White Paper

The On-site Backup Assessment

January 20, 2015

Prepared by:

Matt Russo
Director of Engineering
Data Protection Advisors, LLC

Table of Contents

Why Do An Assessment?

Assessments avoid mistakes. They determine the current status of backup: the good, the bad and the ugly. In medical terms, assessments avoid treating the symptom rather than the cause. To avoid mistakes, doctors conduct examinations and order x-rays. The assessment is an x-ray of your backup environment. Without an assessment precious time and money will likely be wasted.

Information about backup jobs (i.e., metadata) is often locked inside files generated by the backup application. Unfortunately, the metadata is not easily deciphered and takes a considerable amount of time to review.

The backup metadata contains a wealth of information about the jobs, policies, scheduling, and retention times. This information can be viewed via the backup application itself but generally means viewing it policy by policy, or performing searches of the jobs database. Backup applications do provide tools to export this information but time is required to review and manually manipulate the metadata into useful information. However when unlocked with automated tools, these files contain valuable information about backup performance. By collecting and consolidating information the metadata can be easily reviewed, analyzed and compared to an organization's policies to see how the backup environment is performing and whether it is meeting the desired service level. The assessment can also assist in determining problem areas within the environment to investigate.

Supported backup applications are: Symantec NetBackup, IBM TSM, and EMC Networker.

The Objectives

There are four objectives of a backup assessment:

1. To determine the current health of the backup environment.
2. To identify ways to improve performance.
3. To identify ways to lower cost.
4. To determine the impact of new technology.

The Process

To begin the assessment our engineer brings a virtual machine, VM, on a set of DVDs that will be deployed on-site. The VM runs a MS Windows 7 OS configured with all the applications and utilities needed to perform the assessment. The VM runs on a workstation or server with a 64-bit Intel based processor using VMware Player for Windows or Fusion for MAC OS. The VM provides a working environment that is isolated from the rest of the customer environment to process the backup metadata (i.e, data about backup jobs). The backup metadata remains on the VM and does not and cannot leave the customers environment. The VM can also be deployed on an ESX host if the customer prefers.

Once the VM is deployed the next step is to collect the metadata from the backup servers. The customer is provided with one of two scripts based on the OS of the backup application, a batch file for Windows or shell script for Linux/UNIX. These scripts reside on the DVD. The scripts use the backup application's CLI commands to export the backup metadata to a set of files. The customer has the option to execute these commands manually. The CLI commands and options can be viewed in the source script. Once the data has been exported the generated files must be copied to the VM for processing.

After the export data is placed on the VM (e.g. using a CD) our Systems Engineer is provided access to the VM. The majority of the import and analysis is automated but there is some manual intervention required. Backup applications are highly customizable and the options and policies selected alter the fields and their length in certain files. Manual intervention is required to align these fields to ensure the analysis tool captures all data properly. Our on-site Systems Engineer ensures the proper alignment before the analysis begins.

Working entirely within the VM, our engineer runs our tool, which imports the data from the files, analyzes the data, preforms the calculations and generates the reports. Our Systems Engineer reviews the results and reports. After the data and reports are reviewed, the reports are copied from the VM to the customer's chosen location, the virtual server is deleted and the DVD destroyed. This protects the customer's metadata and our intellectual property. The reports are reviewed with the customer and recommendations are made.

Minimum Workstation Requirements for Assessment VM

Memory: 10 GB
Disk: 40 GB
Current version of VMware Player or Fusion.

Duration of the Engagement

Normally, the backup assessment can be performed in 1-2 days. Factors that impact the duration are: the time to scan the DVD and obtain security approvals, the number of clients, and the processing power of the workstation used to run the Assessment VM.

Results

The backup assessment provides deep insight into the health of the backup environment. It provides the following reports:

- Summary of the size and composition of the environment
- Failure rate for all backup jobs
- Top reasons for failed jobs by client
- Backup job throughput
- Clients with long run times
- Clients with long queue times
- Clients with slow throughput
- Impact of job schedules
- Annual growth rate of data
- Daily change rate of data
- Amount of structured v. unstructured data
- Age and amount of data retained
- Average file size
- Backup policy by server
- Retention policy by server.

Example – Health of the Backup Environment

Figure 1 is a view of backup for all clients in the environment. It provides a visual depiction of how successfully backup jobs are completing and the major reasons for failure.

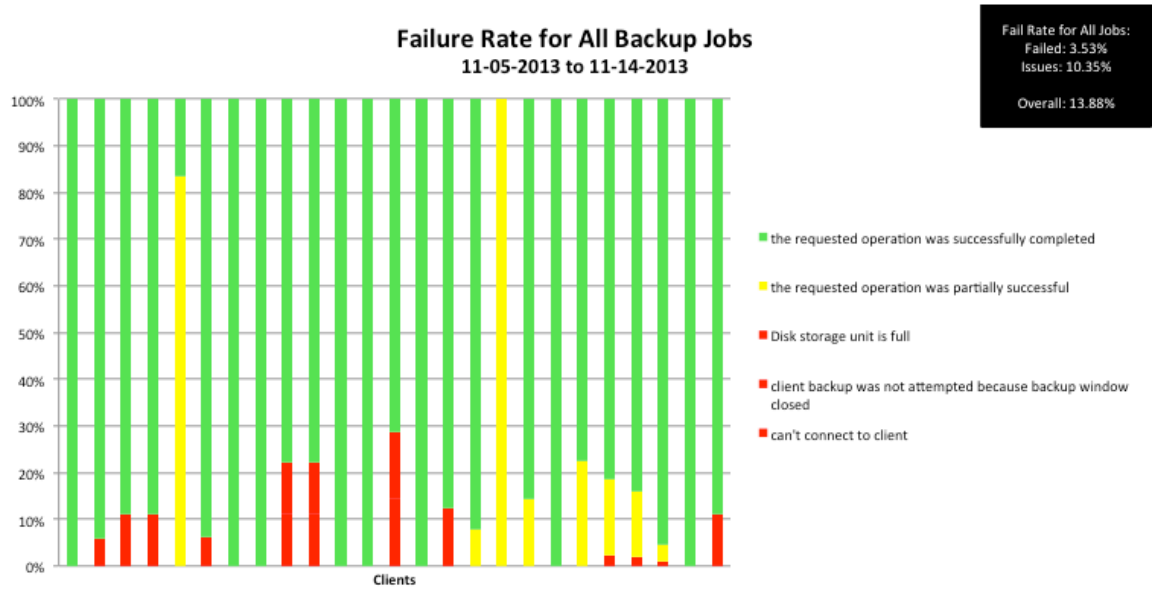

FIGURE 1 – FAILURE RATE FOR ENTIRE BACKUP ENVIRONMENT

The report in Figure 2 shows orphaned clients that have not been backed up in the past 18 months. These clients may have been retired or they may have been removed from a policy temporarily for testing or added to a different policy to rebalance workloads, modify retention, etc. or have been missed and overlooked. There may be valid reasons why they may no longer need to be backed up but it is invaluable to know which clients aren't being backed up. The worst time to determine a client is not being backed up is when the time comes to restore. The list of clients being backed up can be compared to queries against DNS, AD, LDAP, vCenter , etc. to ensure that all systems needing protection are protected.

Last Backup	Last Backup Size (GE	First Backup Da	First Backup Size (GBs)	Largest Size Backed up (GE	Smallest Size Backed up (GE	Avg Size Backed up (GBs)	Number Backu
None since 09/01/2012	0.0000	1/0/00 0:00	0	0	0	0	0
None since 09/01/2012	0.0000	1/0/00 0:00	0	0	0	0	0
None since 09/01/2012	0.0000	1/0/00 0:00	0	0	0	0	0
None since 09/01/2012	0.0000	1/0/00 0:00	0	0	0	0	0
11/14/13 8:42	91.2512	1/8/13 14:24	75	91	75	81	313
11/14/13 8:36	40.8174	1/8/13 14:25	32	42	10	36	311
11/14/13 8:21	8.3117	1/8/13 14:25	12	12	8	10	312
11/14/13 8:21	79.1808	1/4/13 14:24	77	84	72	77	315
11/14/13 8:08	33.5730	1/8/13 14:24	21	34	0	26	312
11/14/13 8:04	66.2097	1/8/13 14:24	64	68	61	65	311
11/14/13 8:02	24.5289	1/8/13 14:25	23	27	23	24	312

FIGURE 2 - CLIENT BACKUPS (CLIENT NAMES NOT SHOW)

Example – Ways to Improve Performance.

Figure 3 shows jobs are queuing for hours before running. Investigation into resource contention will reduce queue times and reduce exposure to data loss.

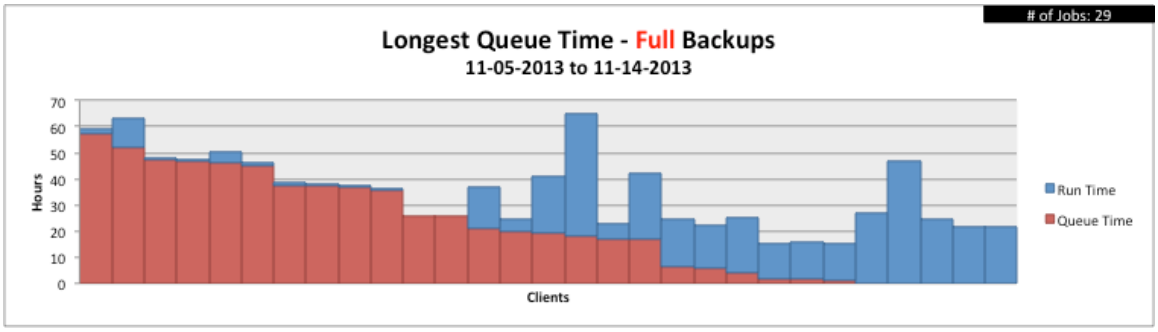

FIGURE 3 – LONG QUEUE TIMES

Figure 4 shows jobs are scheduled with significant concurrency. This can lead to resource contention and reduce throughput. A change in job scheduling can improve overall performance.

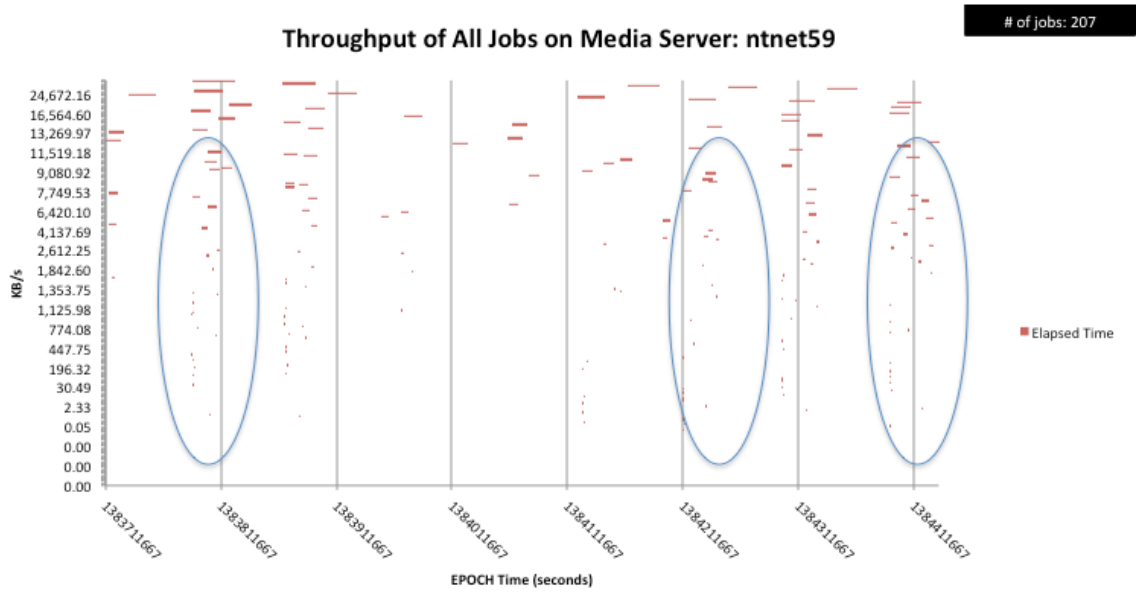

FIGURE 4 – JOB SCHEDULING

Figure 5 shows throughput by client, which is calculated from the backup metadata. There should not be a large variance between the fastest, average, and slowest throughput rate for full and incremental backups of the same client. A large variance could indicate a resource or process contention issue during the backup window and should be evaluated to determine the cause.

Clients with consistently low throughput numbers should be addressed to identify bottlenecks, which could be impeding higher throughput. Is the bottleneck on the client or somewhere along the data path between the client and media/master server? These bottlenecks could be avoided by modifying schedules to reduce contention on shared network segments, media servers, and backup devices. Another source for unexpectedly low throughput can be traced to misconfigured network settings or outdated NIC drivers on the client or outdated firmware or misconfigured port or routing tables on switches.

Throughtput Full Backups in MB/s			Throughtput Incremental Backups in MB/s		
Throughtput based on			Throughtput based on		
Avera	Faste	Slowe	Averag	Fastest	Slowest
12.9	26.7	6.8	3.0	11.5	0.8
30.8	57.1	10.2	3.0	4.2	1.9
23.1	33.9	6.1	3.0	3.4	2.6
10.9	21.7	7.6	3.0	3.3	2.8
7.6	20.0	5.4	2.9	15.9	1.1
10.3	13.4	3.1	2.9	4.0	0.8
9.8	14.5	5.7	2.9	4.5	2.2
22.3	26.9	16.3	2.9	3.5	2.3
12.3	17.7	4.5	2.8	4.2	0.7

FIGURE 5 – CLIENT THROUGHPUT

Figure 6 shows the number of files being backed up for a full backup compared to an incremental backup. The average, largest and smallest number of files for full backups and incremental backups are calculated. If a low percentage of files are changing during incremental backups and you are performing full backups on a regular basis it might be time to consider alternatives to reduce backup times and storage requirements. These alternatives include leveraging deduplication and/or archiving. Deduplication reduces storage requirements by eliminating the need to store static data repeatedly during each full backup. Depending on your backup application, deduplication can eliminate the need to process and move static data across the network to media and/or backup servers and then to backup media, while still providing a full backup image on media. Archiving moves static data from production file systems to long-term storage reducing the amount of data that has to be backed up.

This report also identifies clients with High Density File System (HDFS). HDFS have long backup times due to the time required by the backup application to walk the file system, identify changes to the file system, catalog the data to backup, etc. While backup applications have strategies for minimizing the impact of incremental backups, full backups still require long backup windows. It is typical for clients with HDFS to take multiple days to complete a full backup. By identifying clients with HDFS and using a different backup method, backup times can be decreased by a factor 5x or more.

Largest # of Fil	Smallest # of fil	Avg Number of Files	Largest # of Fil	Smallest # of fil	Avg Number of Files
3,396,734	2,902,830	3,275,650	648	563	615
3,090,413	1,651,077	2,383,474	21,931	6,636	13,531
3,037,600	1,551,739	2,095,240	18,527	3,562	8,812
2,895,969	783,281	1,276,997	86,168	325	17,375
2,132,895	279,899	1,116,702	337,142	0	10,588

FIGURE 6 - FILE COUNT AND AVERAGE FILE SIZES

Example – Ways to Lower Cost.

In Figure 7 below, Oraclesvr1, Oraclesvr2, Notessvr1, and Notessvr2 have overlapping datasets meaning the data is being backed up twice, once as a file system backup and once with an application module. Overlapping datasets prolongs backup times, increases tape and/or disk space requirements, and wastes CPU, memory and network resource on the backup client. If we look at Sqlsvr1, its file system is being backed up but is the SQL database being backed up? If it is being backed up, where is it being backed up and how often?

Host	Data Type	Policy	Policy Type	Active	Data Being Backed up
Oraclesvr1	FileSystem	Database	Standard	yes	/Flash_Recovery_Area,/Grid_Flash_Recovery_Area,/Grid_Archive,/Grid_Oracle,/Oracle,/Archive
Oraclesvr1	FileSystem	OS	Standard	yes	ALL_LOCAL_DRIVES,
oraclesvr2	FileSystem	Database	Standard	yes	/Flash_Recovery_Area,/Grid_Flash_Recovery_Area,/Grid_Archive,/Grid_Oracle,/Oracle,/Archive
oraclesvr2	FileSystem	OS	Standard	yes	ALL_LOCAL_DRIVES,
Notessvr1	FileSystem	DefaultNTServers	MS-Windows	yes	ALL_LOCAL_DRIVES,
Notessvr1	FileSystem	Notes	Lotus-Notes	yes	E:\Lotus\Domino\Data, ,E:\notesdata
notessvr2	FileSystem	DefaultNTServers	MS-Windows	yes	ALL_LOCAL_DRIVES,
notesvr2	FileSystem	Notes	Lotus-Notes	yes	E:\Lotus\Domino\Data, ,E:\notesdata
Sqlsvr1	FileSystem	SQL	Standard	yes	ALL_LOCAL_DRIVES,

FIGURE 7 - CLIENT POLICIES

Example – Impact of new technology

Figure 8 shows what impact deduplication technology can have by calculating the average file size in the customer's environment. Without this insight the potential savings from investment in certain deduplication technology cannot be determined.

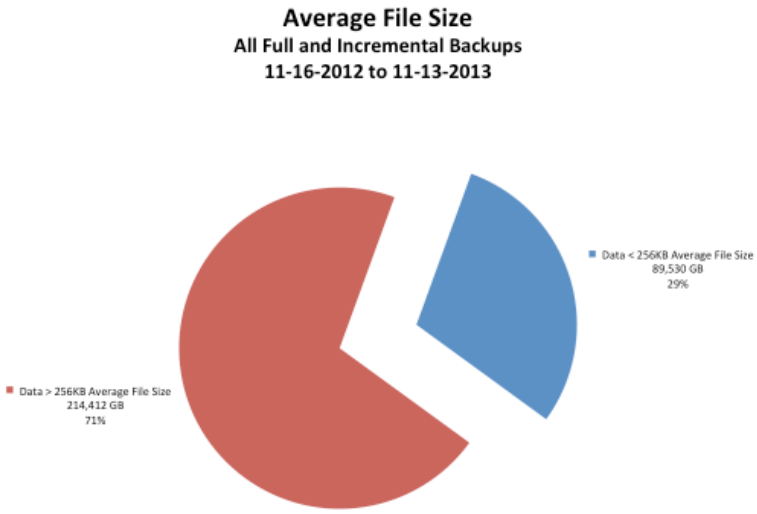

FIGURE 8 – AVERAGE FILE SIZE

Figure 9 shows the estimated client change rate, which is the amount of change between both full and incremental backups. This information can be usefully in determining the amount of network bandwidth required to replicate backup data to an offsite or ***cloud*** location on a daily basis.

As compared between all client fulls					as compared between all client incrs				
0.7%		0.1%	15.6%	-21.6%	-7.1%		-42.2%	30.9%	-485.4%
Normalized Average Daily	Standard Deviation	Average Daily	Max Daily	Min Daily	Normalized Average Daily	Standard Deviation	Average Daily	Max Daily	Min Daily
No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data
0.8%	3.5%	1.1%	16.1%	-9.4%	-3.3%	0.13	-2.2%	21.3%	-19.8%
0.8%	3.1%	0.8%	14.5%	-10.7%	-7.2%	0.36	-13.4%	21.4%	-79.1%
0.9%	3.0%	0.9%	14.2%	-10.5%	-2.6%	0.10	-1.7%	15.7%	-14.5%
0.8%	2.9%	0.8%	14.6%	-9.1%	-8.5%	0.26	-10.9%	25.2%	-59.4%
0.8%	3.1%	0.8%	14.6%	-10.8%	-3.9%	0.11	-2.8%	17.0%	-17.5%

FIGURE 9 - CLIENT CHANGE RATE

Figure 10 shows a report that calculates the growth rate for the entire environment. This information can be used to predict future storage and backup infrastructure requirements by reviewing backup growth rates and files counts.

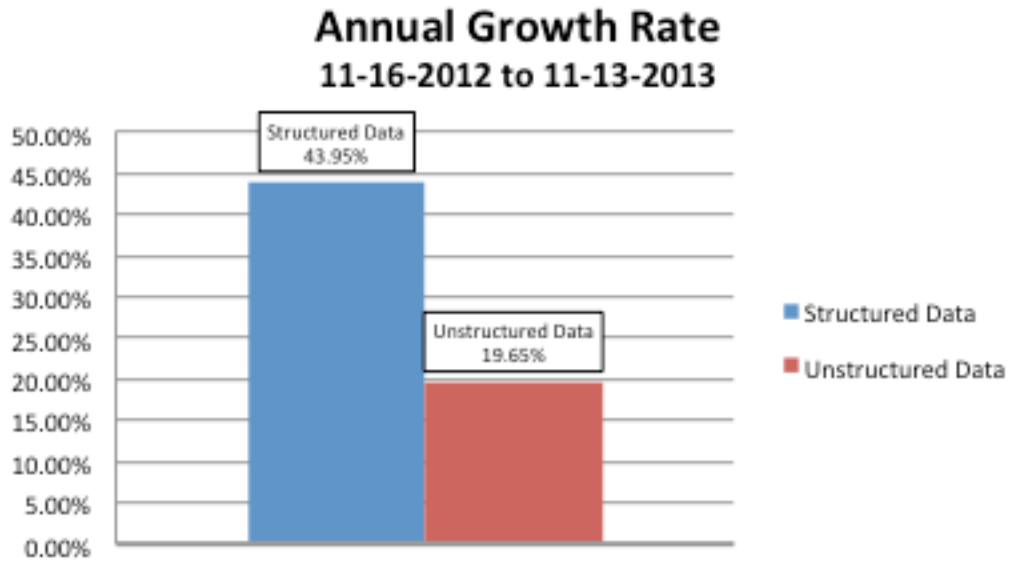

FIGURE 10 – ANNUAL GROWTH RATE

Conclusion

The backup assessment provides an x-ray into the current state of your backup environment. It is a low cost and rapid methodology to gain deep insights. All work is done on-site and in an isolated environment that avoids the risk of any data leaving your site for analysis. The assessment forms the performance baseline to compare the impact of implemented corrective actions and future technology change.