An Auditor’s Guide to Data Analytics

Natasha DeKroon, Duke University Health System

Brian Karp – Experis, Risk Advisory Services
Today’s Agenda

- Data Analytics – the Basics
- Tools of the Trade
- Big Data
- Continuous Auditing Case Study
- Benford’s Law
Data Analytics – the Basics
What is Data Analytics?

- Data analytics is defined as the process of inspecting, cleaning, transforming, and modeling data with the goal of highlighting useful information, suggesting conclusions, and supporting decision making.
  - Various sources

- Data analytics is an analytical process by which insights are extracted from operational, financial, and other forms of electronic data internal or external to the organization. These insights can be historical, real-time, or predictive and can also be risk-focused (e.g., controls effectiveness, fraud, waste, abuse, policy/regulatory noncompliance) or performance-focused (e.g., increased sales, decreased costs, improved profitability) and frequently provide the “how?” and “why?” answers to the initial “what?” questions frequently found in the information initially extracted from the data.
  - KPMG
Not a New Concept

- Late 1980s – generalized auditing software companies form
  - ACL, 1987
  - Caseware, 1988

- Charles Carslaw, Applying Benford’s Law to Accounting, 1988

- Continuous Process Auditing System, AT&T Bell Laboratories, 1989
Common Data Types and Data Structures

- Data is generally organized into files or tables
- A table can be thought of as a two dimensional matrix of data
- Each row represents a single record
- Each column represents a data field
- Each data column, or field, may have a different data type
- Data types determine how data is interpreted, and also what data format is considered valid
  - For example, data can be a date, a number, or plain text
  - Invalid data in a table is often a sign of some other problem
- Often, each record in a table may have a unique identifier, like an employee, customer, or transaction ID
  - When one table uses this identifier to reference records in another table, this is called a relational database
  - Relational databases are a very useful way to organize data
  - Many databases are built using some kind of relational database format
Methodology

- Plan
- Acquire & Understand
- Analyze
- Validate
- Report
Data Analytics
Tools
Desired Features in Analysis Tools

- Able to handle large data sets efficiently
- Wide array of analytical and statistical functions and procedures
- Programmability
- Logging of procedures performed on data
- Ability to easily re-run analysis with minor changes
Types of Tools

- Spreadsheet software
- Databases
  - Desktop software (Microsoft Access)
  - Server-based (SQL/Oracle)
- Generalized auditing software
  - ACL
  - IDEA
- Other tools
  - SAS
  - SPSS
  - Apache Hadoop
Generalized Auditing Software – What is it?

Generalized auditing software (e.g., ACL and IDEA) tools are simple, powerful database tools with features designed for an auditor.

- **Read only** – does not modify source data
- **Audit log** – all commands are traceable and repeatable
- **Scripting** – for automated or repetitive audits
- **Powerful data connectivity** – can access almost any data, even text-based reports and flat files
Tools of the Trade

- Spreadsheet Software
- Desktop Database Software
- Generalized Auditing Software
- Server Database Software
Continuous Auditing
Continuous Auditing Defined

“Continuous auditing is any method used by auditors to perform audit-related activities on a more continuous or continual basis. It is the continuum of activities ranging from continuous control assessment to continuous risk assessment – all activities on the control-risk continuum.”

-The IIA, GTAG 3: Continuous Auditing: Implications for Assurance, Monitoring, and Risk Assessment

- Continuous controls auditing using automated tools is one of the methods of continuous auditing
- Set frequency – daily, weekly, monthly, etc.
Case Study

- ACL CCM technology (now replaced by Audit Exchange)
- ACL as the implementation partner
- Six modules: GL, P2P, Payroll, T&E, PCard and O2C
- Multiple data sources (both in-house and third-party): SAP, ADP, Concur, American Express
- Mostly automated but some manual downloads of source data
- Focus on both master and transactional data
- Fraud detection, control deficiencies, data issues
Implementation Project

• Requirements and design specifications – workshops with ACL, business process owners, IT and audit
• Build – ACL
• Testing – ACL and audit
• Training – administrator and end user
• Go-live
• Continuous controls auditing program
Examples of Analytics

General Ledger (24 tests)
- Critical data fields
- Unauthorized journal entry (JE)
- JEs by unauthorized users
- Duplicate JEs (same account/amount, same JE number/amount)
- Split JEs (single JE/multiple accounts, multiple JEs/single account)
- Segregation of duties (park vs. post, post vs. create account)
- Dormant accounts
- Even dollar JEs
- Suspicious keyword in JE description
- Duplicate GL accounts based on the account description
Examples of Analytics

Purchase to Pay (29 tests)

- Critical data fields (vendor master, requisition, purchase order (PO))
- Split requisitions and POs
- Stale requisitions and POs
- Segregation of duties (requisitioner vs. approver, purchaser vs. receiver, requisition approver vs. PO approver, purchaser vs. vendor master administrator, purchaser vs. AP clerk)
- PO date after invoice date
- Invoice number sequence
- Goods received quantity vs. invoice quantity
- Employee and vendor matches by name and by address
- Duplicate vendors (by name, address, bank account number)
- Duplicate purchases (same vendor same invoice number, same amount same GL account)
Examples of Analytics

Payroll (29 tests)

- Critical data fields (payroll master file)
- Duplicate employees (same bank account or address)
- Employee status not matching the termination date
- Exempt hours worked vs. standard hours
- Non-exempt hours worked vs. expected hours
- Hours worked vs. hours paid
- Employee start date after paycheck date
- Terminations within 14 days of hire
- Invalid pay rates (actual/calculated vs. master file)
- Excessive gross pay
- 401k annual contribution limit, catch-up contribution limit and catch-up age limit
- Job record deletions (data corrections not using effective date)
Examples of Analytics

Travel and Entertainment/Purchasing Card (30 + 32 tests)

- Critical data fields (cardholder master, expense, etc.)
- Invalid cardholder (no matching employee or terminated employee)
- Duplicate cardholders (by employee ID or address)
- Suspicious MCC
- Suspicious keyword in the transaction description
- Declined and disputed transactions
- Split purchases
- Duplicate purchases (same merchant same amount)
- New cardholder watch list/cardholder watch list
- Ghost card activities
- Even/small dollar amount transactions
- Weekend and holiday transactions
- Potential duplicate reimbursements: gas with mileage or PCard with an AP purchase
- Spending limits on transactions (lavish hotel stays, dinners, etc.)
Examples of Analytics

Order to Cash (46 tests)
- Critical data fields (customer master, sales order, etc.)
- Duplicate customers (on name or address)
- Credit limits vs. orders
- Segregation of duties (order entry vs. customer master, order entry vs. product master)
- Unauthorized/excessive commissions
- Delivery quantity vs. sales order quantity
- Shipment/sales order/price change by an unauthorized employee
- Cash receipt vs. invoice amount
- Shipment without a sales order
- Days sales outstanding
Benefits and Costs/Risk

• Benefits
  – Automation = saving time
  – Trending of transactions
  – Red flags
  – Master data issues
  – Control culture (you are being watched)
  – SOX, FCPA and other regulatory requirements

• Costs/Risks
  – Investment
  – Time for review, follow-up and communication of results to management
  – Insufficient understanding of source data (can result in many false positives)
  – Lack of buy-in by management
Big Data
Big Data

What is Big Data?

- Voluminous amounts of structured and unstructured data
  - Structured – currently identifiable by user; e.g., database
  - Unstructured – does not fit easily into traditional relational systems; e.g., email, word processing documents, multimedia, video, PDF files, spreadsheets, social media
- Defined in terms of petabytes and exabytes

“Ever more powerful information technology now allows consumers to carry gigabytes in their pockets and businesses to organize and analyze data on a scale never seen before. People’s willingness to use the new electronic tools to communicate and share information about themselves means that even the most advanced companies are only scratching the surface of the behavioral patterns these troves of data can potentially reveal”

Source: The Financial Times
Shift in Data Sources

The challenge is:

• Approximately 75-90% of data is unstructured (while IT is built for structured data)
• Unstructured data is growing at nearly 10x the rate of structured data
• Less than 5% of unstructured data is proactively managed
The Four “Vs” of Big Data

**Volume**
Amount of data generated or must be ingested, analyzed, and managed to enable business decisions

**Velocity**
Speed at which data is produced and changed; the speed at which data must be received, processed and understood

**Variety**
Both structured and unstructured data generated by a wide range of sources

**Veracity**
The quality and accuracy of received data
**WHAT IS A PETABYTE?**

To understand a Petabyte we must first understand a Gigabyte.

- **1 Gigabyte** = 7 Minutes of HD-TV Video
- **2 Gigabytes** = 20 Yards of Books on a Shelf
- **4.7 Gigabytes** = Size of a Standard DVD-R
- **There are a million Gigabytes in a Petabyte**

"Let me repeat that: we create as much information in two days now as we did from the dawn of man through 2003." (That's something like 5 Exabytes of Data). - Eric Schmidt - Google 8/10

**Twitter:** Over 7TB a Day in Tweets.

**Facebook:**
More than 750 Million Users.
Average user creates 90 pieces of content each month.
More than 30B pieces of content shared each month.

---

**A PETABYTE IS A LOT OF DATA**

- **1 Petabyte** = 20 Million Four-Drawer Filing Cabinets Filmed with Text
- **1.5 Petabytes** = 13.3 Years of HD-TV Video
- **20 Petabytes** = Size of the 10 Billion Photos on Facebook
- **20 Petabytes** = The Amount of Data Per Day Processed by Google
- **50 Petabytes** = Total Hard Drive Space Manufactured in 1995
- **50 Petabytes** = The Entire Written Works of Mankind, from the Beginning of Recorded History, in All Languages

---

**A ZETABYTE IS ONE MILLION PETABYTES!**
Forces Impacting Utilization of Big Data

Technology
- Programming, Infrastructure, Cloud Computing, Integration
Data Generation
- Architecture, Modeling, Extraction
Analytics

Talent & Skill Sets

Data
- Quality vs. Quantity
- Structured vs. Unstructured
- Internal vs. External

Governance & Privacy
- Who
- What
- Where
- How
Benford’s Law
What is Benford’s Law?

Mathematical theory of leading digits. Leading digits are distributed in a specific, non-uniform way.

- Simon Newcomb, 1881
  - Described theoretical frequency that is Benford’s Law
- Frank Benford, 1938
  - Numbers starting with 1, 2, or 3 are more common in nature than those with initial digits 4 – 9.
- Charles Carslaw, 1988
  - Conducts study of tabulated income numbers
  - Concludes that management actively rounds up income numbers so that they look “better”
- Mark Nigrini, 2000
  - *Digital Analysis Using Benford’s Law*
Benford’s Curve

Source: ISACA Website
Benford’s Law

Data Set Criteria

Major Digital Tests

Adjusting the Curve

Reporting

Practical Application
Benford’s Law Resources

• **ISACA Understanding and Applying Benford’s Law:**

• **IIA Putting Benford’s Law to Work:**
  – **Applying Benford’s Law in Excel:**

• **Mark J. Nigrini, Ph.D., Digital Analysis Using Benford’s Law**
Questions?