

Mass Spectrometry Repository

Customer

A leading biotechnology company running a large biological data factory utilizing high throughput mass spectrometry to discover information both for external clients and internal use.

Business Problem

The customer owns and runs a data factory, analysing thousands of samples on the most advanced mass spectrometry machines. The data factory runs 7 days a week, producing mass spectra which are searched against various public domain and proprietary sequence databases.

The management of the vast quantities of data produced, allowing efficient searching and long-term storage and recovery, has rapidly become a major task. With the new generation of mass spectrometry machines promising thousands of times faster data production, a major project was required to ensure that the IT systems were not a bottleneck.

Tessella Solution

A Tessella Project Manager joined the customer's IT team. Initial analysis of the problem showed that the customer might, with the new technology, be producing between 2 and 50 Terabytes of data per year. Of this, approximately one third was required on-line for comparison against sequence databases, and the remainder (the raw files) had to be stored permanently in support of drug applications and patents.

The early analysis also showed that the company needed to develop a cradle to grave data maintenance process. Data had to be captured to a safe storage area from the instrument PCs, checked for consistency and

completeness, then banked into a long-term storage area with controlled access.

The solution designed has several components, to address the different demands. Each instrument runs an application writing MD5 checksums of the data files as they are produced. An automated Data Capture system collects files when they are stable and transfers them, with consistency checking, to a Unix SAN based system. Information about the data is inserted into an Oracle database. Client tools are provided to permit the MS supervisors to resolve any data inconsistencies, and automated scripts then bank the data into the Repository database, from which the customer's search and analysis programs can access the spectra. The raw data files are bundled up and placed in a filesystem, designed to support the billions of files that will eventually be produced. Whenever a spectrum is used, the location of the related raw files is known. The use of MD5 checksumming throughout the process ensures that data recovered is always a true record.

Results and Benefits

By using an experienced PM, the customer reduced the risks associated with such a large data management project. The Repository database will be one of the 10 largest databases in the world within three years.

The problem has been fully identified, and addressed in manageable chunks, with the added benefit that existing staff with appropriate skills have been used very efficiently as temporary members of the team.