AAPG International Conference
Barcelona, Spain
September 21-24, 2003

Jess B. Kozman[1] (1) Schlumberger Information Solutions, Kuala Lumpur, Malaysia

Case Studies of Hierarchical Knowledge Management for Large Datasets

Introduction

Large geotechnical datasets in petroleum exploration have created a continuously growing challenge to traditional data storage, management, and delivery systems. New strategies are required in order for this data to be effectively utilized as an organizational asset to add to proven reserves. Analysis of best practices based on successful information management implementations at over fifty international oil and gas organizations shows that the focus of the solutions has moved from data and information to knowledge. Storage management solutions now recognize three levels of digital datasets; large volume static data files, smaller but more numerous and dynamic user interpretation files, and one-time capture and retrieval files used for snapshots and archives at key milestones and benchmarks in the life of field project lifecycle. In successful implementations, each data type is managed according to business rules derived by onsite analysis of the client's unique organizational workflow.

When disk storage systems for large datasets, especially those required in seismic exploration and development, begin to run at 100% of capacity, interpreters continuously run out of storage space for projects, costs for storage hardware exceed IT budgets, time windows for backup procedures begin to interfere with interpretation, and geoscientists start to spend their time looking for places to put their data instead of for places to drill wells. Over the last four years, many organizations have implemented hierarchical content management systems to alleviate pressures on online disk storage. Hierarchical management uses the access patterns of files to determine their movement between storage devices.

The first step is to determine the usage patterns for different data types. A typical graph of the cumulative volume of data in seismic projects not accessed during a project's lifecycle shows the distinct inflection points that represent changes in usage patterns for different types of files (Figure 1). The history of hierarchical project content management systems has shown a progression through these three levels of files, with associated increases in value added.

Case Study - ChevronTexaco

One of the first successful implementations of hierarchical content management was at Chevron Overseas Petroleum Inc (COPI) in San Ramon, California, in the United States at what is now the ChevronTexaco Upstream Technical Computing center. The management of large static trace files in the interpretation environment began in 2000, with the installation of a 5 terabyte automated tape library to manage application trace files stored on network attached storage for six worldwide business units (Figure 2). At the time, volumes created by data loaders and those created by interpreters were growing at an exponential rate. Business rules were put in place to allow the release to near line media of trace files not accessed in 32 days from over 500 seismic projects containing up to 30,000 physical files. In addition, a second copy of the archived tapes was used to provide backup and disaster recovery capabilities.

The system has since grown to over 11 terabytes and over 100,000 physical files. According to the ChevronTexaco project manager, 7 terabytes of this data exists only on near line tape, and she recently wrote, "I shudder to think of how we would have handled all that inactive data" without the near line system. She indicated the previous system involved hours of work creating offline tapes "destined to be lost in desk drawers" and would have eventually required the purchase of more network attached storage at substantially higher cost than that of tape. The near line system

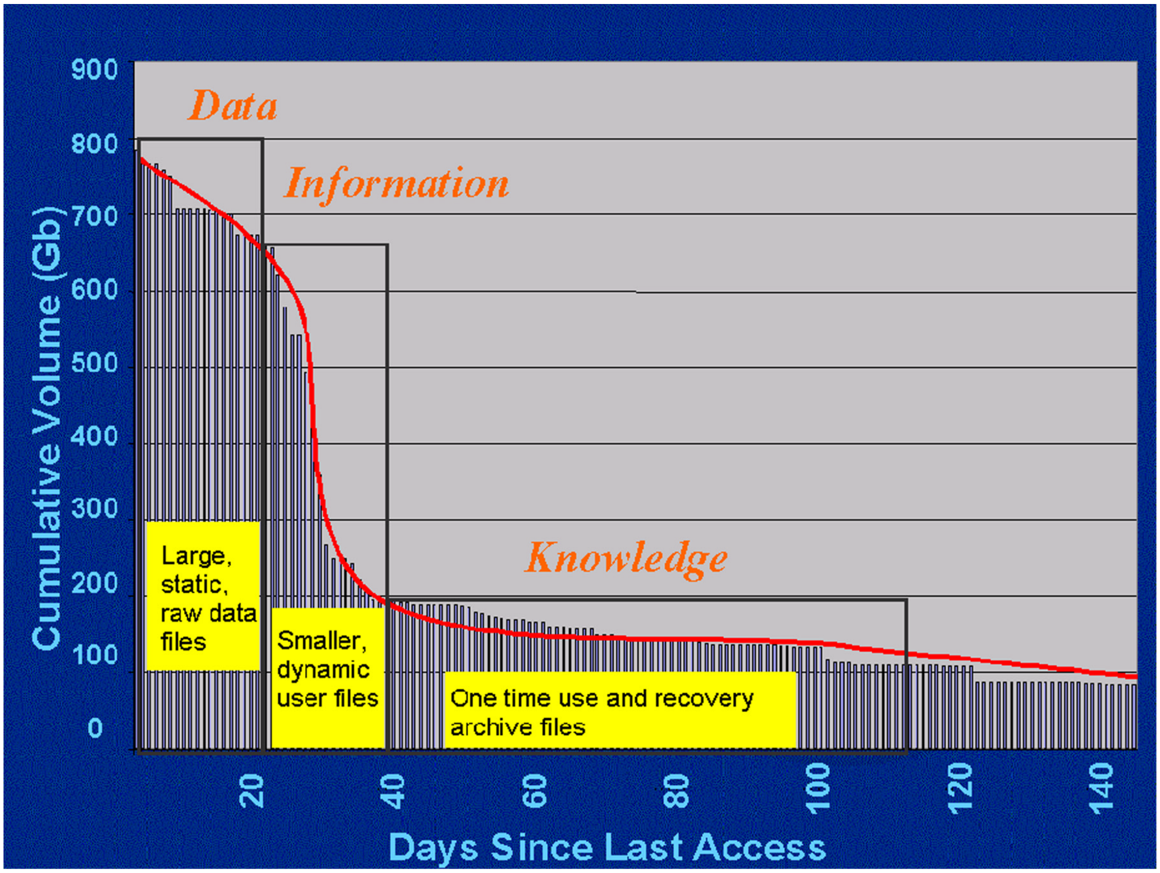

Figure 1. Graph of cumulative volume of files not accessed from disk in a typical geotechnical project data storage environment. A snapshot of file inodes determines date of last access by a user or application, and inflection points mark changes in file type and usage patterns.

also saves money and pressure on the backup system, provides a comfort level for disaster recovery, and makes it easier to retrieve files to a project than from a traditional UNIX backup system. Standard backup tapes at ChevronTexaco are kept for only three months, so projects have been "saved" by having the data on near line tapes after discovering it had been deleted completely from disk at some unknown time. Since the technical center is billed internally for backup services, taking the managed disks out of the backup schedule also saves money. ChevronTexaco has since expanded the use of the system to archiving of multiple data types.

Case Study - Schlumberger Jakarta Powerhouse

At the newly opened Schlumberger multi-client Powerhouse data management center in Jakarta, hierarchical project content management has moved from just data to information and is used to load, store, manage and deliver data from multiple datastores to clients for use in geotechnical projects. The data includes well, log, physical asset, and seismic data. Online storage capacity is augmented by using business rules to release large seismic trace files to over 2 terabytes of near line storage in an automated tape library. In addition, system backups using an incremental dump of file node metadata are stored in a separate tape library. This approach to managing information as well as the raw data with a hierarchical system provides an additional level of security. In a disaster recovery scenario, data managers could be required to provide data delivery services from new hardware at a remote location. The file node backups allow fast re-creation of directory structures on a newly configured file system, and individual files are then staged from near line tape as they are called by users or programs, with the most critical at the head of the queue. To users and applications, all files on the hierarchical virtual disk appear to be available as soon as the directories are re-established. This saves considerable time on a large file system over restoring from a conventional backup where

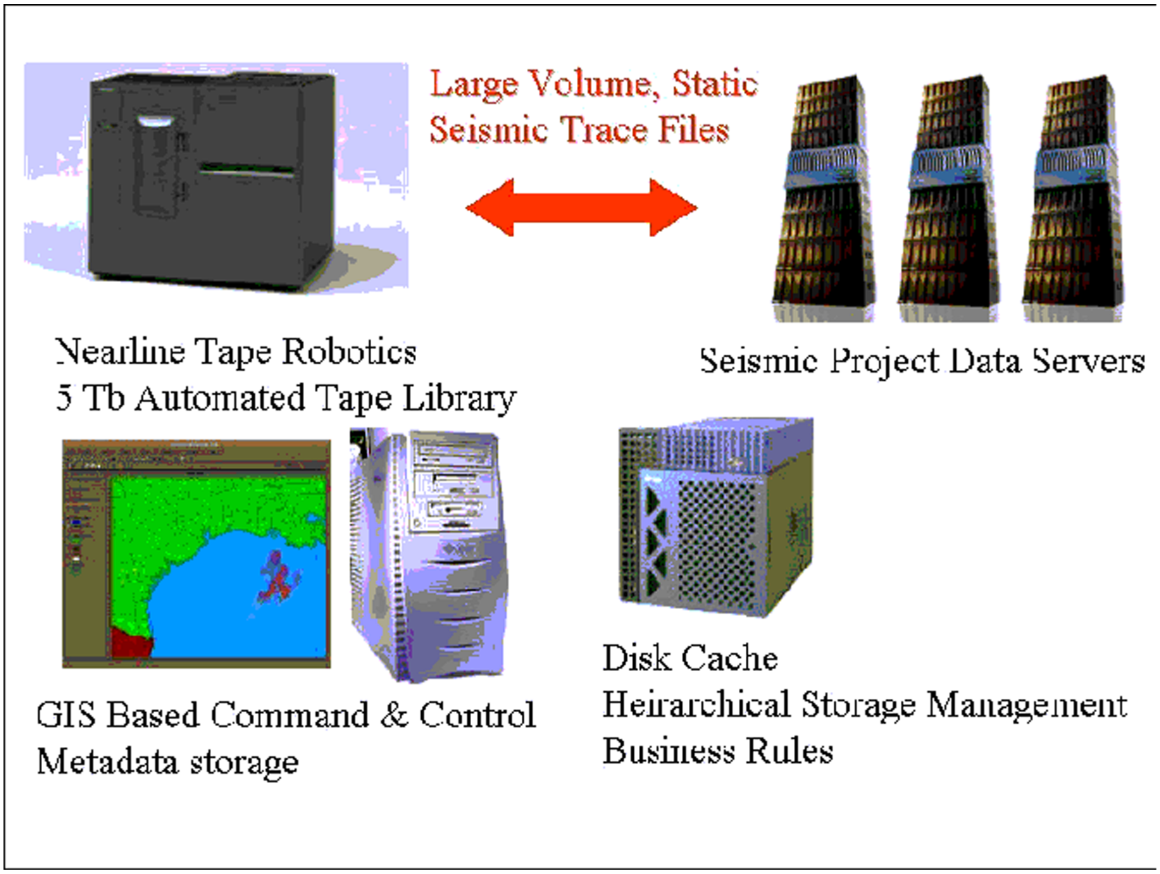
Large Volume, Static
Seismic Trace Files
Nearline Tape Robotics
5 Tb Automated Tape Library
Seismic Project Data Servers
GIS Based Command & Control
Metadata storage
Disk Cache
Heirarchical Storage Management
Business Rules

Figure 2. Schematic layout of a hierarchical management system for geotechnical project content installed at Chevron in 2000. The system was configured to take pressure off of disk storage systems serving data for seismic application projects by releasing non accessed files to near line tape media.

the entire file system must be sequentially restored before users can return to productive work.

User access to the stored client data is provided via the Web using a decision support portal implementation based on three-tier architecture with a thin client interface. Users can log in from any platform using a standard web browser to locate and order data. Interfaces provide both Geographic Information System (GIS) map based and tabular views of multiple repositories, as well as knowledge management or collaboration. Customized views can support and enforce optimized workflows and track key performance indicators relevant to the specific user, and launch applications to analyze and interpret the data using application service provisioning (ASP). The service level agreements for this complete outsourcing of data management and delivery are designed to leverage the shared overhead of the multi-client model, and provide a total cost of ownership that is both predictably based on data volumes and delivery schedules, and demonstrably less than the clients' total budget spent internally on hardware, software, personnel, and infrastructure.

Case Study - Schlumberger Information Solutions (SIS) European Service Center

A similar system at Schlumberger's European Service Center (ESC) in Aberdeen combines hierarchical storage management with redundant servers in a Wide Area Network (WAN) campus cluster configuration. The data storage architecture is built on storage area network (SAN) switches and fiber channel attached disk arrays. There is currently 16 terabytes of disk storage capacity in a RAID 0+1 arrangement at two server rooms in separate buildings, sited 5KM apart. Offsite storage of duplicate tapes from the near line tape robotic system at a separate Schlumberger site has

Figure 3. View of the online storage and automated tape libraries supporting geotechnical project content at the Schlumberger Jakarta Powerhouse, a full service multi-client data management center for loading, management, and delivery of well, log, seismic, and physical asset data. In this implementation, not only raw data but file information is moved to nearline tape robotics according to business rules.

allowed the center to meet client specifications for fail-over capabilities that cover the complete physical loss of either site. In addition, disaster recovery capabilities can be extended to cover the contingency of the destruction of both sites for only the cost of an additional set of tapes. This can be attractive for exploration organizations with multiple sites in a limited area, such as the British Isles, that could be effected by the same interruptive event. By periodically shipping a set of file node dumps and released files to a European mainland location, critical data continuity is guaranteed even with a total loss of U.K. sites. This is a critical component in providing comprehensive business continuity plans that account for both natural and created disasters.

Case Study - German Oil & Gas

The latest step in the hierarchical storage of project content is from data and information to knowledge. A hierarchical knowledge management system (HKMS) manages large static data files, dynamic user files containing project information, and archive files. Archive files are created to capture the knowledge encapsulated in entire projects at benchmarks and milestones in the life of field project lifecycle. These files are also moved between online and near line storage by their own set of business rules, leveraging the same near line hardware and media.

Such a system was installed at German Oil & Gas Egypt (GEOGE) in Cairo, where approximately 2.6 terabytes of interpretation project data is moved automatically according to business rules between online and network attached storage devices and a robotic automated tape library. Files from all three categories are identified by access patterns and segregated onto separate storage partitions and pools of tape media. File usage patterns are continuously monitored to gauge the effectiveness of the background processes and allow tuning of the system. The system provides effective storage management, backup and disaster recovery capabilities, and a method to capture and archive the knowledge contained in projects at key milestones in the project life cycle, including preparation for application upgrades.

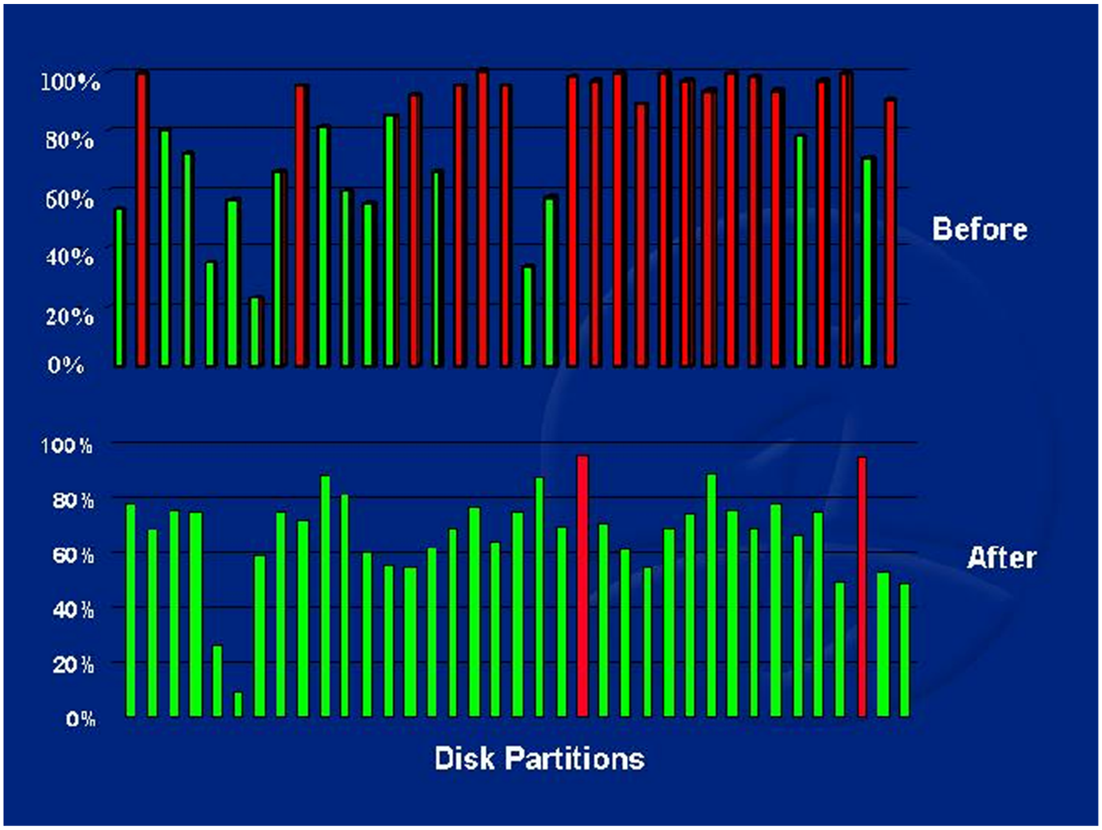

Figure 4. Disk capacity before and after the implementation of a hierarchical knowledge management solution at German Oil & Gas Egypt in Cairo. Available disk space doubled, and usage as an average percentage of available capacity dropped from 95% to less than 70%.

Previous to the installation, GEOGE had more than one terabyte of data on disk with only manual backups of project seismic data and no scheduled backups for data outside of application projects. There was no in-place disaster recovery system. Ninety software application system projects were spread over 36 disk partitions and 30,000 physical files. Most of the disks were 95 to100 percent full. Some seismic trace files in interpretation projects had not been accessed by users in up to nine months, and manually created backup files were occupying more than 200 gigabytes on a high-end, network-attached storage device. Corrupted disks required system administrators to physically move data one project at a time and reload projects from tape, resulting in days of lost work. A 15 percent growth in seismic data over three quarters was predicted, and disk usage had begun to grow exponentially. Some of the multiple versions needing to be managed included AVO and gradient stacks; tau-p filtered and scaled cubes; instantaneous phase, inversion and pore pressure volumes. A proposed integrated reservoir characterization system upgrade required a project archival strategy.

The hierarchical knowledge management solution provides a comprehensive backup and disaster recovery plan that insures security for the corporate asset represented by geophysical interpretation data. High water marks are maintained at 80% for RAID mounted disk storage, and 95% for more efficient network attached storage. Seismic trace files larger than 100 megabytes become candidates for release after 25 days of non-access. Full backups of the projects without static seismic trace data are performed every 15 days, incremental backups are run once a day, and the three most recent full backups are kept in the automated tape library while older versions are removed so the tape media can be recycled. Project archives are user initiated at project milestones or for transfer to other offices. A recent hardware failure allowed GEOGE to test the disaster recovery portion of the solution under mission-critical conditions. All files on an 85% full 18-gigabyte disk were successfully restored using the 24-megabyte per second

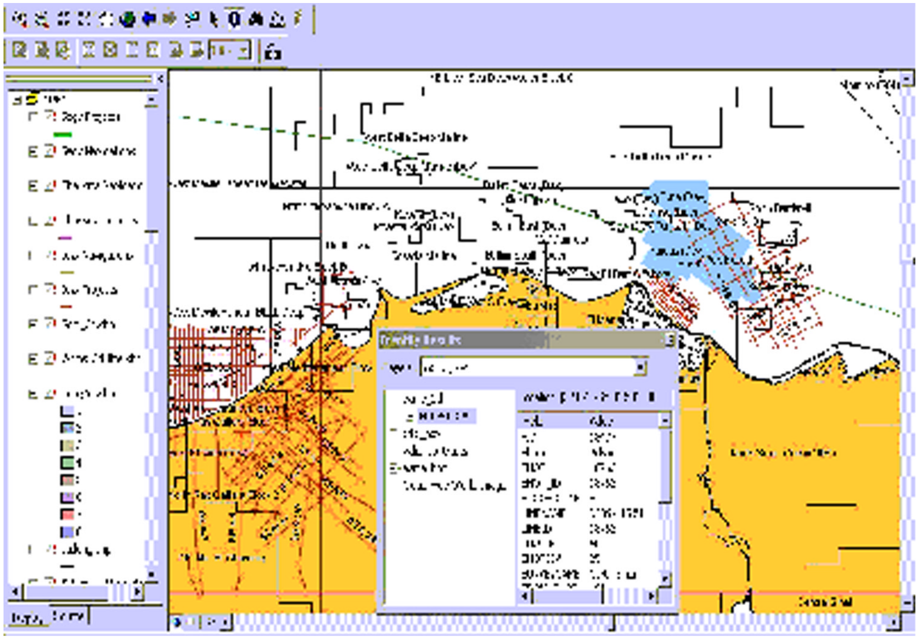

Figure 5. The GIS interface that allows geotechnical users to spatially quality control data loading and query and report on multiple versions of geotechnical data being managed between online and near line storage at GEOGE. (Seismic location data courtesy of WesternGECO)

drives to a point 4 hours prior to the failure with a single command, and according to the network administrator, "the users did not even notice the incident". Network-attached storage is now being used exclusively to deliver large-volume seismic trace files to active interpretation projects. Available disk space has nearly doubled, the average disk usage per partition dropped to less than 70 percent (Figure 4), and a reduction in disk purchase expenditures of 50% has been recommended.

In this implementation, users also have access to a map-based Geographic Information System (GIS) interface for querying and reporting on multiple versions of the data, and data administrators and loaders have an automated workflow for analyzing, validating and cataloging incoming processed seismic data to support exploration decisions and create new trace files in their application projects. The GIS interface allows a single map view presentation of all geotechnical data from multiple datastores including native format seismic on disk, interpretation application format files in seismic projects, navigation for seismic not currently loaded to the system, and well, lease block, and cultural information spatially registered to the same base map (Figure 5). A workflow interface enforces quality control steps such as merging native format seismic with validated navigation points from the corporate data store, visually checking the loaded navigation for spatial errors such as projection and datum shifts or transposition of 3D inlines and crosslines, and accurate determination of scaling and clipping factors applied when loading to interpretation projects. The GIS view provides map or table based searches and queries for drill down mining of metadata about multiple versions of seismic loaded to projects. A web based interface allows users to determine how much data is online or near line at a given moment, and select volumes to be restored for online disk use.

The entire system provides mission critical decision support for GEOGE's drilling by allowing geotechnical users to answer questions such as "What data is available over my area of interest?" "What data has already been loaded to this or other projects?", "What processing flow was applied to the data I am now interpreting?", "Can I be sure my data is loaded in the right place?", and "Am I using the most recent and validated version of the full data set?"

Conclusions

There is sufficient history now to determine trends and best practices in the area of hierarchical management of geotechnical project content. Implementations have progressed from managing only large static data files, through managing information, to a final stage of managing three distinct types of files including those containing the encapsulated knowledge in a geotechnical project. Successful initiatives have focused on setting unique business rules for the transfer of these different file types between online and near line storage, continual evaluation and tuning of those rules to insure the system remains synchronized with changing user behavior, and intuitive GIS map based interfaces that allow users to query, report on, and retrieve the data easily regardless of its storage location. Now that these systems have become integrated into optimized workflows that capture the value of knowledge added to projects during the life of a field, they can be shown to be adding value directly to the bottom line of asset teams. In addition, file usage patterns indicate that another file type exists that is currently not being addressed by these systems. These can be defined as files intended to be globally distributed by a portal implementation and that represent corporate wisdom acquired at benchmarks and milestones during a project lifecycle. The next stage of hierarchical content management will need to develop business rules to address these files as well.

At organizations where these systems have been successfully implemented, they add value to exploration workflows by objectively identifying and efficiently managing geotechnical project files using tuned business rules at all phases of an exploration and development project life cycle. Documented benefits include an increase in available disk space, reduction of costs for additional disk purchase, reduction of backup cost and time windows, ease of restore operations, proven disaster recovery schemes, and reduction in time spent by geoscientists looking for and managing data. This means that asset teams can concentrate on their core competency, and instead of using time looking for data, use it to find oil and gas.