

CRITIKAL

ESPRIT Project Number 22700

Attar Software

Gehe

Lloyds TSB Group

Parallel Applications Centre

University of Stuttgart

D1.10 Consolidated Management Report (Public version)

PAC/CRITIKAL/D1.10 Public Version 1

Chris Scott, Parallel Applications Centre
Akeel Al-Attar, Attar Software
Wolfgang Schneider, GEHE
Dave Nisbet, Lloyds TSB
Tilmann Barth, University of Stuttgart BWI
Holger Schwarz, University of Stuttgart IPVR

5 March 1999

Executive Summary

This report analyses the achievements made in the CRITIKAL project against the original objectives and success factors. Most objectives have been achieved. We conclude that the project has been a success in most respects. However, there have been some areas where the envisaged results have not been achieved. These are discussed.

The technical results of the project are described and the dissemination activities summarised.

The significance of the project is analysed.

This document may be copied and distributed freely in its entirety, provided that due credit is afforded the CRITIKAL Consortium, and all copyright notices remain.

Contents

Executive Summary	1
Contents	2
1 Introduction	3
1.1 General	3
1.2 Structure of the Report	3
1.3 Abbreviations Used in the Report	3
2 Management report	4
2.1 Introduction	4
2.2 High-level objectives and background	4
2.3 Changes of context	4
2.4 Achievements and results	5
2.5 Significance of achievements and results	12
2.6 Commercial exploitation	13
2.7 Management	14
2.8 Deliverables	14
2.9 Publications and dissemination	14
3 Exploitation plans	17
4 Conclusions	17

1 Introduction

1.1 General

This report is the public version of the final report of the CRITIKAL Project (ESPRIT 22700). The project involves the following partners: Attar Software (UK), Gehe (D), Lloyds TSB Group (UK), PAC (UK) and the University of Stuttgart (D).

1.2 Structure of the Report

This document is based on the structure suggested in the EC's "FINAL REPORT OUTLINE" document (ref. Finrep.doc: final report outline – April 1994 -).

1.3 Abbreviations Used in the Report

The following abbreviations are used throughout the report:

SOP: Start of the project

PM1 etc: Month 1... Month 24 of the project

Q1 etc: First quarter of the project ... eighth quarter of the project

Deliverable reference numbers are taken from Table 5.1 in the Technical Annex.

2 Management report

2.1 Introduction

Section 2 reports on the main issues of the project and covers

- the objectives of the project and the background to it at the time the project commenced;
- the major changes in context which have occurred during the project lifetime;
- the major achievements and results of the project;
- the significance of the achievements and results for the partners, the consortium, and for European industry in general;
- a summary of the commercial exploitation opportunities that have resulted from the project;
- a summary of the major dissemination activities and publications resulting from the project.

2.2 High-level objectives and background

The basic premise for the CRITIKAL project was that there is a real business need for tools to enable *in situ* data mining against large databases in a client-server environment.

The high level objectives of the CRITIKAL project were to:

- demonstrate the potential for adding value to the information assets of organisations in different sectors through effective client-server induction from large corporate data warehouses;
- develop and demonstrate an advanced client-server induction system capable of supporting efficient, effective data mining of large databases in business environments;
- generate prototypes of a series of enhancements to Attar Software's XpertRule Profiler product which are enabled by HPC technology and which are commercially exploitable during the lifetime of the CRITIKAL project; and
- generate a body of large-scale data mining experiences, from both the financial and pharmaceuticals wholesale sectors, which will form the basis for generic dissemination and specific marketing activities.

To meet the CRITIKAL project objectives, a key technical objective was the development of an innovative 3-tier rule induction architecture. It was proposed that such an architecture would flexibly support several client-server processing models, and thus be deployable in a wide range of user situations. The architecture would provide scope for optimising mining performance, particularly where the data to be mined is held on parallel HPC platforms.

2.3 Changes of context

As the project was being set up in 1996 the emerging data mining industry was primarily selling algorithms and essentially standalone tools. At the outset, CRITIKAL

encapsulated the vision of making data mining more deployable in a corporate environment by enabling algorithms and tools to be delivered in a way that fits within, and complements, existing IT infrastructures.

Early in the life of the project it became evident that implementing data mining systems in a way that fits with corporate IT infrastructures would not be sufficient to fully address the requirements for corporate adoption of data mining. An additional requirement was identified for a reusable data mining process to enable data mining projects to be planned and executed with confidence.

BWI led the task of developing a data mining process. This process was based largely on the accumulated experience of Attar Software and Lloyds TSB and on information sourced from the literature. After this decision was taken and the work was well underway, the CRISP-DM project aimed specifically at developing a tool independent process for data mining was started.

At the outset of CRITIKAL only IBM among the mainstream IT industry players had a serious data mining offering. During the lifetime of CRITIKAL, data mining has become a component of the offering from other major suppliers such as Business Objects, NCR, and SAS Institute, but the marketplace currently has no dominant players. There has been a steady growth in the number of small and start-up data mining suppliers and there has been an increasing awareness among potential users of the real potential of data mining. During the lifetime of CRITIKAL, therefore, there has been no step-change in the shape or structure of the marketplace for data mining.

2.4 Achievements and results

2.4.1 Analysis

Table 1 contains a comparison of achievements and results against the detailed objectives, envisaged project results and success measures contained in the Project Programme. In the public version of this report some analysis has been removed for purposes of commercial confidentiality.

Objective	Status
1. demonstrate the potential for adding value to the information assets of organisations in different sectors through effective client-server rule induction from large corporate data warehouses	GEHE and Lloyds TSB acknowledge that the technical feasibility and potential for adding value has been demonstrated by the prototype developed in the project.
2. develop and demonstrate an advanced client-server induction system capable of supporting efficient, effective data mining <i>in situ</i> of large databases in business environments	Achieved
3. contribute to the establishment of best practices associated with the deployment of data mining systems and the deployment of data mining results	Achieved
4. generate a body of large-scale data mining experiences, from both the financial and pharmaceuticals wholesale sectors, which can form the basis for generic dissemination and specific marketing activities	The project has contributed to the body of experience concerning the management of 3-tier client-server architectures in general and 3-tier client-server data mining systems in particular. The project has also contributed to best practice in assessing data

	<p>exploitation systems in data warehousing environments.</p> <p>While the CRITIKAL system has had some exposure to business users, the project has not led to the anticipated volume of experiences from a business perspective.</p>
5. generate prototypes of a series of enhancements to Attar's XpertRule Profiler product set which are enabled by HPC technology and which are commercially exploitable starting during the lifetime of the CRITIKAL project	Achieved
6. end-users to add value to their investment in corporate data warehousing strategies for decision support	GEHE and Lloyds TSB acknowledge that the technical feasibility and potential for adding value has been demonstrated by the prototype developed in the project.
7. be able to understand and quantify the business benefits obtainable from the deployment of data mining in support of specific business functions within their respective companies	<p>Lloyds TSB have established a central Data Exploitation and Data Mining Practice. Projects undertaken by the Practice are supported by a business case. The CRITIKAL data mining process emphasises the need to identify and understand the business benefits of a specific data mining project.</p> <p>GEHE are not yet at the point where they are seeking to implement data mining within the business. This point is likely to occur during 1999. CRITIKAL has given GEHE a good qualitative understanding of what is achievable from data mining, but not the quantitative understanding the would be required for the development of a business case.</p>
8. understand the productivity, organisational and strategic implications of different ways of deploying data mining within their respective companies	Both end-users have gained insight into the organisational structures and architectural and systems implications of implementing data mining. It is too early to formulate any substantial analysis of productivity.
<i>Envisaged results</i>	
9. proven prototype of a 3-tier implementation of Attar's Profiler rule-induction system capable of flexibly addressing the data mining needs of large corporate users. The 3-tier architecture will flexibly support a variety of client-server processing models to enable user organisations to implement the technology in a way which best suits their specific user load profiles and IT infrastructure.	Achieved
10. query optimisation methodologies and software to improve the response times of queries issued against parallel RDBMSs by the decision tree builder	Methodologies for optimising performance against Oracle and Teradata have been developed, and experiments conducted. The identified optimisations have not all been implemented into the CRITIKAL system.
11. generic optimisation methodologies and software for decision tree construction against large databases. This will include the development of intelligent software to reduce database accesses by implementing decision tree construction algorithms which use only a limited (sampled) subset of the database for	Methodologies for optimising performance against Oracle and Teradata have been developed, and experiments conducted. The identified optimisations have not all been implemented into the CRITIKAL system.

<p>node branching decisions at, and close to, the root node, and which use the whole database in situ in order to maintain decision tree accuracy near to the leaves.</p>	
<p>12. example tree-node evaluation server based RDBMS query generation modules (for Teradata and Oracle) enabling the SQL which is generated during decision tree construction to exploit the specific performance features of particular RDBMSs</p>	<p>Methodologies for optimising performance against Oracle and Teradata have been developed, and experiments conducted. The identified optimisations have not all been implemented into the CRITIKAL system.</p>
<p>13. a software framework enabling the specification of virtual (composite) attributes and their use in the construction of decision trees</p>	<p>DataPrep is a project output that enables composite attributes to be specified and generated. DataPrep demonstrates the functionality and look and feel that is required of a 3-tier client-server tool. However, DataPrep does not work against data in situ or in a 3-tier architecture.</p>
<p>14. a software framework enabling the analysis of time series (transactional) data using rule-induction data mining and association rules discovery</p>	<p>Support for the analysis of transactional data is provided by the association rules discovery technology. DataPrep also provides a demonstration of the support required for data transformation of time series data.</p>
<p>15. software enabling the generation of the profiles which result from rule-induction as both SQL (enabling their reuse) and as OLE2 objects (enabling their use by other client desktop packages)</p>	<p>Profiles can be deployed as both SQL and Windows Metafile objects.</p>
<p>16. technological developments will be tested and demonstrated against the specific application requirements of the end users in the CRITIKAL consortium</p>	<p>Achieved</p>
<p><i>Success measures</i></p>	
<p>17. Subjective measures of value to individual users will be made</p>	<p>Both end-users have held workshops with users that have yielded subjective assessments of value.</p>
<p>18. show that the resulting technology can be embedded in the decision making processes of the end-users and that it can be effectively used to aid the creation of business solutions. Issues such as tool usability and user acceptance, user perceptions of performance associated directly with operations that specific users wish to carry out, ease of deployment, efficiency of resource consumption, networking demands, reliability, and integrability with existing infrastructures will all figure in the overall judgement as to whether the specific technological results of the project are successful or not.</p>	<p>GEHE and Lloyds TSB acknowledge that the technical feasibility and potential for adding value has been demonstrated.</p> <p>Issues associated with usability, performance, efficiency of resource consumption, networking demands, reliability, integrability with existing infrastructures have been addressed with positive results. However, it has not yet been possible to provide convincing evidence of impact in a business context.</p>
<p>19. be able to identify tasks for which the data mining tools are better suited than existing tools, or tasks which are made possible through the use of data mining tools, but are not possible with other tools</p>	<p>Deliverable D2.3 Guidelines for the Data Mining Process contains a suitability checklist.</p>
<p>20. be able to assess and quantify the costs associated with licensing, installing, maintaining, managing, resourcing and</p>	<p>Achieved for installing, maintaining, managing, resourcing and training via the data mining process.</p>

training for the tool deployment	Not achieved for licensing costs.
21. be able to assess how much of a practitioner's time must be spent using such a tool in order to have a reasonable expectation of useful outcomes	The DM process described in D2.3 provides a basis for estimating human resource requirements for DM projects.
22. there must be clearly identifiable and quantifiable potential business benefits. These benefits could be in terms of speeding up relatively mundane decision making processes, or a demonstrated capability to aid the identification of rare, but significant, "competitive-edge" niche market opportunities, or by a widely agreed perception that the tools were an aid to the stimulation of creativity in decision making processes.	The types of problem appropriate for data mining and the possible outcomes from a data mining approach are widely understood. It is also widely understood that successful data mining projects must be driven by well defined business goals.
23. A successful, widely deployable data mining tool will need to show a high degree of support for inexperienced or occasional users	A usage model that has gained widespread acceptance is one where a relatively small number of specialists in an organisation create models based on specific business scenarios. These specialists have a insight into specific areas of the business and deep knowledge of the data. The models that they create are then embedded in vertical solutions for specific business users. Typically, there are many more users of the results of data mining than there are specialists creating results. While the usage model described above is attractive, most users have not yet deployed data mining so widely. Vertical, embedded data mining solutions are still some time away for most end users.
24. experienced power users must have the possibility to adjust the parameters and options in controlling the data mining process	Achieved
25. Specific requirements and measures of success associated with particular business functions will be identified and described by the end-users as part of the requirements capture process	Achieved. The DM process emphasises the importance of this.
26. one universal requirement will be the capability for confidence level or quality of results information to be supplied to users	Information on statistical significance is available in both the RI and ARD approaches. The DM process includes a validation step. As implemented, the CRITIKAL system does not provide support for integration with metadata or data quality measures.
27. the availability of exploitable project results in the form of HPCN enhancements to Attar's induction technology and new association rules discovery technology	Achieved
28. agreement with the end users in CRITIKAL on best practice for deploying client server data mining technology and deploying the results of data mining	Achieved

Table 1. Comparison of achievements and results against the detailed objectives, envisaged project results and success measures

2.4.2 Software results

The CRITIKAL system contains several software components:

- Support for association rules discovery DLL in CAF
- Association rules discovery DLL
- TCP/IP wrapper for CAF: socket.exe
- Stub DLLs for HP-TNES and Proxy
- NT service .EXEs for Proxy and HP-TNES
- HP-TNES
- Proxy
- Administration client
- Results deployment modules - n.b. these are implemented as options within Profiler, not as standalone DLLs
- DataPrep- n.b. as developed in CRITIKAL DataPrep is a standalone 2-tier client-server tool

The role of these components in the CRITIKAL Phase 2 implementation and CRITIKAL Phase 3 implementation is illustrated in Figures 1 and 2 respectively. Figure 3 provides a detailed architectural analysis of the HP-TNES component.

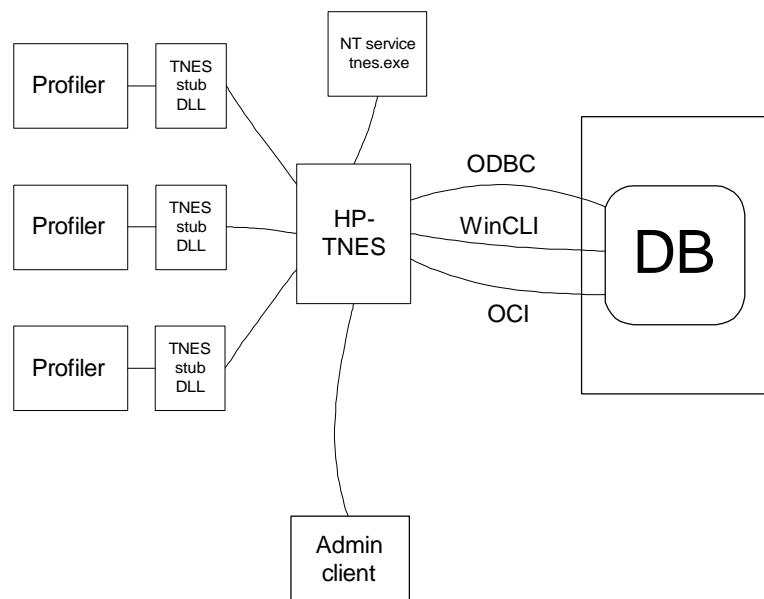


Figure 1. CRITIKAL Phase 2 architecture

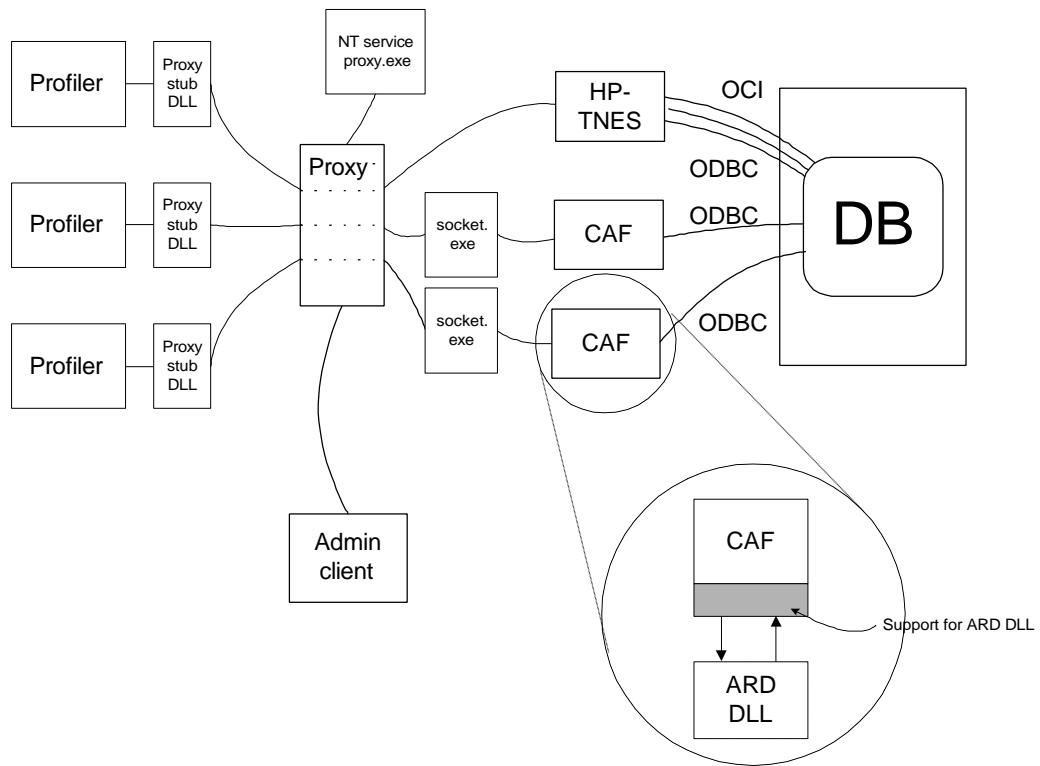


Figure 2. CRITIKAL Phase 3 architecture with expanded association rules discovery (ARD) component

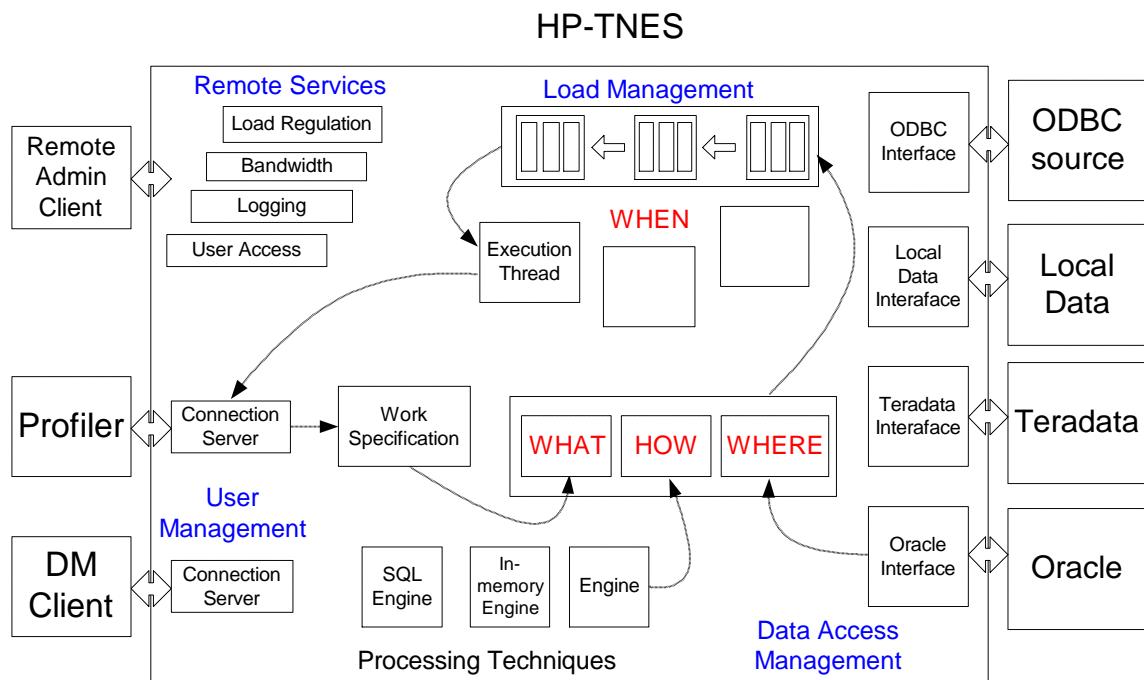


Figure 3. Detail architecture of the HP-TNES component showing an example execution path.

2.4.3 Process results

A data mining process has been developed in CRITIKAL. This is described in deliverable D2.3, which is a publicly available document downloadable from the CRITIKAL website <http://www.pac.soton.ac.uk/critikal>.

2.5 Significance of achievements and results

2.5.1 CRITIKAL partners

Attar Software

The results of CRITIKAL have played a significant role in shaping the next generation of data mining products from Attar Software. The CRITIKAL architecture is now fundamental to the architecture of Attar's products, the CAF has already been productised, CRITIKAL user input has shaped the front-end/usability characteristics of the new products, and the performance optimisation results will be reused.

BWI

CRITIKAL has contributed to BWI's generic knowledge of process and its analysis approaches for process description. BWI have enhanced their reputation in this area, particularly through the publications they have made that describe their work. CRITIKAL has also contributed to BWI's specific knowledge of the data mining process.

IPVR

IPVR has enhanced its reputation for research in data mining through participation in CRITIKAL, particularly through their conference publication. New projects building on the research work carried out in CRITIKAL are being planned (e.g. a German funded data mining project combining data mining and visualisation including VR techniques, the result of this proposal are due Feb. 1999). CRITIKAL has also provided useful knowledge that is being used as a basis for data mining teaching activities.

GEHE

Prior to the CRITIKAL project, GEHE had no direct experience of data mining. CRITIKAL has enabled GEHE to understand

- the potential benefits of data mining,
- how to manage the introduction of data mining,
- how to assess data mining tools,
- the training requirements associated with introducing data mining,
- how to implement data mining projects, and
- how to exploit the results of mining activities.

While CRITIKAL has not led to the immediate uptake of data mining, this is due to where GEHE are in their data warehousing project lifecycle rather than any failing of data mining.

In addition to data mining related issues, CRITIKAL has contributed to GEHE's understanding of the benefits and management issues associated with multi-tier client-server architectures.

Lloyds TSB

At the outset of CRITIKAL Lloyds TSB already had substantial expertise in data exploitation processes and data mining. CRITIKAL has reinforced and augmented these processes and the knowledge associated with them. CRITIKAL has provided particularly high added value for Lloyds TSB

- by aiding understanding of the benefits and management issues associated with multi-tier architectures,
- by providing a data mining project lifecycle, and
- by giving Lloyds TSB insight into trends in the data mining industry.

In addition, Lloyds TSB have benefited from insight into the software development lifecycle for advanced data exploitation products, they have evolved existing software evaluation frameworks to take account of multi-tier architectures, and they have gained experience of the management issues associated with NT in a corporate network environment.

PAC

Participating in CRITIKAL has enabled the PAC to augment its systems design and implementation skills. It has enabled the PAC to establish strong credentials in the introduction and implementation of data mining capabilities. CRITIKAL has provided the PAC with access to a data mining process and it has enabled us to work directly with end users as they experiment with the application of data mining.

2.5.2 European industry

While the CRITIKAL brand name will not persist beyond the end of the project, the technological advances will live on. CRITIKAL will have a most direct impact on the product range offered by Attar, with this in turn enabling Attar's end user customers to deploy data mining more effectively than was previously possible.

More generally, CRITIKAL has contributed to the increasing maturity of the emerging data mining industry. With its emphasis on the balance between fit with infrastructure, process, and tools, CRITIKAL has been in the vanguard of development activities that have addressed the real needs of large companies in deploying data mining.

It was asserted in the proposal to the EC that CRITIKAL was highly innovative in its approach to implementing data mining, and that the consortium was leading the world. The emerging data mining industry has reinforced this positioning, with recent announcements by ISL (with their CHESS product) and Angoss (with their KnowledgeSERVER product and In-Place Mining Drivers) helping to validate the architecture and approach taken in CRITIKAL. The CRITIKAL consortium still believes that it is ahead in terms of the manageability and control (and therefore capability to fit within corporate IT infrastructures).

2.6 Commercial exploitation

The results of the CRITIKAL project have played a significant role in shaping the next generation products from Attar Software. These products have been announced at the Data Warehouse98 Show at London-Olympia in November 1998 and will start shipping in January 1999. The results of CRITIKAL have been exploited in the new products as summarised below:

- XpertRule Miner is Attar's new data mining environment. It supports the data mining process developed in CRITIKAL by supporting all the stages of such a process from data preparation, transformation and exploration to pattern generation, validation, deployment and monitoring. The graphical front end of XpertRule Miner uses the design of DataPrep.
- The architecture of XpertRule Miner uses the multi-tier client server design of CRITIKAL. Attar has implemented a middle tier Proxy server and a number of CAF servers. Attar Software also plans to exploit the performance optimisation techniques researched in CRITIKAL.
- The new data mining Client ProfilerX has been completely redesigned to incorporate the end user feedback on usability, pattern presentation and deployment. ProfilerX is an ActiveX client which can be seamlessly embedded in other front end applications.

Attar Software is also using the Data Mining Process developed within CRITIKAL in its data mining training programmes.

2.7 Management

<<This section removed in public version of report.>>

2.8 Deliverables

<<This section removed in public version of report.>>

2.9 Publications and dissemination

2.9.1 Conference and seminar presentations and workshop participation

Event	Date	Presenter/participant
7th International Workshop on Research Issues in Data Engineering: High Performance Database Management for large Scale Applications: RIDE'97 Conference, Birmingham	7-8.4.97	A Al-Attar, Attar P J Allen & M J Addis, PAC
Unicom Data Mining Conference, London	24.6.97	C Upstill, PAC
IBM SP World, London	28-29.10.97	C Upstill, PAC
CRISP-DM SIG meeting, Amsterdam	20.11.97	A Al-Attar, Attar C J Scott, PAC
HPCnet Effective Knowledge Discovery workshop, Milan	4.7.97	A Al-Attar, Attar B Barham & D Nisbet, Lloyds TSB P J Allen, C J Scott & C Upstill, PAC
UK Industrial IT Forum meeting, Sheffield	25.2.98	A Al-Attar, Attar
EPCC seminar on Data Mining for Financial Services, Edinburgh	1.4.98	A Al-Attar, Attar S Hellberg, PAC
Unicom seminar: Data Mining Business Benefits, London	28.4.98	A Al-Attar, Attar
Operational Research Society Conference, Lancaster	8.9.98	A Al-Attar, Attar
CRITIKAL seminar: Scalable Client-Server Data Mining: Technologies, Management, London	9.12.98	A Al-Attar & H Al-Attar, Attar W Schneider & V Wenzel, GEHE

		B Barham, A Douthwaite & D Nisbet, Lloyds TSB M J Addis, P J Allen, C J Scott & C Upstill, PAC T Barth, BWI
Unicom seminar: Data Mining in the Energy & Process Industries, London	10.12.98	A Al-Attar, Attar
MicroStrategy Congress	12.98	W Schneider, GEHE
'A Multi-Tier Architecture for High-Performance Data Mining' has been accepted for the conference: Fachtagung Datenbanksysteme in Buero, Technik und Wissenschaft (BTW99)	1999	H Schwarz & R Rantzaus, IPVR
Paper on ARD in 3-tier architecture in preparation	1999	H Schwarz & R Rantzaus, IPVR

2.9.2 Internal presentations

Presentation to	Date	Participants
U of Southampton Concurrent Computation Group	24.2.98	M J Addis, PAC
GEHE business users	various	S Weh, GEHE
Lloyds TSB General Insurance	10.12.98	H Al-Attar, Attar D Nisbet, B Barham & A Douthwaite, Lloyds TSB M J Addis, PAC
U of Stuttgart BWI	24.8.98	T Barth, BWI
Internal seminars on data warehousing and data mining	summer 1998	H Schwarz & R Rantzaus, IPVR

2.9.3 Third party briefings

<<This section removed in public version of report.>>

2.9.4 Papers and articles

Al-Attar, Akeel,
CRITIKAL: Client-server Rule Indusion Technology for Industrial Knowledge Acquisition from Large Databases,
in: Proceedings of the RIDE'97 Workshop

Barth, Tilmann; Hertweck, Andreas (1998),
Kognition und Information. Data Mining - intelligenter Helfer in der Datenlawine,
in: *IT-Managment*, o.Jg. (1998), Nr. 10, S. 28-34 (German)

Barth, Tilmann; Hertweck, Andreas (1999),
Das Management von Data Mining Prozessen - strategische und operative Aspekte (provisional title),
in: *Computerworld Schweiz*, to be edited in 1999 (German)

Rantzaus, Ralf; Schwarz, Holger,
A Multi-Tier Architecture for High-Performance Data Mining
to appear in: Proceedings of BTW99

2.9.5 Publicly available reports

At the time of writing, CRITIKAL report “Guidelines for the Data Mining Process” (D2.3) was available for download from <http://www.planung.bwi.uni-stuttgart.de/> and <http://www.pac.soton.ac.uk/critikal/>.

CRITIKAL reports “Data Warehouse Design: the Impact of Data Mining” (D2.4) and “Best Practice” (D8.3) will be available shortly from <http://www.pac.soton.ac.uk/critikal/>.

2.9.6 Project poster and flier

The PAC produced a CRITIKAL poster and flier (see Figure 4) for use at the CRITIKAL seminar.



Figure 4. CRITIKAL poster and flier

2.9.7 Website

The PAC has developed a website for CRITIKAL at <http://www.pac.soton.ac.uk/critikal/>. This will continue to be supported for at least six months after the end of the project.

2.9.8 Demonstrator

Attar and the PAC have implemented a Web-accessible demonstration of the CRITIKAL system. This was publicly demonstrated for the first time at the CRITIKAL seminar on 9 December 1998. It can be accessed on <http://194.164.24.39/critikal>.

3 Exploitation plans

<<This section removed in public version of report.>>

4 Conclusions

In CRITIKAL, the partners sought to establish a position of leadership in the data mining industry. They have successfully achieved most of the objectives of the project. In particular, the technical feasibility, manageability and usability of multi-tier client-server mining against large data warehouses within the IT infrastructures found in large companies have been successfully demonstrated. No one else has yet demonstrated the capabilities that have been demonstrated in CRITIKAL. With recent announcements from vendors like Angoss and ISL, the vision represented in the CRITIKAL project seems to be receiving validation.

Attar Software has already announced product technology featuring developments made in CRITIKAL.

All partners in CRITIKAL have acquired significant knowledge from participation. This knowledge is being reused in a wide range of activities including research, teaching, process improvement, system engineering and project management.

The overall success of the project is, however, moderated slightly by the fact that no substantial, real-world business problem has been undertaken using the CRITIKAL system.