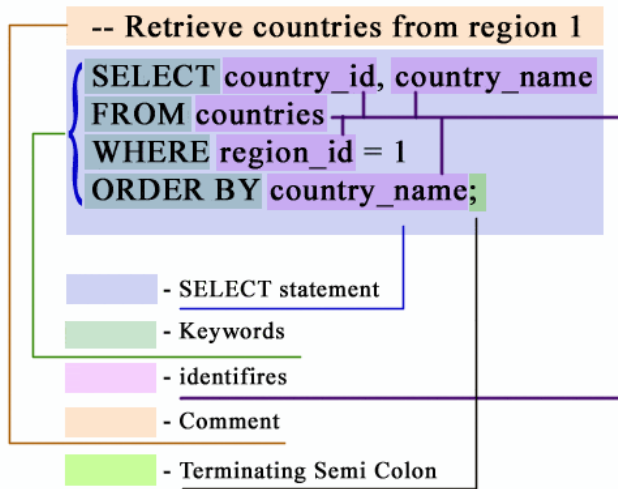
¹¹ "Which languages should you learn for data science?" Peter Gleeson (2017), <https://www.freecodecamp.org/news/which-languages-should-you-learn-for-data-science-e806ba55a81f/>

¹² "Which languages should you learn for data science?" Peter Gleeson (2017), <https://www.freecodecamp.org/news/which-languages-should-you-learn-for-data-science-e806ba55a81f/>



Structured query language (SQL) is used in processing data before storing it into a relational database. Data is kept in the database for easy retrieval and access.¹³

SQL Language Elements



[<https://www.w3resource.com/sql/sql-syntax.php>]

Other Technology



Hadoop

Hadoop is mainly used for Big Data storage. Hadoop stores both structured and unstructured data. Hadoop integrates with other technologies.¹⁴

APACHE SPARK

Hadoop can be used in conjunction with NoSQL databases and Apache Spark. Spark is a general-purpose distributed processing engine that performs high performance, in-memory processing of large data sets.¹⁵

¹³ “Which languages should you learn for data science?” Peter Gleeson (2017), <https://www.freecodecamp.org/news/which-languages-should-you-learn-for-data-science-e806ba55a81f/>

¹⁴ “WHAT IS APACHE HADOOP?” <https://mapr.com/products/apache-hadoop/>

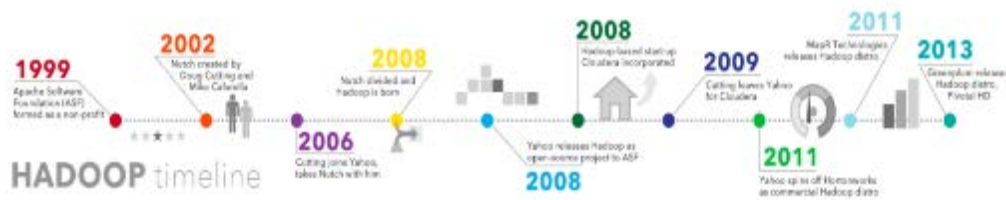
¹⁵ “WHAT IS APACHE HADOOP?” <https://mapr.com/products/apache-hadoop/>

MapReduce

MapReduce is a processing technique and a program model for distributed computing.

MapReduce programming in conjunction provides valuable insights on Big Data such as:

- Scalability. Processing large data sets in the Hadoop Distributed File System (HDFS).
- Flexibility. Enabling easier access to multiple sources and types of data.
- Speed. Minimal data movement with parallel processing. Hadoop offers fast processing of massive amounts of data.
- Simple. Developers can write code in a choice of languages, including Java, C++, and Python.



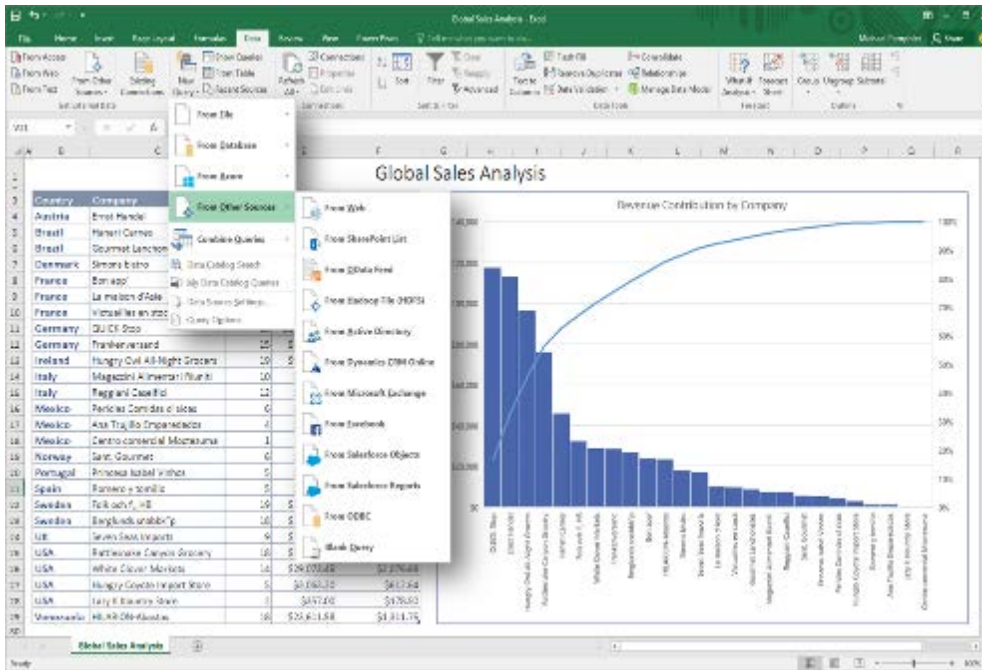
[https://www.sas.com/en_us/insights/big-data/hadoop.html]

Visualizations and Dashboards

Excel

Microsoft Excel is frequently used for data analysis, is the most commonly used tool in the profession and is used to capture everything from a reconciliation of prepaid expenses to capturing thousands of rows of data generated for analysis purposes. EXCEL has strong computing capability and oftentimes is the most comfortable, user friendly tool to use. Microsoft EXCEL is the “ole standby” in the industry, however, is not fail-proof. Untrained users can make material errors as there are no true “checks” to the analysis performed other than sum formulas. Excel offers basic operations reviewing and reading data into Excel using various data formats, organization, and manipulation.¹⁶ Basic tasks such as filtering, sorting, creating pivot tables, and writing “IF” formulas are possible with EXCEL. The platform works well, however, does not have the same processing engine as an IDEA or ACL technology for large data sets.

¹⁶ “15 Excel Data Analysis Functions You Need to Know”, <https://excelwithbusiness.com/blog/15-excel-data-analysis-functions-you-need-to-know/>



Tableau

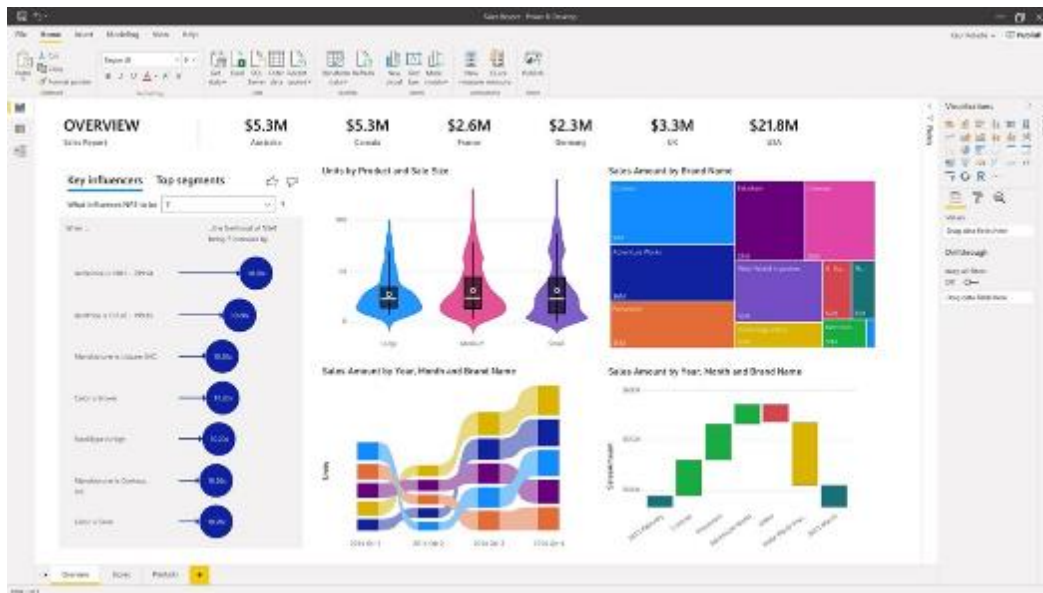
Tableau is a data visualization tool used in data science and business intelligence. Tableau provides insightful and colorful visualizations in an interactive and impactful way.¹⁷



¹⁷ Pandey, Parul. "Data Visualisation with Tableau." DataCamp Community, September 24, 2018. <https://www.datacamp.com/community/tutorials/data-visualisation-tableau>.

Microsoft Power BI

Power BI is a suite of business analytics tools that deliver insights throughout an organization. Power BI is used to connect to hundreds of data sources, simplify data prep, and drive ad hoc analysis. The tool is used to produce reports, then publish them for an organization to consume on the web and across mobile devices. You can create personalized dashboards with a unique, 360-degree view of a business.¹⁸ Microsoft Power BI is utilized to gather data and visualize it in such a way that the data can be useful.



Computer Aided Audit Tools (CAAT)¹⁹

Computer Aided Audit Tools (CAATs) are packages designed for data analysis. CAATs are typically used to analyze and manipulate data. In the context of accounting, CAATs are additionally used to compare both financial and nonfinancial data. CAATs are typically used to streamline the audit process.²⁰

CAAT Features

Audit functions that are normally performed manually can now be standardized by accounting/CAAT software. Once the data is verified using the CAAT system, the data is retained so it can be reused in the audit process to identify errors and segregate transactions within accounts.

¹⁸ <https://powerbi.microsoft.com/en-us/>

¹⁹ <https://www.auditnet.org/audit-library/computer-assisted-audit-tools-and-techniques-caatt>

²⁰ A Guide to Computer Assisted Audit Techniques (2006), http://www.mtc.gov/uploadedFiles/Multistate_Tax_Commission/Audit_Program/Resource/caat.pdf

CAATs generate both customized reports of the findings and produce a standard audit trail consistent with the Generally Accepted Accounting Principles (GAAP).²¹

The use of data analytics in CAATs is useful to:

- Identify anomalies;
- Spot patterns and risk indicators;
- Analyze datasets instead of population sample;
- Analyze data from disparate datasets; and
- Discover multiple systems or locations of data.²²

Common CAAT Software



IDEA® software is a powerful and comprehensive data analysis tool that allows professionals to gather evidence, discover trends, assess risk, and provide the business intelligence needed to make informed decisions from multiple data sources.²³

IDEA is an advanced data analytics tool built for internal/external auditors which provides full functionality from importing complex files to performing extraction and interrogation of data. The program can read flat or relational databases, spreadsheets, print files, and many more file types. Commands are both pre-programmed (click of a button), as well as customizable with the scripting feature.

IDEA also has several add-ins including “smart-analyzer” and a data visualization add-in. The smart analyzer allows the user to have several pre-built scripts/functions to run common fraud and other audit related tests such as a journal entry test or an interrogation of an accounts receivable subsidiary ledger. The data visualization tool allows the data to be mapped pictorially to provide an added level of analysis and documentation for the audit file.

²¹ A Guide to Computer Assisted Audit Techniques (2006), http://www.mtc.gov/uploadedFiles/Multistate_Tax_Commission/Audit_Program/Resource/caat.pdf

²² <http://www.bcscpa.com/cpe-presentations/CAATs1.pdf>

²³ <https://www.audimation.com/wp-content/uploads/2019/12/caseware-idea-product-sheet.pdf>

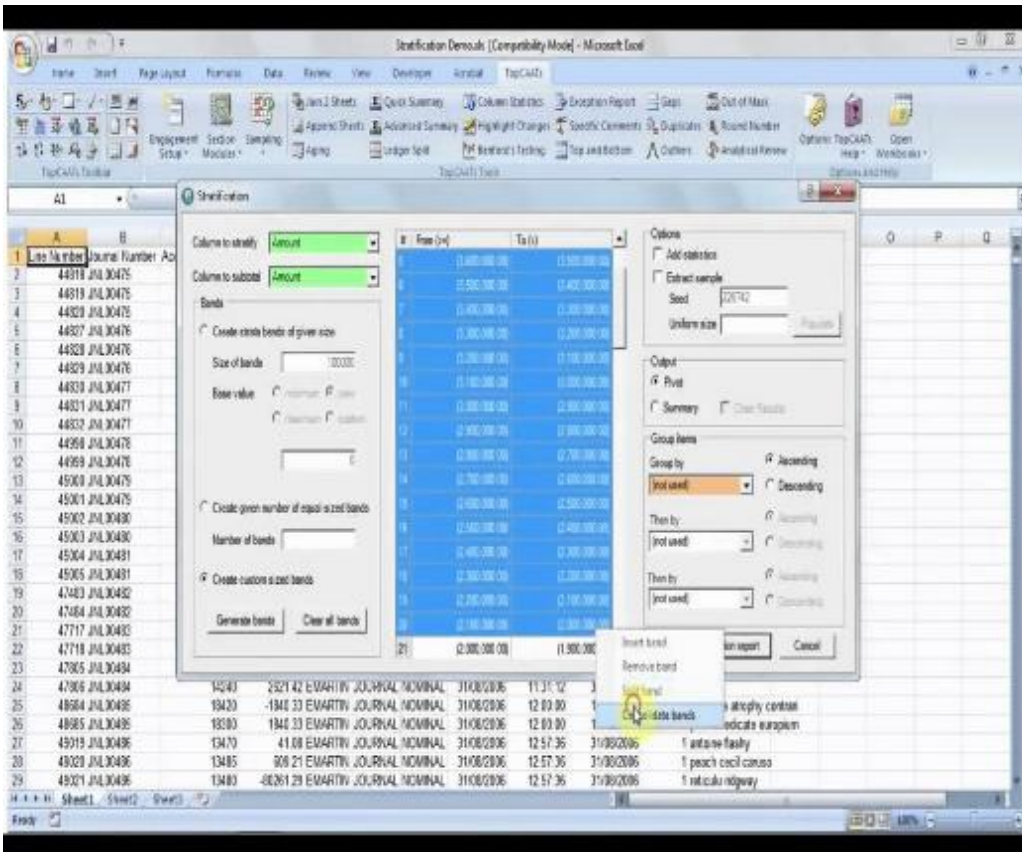
ACL²⁴

ACL is an advanced data analytics tool built for internal/external auditors which provides full functionality from importing complex files to performing extraction and interrogation of data. ACL was built on a computer programming light platform which provides for functions to be developed and written to standardize and streamline processes and functions to create efficiency and effectiveness. ACL has significant advantages when automating tasks using the scripting function. Additionally, ACL can handle large data sets and has strong processing power.

Like IDEA, the program can read flat or relational databases, spreadsheets, print files, and many more file types. Commands are both pre-programmed (click of a button), as well as customizable with the scripting feature.

TeamMate Analytics

TeamMate Analytics is based in Microsoft EXCEL as an add-in. TeamMate Analytics, is widely used by sole practitioners and Big Four accounting firms. TeamMate offers a suite of more than 150 CAATs that assist auditors with performing data analysis and delivering significant value within their organizations, internal and/or external clients.²⁵



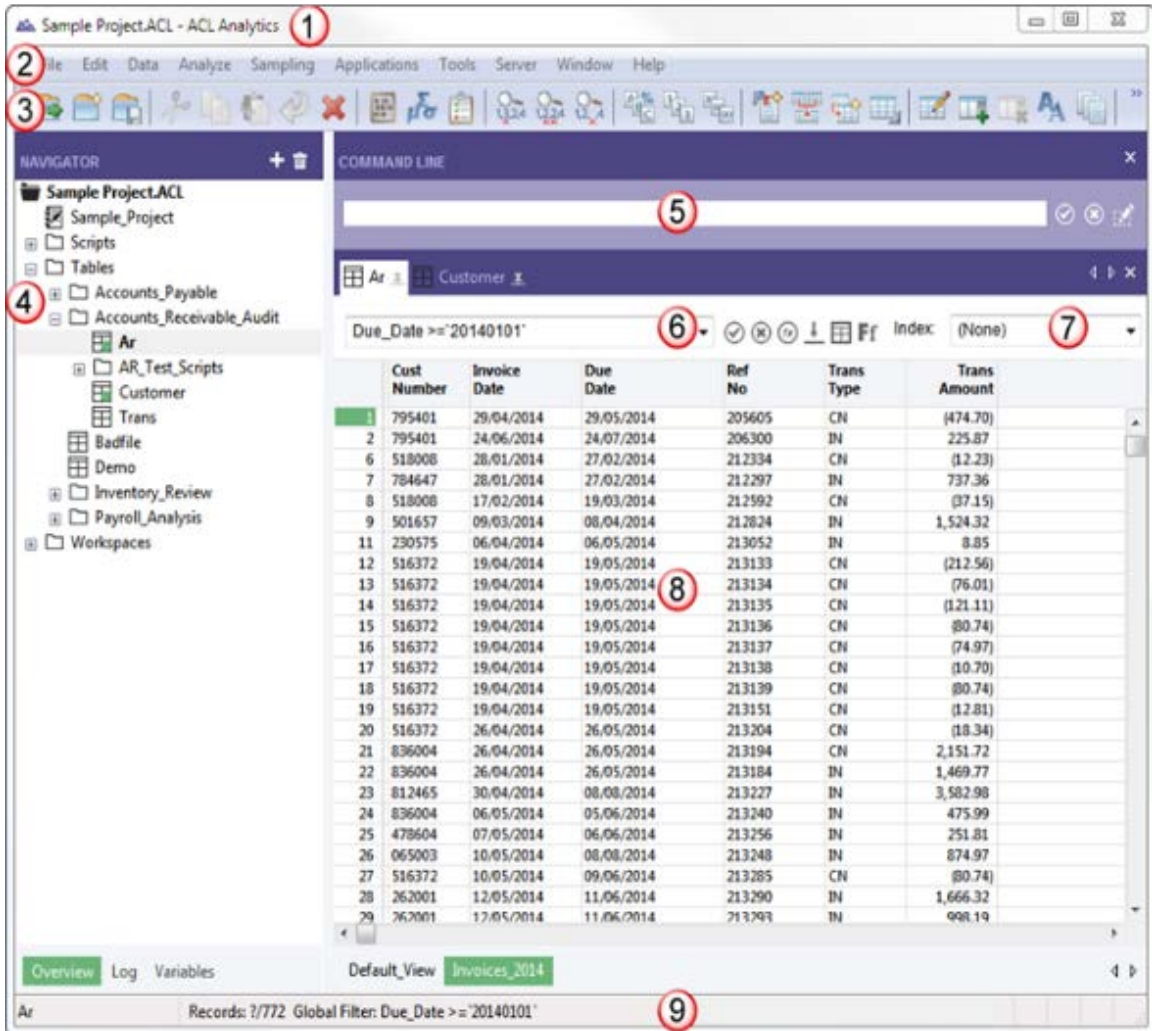
²⁴ <https://www.acl.com/>

²⁵ <http://www.teammatesolutions.com/data-analytics.aspx>

Galvanize Analytics



Galvanize's audit analytics suite of software products offers data extraction, analysis, and fraud detection features. Galvanize has comprehensive audits to identify errors, fraud, and anomalies to support of regulatory, operational compliance, and efficiencies.²⁶



Validis²⁷

Validis is an interactive database that enables small and medium enterprises (SMEs) to share financial data history directly from the software in a standard format. Validis can be used to speed up loan decisions, initiate audits, and foster business intelligence.²⁸ Validis has developed technology which connects directly with accounting packages and extracts core data such as the general ledger, accounts

²⁶ https://help.highbond.com/tech_briefs/ax/ax_54x_tech_brief.pdf

²⁷ <http://validis.com/>

²⁸ https://validis.com/wp-content/uploads/2019/08/US-DataShare-2pp_Ext-Mar191.pdf

receivable subsidiary ledger, and accounts payable subsidiary ledger. This tool allows auditors to receive a standard set of data with pre-written reports that can be utilized in the audit process. The tool is designed to alleviate the pain of tracking down data and putting it into an easy to use format (cloud view or EXCEL).

Validis is currently compatible with the following accounting packages in the U.S.:

- QuickBooks
- QuickBooks Online
- Xero
- Sage 50 (Peachtree)
- Sage ACCPAC
- Microsoft Dynamics GP
- Microsoft Dynamics NAV

This technology is an advancement in the way to approach an audit. Historically, an auditor can spend a significant amount of time working with data to normalize it and put it into a format that is usable. This tool takes that normalization and time away by generating that data for the auditor. The tool is secure and requires authorization of the client to release the data.



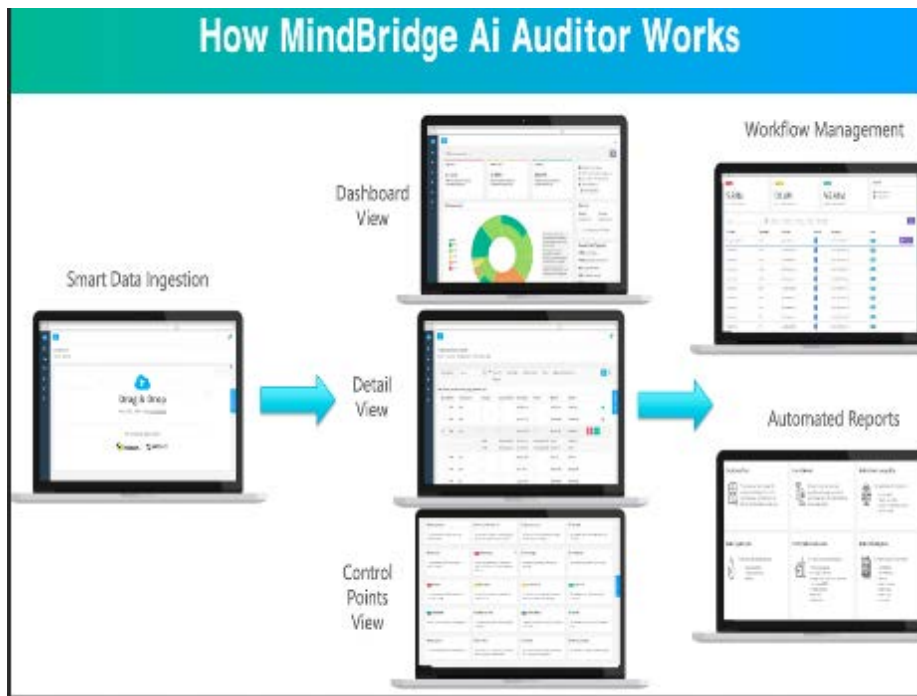
MindBridge AI Auditor was the world's first AI-powered financial auditing solution. MindBridge offers the following features:

- Analytics: Improving testing and audit planning using AI-driven insights through the client's dataset, such as monthly sales and cost of goods sold.
- Journal entry testing: Performing faster, complete, and comprehensive testing of the client's general ledger.
- Search: Extraction of transaction details within a general ledger, notwithstanding the level of flagged risk.
- Risk identification: identifying high-risk vendors by analyzing the standalone risk of invoices.²⁹

²⁹ Doeren Mayhew improves audit engagements using AI, Mindbridge, <https://www.mindbridge.ai/wp-content/uploads/2019/10/MindBridge-AiAuditor-CaseStudy-DoerenMayhew.pdf>

MindBridge AI Auditor: “The Solution”

- Uses Artificial Intelligence AND Machine Learning.
- Purpose built for Accounting industry and exceeding standards.
- Detects Anomalies, Irregularities, and Outliers in Financial Data that are missed with the human eye or a rules based analytics solution
- An Extensible AI platform to support Audit & Assurance, Forensics, Advisory Services, etc.
- Speed, Ease of Use, Granularity & Completeness
- Greater Insight, Cost Savings and Risk Reduction / Assurance



[<https://s3.amazonaws.com/primeglobal-assets/downloads/MindBridge-PPT-PrimeGlobal-August-23-2018-Final.pdf?mtime=20180904145053>]

PDF Conversion Tools

One of the most common challenges faced in the accounting profession is obtaining usable data. This is especially true when it comes to print or PDF type files. Tools such as IDEA, ACL and Teammate Analytics come with converters/conversion wizards that guide you through the process. However, if you do not have access to those tools, other tools exist which can help with software conversion. Assuming you do not have the advanced tools, consider using the Adobe built-in tool.

- Adobe 10 – Click on File>Export to>Spreadsheet>Microsoft Excel workbook
- Adobe pre-10 – File>save as>label as an Excel spreadsheet as the file type to save as

The data will export to an Excel spreadsheet. Warning: Occasionally, there will be required cleanup based on the data exported and the complexity of that data.

DATA ANALYTICS IN THE ACCOUNTING FIELD

CPA Exam

AICPA recognized that the incorporation of data analytics in the audit process is inevitable. As a result, data analytics was added to the CPA exam in 2019, affecting both the BEC and AUD sections of the exams.³⁰

Public Accounting

The Big Four have all welcomed AI initiatives. They generally offer AI services that fall into one or more of these categories:

- AI products that provide the associated end-customer benefits
- AI processes where AI technology is used to streamline daily workflow and operations in order to automate and increase day-to-day productivity
- Intelligence or insight services in which AI is used to provide clients targeted advice to support informed and strategic decisions³¹

Deloitte

Deloitte is part of an AI initiative called Catalyst. Deloitte provides Catalyst companies with funding and access to their client base in order to partner and translate AI technologies into practical business solutions for their clients.³²

Additionally, Deloitte has partnered with IBM Watson to develop LeasePoint, an AI-enabled system used in the leasing industry to teach and develop an end-to-end leasing portfolio.³³

³⁰ "Uniform CPA Examination Blueprints", AICPA (2019), <https://www.aicpa.org/content/dam/aicpa/becomeacpa/cpaexam/examinationcontent/downloadabledocuments/cpa-exam-blueprints-effective-july-2019.pdf>

³¹ Cognitive technologies: The real opportunities for business, David Schatsky (2015), <https://www2.deloitte.com/us/en/insights/deloitte-review/issue-16/cognitive-technologies-business-applications.html>

³² AI in the Accounting Big Four – Comparing Deloitte, PwC, KPMG, and EY, Daniel Fagella (2019). <https://emerj.com/ai-sector-overviews/ai-in-the-accounting-big-four-comparing-deloitte-pwc-kpmg-and-ey/>

³³ Cognitive technologies: The real opportunities for business, David Schatsky (2015), <https://www2.deloitte.com/us/en/insights/deloitte-review/issue-16/cognitive-technologies-business-applications.html>

EY

EY is currently implementing AI to automate the auditing process and leave employees with more time to participate in the judgment and analytical part of the process.³⁴

PWC

PWC is developing an AI audit tool that uses reinforcement learning and becomes more capable to make decisions without interference with every audit (a common capability for ML applications).³⁵

KPMG

KPMG has built a suite of AI tools called KPMG Ignite that include the following:

- Call Center Analytics Engine
- Event Prediction Tool
- Document Compliance Tool³⁶

BDO

The national CPA firm BDO uses Big Data analytics to identifying risk and fraud during audits.³⁷

Private Accounting/Sole Practitioners

Clients are unique, with different goals and values. CPAs are now incorporating data analytics in their practice to better serve their clients. As discussed in previously, many CAAT and tax software aggregate client data show patterns and irregularities. Data analytics can help CPAs discover multiple opportunities and provide value to clients through various strategies. Three main areas where analytics can be instrumental in client services are:

- Financial planning;
- Asset protection; and
- Tax planning and preparation³⁸

³⁴ Cognitive technologies: The real opportunities for business, David Schatsky (2015), <https://www2.deloitte.com/us/en/insights/deloitte-review/issue-16/cognitive-technologies-business-applications.html>

³⁵ Cognitive technologies: The real opportunities for business, David Schatsky (2015), <https://www2.deloitte.com/us/en/insights/deloitte-review/issue-16/cognitive-technologies-business-applications.html>

³⁶ Cognitive technologies: The real opportunities for business, David Schatsky (2015), <https://www2.deloitte.com/us/en/insights/deloitte-review/issue-16/cognitive-technologies-business-applications.html>

³⁷ "10 companies that are using big data", Eleanor O'Neill (2016), <https://www.icas.com/thought-leadership/technology/10-companies-using-big-data>

³⁸ Leveraging Data Analytics (2019), Ernie Guerriero, <https://www.cpajournal.com/2020/01/15/leveraging-data-analytics/>

SUMMARY

In this unit, we went over the various tools and technology used in the data analytics process. The data analytics process is detailed and time intensive. As a result, organizations employ a suite of tools and technologies to effectively manage the process.

NOTES

Unit 3

Data Analytic Techniques

LEARNING OBJECTIVES

After completing this unit, participants will be able to:

- › Use data analytics in sampling techniques
- › Use data analytics to test journal entries
- › Use data analytics to organize data for review and analysis

INTRODUCTION

Data analytical procedures can be utilized on a variety of accounting/attest related engagements, including but not limited to:

- Audits
- Reviews
- Forensic and Litigation Engagements
- Due diligence projects
- Consulting engagements

Regardless of whether you work in industry or in public accounting, data analytics can and will apply. Organizations need to take advantage of their data to answer unique business questions.

ACQUIRING DATA

Before any function can be performed, we first need to understand the basics of acquiring data. The majority of projects and initiatives fail not because data exists. Typically, it is either data cannot be

accessed or data is maintained in silos and is not tapped and combined to produce meaningful information. Data can be a very powerful tool if harnessed appropriately.

The reporting that is required to execute any function using technology operates on a simple principle: garbage in garbage out (GIGO). You can have the most powerful tools offered and you can provide them to all staff. However, if you cannot successfully acquire and validate data, you will have no use for those tools. Additionally, if the data you acquire is poorly kept and maintained, then the tools also will serve minimal utility except to tell you that the data cannot be relied upon (which is not the point!!).

Client data entry is extremely significant. Advising your client on data entry is an important first step to receiving data that you can work with. If you are requesting a manual report that a client needs to prepare using data from multiple sources, typically that means data is not being captured in a usable manner. Data cannot always be captured to remove manual tasks; however, many times there are solutions. Understanding what data your client captures and what data they do not capture (but could) is usually identified in walkthroughs or in testing. How great would it be if you could not only save yourself time (by having data available to perform better analysis), but also to help your client enter data appropriately so it saves them time and creates more usable information!

In an audit environment, data comes from various sources. Each engagement carries different challenges, different systems, and different sets of data from a multitude of software packages. When making a request for data, you first need to understand the capabilities of the software platform being used. In a typical audit, you will obtain a main transaction register (general ledger) and the associated subsidiary ledgers or journals. The subsidiary ledgers may be integrated (direct feed to the general ledger) or not integrated. The distinction is important because for instances where systems are not integrated, the auditor needs to identify how the systems interface and/or how the data goes from one system to another.

When a subsidiary ledger is directly integrated, your job typically is easier because there are less considerations and by doing a walkthrough, you can rely upon those automated IT controls. However, when a subsidiary ledger is not integrated with the main transaction register, there are many other considerations including but not limited to:

- How do the systems communicate with each other?
- Do the systems integrate using a manual process or automated process?
- Are IT general controls in place to capture all the data?
- Does the organization have appropriate controls in place to reconcile the data between the systems?
- Are the systems managed by the same people or different people?
- Are both systems following the same rules (for example, is one cash basis and one accrual basis)?

In practice, we see the revenue or billing ledger is oftentimes not integrated with the general ledger. This is especially true in the startup/tech company environment because the software utilized is typically internally developed (which creates its own set of challenges).

EXAMPLE

Not-For-Profit Organization

Many not-for-profit organizations utilize QuickBooks for their main general ledger accounting package. The development (fundraising) department utilizes a separate contribution tracking/CRM system to capture donations. Through inquiry and walkthrough procedures, we identify that the development department and the accounting department do not talk to each other. We also noticed that the development department is the first one to receive the checks, update their contribution tracking/CRM system, and then the information is provided to the accounting department (only with checks). What controls need to be in place to prevent an error?

- Daily cash receipts totals should be matched between the two systems
 - Reconciliations should be prepared to identify differences between the basis of accounting utilized
 - Accounting and Development should have a periodic meeting to understand the data being captured (cash basis vs. accrual basis)
 - Accounting should speak with the software providers to identify a way to integrate the data such that there are not manual postings (duplicate effort)
-

Data File Type or Format

Data can be exported from systems in a multitude of formats. Some of these formats are readable by a human (machine printed, PDF, text, EXCEL) and some are only readable through computer software (Delimited Text, database files, etc.). Additionally, files may be readable, however, their utility may be limited if the files cannot be searched or the data cannot be extracted into a structured format.

Therefore, to utilize technology, one must determine which of the formats is compatible with your technology tool (IDEA, ACL, etc.). If you spend 50% of your data analysis time on importing the data and 50% on carrying out the analytics tasks, the efficiency may not outweigh the time cost of using certain software. Therefore, it is important to fully understand the data and acquire the data in the format that will take the least amount of time to begin working with it. If the client has an EXCEL file and it is very messy (not structured in a table) or a PDF file, the question one should ask is “which takes less time to convert into a table”?

If you have a continuous task, such as an annual task, monthly task or weekly task, scripting the process may be worth the investment. Certain technologies, such as IDEA or ACL, allow the user to write commands that can be executed again such that a consistent file format and file is provided (regardless of the data on that file). Therefore, running a task the first period may be time consuming and costly, however, running the task for a second period can result in significant time savings if data can be scripted. This is common when utilizing ACL software to script the import function and save

those scripted functions. The import process is arduous the first time around, however, upon creation of a script, the benefits are significant.

In a typical audit engagement, the following files are obtained from a client's master accounting package and subsidiary ledgers:

- Trial balance at year end
- General ledger ("GL") for period under audit
 - Account number
 - Account description
 - Sub account number
 - Sub account description
 - Batch #
 - AJE# if different than batch
 - Journal source code
 - Transaction date
 - Period posted (if different than transaction date)
 - User posted by
 - Time posted
 - Description of transaction
 - Debit amount
 - Credit amount
- General ledger for subsequent period (same information as above GL)
- General journal entry posting file
- Cash disbursement (or check) register for period under audit
- Cash disbursement (or check) register for subsequent period
- Cash receipt journal for period under audit by customer by invoice

- Cash receipt journal for subsequent period
- AP aging schedule at year end by vendor by invoice
- AR aging schedule at year end by customer by invoice

Technology Alert! There are now tools available and ones being built which can create a direct bridge into an accounting package and extract out all the reports, producing them in a readable format with the click of a button! See previous section for examples of these tools.

Many times, the direct contact person at a client does not know of the capabilities of systems. If the client is agreeable to call the software vendor with the auditor, you can talk through the files that can be produced and you may find out that certain file types are accessible.

One last point of note. The general ledger oftentimes contains substantial data usually in either a pre-built report or one customized by the IT department staff to run specific fields. Depending on the size of the general ledger, this can create a daunting and extraneous burden on the network when running these reports; therefore, when requesting reports, be flexible and request them in advance. It is important to be conscious of this fact and potentially consider running reports at off peak hours to avoid crashing a server. The better your communication lines and candor, the better the success.

SAMPLING

All audits incorporate an element of sampling. Considering an audit provides “reasonable” and not “absolute” assurance, auditors must accept a level of sampling risk. AU-C section 330, *Performing Audit Procedures in Response to Assessed Risks and Evaluating the Audit Evidence Obtained*, and AU-C 530, *Audit Sampling*, recognizes that auditors are often aware of items in account balances or classes of transactions that likely contain misstatements. Auditors consider this knowledge in planning procedures, including audit sampling. They usually will have no special knowledge about other items in account balances or classes of transactions that, in their judgment, will need to be tested to fulfill the audit objectives. Auditors might apply audit sampling to those account balances or classes of transactions. AU-C section 530 provides guidance for planning, performing, and evaluating audit samples using two approaches: nonstatistical and statistical.³⁹

According to AU-C section 530, *Audit Sampling*, audit sampling is “the selection and evaluation of less than 100 percent of the population of audit relevance such that the auditor expects the items selected (the sample) to be representative of the population and, thus, likely to provide a reasonable basis for conclusions about the population”.

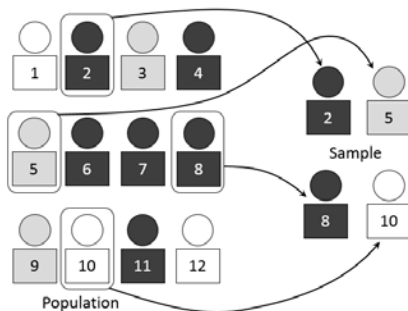
In an audit environment, there are two types of sampling: statistical and nonstatistical. Of these types, there are several approaches that can be taken. Let’s look at a few different types of sampling methodologies and how a computer system can help.

³⁹ Derived from AICPA Audit Sampling Guide

Random Sampling

Random sampling involves selecting items in a population whereby all items have an equal chance of selection. When testing internal controls, it is common place to perform a random sample because it does not allow for bias (each item has an equal weight to be selected). Many software, including Microsoft EXCEL, have the capability of performing a random sample. It is important that when defining the sampling approach as statistical that a computer system is utilized because it allows for a true random, or a statistically valid, sample. Often, we may define our sample as random, but we are really doing it haphazardly because our selections are made with our eyes (most humans gravitate towards large numbers in a population).

Here is an example of a random sample. There is no logic behind the selection except a randomly generated selection.



Random sampling using technology (ACL input screen):

The screenshot shows the 'Sample' dialog box in ACL. The 'Main' tab is selected. The 'Sample On...' button is visible. The 'Sample Type' section has 'Record' selected. The 'AMOUNT' dropdown is set to 'Sample Parameters'. Under 'Sample Parameters', 'Random' is selected. The 'Size' field is set to 60, 'Seed' is empty, and 'Population' is set to 408875. The 'Algorithm' is set to 'Mersenne Twister'. There are 'If...' and 'To...' buttons with empty text boxes. The 'Local' and 'Use Output Table' checkboxes are checked. The 'OK', 'Cancel', and 'Help' buttons are at the bottom.

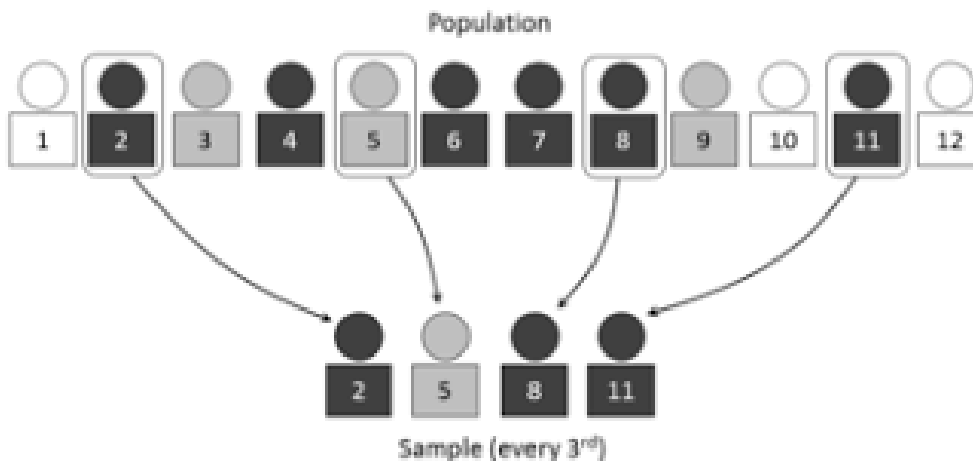
Interval or Systematic Sampling

This type of sampling involves selecting items in a population based on a random starting point and a fixed interval. The starting point is chosen by the auditor and the fixed interval typically is chosen based on the desired sample size. In systematic sampling, you are selecting every “nth” item in the population. The “n” would be the interval.

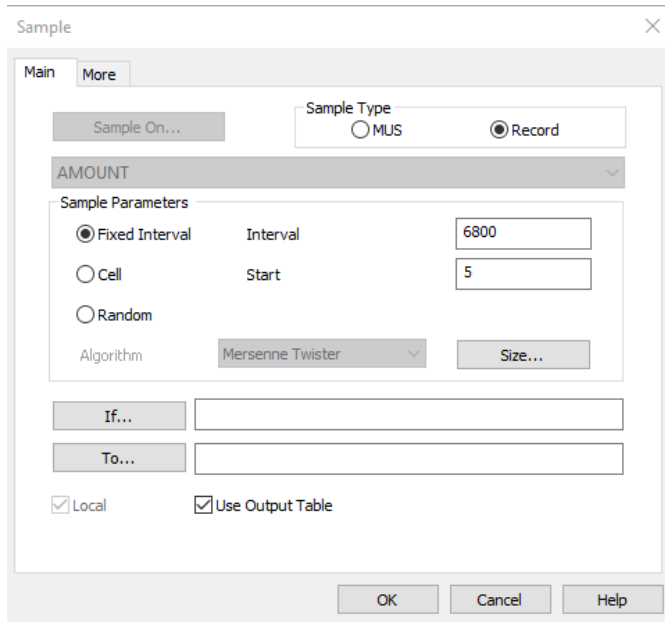
EXAMPLE

You are assigned to perform a test of controls over purchasing approval. The purchase journal utilized has 1,000 unique items, or sampling units. You have determined that a sample size of 60 is appropriate. You have decided to start at item #1 in the population as the starting point. In this case, you would select every 16.667 items (in practicality, you would round to every 16 items). So, you would pick #1, #17, #33, #49, ..., #945.

Example picture of systematic or interval sampling:



Systematic or interval sampling using technology (ACL input screen):

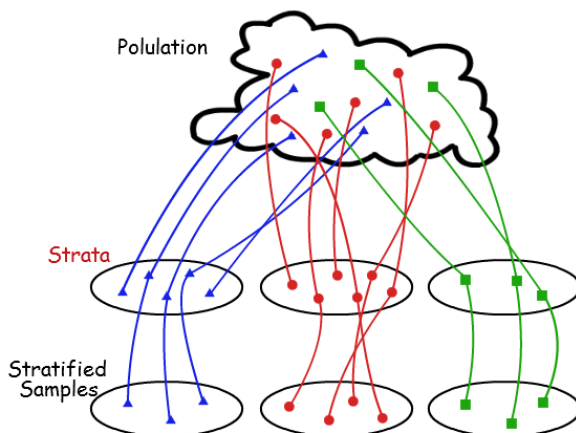


*Population of 408,000, sample size of 60, selection of every 6,800 items starting at 5.

Stratified Sampling

Stratified sampling involves breaking apart a population into sub-populations. This type of sampling can help to focus risk and sample sizes based on population characteristics. The goal of stratified sampling is to break apart the population into closely related or statistically similar items.

Example of stratified sampling:



EXAMPLE 1

Let's assume a manufacturer sells 3 products, each with a different price (A - \$50, B - \$100, C - \$1,000). In this case, the auditor may determine that the sampling should be done such that each product has an equal chance of being selected. In this case, A, B, and C would be dissected into strata and used as the basis for sampling. From that point, the auditor may evaluate the population size in each stratum and weigh the sampling methodology such that we selected more from the larger population. That is where technology comes in.

EXAMPLE 2

A not-for-profit organization receives funding from the Federal government. The auditor has been engaged to perform a single audit over the major federal program. The auditor has determined that allowable cost is the most relevant item to test since the grant is paid to the organization based on allowable expenditures incurred. The budget has 5 main categories: 1) payroll and benefits, 2) consultants and professional fees, 3) occupancy, 4) other direct costs, 5) indirect costs.

A common way to stratify these populations is to do so based on the budget category because each population may have different sets of controls. In the above example, Category 1, Categories 2-4, and Category 5 were determined to be the strata because each population is required to be tested differently. The auditor then pulled a stratified sample of 40 from Category 1, 40 from Categories 2-4, and no sample from Category 5 as the population can be tested substantively/analytically.

Why Stratify?

By approaching the sampling from this methodology, we can break apart populations into smaller segregated groups which provide more effective and focused samples, allowing the auditor to be more efficient with testing and analysis. Also, when an error is found, the auditor now can extrapolate among the strata and not based on the whole population to provide for more effective and precise extrapolation of error.

When pulling from a general population in total, large items may skew the data. If an error is found on the large item, it may not be indicative of smaller items. Therefore, stratification aids in more *effective* sampling.

Other Examples in Practice:

1. Single Audit Allowable Cost Testing

- a. The client tells you they do not keep receipts for expenditures less than \$25. The approval process is still in place, but there is not much to look at beyond an approved expense report. In doing your testing, you may consider eliminating or segregating the bucket of expenses less than \$25. If you include that as a stratum, you can identify the total \$ amount of those transactions and realize it is insignificant to the program as a whole and exclude them, focusing your efforts on the more significant transactions.

2. Check Disbursement Stratification
 - a. Assuming the client has a check signing policy where all checks > \$5,000 require two signatures, you may want to stratify the population over/under that amount to get an even distribution.
 - b. Client has an approval policy of >\$1,000 requires secondary approval. You may want to create a stratum of \$900-\$999 as there may be a fraud risk that expenditures are passing under the radar of approval and carry a heightened risk.
3. Testing by Office or by State
 - a. If a company has 3 separate offices, all segregated by a cost code, does it make sense to sample globally or stratify by office? You could justify this either way; however, stratifying on an office level would help you to more closely identify if an issue is isolated to an office rather than an organization as a whole.
4. Revenue/Receivable Stratification
 - a. Consider stratifying by the average population value. If you notice the average # amount in the population is low, you may realize the population is mostly smaller items. Those smaller items tend to carry different characteristics than the large transactions or medium-sized transactions in which case you could address this risk through stratification.

Utilizing specialized software:

Select 60 items utilizing a stratified approach. Assume the following stratum:

- Negative items
- Items from \$0-\$250
- Items from \$250-Program ISI (\$50,000)
- Items greater than Program ISI

Stratified sampling using technology:

Stratified Sample ×

STRATA	----- RECORD -----		AMOUNT	ENTER SAMPLE SIZE
	COUNT	COUNT %		
< 0.00	125887	30.79%	-467,547,697.47	<input type="text" value="18"/>
0.00 - 250.00	196408	48.04%	14,018,308.57	<input type="text" value="29"/>
250.01 - 50,000.00	85129	20.82%	224,675,285.33	<input type="text" value="13"/>
> 50,000.00	1451	0.35%	228,854,103.58	<input type="text" value="1451"/>
<<< Not In Use >>>				<input type="text"/>
<<< Not In Use >>>				<input type="text"/>
<<< Not In Use >>>				<input type="text"/>
<<< Not In Use >>>				<input type="text"/>
<<< Not In Use >>>				<input type="text"/>
<hr/>				
TOTALS	408,875	100.00%	0.01	
<hr/>				
Enter a name for the sample output file	<input type="text"/>			
				<input type="button" value="OK"/>
				<input type="button" value="Cancel"/>

The benefit of the data analytics software is that the user is provided with a closer view at the data in a different way. Once you set your strata, you will be able to see how many (count) transactions are in each bucket as well as the total dollar amount of each bucket. The computer system can automatically calculate the defined strata by both % and dollar amount. In the above example, we selected everything over scope (1,451), and sampled 60 among the remaining 3 strata based on relative number items in each stratum.

ACL and IDEA both have stratification functions whereby you can set the strata and then see how the population count and dollar amounts look. By running/re-running using different strata inputs, you may identify data you normally would not identify using a manual non-CAAT approach.

Monetary Unit Sampling

Monetary unit sampling (MUS), also known as probability proportional to size or dollar unit sampling, is a sampling methodology whereby each dollar in an item is considered a separate sampling unit. MUS is a method of statistical sampling used to assess the amount of monetary

misstatement that may exist in an account balance.⁴⁰ Monetary unit sampling requires the following inputs:⁴¹

- Population Dollar Value – Total dollar amount included within the population.
- Confidence Level – How dependent are you on this sample? The higher the confidence level, the more reliance you are placing on the sample. In general, high confidence = large sample size.
- Tolerable Error – Precision of the sample; the higher the tolerable error utilized, the less precise the sampling estimate. In practice, auditors will consider setting this at less than the materiality level (assume you use 2% materiality, consider setting to 1.5% or 75% of materiality). The higher the tolerable error, the lower the sampling size. This is the highest amount the auditor can live with before the financial statements are considered materially misstated.
- Expected Error – The error amount that you expect to see (recorded amount compared to actual amount). In general, increasing the expected error increases the sample size.

Input	Increase input: Effect on sample size	Decrease input: effect on sample size
Confidence Level	Increases	Decreases
Tolerable Error	Decreases	Increases
Expected Error	Increases	Decreases

Technology greatly assists the auditor in performing a monetary unit sample because the tools will calculate the sample size. Additionally, the auditor should consider extracting individually significant items from the population and not include within the sample as high dollar value items are typically outliers and not subject to statistical projection of error.

⁴⁰ <https://www.cpajournal.com/2017/10/20/greatest-hits-monetary-unit-sampling-using-microsoft-excel/>

⁴¹ <http://www.audimation.com/Resources/Video-Gallery/emodule/1675/eitem/44>

Example of monetary unit sample planning in IDEA:

Monetary Unit Sampling - Plan

Total Value of Sampled Population

Use values from database field: CREDIT

Positive values Negative values Absolute values

Value of the sampled population: 30,646,318.82

Settings

Confidence level (%): 95.00

Amount Percentage

Tolerable error: 5.00 %

Expected error: 1.00 %

Change basic precision pricing (BPP) from 100% : 100

Approximate sample size: 90

Sampling interval: 340,514.65

Sum of tolerable sample taintings: 90.00 %

You may accept the population at the 95.00% confidence level when no more than 0.900000 total taintings are observed in a sample of size 90.

This is the minimum sample size that will allow you to draw the above conclusion.

Buttons: Estimate, Accept, Print, Cancel, Help

Monetary Unit Sampling - Extract

Extraction type

Fixed interval Cell selection

High value handling

High values in sample as aggregate High values in database

High value file name: High Values

Numeric field to sample: CREDIT

Sample interval: 340,514.65

Random starting point: 245,826.84

Change high value amount: 340,514.65

There are 6132 items with a value of 0. Items with a value of 0 will have a 0% chance of being selected.

	Total	Records
<input checked="" type="radio"/> Positive values	30,646,318.82	5,600
<input type="radio"/> Negative values	0.00	0
<input type="radio"/> Absolute values	30,646,318.82	5,600

File name: Monetary Sample

Create a virtual database

Buttons: OK, Fields, Cancel, Help

JOURNAL ENTRY TESTING

The journal entry testing process is utilized for multiple purposes including analyzing a large amount of data, extracting out specific items based on a set of parameters, and allowing the user to focus on a specific risk or account or other criteria. The journal entry, or transactional level, test becomes beneficial as the data sets you are working with become larger. Smaller data sets do not always have the same benefit as the data in certain cases can be analyzed and tested with a manual scan in a quicker, more effective manner. In order to effectively perform journal entry testing in data analytics software, you will need to identify what data is available and with the data, which fields are available.

The most important aspect of performing an advanced transactional level test is having a goal in mind and clearly understanding the inputs involved. Oftentimes we make the mistake of placing too much reliance upon our technology without fully understanding what we are trying to achieve. The user should not say “I want to use ACL software and do a journal entry test”; instead a more effective point would be “I want to use ACL software to identify all journal entries > ISI, all entries posted by a user other than Sally, and all journal entries that fall outside normal monthly closing entries”. By placing focus on your inputs, you will achieve a much better, and more focused, output.

Many of these testing functions can be performed in IDEA, ACL, Teammate Analytics and even EXCEL. The purpose of a transactional level journal entry test is typically to address the risk management override or fraud. Additionally, these searches are to extract and isolate transactions for testing to document the auditor’s evaluation of significant journal entries. During an audit planning meeting, the audit team should talk through significant risk areas and identify how technology can help.

The most common of queries or “tests” performed in the journal entry test using technology are as follows:

Test	
999 amounts	<ul style="list-style-type: none"> ■ A search that identifies all transactions whereby the last 3 digits before the decimal place are 999. ■ Utilized to detect entries which are just below tolerable approval limits and transactions below certain dollar thresholds.
Large amounts	<ul style="list-style-type: none"> ■ Extraction of all transactions or individual transaction postings > a defined input. ■ Valuable test to identify significant transactions and entries and uncover non-standard entries, which typically carry larger dollar values.
Out of balance entries	<ul style="list-style-type: none"> ■ Utilized to identify and isolate journal entries which do not balance across journals or among the entry ID.
Rounded amounts	<ul style="list-style-type: none"> ■ Utilized to identify transactions ending in a round even number (‘000). ■ Significant items such as purchase agreements, large transfers, and other non-standard transactions are typically identified.
Specific dates	<ul style="list-style-type: none"> ■ Search for transactions posted on typical days off such as Federal or company holidays.

Test	
Weekends	<ul style="list-style-type: none"> Search for transactions posted on weekend dates. This is useful for organizations whereby the employees do not work on weekends. Transactions posted on non-working days may be indicative of fraud.
Keyword search	<ul style="list-style-type: none"> Extraction of all transactions posted with an identified word included within the description or memo field. Typical searches are for words indicative of manual override such as: “error”, “quota”, “reclass”, “override”, “mistake”, “correction”, “estimate”.
Unusual times	<ul style="list-style-type: none"> Extraction of the transactions posted at time periods defined by the user. Typically, this would include a search for times outside of normal working hours.
Unusual postings	<ul style="list-style-type: none"> Focused search to identified transactions to accounts that are out of place. Common examples include transactions which are debits to liabilities and credits to income, transactions posted to revenue whereby the other side of the transaction is other than receivables or deferred income, and re-classifications among profit and loss accounts to distort classification (manual adjustments to EBITDA accounts; inflating revenue/expenses to increase top line, etc.).

JOIN OR MATCH FUNCTIONS

Joining databases is an important function when working across technology platforms. The join function works when there is a common field among two databases and layers an additional column or columns onto a database. The two databases are matched by using a common field. That common field will need to be a character (text) and should be unique to a particular item. This field used for the join can be called the “primary key” or “unique key”.

There are several types of joins that can be performed; however, the most common are as follows:

One-to-One

A one-to-one join assumes there are no duplicate records in a database. If you are combining two transactional level databases (for example, a general ledger and a sales journal), you would not perform a one-to-one match. An example of a one-to-one match could be combining an employee master file by employee number with the logical access assigned to those individuals (each will not have duplicates).

Example of one-to-one:



Many-to-One

A many-to-one join can be performed when combining a transaction register with a master file. Many-to-one joins exist when bringing a field from one flat database over to a master database. This is the most common type of join function in auditing.

EXAMPLE 1

You are auditing a not-for-profit organization that receives federal funding and are performing testing procedures over allowable cost. The general ledger has expense categories and cost centers with a logical coding structure as follows:

- 50000 – payroll and benefits
- 52000 – consulting and professional fees
- 53000 – occupancy
- 54000 – other direct expenses
- 55000 – indirect costs

In this example, there are multiple payroll accounts, including 50001, 50002, 50003, 50004, etc. You would like to perform your sampling based on budget category, whereby the budget has 5 main categories: 1) payroll and benefits, 2) consultants and professional fees, 3) occupancy, 4) other direct costs, 5) indirect costs. A many-to-one join can be performed from the transaction register to the budget categories to create a new field (column) within the database to include these categories. A stratified sample can then be performed which then utilizes the budget category as the strata.

Example of many-to-one join:

account	date	name	amount	Prefix	Category
50001	1/1/2017	sam smith	1.00	50	Payroll and benefits
50001	1/1/2017	pete smith	38.00	52	consulting and professional fees
50002	1/1/2017	sam smith	75.00	53	occupancy
50002	1/1/2017	pete smith	112.00	54	other direct expenses
52000	1/1/2017	plumbers llc	334.00		
52001	1/1/2017	electrician llc	371.00		
52001	1/1/2017	tony's pizza	408.00		
52001	1/1/2017	vinny's pizza	445.00		
53001	1/1/2017	bounce castle	741.00		
53001	1/1/2017	bounce castle	741.00		
53001	1/1/2017	bounce castle	741.00		
54001	1/1/2017	tony's pizza	408.00		
54001	1/1/2017	vinny's pizza	445.00		
55002	1/1/2017	tony's pizza	408.00		
55002	1/1/2017	vinny's pizza	445.00		

Many-to-Many

This is a join procedure whereby we are taking two discrete databases and joining them together. These databases contain multiple instances of each primary key. This match is useful if you want to identify a match as the outlier. For example, let's assume we want to compare the employer address field in the employee file to the vendor address in the vendor master file. We don't expect to find a match, however, a many-to-many join will help us to find a match.

In certain instances, you may consider summarizing data prior to performing a join. Many-to-many joins can be very messy and hard to analyze, therefore, summarizing data prior to joining could provide for a better matchup or a one-to-one matchup as opposed to a many-to-many matchup.

Example of many-to-many join:

CHECK_NUMBER	VENDOR_ID	DATE	PAYMENT	Name	Address
11817	John Smith	1/3/2017	22020.00	John Smith	123 Easy Street
11818	John Smith	1/7/2017	311.28	John Smith	1234 East Street
11819	John Smith	1/7/2017	61.20	Harry Jones	13 Snowman Way
11820	John Smith	1/7/2017	201.66	Harry Jones	14 Snowman Way
11821	John Smith	1/7/2017	854.51	Harry Jones	13 Snowman Way
11822	John Smith	1/7/2017	114.70	Executive Director	35 pkwy 18
11823	John Smith	1/7/2017	124.56	Executive Director	35 pkwy 18
11824	John Smith	1/7/2017	641.27	Executive Director	96 Tpk dr
11825	John Smith	1/7/2017	306.54		
11826	John Smith	1/7/2017	476.80		
11827	John Smith	1/7/2017	801.21		
11828	Harry Jones	1/7/2017	346.84		
11829	Harry Jones	1/7/2017	331.94		
11830	Harry Jones	1/7/2017	158.45		
11831	Executive Director	1/7/2017	91.80		
11832	Executive Director	1/7/2017	345.20		
11833	Executive Director	1/7/2017	87.00		
11834	Executive Director	1/7/2017	800.99		

Example Joins in Practice:

When performing an ERISA audit of a defined benefit plan, we commonly are looking for employees that were either added to the participant valuation report or subtracted from the participant report to test the newly eligible or terminated employees from the plan for compliance. A very simple procedure that can be done is combining the prior year participant valuation report with the current year participant valuation report to identify matches (people that were in the plan last year and in the plan this year) versus non-matches (which can be people in prior year not in current year OR people in current year not in prior year). This allows the auditor to very easily test existence of the data because we can hone our sample population and ignore the matches.

Looking another way, let's assume our relevant assertion is completeness. Our matching procedures would most likely be from different databases in the example above. One common procedure is to take a participant valuation report and perform a join with the payroll register. The payroll register commonly has important details such as "date hired" and "date terminated". This will allow the

auditor to identify employees that became newly eligible or identify employees that were terminated to identify if they were marked consistently across platforms.

Common Joins in Practice:

- Comparison of employee master file to vendor master file for address matches
 - matching key: address
- Comparison of participant valuation report contributions to payroll file contributions
 - matching key: employee ID
- Identification of related party transactions by combining one of the following: the cash disbursement journal, the sales journal, expense ledger, general ledger with a listing of known related parties
 - matching key: name
- Matching summarized bank data by date with summarized cash transactions in the general ledger by date
 - matching key: date

APPEND FUNCTION

In certain cases, files are exported into multiple sheets in one workbook (i.e., a general ledger exports to six different tabs in one EXCEL file). The append function will come in handy if this occurs and allows for multiple files to be merged together in order to create one master database. In practice, this is seen on files with weekly data on a different tab. The goal is to collate this data so you have 1 database, not several workbooks with discrete data.

Consider the following:

Client provides you a payment register by week on a separate EXCEL tab due to size of data. Appending the database allows you to take all those tabs and create 1 continuous database to be able to perform better data analytics and other procedures.

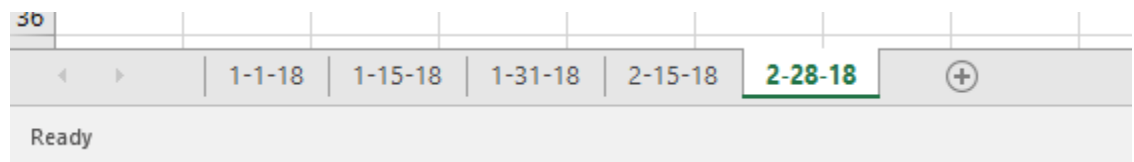


Image: EXCEL workbook with weekly tabs

BENFORD’S LAW

Benford’s Law predicts the expected digit frequencies in a list of numbers. In the book “Digital Analysis Using Benford’s Law” by Mark J. Nigrini, PhD., the concept of Benford’s Law was developed by physicist Frank Benford who made the observation while examining the logarithm books. These books were used years ago to multiply large numbers before computers or calculators were utilized. The observation he made was that the pages of the logarithm book beginning with 1 and 2 were more worn than the pages with 8 and 9. The theory of Benford’s Law was that the numbers 1, 2 and 3 were dealt with much more often, however, with each succeeding digit, the amount of time 1, 2, and 3 were used decreased.

Benford formulated that the expected frequencies of the various digits in lists of numbers using a mathematical assumption are based on geometric sequences. These expected frequencies, known as Benford’s Law, always give a higher probability to the lower digits (such as ‘0’, ‘1’, and ‘2’), but from about the fourth position onwards, the probabilities are much lower and closer together. The research done by Benford was published in a paper titled “The Law of Anomalous Numbers.”⁴² A Benford’s Law test that is run on the first digit is referring to the left most digit (in the number 1,234, the “1” is the first digit). The second digit in a Benford’s Law test refers to the second digit from the left (in the number 1,234, the “2” is the second digit).

The following table shows the expected digital frequencies and the probability based on the digit:

Digit	1 st	2 nd	3 rd	4 th
0	N/A	.11968	.10178	.10018
2	.30103	.11389	.10138	.10014
3	.17609	.10882	.10097	.10010
4	.12494	.10433	.10057	.10006
5	.09691	.10031	.10018	.10002
6	.07918	.09668	.09979	.09998
7	.06695	.09337	.09940	.09994
8	.05799	.09035	.09902	.09990
9	.05115	.08757	.09864	.09986

According to the Association of Certified Fraud Examiners (ACFE) 2015 fraud examiners manual, one of the goals of a Benford’s Law test is to identify fabricated numbers. An important point of note on Benford’s Law is the concept of “natural” versus “non-natural” numbers. Natural numbers are those that are not ordered in a particular numbering scheme and are not human-generated or generated from a random numbering system. Non-natural numbers are designed systematically to present information that restricts the natural nature of number (for example, employee ID numbers,

⁴² Digital Analysis Using Benford’s Law, by Mark J. Nigrini, © 2000

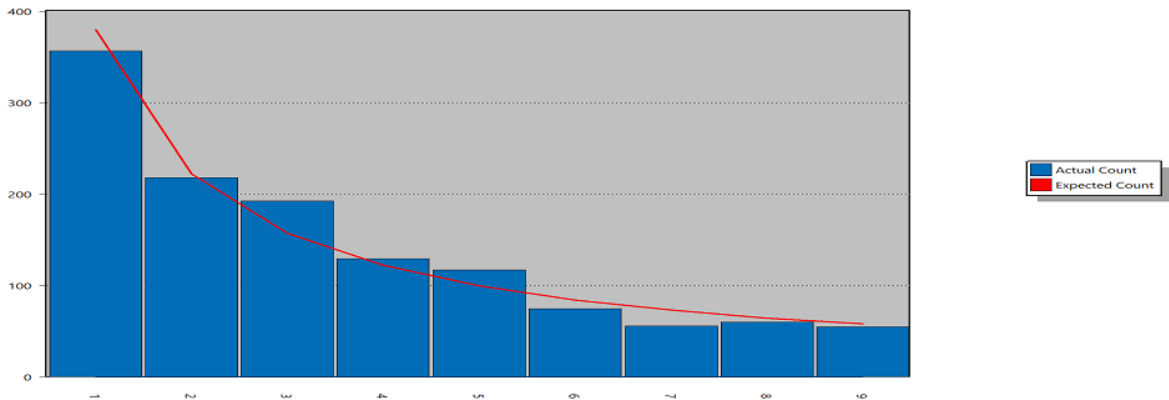
or zip codes). Testing data sets for the occurrence or non-occurrence of the predictable digit distribution can help identify included numbers that are not legitimate.

In the book, *Digital Analysis Using Benford's Law*, as noted above, the following criteria must be met for a data set to conform to Benford's Law:

1. The data set should describe the sizes of similar phenomena. As an example, the revenues of companies on the NYSE.
2. There should be no built-in minimum or maximum values in the data set. As an example, if rent had a minimum floor of \$4,000, a data set of rent would have an excess of first-digit 4's. This would create a skew of the data and an expected pattern whereby there would be no transactions with first digit 1, 2, 3.
3. The data set should not be made up of assigned numbers. Assigned numbers are numbers given to things in place of words such as social security numbers, bank account numbers, telephone numbers, etc.
4. The data set should have more small items than big items.

Additionally, the numbers in data sets should have four or more digits for a good fit for a Benford's Law test, however, not required.

Example Benford's Law graphical display looking at the first digit in a check register (sample data):



SPECIFIC AUDIT AREAS

Accounts Payable

Common file requests – Accounts payable detail file at year end; check register; general ledger; vendor master file.

With accounts payable, you can utilize technology to focus on the following items:

- Fictitious vendors or conflict of interest – summarization or aggregation of data over a multiple year period could identify this. Specifically, if you summarize by vendor, you can now identify when a vendor came aboard via their first invoice which could then trigger you to isolate your testing to new vendors.
- Fictitious, inflated and/or duplicate invoices – performing a duplicate test of vendor and amount could identify an issue. Looking specifically at the count over a period of time (for example, assume you look at year 1, year 2 and year 3. Year 1 you have 12 invoices, year 2 you have 12 invoices and year 3 you have 15 invoices). By identifying that year 3 has an anomalous number of invoices, you quickly identified that you need to investigate year 3 as there may be duplicates (not by number but by service) or errors (invoice entered twice).
- Write off payables – summarization of activity by vendor in accounts payable by type (invoice, journal entry, cash payment, currency adjustment) would identify for you whether there are any adjustments to invoices that require investigation. A client may decide to write off an old liability, which would be identified through this simple analytic.
- Compare accounts payable aging data – take an accounts payable aging schedule for year 1, year 2 and year 3. Combine those schedules by vendor and identify issues such as non-payment of vendor bills (could be fictitious transactions), inconsistency or issues of missing payables, and patterns which can be used to explain overall fluctuations.
- Subsequent payments – combine the listing of subsequent accounts payable payments to the listing of accounts payable at year end to identify what was paid and what was not. Sometimes non-payment could imply a contingency or estimate in accounts payable requiring further risk and procedures to address.

Accounts Receivable

With accounts receivable, you can utilize technology to focus on various items.

Common file requests – Accounts receivable detail file, cash receipt journal, sales journal, general ledger, customer master file.

The following can be done:

- Join the sales journal and accounts receivable journal by customer to identify any significant concentrations of A/R to sales to identify instances of AR collectability. Are there any clients where there are 100% AR compared to sales that may warrant confirmation or further testing?
- Roll forward – summarization of activity by customer in accounts receivable by type (invoice, journal entry, cash payment, currency adjustment) would identify for you whether there are any adjustments to invoices that require investigation. A client may decide to write off an invoice to an account other than bad debts to hide it, which would be identified through this simple analytic.

- Join the accounts receivable journal at year end to a subsequent payments listing to match payments up to the specific invoices in which they relate. This will allow you to identify which invoices were collected after year end in the system to better focus your AR confirmation testing.
- Sampling – stratify the accounts receivable invoice population for confirmation testing.
- Age the accounts receivable listing by invoice date to verify that the aging schedule provided was not manipulated.
- Combine 3 years of aging history to identify instances of slow moving customers or payments and to help perform a retrospective review of the allowance for doubtful accounts.

Payroll

Significant data typically exists around payroll.

Common file requests – Payroll processor report by pay period; timekeeping data; employee master file, general ledger.

This data can be used for many purposes including but not limited to:

- Importing – script the import of payroll processor master file data.
- Duplicate or unauthorized payments – identify instances where employees are getting paid twice for the same service. Obtain payroll processor data and summarize by employee for the year and compare to payroll authorizations. The payroll authorizations can be joined against the summarized data and compared to identify instances where employees are being paid more than authorized amounts.
- Individuals set up as vendors and W-2 employees – join the payroll master file address list with the vendor master file address list and identify any instances with a duplicate address. The same can be done with the listing of employee names and vendor names to identify matches. Performing a “fuzzy” match can help alleviate common issues of names being close but not the same (Bill Johnson vs. William Johnson vs. Billy Johnson).
- Unapproved overtime hours – summarize the hours by week by employee and create a database for comparison (such as approved hours by person by week). In certain cases, companies may have manual sheets of approved time or may indicate that no overtime is allowed – each case would be analyzed and understood to determine the appropriate course of action. By looking at # of hours worked by pay period, you can easily identify whether payroll hours are being logged in excess of approved amounts.
- Unauthorized salary increase – assume a client has a policy whereby salary increases are granted annually. Obtain the payroll processor data, summarize by pay period and perform an extraction whereby each employee has more than XX number of changes in their compensation.

Travel and Entertainment Expense Reimbursements

Travel and entertainment expenses are an often abused area and an area of high risk for large organizations with many employees. There often are multiple systems involved and many individuals with approval authority.

Common file requests – General ledger; credit card detail reports; travel and entertainment tracking system data; check register.

The following can be done:

- Split purchases – run an analytic function to summarize transactions (same employee, expense type, date and amount). Perform a duplicate test, then test to see if those items extracted are below an authorized limit which may be indicative of sliding expenses through under an approval limit.
- Duplicate submissions for the same bill – this occurs when an employee submits an invoice for approval and then also includes that within their expense report. This may involve combining multiple reports depending on how the company tracks these types of expenses. For example, travel reimbursements may go through an expense system such as Tallie or Concur. But invoices for vendor payments may go through the general ledger AP module directly without being entered into Tallie or Concur. Or even more difficult, a corporate card may be used for purchases OR an employee submits through a central processing arm and the information runs through a secondary system. All of these possibilities create an environment where duplicate submissions can occur. This can be done by performing an extraction of all payments out of AP to employees and summarizing by employee by invoice. Then, perform the same extraction and summarization out of the expense reimbursement system (or credit card transaction report, whichever is used by client). Compare the data you have and identify duplicates.
- Abuse of spending – this occurs when the company does not have a handle over their expense reimbursements and do a poor job of tracking travel and expense account spending. Consider summarizing the data by person by month or by department by month and analyze visually to see if there are any unusual spikes or comparisons that should be investigated.
- Unauthorized expenses or expenses approved above authorization limits – combine the master file of expense authorization with the expense reporting system. Join the limits associated with the specific people that are limited and perform an extraction of all items over limits identified to determine. Consider summarized by person by date before doing this join to capture individuals that split their reimbursement up to avoid approval being required.

Fixed Assets

Fixed asset databases can be small or large depending on the type of company. Manufacturing companies tend to have significant improvements/purchases whereas a community based not-for-profit may only have a few assets. Reconciling databases is important when considering completeness and fraud risks.

Common file requests – General ledger; physical inventory count with tag ID; fixed asset register with depreciation calculations; shipping/receiving logs.

- Join the physical inventory count with the fixed asset register by Tag ID. IF both databases do not have a common factor, consider advising or creating one so that future reconciliations can be done with the click of a button.
- Analyze the receiving log close to cutoff with the physical count or date placed in service on the fixed asset register.
- Compare additions and deletions in the general ledger to the fixed asset rollforward or directly to the fixed asset register. Matching keys and coding the data are very important to execute these tasks with technology.
- Identification of purchases/assets received not included in fixed assets (potential theft).

Check Register and Cash Disbursements

In general, misappropriation of assets at an entity will generally involve cash whether on the cash receipts side or the cash disbursements side. Therefore, various fraud tests can be run to isolate certain items from a check register to identify irregularities, or to isolate specific checks.

Area	Considerations
Gap Detection	Gap detection on check number; investigate significant gaps in checks
Significant Vendors	Summarize by vendor – document understanding of significant vendors and select on test basis
Significant Checks	Extract checks > a certain dollar value or approval limit
Unusual Vendors	Last names of employees; vendors with 1 check or minimal checks, round even numbers (shell company schemes, pass-through entity issues)
Related Party Consideration	Obtain listing of board members; obtain listing of board member affiliations if applicable (company they work for; ownership interests; may be needed for 990 disclosure)
Completeness	Compare CD journal to disbursements indicated in general ledger – are you missing a portion of the population (wire transfers?)
Compare	Obtain a check register over a multiple year period and join that data together. Once joined, you can compare vendors' year over year to identify any significant or unusual trends. Consider filtering the data to identify instances whereby vendors started, vendors dropped off, or vendors have significant fluctuations in activity

Revenue Considerations

Revenue is often the most significant risk area. The following are procedures that can be performed and questions asked to identify data analytics that can be performed.

Area	Considerations
Internal Controls	What are the organization's internal procedures surrounding revenue? Do they use separate software? Does the sales staff talk to the accounting staff? The controls will significantly impact the data available and the testing that can be performed.
Completeness	If the systems permit – match the revenue from the billing system to the general ledger revenue (reconcile differences such as basis of accounting, cash vs. non-cash, items not recorded, certain bequests with no valuation). Does the accounting staff corroborate the data / contracts provided by the sales staff?
Stratification	Stratify the population in a variety of ways: <ul style="list-style-type: none"> ■ Extract all amounts over a threshold (individually significant) ■ Segregate by a number (for example, average dollar value of transactions) or by a character (for example, sales by state) ■ Segregate outliers such as related party customer transactions from the population using key word extractions
Summarization	Summarize the transaction register in a variety of ways: <ul style="list-style-type: none"> ■ Sales by month ■ Sales by product line ■ Sales by customer
Join	<ul style="list-style-type: none"> ■ Join the sales journal and the purchases journal to do comparisons of margin by product line within a period of time. ■ Join the A/R journal and the sales journal and identify instances whereby current sales are a significant portion of AR (indicative of slow collection).

DATA VISUALIZATION

Data visualization is not a new concept. We have been seeing bar chart presentations for over 20 years. However, as with any art form, the presentation used can make or break the point you are trying to make.

Data, at its core, is black and white. When a layperson thinks of data, they envision the image of the matrix. Although green and black, it's the binary image of 1's and 0's running down a page and the average person cannot read it.



If you cannot tell a story effectively with data, the context or point you are trying to make may be lost with the user/reader. In the book, “Storytelling with Data”,⁴³ there are six key lessons to effectively show data visually:

1. Understand the context and know your audience
2. Choose the right display/tools
3. Eliminate clutter and other unnecessary detail
4. Focus attention where you want it
5. Think like an artist or designer
6. Tell a story

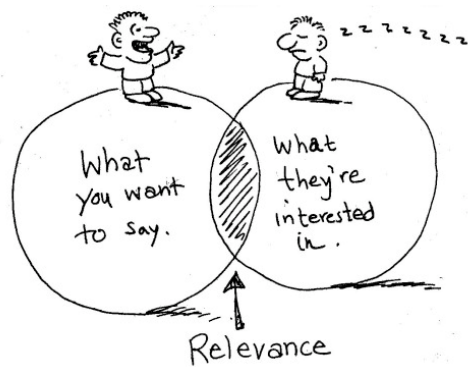
Let’s look at a few of these.

Understanding the Context

Before you decide what the presentation should look like, you first should understand who the audience is and what you are trying to portray. The result should be created with the audience in mind. Is the presentation supposed to be exploratory or explanatory? It is important to draw a distinction between presenting data to find a solution OR presenting the solution. Explanatory data is typically where data visualization comes into play because you are trying to convince the audience that the content is real.

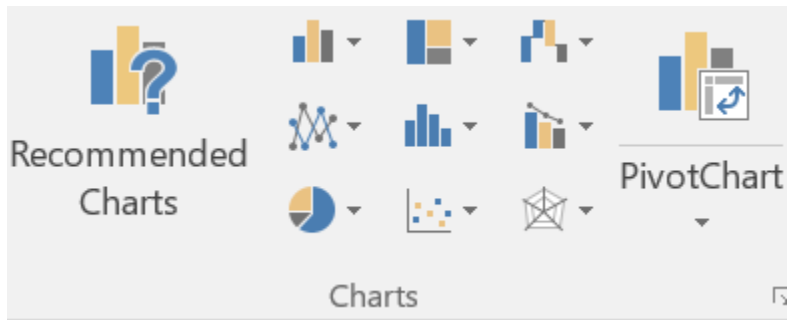
Knowing the audience is an important aspect of data visualization. In certain cases, this is a complex process to get to know the wants and needs of the client. In other cases, this may be direct and a matter of picking the right color. Each project will be different and each client will have a different want or need. Once you know that, you can then move towards how the audience will perceive the materials and focus on an audience takeaway. Each presentation carries different takeaways and the highlight of the presentation should be the key takeaway. What are you trying to portray?

⁴³ Story Telling with Data, by Cole Nussbaumer Knaflic, © 2015



Types of Displays

There is a myriad of different types of visual displays that can be used to present data to an audience. In this manual alone, we've utilized simple text, tables, vertical bar graphs, etc. For the everyday user, Microsoft products contain some of the basic graphs that are used in many presentations. Depending on the presentation and the data, one type may be more appropriate than another.



These examples include:

- Bar charts
- Tree maps
- Sunbursts
- Waterfalls
- Funnels
- Line graphs
- Histograms

- Scatter charts
- Pie charts
- Heat maps

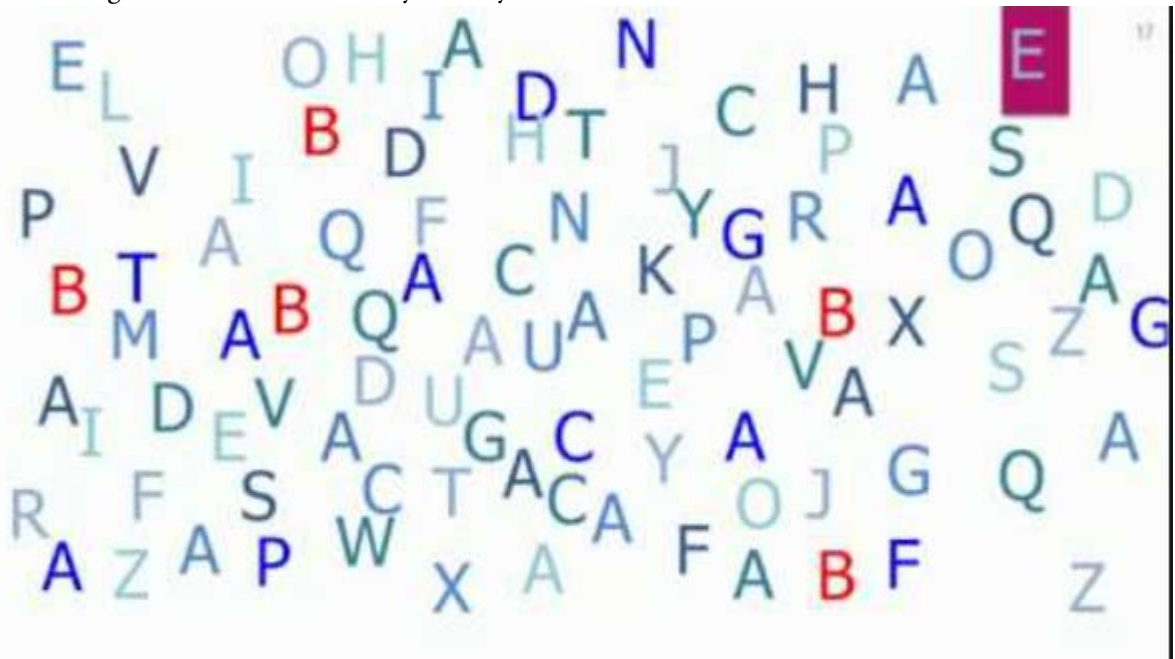
Eliminate Clutter

The brain can only process so much at one time. Certain items take precedence over others when presented on a page or a screen. Where will you get the most value? Attention spans are short and everyone is tired of “death by PowerPoint”. Have you ever sat through a presentation where all you see is words on a screen in small type set font? How about a scenario where the screen was packed with more than 10 different points? Both are set to fail prior to even delivering the pitch. Clutter in a presentation can occur when there are images or pieces of information on a page that don’t add value to the presentation.

Focus Attention Where You Want It

When you review the following picture, how many B’s do you see?

Now look again and tell me how many S’s do you see?



Using visuals is a powerful way to focus your attention on important facts while leaving other facts or less important facts (that aid in pointing out a fact) in the background. The items to call out vs. the items to display and not call out is a matter of judgment. In the above image, one may show this picture to prove the point that the brain can process items that stand out with color much faster and more succinctly. There are several ways to emphasize points pictorially or via simple text. Some of the more common methods include using: bold, italics, colors, sizing, underlining, segregating information on a page, and using bullet points.

Tell a Story

Ultimately, every presentation is done to tell a story. In the book, *Talk Like Ted*, the author talks about the best way to tell a story and that each story must have 3 elements. The story must be emotional, novel and memorable. In order to do that, the presenter must find a way to captivate the senses by making a story emotional and touch a heart string (or another emotion). The presenter must be novel and teach the audience something through the delivery. Lastly, the presenter must be memorable by providing details and facts in a way that you remember them when you leave. The same holds true with a pictorial presentation. We have all sat through a presentation where your jaw drops at a fact or a statistic or a revelation the presenter had. Those are the details we want to capture and want to present in a way that the audience can remember.

Have you ever walked into someone's house and been speechless at the décor? What causes that speechlessness? Is it the cost of the items included within the room? Is it the location or feng shui? All of these variables play into a visual.

I LOVE DATA



I LOVE DATA

I LOVE DATA

I LOVE DATA

Which picture above presents the fact that “I love data”? The illustration shown above may be misleading without the theatrics or the words behind it because the audience may not know what that picture represents.

SUMMARY

In this unit, we reviewed important concepts to understand how to use data analytics in different sampling techniques, to test journal entries and other key areas important to auditors. We also reviewed how data analytics can be used to aid in organizing data for review and analysis.

NOTES

Case Studies

CASE STUDY #1 - NEWARK WATERSHED

Fraud

In an investigative report issued by the State of New Jersey, Office of the State Comptroller in 2014 on the “Newark Watershed Conservation and Development Corporation”, it was indicated that significant frauds were identified, however, were not caught by the auditor.⁴⁴ This leads to believe that general audit procedures and manual procedures do not suffice in the 21st century. As auditors, we need to take our work to the next level and leverage the technology we have to bring a benefit to society and protect the public (while increasing efficiency and effectiveness).

Let’s look at what went wrong here and discuss how it could have been avoided utilizing a data analytics approach.

The Office of the State Comptroller (OSC), among other things, found:

1. The executive director continually engaged in wasteful and abusive spending practices with city monies in the following ways:
 - a. Wrote unauthorized payroll checks
 - b. Maintained an actively trading brokerage account making risky investments
 - c. Handed out contracts to related parties with no formal bidding procedures
 - d. Gave no-interest loans to an NFP which was operated by a trustee of the entity issuing the loan
 - e. Gave improper bonuses and advances
 - f. Distributed hundreds of thousands of public dollars to community organizations without approval
 - g. Acted recklessly in use of petty cash
2. The board approved a payout to the executive director and failed to institute appropriate contracting and spending policies and performed minimal to no sub-recipient monitoring.

Let’s start at the top and work our way down. Since not-for-profit spending is always under scrutiny both from a volume standpoint and from an “efficiency of the spend” standpoint, auditors have a

⁴⁴ State of New Jersey, Office of the State Comptroller, Investigative Report, Newark Watershed Conservation and Development Corporation, February 19, 2014

responsibility to understand the spending policies and practices of the not-for-profit and perform testing to be comfortable the entity is in compliance with donor wishes, funding source requirements and other interested parties.

Writing unauthorized payroll checks

This is an easy one to catch, but requires some understanding. First and foremost, most companies utilize a payroll processor. Furthermore, most payments are made via direct deposit from that payroll processor. Therefore, it is in plain sight to see the payroll checks made via the payroll processor as these companies can generate summary reports. So how do you catch a check that doesn't go through the normal processing channels? Prepare a payroll reconciliation from the general ledger to the payroll tax returns. If they check out, you know there are no "unusual" or "misplaced" postings in the payroll expense accounts in the general ledger. But your job is not done.

Obviously, data analytics cannot catch a fraud, but they can be used to identify anomalies and unusual issues. In this case, what one could do is obtain the payroll processor's report over a 2-3 year period, **JOIN** the databases together by a *matching key* such as the social security number or in the case of a small organization, the name. This will point out consistency in pay or significant increases, new payees, or payees you have never seen before (maybe an employee that doesn't exist). Upon identification, you will need to pull the contracts, salary approvals and have the appropriate discussions with the right people. But, that would not be as clear if you did not analyze the data over a 3-year period which the data analytics software made possible by joining the files.

Brokerage account with risky trades

This is caught by obtaining the brokerage account statements in electronic format. First step would be to identify what the company is invested in. This can be done by reviewing the year-end statement. If nothing stands out, but you know of a risk, obtain a PDF or EXCEL copy of all transactions consummated from the investment house. Margin trading will show up when there is a short number of days between each buy and sell, therefore, you can first run a differential between date purchased and date sold and **EXTRACT** all transactions with sales trades within 30 days of purchase.

Secondly, you can perform a **SUMMARIZATION** on the company or investment being traded, run a count of the number of times the investment is traded, and sort either by frequency, dollar amount traded, or dollar amount lost. These simple functions will help you identify if the company is wisely investing or executing unusual trades.

Handing out contracts to related parties with no formal bidding procedures

Gave improper bonuses and advances

Distributed hundreds of thousands of public dollars to community organizations without approval

This is a prevalent issue as most NFPs engage in some type of contract for work to be done and write out checks/wires for most if not all the transactions they consummate. This issue is twofold: 1)

engaging in unauthorized contracts, and 2) related party contracts. To execute this task, you will need both the AP listing and the check register (both because if there is a contract entered but not paid, it will not show up in the check register). First task, you will need to perform a SUMMARIZATION of the check register for the year. Upon summarizing, you can sort by either name or dollar amount. Next, there are a few things to extract using the EXTRACTION function:

1. Round even transactions
2. Transactions with minimal count (for instance 1 check for the year above a certain dollar threshold)
3. Related party names
4. Payments made out to “cash”
5. Payments individually close to or at the approval limit

Extracting 1 & 2 are easy because they involve a simple formula in any data analytics software, but how do we identify related party names? First, you will need to understand who a related party of the entity is. Typically, these include board members, employees, board member’s business ventures and family members of both board members and employees. To execute this task, you can either perform it manually (scan the list), via “search” (search each name individually), or perform a LOOPING WORD SEARCH. This function is a complex equation which will search a database for instances of a word, or multiple words based on another database. Upon finding a match (or using a fuzzy match, meaning a close match), the software will EXTRACT the transactions with those parties. This will automate the process of searching for the related parties such that you have the related party reference file. Upon identification of related parties, transaction support should be requested and tested.

Searching number 5, payments at or around the approval limit, would require an extraction of payments ending in ‘999 and ‘000 which requires the use of an expression in the data analytics software.

Gave no-interest loans to an NFP which was operated by a trustee of the entity issuing the loan

No technology solution here beyond obtaining the general ledger or check register, summarize by “name” and identify the related parties. Upon investigation, you would find out that the transaction is a loan and not an expense. In searching the general ledger, extract specific accounts (loan accounts, exchange accounts, intercompany accounts) in order to isolate and reduce the population being searched.

The board approved a payout to the executive director and failed to institute appropriate contracting and spending policies and performed minimal to no sub-recipient monitoring.

A strong internal audit or other monitoring procedure is required to capture this. This is where an ACL or other specialized software comes into play whereby procedures can be automated. The data specialist can create scripts which can be used to identify common items and to perform “tests” of the data and produce graphical depictions, extract anomalies, and provide the internal auditor with all they need to investigate high risk items identified with efficiency. The time is spent on the front end and can significantly aid in addressing sub-recipient monitoring as well as identify issues with spending.

Conclusion

The issues raised within the investigative report indicated several instances of lack of oversight. Technology works very well to investigate and catch those issues, however, that is only part of the concern. In this type of environment, you also need to know how to address the issues with the parties.

CASE STUDY #2 - DOV-Q, CUSTOMER ATTRITION, AND PREDICTIVE ANALYTICS

Dov-q Industries is a publisher of several regional web-based and app-based newspapers and local news aggregation products in central Ohio, and has been experiencing more “churn” than usual over the past sixteen months. Churn rate is another name for customer attrition or the amount of customers who will stop subscribing to a business service or who will cancel their subscriptions. The actual formula is the number of cancelling customers divided by the number of current customers.

Dov-q believes that some of the churn may be based on COVID-19 but can't be sure, so management has asked your firm to produce a report which can predict which customers will cancel their subscriptions in the next year or two so that management can concentrate on these customers with incentives/discounts/ or other incentives for retention purposes.

The firm has lost nearly a million dollars in advertising revenue because of the churn - ad rates are partially based on total verified subscribers. Again, while COVID-19 may have played a role in the churn, Dov-q needs to identify who is going to cancel so their ad rates and revenue can ideally get back to where they were in 2019 or at least to stabilize and retain their current subscription base. While some revenue is driven by user subscriptions, several of the apps are free for customers to use for basic news items.

You are assigned to the Dov-q engagement. Your firm has decided to use the CRISP-DM process to perform the predictive analysis.

CRISP-DM Model as applied to Dov-q:

STEP ONE: Business Issue Understanding:

- Define Business Objectives
- Gather Information Required
- Determine Appropriate analysis method
- Clarify scope of work
- Identify deliverables

STEP ONE: Business Issue Understanding:

What are Dov-q's business objectives?

How will Dov-q gather required information?

What will be the appropriate analysis method?

What is/are the deliverable/s?

STEP ONE: Business Issue Understanding:

Answer:

Predict the likelihood or probability that a subscriber will cancel a subscription.

Predict a subscriber's value (purchases multiple paid subscriptions/is a social influencer on social media, etc). Dov-q will want to retain high-value subscribers, so how will they incentivize this group? More frequent contact? Proactively contact? Filter subscribers based on this value?

Gather historical subscription data to determine the high-value subscribers. As for analysis, if the historical data is defective in some way, run a Mobile A/B testing analysis to derive results about user interface and user experiences, and content/delivery.

Mobile A/B Testing involves pushing two separate apps (original version and enhanced or updated version) and then tracking use (i.e. minutes spent viewing, amount of interaction, etc.). Results are aggregated and analyzed to support the use of one of these apps.

STEP TWO: Data Understanding:

- Collect initial data
- Identify data requirements
- Determine data availability
- Explore data and characteristics

STEP TWO: Data Understanding:

How will Dov-q collect the initial data?

How will Dov-q identify data requirements?

What data is available?

When obtained, how will the data be explored and what will be its characteristics?

STEP TWO: Business Issue Understanding:

Answer:

Can we use a tool to scrape data from existing sources? If the data is unstructured, how difficult will it be to structure it for our analysis? Is the data in-house? Do we have to get it from a Dov-q vendor? How much will this cost?

First take a look in-house for accounts payable data. Also obtain any data showing customers who recently upgraded from free to a fee-based service – that's the target for retention. Is there anything else out there – possibly data which has a reason that the customer gave for leaving?

If none of this data exists, as noted in Step One, run a Mobile A/B testing analysis to derive results about user interface and user experiences, and content/delivery.

Mobile A/B Testing involves pushing two separate apps (original version and enhanced or updated version) and then tracking use (i.e. minutes spent viewing, amount of interaction, etc.). Results are aggregated and analyzed to support the use of one of these apps.

STEP THREE: Data Preparation:

- Gather data from multiple sources
- Cleanse
- Format
- Blend
- Sample

STEP THREE: Data Preparation:

How will Dov-q collect the initial data?

How will Dov-q identify data requirements?

What data is available?

When obtained, how will the data be explored and what will be its characteristics?

STEP THREE: Data Preparation:

How will Dov-q gather data? Are there multiple sources?

How will Dov-q cleanse the data?

How will Dov-q format the data?

How will Dov-q blend the data?

How will Dov-q sample the data?

STEP THREE: Data Preparation:

Answer:

Assuming that Dov-q can “pull” or obtain some raw data on the topics noted in steps 1 and 2, it will need to cleanse, blend, and prepare it for a predictive modeling algorithm. These steps include removing misleading, incorrect, multiple-copy, or useless data, and merging the useful data into a

single database. If required, sampling would be used to derive the customer data that we want in a subset instead of the entire population, but here it would make more sense to decline sampling.

Preparation encompasses converting the raw historical data, which is usually comprised of tables (including.csv files), text, or images, into a format which the algorithm will understand. This can involve converting the data into IEEE-754 floating point numbers.

STEP FOUR: Perform Exploratory Analysis and Modeling:

- Develop a methodology
- Determine the important variables
- Build the model
- Assess the model

STEP FOUR : Perform Exploratory Analysis and Modeling

Answer:

Run the formatted data in the algorithm - it's best to use test data when doing this.

Select a Model that you want to use – we can use a classification model for Dov-q. The usual case is to use several models and select the one with the best performance.

The algorithm will then “train” the model through a process called Machine Learning (ML) or Deep Learning. Software can do much of this work with a few mouse clicks.

Are we seeing something we expected for model output based on our test inputs (our customer objectives mentioned earlier) from the testing? This is our assessment sub-step.

STEP FIVE: Validation:

Evaluate the results

Review the Process

Determine next steps

Look at the results – are they valid? are they invalid?

STEP FIVE: Validation

Answer:

This step incorporates the input-output comparison based on test data noted in the prior step. The initial test data output will establish a baseline for the prediction metrics (e.g. which customers, especially high-value customers, are ready to cancel their subscriptions over the next few months). If

we are getting invalid results in the output, we will need to return to the previous steps and start again.

Review of the Process/Lessons learned here are based on the quality of the test output. If we have used several models, we can base our final selection on output quality and the model which is generating the least amount of errors.

STEP SIX: Visualization and Presentation:

- Communicate Results
- Make recommendations

STEP SIX: Visualization and Presentation

Answer:

When we have finally found a model which produces valid output that predicts what we were original looking for based on our inputs, we can present our work to Dov-q.

The presentation phase is more art and communication soft skills than science –remember that the Dov-q audience is not entirely composed of data specialists, so consider using the least technical terms if possible, and employ images to show your work (the “visualization” of the work).

Another visualization tip is to use the six step method that we have just gone through, absent the most technical terms, in your presentation.

CASE STUDY #3 – NOT-FOR-PROFIT AUDIT

Example Entity Worldwide, Inc. (Sample Entity) is a not-for-profit organization that receives revenue from multiple sources including the U.S. Federal government, the general public, and corporations/foundations. Total annual revenue is approximately \$28 million with expenses of approximately \$24 million. The client has engaged Independent Public Accountant, *CPA Firm, PC*, to perform the June 30, 2017 audit. Specifically, we will be performing a single audit in accordance with Uniform Guidance.

To start, you interviewed the client to understand how data was captured and stored. It was identified that their general ledger is the main source of record and contains everything the client uses to report to the federal government. The client indicated to you that the general ledger is set up in such a way that each contract/source is identified with a separate cost center code, which is part of the account number string. The client walked the accountant through the materials.

As part of the audit, the client has provided the CPA firm with a listing of the cost center identifier pertaining to each contract/program. The client indicated that it would be impossible for them to print out a manual transaction register by cost center, and they do not keep copies of printouts during the year. As a client advocate, we want to (i) maintain professional skepticism, (ii) perform the role of auditor, and (iii) provide exemplary service to the client. Finally, we must keep billable hours down with the same level of effectiveness or the partner is going to disallow the dinner expenses.

Functions Involved

- Importing
- Direct extractions
- Summarization
- Pivot table
- Sampling
- Join

Let's use an example to display this point and start by the partner saying "use data mining software on the engagement". Your first response is probably "how?" That's a great first response because it is easy to get lost in the data. What are we trying to do? What are the first steps? To make it simple, we are really looking to isolate and identify transactions directly related to specific major programs we are required to test. The issue is that the client has not provided a schedule of expenditures of federal awards yet, therefore, we have no idea what major programs we will be required to audit. We just identified an area of use! We need to have a good picture of the client's expenditures incurred by grant in the planning phase so that we can select the programs we need to test to issue our opinion on compliance.

Most not-for-profits are structured in such a way that the general ledger is organization by account number, sub account number, fund, etc. The sub account or cost center is of specific importance in our example because it represents either a specific contract, specific program or both. The objective now is to extract and summarize from client data (the general ledger is ideal) the revenues and expenditures by sub-account for purposes of identifying significant grants and contracts. Once you obtain the general ledger, you first will need to test the clerical accuracy of the general ledger and verify the data the client gave us is complete.

In the case of a general ledger, it is often beneficial to summarize the transactions by account number, add prior year beginning balance sheet balances from the trial balance, and compare the calculated numbers to ending balances on the year end trial balance to verify the general ledger rolls forward from prior year (complete data set). This procedure will help establish a base line understanding that no line items have been deleted in entirety.

Once you are comfortable with the completeness component, you now need to understand the coding structure with the data provided and how to extract the specific information.

Simple structure:

10000-100

10000 – account number

100 – cost center (grant or contract)

By having a discussion with the client, you were able to identify that cost centers numbered between 1001 and 1050 are all the active grant contracts. Therefore, your goal is to extract out cost center 1001-1050 for the revenue and expense only (client indicated they don't differentiate cost centers on the statement of financial position).

- In order to isolate the account number from the cost center, you can use an @split function in your data analytics software OR you can utilize EXCEL and perform a text to column on the field.
- Once split, the new column generated can be labeled "sub_acct". Now you are rocking and rolling.
- Next, since our objective is to extract the revenue and expense only, we are going to run a direct extraction on the data and pull into a separate database the revenue and expense accounts only. We know that the structure of account numbers is as follows: 4000 = revenue, 6000-8999 = expenses. Therefore, we can run a direct extraction on the account number to indicate: account number >= "40000". This extraction will pull a piece of the pie out of the whole picture. In EXCEL, you can run a filter and copy the data to a separate tab once filtered to work without having hidden rows.

Direct Extraction



Once you have an extracted revenue and expense report, you can run a pivot table on the data to generate a table that shows the account on the left, sub_acct in the column header, and amount fields in the data section. By performing this function, you just generated a revenue and expense schedule by cost center which can be used for multiple purposes including, but not limited to, identification of improper revenue recognition, identification of significant grants and contracts and identification of new vs. continuing projects (if you compare to prior year). This will allow you to more accurately identify major programs early in the process.

Major Program Determination and Extraction

During your major program determination process, you determined that grant number 1018, 1029, and 1049 are your major programs which require testing. Therefore, your next step is to extract out that specific program from the general ledger for purposes of sampling, testing and reporting.

When your data is in EXCEL, you can add a filter (Date > Filter) to the data and filter out the cost center related to the projects being audited. Next, you can copy that filtered data to a new tab. Generate a Pivot Table off that data. In the pivot table fields, use the account and description in the rows, the sub_acct in the column and the amount in the values section. This will generate a profit and loss statement for the specific cost center.

When your data is in a specialized tool such as IDEA or ACL, utilize the extract function to specifically extract the data using the sub_acct field. This will extract out the transaction register for those projects. If you require doing each individually to select individual samples, the extract would be one by one rather than in the same function.

Sampling

Now that you have an expenditure general ledger extracted and segregated by major program, the expenditure testing sample can be pulled from the data. The auditor should consult with the appropriate sampling guide to determine the number of items to select prior to or during the selection process. Once the number of items to select is established, the auditor will need to determine the sampling method: systematic, random, stratified or another.

Below is an example of each method utilizing the expenditure database created above.

Systematic or Fixed Interval

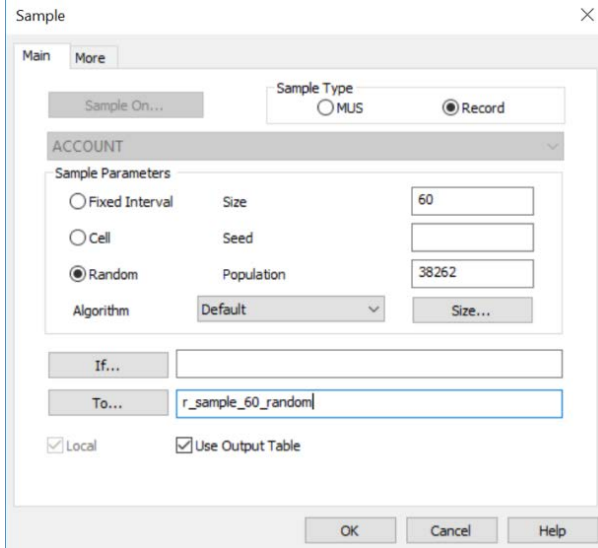
Systematic or fixed interval sampling is a form of sampling where you select every nth item in the population. Select 60 items utilizing the systematic approach.

The screenshot shows a 'Sample' dialog box with the following configuration:

- Sample Type: Record (selected)
- ACCOUNT (dropdown menu)
- Sample Parameters:
 - Fixed Interval (selected): Interval = 637, Start = 1
 - Cell (unselected)
 - Random (unselected)
 - Algorithm: Mersenne Twister
- If... (empty field)
- To...: r_sample_60_interval
- Local: checked
- Use Output Table: checked
- Buttons: OK, Cancel, Help

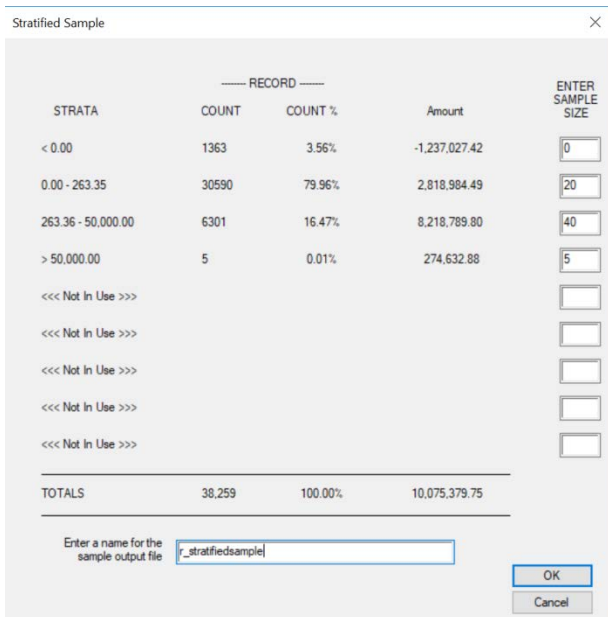
Random

Random sampling implies that a random number will be generated from a computer and used as a basis for sampling. A “random number seed” will be applied, signifying the ordering for that sample. If you re-run, the random number seed will allow you to pull the original sample. Select 60 items utilizing the random approach.



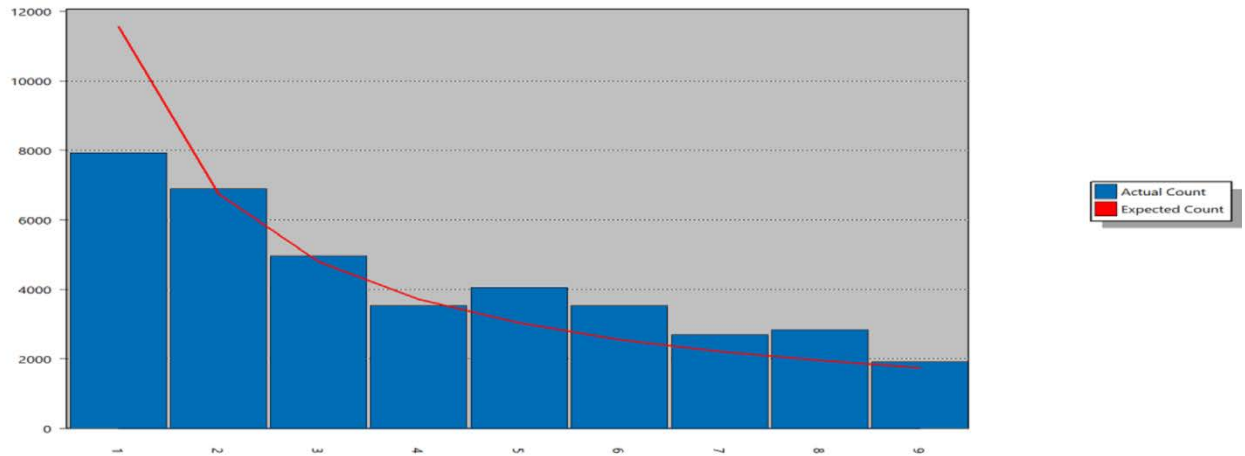
Stratified

Select 60 items utilizing a stratified approach with the average dollar value as the split between strata (total population amount/count) and all items over ISI (50000).



Benford's Law

To assess the population and look for anomalies, we ran a Benford's Law test on the first digit and produced the following graph.



Leading Digits	Actual Count	Expected Count	Zstat Ratio
<u>1</u>	7902	11517	40.284
<u>2</u>	6869	6737	1.766
<u>3</u>	4947	4780	2.576
<u>4</u>	3530	3708	3.060
<u>5</u>	4047	3029	19.259
<u>6</u>	3524	2561	19.684
<u>7</u>	2696	2219	10.430
<u>8</u>	2824	1957	20.108
<u>9</u>	1919	1751	4.108

In analyzing this graph, it was noted that the “1’s” were much lower than expected and the “8’s” were much higher than expected. After further analysis on the data set, it was noted that we included payroll in the data. After further review, many individuals received had an allocated salary in the “\$80” range each pay period which skewed the data. As a further procedure, we selected a few of those transactions to verify their accuracy and identify if any further risk exists.

Reporting

Through a conversation with the client you learned that the reporting is based on specific budget categories which compiles the data that is included within the expense general ledger. Based on this information, you can request and obtain a mapping file for how the client specifically bridges the two reports. From here, you noted that the client maps the file using letters (A to G) which signify the budget categories.

A – Salaries and Benefits

B – Consultants and Legal

C – Occupancy

D – Travel

E – Other

F – Subcontracts

G – Indirect Cost

You were able to ask and receive from the client the mapping file which had a letter next to each general ledger account number. Using this file, you can do a match between account number in the major program expense file with the account number and bring over the new column denoted “category”. In EXCEL, you can use a “vlookup” formula to pull over the category next to the data.

Next, summarize on the category by expense account or create a pivot table in EXCEL. The following is the result:

Category	1018	1029	1049	Total
A	982,256.11	277,579.76	3,350,917.60	4,610,753.47
B	584,611.40	129,474.01	1,141,182.22	1,855,267.63
C	624.00			624.00
D	460,508.26	165,509.74	638,866.28	1,264,884.28
E	65,217.43	37,226.89	332,312.78	434,757.11
F		92,214.73	336.00	92,550.73
G	412,586.66	162,750.90	1,241,202.24	1,816,539.80
Expense Total	2,505,803.87	864,756.04	6,704,817.12	10,075,377.02
Revenue	(2,505,803.87)	(842,000.00)	(6,500,000.00)	(9,847,803.87)
Surplus (Deficit)	(0.00)	22,756.04	204,817.12	227,573.16

Documentation

The following table is an example of documentation that can be done to show the results of each test and how we met our audit objectives.

Process	Function	Result
Obtained “FYXX GL Detail Expenses.pdf” from client	Import into ACL	Workable file; export to EXCEL
Add fields to file	Field manipulation:	Database has additional columns including 1

Process	Function	Result
	<ol style="list-style-type: none"> 1. Data>Append> add “Amount” field = equation editor> debit-credit 2. Data>append>add “sub_acct” field > equation editor > @split(account_number,”-“,”-“,1,0) 	amount field and the sub account required for selection and testing stripped out of the long sub account
Test completeness	<ol style="list-style-type: none"> 1. Summarization of expenditure G/L by “account_number” 2. Compare expense G/L to trial balance provided 	Verified G/L is complete and agreed to trial balance provided
Extract major program (XX.XXX) sub codes from the general ledger	<ol style="list-style-type: none"> 1. Identify the grant sub codes required for testing 2. Perform a direct extraction as follows: Extraction criteria: sub_acct = "1018" .OR. sub_acct ="1029" .OR. sub_acct="1049" 	Child database which includes only the sub codes associated with major program
Generate an expense summary by sub code w/account description	Generate a pivot table from the expense general ledger. From the pivot table, use the account_number and account_description as the “row area”. Use the sub_acct field as the “column area”. Use the amount field as the “data area”	Summary by account number and account description by sub account code; essentially, a summarized expense trial balance by grant
Carve out expense transactions which will not be used in the sample, if applicable. In the case of Example Entity Worldwide, Inc., the indirect allocation (account 8504) is tested separately from the expenditures, therefore, they will be excluded from the sample.	Perform a direct extraction: Extraction Criteria: account_number <>"8504".	Separate data base r_direct expenses
Sampling	Perform a random sample on the population. Sampling>Random>enter “60”	Grant sample – emailed to client as selections (r_sample_60)
Identify significant transactions based on materiality (transactions > \$50,000)	Direct extraction – amount > 50000	r_significant_transactions file sent to client for supporting documentation

NOTES

Take Advantage of Diversified Learning Solutions

We are a leading provider of continuing professional education (CPE) courses to Fortune 500 companies across the globe, CPA firms of all sizes, and state CPA societies across the country, as well as CPA associations and other financial organizations. Our efficient and flexible approach offers an array of customized cutting-edge content to meet your needs and satisfy the priorities of your business. Select from live classes, live webinars, conferences, or online training, including Nano courses, based on your preferred method of learning.

Meet your CPE requirements, increase productivity, and stay up-to-date with relevant industry trends and mandatory regulations with collaborative live or online learning.

Live Training Topics	Online Training Topics
Accounting and Auditing	Accounting and Auditing
Employee Benefit Plans	Business Law
Ethics	Business Management and Organization
Information Technology	Economics
Governmental and Not-For-Profit	Ethics
Non-Technical (including Professional Development)	Finance
Tax	Information Technology
	Management Services and Decision Making
	Personal and Professional Development
	Tax

“We have enjoyed [your] programs and have found the content to be an excellent learning tool, not only for current accounting and management issues, but also how these issues apply to our company and affect how our business is managed.”

—Debbie Y.

Unauthorized reproduction or resale of this product is in direct violation of global copyright laws.

Reproduced by permission from Kaplan.



© 2020 Kaplan, Inc. All Rights Reserved.