Identifying Trends in Enterprise Data Protection Systems

George Amvrosiadis, University of Toronto
Medha Bhadkamkar, Symantec Research Labs

June 10, 2015

Dear Applicant,

Due to unforeseen reasons, the U.S. Consulate in Toronto will not be able to process your visa application on June 24th. Please note, the Consulates are currently experiencing technical problems with visa systems. This issue is not specific to any particular country or visa category. We apologize for the inconvenience and are working urgently to

June 24, 2015

U.S. Visa-Processing Glitch Is Partially Fixed

World-wide, U.S. posts issuing visas handle 50,000 applications a day.

A computer problem that crippled U.S. visa processing around the globe for two weeks has been partially solved, the government said

"This has become a multimillion-dollar loss," [1]

June 25, 2015

In parallel, we are continuing to restore data from backups and overseas post databases. This process is ongoing. [2]

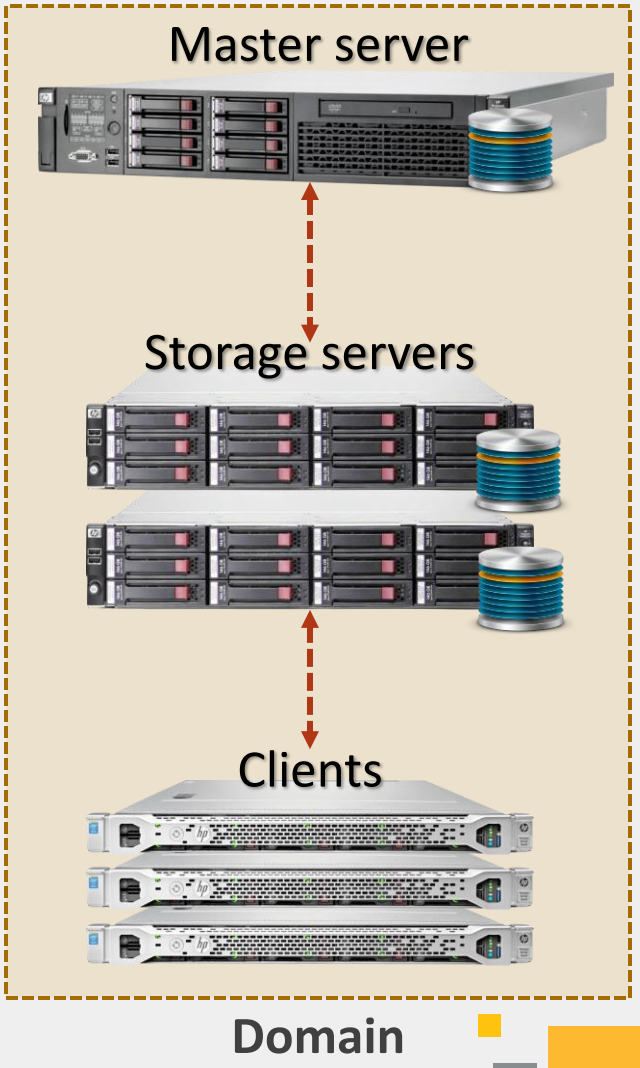
We need to fix Backup Systems

- Too many parameters to fine-tune
  - Top 3 commercial products come with 1000-page admin guides
  - Businesses experience problems recovering from backups 1 in 6 times [3]
- Too much data to backup
  - 94% of businesses backup more than just mission-critical data, and 40% backup everything [4]
  - Only 28% of businesses complete all backups on time [5]
- Complexity and missed goals lead to frustration
  - 55% of businesses plan to change backup tools within 24 months

Study goal: Use customer data to help researchers understand and improve data protection systems

Anatomy of Modern Data Protection Systems

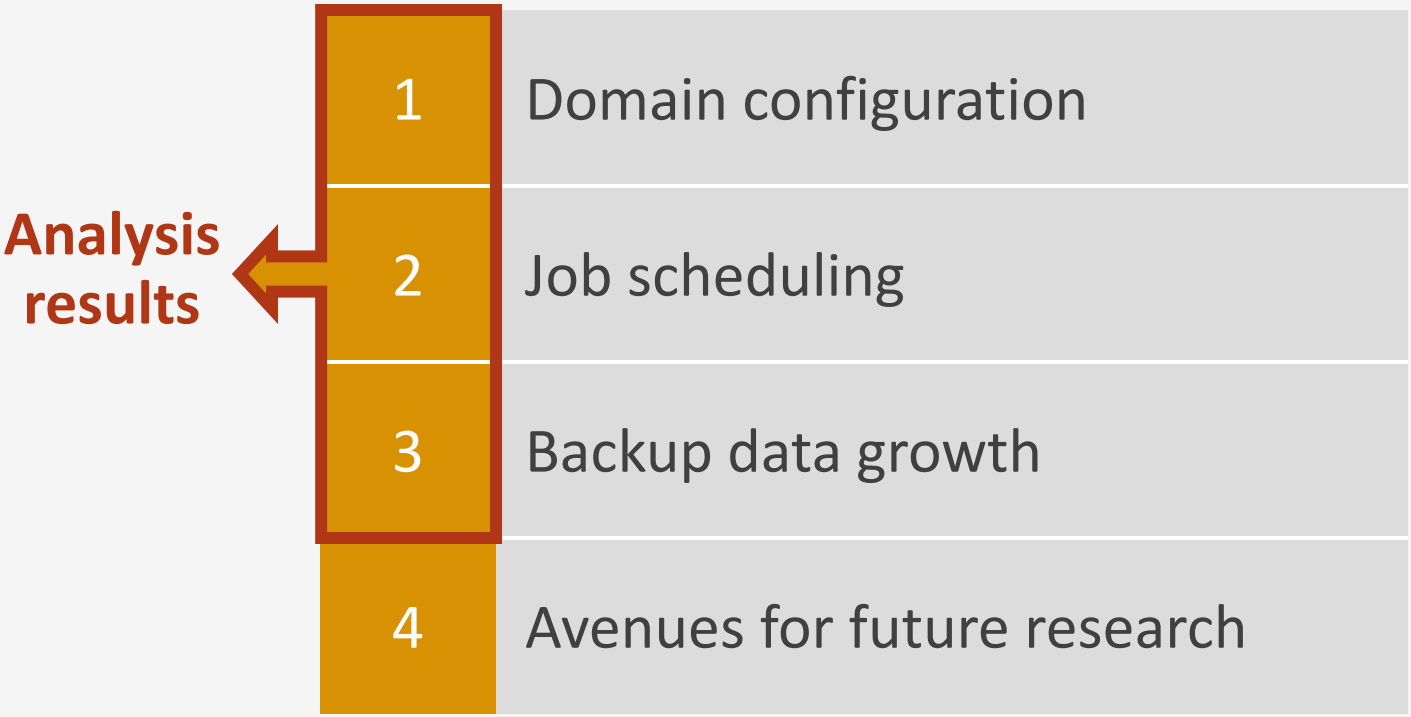
- Data protection systems are multi-tiered **domains**
  - **Master server:** job scheduling, backup image metadata
  - **Clients:** transmit backup data
  - **Storage servers (optional):** backup storage management
- **Backup policies** specify clients' backup schedules
  - E.g. "weekly full, daily incremental backups"
  - **Policy types** tailored to applications
    e.g. Oracle, VMware, Microsoft Exchange



Study Dataset

- Customer domains periodically transmit telemetry
  - Collected from consenting Symantec NetBackup customers
  - Weekly reports of runtime and configuration statistics
- Telemetry allows us to study how domains evolve
  - Reports can be grouped and analyzed as time series
- Dataset represents large, diverse domain population
  - 1M telemetry reports from 40,000 domains, collected over 3 years
  - 35% of domains 3-tiered, rest 2-tiered
  - 31% of domains use dedicated backup appliances

Outline



Outline

1 Domain configuration

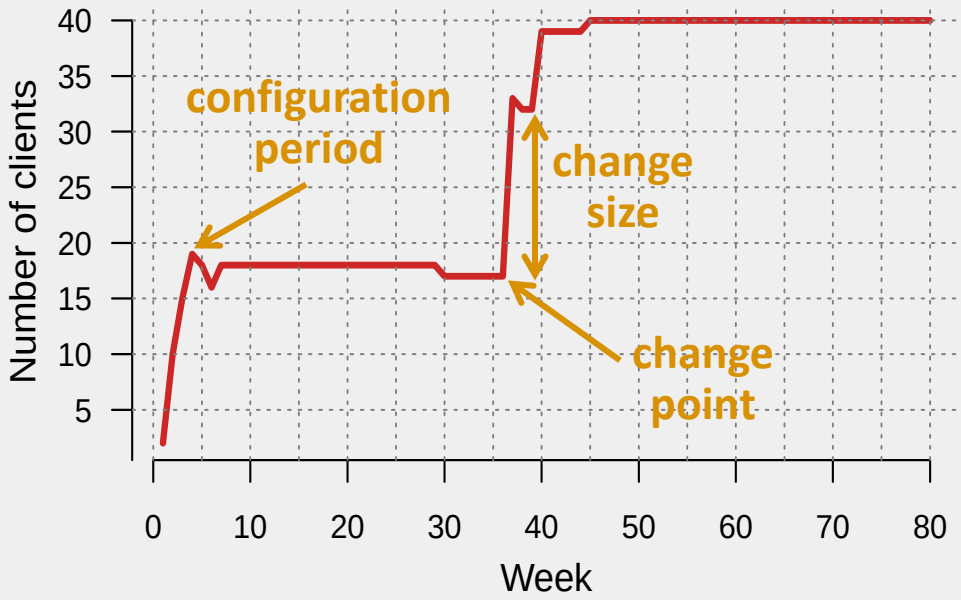
2 Job scheduling

3 Backup data growth

4 Avenues for future research

Domain configuration: Clients

- Client populations rarely shrink
  - Client population reaches stable state after first 3 weeks
  - 93% of changes attributed to growth

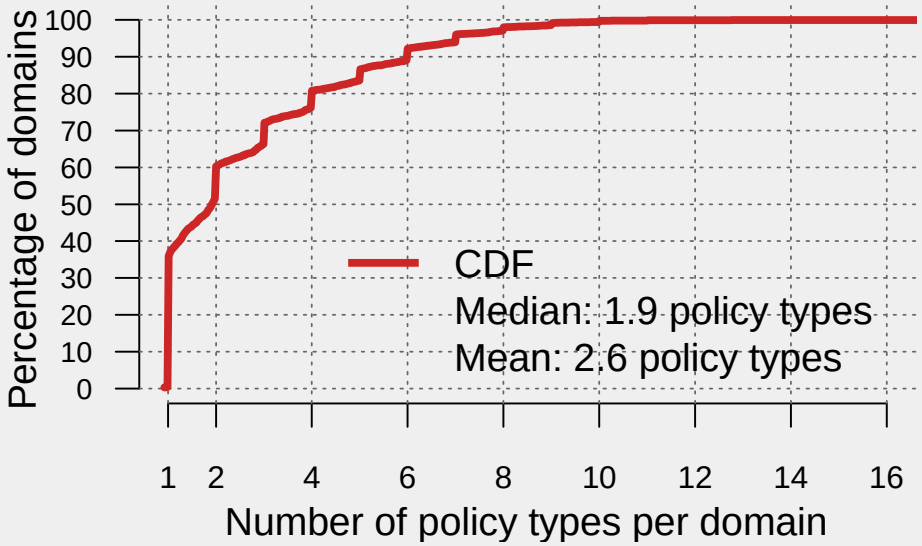


- Clients are introduced every 3 months, in groups
  - Low variability across changes, growth bursts 5% of the time

Takeaway: For resource provisioning, keep in mind that clients are added in bursts

Domain configuration: Backup policies

- Domain components remain unprotected at times
  - 16% of clients spend time unprotected
  - Only 32% of domains use a policy to protect master server state
- Domains typically protect fewer than 3 application types
  - 36% of domains deploy policies targeted to a single application
  - Number of policies stays fixed after first 3 weeks



Takeaway: Domains are homogeneous wrt. client policies, making policy auto-configuration a feasible goal

Outline

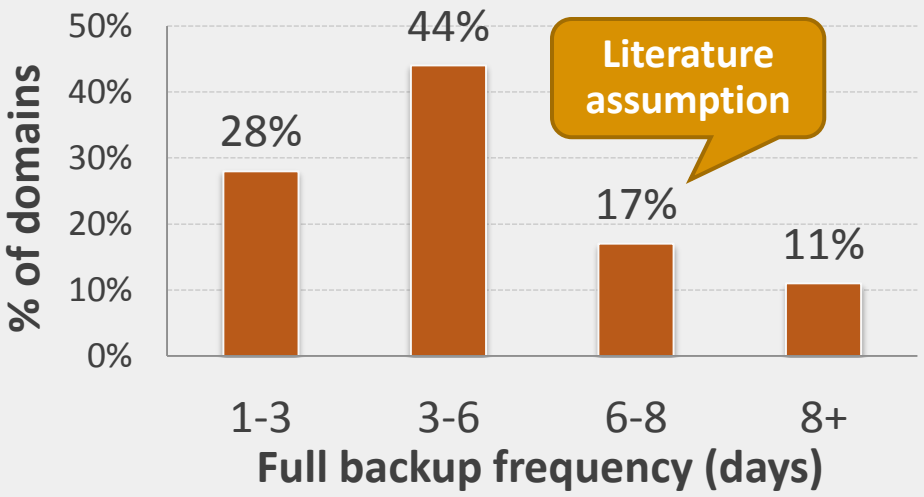
1	Domain configuration
2	Job scheduling
3	Backup data growth
4	Avenues for future research

Job scheduling: Frequency

- Recoveries are rare and sparse
  - Occur in 1275 domains (3.1%)
  - 337 domains (0.8%) recover frequently as part of testing

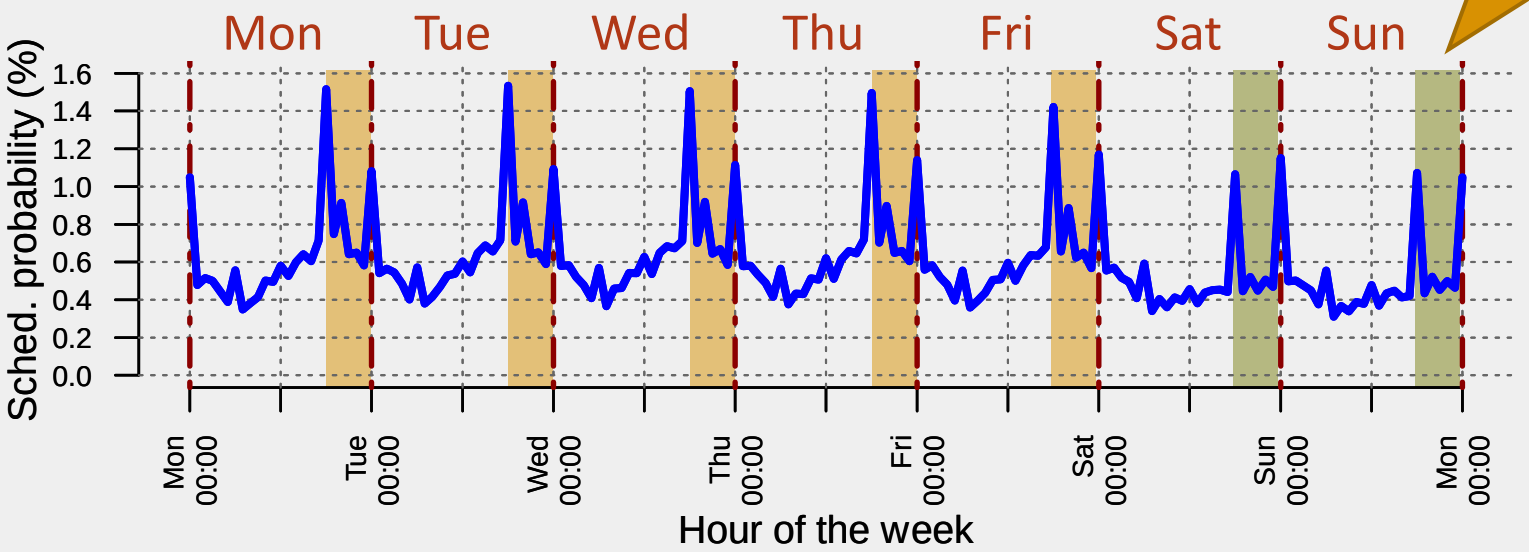
Domains	Recovery events	Avg. event frequency
938	< 5	2 months
337	≥ 5	2 weeks

- Frequent full backups are preferred to incremental ones
  - Full backups are rarely weekly events
  - Only 33% of frequent full backups are complemented by incremental ones



Takeaway: Recoverability of images is rarely tested, and frequent full backups are preferred to incremental ones

Job scheduling: Timing



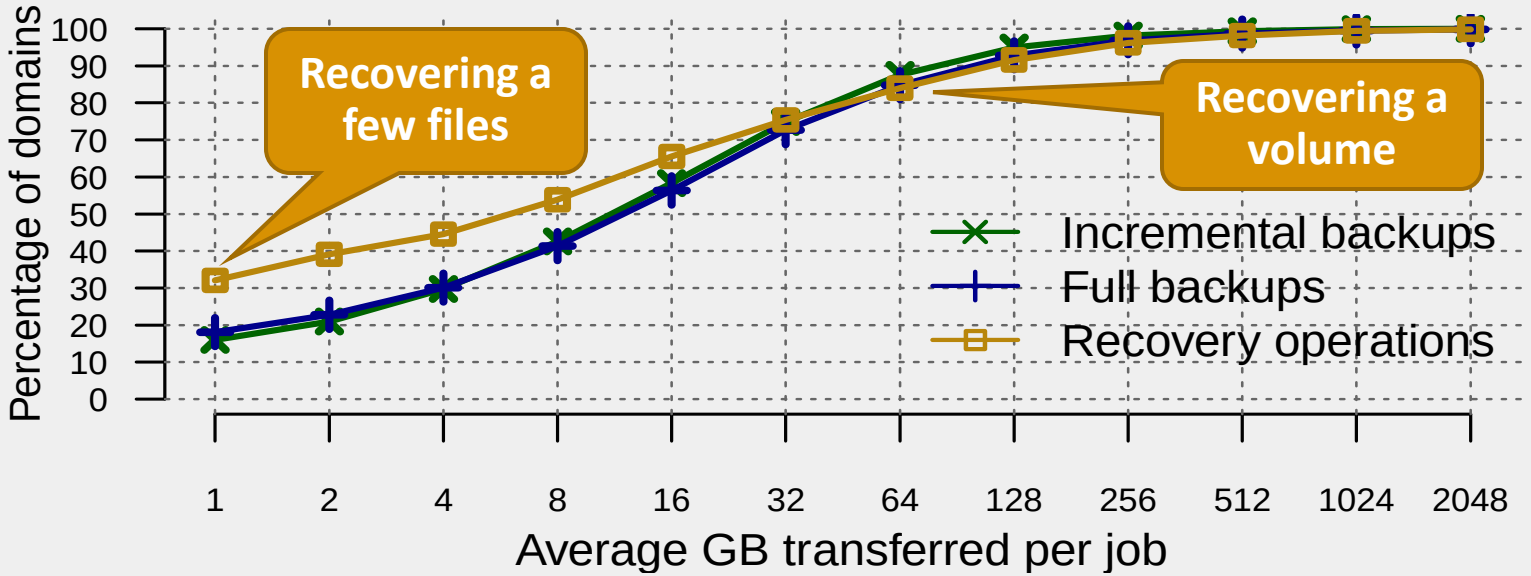
- Default scheduling windows are popular
  - Activity spikes at beginning of scheduling windows (6pm, 12am)
  - Administrators schedule fewer jobs during the weekend

Takeaway: Consistently using the same/default scheduling window creates job bursts

Outline

1	Domain configuration
2	Job scheduling
3	Backup data growth
4	Avenues for future research

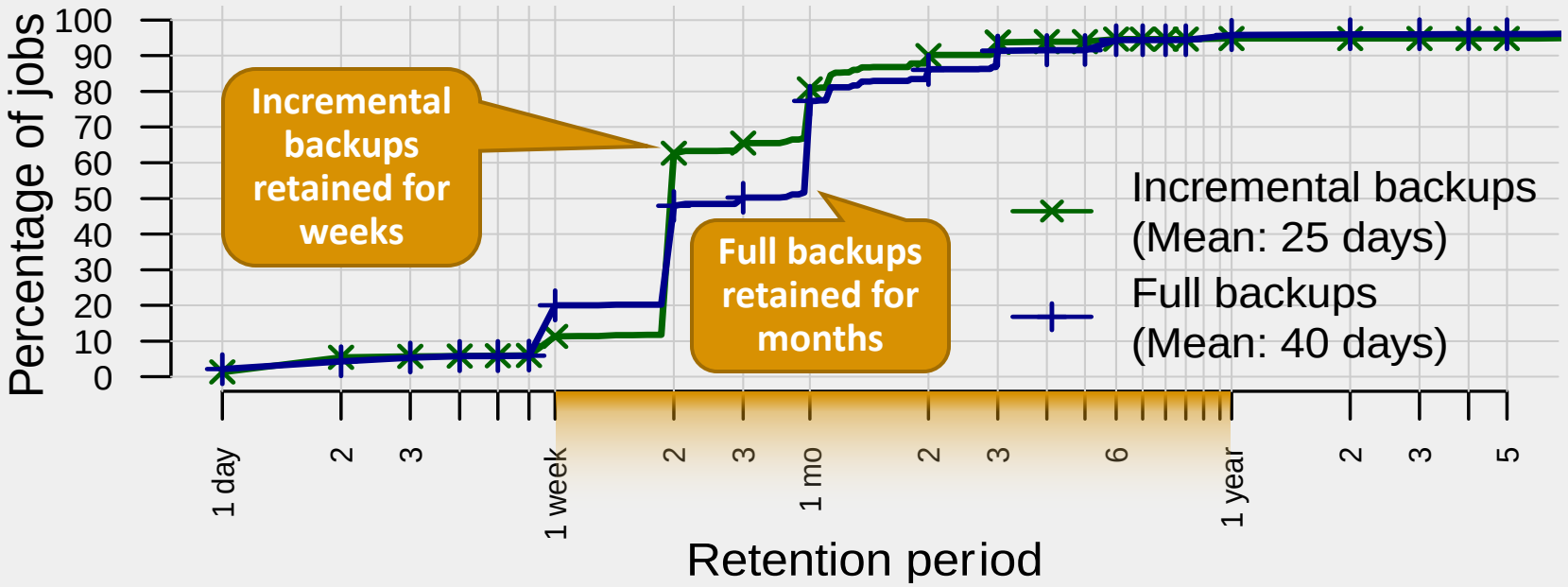
Backup data growth: Job sizes



- Incremental backups resemble full backups in size
  - Deduplication reduces full backup sizes by 89% on average
  - Incremental backups can be larger due to misconfigurations, timestamp modifications by maintenance tasks

Takeaway: Deduplication may obsolete incremental backups. Recovering only a few files is not uncommon.

Backup data growth: Retention periods



- 94% of retention periods picked from preset values
- Retention time is proportional to backup frequency
  - Less frequent full backups are retained longer ($\rho_{x,y} = 0.53$)

Takeaway: Retention periods are selected with backup storage capacity in mind

Outline

1	Domain configuration
2	Job scheduling
3	Backup data growth
4	Avenues for future research

In summary: Avenues for future research

- Auto-configuration and self-healing backup systems
  - Clients are introduced in bursts, but may be left unprotected
  - Domains are homogeneous wrt. policies protecting clients
  - Default scheduling windows are preferred, causing job bursts
- Improve rehydration time of deduplicated backup images
  - Deduplicated full backups are preferred to incremental ones
- Revisit backup retention as a need-based feature
  - Dedicated backup appliances are widely used
  - Retention periods are picked with storage capacity in mind
- Re-examine techniques for instant recovery
  - Recovery events made up of few files are not uncommon

Thank you:

Symantec Research Labs,
Symantec Backup and Recovery group,
University of Toronto SysNet and CSL groups,
Fred Douglis and anonymous reviewers

References

[1] M. Jordan, “U.S. Visa-Processing Glitch Is Partially Fixed”. Wall Street Journal. Updated June 24, 2015.

[2] Bureau of Consular Affairs, “Technological Systems Issue”. U.S. Department of State. Updated June 25, 2015.

[3] Veeam Software, “Virtualization Data Protection Report 2013”. Analyst report, 2013.

[4] Iron Mountain, “Data Backup and Recovery Benchmark Report”. Analyst report, 2013.

[5] Dimensional Research, “The state of IT recovery for SMBs”. Analyst report, 2014.