

Coping with the data explosion

Stephen Rutherford, Eldas Project Leader, Edikt, details solving data access and integration problems in the 21st Century...

The huge increase in the volume of data produced in research and commercial environments in the last decade has led to the introduction of such phrases as 'data avalanche' and 'data explosion'. The explosion of data has been as a result of improving digital technologies. These improvements include more powerful computation resources, faster data streams and more storage resources. In many cases, the data is distributed across international organisations, and both the number and size of these globally distributed data sources is growing very quickly. While this expansion is led by science, all communities – whether they are commercial, public sector, healthcare, engineering or entertainment – are following suit.

This data is not only on a massive scale. It is also heterogeneous and dynamic. Its distributed nature means that it needs to be shared across both organisational and geographical boundaries, and storage tends to be independently managed. All this makes data processing increasingly challenging. The ease and speed with which this data is generated and changed also makes it increasingly difficult to ensure its quality.

Why is all this data so important? Data can be thought of as simply a collection of facts. But without such facts, it is impossible for individuals or organisations to make decisions, draw conclusions, or undertake any analysis or calculation. Data is hence evidence, and this evidence is the fulcrum around which critical business decisions are made and research conclusions are drawn.

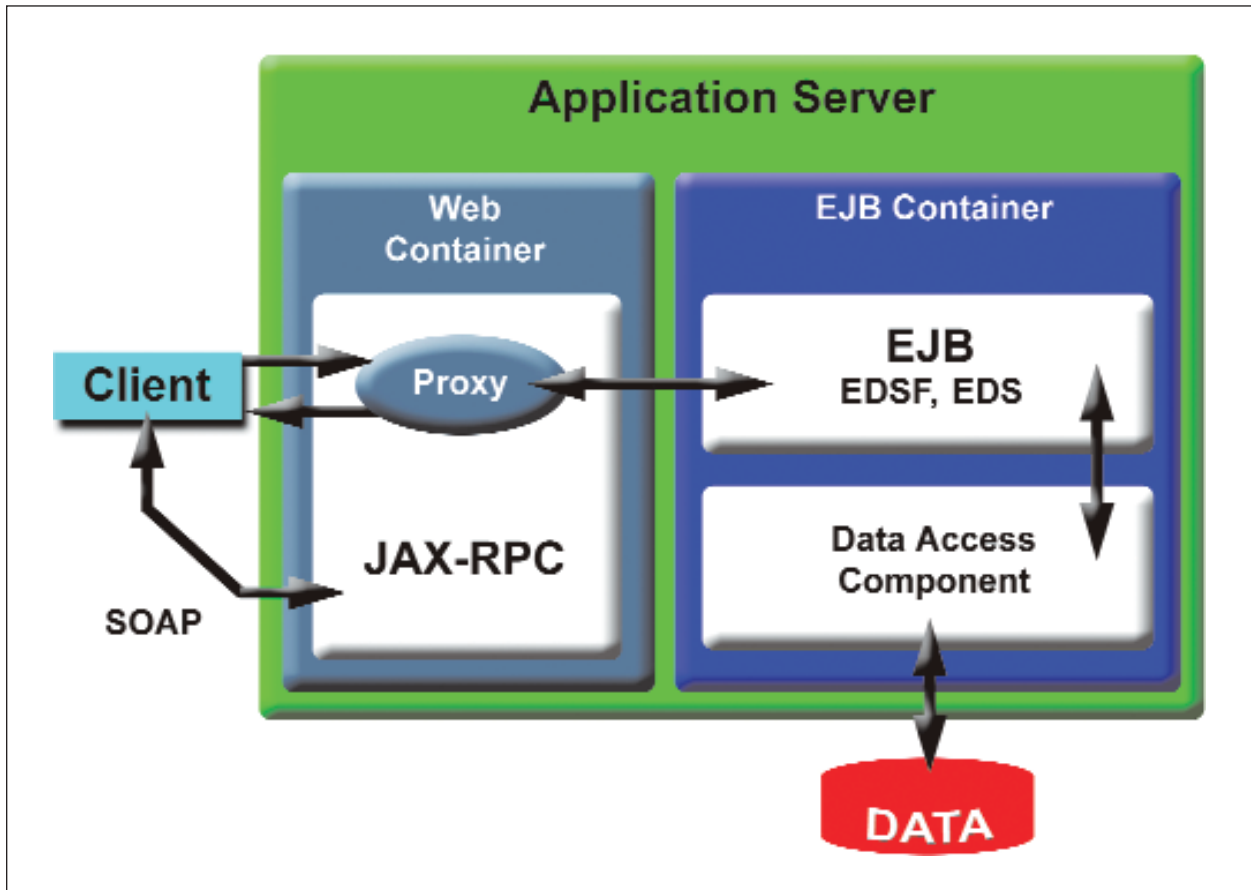
On an individual level, creativity, imagination and effort are required to make sense of recorded evidence. The very same disciplines are needed on an organisational level, but both the body of evidence and the resources available to gather and make sense of it are on a larger scale. In the electronic age, this scaling is amplified enormously and the evidence itself tends to be much more widely distributed. Digital data itself is, of course, little more than a collection of uninterpreted bits and bytes. But if this data can be equipped with meaning, it then becomes information. And if then used to make a decision, or solve a problem, this information becomes knowledge. The data explosion presents great challenges in finding the nuggets of

information and knowledge buried under this avalanche of bits and bytes. Hence, reliable tools to access and integrate legacy, raw and derived data, and manage its transformation into knowledge, are in high demand.

'Digital data itself is, of course, little more than a collection of uninterpreted bits and bytes. But if this data can be equipped with meaning, it then becomes information.'

All business and academic communities need to be able to make informed and knowledgeable decisions. The better they understand their data, the better these decisions are. The different groups needing to make these decisions are almost as numerous, varied and widely distributed as the data itself:

- Astronomers want to combine data from infrared, ultra-violet and radio scanners to paint a complete picture of the Universe;
- Financial analysts want to integrate data from a bank's mortgages and loans department with credit ratings from a newly acquired division. Another challenge is accessing legacy data as a result of mergers;
- Aircraft manufacturers run whole system simulations. This requires vast amounts of aerodynamic data for different components, such as fuselage, wings and engines. The wide distribution of this data is compounded by the fact that modern aircraft manufacturing is a collaborative, international concern with different components often constructed by different companies;
- Supermarket chains, telecommunications companies and internet vendors are all organisations wanting to maximise revenue. This is achieved through the gathering of transactional data leading to better knowledge of their customers;
- In the field of bioinformatics, there has been an exponential growth in the amount of research data. This has been partly fuelled by genetic studies, where the



relevant data may be distributed across continents, presenting a significant integration challenge.

'Grid computing has emerged as a major development in recent years. The UK Government is investing £250m over five years in the development of Grid technology and the United States is investing over \$500m in 2004.'

The process of transforming data into knowledge requires access and integration. These are problems common to all disciplines across academia and commerce. Accessing the data is patently the first stage. With the increasing ubiquity of the internet and the globalisation of data, however, integration is the next. If data sources can be successfully federated and geographical boundaries virtualised, then access and integration becomes transparent to the user. Specific data resources simply become 'virtualised data', their physical locations irrelevant, accessible and analytical on request. Naturally, however, there are some specific complications, including:

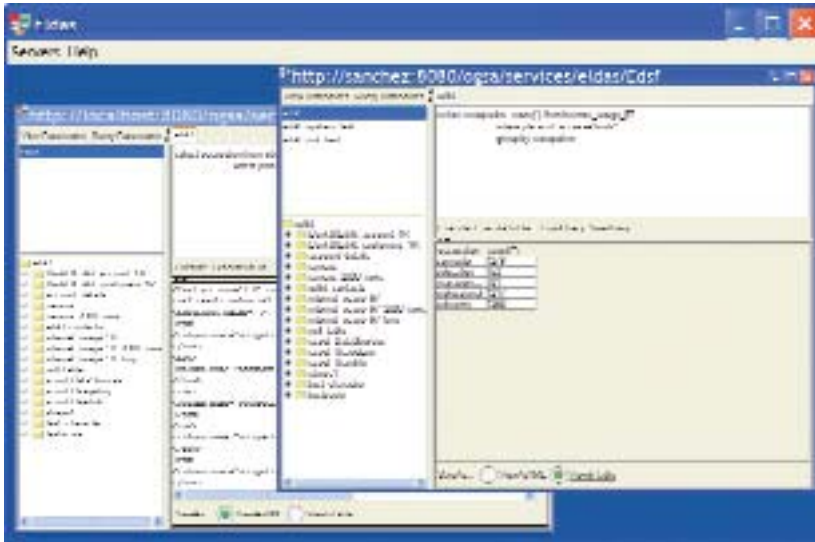
- How is security handled when integrating data from a combination of public, project and private level sources?
- How is integration of heterogeneous data sources handled?

- How can multiple, distributed databases be accessed using a single query?
- How is inconsistent and tarnished data handled?
- How is 'updated' data handled?
- How is all this made 'user-friendly'?

All of this requires automation. Automation requires the correct framework and architecture, the correct design and the correct tools.

Where science leads, commerce will follow. The e-science community consists of scientists using and developing software tools and applications to support collaborative research involving very large datasets in diverse disciplines such as astronomy, biological sciences, physics and geosciences. To quote John Taylor, Director General of Research Councils, Office of Science and Technology, "e-science is about global collaboration in key areas of science and the next generation of infrastructure that will enable it. e-Science will change the dynamic of the way science is undertaken".

The Grid has become the first infrastructure for e-science. The Grid (analogous to the electrical grid) envisions on-demand computing, where anyone, anywhere, can access whatever computing power and data that they need, so long as they can 'plug in'. Grid computing has emerged as a major development in recent years. The UK Government is investing £250m over five years in the development of Grid technology and the United States is investing over



\$500m in 2004. Building the infrastructure to support this vision is a complex problem, requiring several co-ordinating bodies and international collaboration.

Edikt (e-Science Data Information & Knowledge Transformation) is a project based at the National e-Science Centre, University of Edinburgh, UK. Its remit is to provide software solutions for the e-science community using both existing and emerging technologies. The focus is encompassed by Edikt's very name – to take e-science problems, and convert their electronic data into information and, subsequently, knowledge. This transformation process is provided by Edikt's software solutions. Within this remit, the project's objectives are to:

- Transfer existing research to scientific software applications;
- Exploit existing industrial software technologies;
- Encourage commercialisation of novel techniques developed.

Ultimately, this should allow the development of generic spin-off technologies that may have commercial applications beyond scientific research. For this reason, Edikt has a dedicated commercialisation manager who can push out the knowledge, expertise and software to industry and business.

Eldas overcomes many of the problems of access and integration arising as a result of the explosion in heterogeneous and distributed data. Data integration that currently takes months or years will be compressed to a fraction of the time, with enormous scope for reducing costs.

Dr Mark Parsons, Commercial Director of EPCC and a member of the Globus Alliance, said: "Eldas is the first bridge between the Grid for science and the Grid for business. It delivers truly vendor neutral data access and integration services in an easy to install package. It gives business and science the opportunity to leapfrog their

competitors by transforming their ability to extract and analyse data from across their organisation."

Edikt have successfully addressed a number of issues that could have prohibited wide adoption of the Eldas software. Unless software is easy to install, deploy and use, then people will look elsewhere. Consequently, a great deal of time and effort has been invested in ensuring that Eldas is quick and simple to deploy, and easy to use. Robustness is also absolutely critical in ensuring that Eldas retains and extends its user base. Hence, Edikt have been developing robust software as a result of high software

engineering standards. The heterogeneity of data sources is such that Eldas has been carefully designed to provide access to data such as Relational Database Management Systems (RDBMS), XML databases, ASCII files and binary files. The distributed nature of the data means that data sources are inevitably hosted on a variety of machine architectures and operating systems. Consequently, Edikt have ensured that Eldas is truly independent of the machine in which it is to be deployed. To ensure that data sources are widely accessible, it is advantageous to support as many access methods as possible. For this reason, the Eldas architecture has been designed to adapt easily to different interfaces, allowing different methods to be supported.

Edikt is a collaboration between the University of Edinburgh and the University of Glasgow, and works closely with EPCC and the National e-Science Centre (both located in Edinburgh). Grid-based data access was pioneered by EPCC at the University of Edinburgh. Edikt is funded by the Scottish Higher Education Funding Council (SHEFC). Eldas is released under a freeware licence, allowing unrestricted academic use but commercial exploitation must be authorised. Downloads and details of current and future functionality, including the supported data resources and interfaces, can be found at www.edikt.org/eldas. Technical publications can be found at www.edikt.org/publications/eldas/pub.htm.

Stephen Rutherford
Eldas Project Leader

Edikt
Old College
South Bridge
Edinburgh EH8 9YL

Tel: 0131 651 4038

info@edikt.org
www.edikt.org



edikt