# SIT719

# SECURITY AND PRIVACY ISSUES IN ANALYTICS

## ASSIGNMENT-2

## DEAKIN UNIVERSITY

**NAME-SHANTANU GUPTA**

**STUDENT ID-218200234**

## Executive Summary

We know, that we are living in a data-rich or big data society. Today's modern businesses wanted to gain competitive advantage and for this to remain innovative and we can achieve this by using advanced analytics and machine learning. But on the other hand, people do have concerns whether these pattern mining technologies may violate people's privacy or security. Dumnonia Company is one of Australian's leading insurance companies which is focusing to take a comprehensive approach to privacy, by understanding and minimising privacy risks, at the same time maximising data utility.

This report provides an analysis and evaluation of the current and prospective privacy technology solutions in the world and how we can implement these technologies within an organization. Methods of analysis to protect privacy include K-anonymity, different extensions of k-anonymity like L-diversity and its flavours, differential privacy as well as such as common privacy risks within the organization and how to protect them. All the implementation guide is in reference with Dumnonia Company.

The report also investigates the fact how we change from traditional enterprise storage system to the cloud storage system. The report also evaluates different AWS cloud security solution like AWS Cloudwatch, different security IT products regarding cloud security, and how to move your old data to the cloud system.

This report will not only evaluate all the possible design solution to the privacy and security problem but will also evaluate a commercially viable solution to the problem.

The report is intended to provide the executive board with background information to assess the feasibility of a proposal they have been asked to assess concerning the design of a comprehensive approach to privacy.

## Organisational Drivers

- **The Pace of innovation:** Growing adoption of Artificial Intelligence, internet of things, data mining, machine learning etc. will help Dumnonia Company to radically innovate and experiment with this new business models. For example, risk management is enabled by analytics that prevents wasted time with policyholders, also reduce Fraud, and also help

in managing pricing premiums. For example, an agent can monitor data in real-time from various social media platforms to see if a policyholder might be engaging in fraud.

- **Customer Expectations:** The widespread use of new customer technologies has created greater demands for new insurance solutions and better interaction channels. Previously Dumnonia is using techniques for evaluating risks of ensuring that person by actuarial data, claims data, and a bunch of other techniques to create better products. But now, after moving to the big data system they design products that would be more customer-centric. For example, intelligent management platforms feature smart dashboards that you can access as an agent to get a complete overview of each client's portfolio
  - Faster claims processing leading to improved customer satisfaction
  - Personalized products and offerings
- Regulatory Pressures / Cyber risk regulations from government. Different policies like GDPR, compliance factors etc.
- Website/App traffic/ Customer based portal interaction
- Organization Reputation
- Old data and real-time data help Dumnonia how profitable their insurance business and accordingly change their sales practices to improve those profits. In return, they get maximized returns from investments that are impacted by the low-interest regime, control / prevent fraudulent claims. A Real-time analysis will help this company to predict the probability of a policyholder being involved in an accident or having some kind of disease.

## Technology Solution Assessment

Before diving into real implementation tools for protecting privacy, I wanted to mention some technical ways where we guarantee privacy.

### Technical Approach Guarantees for Privacy

- Access Control
- Perturbation based scheme
- Secure multi-party computation like homomorphic encryption, quantum encryption
- Anonymity
- Query control

Besides these approaches, you need follow Code of Conduct rules and following good ethical practices in designing privacy policies. Follow Privacy by design approach for making privacy policies for the company and policies should be not complicated like Twitter, Coursea, as it will find difficult for understanding a naïve user. Privacy policies should be simple to understand. For Example-Facebook has very good written privacy and data policies. So, every company should treat Facebook policy as a benchmark for them.

**Access Control:-**We can prevent this by using password control, anti-virus, firewalls, tokens, and two-factor authentication etc.

**Perturbation based Scheme:-**

There are two main types of data perturbation scheme. The first type is known as the probability distribution approach in which we take the data and transform to some other probability distribution curve without affecting original data points, this is also called a synthetic data and the second type is called the value distortion approach by adding some type of noise like multiplicative, speckled or additive noise, or other randomized processes

It is relatively easy and effective technique for protecting sensitive health data from unauthorized use.

**Query control: -** You restrict some of the queries you can ask over the data. So this is instead of just handing it over, you put it behind an API, and try to be careful on what kind of questions people are asking.

**Anonymity Model:-**

1) **ARX-Data Anonymization Tool**

- Open source software for anonymizing sensitive personal data
- Developed in close cooperation between the Chair for Biomedical Informatics, the Chair for IT Security and the Chair for Database Systems at Technische Universität München (TUM), Germany
- Supports a wide variety of data:
  - Privacy and Risk Models
  - Transforming Data
  - Analysing usefulness of output data
- Software used in variety of applications:
  - Commercial Big Data Analytics platform
  - Research Projects
  - Data Sharing

- Training purpose
- Handle large datasets
- Codes and API is written in Java
- Tool transforms datasets into syntactic, statistical and semantic privacy models that mitigate attacks leading to privacy breaches.
- Supports all privacy model, data transformation model and quality model
- Graphical frontend of ARX provides various visualizations, wizards and a context-sensitive help.
- ARX 3.7.1 version is currently used
- Platform independent java application
- Good Documentation
- Download Free software from official website
- This tool also provides built-in data import facilities for relational databases (MS SQL, DB2, SQLite, MySQL), MS Excel and CSV files.

**2) UTD Anonymization Toolbox**

- Developed by University of Texas, Dallas
- Toolbox currently contains 6 different anonymization methods over 3 different privacy definitions
- Last stable version released in 2012
- Documentation is available but not exhaustive

**3) SECRETA**

- Evaluating and comparing anonymization algorithms for relational, transaction and relational-transaction datasets.
- 9 anonymization algorithms & 3 bounding methods
- Interactive Visualization
- Operates in two quality metric: Evaluation and Comparison
- Backend is written in C++
- Good Video presentations how to use this tool
- No link for downloading

Other tools like Cornell Anonymization, TIAMET, and open anonymizer are currently not in used because they stopped working after their first product release. These software's don't have the official website. Not much information about these software's.

Several companies/labs are working to create technologies that efficiently extract useful information from any data without sacrificing privacy like Annon Tool, µARGUS, sdc-MICRO. These tools are handling specific algorithms with the privacy. They are not the complete software and there

is lot of research is going in these software to make complete, powerful and robust privacy software. For example, sdc-Micro is an R package used for the generation of anonymized (micro) data.

### 4) Privacy Consulting Companies

- Companies like Privacy analytics, Privitar, Imperva and many more security consulting companies
- These are the companies that are giving their products to the customers to leverage data with an uncompromising approach to data privacy, comply with regulations and reduce risk and privacy-preserving data operations.
- For example like Privitar has three products like Privitar Publisher, Privitar lens, Privitar Secure link while Privacy analytics have product like Privacy analytics eclipse, Privacy analytics lexicon
- Product selection depends on customer requirements and budget
- For example Privitar publisher products have features like all the anonymization algorithms, Tokenization, encryption, masking, lookup etc. In addition to these technologies they are giving protected data domains, watermarking and meta-tagging technologies.
- Privacy analytics eclipse software are totally based on health data. They are protected health information with useful and high quality data information. Their main aim is to protect individual privacy. Their algorithms is verified by the industry experts. Their products is based on more de-identification process.
- Privitar's & privacy analytics patented software solution allows the safe use of sensitive information enabling organisations to extract maximum data utility and economic benefit.

### 5) Cloud Security in AWS

First we investigate the different problems in Cloud Security:-

- Data Access-Who manages and have access to the data.
- Laws & Regulation-Where data is stored and what laws apply
- Contract termination-Will data remain in the cloud after termination of service.
- Data Breach-Will you know when there is a data breach.

Threat Identification is done in 3 stages in the AWS cloud.

- Monitoring Data-Using Machine learning algorithms you can set your monitoring application to flag an event.

- Gaining visibility-Once you get to know something wrong is going on, you would want to know when & where.
- Managing Access-With managing access you will have a list of users who have access, and hence wipe the culprit out of the system.
- **AWS Cloudwatch:** Cloud Monitoring Tool
  - Monitor EC2 and other AWS resources
  - The ability to monitor custom metrics
  - Monitor and store logs
  - Set Alarms
  - View Graphs and Statistics
  - Monitor and react to resource changes
- **AWS CloudTrail-**Gaining visibility. To track the activities that is happening in your AWS account
  - CloudTrail is a logging service which can be used to log the history of API calls.
  - It can also be used to identify which user from AWS management console requested the particular service.
  - This is the tool where you will identify the notorious “hacker”
- **AWS IAM-**For managing access
  - Granular Permissions to employee
  - Secure access to application running on EC2 environment
  - Free to use

Amazon also gives the option to manage the cloud security by the organization itself. In this there are two options.

- **Option1-**Where organization control the encryptions method and the entire key management infrastructure(KMI)
- **Option2-** Where organization control the encryptions method and AWS provide you the entire key management infrastructure (KMI) and you provide the KMI management layer.

Depending on which option you select depends on the company’s capabilities.

There are some security IT consulting companies like Gemalto, Capgemini, IBM Security, Imperva etc. that also provide their security solutions for the cloud. For Example, Gemalto safe net data protection services like Authentication and access management, encryption, tokenization, key management, cloud and big data security, compliance and data privacy etc. All other companies are providing more on these same features.

## 6) Advanced Cryptography algorithms

Secure multi-party computation like homomorphic encryption, quantum encryption.

**Homomorphic Encryption-**In 2010 Craig Gentry, graduate student thought a new way to protect data which is called Fully Homomorphic encryption. It is a way of process/encrypt data without ever decrypting it. As data & computation moved to the cloud, fully homomorphic encryption would allow data to be processed without having given access to it. There is lot of research is going in this field on how we can implement practically.

**Quantum Encryption-**Privacy is one of society's most valued qualities. All the encryption algorithm depend on the math like random numbers, prime numbers etc. .Instead of math, this technology relies on the law of physics called Heisenberg uncertainty principle. The principle is like that you cannot know everything about the atom. Every encryption algorithm is based on the key. The law of quantum physics is helpful in knowing the key in the form of stream of photons or light particles. Quantum nature of atoms prevent data privacy. Photons have a property called spin which can be changed when it is passed through filter. Quantum cryptography to work in real world is not so easy as physicist only send quantum keys around 200km.If this technology is successful then we need to rebuild all our internet once again.

## 7) Anonymizing Server's Log Data

In order to prevent companies does not track cookies of the users then the companies need to anonymize the entire IP address or maybe removing the last octet of the IP address, or in other words, we put zeros into the last eight bits of a 32-bit IP address. For this you need to write your algorithm and the main idea is that by transforming the cookie identifiers with an HMAC (keyed hash function) — using a randomly generated, ephemeral key for each day of logs.

Most of solution that exist in the internet is from the user side like CyberGhostVPN, Privacy Badger, Ad Block, or browser like TOR browser. But all these tools users need to download and know some basic understanding what the function is and which threats they are preventing from. It's actually surprisingly easy. There are some very effective bits of software you can install to block trackers, encrypt your website connections,

or stop spying ads from running - all of which can make a big difference to your privacy.

## 8) Moving Old Data to the Cloud

### AWS Snowball

- Snowball is a petabyte-scale data transport solution that uses devices designed to be secure to transfer large amounts of data into and out of the AWS Cloud.
- Today customer use AWS snowball to migrate any type of data like images, video, text etc.
- Don't need to write any code or purchase any hardware to transfer your data.
- One-fifth the cost of transferring data via high-speed Internet.
- Features like Fast data transfers, Encryption, End-to-end tracking, Tamper-resistant etc.
- 10 days free usage
- Snowball 50TB- $200,Snowball 80TB- $320

I think from all the solutions that I described above the best solution for Dumnonia Company is to use ARX De-anonymizer tool and buy cloud services from AWS. First observe the results from 6-months and then take further decision. Take this as a pilot project and change decision according to that. If they face any problem in implementation of these software's, or any other issue then they take the help of consulting companies/experts that I have mentioned above. Obviously, consulting companies cost more but then you did not worry about implementation, issues because they take care all this problem as they are experienced and Dumnonia company need to focus only on insurance business.

## Implement Software

**Aim:**

- Implement high quality Big data system
- Design better customer products
- Work effectively in scheduled time and work efficiently in the current and planned information technology infrastructure
- Cost effective

Our Software implementation is done in four stages. They are:-

# 1-Feasibility & Planning

The purpose of this first phase is to find out the scope of the problem and determine solutions. I am focusing here on resources, feasibility, issues, cost, time and benefits.

**Resources:**

- Good IT Infrastructure
- Well trained Staff
- Social Media

**Cost:**

- **ARX-Data Anonymization Tool:** Free of cost
- **AWS CloudTrail:**
  - 90 days free of charge
  - Charges are defined on how much we are using AWS service
  - **Management events-** This provide insights into the management and performed operations on resources in your AWS account. First copy within each region is delivered free of charge. Additional copies of management events are charged $2.00 per 100,000 events.
  - **Data events-** Data events provide insights into the resource operations performed on or within the resource itself. Data events are charged at $0.10 per 100,000 events.
  - **Usage related charges-** Once a CloudTrail trail is set up, Typical Amazon charges are less than $3 per month for most accounts.
- **AWS Cloudwatch:**
  - Every month you will receive some features (10 metrics, 10 alarms, 5 GB of ingested log size, 5 GB of archived log size, 3 dashboards and 1 million API requests), each month at no additional charge.
  - Prices vary by region. After consumption of free services the Detailed Monitoring cost you $2.10- $2.50 per EC2 instance, prorated by the hour
- **AWS Snowball:** Snowball 50TB- $200,Snowball 80TB- $320
- Consulting fees, training fees, maintainence fees

Therefore, I say that the anticipated Return on Investment (ROI) for this project is very high. I say “anticipated” because this calculation is based on assumptions and those assumptions may or may not happen.

**Feasibility Study:**

**Technical Staff:** There is no problem of finding technical staff.

**Economic:** It is not a costly project as for anonymizing the dataset they are using free software, they only need to spend the money only buying the services from AWS, consulting fees and fees involved in training employees and users.
**Legal:** The Company is following all the regulation that are provided by Australian government.
**Operational:** There is high ROI, so I don't think so there is a problem with operational cost.
**Schedule:** I think it will take 6-8 months to see any changes.
**Benefits:**

- IT flexibility
- Cost reduction
- Competitive in the insurance market
- Get more business value from sensitive data – while enhancing privacy protection.
- Gain faster insights
- Overcome trust barriers
- Enable data driven innovation without compromising on privacy.
- Reduce privacy risks
- Simplify compliance
- Enable multi-cloud deployment

## 2-System design, analysis and requirements

The second phase is where we describe necessary specification, features, operations that will satisfy the functional requirements of the proposed system.

Define and document all the software requirements, business assessment & challenges, contingency plan in the Software requirement specification document (SRS) and get them approved from the senior members.

Once it gets approved, our next step is to make a team with Project manager, functional leads, project sponsor, technical leads and most importantly senior business unit managers with a vested interest in the success of the project.

Project manager had the following duties. They are:-

- Announce who is filling each role on the project
- Motivate the team and promote the project within the organization
- day-to-day execution

All the team member should brainstorm on these topics like:

- Identify key resources
- Vendor Selection
- Send out the functional leads to meet with the business process experts, consultants, field experts to analyse the processes to see if any further improvements can be made to the project or not.

After refining the business processes, it is time for implementing the software in the company.

## 3-Development & Implementation

The third phase is where the project is put into the production by moving the data and the components from old system and placing them in the new system.

High Level Design-Architecture of software
LLD-How each and every feature of the product work

**Project Kick-off-** Start identifying all business processes and how they were transformed to reside within the software, the technical team begins the work in programming and configuring the meet the business needs. Note all of the passes and fails of each step and also be sure to note any lack of functionality, software bugs or key usability issues.

It is time to check if the team is on track with its configuration of the software. Face to Face Training for all employees to demonstrate the process in the new software.

Conduct a thorough end-to-end testing like white-box testing, black-box testing of the software with every business process from every function of the company and gets approved from quality assurance engineer.

Time to begin the process of transitioning from one system to a new one. The organization needs to plan, prepare and execute the cutover, by creating a cutover plan to describe all cutover tasks to be completed before go-live. This plan is designed to minimize the ongoing risk of moving from one system to another.

Final User Training Developing a user support plan then allow the users to complete the task individually. Be sure to provide a job aid and data sheet for each process to be tested so each user knows where to click and what the expected results are.

## 4-Operations and maintenance

The last phase is when end users can fine tune the system, boost performance, add new capabilities or meet additional requirements.

Providing in-person support to customers and know the new system and allow users to interact. Use Google Forms to create an online questionnaire to ask the consumer for new data products. Allow to create questions under each topic to be answered on a Scale from 1 to 5 where 1 = Strongly Disagree and 5 = Strongly Agree.

- Final project status reports complete
- Completion and storage of project file

Maintenance should be done as per service level agreement from other companies.

**References:**


Madeleine Dean (2017), **the best 4 data anonymization software to use** Available at: https://windowsreport.com/data-anonymization-software/ (Accessed 16 September 2018)

Carlos L.Aguilar (2018), **10 Steps Your Software Implementation Should Have** Available at: https://www.lisoblog.com/10-steps-your-software-implementation-should-have/ (Accessed 16 September 2018)

Case Writers (2018), **Imperva**, Available at: **https://www.imperva.com/** (Accessed 16 September 2018)

Case Writers (2018), **Privitar**, Available at: https://www.privitar.com/ (Accessed 16 September 2018)

Case Writers (2018), **Gemalto**, Available at: https://www.gemalto.com/ (Accessed 16 September 2018)

Bill Howe/Coursea (2017), Communicating Data Science Results Available at: https://www.coursera.org/learn/data-results (Accessed 16 September 2018)

Case Writers (2018), Innovative architects**,** Available at: **The Seven Phases of the System-Development Life Cycle** Available at: https://www.innovativearchitects.com/KnowledgeCenter/basic-IT-systems/system-development-life-cycle.aspx (Accessed 16 September 2018)

Case Writers (2018), Technopedia**, Data Perturbation** Available at: https://www.techopedia.com/definition/25013/data-perturbation (Accessed 16 September 2018)

Case Writers (2018), Uni Learning, **Good and poor examples of executive summaries** Available at: https://unilearning.uow.edu.au/report/4bi1.html (Accessed 16 September 2018)

Case Writers (2018), **ARX**, Available at: https://arx.deidentifier.org/overview/related-software/ (Accessed 16 September 2018)

Case Writers (2018), **University of Texas, Dallas (Data Security and privacy lab)** Available at: http://cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php?go=home (Accessed 16 September 2018)

Case Writers (2018), **SECRETA**, Available at: http://users.uop.gr/~poulis/SECRETA/index.html (Accessed 16 September 2018)

Case Writers (2018), **Edureka, Cloud Security Tutorial | Cloud Security Fundamentals | AWS Training | Edureka**

Available at: https://www.youtube.com/watch?v=0lw4KU5wHsk (Accessed 16 September 2018)

Matthew warren (2018), Future learn, **SIT719 Security and Privacy Issues in Analytics,** Available at: https://www.futurelearn.com/your-programs/security-privacy-analytics/3 (Accessed 16 September 2018)

Case Writers (2018), Amazon, Available at: https://aws.amazon.com/ (Accessed 16 September 2018)

Case Writers (2018), Amazon, Available at: https://aws.amazon.com/snowball/ (Accessed 16 September 2018)

Case Writers (2018), Amazon, Available at: https://aws.amazon.com/cloudwatch/ (Accessed 16 September 2018)

Case Writers (2018), Amazon, Available at: https://aws.amazon.com/cloudtrail/ (Accessed 16 September 2018)