Should you replicate databases locally?

When people are working on the same data in different places around the world, should the data always be stored and accessed from a central server (like with webmail), or a copy downloaded onto a local device (like with Blackberry e-mail)? By Philip Neri, VP marketing with Paradigm

Companies managing assets over large, if not global, geographical areas often operate these assets from multiple locations, with at least one office positioned in the assets' state or country, and with activity taking place at a centralized technical location, be it the company's headquarters, regional main office or elsewhere.

For the best efficiency, it is important that each geoscientist, engineer or data administrator be looking at the same collection of data wherever they are working from, inclusive of all the most recent edits, transformations and metadata that may have been contributed at any of their locations.

Data strategies may be relatively easy to implement if all locations are in close proximity, and in a region where network infrastructure offers good bandwidth and reliability. But processes become more involved if (for example) the operating unit is periodically disconnected from a wide area network, and on a low-bandwidth connection.

Background

In the 21st century, technology has brought together distant places at a pace that was not anticipated some 15 or 20 years ago.

In the energy industry, many companies had trans-regional or international operations from the very early days of oil and gas exploration and production.

However there was not an expectation of frequent and intensive communication and collaboration, and remote or regional offices would operate as independent entities in terms of their data, their activity and their resources.

Two factors concurrently put pressure towards a change in the way operations are conducted: the increase in reservoir complexity, and the need for more advanced technologies in order to be successful.

It became increasingly unrealistic to populate each operating unit with all the highly-skilled resources that would be required oftentimes only on an occasional basis. Putting experts from a central pool onto airplanes on an on-demand basis was the alternative, but this was not a very effective use of such high-value persons' time. The solution that emerged as the most effective was to have teams working both on location, close to assets, and in one or more central facilities where specialists can cover a wide diversity of specialized tasks for numerous field locations. Satellite links, video-conferencing systems and improving infrastructure made it easier and easier for such geographically-separated teams to work together.

Satellite links and video-conferencing systems made it easier for geographically-separated teams to work together

With two or more teams working on different aspects of a same asset at the same time, the one element that still needed to be addressed was the synchronization of all the data being worked on such that at all times each team would see and be able to use information, knowledge and results emanating from the other team.

The two main options for how people work with data are (i) multiple views on a single database (like looking at your hotmail from anywhere in the world); or (ii) replicated databases with synchronization (like how your Blackberry can download your e-mail, but the main e-mail database is somewhere else).

Single database multiple views

With a single database multiple views architecture, all the data, the applications and the compute power are located in a centralized server facility.

Users log in remotely, and execute the applications remotely. The graphical interface and the display of results is transferred over the network to show up on the user's screen, wherever he or she may be.

From a data management point of view, the only vital requirement is a very rigorous management of data ownership. As a project moves forward, the interpretations and results generated by each user must be clearly labeled as such. There must be flexible capabilities to establish and enforce policies that define who can see, re-use, edit or delete different types of data created by other members of the geographically-distributed team.

This centralized architecture is of course very dependent on the quality, speed and reliability of the network, and for mission-critical activities many companies still see a risk factor for remote users either operating on or close to drilling facilities (i.e. away from major urban areas), or for operations in developing countries where network infrastructure is not yet fully established and reliable.

Replicated databases with synchronization

But many companies choose to have a copy of the complete project dataset running locally, in order to shield each operational location from any wide area network (WAN) interruption or performance degradation.

They end up replicating many hundreds of megabytes, or even terabytes, of data multiple times.

The relatively low cost of storage makes this a reasonable choice when set against tight work schedules and other business priorities.

The challenge is to ensure that at all times updates made at any one of the operational locations becomes available to the other locations within a reasonable time-frame.

While any granularity of an updating process would be envisaged, in most cases a daily refresh is considered satisfactory, especially if the locations are linked by weak network connections.

The data management infrastructure must then run a synchronization process that will make it possible to compare activity on the different versions of the dataset since the last update took place, and perform an update of all instances such that once the synchronization process has run all the most recent edits, modifications, new objects or other changes to data are reflected and identical in each place.

Portable devices

With increasingly high-powered portable computing resources, many companies now entrust field geoscientists and engineers with increasing amounts of data and sophisticated applications that allow them to integrate new data, perform comparisons with initial models or older data, review surrounding or regional data, and perform other tasks that involve both large data collections and substantial compute capabilities.

Field locations, often at a distance from major communication networks, favor a stand-alone operational model where the portable computer is assumed not to be connected to any remote resources.

On returning to a more substantial facility, there will be the same need for synchronization between the portable device and the company database.

Data ownership

In all scenarios, remote access to shared systems, synchronization of replicated databases or dealing with offline portable devices, probably the most important feature of a data management infrastructure is the ability to label and control the access / edit permissions of each data item for individual users or groups of users.

Rigorous control systems are often perceived as inhibiting productivity, but on the long run they actually act as a facilitator in ensuring project data integrity and in solving inconsistencies in results.

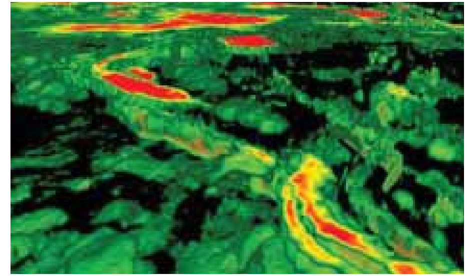
For example, if two or more users are editing a same data instance, it is better that they each work on their version of the item, while being able to see the progress their colleagues are making.

Selective copy and paste utilities can help take on and leverage the work of others, but each version is under the complete control of the respective users.

Synchronization processes will keep all sites updated, but will not result in overwriting any one user's data with the version or versions of other colleagues.

At some point, when work is close to completion, decisions can then be made to merge different user contributions into a single final entity.

Should the final version later be challenged, it is possible to go back to the individual contributions and check to see if the best choices were made when merging.



Modern high-quality data is rich in information; this illustration of a 3D rendering of a channel system is of interest to many disciplines: geology, sedimentology, stratigraphy, reservoir characterization and drilling engineering

Maturity

The management of terabytes of geoscience and engineering data over the many years of an oil and gas asset's life cycle is critical in order to ensure that the accrual of knowledge and information is always accessible and put to use for future operations.

The maturity of the data infrastructure that supports the asset, and its alignment with the business model of a distributed global industry, will impact significantly the efficiency of the teams working towards optimal exploration and production outcomes.