EU GDPR & US Privacy Shield, Who is Most at Risk? With such Large Fines, Will Companies Sue for Damages if they Fail an Audit?

Top Management Consulting Firms, Software Companies, even The EU and US regulating offices may be targets when a fine of 4% of last year's receipts is on the line and perhaps the company's future.

All participants in the alarming and complex world of regulating Personally Identifiable Information are breaking new ground as PII has never been governed to this extent before. The risks are obviously high for companies doing business with EU citizens if they fail to protect PII data, Indirect Identifiers, or fail to remove PII data for an individual requesting their right to be forgotten. But what are the risks for companies providing guidance and direct assistance to those companies preparing for an audit by EU or US regulators? Will Consulting firms prudently change, limit or stop offering certain guarantees?

One thing we know for sure, the European Union is serious about protecting its citizens and enforcing its regulations. Today, June 27, 2017 the EU announced a **$2.7 Billion dollar fine against Google** for using data in a way that was contrary to the best interests of EU citizens. This may be a strong indicator that when the EU begins assessing fines for breaching GDPR regulations, they won't be lenient. GDPR regulations have already been in effect for over a year; the EU has just delayed assessing fines and penalties until May of 2018 as a grace period.

What expectations will a company have when they hire a Consulting Firm to provide guidance, technical assistance, software and process re-engineering to prepare for an audit? If a company unexpectedly fails an audit will they seek compensation from the consulting firm they relied upon to avoid such a situation or the consulting firm seek compensation from the software vendor or both? I would like to highlight several technical areas of concern that a consulting or software firm needs to address before engaging clients in this new era of EU GDPR regulations;

1. In today's IT world, companies have enormous volumes of data. Will your solution interrogate every record/row of a client's database, or just a subset? Will the software used by EU regulators be superior to your solution and scan every field in every record/row and discover PII data that you missed? Will they use algorithms you did not think about or code properly?
2. Will your solution search for all the various forms of PII data in your client's data base and will it equal the number of PII data types being searched for by EU GDPR software and services experts?
3. Will your solution discover Indirect Identifiers? If EU Auditors use applications and methodologies with more advanced capabilities both you and your client may be vulnerable.
4. Will your solution have the ability to search every file/record/row and column of data for all information associated with a person that has requested their right to be forgotten? And continue to repeat this process daily or as new data enters your file systems?
5. Will your solution be effective across all the major data platforms in use today, IBM, Oracle, Microsoft SQL Server, Teradata, Hadoop, Word, PDF, and other **Unstructured data**? If EU regulators are expecting to audit many different firms they certainly expect to audit data residing in all of the platforms and may have the capability to do so.
6. Perhaps **the MOST IMPORTANT Question to consider is, how do you keep PII data from filtering back into a client's database? Remember Data is rarely Static.**
   - Imagine your client purchases a large dataset containing emails, or names and addresses just a month after your engagement. Do you expect your client to re-engage your services to discover and encrypt the PII data contained in that dataset? Why would a client initially engage your services if you have no plan for continued ongoing surveillance?
   - If an individual asks for all their PII to be removed and never captured again, how do you ensure it won't be reintroduced from various others sources? An Ongoing procedure to scan for PII data for all individuals that have asked to be 'forgotten' must be in place.

7. Finally, how do you find enough time in a day to keep processing a client's data stores to encrypt and sequester PII and to remove PII for individuals requesting to be forgotten?

    - The Information world has changed dramatically from the days when files where relatively static and evening jobs could be run to perform various tasks in a 6-10-hour window. Modern practices demand 24/7 real time processing with data continuously streaming into a database, data lake or data warehouse.

Over the past two years BigDataRevealed (BDR) has dedicated its resources to building a software solution and developing a methodology that directly addresses EU GDPR and most any data compliance. Let me explain;

In order to implement a GDPR software solution for a client it will be necessary to incorporate all their data sources into the solution. With the numerous legacy SQL and Mainframe platforms, Word documents, e-mails, PDFs and other forms of unstructured data prevalent in the business world today, you will need to be a master of many programming languages and understand all the databases in use. Seems as if that is a daunting task! And we believe you are absolutely correct, and that is why BigDataRevealed selected Apache Hadoop as a staging area for all GDPR data discovery, remediation and intelligent catalog/metadata creation. These are the reason why;

1. Hadoop can provide incredible processing power to complete the discovery, sequester/encryption and data removal functions.

2. Hadoop can provide enormous storage capacity and be easily expanded if required.

3. Hadoop will accept any digitized data, no matter what platform the source data is coming from, even unstructured sources of data.

4. Only one programming language is required to process any source data.

5. Hadoop is FREE. No need to pay license fees for each type of data source processed.

6. An ETL tool will be necessary to swiftly connect data sources to software running within the Apache Hadoop environment. However, many quality, inexpensive ETL tools are available in the market place.

These points support my philosophy and methodology of using Apache Hadoop, or some form of common staging platform, to assemble disparate data sources before processing. Using well designed software that is totally embedded within the staging platform will most efficiently use resources to complete all phases of GDPR compliance processing, such as PII identification, sequestering/encryption and PII removal for individuals exercising their 'Right to be forgotten'. A strong software product could also provide additional functionality when not engaged in GDPR violation discovery, such as the creation of an Intelligent Catalog and Metadata that can be used in a collaborative manner for analysis, BI reporting and Outlier evaluations.

Back to the original topic of this article. Where does the buck stop when regulators discover GDPR violations? Is it based on when the PII data exposure was found compared to the date of final remediation? Who will really know when the problematic data entered the data base? Perhaps hackers will intentionally insert PII data into legacy systems just for the thrill of causing problems. How much new data will stream into a data base from third party marketing data, IOT or even a disgruntled employee. What advice and assistance do you provide for your client?

In order to build a reasonably failsafe process and get as close to Zero Latency as possible in your remediation process you could consider:

1. Running Discovery processes 24x7 knowing that it may take several days to evaluate all the data.
2. Run real-time discovery when porting data into your data lake or legacy systems
3. Write your own intake processes to discover and remediate on the fly before PII and Private data is written into your file systems
4. Run real-time discovery of Right to be Forgotten groups before data is written into your data files

There appears to be no straight forward answer to the question of who may be at fault when a client is assessed a crippling GDPR fine. We may need to wait for Cases to be resolved in court before the answers present themselves. We hope our methodology summary presented here will protect your firm from suffering loss of business or reputation as the result of a client receiving heavy GDPR fines or fighting unanticipated court challenges.

More about BigDataRevealed:

We believe our choice of Apache Hadoop is the most reasonable platform for the task of preparing for EU GDPR compliance. Below are the features available within BigDataRevealed (BDR) that make us unique.

1. BDR is an application that is collaborative, reusable, and able to Discover all forms of PII data patterns while being easily augmented to include other Patterns desired by the user company.
2. BDR stores and create an intelligent catalog with extensive metadata for file and columnar naming and business classifications of all files/columns
3. BDR uses a powerful Regular Expressions engine, NLP (natural language process), data mining algorithms. All these must be extensible, sharable, collaborative, schedulable and able to send warnings when issues arise.
4. BDR is able to Sequester Originating files in Hadoop Encrypted Zones, Encrypt file columns based on User rules or processes, process each and every record so not to miss those random social security numbers buried in text blocks storing IP Addresses.
5. BDR can create and search for User defined and Regulatory defined ‘Indirect Identifiers’ in accordance with the EU GDPR Regulations. An example is to not allow a Zip Code, Birthdate and Gender to appear in a single row because of the small number of individuals that would have all those values.
6. BDR uses Free Apache Hadoop as a Staging area for all the numerous sources of legacy data, IoT streams and unstructured data, all consolidated.
7. BDR is 100% embedded within Apache Hadoop and able to utilize the massive processing power and speed of Hadoop.
8. BDR has callable API’s so that other technologies, such as ETL and Data Movement tools can work in unison with BigDataRevealed.

We would appreciate comments from any and all readers so that we can improve our methodology and form a consensus on what constitutes best practices. We value comments concerning software products that you feel deliver value for the coming era of PII regulation. We will attempt to share these ideas in future posts with our regular readers.

Tell us some of your experiences and please don’t hold back. An unsuccessful endeavor, as well as a successful one provides insight. Ask us questions or propose thoughts on what you expect to happen with EU GDPR and US Privacy Shield.

Let’s collaborate.

Steven.meister@bigdatarevealed.com – info@bigdatarevealed.com - EUGDPR Requirements