日本オペレーションズ・リサーチ学会
2019年 秋季研究発表会

データベース更新形態が特殊な場合のバックアップ最適方策

01014433 金城学院大学　　*中村　正治 NAKAMURA Syouji
01400043 愛知工業大学　　中川　覃夫 NAKAGAWA Toshio

1. はじめに

近年の各種のデータ量は爆発的に増えていることから，システムデータのバックアップは重要課題である．データ量の増加の予測は困難化している現状で格納に必要なストレージ容量を抑えて，より経済的にバックアップを行うことが一層重要な課題となっている．バックアップを効率的に実施するために同一内容のデータ部分を検出・削除する重複排除機能により，バックアップデータ量の大幅な削減を実現している．

データベースのバックアップ時期にはいくつかの計画された時間で行われる．データベースのバックアップはデータベース更新時に障害発生したとき，データベースの更新が設定された回数に達したとき，または，データベース更新時に障害発生したときと計画された時間がきたときにバックアップが実施されると仮定する [1],[2],[3]．再生過程を適用して，バックアップ，データベースリカバリーの期待費用を与えて，データベースバックアップ費用を最小とする最適バックアップ実施間隔を確率モデルにより解析的に導出する [3] し，最後に数値例を示し議論する．

ここでは，データベスの更新は分布 $F(t)$，密度関数 $f(t) \equiv dF(t)/dt$，平均値 $\mu = \int_0^\infty \overline{F}(t)\mathrm{d}t$ の再生過程とする．ただし，任意の $\Phi(t)$ に対して $\overline{\Phi}(t) \equiv 1 - \Phi(t)$．すなわち，$(0,t]$ 区間でデータベースの更新が j 回発生する確率は $F^{(j)}(t) - F^{(j+1)}(t)(j = 0,1,2,\ldots)$, ただし $F^{(j)}(t)$ は $F(t)$ の j 重畳み込みで $F^{(0)}(t) \equiv 1$ for $t \geq 0$.

さらに，k 回目の更新での総更新量は

$$S_0 + S_1 + S_2 + \cdots + S_{k-1} = \sum_{i=0}^{k-1} S_i \quad (k = 1,2,\ldots). \tag{1}$$

$c_1 + c_2 x$ をデータベースの更新量が x である場合の費用としたとき，k 回更新の期待費用は

$$C_k = c_1 + c_2 \sum_{i=0}^{k-1} S_i \quad (k = 1,2,\ldots). \tag{2}$$

ただし，$C_0 \equiv 0$，そして C_k は k に関して C_∞ まで単調増加である．

各更新に対して，失敗と成功の確率を $p(0 < p < 1)$ と $q(q \equiv 1 - p)$ であらわすとき，時刻 t の故障確率は

$$F_p(t) = \sum_{j=1}^{\infty} pq^{j-1} F^{(j)}(t) = 1 - \sum_{j=1}^{\infty} q^j [F^{(j)}(t) - F^{(j+1)}(t)].$$

2. 更新 N 回目でフルバックアップ

フルバックアップは $N(N = 1,2,\ldots)$ 回目の更新時または故障が起きたときのどちらか早い方で実施する．そのとき，フルバックアップまでのの平均時間は

$$Nq^N \int_0^\infty [F^{(N)}(t) - F^{(N+1)}(t)]\mathrm{d}t + \sum_{j=1}^{N} pq^{j-1} \int_0^\infty [F^{(j)}(t) - F^{(j+1)}(t)]\mathrm{d}t = \mu(Nq^N + \sum_{j=1}^{N} pq^{j-1}) = \mu \sum_{j=0}^{N-1} q^j. \tag{3}$$

c_N を N 回目の更新時におけるフルバックアップの付加費用とする．フルバックアップまでの総期待費用は

$$q^N \sum_{k=1}^{N} C_k + \sum_{j=1}^{N} pq^{j-1} \sum_{k=1}^{j} C_k + c_N = \sum_{k=1}^{N} C_k q^k + c_N. \tag{4}$$

したがって，単位時間当たりの期待費用は

$$C(N) = \frac{\sum_{k=1}^{N} C_k q^k + c_N}{\mu \sum_{k=0}^{N-1} q^k}. \tag{5}$$

$C(N)$ を最小とする最適な N^* を導出する．不等式 $C(N+1) - C(N) \geq 0$ から，

$$\sum_{k=1}^{N} (C_{N+1} - C_k) q^k \geq c_N, \tag{6}$$

左項は N に関して $\sum_{k=1}^{N} (C_\infty - C_k) q^k$ まで単調増加する．したがって，もし $\sum_{k=1}^{N} (C_\infty - C_k) q^k > c_N$ ならば，式 (6) を満たす有限でただ一つの最小となる $N^*(1 \leq N^* < \infty)$ が存在する．また，N^* は q に関して減少する．

3. T 時点でフルバックアップ

フルバックアップを時刻 $T(0 < T \leq \infty)$ または更新時に故障したときのどちらか早い方で実施する．そのとき，フルバックアップまでの平均時間は

$$\sum_{j=1}^{\infty} pq^{j-1} \int_0^T t\mathrm{d}F^{(j)}(t) + T \sum_{j=0}^{\infty} q^j [F^{(j)}(T) - F^{(j+1)}(T)] = \sum_{j=0}^{\infty} q^j \int_0^T [F^{(j)}(t) - F^{(j+1)}(t)]\mathrm{d}t. \tag{7}$$

c_T を T 時点のフルバックアップ費用とする．フルバックアップまでの総期待費用は

$$\sum_{j=1}^{N} pq^{j-1}F^{(j)}(T)\sum_{k=0}^{j-1}C_k pq^{j-1} + \sum_{j=1}^{\infty} q^j[F^{(j)}(T)-F^{(j+1)}(T)]\sum_{k=1}^{j}C_k + c_T = \sum_{k=1}^{\infty}C_k q^k F^{(k)}(T) + c_T. \quad (8)$$

したがって，単位時間当たりの期待費用は

$$C(T) = \frac{\displaystyle\sum_{k=1}^{\infty}C_k q^k F^{(k)}(T) + c_T}{\displaystyle\sum_{k=0}^{\infty} q^k \int_0^T [F^{(k)}(t)-F^{(k+1)}(t)]\mathrm{d}t}. \quad (9)$$

$C(T)$ を最小とする最適な T^* を求める．$C(T)$ を T に関して微分し 0 とおくと，

$$Q_1(T)\sum_{k=1}^{\infty} q^k \int_0^T [F^{(k)}(t)-F^{(k+1)}(t)]\mathrm{d}t - \sum_{k=1}^{\infty}C_k q^k F^{(k)}(T) = c_T, \quad (10)$$

ただし，$Q_1(T) \equiv \dfrac{\displaystyle\sum_{k=1}^{\infty}C_k q^k f^{(k)}(T)}{\displaystyle\sum_{k=0}^{\infty} q^k[F^{(k)}(T)-F^{(k+1)}(T)]}$.

もし，$Q_1(T)$ が T に関して $Q_1(\infty)$ まで単調増加し，$\displaystyle\sum_{k=0}^{\infty} q^k[\mu Q_1(\infty)-C_k] > c_T$ ならば，式 (10) を満たす有限でただ一つの $T^*(0<T^*<\infty)$ が存在する．

4. 数値例

4.1. $S_i = S/(i+1)$のとき

$S_i = S/(i+1) \quad (i=0,1,2,\ldots)$ のとき，

$$C_k = c_1 + c_2 S\sum_{i=1}^{k}\frac{1}{i} \quad (k=1,2,\ldots), \quad (11)$$

$C_0 \equiv 0$，C_k は k に関して c_1+c_2S から ∞ まで単調増加する．

Table I は $F(t)=1-\mathrm{e}^{-\lambda t}$ を仮定し，p と S に対して $\lambda=1.0, c_1=10.0, c_2=1.0, c_N=50.0$ のときの最適な N^* と $C(N^*)$ をあらわしている．更新量が増大するにつれて，更新回数が少ない時にバックアップした方が良いことを示している．しかし，更新量が増大するればバックアップの費用は増大する．また，更新時の失敗確率が増大すれば，なるべく更新回数を多くしてバックアップした方が良い．

Table I: 最適値 N^* と $C(N^*)$，$\lambda=1.0, c_1=10.0, c_2=1.0$ と $c_N=50.0$

p	$S=5.0$		$S=10.0$	
	N^*	$C(N^*)$	N^*	$C(N^*)$
0.01	13	26.132	7	36.964
0.02	14	26.305	7	37.164
0.03	14	26.487	8	37.352
0.04	15	26.680	8	37.551
0.05	15	26.887	8	37.762
0.06	16	27.105	8	37.987
0.07	17	27.339	8	38.225
0.08	18	27.590	9	38.463
0.09	19	27.859	9	38.711
0.10	20	28.149	9	38.975

4.2. $S_i = \alpha^i S \quad (0<\alpha<1)$のとき

$S_i = \alpha^i S(0<\alpha<1; i=0,1,2,\ldots)$ のとき，

$$C_k = c_1 + c_2 S\frac{1-\alpha^k}{1-\alpha}(k=1,2,\ldots),$$

ただし．$C_0 \equiv 0$，C_k は k に関して c_1+c_2S から $c_1+c_2S/(1-\alpha)$ まで単調増加する．

Table II は，$F(t)=1-\mathrm{e}^{-\lambda t}$ を仮定し，p と S に対して $\lambda=1.0, \alpha=0.9, c_1=10.0, c_2=1.0, c_N=50.0$ のときの最適な N^* と $C(N^*)$ をあらわしている．$S_i = S/(i+1)$ の場合と同様なことを示している．

Table II: 最適 N^* と $C(N^*)$，$\lambda=1.0, \alpha=0.9, c_1=10.0, c_2=1.0, c_N=50.0$

p	$S=5.0$		$S=10.0$	
	N^*	$C(N^*)$	N^*	$C(N^*)$
0.01	5	33.371	4	45.338
0.02	6	33.580	4	45.561
0.03	6	33.789	4	45.794
0.04	6	34.010	4	46.036
0.05	6	34.241	4	46.288
0.06	6	34.484	4	46.550
0.07	6	34.740	4	46.823
0.08	6	35.008	4	47.106
0.09	6	35.290	4	47.401
0.10	6	35.586	4	47.708

5. おわりに

フルバックアプを更新回数 N または故障の場合，時刻 T または故障の場合に実施する 2 つの場合について，バックアップにデータの更新時に重複がある場合を想定した確率モデルを構築した．そのとき，バックアップ費用が最小となる最適バックアップ実施間隔を解析的に導出した．最後に数値計算では，更新回数 N でバックアップする場合を示し種々議論した．

謝辞

本研究の一部は，文部科学省科学研究費基金（基盤研究 (C)）課題番号 (18K01713)(2018-2020) による補助を受けている．

参考文献

[1] 羽島正明 (2013) : これから始める「次世代ストレージ」ビグデータを受け止めろ　階層化と重複排除が主流に, 日立評論, 2010. 10 , pp.20-30.

[2] 中村正治, 趙旭峰, 中川覃夫 (2014) : 重複排除処理によるバックアップ効率化を考慮した最適方策, 日本オペレーションズ・リサーチ学会研究発表.

[3] 中村正治, 中川覃夫 (2018) : ビッグデータ保存サーバの最適データ配置, 日本オペレーションズ・リサーチ学会研究発表.