

Kimball Design Tip #28: Avoiding Catastrophic Failure Of The Data Warehouse

By Ralph Kimball

The tragic events of September 11 have made all of us re-examine our assumptions and our priorities. We are forced to question our safety and security in ways that would have seemed unthinkable just weeks ago.

We have been used to thinking that our big, important, visible buildings and computers are intrinsically secure, just because they are big, important, and visible. That myth has been shattered. If anything, these kinds of buildings and computers are the most vulnerable.

The devastating assault on our infrastructure has also come at a time when the data warehouse has evolved to a near production-like status in many of our companies. The data warehouse now drives customer relationship management, and provides near real time status tracking of orders, deliveries, and payments. The data warehouse is often the only place where a view of customer and product profitability can be assembled. The data warehouse has become an indispensable tool for running many of our businesses.

Is it possible to do a better job of protecting our data warehouses? Is there a kind of data warehouse that is intrinsically secure and less vulnerable to catastrophic failure?

I have been thinking about writing on this topic for some time, but suddenly the urgency is crystal clear. Let us list some important threats that can result in a sustained catastrophic failure of a data warehouse, and what kinds of practical responses are possible.

Catastrophic Failures

Destruction of the facility – A terrorist attack can level a building or damage it seriously through fire or flooding. In these extreme cases, everything on site may be lost, including tape vaults, and administrative environments.

Deliberate sabotage by a determined insider – The events of September 11 showed that the tactics of terrorism include the infiltration of our systems by skilled individuals who gain access to the most sensitive points of control. Once in the position of control, the terrorist can destroy the system, logically and physically.

Cyberwarfare – It is not news that hackers can break into systems and wreak havoc. The recent events should remove any remaining naïve assumptions that these incursions are harmless, or “constructive” because they expose security flaws in our systems. There are skilled computer users among our enemies, who are actively attempting today to access unauthorized information, alter information, and disable our systems.

Single point failures (deliberate or not) – A final general category of catastrophic failure comes from undue exposure to single point failures, whether the failures are deliberately caused or not. If the loss of a single piece of hardware, a single communication line, or a single person brings the data warehouse down for an extended period of time, then we have a problem with the architecture.

Countering Catastrophic Failures

Distributed architecture – The single most effective and powerful approach for avoiding catastrophic failure of the data warehouse is a profoundly distributed architecture. The “enterprise data warehouse” must be made up of multiple computers, operating systems, database technologies, analytic applications, communication paths, locations, personnel, and on-line copies of the data. The physical computers must be located in widely separated locations, ideally in different parts of the country or around the world. Spreading out the physical hardware with many independent nodes greatly reduces the vulnerability of the warehouse to sabotage and single point failures. Implementing the data warehouse simultaneously with diverse operating systems (e.g., Linux, Unix, and NT) greatly reduces the vulnerability of the warehouse to worms, social engineering attacks, and skilled hackers exploiting specific vulnerabilities.

Parallel communication paths – Even a distributed data warehouse implementation can be compromised if it depends on too few communication paths. Fortunately, the Internet is a robust communication network that is highly parallelized and continuously adapts itself to its own changing topology. The Internet is locally vulnerable if key switching centers (where high performance web servers attach directly to the Internet backbone) are attacked. Each local data warehouse team should have a plan for connecting to the Internet if the local switching center is compromised. Providing redundant multi-mode access paths such as dedicated lines and satellite links from your building to the Internet further reduces vulnerability.

Extended storage area networks (SANs) – A SAN is typically a cluster of high performance disk drives and backup devices connected together via very high speed fiber channel technology. Rather than being a file server, this cluster of disk drives exposes a block level interface to computers accessing the SAN that make the drives appear to be connected to the backplane of each computer. SANs offer at least three huge benefits to a hardened data warehouse. A single physical SAN can be 10 kilometers in extent. This means that disk drives, archive systems and backup devices can be located in separate buildings on a fairly big campus. Second, backup and copying can be performed disk-to-disk at extraordinary speeds across the SAN. And third, since all the disks on a SAN are a shared resource for attached processors, multiple application systems can be configured to access the data in parallel. This is especially compelling in a true read-only environment.

Daily backups to removable media taken to secure storage – We’ve known about this one for years, but now it’s time to take all of this more seriously. No matter what other protections we put in place, nothing provides the bedrock security that offline and securely stored physical media provide.

Strategically placed packet filtering gateways – We need to isolate the key servers of our data warehouse so that they are not directly accessible from the local area networks used within our buildings. In a typical configuration, an application server composes queries which are passed to a separate database server. If the database server is isolated behind a packet filtering gateway, the database server can be configured to only receive packets from the outside world coming from the trusted application server. This means that all other forms of access are either prohibited, or they must be locally connected to the database server behind the gateway. This means that DBAs with system privileges must have their terminals physically attached to this inner network, so that their administrative actions and passwords typed in the clear cannot be detected by packet sniffers on the regular network in the building.

Role enabled bottleneck authentication and access – Data warehouses can be more easily compromised if there are too many different ways to access them, and if security is not centrally controlled. Note that I didn’t say centrally located, rather I said centrally controlled. An appropriate solution would be an LDAP (Lightweight Directory Access Protocol) server controlling all outside-the-gateway access to the data warehouse. The LDAP server allows all requesting users to be authenticated in a uniform way, regardless of whether they are inside the building or coming in over the Internet from a remote location. Once authenticated, the directory server associates the user with a named role. The application server then makes the decision on a screen by screen basis as to

whether the authenticated user is entitled to see the information based on the user's role. As our data warehouses grow to thousands of users and hundreds of distinct roles, the advantages of this bottleneck architecture become significant.

There is much we can do to harden our data warehouses. In the past few years our data warehouses have become too critical to the operations of our organizations to remain as exposed as they have been. We have had the wakeup call.

I have written extensively on the above topics. The design of distributed architectures and the discussions of packet filtering gateways and role enabled security are covered comprehensively in the Data Warehouse Lifecycle Toolkit (Wiley, 1998). The application of SANs to data warehouses is described in my IE article of March 8, 2001 "Adjust Your Thinking for SANs" which can be found in the article archive on my web site at www.kimballgroup.com.