

The Resilient and Conservation Database

Victoria E Holt

The demand on data availability and resilience is a critical factor when providing agile data anytime in this system of interest. The need to reduce complexity of applications to provide federated mission critical secure systems which are scalable, auditable and recoverable is of key importance. There are 5 closely related environments which are part of the holistic picture. These being interlinked areas

- Resilience is defined as the capacity of ecosystems and populations to return to a previous state of after they have been disturbed. Thus for a database there are various archetypes which need to be followed to ensure resilience and conservation of data.
- A backup is a physical copy of data a file is held on a physical hard disk which is to be used for restore and recovery purposes in the event of the current data being lost or corrupted.
- Archiving data where a copy of the file is held for safety on a secure stable storage medium often at a different geographic location for long term recovery or retrieval of data for historic purposes.
- Cloud computing which is a subset of grid computing, a paradigm in which tasks are assigned to a combination of connections, software and services access over a network. This is a set of shared virtual resources which are accessible in a secure and scalable manner.
- Grid computing is a technology approach to manage a cloud but not all grids manage a cloud. Sometimes known as the computational grid resources of many computers in a network are applied to a single problem at the same time. Examples of this are SETI (Search for Extraterrestrial Intelligence) @Home project or to deal with the data centric problems of the Large Hadrom Collider at CERN.

High availability for mission critical systems has an architecture designed in such a manner to aid in system uptime and predictable performance. A rich picture of a resilient and conservation database system is in figure 1. The rich picture shows a variety of data required to be stored, the retention, 24/7 365 days a year required access, the need for secure systems available for analysis and the people involved requiring and sharing data all of whom have different viewpoints on what is most important.

strategies will cover how to ensure the data is visible through applications if required.

- Regular monitoring and maintenance of the database ensure excellent system performance remains achievable. Every database server contains data and should have the ability to fix some internal corruptions without administrators intervention to ensure the large amount of data needed by businesses is continually available.
- The ability to provide disaster recovery for databases adds resilience to the system and if another database or backups are held remotely at another location the level of resilience increases.

Disaster Recovery (DR) is a process within business continuity planning. This is the process of restoring a previous copy of the data and other necessary processes to ensure the data is brought to a known point of consistency.

Business Continuity Planning (BCP) is best described as the process, policies and procedures that are carried out by an organisation to ensure that essential business functions continue to operate during and after a disaster. This planning allows organisations to protect their mission critical services and give themselves their best chance of survival after a natural or human-induced disaster, system failure or infrastructure failure. The resilient system is required to stop lost of revenues, loss of productivity, the lost of the companies reputation etc.

The Resilient System

The root definition for a resilient system is

A system to ensure data is available all the time for the public or organisations whether it is provided locally or by multiple servers or remotely at geographic spatial sites.

The requirement for resilient disaster recovery is determined by the organization's need to protect against site failure, network failure, storage failure or data loss caused by faulty code. All of these eventualities should be covered within a documented plan. The plan for backups asks two questions to determine the level of cover required.

- How long a can the system be down?
- How much data loss is acceptable?

The resilient system has three types of standby solution available depending on how long the system can be down Hot -> Warm -> Cold. Resilience is explained in figure 2.

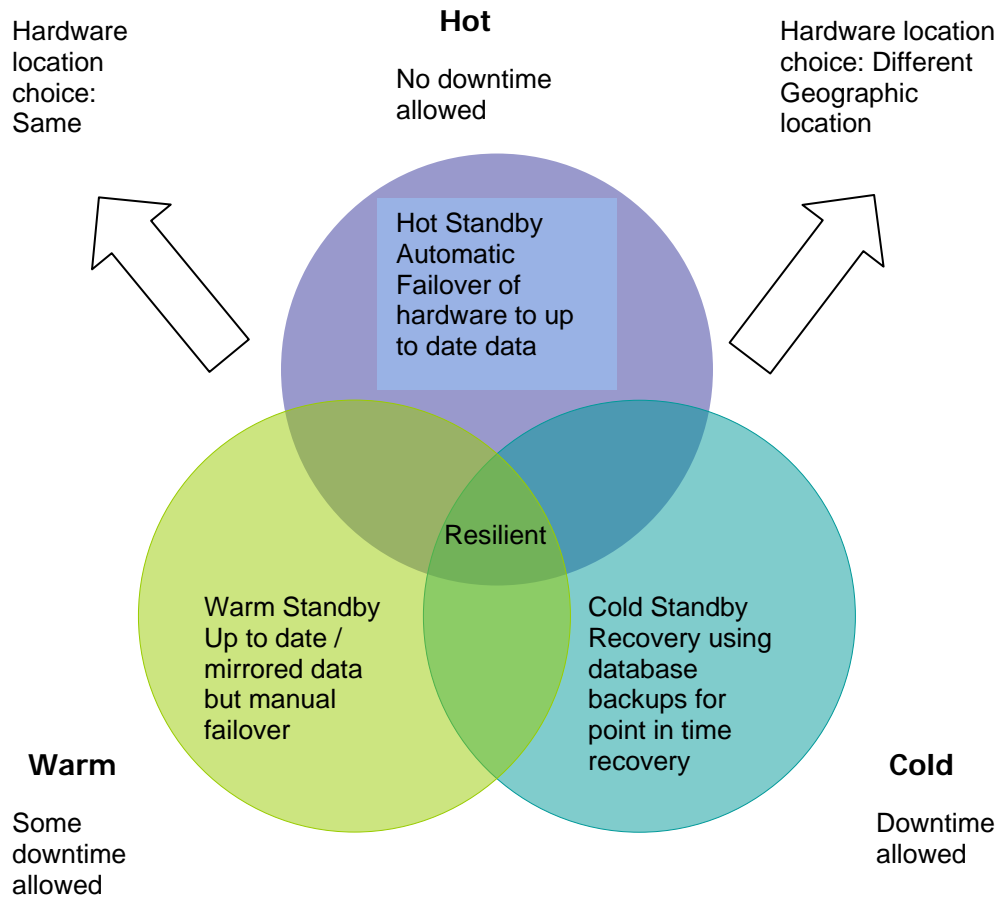


Figure 2 System recovery for a resilient high availability solution with additional disaster recovery

The additional disaster recovery solution is also available in case the above resilient system fails or becomes unavailable due to other issues. However the espoused theory (what people say) and theory in action (what people do) is often divergent and issues arise.

The Backup System

To reduce the risk to the production database data and ensure the data is available when it is required it is necessary to carry out database backups. Backing up the database ensures that the data loss is always kept to a minimum and this can help keep the database in a good state of health. Many organisations have legal requirements to maintain the data for a certain number of years and although this data may not be

required to be continually accessible there may be a need to recover this at some point in the future. It is necessary to establish a backup strategy that follows the organizational requirements.

Figure 3 shows a backup system.

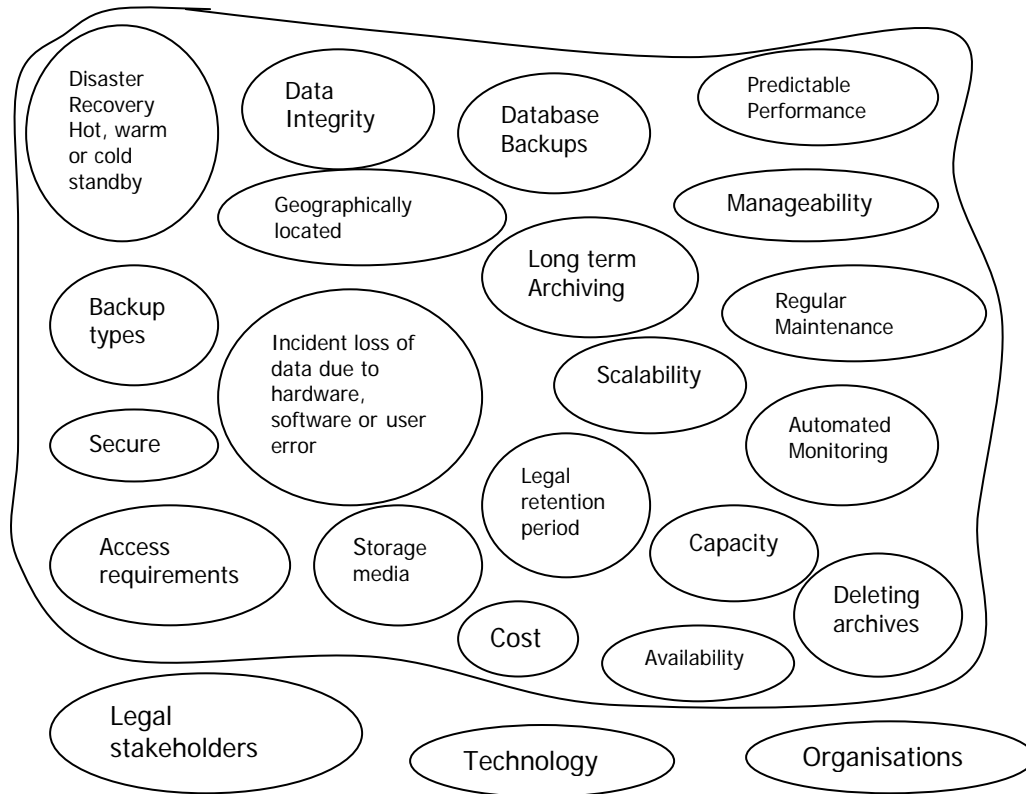


Figure 3 A systems map for database backups

The size of the database affects the level of work required. Very large database infrastructures require a huge amount of maintenance to ensure they remain healthy which ideally should be performed following a best practice maintenance schedule. To reduce the risk of failure it is best to have a maintenance strategy which is divided into small tasks which are automated where possible.

A root definition of the backup system is

A system to protect data for organisations and the public by the Database Administrator by regularly creating backups using a database software application which stores the files in a different geographic location from the original

For each system there are stakeholders such as organization, legal and technical people and the public users of the system. The backup and restore strategies each have their own strengths and weaknesses.

The database sizes and frequency of data modification determine which database backup strategy is used. They are full database backups, differential backups and the transaction logs backup strategy. These ensure databases are backed up throughout the day. In case of failure, it is possible to recover all backed-up complete transactions up to the point in time of a failure. Only uncommitted transactions will be lost.

Database backups are usually stored on storage media and held at the organizations defined locations which for most organizations this will be a different site than that of the production database in case of disaster. It is unlikely that you will ever use the archived backup to restore systems but it maybe required so plans should be put in place. It should be noted that not all data is held in a database so other plans need to be put in place to backup that data.

The design of the backup strategy should consist of simple procedures that are regularly carried out to ensure that the databases are backed up and can be restored to the server. The planning and implementation of backup and recovery plans require careful consideration.

A standard maintenance plan

A standard maintenance plan is a written document that details a proposed program of work. It describes details the strategies, establishes objectives, methods and processes that must be followed to achieve the objectives and goals. It captures the vision, current status, expected needs, projected results and financial needs for the business.

This could consist of

- Creating backups of user and system databases
- Monitoring for any failures of full backups of databases, differential and or transaction log backup failures on production servers primarily.
- To regularly test the backup procedures and verify the quality of the backups taken to make sure that the backups restore accurately.
- Have a plan on how to restore historical backups
- Identify how the integrity and optimization of backups are maintained in the database with supporting applications
- Ensure the database is not fragmented, files not larger than they need to be and indexes rebuilt to ensure the core databases are not corrupt.
- Devise a plan of how many backups and types of backups need to be taken everyday.
- Define the retention period for backups before they are archived for the long term.
- Ensure backups can be transported off-site or protected away from the servers for increased security.

A key factor to ensure resilience can be maintained is the regular testing of restorations on a different cross section of databases from different servers to restore backups taken. The test to include:

- Ensure the backup is readable
- Restore a full database backup to a test server
- Test the transaction log files by restoring the backups to a specific point in time on a test server
- Run some basic tests to ensure the database can be accessed and is functioning correctly.

Very large databases backups are more likely to have errors so it is best to check whether the file is damaged before placing the backup in an archived location or spending time trying to restore a faulty backup.

Long Term Conservation

Mullins defines Database Archiving as " the process of removing selected data records from operational databases that are not expected to be referenced again and storing them in an archive data store where they can be retrieved if needed. "

This archiving of data for long term conservation fulfils a need for total digital preservation within the database. Once data has been migrated to this archived database, backups can be moved to long term storage facilities off site that are not accessible generally. However, data may still be required to be accessed in some form.

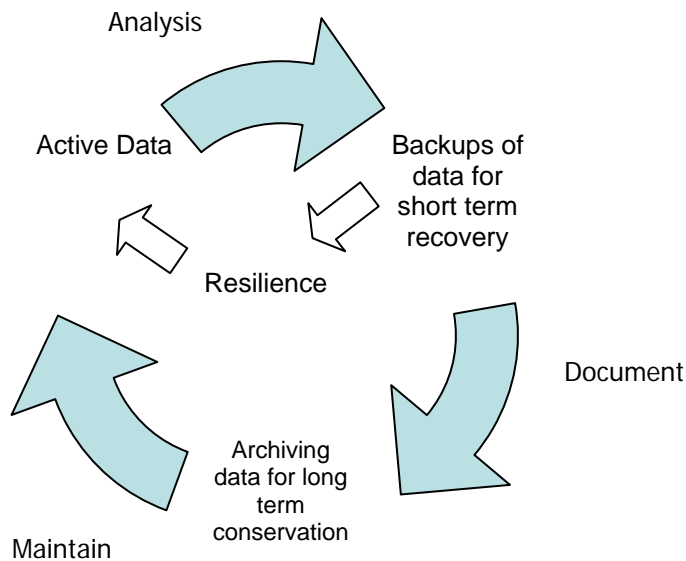


Figure 4 Archiving Data Lifecycle

Data exists in many formats and database archiving is only a small part of the data archiving plan. The government and organizations continually introduce new regulations which require data to be conserved. Figure 4 shows the iterative process of archiving and how this leads to providing more resilience to the holistic system and how even archived data can still be linked to active data. Each type of data may require a different archiving process depending on the physical nature of the software.

An example of archiving is the internet archive. The Internet Archive is building a digital library of Internet sites and other cultural artefacts in digital form to prevent the Internet, items of historical significance and other "born-digital" materials from being lost.

The archiving, preservation, best practice and management of digital data is being researched holistically by the Digital Curation Centre (DCC). They are looking to preserve data for future generations, over its life cycle, the processes for producing good data and management. There is the need for standards to be put in place to assist with conserving data on a huge scale, the complexity of which needs to be addressed. Will the data after 30 years mean anything or has the structure, the purpose of data, the meta data been catalogued to go with the backups. The implications of saving our history for the future are high.

When it comes to archiving data there are various things which need to be determined. These are:

- It is necessary to understand the industry requirements such as legal and compliance standards requirements
- Data analysis to review application data to determine the key tables and dependencies between the tables when archiving.
- It is necessary to understand how the data is used and accessed to see what additional data and applications can be archived
- Justify a business case for archiving the data is essential.
- Regular backups restore quickly, are easier to manage and help increase performance.
- Set up automate data archiving to occur regularly on a daily, weekly, monthly or quarterly basis
- To make it standard practice for all companies
- Determine what is trusted data

A different approach is required for different lengths of data storage, see figure 5.

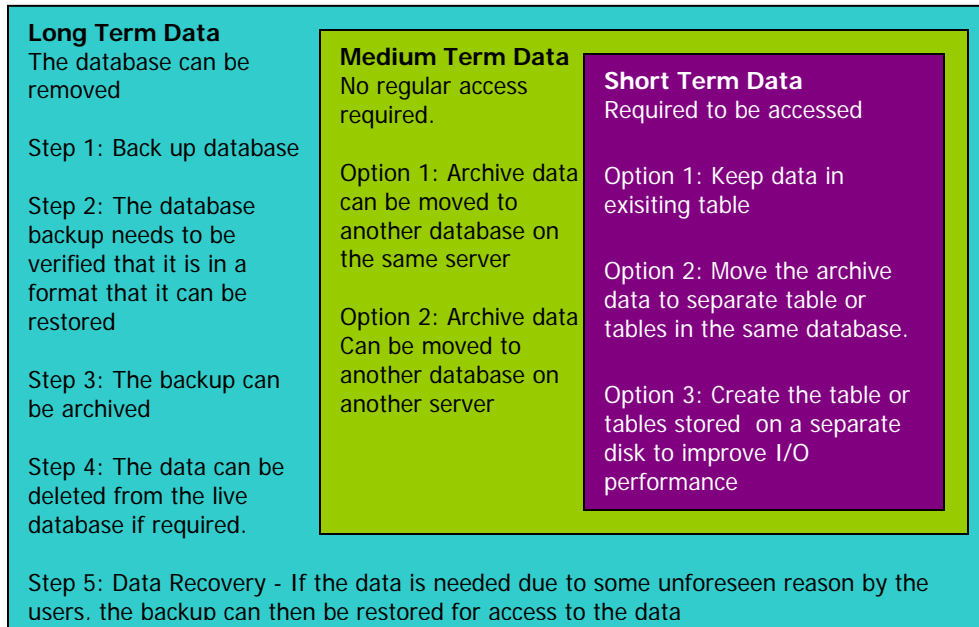


Figure 5 Archiving choice

A permanent archiving solution should be automated and a documented plan created to ensure a system is established for the long term archiving needs. This part of the plan may cover

- Decide on online or offline storage for long term archiving of backups.
- Evaluate the systems to determine possible data for archiving.
- Test the archive process
- Ensure sufficient time is available to archive the data.
- Backup the server configuration every time it is changed in case it is required when a restore takes place.
- Take full backups
- Establish an automated process. Check the execution result of the process for success or failure.
- Make sure the users understand the archiving process and check the archive process works. To ensure users don't break the system move a copy of the backup off-site.

The need for continuous data availability highlights the need for resilience and retrievable short and long term data recovery. The data explosion of digital data poses new issues when transporting this globally. Grid technology could mean internet traffic can move data around at 10000 faster will alter what and how much data is stored This makes the need for future resilience more important. This alternate internet based on grid technology could mean people store data which is accessible from anywhere anytime. This is currently being referred to as cloud computing.

The grid computability developed at CERN has been built with dedicated fibre optic cables and modern routing centres to handle the data explosion. The big ideas behind the grid are

Resource sharing	On a global stage. The grid philosophy
Secure access	There must be a high level of trust. Secure access obtained Access Policy – what is shared? who is allowed to share ? can sharing occur? Authentication – How do you identify a user or resource? Authorization – How do you determine whether a certain operation is consistent with the rules?
Efficient use of resources	a mechanism to allocate work efficiently and automatically among many resources
The death of distance	ensure distance makes no difference to efficient access to computer resources
Open standards	Common standards to which everyone can constructively contribute

The emergence of the cloud to address the systemically desirable and culturally feasible changes will affect this bounded system. Thus the need for reliable, resilient, conserved data at speed with a never ending volume of new data that is easily accessible requires new systems and procedures to be established to deal with the holistic nature of data.

References

Essential SQL Server 2000 An Administration Handbook Buck Woody

The Handbook for Reluctant Database Administrators Josef Finsel

SQL Server Backup and Recovery Tool and Techniques Frank McBath

Archiving Data in SQL Server and Maintenance tasks: automating the restore verify process <http://www.mssqltips.com/tip.asp?tip1121>

The Internet Archive <http://www.archive.org/index.php>

Digital Curation Centre <http://dcc.ac.uk/>

Real World SQL Server Disaster recovery A survival toolkit for the DBA by Brian Knight

Checklist: how to archive SQL Server backups by Greg Robidoux Edgewood Solutions 11/5/06

SQL Server disaster recovery: Recreating historical data by Greg Robidoux Edgewood Solutions 14/9/06

Database backup and restore corruption you haven't considered by Greg Robidoux Edgewood Solutions 14/9/06

D-lib Magazine Dec 2006 volume 12 Number 12

Using the Audit checklist for the certification of a trusted Digital Repository as a framework for Evaluating Repository Software Applications.

Legal Requirements to Archive Database Data , Database Archiving, The Impact of Data Volume on operational databases, The Trend toward long-term data retention Craig Mullins

<http://www.dbazine.com/blogs/blog-cm/craigmullins/blogentry.2007-01-02.3463283692/sbtrackback>

Anatomy of an Archiving Project - seven essential components for managing your enterprise data Dec 2006 White paper Princeton Softech

Science and Technology Dictionary Chambers

The Oxford Companion to Philosophy

Coming soon: superfast internet by Jonathan Leake Science Editor The Sunday Times 6 April 2008-04-07

Grid Café <http://gridcafe.web.cern.ch/gridcafe/whatisgrid/whatis.html>