УК БАН '20

Устойчиви Стратегии за Архивиране

Автор: Никола Стайков

Ментор: Явор Папазов

9 март 2020 г.

Съдържание

1 Въведение **2**

2 Теоретична постановка **3**
2.1 Възстановяване 3
2.2 Съхранение 3

3 Цена на възстановяване **4**
3.1 Пълни архиви 4
3.2 Инкрементални архиви с работещ пълен архив 4
3.3 Очаквана цена на възстановяване 5
3.4 Симулация Монте Карло 6

4 Цена на съхранение **7**

5 Резултати **7**

6 Бъдещо развитие **8**

7 Благодарности **8**

Абстракт

Архивите представляват резервни копия на данни, които да бъдат възстановени в случай на злополука. Те са основното средство за защита срещу рансъмуер и други видове вируси, както намаляват рисковете от критични щети в случай на природни бедствия. На свой ред обаче могат да представляват съществен разход за големите компании поради огромното количество данни, които трябва да бъдат подсигурени. Това на свой ред поражда нуждата те да бъдат внимателно планирани. Настоящият проект разразглежда модел за архивиране на данни, състоящ се от пълни и инкрементални архиви, изчислявайки очакваната цена за възстановяване на данните и цената на съхранение. Процесът по възстановяване е пресъздаден и анализиран чрез визуализация на python и Монте Карло симулация. Моделът намира оптимална стратегия за архивиране при предварително въведени параметри, характеризиращи работата на конкретния клиент.

1 Въведение

Настоящият проект изгражда теоретичен математически модел с цел намирането на оптимална стратегия за архивиране. Моделът включва два вида архиви, пълни и инкрементални. Параметрите, които оптимизираме, са интервалите в дни между видовете архиви. Структурата от архиви между два пълни архива образува цикъл, който се повтаря неопределено. Той се дефинира изцяло от два интервала - между пълните и между инкременталните архиви. Получените резултати се базират на два основни компонента- възможността за провал при възстановяваване на данните (различна за видовете архиви) и цената за съхранение на данните. Ефектите на двата компонента се изчисляват самостоятелно като след това се обединяват посредством константа, показваща отношението между цената на данните и техния размер. Подобни модели са правени и в миналото [13], [12], но те разглеждат само цената за съхранение, при неконстантна скорост на генериране на данни. Там интервалът между два последователни архива не е константен. За пълното разбиране на проекта са нужни базови знания по статистика и вероятности, използвани са стандартни означения и дефиниции.

2 Теоретична постановка

Настоящият модел е създаден с цел да изчисли и оптимизира очакваната цена като сума на два компонента, цена на възстановяване и цена на съхранение. Двата за разгледани поотделно като очакваната цена за всеки от тях е изчислена при различни стратегии на архивиране. Моделът обединява двата компонента чрез безразмерна величина, отразяваща отношението между стойността на генерираните данни и цената за тяхното съхранение.

2.1 Възстановяване

Ще разглеждаме архивът като структура от данни със следните компоненти:

$$b \begin{cases} d\text{: датата на създаването на архива като разлика в дни от първия архив} \\ p\text{: вероятността възстановяването да е неуспешно} \\ r\text{: цената за опит за възстановяване} \end{cases}$$

Разгледани са два вида архиви:

1. Пълен архив: запазва копие на цялата база данни до момента на създаване
2. Инкрементален архив: запазва само промените спрямо последния архив

Архивите от конкретен вид имат обща вероятност за провал и цена за опит за възстановяване.

За да може един инкрементален архив да е успешен, трябва всички инкрементални архиви преди него до успешен пълен архив също да са успешни, както и самият пълен архив.

Важно е да се отбележи, че цената на данните в случая не съвпада с пазарната им такава. Това е така поради субективната им важност за компанията. Това е и причината в описания модел цената на данните да се разглежда като вечно увеличаваща се величина и за целите на изследването "скоростта на работа"на компанията се счита за константа. Ще я означим с w.

Цената на процеса по възстановяване ще разглеждаме като сума от два фактора:

1. Цената на изработването на загубените данни, означена с W
2. Цената на процеса по възстановяването, означена с R.

Дефинираме $W = \Delta t.w$, където с Δt означаваме разликата в дни между датата на създаване на последния успешно възстановен архив и датата на възстановяване, и $R = \sum_{i=1}^{n} r_i$, където броят опити за възстановяване е n. Нека разликата в дни между първия направен архив и датата на възстановяване е T. В случай, че никой от пълните архиви не се окаже успешен, въвеждаме променлива W_T, съответстваща на цената на преработване на всички съществуващи файлове. Ясно е, че $W_T > T.w$

2.2 Съхранение

В даден момент всеки архив генерира цена за съхранение в зависимост от размера си и колко дълго е съхраняван. Приемаме, че размерът на генерираните данни, S, е пропорционален на стойността на генерираните данни W, описана в предишния параграф. Тогава $S = Wk$, където k е константа. Цената за съхранение на количеството

данни, които се генерират за един ден за един ден означаваме със s. Така приносът на всеки архив към цената на съхранение може да се опише чрез формулата:

$$P_S = W\Delta ts,$$

където с Δt е означено времето, през което е бил съхраняван архивът.
Пълните и инкременталните архиви имат различен принос, тъй като размерът на едните е фиксиран, докато на другите постоянно расте.

3 Цена на възстановяване

В тази част е описана компонентата от модела, свързана с цената на процеса по възстановяването на данни и последващата преработка на загубените данни.

3.1 Пълни архиви

Когато разглеждаме само пълни архиви, моделът е разпределение на Бернули с краен брой опити, а именно броят пълни архиви. Спираме, когато успеем да намерим успешен архив, започвайки от последния и вървейки към първия. Да дефинираме свойствата на пълен архив(b_F):

$$b_F \begin{cases} p_F : \text{вероятността за провал} \\ r_F : \text{цената за опит за възстановяване} \\ t_F : \text{дните между два последователни пълни архива} \end{cases}$$

Нека k е броят пълни архиви направени преди деня на атаката. Тогава:

$$k = \left\lfloor \frac{T}{t_F} \right\rfloor + 1$$

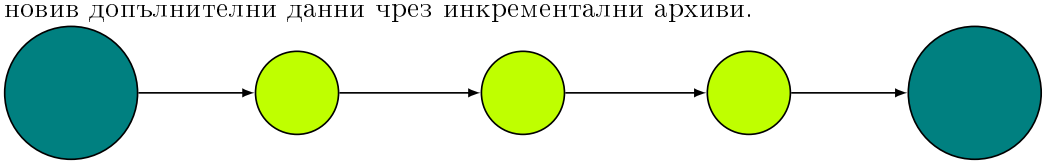
Сега можем да дефинираме очакваната цена на възстановяване:

$$E(T) = p_F^k \left(W_T + k.r_F\right) + \sum_{i=0}^{k-1} (1-p_F).p_F^i \left(\left(\left\{ \frac{T}{t_F} \right\} + i \right) t_F.w + (i+1).r_F \right) \qquad (1)$$

Направените изчисления са ключови поради невъзможността инкременталните архиви да бъдат възстановени без работещ пълен архив и следователно намирането на такъв е първият ни приоритет. Сега можем да разгледаме инкременталните архиви при работещ пълен архив.

3.2 Инкрементални архиви с работещ пълен архив

Ще разгледаме случая, когато имаме работещ пълен архив и се опитваме да възстановив допълнителни данни чрез инкрементални архиви.



Нека дефинираме свойствата на инкременталните архиви (B_I) по подобен начин:

$$b_I \begin{cases} p_I : \text{вероятността за провал} \\ r_I : \text{цената за опит за възстановяване} \\ t_I : \text{дните между два последователни инкрементални архива} \end{cases}$$

Нека с T_F да означим разликата в дни между денят на атаката и датата на успешния пълен архив и с l да означим броят инкрементални архиви, които трябва да разгледаме. Имаме две опции за l в зависимост от това дали последният пълен архив е бил успешно възстановен:

$$l = \begin{cases} \left\lfloor \frac{T_F}{t_I} \right\rfloor, \text{ако } T_F < t_F \\ \left\lfloor \frac{t_F}{t_I} \right\rfloor - 1, \text{ако } T_F > t_F \text{[1]} \end{cases}$$

Да отбележим, че това последният пълен архив да е успешен е еквивалентно на $T_F < t_F$.

Сега сме в точно обратната ситуация спрямо миналата част. Процесът по възстановяване на инкрементални архиви продължава докато не се натъкнем на провал, тъй като това би означавало, че всички следващи инкрементални архиви също са неизползваеми. С това намаляме W, тъй като в началната позиция сме готови да преработим данните до датата на успешния пълен архив. Сега сме готови да изчислим очакваната цена:

$$f(T_F) = (1-p_I)^l.((T_F - t_I.l).w + r_I.l) + \sum_{i=0}^{l-1}(1-p_I)^i.p_I((T_F - t_I.i)w + r_I.(i+1)) \quad (2)$$

Сега знаем колко ще намалее цената на възстановяването, когато използваме инкрементални архиви, и можем да построим цялостния модел, използвайки уравнения 1 и 2.

3.3 Очаквана цена на възстановяване

Към всяко събираемо в уравнение 1 трябва да добавим ефектът на инкременталните архиви и получаваме нови събираеми от вида:

$$P(W + R),$$

където P е вероятността определена комбинация от събития да се случи, W е цената на данните, които трябва да бъдат създадени наново, а R е цената на процесът по възстановяването. Инкременталните архиви намаляват цената на данните, които трябва да бъдат създадени наново, но увеличават R. Както споменахме по-горе, има само един случай, вкойто броят инкрементални архиви, които трябва да имаме предвид е различен и той е именно този, в който последният пълен архив е възстановен успешно. Ако i-тият пълен архив е успешен[2]:

$$T_F = t_F\left(\left\{\frac{T}{t_F}\right\} + i - 1\right)$$

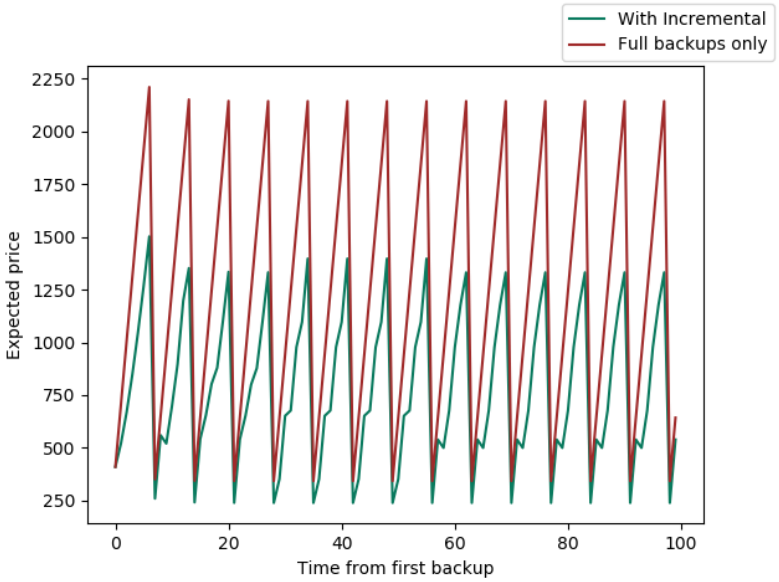
Комбинирайки уравнения 1 и 2 получаваме:

$$F(T) = p_F^k(W_T + k.r_F) + \sum_{i=0}^{k-1}(1-p_F).p_F^i\left(f(T_F) + (i+1).r_F\right) \quad (3)$$

Използвайки уравнения 1 и 3, можем да построим графика на очакваната цена с и без използването на инкрементални архиви.

[1] Можем да опитваме да възстановим само инкрементални архиви предхождащи следващият пълен архив

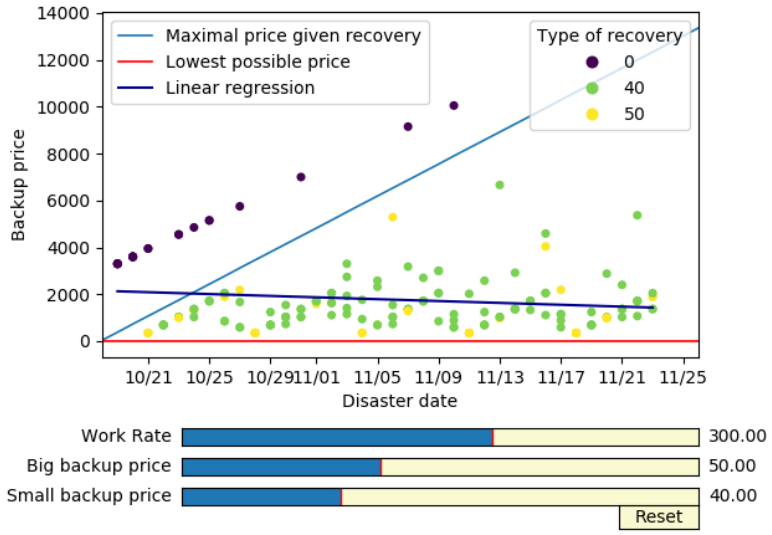
[2] Това съответства на $i-1$-вото събираемо в сумата от уравнение 3



Фигура 1: Full only and Whole model

3.4 Симулация Монте Карло

Направена бе симулация от тип Монте Карло на Python, която генерира случайни процеси на възстановяване на данни с описаната структура на архивите. Цената на възстановяването беше направена на графика спрямо датата на атаката:



Фигура 2: Симулация Монте Карло

Цветовете във фигура 2 показват вида на последния успешно възстановен архив, пълен, инкрементален или несъществуващ.

Във фигура 1 и фигура 2 данните са за седмичен пълен архив и ежедневен инкрементален

Линейна регресия на данните беше генерирана, която показва как ефектът от начално неподсигурените данни намалява с времето, тъй като цената при провал се изчилсява като цената за преработването на всички данни, които компанията е генерирала.

4 Цена на съхранение

Цената на съхранение се генерира от вече създадените архиви като за всеки индивидуално зависи от вида му и датата на създаване. За всяка от стратегиите изчисляваме цената за съхранение на данни от t дни, поради съображения, че размерът на важните данни, които бихме съхранявали по-дълго, е незначителен. Цената за съхранение на определено количество данни за един ден ще означим със s. За целите на модела приемаме, че размерът на данните S е пропорционален на цената за създаването им. За да отразим това взаимоотношение въвеждаме константата $c = \frac{S}{W}$.

Нека в даден момент от създаването на първият архив искаме да изчислим приносът на пълен архив, създаден d дни след направата на първия архив. Броят дни на генериране на данни преди първия архив ще означим с T_0. Тогава архивът има размер $S = cW = cw\Delta t = cw(T_0 + d)$, където w е скоростта на работа предишната компонента. Цената за съхранението му за ден е s и сме го съхранявали $T - d$ дни, поради което конкретният пълен архив е генерирал цена за съхранение:

$$cw(T_0 + d)s(T - d).$$

Инкременталните архиви от друга страна имат фиксиран размер cwt_I. Така приносът към цената за съхранение на инкременталните архиви се променя само от времето, през което са били съхранявани, а именно $T - d$. Окончателно приносът на инкрементален архив, направен d дни след първия архив е:

$$cwt_I s(T - d).$$

Ако с B_I означим множеството на инкременталните архиви, а с B_F множеството на пълните архиви, то цената за съхранение е:

$$\left(\sum_{b \in B_F} cw(T_0 + d_b)s(T - d_b)\right) + \left(\sum_{b \in B_I} cwt_I s(T - d_b)\right). \tag{4}$$

За $b \in B_F$, d_b приема стойности:

$$lt_F |\ l \in \left[0, \left\lfloor \frac{T}{t_F} \right\rfloor t_F\right],$$

а при $b \in B_I$:

$$d_b = mt_F + nt_I |\ m \in \left[0, \left\lfloor \frac{T}{t_F} \right\rfloor\right], n \in \left[1, \left\lfloor \frac{t_F}{t_I} \right\rfloor\right].$$

Замествайки в 4 получаваме

$$t_F \tag{5}$$

5 Резултати

Беше построен модел за изчисляване на очакваната цена при възстановяване на данни. Нещо повече, ефектът от инкременталните архиви беше показан в сравнение със стратегия използваща само пълни архиви. Беше направена и анализирана симулация от тип Монте Карло, коятодемонстрира реалния процес на възстановяване.

6 Бъдещо развитие

Авторът разглежда няколко посоки за бъдещото развитие на проакта, а именно:

- разглеждане на неконстантна скорост на работа за модела за архивиране
- разширяване на модела за откуп в посока описване на по сложни начини за разпространение на рансъмуера.
- използване на резултатите и базите данни на подобни проучвания с цел подкрепянето на модела с реални данни.[4]
- използване на динамичен модел за оценка на откупа

7 Благодарности

Искам да благодаря на своя ментор, Явор Папазов, и на Константин Делчев за безотказната помощ в избора на темата на проекта и последващото му развитие, за снабдяването ми с всички нужни материали за запознаването ми с темата, както и за изслушването на въпросите ми. Искам също да благодаря на Станислав Харизанов за професионалните съвети.

Литература

[1] Danny Yuxing Huang, Maxwell Matthaios Aliapoulios, Vector Guo Li, Luca Invernizzi, Elie Bursztein, Kylie McRoberts, Jonathan Levin, Kirill Levchenko, Alex C Snoeren, and Damon McCoy. Tracking ransomware end-to-end. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 618–631. IEEE, 2018.

[2] Tristan Caulfield, Christos Ioannidis, and David Pym. Dynamic pricing for ransomware.

[3] A Cartwright, Julio Hernandez-Castro, and Anna Stepanova. To pay or not: Game theoretic models of ransomware. In *Workshop on the Economics of Information Security (WEIS), Innsbruck, Austria*, 2018.

[4] Masarah Paquet-Clouston, Bernhard Haslhofer, and Benoit Dupont. Ransomware payments in the bitcoin ecosystem. *Journal of Cybersecurity*, 5(1):tyz003, 2019.

[5] Kurt Thomas, Danny Huang, David Wang, Elie Bursztein, Chris Grier, Thomas J Holt, Christopher Kruegel, Damon McCoy, Stefan Savage, and Giovanni Vigna. Framing dependencies introduced by underground commoditization. 2015.

[6] Amin Kharraz, William Robertson, Davide Balzarotti, Leyla Bilge, and Engin Kirda. Cutting the gordian knot: A look under the hood of ransomware attacks. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 3–24. Springer, 2015.

[7] J Michael Harrison, N Bora Keskin, and Assaf Zeevi. Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. *Management Science*, 58(3):570–586, 2012.

[8] Julio Hernandez-Castro, Edward Cartwright, and Anna Stepanova. Economic analysis of ransomware. *Available at SSRN 2937641*, 2017.

[9] Aron Laszka, Sadegh Farhang, and Jens Grossklags. On the economics of ransomware. In *International Conference on Decision and Game Theory for Security*, pages 397–417. Springer, 2017.

[10] Miguel Sousa Lobo and Stephen Boyd. Pricing and learning with uncertain demand. In *INFORMS Revenue Management Conference*, 2003.

[11] Michael Rothschild. A two-armed bandit theory of market pricing. *Journal of Economic Theory*, 9(2):185–202, 1974.

[12] S Nakamura, C Qian, S Fukumoto, and T Nakagawa. Optimal backup policy for a database system with incremental and full backups. *Mathematical and computer modelling*, 38(11-13):1373–1379, 2003.

[13] Cunhua Qian, Yingyan Huang, Xufeng Zhao, and Toshio Nakagawa. Optimal backup interval for a database system with full and periodic incremental backup. *JCP*, 5(4):557–564, 2010.