Intelli-Restore as an Instantaneous Approach for Reduced Data Recovery Time

Leon Mugoh, Ismail Lukandu Ateya, Bernard Shibwabo Kasamani

Faculty of Information Technology

Strathmore University, Nairobi

Kenya

lmugoh@iom.int, iateya@strathmore.edu, bshibwabo@strathmore.edu

Abstract: *Due to the competitive and regulatory pressures and the high demands and dependence placed on data, there is need for higher data availability and a faster means of recovering the data in case it becomes corrupted or lost. Based on results provided on the reasons behind the long / high data recovery times by Kenyan SMEs this paper provides a solution that reduces the data recovery time. In order to solve the problem of high data recovery times, an instantaneous data recovery strategy based on an existing Continuous Data Protection (CDP) architecture is introduced as an important component of a well-rounded backup and recovery strategy. CDP is a disk based backup solution which ensures that data is retrieved at a much faster rate during recovery. The solution presented in this paper could help organizations adopt or complement existing data recovery strategies.*

Key words: Intelli-Restore, automatic data recovery, continuous data protection, data management.

1. Introduction

Large volumes of data are continuously being stored in data repositories around the world. Enterprises are increasingly interconnected exposing information to a growing number and a wider variety of threats and vulnerabilities. With data storage cost per megabit reducing, data should even in the present be stored at a higher rate than ever before (Shibwabo and Ateya, 2011).

The ability to make effective decisions is crucial to an organizations survival in today's tumultuous business environment. In order for firms to evaluate alternatives and make informed choices they must have reliable and timely data upon which to make their decisions (Mugoh, Ateya and Shibwabo, 2011; Anandarajan, Anandarajan and Srinivasan, 2004). Data is considered to be an irrepleceable strategic asset to an organization that should be safeguarded (Coombs, 2008).

Enterprises, in accordance with the ISO/IEC 17799:2005 standards, use various ways to recover important information that was lost. They can use some data recovery software to try and retrieve the lost data, use some data repair software for repairing corrupted data, use external data recovery services or restore data from a data backup. Data recovery software alone is not enough to restore lost or corrupt data because it does not fully guarantee that the data will be fully recovered, and it may end up being a long and very tedious exercise trying to restore the data using recovery software when you could just simply copy back the lost or corrupted data back (ISO/IEC, 2005).

An existing CDP architecture provides for simultaneous backup in a central backup server environment where there is backup of all production servers at the same time without having to for one backup to finish before going to the next server. CDP continuously backs up data in real time therefore there are no more backup windows, which is the period of time the backup takes place and renders the systems and data inaccessible (Mugoh, et al., 2011).

There is need to complement any CDP architecture with an automatic response mechanism that can further reduce data recovery time through eliminating the human involvement in data recovery by having the system automatically detect the data loss or corruption and instantly request the backup server to restore the data.

2. Literature Review

According to Wendt (2009), CDP is described as a methodology that continuously captures or tracks data modifications and stores changes independent of the primary data, which then enables instant recovery from any non-predetermined point-in-time. This ability to do recoveries at any previous point-in-time is what distinguishes CDP from other data protection approaches such as snapshot technology (Hanavan, 2007).

CDP presents a major breakthrough in data protection and dramatically changes data protection focus from backup to recovery (Mugoh, et al., 2011). The article described by Dong, et al. (2007) acknowledges that the traditional systems for backing up data are time-consuming and this research, just like (Mugoh, et al., 2011), proposes to introduce the use of continuous data protection (CDP) to reduce the recovery time objective (RTO) of data recovery. The article by Dong, et al. (2007) has focused on data protection and recovery on web databases and is looking at improving the recovery point objective (RPO), while (Mugoh, et al., 2011) is focusing more on reducing the RTO of the recovery of any lost or corrupted data by SMEs in Kenya.

The paper by Mugoh, et al., (2011) describes a CDP architecture for Kenyan SMEs based on the findings of a survey. Mugoh, et al., (2011) targeted the system administrators of small and medium sized enterprises (SMEs) based in Nairobi, Kenya. This is because it is most likely that it is the system administrator who is responsible for data backups and recoveries in most of the organizations.

Simple random sampling technique was used to select SMEs. This was done by listing all the elements in a table then assigning them unique numbers, determining the sample size n which was calculated, using the formula obtained from Creative Research Systems (2010), to be 52 where the size of the population was 250 selected randomly from the Yellow Pages of Nairobi City. The Interval level rate was 10%, expected response rate was 60% and the confidence level rate was 90%, then used a random number generator and picked the elements with the numbers that match the random number generated until the desired sample size was obtained (Mugoh, et al., 2011).

Table 1 shows the findings on the size of data backed up.

Table 1 Size of Data Backed up Daily (N=40) (Adapted from Mugoh, et al., (2011))

Size of data	Frequency	Percentage	Cumulative percentage
Less than 1GB	3	7.5	7.5
1GB to 50GB	28	70	77.5
51GBto 100GB	4	10	87.5
Over 101GB	5	12.5	100

It was further found that, according to and Table 2, Majority of the organizations (52%) carried out their backups daily**.**

Table 2 Frequency of Data Backups (N=40) (Adapted from Mugoh, et al., (2011))

Period	Frequency	Percentage	Cumulative percentage
Hourly	4	8	8
Daily	27	52	60
Weekly	8	15	75
Every 2 weeks	1	2	77
Monthly	7	13	90
Every 3 months	4	8	98
Every 6 months	0	0	98
Yearly	1	2	100

Fig. 1 presents findings on the data backup media used to back up server data, where a majority of the SMEs (47%) data tape and the least used media to backup server data was found to be CD/DVD (6%).

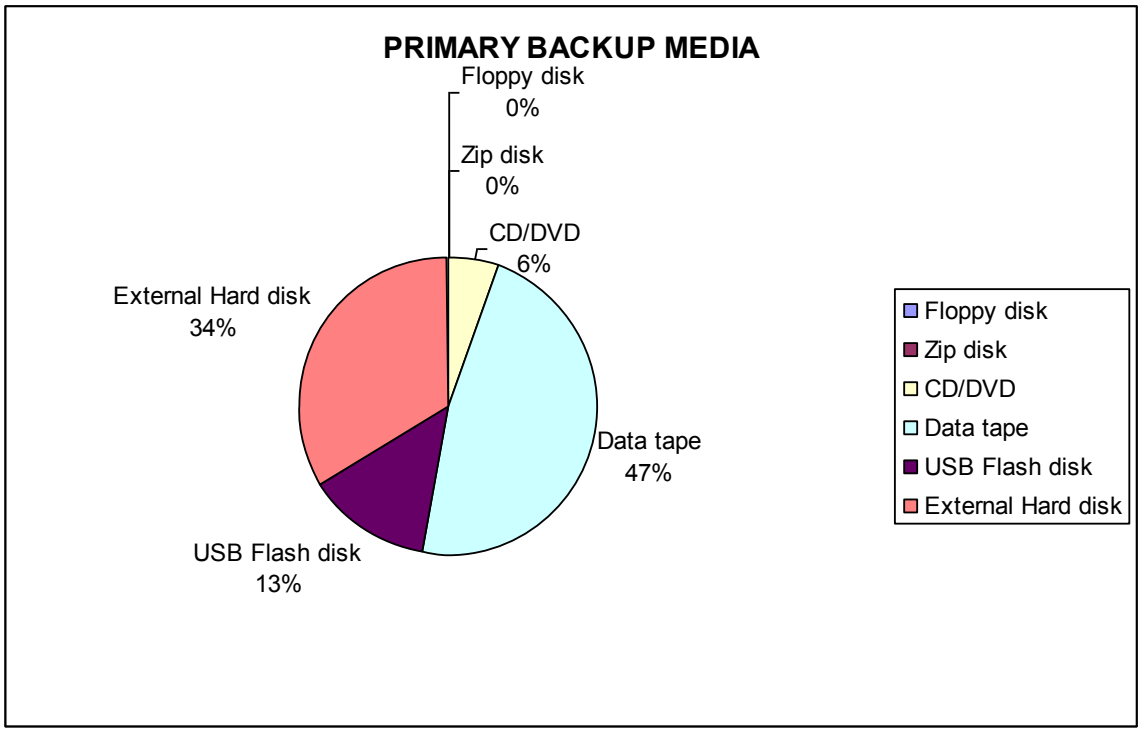

Fig. 1 Server Data Backup Media (N=40) (Adapted from Mugoh, et al., (2011))

Most SMEs use external hard disk (43%) to backup data stored in Client Workstations as shown by Fig. 2.

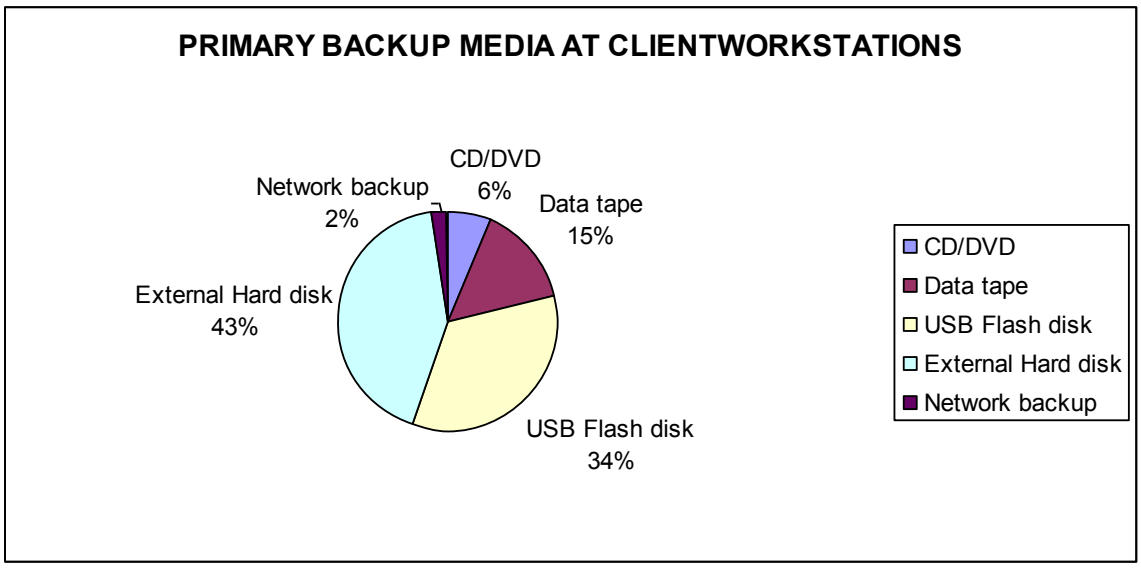

Fig. 2 Client Workstation Data Backup Media (N=40) (Adapted from Mugoh, et al., (2011))

Fig. 3 shows a routine backup process that also serves to review the backup and test restore logs and make corrections where required.

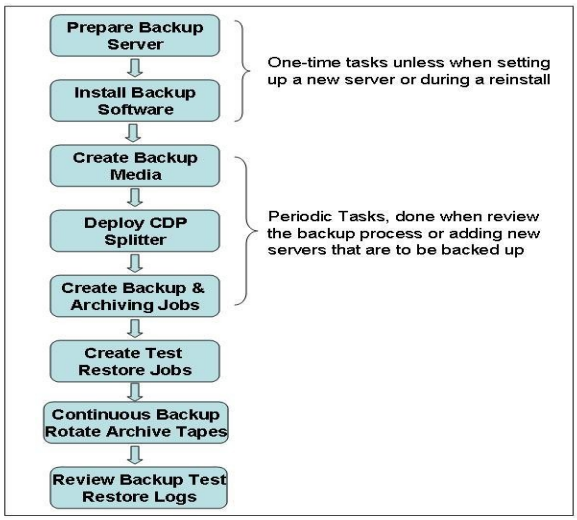

Fig. 3 Data Backup Process (Adapted from Mugoh, et al., (2011))

The proposed architecture by Mugoh, et al., (2011) has been designed to address the problems associated with data backup while little if any attempt is directed towards the recovery process. We introduce Intelli-Restore as a solution to address the automatic recovery need.

3. Intelli-Restore: The Concept of Automated Data Recovery

3.1 Overview of Intelli-Restore

It simply takes too long to get a file back the traditional way, which is to call the Information Technology Support Centre and request a file to be restored. Typically this involves the IT Administrator searching through the catalog to find which tape the file is on, which means multiple tapes to search. The restore might take two or three tapes to get the data back and there may be no guarantee that the tape is even on-site (Greene, 2005). CDP helps to reduce the time taken to recover data as it lets the users restore their own files with End User Restore. CDP solutions now enable the end user to retrieve files from a safe, secure location on the network using a standard Web browser.

Intelli-Restore is a concept that aims to complement CDP by introducing automated data recovery that hopes to further help reduce the time taken to recover data by eliminating the human element involved in data recovery where instead of the user or IT Administrator detecting the data loss and requesting for a restore it is Intelli-Restore that detects and automatically requests the backup server to restore the lost or corrupt data and finally verify the recovered data.

This idea was borrowed from the concept of the Intrusion Prevention System (IPS) as an improvement to the Intrusion Detection System (IDS). IDS is tool that can be used to detect inappropriate, incorrect or abnormal activity to help determine if a computer network or server has experienced an unauthorized intrusion and IPS is a tool used to actively drop packets of data or disconnect connections that are involved in abnormal, incorrect or inappropriate activities (Holland, 2004).

The IDS detects any abnormal behavior and alerts the system administrator to take corrective action incase of a violation while the IPS will detect any inappropriate activity and actively take corrective actions like disconnect connections or drop packets. In relation to this the Intelli-Restore agent detects the data error and takes corrective action by initiating a data restore.

Fig. 4 shows the overall picture of the Intelli-Restore concept where a host computer is directly connected to a primary storage and the Intelli-Restore system, this system is responsible for detecting the data loss or corruption in the host computer and sending a data restore request - over the Local Area Network (LAN) - to the backup server.

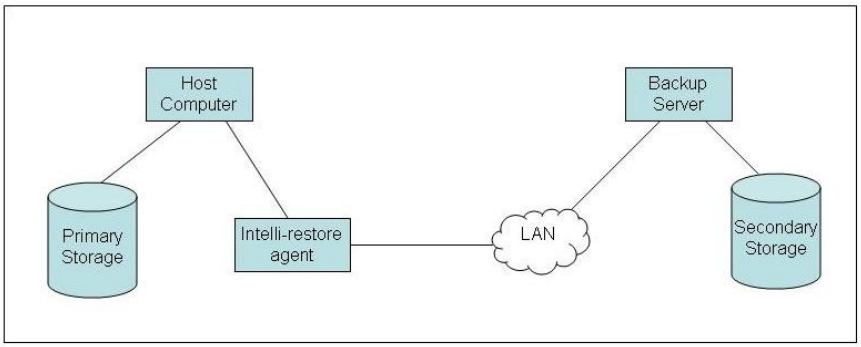

Fig. 4 Overview of the Intelli-Restore environment

Fig. 5 illustrates the system environment of the host computer that is being protected using CDP and shows where the Intelli-Restore agent will be located to be listening on to the system calls that take place and trap the errors that indicate a data loss or corruption has occurred.

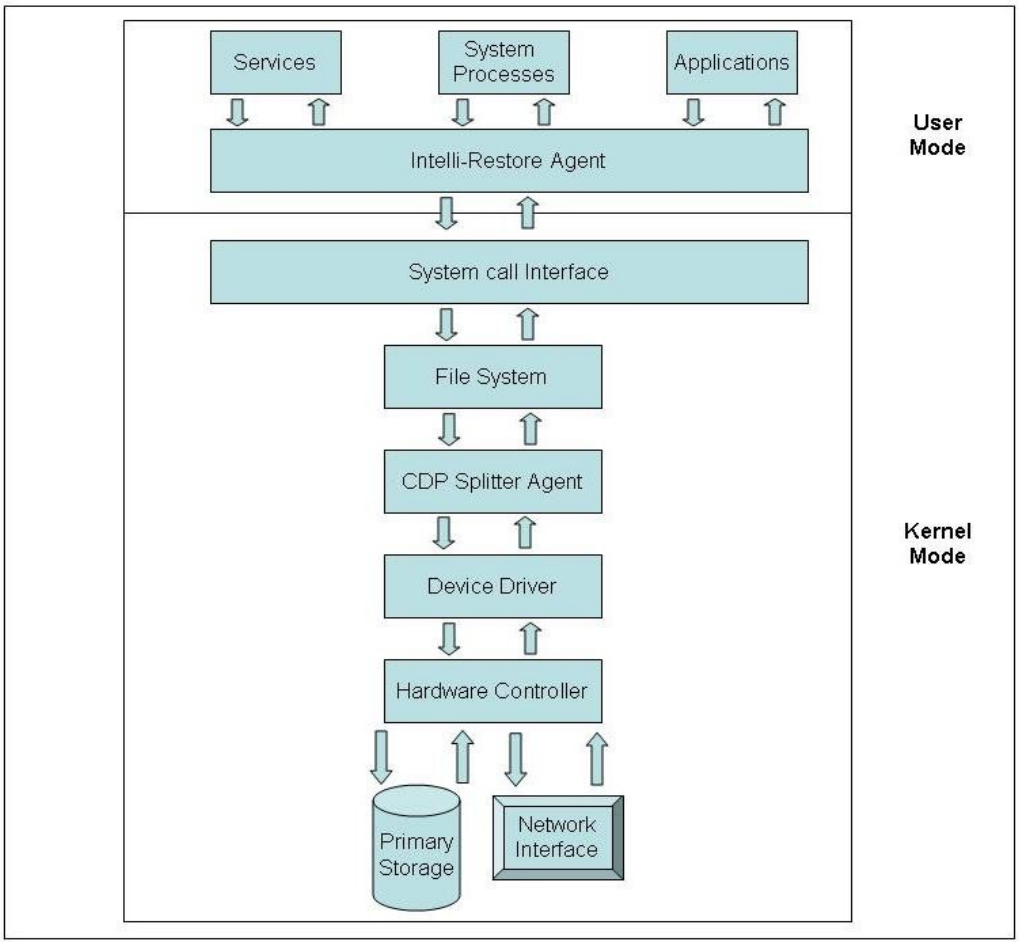

Fig. 5 System environment of host computer

3.2 Components of the Intelli-Restore Agent

Fig. 6 is a block diagram illustrating the components that make up the Intelli-Restore agent that will be responsible for detecting data losses and automatically requesting for a data restore.

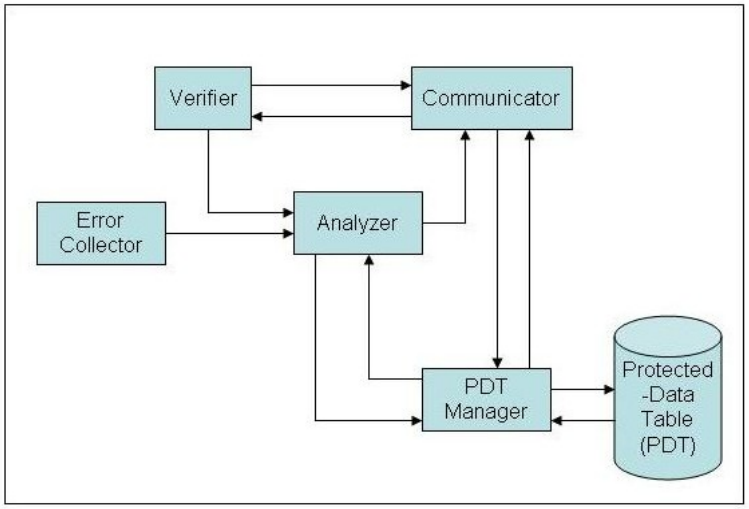

Fig 6 Components of the Intelli-Restore agent

The error collector module is responsible for listening out in the system environment with the intention of trapping the errors including information on the data involved in the incident and sends the error plus the meta-data information to the analyzer.

The analyzer filters out the error checking for errors that reflect either lost or corrupt data, then using the Protected-Data Table (PDT) Manager, it verifies whether the data in question is actually being backed up (Protected) and has been set for instant restore.

If the data is protected then the Analyzer through the Communicator sends a request to restore the lost or corrupt data to the backup server and after the restore is complete the server through the communicator requests the verifier to verify the restored data.

3.2.1 Error Collector

The error collector will trap the system error and check if it is an input / output (I/O) error as there are many different errors passed in the system. If it is an I/O error then the error collector gathers meta-data information about the data that is tied to the error that has been trapped and sends the error plus the meta-data information to the analyzer as shown by the flow chart in Fig. 7 and a description of the processes is shown in Table 3.

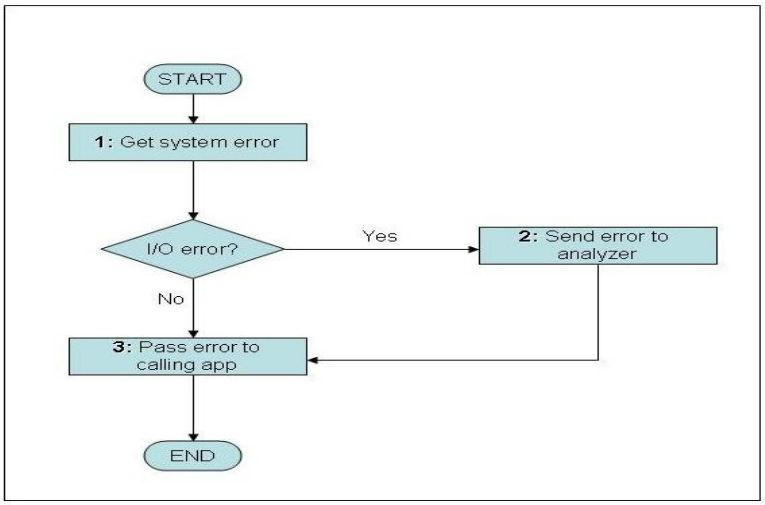

Fig. 7 Error Collector flow chart

Table 3 The Error Collector processes

Process ID	Process	Process Description
1	Get system error	The error collector scans the results of the system calls that are sent back to the calling application and looks for any error returned by the system call interface e.g. error ID 5 and its description could be Access Denied
2	Send error to analyzer	Once the error collector detects a system error, the error ID and its description is passed on to the analyzer to determine if data needs to be restored e.g. analyze error 5, Access Denied
3	Pass error to calling application	After the error collector has requested the analyzer to analyze the error to determine if data should be restored it then sends the error on to the calling application.

3.2.2 Analyzer

Fig. 8 shows how the Analyzer functions after it is notified by the error collector about an error that has occurred.

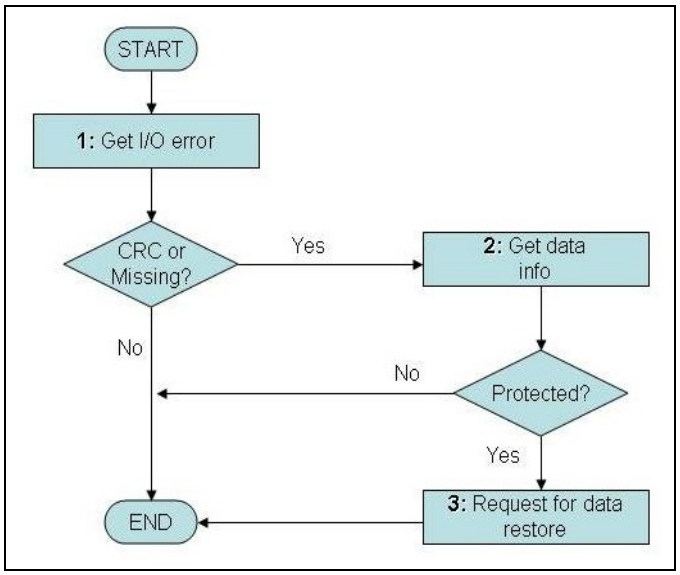

Fig. 8 Analyzer Flow chart

It first checks if the error indicates that the data is missing or corrupt and if it is a CRC or missing file error, it uses the PDT to check on the protection status of the lost of corrupted data, if the PDT returns that the data is protected then the analyzer, through the communicator, will request the backup server for a data restore. Included in the request is the meta-data information on the lost data. Table 4 gives a description of each of the processes included in the Analyzer.

Table 4 The Analyzer processes

Process ID	Process	Process Description
1	Get I/O error	The analyzer receives a request to analyze an error sent by the error collector, the parameters passed could included; the error ID, error description, and block address or file ID, file name and file path e.g. error 23 Data Error (cyclic redundancy check), FILE044, d:\data\stocks.mdf.
	Get data information	The analyzer uses this process to gather information about the data that needs to be restored, the information needed is the protection status of the data which is used to indicate whether the data is protected and has a backup copy in the backup server and that a restore is possible. The analyzer uses the PDT manager to gather this information by sending the block address or file ID and maybe file name plus file path e.g. FILE044, d:\data\stocks.mdf. Other information obtained includes the restore bit and the restore version or last point in time the restore was last done if there was an attempt to restore the file before and the restore status.
3	Request for data restore	The analyzer with the help of the PDT manager determines whether a file is protected or not. Here a protected file means a file that is being backed up to the backup server and can be recovered incase the data is lost or corrupted. Depending on the status returned by the PDT manager if the data is protected then the analyzer will request the backup server, through the communicator, to restore the data. It will pass on to the communicator the meta-data information obtained from the PDT manager.

3.2.3 PDT MANAGER

Fig. 9 shows that the PDT Manager is responsible for maintaining the Protected-Data Table as it is anticipated that there will always be changes to what data gets protected.

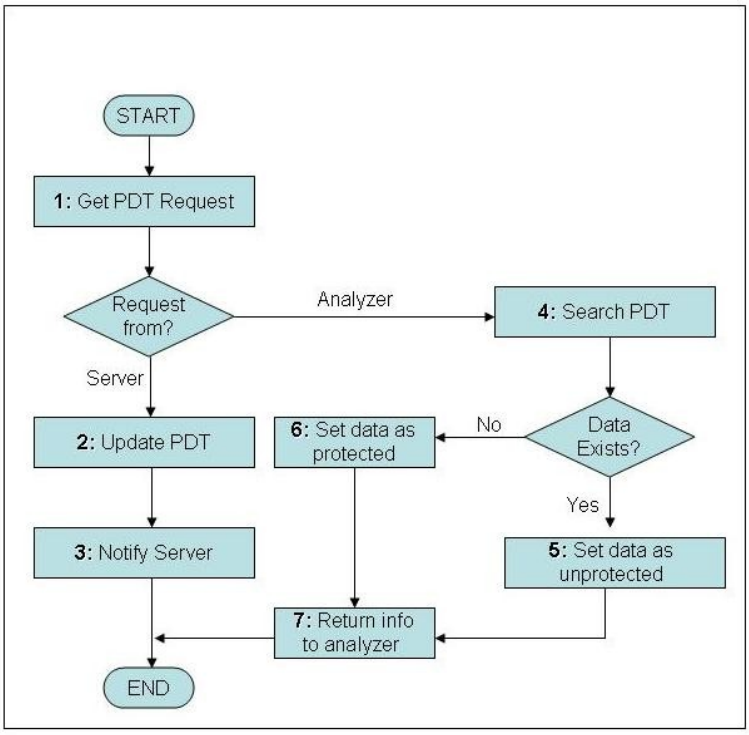

Fig. 9 PDT Manager flow chart

The PDT listens out for requests and checks whether they are from the server or from the analyzer, if it is from the analyzer it means that it is a request to return the protection status of the data that needs to be restored, so using the parameters, which contain information about the lost or corrupted data, the PDT manager searches the PDT and if a record of the data that needs to be restored exists, then the PDT manager returns that the data is protected otherwise it is unprotected. If the request is from the server then it means that this is a maintenance request which could involve adding, deleting or modifying a record in the PDT. Table 5 contains a list of the processes involved in the PDT manager and their description.

Table 5 The PDT manager processes

Process ID	Process	Process Description
1	Get PDT Request	The PDT manager receives a request from the server to update the PDT or from the analyzer to retrieve information from the PDT
2	Update PDT	This happens when a new file has been selected at the backup server to be protected or if existing data is no longer to be protected. Updating the PDT may involve adding, modifying or deleting a record in the PDT e.g. delete FILE0054
3	Notify Server	The PDT manager through the communicator just sends to the server the result of the update request, it could be successful or unsuccessful e.g. 0, update successful
4	Search PDT	This process happens when the analyzer, through the PDT manager, requests for information about the data that is to be restored. The PDT manager searches the PDT and if the record is found then the protected flag is set to true and results of the search are sent back to the analyzer otherwise the protection status is set as false and the data is considered unprotected therefore no data restore takes place.
5	Set data as unprotected	The PDT manager searches the PDT and if the record is not found then the protected data status is set to false which means that the data that needs to be restored is not protected and therefore no data restore will take place.
6	Set data as protected	The PDT manager searches the PDT and if the record is found then the protected data status is set to true which means that the data that needs to be restored is protected and therefore the analyzer can go ahead and request the backup server to restore the data.
7	Return information to analyzer	The PDT manager returns the protection status of the data that needs to be restored which could be protected or unprotected and if the data is protected it returns in addition to the status, the restore bit and the restore version or last point in time the restore was last done if there was an attempt to restore the file before and the restore status.

3.2.4 Protected-Data Table (PDT)

This is a log or database that keeps a record of all the data located in the host computer that is to be protected, which will be restored automatically in case they get deleted or corrupt. The PDT would contain such information like name or the block address of the protected file, the path it is located, the size of the file, the CRC number of the file, the location of the backup copy in case there is more than one backup server, the restore bit to indicate if the file has been restored before and the last restore

point which is the point in time the last restore was done and the restore status (successful or unsuccessful).

3.2.5 Communicator

The communicator is the interface between the Intelli-Restore agent and the CDP backup server as shown in Fig. 10.

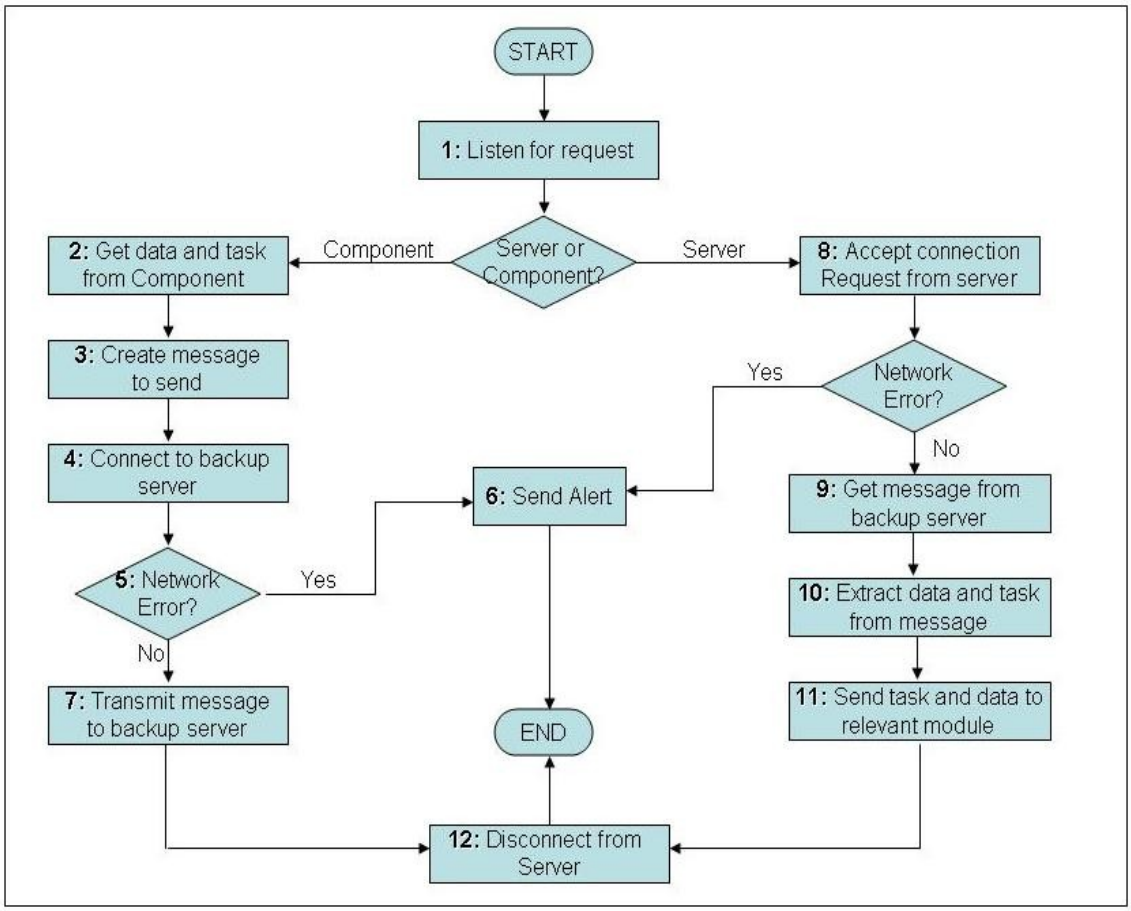

Fig. 10 Communicator flow chart

It is used by the analyzer to send restore requests to the backup server, the PDT receives requests from the server to update the PDT using the communicator and the verifier receives requests from the server to verify the restored data. The communicator will connect to the server and if there is no network error it will either send a message to or receive a message from the backup server.

Receiving involves getting the message from the server, extracting the data and task from the message and sending them to the desired Intelli-Restore component (e.g. Analyzer) and finally if there are no more messages to send or receive it disconnects from the server.

Sending involves obtaining the task and data from one of the Intelli-Restore components, creates a message and appends the data and tasks to it plus attaches the host information like host name and address, it then transmits the message to the backup server and finally if there are no more messages to send or receive it disconnects from the server. Table 6 describes the processes that are included in the communicator.

Table 6 The Communicator processes

Process ID	Process	Process Description
1	Listen for request	Communicator is waiting for a request from either a component to communicate to the server or listen for any connection requests from the backup server to receive messages. The component will request the communicator to send a task and the parameters for the backup server to carry out e.g. the analyzer can request the communicator to send the request to carry out a restore with the information about the data as the parameters(file name or block address), other parameters can include the location of the backup copy, the restore bit to indicate if the file has been restored before, restore status and the last restore point which is the point in time the last restore was done.
2	Get data and task from component	The calling component passes the task to be carried out plus the parameters, for example the verifier would signal the backup server to end the restore process and the parameters would include the restore session ID and the resultant status of the restore job whether it was successful or not.
3	Create Message to be sent	In this process the communicator generates a message ID, specifies the message type which could be a 0 for call or a 1 for reply, and adds to the ID and the type the message body which contains the task to be carried out plus its parameters. The host and backup location information is also attached to the ID, type and body an example would be MSG004 as the message ID, 0 as the message type, message body would be the request to restore data plus the parameters which could include the file name or block address of the data to be recovered, the restore bit and the restore version or last point in time the restore was last done if there was an attempt to restore the file before and the restore status. The host information will include the host address and the backup location will be the address of the backup server.
4	Connect to Backup Server	The communicator requests to connect to the backup server the command is to connect and the destination server address and port number are specified as the parameters e.g. 10.1.0.4 as the server address and the port number can be 9005.
5	Network Error	This involves checking the status returned by the server after the communicator request to connect. If there is no error then the communicator will continue to either receive or send a message. This status will contain the error ID and the error description an example of a returned status could be 10060 Connection timed out.
6	Send Alert	The communicator notifies the calling component in case an error occurs while connecting to the backup server, The information passed will include the error ID and the description, for example 10060 as the ID and Connection timed out as the description.
7	Transmit Message	Communicator delivers the created message to the backup server it could be a request to initiate a data restore, so it will contain the command to start the data restore and the parameters will be the information about the data that is to be restored which could be filename or block address, the restore bit to indicate whether there was an attempt to restore the data before, the resultant status of the previous restore and the point in time where the data was restored or the restore version.

Process ID	Process	Process Description
8	Accept connection request	This is the communicator's response to a connection request from the backup server, which could be a message that the communicator acknowledges the connection request and therefore data transmission can commence after the connection succeeds.
9	Get message from backup server	The backup server sends a command and its parameters to the relevant component through the communicator, the communicator receives the message which is made up of the message ID, message type, message body, source address and source port e.g. MSG098, 0, verify RST023 d:\data\stock.mdf, 10.1.0.5 9005
10	Extract task and data from message	The communicator retrieves the command to be carried out and its parameters, in this case the command is to verify a restore job which contains parameters such as the restore session ID, the file name or block address and the restore path. RST023, d:\data\stock.mdf.
11	Send task and data to relevant component	The communicator does this by calling the appropriate component based on the command that was sent by the server and passing the parameters along. It could be a command to verify the restore job in which the communicator will call the verifier or update the PDT which will mean that the communicator will have to call the PDT manager.
12	Disconnect from server	The communicator sends a request to the backup server to terminate the connection, this includes the command to close the connection and parameters could include the backup server address. The communicator sends a BYE request to 10.1.0.5 after it gets an acknowledgement from the backup server it sends a SHUTDOWN requests and gets an acknowledgement from the server.

3.2.6 Verifier

The verifier has the task of confirming to the backup server that the restore was successful and that the restored file is accessible. Fig. 11 illustrates the processes involved in verifying a data restore.

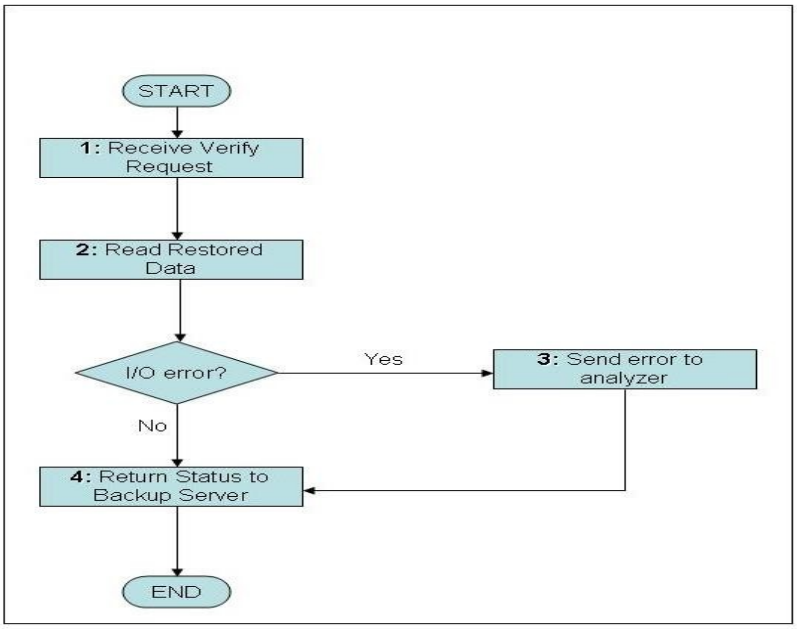

Fig. 11 Verifier Flow chart

The verifier receives a verify request from the backup server through the communicator and using the data information passed through in the request it reads the restored data to check it for errors. If an error occurs then the verifier requests the server to restore the next version of the data by sending the error to the analyzer else it notifies the server that the restore was a success. Table 7 shows a list of the processes involved in the Verifier and their descriptions.

Table 7 The Verifier processes

Process ID	Process	Process Description
1	Receive verify request	The backup server, through the communicator, requests the verifier to verify the restored data, to confirm that the data restored is accessible. The backup server sends the verify command with the restore session ID, the file ID or block address, file name and the restore path. e.g. verify RST023, FILE0056,d:\data\stock.mdf.
2	Read restored data	The verifier given the information about the restored data will attempt to open and read the restored data and trap any errors that it encounters. Read FILE0056.
3	Send error to analyzer	In case an error is detected when reading the restored data, the verifier notifies the analyzer to request for another data restore by passing the information about the restored data and this time setting the restore bit to 1 to indicate that there was already an attempt to restore the data. The error ID and its description plus the meta-data information about the data that is to be restored (which includes the file id, file name, file path) is passed on to the analyzer to determine if data needs to be restored e.g. analyze error 23 Data Error (cyclic redundancy check), FILE044, d:\data\stocks.mdf, 1.
4	Return status to backup server	The verifier has to notify the backup server on the outcome of the verification process, the status returned could be 0 for successful data restore or -1 for an unsuccessful data restore.

4. Conclusion and Limitations

This paper presents Intelli-Restore as an automated approach to data recovery based on Continuous Data Protection. The presented approach is a platform independent design that integrates the critical components for automated data recovery. The accomplished goal is an architecture for the implementation of CDP to help reduce the time taken to recover data.

By deploying CDP, organizations do not replace traditional backup but rather add or introduce an important component of a well-rounded backup and recovery strategy. CDP complemented with Intelli-Restore is expected to further reduce data recovery time as it attempts to eliminate the human involvement in data recovery by having the system automatically detect the data loss or corruption and instantly request the backup server to restore the data.

The existing backup model uses data tapes as its primary backup media and according to Mugoh, et al., (2011), Tape based backup systems lacks the speed, reliability, flexibility and simplicity that many organizations need today in a data protection solution also backing up to tape alone is no longer adequate and it is difficult to administer for backups and recoveries.

This paper mainly focused on the recovery of data stored in Servers and Client Workstations. The paper was based on an initial study by Mugoh, et al., (2011) that was also focused on small and middle sized companies in the Nairobi area. Intelli-Restore is limited to file or block level Data protection cannot handle application level data protection.

This means that Intelli-Restore cannot detect any data loss or corruption in applications for example if a row was deleted in an excel sheet or a table within a database was dropped. The Servers and Workstations are expected to be constantly connected in the Local Area Network. Therefore, Intelli-Restore does not cover mobile computing which connect over intermittent networks.

CDP is a disk-based solution, this in itself makes data recovery faster as it is faster to retrieve data from a disk than from a tape and it further reduces the data recovery time by simplifying the data recovery process where the steps taken to recover data are reduced.

Additional benefits of implementing CDP include the possibility of having simultaneous backup in a central backup server environment where the solution can backup all production servers at the same time without having to wait for one backup to finish before going to the next server. CDP continuously backs up data in real time therefore there are no more backup windows, which is the period of time the backup takes place and renders the systems and data inaccessible. Intelli-Restore has been introduced to manage the data recovery process after backup.

References

Anandarajan M., Anandarajan A., Srinivasan C., 2004. *Business Intelligence Techniques,* Berlin-Germany: Springer-Verlag

Coombs, W.T. 2008, *PSI Handbook of Business Security*, Greenwood Publishing Group, Westport, Connecticut

Creative Research Systems 2010, Sample Size Formulas for our Sample Size Calculator, viewed 20th March 2010, http://www.surveysystem.com/sample-size-formula.htm

Dong, G., Lin, X., Wang, W., Yang, Y. & Yu, J.X. 2007, *Advances in Data and Web Management*, Springer, Enterprise Square, Hong Kong

Greene, B. 2005, Continuous Data Protection a Better Backup Option, viewed 19th February 2010, http://www.technewsworld.com/story/45721.html

Hanavan, P., 2007. An Overview of Continuous Data Protection, viewed 15th February 2010, http://www.infosectoday.com/Articles/Continuous_Data_Protection.htm

Holland, T., 2004, Understanding IPS and IDS: Using IPS and IDS together for Defense in Depth, viewed 17th March 2010, http://www.sans.org/reading_room/whitepapers/detection/understanding_ips_and_ids_using_ips_and_ids_together_for_defense_in_depth_1381

ISO/IEC 2005, International Standard - Information Technology - Security Techniques - Code of Practice for Information Security Management, Viewed 16th March 2009, http://www.securitycn.net/img/uploadimg/20060313/ISO_IEC_17799_2005_PDF_version_(en).pdf

Mugoh L., Ateya I. L., & Shibwabo B. K., 2011 Continuous Data Protection Architecture as a Strategy for Reduced Data Recovery Time. *Journal of Systems Integration,* 2 (4): 54-69

Shibwabo B. K., & Ateya I. L., 2011. Repository Integration: The Disconnect and Way Forward through Repository Virtualization Supporting Business Intelligence. *International Journal of Current Research*

Wendt, M.J., 2009. Achieve Near-Real Time Backup and Recovery With Near-Zero Ongoing *Administrative Effort*, viewed 17th February 2010, http://viewer.bitpipe.com/viewer/viewDocument.do?accessId=11624743

JEL Classification: L60, M15