Young Talents 2020

Resilient Backup Strategies

Author: Nikola Staykov

Supervisor: Yavor Papazov

April 6, 2020

Contents

Abstract

Backups constitute copies of data, which are to be recovered in case of need. They represent the most efficient means of precaution against ransomware attacks or natural disasters. However, they can be the source of significant costs, especially when it comes to big organizations, which keep enormous amounts of data. Therefore, backups should be carefully planned. The current project considers a theoretical model, which finds the optimal backup strategy, consisting of full and incremental backups, considering the cost of the data recovery and storage processes. The recovery process is recreated and analyzed via Python and a Monte Carlo simulation. The model finds an optimal backup strategy given parameters, which characterize the clients' work pattern.

1 Introduction

The current project builds a theoretical mathematical model in order to find an optimal backup strategy. The model considers two types of archives, full and incremental. The parameters, which we optimize, are the intervals in days between the types of archives. The structure of the archives between two full archives forms a cycle, which is then repeated indefinitely. It is defined entirely by two intervals- between two full and between two incremental backups. The results are based on two main factors, the probability of recovery failure and storage cost. The effects are calculated separately and then combined via a constant, showing the relation between the size and value of the data. Similar models have been built in the past [1], [2], but they consider only the storage price with non-constant generation speed. There the interval between two consecutive archives is non-constant. The possibility for correlation between the lost and recovered data is also considered. For the full comprehension of the project basic knowledge of probability and statistics is needed. The used definitions and notations are standard.

2 Theoretical setting

The idea behind the described model is to calculate and optimize the expected price as a sum of two components, recovery price and storage price. The two are considered separately and the expected price for each is calculated according to the backup strategy. The model combines the two via a constant, reflecting the connection between the size and value of the data. As this varies from company to company, it allows for the results to fit the user's needs.

2.1 Recovery

We will consider a backup as a structure, containing the following properties:

$$B \begin{cases} d\text{: the date on which the backup was made, as a day difference from a starting point} \\ p\text{: the probability that the recovery is unsuccessful for any reason} \\ r\text{: the price of trying to recover the data from the given backup} \end{cases}$$

Two types of backup will be considered:

1. Full backup: a backup of the whole database
2. Incremental backup: only saves the changes from the last backup

The backups from a certain type share common probability of failure and price for a recovery try.

In order for an incremental backup to be successful, all the incremental backups which precede it up to a full backup need to be successful as well as the full backup itself.

In this case data value should clearly be taken into account from a subjective point of view. Even though on the market some data may not be worth a lot, if it is essential for the functioning of a given company, it is clear that it will be willing to pay a lot to regain access to it immediately. Therefore, in the described model data value is considered as an ever-increasing amount, for the purposes of the research the "work rate", namely the data value generated in a day, of the company is taken as a constant. We will denote it with w.

The cost of a backup recovery will be considered as a sum of two factors:

- The cost of redoing the lost work, denoted with W
- The cost of the recovery process itself, denoted with R

. We define $W = \Delta t.w$, where with Δt we denote the difference in days between the successful backup and the disaster date and $R = \sum_{i=1}^{n} r_i$, where the number of attempted backups is n and

$$S = \Delta t.w + R,$$

Let the difference in days from the first backup to the disaster date be T. In case none of the backups are successful, we consider a variable W_T, corresponding to the price of redoing the whole work the company has done from the beginning. It is clear that $W_T > T.w$

2.2 Storage

At a given moment each existing archive generates storage cost in relation to its size and the time it has existed. We assume that the size of the generated data, S, is proportional to the value of the generated data, W, described in the last paragraph. Then $Sc = W$, where c is a constant. The storage price for the data generated in one day, for one day, we denote with s. Then the contribution to the total price of each individual archive can be described as follows:

$$P_S = S\Delta ts$$

where with Δt we denote the time, in which it has been kept.
Full and incremental backups have different contribution as the size of full backups is constantly growing and the size of incremental backups is fixed.

3 Recovery price

This section describes the component of the model, related to the recovery process and the following re-make of lost data.

3.1 Full backups only

When we only consider a set of full backups, the model is simply a Bernoulli distribution with finite trials, namely the number of full backups. We stop when we find a successful backup, starting from the latest and going to the last. Let us define the properties of a full backup(B_F):

$$B_F \begin{cases} p_F : \text{the probability of failure} \\ r_F : \text{the recovery trial cost} \\ t_F : \text{the days between two consecutive full backups} \end{cases}$$

Let k be the number of full backups made before the disaster date. Then:

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

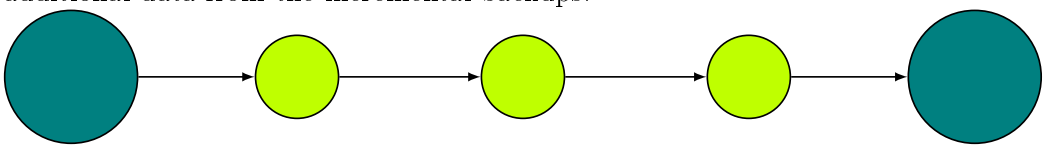
We can now define the expected backup cost:

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

This calculation is essential as incremental backups can only work when there is a working full backup and therefore the first thing we need to do is find the latest one. We can now move on to considering the incremental backups given a working full backup.

3.2 Incremental backups with a working full backup

We will now consider the case when we have a working backup and we are trying to recover additional data from the incremental backups.



Let us define the the properties of the incremental backup(B_I) in a similar fashion:

$$B_I \begin{cases} p_I : \text{the probability of failure} \\ r_I : \text{the recovery trial cost} \\ t_I : \text{the days between two consecutive incremental backups} \end{cases}$$

Let T_F denote the difference in days between the disaster date and the successful full backup and l denote the number of incremental backups we have to consider. We have two options for l depending on whether the latest full backup was successful:

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

Note that the last full backup being successful is equivalent to $T_F < t_F$.

We are in the exact opposite situation with respect to the previous subsection. The process of recovering incremental backups continues until we conduct an unsuccessful attempt to recover the data, as this will mean none of the following backups can be used either. Note that we are reducing W since in the initial position we are willing to redo the work up to the working full backup. That being said, we are ready to calculate the expected price:

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

Now we know how much the price will decrease when we use incremental backups and can build the whole picture using equations 1 and 2.

3.3 Overall expected price

For each summand in 1 we should add the effect of incremental backups, so we get new summands of the type:

$$P(W+R),$$

where P is the probability of a certain combination of events occurring, W is the cost of the data that has to be reworked and R is the cost of the recovery process. Incremental backups lower the cost of the data that has to be reworked but make R bigger. As mentioned before, there is only one case when the number of incremental backups we have to consider is different and it corresponds to the first full backup being successful. If the i-th full backup is successful[2]:

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

By combining equations 1 and 2 we get:

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

3.4 Correlation of data

As there can be a connection between the recovered and lost data, we introduce a function to show the relation between the work rate(w) and the recovery rate(r) as a function of the fraction of data lost x2:

$$\frac{r}{w} = g(x) = e^{1-x}$$

[1] We can only try to recover incremental backups preceding the next full backup

[2] This corresponds to the $i-1$-th summand in the sum from equation 3

The more data we have lost, the closer the recovery rate to the work rate. If the recovery takes place at time T and the closes successful backup was made on a date b_d, the lost data are from $\Delta t = T - b_d$ days and $x = \dfrac{\Delta t}{T + T_0}$, where T_0 is the number of days before the first backup.
If the last successfully recovered incremental backup is the $i-$th in its cycle, we have:

$$x = \frac{T_F - t_I i}{T + T_0}.$$

We can now substitute w with $r = wg(x)$ in 2, respectively in 3:

$$f(T_F) = (1-p_I)^l((T_F - t_I l)wg(x) + r_I l) + \sum_{i=0}^{l-1}(1-p_I)^i p_I((T_F - t_I i)wg(x) + r_I(i+1)). \quad (4)$$

3.5 Monte Carlo simulation

A Monte Carlo simulation has been build with Python to generate random recovery processes with the described conditions of backup structure. The price of the recovery has been graphed with respect to the disaster date3

4 Storage price

The storage price is generated from the existing archives as for each individually depends on its type and creation dates. For the purposes of the model we assume that the size of the data, S, is proportional to the price of their creation. We introduce the constant $c = \dfrac{S}{W}$.
At a given moment, T, we want to calculate the the contribution of a full backup, created d days after the first backup. The number of days of generating data before the first archive we will denote with T_0. Then the archive has size $S = \dfrac{W}{c} = \dfrac{w\Delta t}{c} = \dfrac{w}{c}(T_0 + d)$, where w is the work rate from the previous component. The storage price for a day is $s = \dfrac{w}{c}$ and we have kept it $T - d$ days. Therefor the backup has generated storage price:

$$\frac{w}{c}(T_0 + d)s(T - d).$$

Incremental backups, on the other hand, have fixed size cwt_I. Their contribution to the storage price is only change by the time they have been kept, namely $T - d$. Therefore the contribution of an incremental archive, created d days after the first backup is:

$$\frac{w}{c}t_I s(T - d)$$

Let B_I be the set of all incremental backups and B_F that of all full backups. The the storage price is:

$$\left(\sum_{b \in B_F} \frac{w^2}{c^2}(T_0 + d_b)(T - d_b)\right) + \left(\sum_{b \in B_I} \frac{w^2}{c^2} t_I (T - d_b)\right). \quad (5)$$

For $b \in B_F$, d_b takes values:

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

and for $b \in B_I$:

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

Substituting in 4 we get

$$\left(\sum_{l=0}^{\left\lfloor \frac{T}{t_F} \right\rfloor t_F} \frac{w^2}{c^2}(T_0 + lt_F)(T - lt_F) \right) + \left(\sum_{m=0}^{\left\lfloor \frac{T}{t_F} \right\rfloor} \sum_{n=1}^{\left\lfloor \frac{t_F}{t_I} \right\rfloor} \frac{w^2}{c^2} t_I (T - mt_F + nt_I) \right) \tag{6}$$

With that we ca find the storage price of the generated data. Changing the ratio between the storage size and the value of the generated data (c), we can find optimal backup strategies with respect to the intervals between consecutive full and incremental backups.

5 Results

The results from the model are systemized in the appendix. The two cases for data correlation are compared [8.2, 8.2]. The respective best strategies are extracted. The backups are stored for a given period of time, which depends on the type of data. It is known, however, that the mean storage interval is from 1 to 3 months [3], [4]. This is why the strategies are evaluated considering their mean score for values of T in this interval. As the model aims to find the best strategy, we may assume that w is unity and to measure the result in days of work. This is to say that the price drawn from the model is actually the price of the data we would have created for the respective number of days. The relation between the scores of the best strategies, respectively in the two cases, is also shown for different values of c 5. It is important to note that these strategies are not the same but for a given c the optimal strategy when the data is correlated always performs better that that without data correlation.

A model was built to estimate the expected price of recovering data and its storage with and without correlation of the stored data. A Monte Carlo simulation was created and analyzed in order to demonstrate the actual recovery process.

6 Further development

The author considers several future development directions for the project, namely:

- non-constant work rate for the
- complex connection between data value and storage price
- Data correlation for easier recovery from partial leftover files

7 Acknowledgments

I want to thank my mentor, Yavor Papazov, and Konstantin Delchev for the enormous help with the choice of the research subject and for providing me with all the necessary material to get familiar with the topic, as well as listening to my questions along the whole way. I extend my gratitude towards HSSIMI and SRS for the opportunity to develop this project and the irreplaceable atmosphere of dedication and concentration.

References

[1] Cunhua Qian, Yingyan Huang, Xufeng Zhao, and Toshio Nakagawa. Optimal Backup Interval for a Database System with Full and Periodic Incremental Backup. *Journal of Computers*, 5(4), apr 2010.

[2] 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, dec 2003.

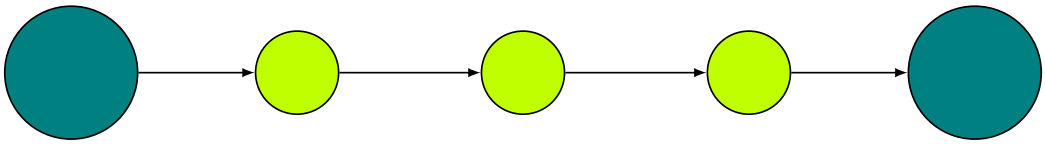
[3] Mikhail Gloukhovtsev. Technical report, 2014.

[4] J Schepers and P Huisken. Backup and Restore Backup alternatives for Network Appliance filers. Technical report, 2001.

8 Appendix

8.1 Plots and graphics

The graphic below symbolizes the way we search through existing backups for the last available to recover.



Using the described equations 1 and 3, we can construct a graph of the expected price with and without incremental backups included.

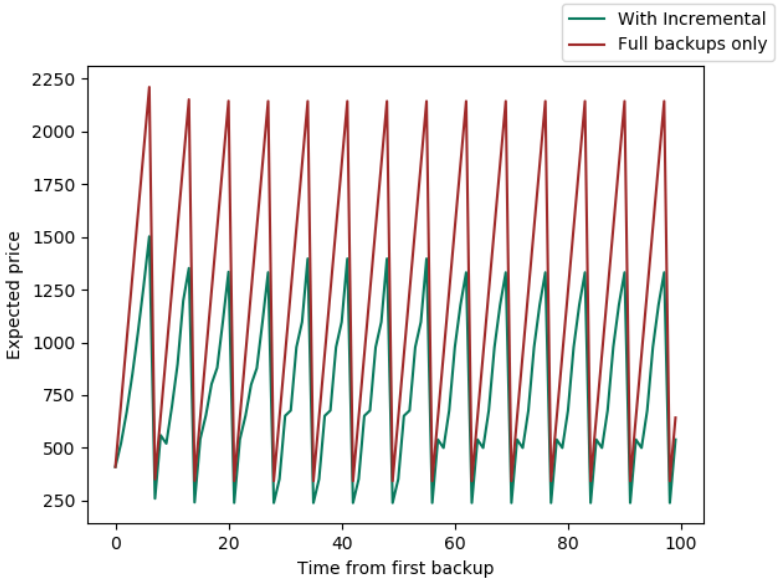


Figure 1: Full only and Whole model

The next plot shows the relation between the work rate and recovery rate of data with respect to the amount of data lost:

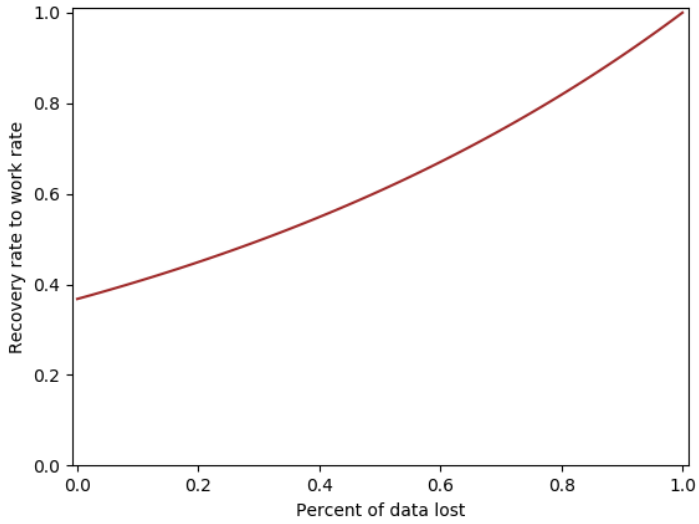


Figure 2: Connection between w and r

The graphic below is generated from the Monte Carlo simulation and the stochastic recovery processes.

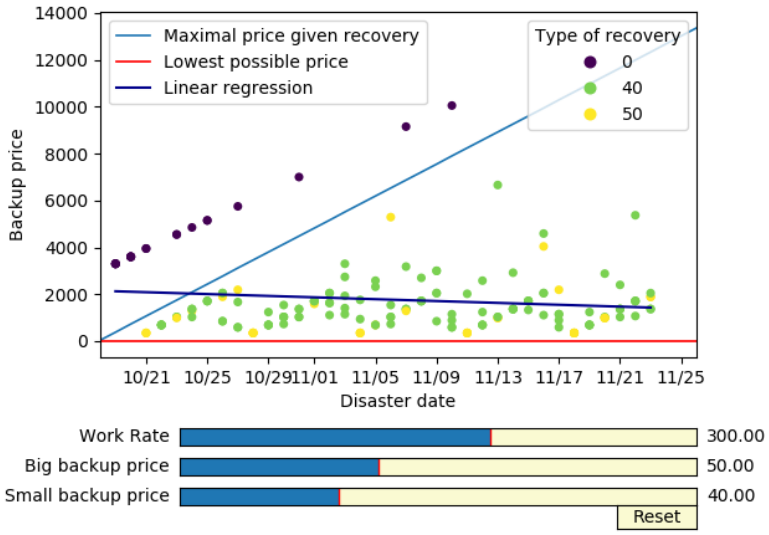


Figure 3: Monte Carlo simulation

The colors in Figure 3 represent the type of the last backup, which was successful during the recovery, full, incremental or non-existing. A linear regression has been made of the data generated, which is to show that the effect of initial unsecured data fades with time, as the price of failure is calculated as the price to redo the whole work from the creation of the company.

The plot below show the results of some of the possible strategies when $c = 1000$

In both Figure 1 and Figure 3 the data showed is for a weekly full and daily incremental backups

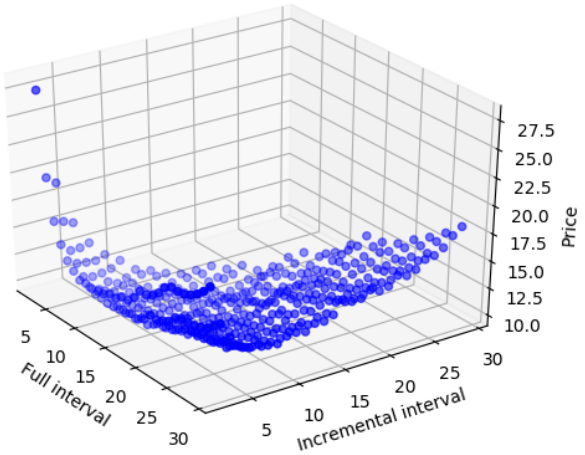


Figure 4: Visualization for $c = 1000$

The next plot shows the prices of the best strategies with and without data correlation for different values of c.

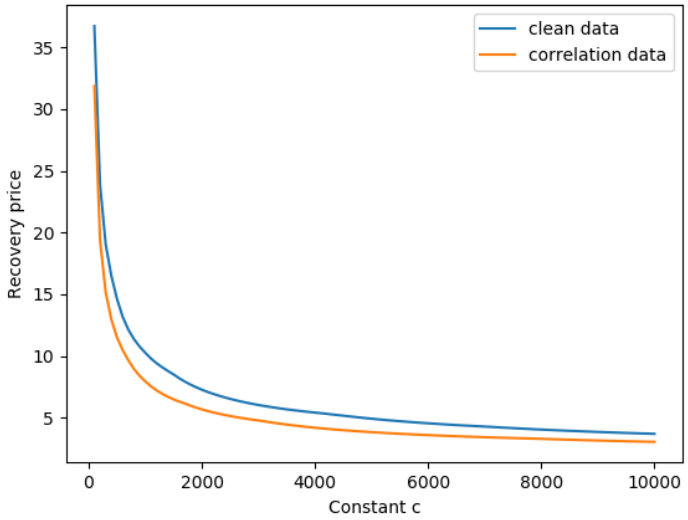


Figure 5: Comparison of the prices in the two cases

We notice that for small c the difference is big and then fades. This is due to the fact that when c is bigger the optimal strategies use more frequent backups and therefore the chance to lose a significant part of the data vanishes.

8.2 Results

The tables below summarize the results for the best strategies for different values of c in the two considered cases.

With data correlation:

Constant c	Full interval	Incremental interval	Price
100	29	15	31.86
200	29	16	19.40
300	27	14	15.16
400	27	14	12.97
500	26	9	11.5
600	21	12	10.49
700	16	9	9.66
900	16	9	8.39
1000	16	6	7.94
1600	16	6	6.31
1700	11	6	6.15
3000	11	6	4.78
3100	8	5	4.71
7899	8	5	3.31
8000	6	3	3.29
10000	6	3	3.05

Without data correlation:

Constant c	Full interval	Incremental interval	Price
100	29	16	36.72
200	29	10	23.9
300	26	9	19.04
400	21	8	16.49
500	17	6	14.62
600	16	6	13.18
900	16	6	10.76
1000	13	5	10.26
1300	13	5	9.08
1400	11	3	8.79
1500	8	3	8.51
4100	8	3	5.37
4200	6	3	5.32
4500	6	3	5.17
4600	5	2	5.11
6999	5	2	4.29
7100	4	2	4.26
10000	4	2	3.7