

WHITE PAPER

Decrease Tax Fraud Using Validation Through Taxpayer Pattern

White Paper

By Hitachi Vantara

Introduction

One of the greatest challenges facing financial institutions and main-street America is identity fraud (IDF). A recent study indicates 13.1 million Americans were victims of IDF in 2013, or one victim every two seconds, resulting in a direct economic loss of \$18 billion (Teller Vision, 2014). It appears there is a major data breach monthly that exposes consumer financial information placing them at increased risk. Home Depot, Target and Staples are major retailers that have experienced significant data breaches. IDF is real and the trend indicates it will only get worse.

Unfortunately, the Internal Revenue Service (IRS), one of the world's largest and most respected financial institutions, is not immune to the phenomenon. The IRS has felt the impact of IDF manifesting itself in the form of fraudulent tax returns, affecting billions of dollars (GAO, 2014; IRS; Rosen, 2014). In particular, the most serious problem stems from criminal use of stolen names and social security numbers to file tax returns. According to an August Government Accounting Office Report (2014), the Internal Revenue Service estimates that the agency paid over \$5 billion in fraudulent claims of this type in 2013, but was able to prevent close to \$25 billion in losses.

The full extent of IDF will probably never be known because it is difficult to detect, particularly since tax returns are filed on an "honor system." This means that the data is not available to validate returns before they are processed because earning statements are not aligned with the filing process (GAO, 2014). This places the agency at a significant disadvantage, and criminals continue to exploit information because they are keenly aware that a viable validation system does not exist. Fortunately for tax payers, the agency is introducing improved technology, such as analytics. They are also working with Congress and the Department of Treasury to update policy legislation to deal with this growing problem.

Why Is It So Easy To Commit Identity Tax Fraud?

Several key factors make IDF easy.

Policies

- Employer wage data is not available when most fraudulent tax returns are electronically filed.
- Employer wage data is sent directly to the Social Security Administration (SSA), not the IRS.
- Refunds are directly deposited into bank accounts or debit cards; many fraudulent returns can be deposited to the same account or debit card.
- Social Security numbers from the Social Security Death Index are readily available and easily obtainable by the public.

HOW DOES TAX IDENTITY FRAUD WORK?

A violator steals a social security number from a taxpayer and files a tax return under the taxpayer's name using false, unvalidated wage data (W-2) before the taxpayer has filed their return. The criminal collects a refund from the IRS before the taxpayer's wage data can be validated by an employer.

Lack of Collaboration Between Federal Agencies

- Siloed systems in multiple federal agencies lack timely data-sharing capabilities.
- Government programs are often run by completely different groups, each with its own audit, investigations and compliance functions.

The IRS needs to create a real-time tax system: a cross-functional task force to transition into a system in which the IRS matches tax returns against third-party reports (for example, employer wage data, Social Security Administration Death Index, financial institutions, other federal agencies) before paying out refunds, and makes this information available in electronic form to taxpayers for assistance in preparing their returns.

Solutions

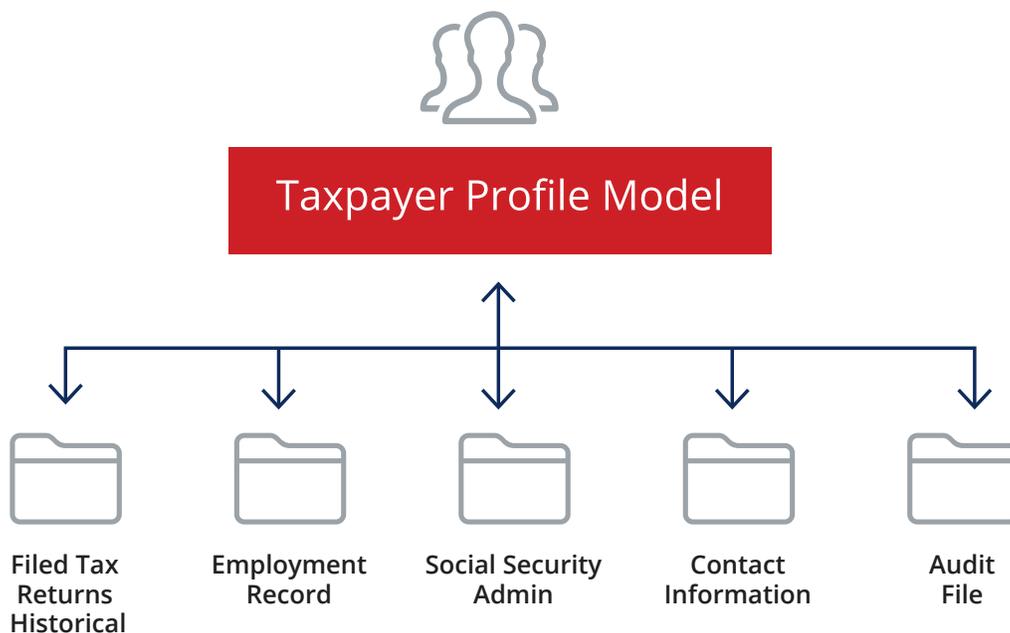
Our solution centers on an “inferred validation” concept: In the absence of actual data to substantiate taxpayer earnings and withholdings, such as employer wage data, new returns will be screened by comparing them against historical taxpayer filing patterns, Social Security Administration (SSA) data and financial institution accounts information. If the new return exhibits similar patterns based on historical data, the IRS can infer that a tax return is valid. If inconsistencies are found, the IRS can infer an invalid or fraudulent tax return that would require review and interaction with the taxpayer before processing and issuing a refund.

This solution can complement current IRS efforts through utilizing the vast resource of data not available to those who seek to exploit gaps in the tax-return filing process. That data, as mentioned in the previous paragraph, will be aggregated in a meaningful and timely way to allow the agency to “connect the dots” among taxpayer historical filing patterns, SSA vitals, financial accounts data and a new return, similar to the use of DNA in forensics. Metaphorically speaking, if there is no DNA match between the historical data and a new return, one could conclude with confidence that the return is fraudulent.

There are two major components to implement this solution. The first is creating a taxpayer profile as a validation mechanism for new returns. The second is a screening process based on a predetermined but dynamic set of rules to compare the new return and infer a processing decision based on scoring using predictive analytics.

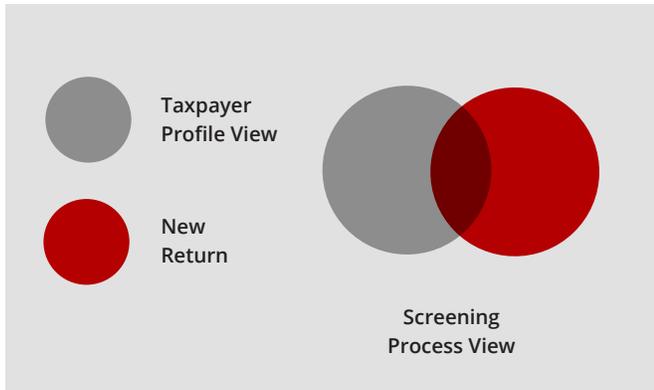
Taxpayer Profile

Creating a profile for each taxpayer is the first step to combating fraudulent returns. Using existing historical tax filing records and vital information from the SSA, the IRS can gather intimate insights on each taxpayer that paints a more holistic understanding of individual filing patterns. Armed with this information, the agency can shift to more of a fraud-prevention approach by having the ability to predict the patterns taxpayers should follow based on their profile. This is not a new concept: Retailers and marketing firms examine customer consumption patterns to better understand how and when to target specific customers (Dahlström, & Edelman, 2013; Ting, 2013). Imagine the possibilities that could flourish by mining the massive amounts of data to build a unique profile for each filer, something very difficult for a criminal to replicate.



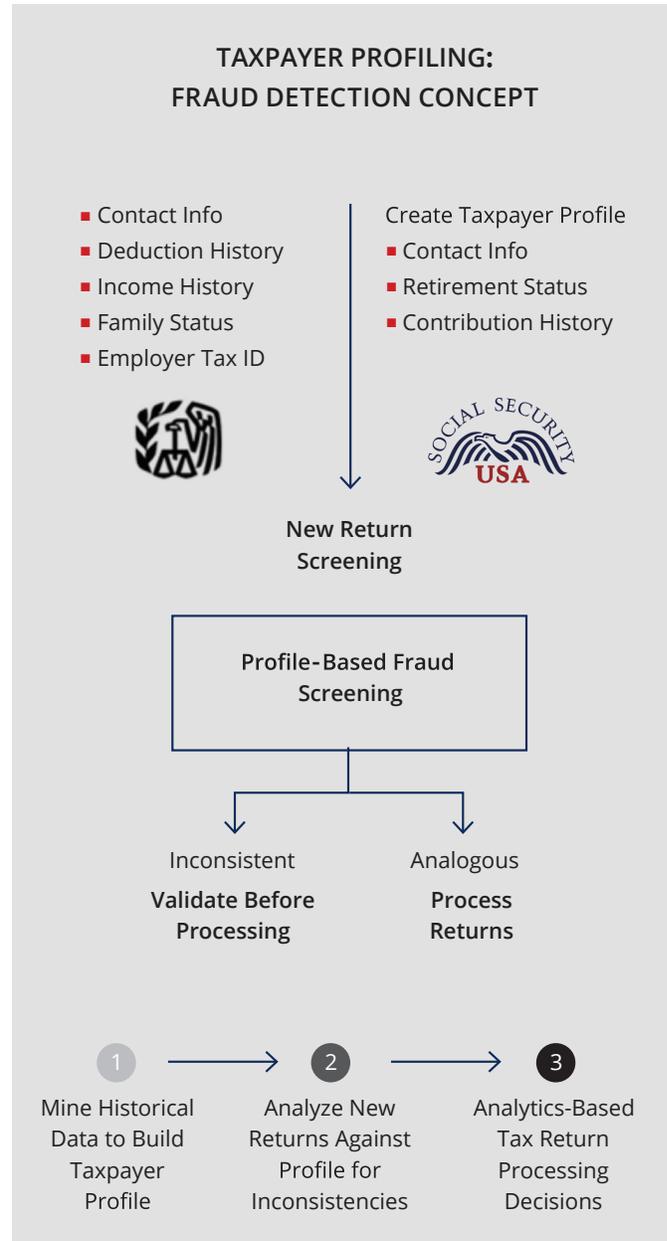
New Return Screening

The second component is an automated screening application that compares the taxpayer profile with the new return to look for inconsistencies in taxpayers' filing patterns. The agency will have the ability to "tune" or adjust the parameters to account for changes that typically occur during each tax year. For example, a taxpayer profile may show an average gross income of \$23,000 and one exemption. To account for changes, the agency may adjust the screening criterion that includes an income range of +/- 15%, and up to three exemptions. As a new return arrives for this taxpayer, as long as the gross income falls within the 15% and the exemptions don't exceed three, the IRS can infer a valid tax return. Consider the scenario in the chart below where there are several inconsistencies.



An evaluation of the screening process view reveals several inconsistencies between the historical taxpayer profile view and new return. There is a significant income increase; the tax payer is claiming three additional exemptions along with a \$17,000 deduction for mortgage interests; and the address and phone number changed. Because there is a significant difference between the profile and the new return, this record would be flagged and require validation with the filer to confirm the changes before processing. Of course, there is a possibility the taxpayer moved to Portland for a higher paying job, got married and had two children, and purchased a home that is unaffordable, based on newly reported income.

The solution will collect pertinent data from IRS and SSA records to build a unique profile for each taxpayer. This profile will present a holistic view of the taxpayer's historical filing patterns and current social security status. As the taxpayer files a new return, it will be compared or screened against their personal profile to look for inconsistencies based on rules set by the agency. Based on the screening, the new return will either be validated for processing because it is consistent with past filing patterns, or it will require the agency to further validate the return due to inconsistencies. The figure below illustrates how our solution will work.



How Is This Concept Different From Other Efforts?

There are two key attributes of this concept that make it different from other initiatives. First, it seeks to holistically connect the dots between the unique and dynamic historical information about each taxpayer with future returns. The agency will have a 360-view of the taxpayer, and not just two or three disparate data points, to make analytic decisions within a full range of contextual situations. In our assessment, this will reduce return screening errors that may inconvenience the taxpayer, or worse, lead to processing fraudulent returns. In both situations, there are no real winners.

Second, because the capability will be designed for functional users to dynamically set the profile-matching parameters for new tax return screening, the agency will have the flexibility and agility to address emerging threats. In other words, there will be no requirement to have a team of programmers or analysts because these capabilities will be built into the system. Based on our research of the problem, our team now has an appreciation of the dynamic nature this presents the IRS. We understand the need for a solution that stays one step ahead of the threat until tax policy and legislation can be more closely aligned with the operational needs of the agency.

CUSTOMER EXAMPLE: FINANCIAL REGULATORY BODY

Goal

Empower analysts to identify suspicious patterns among billions of market transaction records per day.

Pentaho Solution

Users explore summary data with the ability to request detailed data sets on the fly for drilldown through multi-dimensional models.

Architecture

Leverages Hadoop with Amazon Elastic MapReduce and Hive; uses Amazon RedShift as a high-performance analytical database in the cloud.

The Streamlined Data Refinery Solution

A streamlined data refinery architecture meets all of the core requirements of the use cases described above, providing a user-driven, trusted, data-delivery process. At its core, the design pattern accommodates an on-demand process of user-initiated data requests. It blends and refines any data, automates analysis schema generation, and publishes analytic data sets in the format of choice.

Scalable Data-Processing Hub

Usually Hadoop, this store is meant to house and manage a variety of structured and unstructured data from across the organization. In the diagram, Hadoop serves as the landing zone for data across the web, social media, transactional systems, and machines or sensors.

High-Performance Database

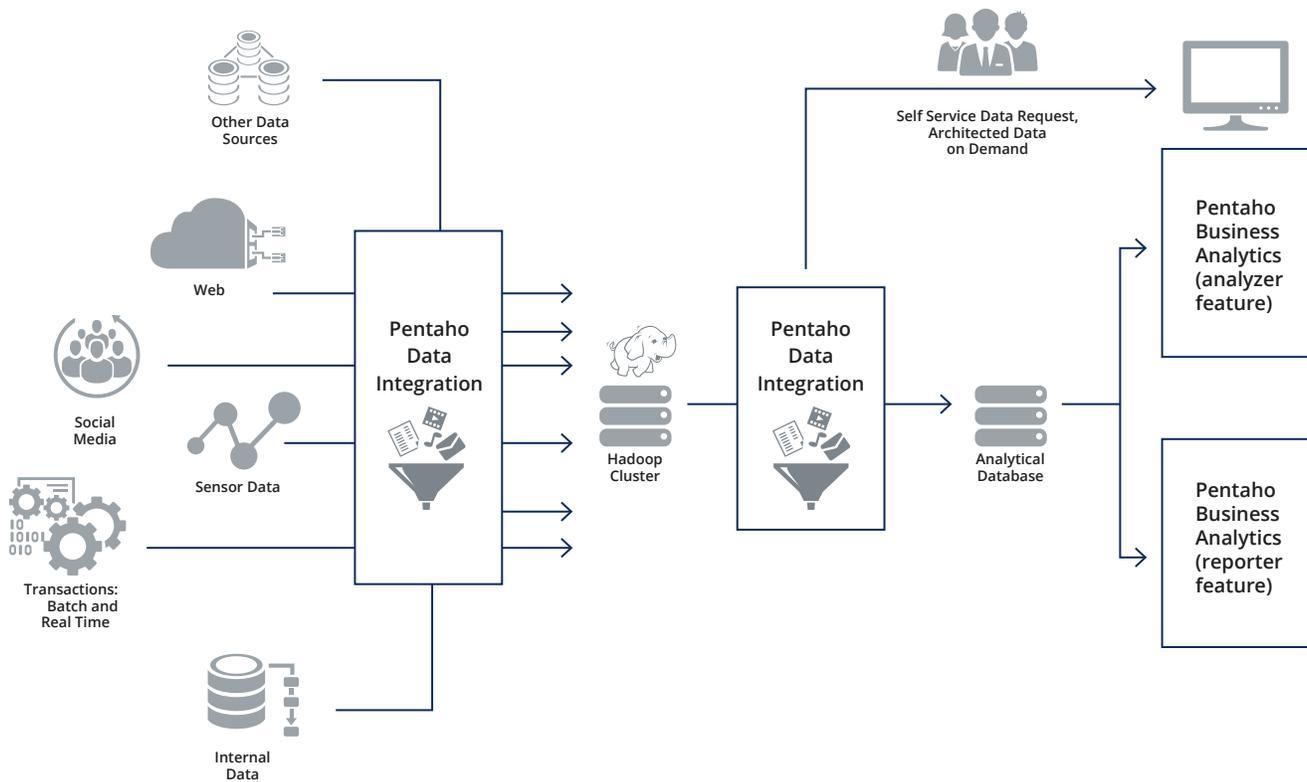
The database chosen must facilitate high-performance queries for analytics and visualization. When scale is required, an analytical database such as HP Vertica is a solid choice.

Pentaho Data Integration (PDI)

The Pentaho platform's highly scalable data integration engine, managed through its intuitive end-user interface, provides the "glue" between the different data sources and stores in this architecture. The entire process outlined here can be triggered on-demand via PDI:

- **Blending and orchestration:** PDI ingests data from virtually any data source, including both traditional systems and Big Data stores. Then, it processes, cleanses and blends the data in the required combinations to drive insight.

STREAMLINED DATA REFINERY ARCHITECTURE DIAGRAM



- **Automatic modeling and publishing:** As part of the data orchestration process, PDI automatically creates an OLAP schema and publishes it to the Pentaho Business Analytics server for end user exploration and visualization.
- **Governance:** PDI's robust functionality enables IT to quickly and easily validate data sources being blended at the source. It allows for the right measure of control, without creating unnecessary frictions to end user data access.

Conclusions

In this discussion, we highlighted three core data delivery needs that are only being met on a limited basis in the market today:

- Orchestrate on-demand processing, blending and modeling of user requested data sets in order to accelerate time to value in complex analytics initiatives.
- Ensure proper data governance during the delivery process so that risk is minimized and confidence is increased in data-driven decisions.
- Provide blended and enriched data in the end user format of choice so that business users can be more productive in deriving insight from diverse data.

The Streamlined Data Refinery's User-Driven, Governed Process



Indeed, these challenges cut across a variety of sector-specific use cases discussed, including “deep data” exploration by researchers, forensic analysis of unexpected events, compliance assurance in regulated industries, and delivery of data to key customers and partners “as a service.”

The streamlined data refinery provides a well-defined architecture to address these needs in a fashion that both leverages existing organizational competencies and ensures that the on-demand data delivery process can quickly adjust to changes in the data environment.

References

Dahlström, P., & Edelman, D. (2013). The coming era of “on-demand” marketing. *Mckinsey Quarterly*, (2), 24-39.

Government Accounting Office. (2014) Identity theft: Additional actions could help IRS combat the large, evolving threat of refund fraud (GAO-14-633). Retrieved from <http://www.gao.gov/products/GAO-14-633>.

Internal Revenue Service (n.d.). Internal Revenue Service strategic plan FY2014-2017. Retrieved from <http://www.irs.gov/pub/irs-pdf/p3744.pdf>.

Rosen, I. (Producer). (2014, September 21). 60 minutes — Biggest IRS scam around: Identity tax refund fraud [Television broadcast]. New York, NY: Central Broadcasting Service.

Social Security Administration (n.d.). Agency strategic plan 2014-2018. Retrieved from <http://www.ssa.gov/agency/asp/materials/pdfs/plan-2014-2018.pdf>.

Teller Vision (2014). Identity fraud creating a new victim created every two seconds (cover story). *Teller Vision*, (1440), 1-2.

Ting, R (2013, March 11) The customer profile: Your brand's secret weapon (BLOG). *Harvard Business Review Blog*. Retrieved from <http://blogs.hbr.org/2013/03/the-customer-profile-your-bran/>

Learn more about Pentaho Data Integration and Pentaho Business Analytics at HitachiVantara.com



Hitachi Vantara at a Glance

Your data is the key to new revenue, better customer experiences and lower costs. With technology and expertise, Hitachi Vantara drives data to meaningful outcomes.

Hitachi Vantara

Corporate Headquarters
2535 Augustine Drive
Santa Clara, CA 95054 USA
hitachivantara.com | community.hitachivantara.com

Contact Information
USA: 1-800-446-0744
Global: 1-858-547-4526
hitachivantara.com/contact



HITACHI is a registered trademark of Hitachi, Ltd. All other trademarks, service marks and company names are properties of their respective owners.

P-025-B BTD August 2019