COMMIT

WORKPLAN

WORKPACKAGES

DELIVERABLES

BUDGET

SPATIOGTEMPORAL DATAWAREHOUSES FOR TRAJECTORY EXPLOITATION (P19)

Projectleider prof.dr. Martin Kersen

1. Background

Internet technologies have created a fabric for information exchange far beyond the original intent to share multimedia information between humans. It has become a landscape where entities (people, sensors, devices, observatories) generate events bound in time and by geo-spatial locations, carving out event trajectories of interest. Collecting, archiving and annotation of these trajectories forms the basis for knowledge extraction, e.g., for fleet optimization to reduce traffic congestion and energy consumption, manage social communities, e.g., to find similar cases quickly for emergency handling (spam, DDOS), and to organize the community by mutual interest, e.g., holiday entertainment. Likewise, large-scale trajectory collections become the prime target for scientific discovery and natural disaster management, e.g., seismic and remote sensing.

In scientific terms, the above calls for major technological advances in the design, development and deployment of a spatial-temporal datawarehouse. The technical challenges are focussed on scale and responsiveness in the face of massive observations. One of the salient outcomes of this project is an open-source reference platform, called the *Time-Trail Vault*, which supports experimentation with datamining algorithms over trajectories, query processing over scalable databases derived from real-life trajectory use-cases, and community-centric trajectory tagging and profiling.

The consortium is led by CWI, which has a long-standing, successful tradition to pursue frontier database research with the creation of spinoffs and broad experience in dissemination to the market. The prime partners University of Twente, Utrecht University, and TomTom research laboratory represent leading research groups in their domains of interest: data mining, interoperability, and route information management. The partners Hyves, KNMI, and Arcadis provide the use-cases from real-life application frameworks and data sets covering domains: large scale system networks, environmental networks, and social networks. Small high-tech companies MAG/EuroCottage and MonetDB are involved as contributors to improve market uptake and dissemination efforts.

This project addresses the fundamental challenges emerging from databases populated with trajectory data using novel techniques and scientific insights, such as automatically evolving distributed database architectures [15][20] and self-tuning database techniques [1][2] to cope with the ever growing data volumes. Scaling towards the data volume and breadth of applications requires major innovations in the hitherto solutions developed for datawarehouses [18][19]. It is a key topic in the major database conferences, e.g., VLDB, SIGMOD, ICDE and EDBT. MDL-based pattern mining such as [3][4][14] forms the centre of a wide variety of algorithms and applications that all depend on models consisting of small sets of patterns. Generalizing these techniques to spatiotemporal and sensor data streams at a scale exemplified by the use cases is an open research topic at conferences such as KDD and PKDD. The realization of dataspaces [5] around

volunteered information requires (geospatial) entity recognition [8][9][12][16][17], implicit handling of the uncertainty around vague objects [7] and disambiguated conflicts [10], and user-guided data cleaning [6][11].

The project extends the Bricks project on scalable database technology for the sciences and the successful work on databases in MultimediaN, which led to world-wide use of award-winning open-source technology MonetDB and three Bricks spin-off companies. The project addresses specifically the scientific and technological requirements stemming from both (mobile) sensor (TomTom) and environmental science applications (KNMI) as captured by trajectory information, but generalizes the target towards plural deployment in all domains where large-scale databases, data management, and decision support are prevalent.

The work described in this project is aligned with international FP-7 projects (EMILI, TELEIOS, LOD2, Planetdata, NERA, and EPOS) and eScience cooperation (UK, USA). This collaboration will lead to an unique international exposure of datawarehouse technology captured in the SciLens portal. The project aims at providing technology and expertise on database management in building strength within the COMMIT consortium, especially P5 (Sensor Content for Wellbeing), P12 (Dependable cooperative Systems for Public Safety) and P23 (Biobanking with imaging).

COMMIT themes: Middleware (databases), Science (generic, beta), I-services (mobility), Analysis (methods, mining), User aspects (interaction), public safety.

2. Problem description

TomTom helps building a trustworthy and efficient road infrastructure for mobility. They face the challenge to improve their real-time information about the anticipatory state of the network, better estimates of delays and times of arrival, managing fleets of vehicles and emergency services, planning and maintenance, and profiling for better location-based information brokerage. The insights provide opportunities for significant economical savings decrease of (fossil) fuel consumption, reduction of environmental pollution, and improved user satisfaction. The quality and public acceptance of infrastructural projects can be significantly improved with an open dialogue among the community of travellers, citizens, and professionals involved in those projects. Such community orchestration has become an integral part of companies involved in infrastructure development and maintenance, e.g., Arcadis.

Malignant behaviour of individuals in and against large social networks, such as managed by Hyves, hinder community collaboration. Providing social networking technology protected against spam and DDOS attacks is a major challenge.

Poor data quality on seismic behaviour and distributed sensor reading integrations in a timely manner may worsen environmental damage. KNMI with ORFEUS are an international clearing house for seismic data, which faces severe scalability issues.

The major scientific technological problems exhibited by these diverse problem areas stem from the avalanche of trajectory information coming from mobile devices, e.g., from persons and vehicles, access to internet sources, e.g., click streams, and sensory observations about physical systems, e.g., seismic and remote sensing. Hundred of millions mobility events a day, thousands of seismic sensor readings per minute, and hundred of thousands of web clicks in a short burst, are to be dealt with efficiently. The exponential growth calls for efficient storage, innovative data management, community engaged annotation and scalable data exploration and datamining algorithms. It is important to continuously organize, summarize and distribute the bulk of event for decision making. Exploitation requires a long living and robust data management solution. The trend to keep all event trajectories for *a posteriori* analysis cannot be maintained at the growth rate currently seen.

Performance is just one of the dimensions. Equally important is to recognize and group behavioural patterns for decision making. To illustrate, the system's characterization of each single user context is called a *spatiotemporal-profile*. It represents the system's understanding of the user's activities in space and time, gradually built up from valuations and other communications with that user. Annotation is also the result of enrichment of the trajectories through classification, learning and pattern mining of sensory or observational data. Such can be handled in a Time-Trail Vault and provides information for, e.g., MAG/EuroCottage.

It is important to keep a firm grip on data quality for the inherently imperfect processes above, where ambiguity and conflicts occur naturally. For example, seismic sensors may produce imprecise or even faulty information due to physical defects or observations anomalies. Quality control in mobility, seismology and communities each take a different stance. In the context of user-supplied annotations, we need to adequately deal with accuracy and completeness of and trust in volunteered information, geo-location of not explicitly geo-referenced information, and matching this against known spatial information from executed surveys and *in situ* sensor nets. Volunteered spatial-temporal data will come to us in large quantities as semi-structured reports describing local context. These volunteers may not have positioning devices and thus such data may lack precise geo-location. In temporal and thematic details, the reports may be similarly imprecise.

The scientific challenges addressed within this project are summarized along four dimensions:

1. The Trajectory *datamining* challenge: How to abstract and generalize the diverse application-specific trajectory mining tasks into formal scientific problem descriptions in order to develop sound algorithms as basis for widely applicable solutions.
2. The *Spatiotemporal-profiling* challenge: How to obtain clean data and orchestrate the tagging for and by improved social interaction.

3. The *Trajectory datawarehouse* challenge: How to deal efficiently with the data explosion of trajectory events generated in the web-of things and be able to produce timely responses to user queries and finding patterns of interest over trajectory events.
4. The *Open-source* challenge: How to capitalize and extend one of the Netherlands best database technologies to form a strong basis for science datawarehouses, thereby enabling a higher chance of scientific breakthroughs.

3. Objectives

Project's goal

Collecting, archiving and annotation of *event trajectories produced by humans and moving objects in general, forms the basis for knowledge extraction, e.g., for fleet optimization to reduce traffic congestion and energy consumption, manage social communities, e.g., to find similar cases quickly for emergency handling (spam, DDOS), and to organize the community by mutual interest, e.g., holiday entertainment. Likewise, large-scale trajectory collections become the prime target for scientific discovery and natural disaster management, e.g., seismic and remote sensing. In scientific terms, the above calls for major *advances in the design, development and deployment of a spatial-temporal datawarehouse for improved user profiling using large-scale datamining methods. The technical challenges are focused on scale and responsiveness in the face of massive observations. One of the salient outcomes of this project is an open-source reference platform, called the Time-Trail Vault, which supports experimentation with datamining algorithms over trajectories, query processing over scalable databases derived from real-life trajectory use-cases, and community-centric trajectory tagging and profiling.

Planning of all dimensions

The project is organized around 3 themes covering 2 workpackages each for the critical mass required. Two supportive workpackages look at long term research and system management facilities in the short term.

Overall goals:

- Datamining methods for better understanding seismic behavior
- Datawarehouse methods to improve analysis of time-series at 100s
- GB scale
- Open-source benchmark for geo-spatial events based on routing
- Datawarehouse methods to manage >TB scale geospatial events
- Geo-spatial portal for improved user community interaction
- Geo-spatial user profiling methods for improved *interaction
- Multi-scale query processing for informative and speedy user
- interaction
- Datawarehouse reference platform for geo-spatial temporal event
- Handling

Results

The main outcome is a reference platform of the Time-trail Vault, which includes sample data from KNMI, TomTom, EuroCottage, Arcadis and Hyves. This platform is used to derive knowledge extraction algorithms for time series in object routing, seismic observations, and geospatial profiling. All are critical components in emerging business intelligence applications. The software developed becomes part as (optional) components of the MonetDB open-source project, whose grand total distribution is expected to reach >200.000 in the course of the project. The benchmark dataset for studying mobility is developed and disseminated. The results on scientific data management *will be disseminated as real-life cases through the SciLens portal.

Deliverable Impact and Valorization

The MonetDB software is distributed worldwide. Within the project it leads to potential *cost reduction and improved performance at the partners KNMI, Hyves, TomTom. The number of downloads is based on 3 feature releases and 9 bug-fix released/year. The number of downloads of the new modules within this year will be around 50.000 (based on availability of the source and the quality assurance schedules) The *geo-portal applications will be made available, which should read 1000s of user interactions. The time-trail vault architecture will be integrated with the SciLens portal, which provides a showcase for large escience warehousing challenges. Yearly hands-on-workshops are organised around the Time-vault warehouse to stimulate further take up within and outside of COMMIT.

Deliverable Dissemination

The knowledge generated will be presented to the international scene and interested laymen in science databases as part of the SciLens portal. The MonetDB software package is used for both education and commercial use worldwide. Consultancy is provided through the MonetDB company. Hands-on workshop experiences will be provided to project partners and their affiliated institutions.

International Imbedding

The activities are internationally embedded through the research lab of TomTom, the European projects undertaken by KNMI, the European projects undertaken by CWI, and the UK eScience program. The Time-Vault datawarehouse unifies trajectory management techniques in a broad scale of application domains, including astronomy, traffic management, social network interaction, remote sensing, emergency management and semantic web techniques. It is showcased through the SciLens portal on a worldwide leading database platform for data intensive research.

Deliverable Synergy

P5. Synergy with other projects will be searched by offering direct technical consultancy in using the MonetDB software, e.g. by regularly organised hands-on-workshops. Close interaction with projects using database technology is envisioned projects with its emphasis on sensor information, and/or data warehouse requirements.

4. Economic and social relevance

Social-economic relevance: Data warehouses with spatiotemporal events manifest themselves in many information systems. Their social relevance spans global applications, e.g., KNMI is involved in the verification on international acts on nuclear activities, and University Twente / ITC in the UN's Millennium Development Goals to improve health, poverty, education, gender equality and environmental sustainability in less developed countries. The economic relevance stems from the prime business partners TomTom, and Arcadis, who all address trajectory information and whose analysis has important impacts on short/long term planning of their customers. The notion of locality and time in the context of Hyves illustrates the generic nature of the underlying technical problems, where the effects translate into better energy management in data centres and early warning systems for invasive spam activity. Together with the smaller companies, it shows that the socio-economic relevance spreads way beyond the partners' core business activities.

The technology developed within this project serves several socio-economic purposes directly:

- The trajectory mining algorithms form a basis for iServices in areas as broad as mobility (TomTom), environmental monitoring (KNMI) and distributed system management (Hyves), all of which lead to direct economic benefits, while the open-source packaging secures the widest possible indirect economic benefit.
- The geo-tagging algorithms lead to economic benefits by improved data quality, whose exploitation is tested by Arcadis and MAG/EuroCottage, but whose solutions are spread through ITC as open-source to reach the developing countries as socio-economic contributions. It serves iServices.
- The Time-Trail Vault platform provides a Middleware open-access infrastructure to study solutions against real-world use-case databases. In particular, it provides a cost-effective solution for the sciences to build virtual observatories with trajectory information and geo-annotated information using a common, well-supported and easily accessible platform.

Economic effects: Aside from the direct positive effect on the service offerings of the partners investing in this common goal, we can assess the indirect effect on the economy using the valuation model developed in MultimediaN. Since the Time-Trail Vault extends the MonetDB platform already on the open-source market, we can safely predict that this new platform will draw hundreds of users worldwide daily, and that the open-source distribution over the lifetime of the project will reach 500.000 copies. Even with a modest implicit price tag of 100 euro per

copy, this leads to an implicit economic value of over 50 million euro. Such a success also provides the basis for the MonetDB company to expands its business into consultancy.

Urgency: Bringing together such a variety of partners for the development of generic technology is unique. For each and everyone, the investments necessary to even come close to a solution are too high. The urgency is felt by all, because they are currently locked into non-flexible, proprietary solutions. The concerted action led by a group known for its achievements in database system technology worldwide, forms the best guarantee for lasting results. The end result of the partners' investments is a contribution to society of use-case abstractions and an industrial strength platform to steer technological innovation in the Netherlands.

5. Consortium

Collaboration is organized around three dimensions: within project P19, within COMMIT, and internationally. Within P19 the commercial partners TomTom, KNMI, Hyves, Arcadis, MonetDB and MAG/EuroCottage, provide representative application scenarios and real-life data from various areas of modern socio-economical life. Despite their diversity and application-specific characteristics, these scenarios pose common data management challenges that go beyond the solution provided by existing technology. Any individual application-specific extension of existing datawarehouse technology can only provide a short-term solution. The goal of this project is to capitalize on the synergy and cross-fertilization effects of combining the diverse challenges into a larger framework. Providing generalized scientific solutions forms a long-term basis for addressing the data management challenges not only within this project but also far beyond.

The Algorithmic Data Analysis group at Utrecht University has a long-standing track-record in pattern mining base on Minimum Description Length (MDL). The Krimp algorithm [3][4][14] is the centre of a wide variety of algorithms and applications that all depend on models consisting of small sets of patterns. A research direction is to generalize this towards all dimensions of "Who, What, When, Where event" databases, including both the use of constraints and patterns on spatiotemporal data.

The University of Twente and ITC contribute the expertise on sensor data management, management of uncertainty in data, data cleaning, and geospatial data infrastructures and technologies.

The database group of CWI acts as a pivot in this project. CWI hosts one of the worlds leading database research groups that has a long track-record in pioneering highly innovative hardware-aware column-oriented database technology. Its award-winning open-source database system MonetDB provides a world-wide unique basis for developing novel database technology to

support and integrate the datamining as well as data cleaning and enriching techniques developed by the two universities.

TomTom research plays a central role as link between the (non-)profit and research partners. Not constrained by short-term commercial goals, but familiar with real-life requirements, it helps with assessing the techniques developed within this project by designing and implementing a generic real-world application framework. Hyves is by far the largest and still growing social network in the Netherlands with more than 5 million people registered. KNMI with Orfeus manages the largest sensor-enabled infrastructure for seismology.

The current innovation market is based on open-innovation where research institutions and companies together develop technology for the years ahead. The worktable structure around a common reference platform has proofed a highly successful and natural setting to enable knowledge and technology transfer. TomTom, Hyves and MAG/EuroCottage are relatively young companies. There is no legacy research department, but a need to partner with the leading institutes. The project timing and scope is urgent. The technological basis is already stressed to the edge, but also within the current economic climate, investments are even more important than during high conjuncture. The investment has both a mid-term and long-term effect. The mid-term effect for TomTom is partly securing their leadership in navigation technology in their R&D department. For Hyves it is a necessary investment to avoid a system collapse before it reaches the world at large.

Collaboration beyond P19 is emphasized by international projects. The work in EMILI is focussed on emergency management, from which we derive event processing technology. The work in TELEIOS aims at remote sensing for natural disaster recognition and management, which provides the cornerstone to extend relational technology into image processing and management. The work in LOD2 and Planetdata provide the hooks to weave information as pursued in the semantic web at a global scale. All these activities aim for a real-life demo made available worldwide through the SciLens portal.Many of the COMMIT projects also rely on database and datawarehouse technology, but more in terms of commodity functionality provided by off-the-shelve solutions then stretching their functionality. As such, the approach taken in P19 is to provide for the necessary technical expertise on a case-by-case basis and hands-on-workshops using the MonetDB platform. In addition, we selectively interact with projects that exhibit new datawarehouse requirements, Direct interaction in the form of working meetings organized by P19 is foreseen with P5 (Sensor Content for Wellbeing), P12 (Dependable cooperative Systems for Public Safety), P20 (e-Infrastructure) and P23 (Biobanking with imaging).

The project is centred around leading senior researchers, including Martin Kersten (CWI) a visionary database architect, Stefan Manegold (CWI) a researcher system architect with strong

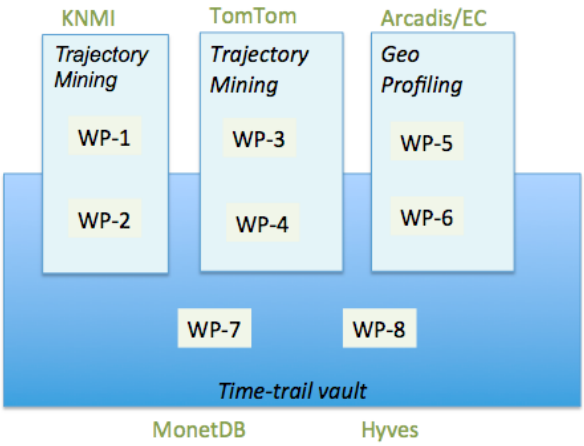
quality assurance drives, Arno Siebes (UU) a world expert on knowledge discovery in databases, Ralf Shafer, leader of TomTom research, Maurice van Keulen (UT) an expert on probabilistic data integration, and Rolf de By (ITC) an expert on spatial data infrastructures. The newly acquired personnel will be monitored and guided by these individuals. CWI is in the lead as it comes to Time-Trail Vault architecture, development, packaging and dissemination. With sufficient take-up, we expect a spinoff company to be formed in the wake of the project to secure future developments and maintenance.

6. Workplan

The workplan is organized around three workpackage clusters for manageability.

- Trajectory Mining led by UU and TomTom research lab, supported by CWI, aimed at Stream Mining, Trail Mining, and Streaming Query Processing, with use cases derived from TomTom and KNMI.
- Geo Profiling led by UT, involving CWI & ITC, geared at the projects Community Tagging, User-centric Geo-profiling with use cases derived from Arcadis, MAG.
- Time-Trail Vault led by CWI, involving MonetDB and Hyves consists of the projects Stream Query Processing, Time-Trail Architecture, Multi-Scale Query Processing, and Multi-Scale Datawarehouse Virtualization, based on use-cases from TomTom, KNMI, Hyves, Arcadis.

The workplan is detailed in the appendix. It roughly covers a start-up phase of three months, internships as of day one for all junior researchers, monthly project meetings, and cooperative development of demos on a biannual basis to exhibit results to the corporate management. The relationships between the workpackages are illustrated below.



The Time-Trail Vault reference architecture forms the commonly shared, publicly available back-bone that integrates the two flavours of the data warehouse platform. A prototype version to experiment with the requirements posed is made publicly available within the first year as a datawarehouse service. It will evolve in 6-months release cycles pushing scientific results into the open-source community for early take-up.

The starting point for the reference architecture is the combined requirements derived from the use-cases, which cover a wide area of scientific and technological challenges. The TomTom input provides an experimental base for addressing real-life, and scalable datamining. The first step in that direction is to adjust the Krimp technology to become applicable to point and

trajectory information. The database layer is expanded to cater for the complementary challenges and requirements of the three working-tables' application areas help to develop more robust and generally applicable shared datamining algorithms and database technology in the Time-Trail Vault. Both working-tables' demonstrators evaluate not only the working-table specific work, but also the commonly shared issues through the Time-Trail Vault.

WORKPACKAGES

Project number P19	
WP title & acronym	WP1: Trajectory Mining
WP leader	Arno Siebes, Universiteit Utrecht

Objective:

The goal of this project is to extend the Krimp data mining algorithms to mine the patterns in large Time-Trail Vault data warehouses efficiently, this is also known as *trajectory mining*.

The key subjective criteria are *geographic* cohesion and *semantic* (as denoted by the attributes) meaningfulness, e.g., patterns that describe geo-temporal unrelated events are deemed unimportant. *Mining constraints* are used to enforce the subjective application criteria. The objective criteria, based on Minimum Description Length principle (MDL), aim to ensure that the data analyst is confronted with statistically significant results, only.

Krimp has a proven track record to select the statistical meaningful item sets from the large set of frequent item sets. Moreover, the resulting *Code Table* is known to provide a very precise characterization of the underlying data distribution.

The three main challenges of this project are 1) *Krimp* has to be generalized to spatiotemporal data, this includes the non-trivial generalization to real-valued data, 2) user defined *constraints* have to be integrated with the MDL-based framework, together they provide the technical means to derive meaningful patterns, and 3) the efficiency and scalability of the developed algorithms are to be shown on seismic data provided by KNMI and managed by MonetDB (WP-2, WP-4, WP-7, WP-8). The relevance of the discovered patterns is, of course, a driving force in the algorithm development, next to the technical means, joint research with WP-2, WP-3 & WP-4 is integral to this project. That is, the algorithms will be tested and demonstrated on data sets provided by TomTom and integrated by WP-2 & WP-4 in the Time-Trail Vault.

Project number P19	
WP title & acronym	WP2: Time-Trail Warehouse Architecture
WP leader	Martin Kersten, CWI

Objectives: Design and develop a reference architecture for database-centric analysis of large spatiotemporal data streams comprised of trajectory events.

Background: Efficient content-based analysis of time trails, i.e., streams of (Who, What, When, Where) events, poses new challenges for query optimization. It is the key component to achieve the required scalability into the hundreds of terabytes of events that should be mined and inspected at near real-time requirements.

Description of work: The driving applications are the trajectory streams from TomTom and KNMI. They both call for innovative techniques to handle both bulk loads, enable short circuit responses, and integrate seamlessly with large trajectory repositories. The major task is to tackle the challenge posed by using domain specific optimization techniques with “low-level” dynamic optimization techniques and transparent access to foreign file repositories. Traditional cost-based query optimization techniques are not applicable is in this context as the shape and characteristics of a good query execution plan depend on the actual data distribution. Instead, these characteristics change over time. The query plan is changed accordingly, preferably in an automatic self-organizing way. These optimization techniques form a significant contribution to support the content analysis in WP-1, WP-3. In turn, the instant stream mining techniques as developed in WP-1 and applied in WP-3 will deliver valuable input to novel data-dependent self-tuning optimization techniques in particular of continuous queries over data streams in general.

Project number P19	
WP title & acronym	WP3: Trail Mining
WP leader	Ralf-Peter Schäfer

Objectives: Design and implement a generic template for a real-world application framework, and provide sample instantiations.

Background: Real-world application scenarios form both a challenge that provides the data for the research done in work packages WP1, WP2, WP4, as well as a gauge to assess the applicability and socio-economical impact of the of the algorithms, techniques and reference architectures developed in this project.

Description of work: This work package will design and implement a template for a real-world application framework, provide requirements and guidelines how to instantiate the template for various application domains and usage scenarios. Using traffic data of various sources, trip and statistical data from navigation communities. This framework serves as a platform to perform feasibility studies and performance analysis of existing and newly developed spatiotemporal mining techniques as well as the respective extended database technology developed in the course of this project. Given the real-world data and application scenarios, we will develop suitable benchmarks for feasibility studies and performance assessment. Traffic data of various sources, trip and statistical data from navigation communities, as well as usage scenarios from WP1 (KNMI) will be used to demonstrate a sample instantiations of the framework template.

A *Route & Trail mining demonstrator* will show how the techniques developed help to discover patterns and relationships in and derive information from GPS probe data that then serve as valuable input to develop robust routing.

Project number P19	
WP title & acronym	WP4: Trajectory Stream Processing
WP leader	Stefan Manegold, CWI

Objectives: Design and develop trajectory-based stream storage and query processing techniques to support real-time trajectory exploration and mining algorithms.

Background: Acting as the common back-bone to support the vast variety of trajectory mining challenges stemming from diverse use case scenarios as addressed in WP-1, WP-3, WP-5, WP-8 this component of the Time-Trail Vault advances stream database processing towards spatiotemporal events. The variety of data types and the complexity of spatiotemporal pattern detection call for innovations in the lower parts of the software architectures, i.e., the core of the database engine that powers the data mining system. In particular, it should deal with partial time series possible stored in auxiliary file-repositories.

Description of work: The driving applications are mostly the trajectory streams from TomTom and KNMI. They both call for innovative techniques to handle both bulk loads, enable short circuit responses, and integrate seamlessly with large trajectory repositories. Novel data exploration techniques have to be developed to meet the special characteristics of data that is "only passing by". The challenges range from fast outlier detection, on-line clustering and aggregation of data streams to continuous analysis of the data stream(s) to derive information "on-the-fly" and support instant decisions [3][4][14]. Furthermore, existing stream processing systems do not provide the flexibility and processing capacities to efficiently implement and perform complex data mining tasks on high-volume data streams. The envisioned solution is to leverage the flexibility and scalability of existing high-performance database technology. Enriching it with novel efficient stream-mining primitives yields the highly scalable back-end support for a large-scale high-performance stream mining architecture

Project number P19	
WP title & acronym	WP5: Community Tagging
WP leader	Andreas Wombacher, Universiteit Twente

Objectives: Design and develop trajectory-based stream storage and query processing techniques to support real-time trajectory exploration and mining algorithms.

Background: Acting as the common back-bone to support the vast variety of trajectory mining challenges stemming from diverse use case scenarios as addressed in WP-1, WP-3, WP-5, WP-8 this component of the Time-Trail Vault advances stream database processing towards spatiotemporal events. The variety of data types and the complexity of spatiotemporal pattern detection call for innovations in the lower parts of the software architectures, i.e., the core of the database engine that powers the data mining system. In particular, it should deal with partial time series possible stored in auxiliary file-repositories.

Description of work: The driving applications are mostly the trajectory streams from TomTom and KNMI. They both call for innovative techniques to handle both bulk loads, enable short circuit responses, and integrate seamlessly with large trajectory repositories. Novel data exploration techniques have to be developed to meet the special characteristics of data that is "only passing by". The challenges range from fast outlier detection, on-line clustering and aggregation of data streams to continuous analysis of the data stream(s) to derive information "on-the-fly" and support instant decisions [3][4][14]. Furthermore, existing stream processing systems do not provide the flexibility and processing capacities to efficiently implement and perform complex data mining tasks on high-volume data streams. The envisioned solution is to leverage the flexibility and scalability of existing high-performance database technology. Enriching it with novel efficient stream-mining primitive's yields the highly scalable back-end support for a large-scale high-performance stream mining architecture.

Project number P19	
WP title & acronym	WP6: User-centric Geo-profiling
WP leader	Maurice van Keulen, Universiteit Twente

Objectives

Objectives: Design and implementation of a framework for geoprofile-driven content enrichment and data quality improvement on the basis of user-volunteered content and open spatial data services.

Background: The domain-of-activity is often the catalyst in the formation of a societal network, but standardized approaches to their IT support are inadequate. The where-and-when of network member activity is a dominating information factor, and its proper characterization (in geoprofiles) allows to optimize communication between the members.

Description of work: This WP aims to provide a toolset for building support systems for a networked community whose members display similar activities in similar locations and want to share raw multimedia content, valuations and experiences about those. Scientific challenges encountered include geo-referenced entity resolution, automatic data integration for content enrichment, information extraction, data quality improvement, uncertainty management and data quality improvement, domain-of-activity specification, understanding user activities and routes (geoprofiling), and geo-data processing. An XML-based ETL-architecture will be developed which is based on a spatially enhanced XML DBMS and which includes a toolset to provide for development support with the above challenges from user-volunteered content.

Two demonstrators are developed in this context: (1) EuroCottage portal improvement for community building and holiday home location profiling and (2) international development collaboration in agriculture. The first is developed in stages: we first focus on content enrichment for the EuroCottage portal for early dissemination purposes and as a basis for a later geoprofile-driven enhancements. Furthermore, we also plan to quickly develop a EuroCottage holiday mash-up service to attract and gather user-volunteered content to be used for the geoprofiling research and validation. The second is developed in year 4 as an additional demonstrator to increase general impact and dissemination and to validate domain flexibility.

Project number P19	
WP title & acronym	WP7: Multi-Scale Query Processing
WP leader	Stratos Idreos, CWI

Objectives: Design and develop multi-scale query processing technologies for efficient and timely processing of ad-hoc queries over huge data volumes.

Background: A key assumption underlying contemporary database architectures is to prepare the database storage and processing scheme for instant delivery of precisely articulated queries (e.g., fetch based on key or simple (join) predicates). However, ad-hoc querying hardly ever is precise, i.e., based on a formulation that can be answered quickly. Instead, aggregate queries and summaries over large subspaces are the scheme for the user to zoom in onto the answer set. Consequently, the query interaction paradigm has reached its end of life in the context of extreme large databases constructed from scientific experiments (e.g., astronomy) and large scale distributed sensory systems (e.g., health, surveillance, logistics).

Description of work: Current systems cannot cope with such ad hoc queries over large datasets. We plan to investigate a completely new query processing paradigm that removes many of the current restrictions. For example, modern systems need to load and analyze all data before being able to process it. Our architecture aims at an incremental data consumption driven by processing needs, i.e., no need to load, analyze and store data not used by queries. Another pitfall of modern systems is the need to build a priori and complete query accelerators, i.e., indexes. This is prohibitive with huge datasets as it would cost an enormous amount of time unless we know exactly which portions of the data are relevant. However, this is not the case for ad hoc query processing. We aim for a scheme that summarizes data portions succinctly, replace them by analytical models, and adaptively partition the database storage using the “database cracking schemes” pioneered at CWI [1,2]. The idea is that all necessary accelerators are built incrementally, automatically and in a transparent way to the user and the challenge is to achieve this with ad hoc queries in huge datasets. To achieve the goals described above, the query execution engine is replaced by a new one that can detect answer convergence and also can exploit randomized sub-plan evaluation to tackle the query load.

The approach taken is evaluated against the databases provided by KNMI, TomTom and Hyves in particular, but extended to cover other science databases as well. These offer good examples of our motivation scenarios. For example, TomTom contains a massive database that is queried continuously by numerous users. Given the vast amount of users, queries are ad hoc, continuously changing focus and given the nature of the application, answers need to be formulated fast otherwise they are not useful. Our approach aims to tackle exactly these kinds of problems with fast answers of ad hoc queries over huge data.

Project number P19	
WP title & acronym	WP8: Datawarehouse Virtualization
WP leader	Fabian Groffen, CWI

Objective Design a distributed spatio-temporal warehouse geared at system elasticity using DBMS virtualization.

Background: Virtualization of a database application in a cloud requires great care in the amount of resources to be dedicated, e.g., RAM, disk and CPUs. The state-of-the-art merely mimics a raw device onto a hardware platform, where it has to compete with others for resources. It easily leads to suboptimal performance in terms of resources and energy consumption compared to dedicated distributed systems.

Description of work: One of the benefits of a database system is that it has the knowledge to dynamically (re-) partition the database to minimize the number of virtual servers needed. The underlying technology has been pioneered in the context of a centralized solution and is known in the literature as "database cracking" and "sharding". A variation closely related to hardware infrastructures has been recently studied in the context of the DataCyclotron [20]. It provides an outlook on significant throughput improvement over complex queries. All schemes innovate over existing commercial solutions by adapting the storage layout as part of the query processing requests. Research has shown that this approach carries a lot of potential They aid in the creation of a fully-self-organizing distributed version of MonetDB.
Key scientific challenges are a) the level of database cracking/partitioning in relation to the time required to instantiate a virtual server or data access latency, b) the number of replicas to maximize response time, c) the scheduling of up/down time of virtual servers, and d) the system hardware topology.
The technology will be evaluated against examples provided throuhg the WP-1, and WP-3 and the Hyves infrastructure. Hyves already have a cluster comprised of several thousands of machines, which includes tens of MySQL database servers working in a sharded setup. This set up has clear limitations in terms of maintenance, and is consider inadequate to reach out into business intelligence application settings. The approach taken here is to challenge the evolving datawarehouse system with selected use-cases encountered in the Hyves setting. The focus is initially on fast stream processing. This includes massive click stream with trajectory events that should become input for both malicient behavior detection, e.g., spam and DDOS, and user profiling.
Further evaluation will be performed in the context of astronomy, e.g., LSST, and commercial applications, e.g., Hyves, Nozhup BV a client of MonetDB BV.

8. Deliverables

Important conference contributions

?

Important conference contributions

40

Products

1. GeoPortal Application

   Design and implementation of a framework for geo-based tagging of information by different user groups of a community and a sample use case for Arcadis. Consensus of a community on a development project in public space involves several interest groups. The consensus process is based on discussing on various design alternatives in a spatial representation (e.g., a map). To keep the spatial representation readable, aggregations of contributions have to be provided according to topic, space and time dimension. The main challenge is to provide information in a fast and flexible way without making many assumptions on the way users are going to ask queries. Every query considering the when, where, and what aspect of a fact requires aggregation on the different dimensions to enable visualization. Pre-aggregating all possible dimensions and levels consumes quite some storage with no guarantee of having pre-aggregated the right level.The idea is to observe the query behaviour of users and self-organize the pre-aggregation levels. I.e., based on the observed user behaviour and a storage boundary the optimal pre-aggregation levels are determined in a streaming fashion minimizing the query response time. Due to changes in the user behaviour over time the originally set up pre-aggregation levels may be not optimal anymore and may require adjustment.

   - WP 5 YP 2013

2. GeoTagging Application

   Design and implementation of a framework for geoprofile-driven content enrichment and data quality improvement on the basis of user-volunteered content and open spatial data services. We provide a toolset for building support systems for a networked community whose members display similar activities in similar locations and want to share raw multimedia content, valuations and experiences about those. Scientific challenges encountered include geo-referenced entity resolution, automatic data integration for content enrichment, information extraction, data quality improvement, uncertainty management and data quality improvement, domainof- activity specification, understanding user activities and routes (geoprofiling), and geo-data processing. The results will be presented using the Eurocottage portal.

   - WP 6 YP 2013

3. SciLens Applications

   The challenges posed by large-scale scientific databases are illustrated through the SciLens portal (http://www.scilens.org). The work on seismic data within this project complements ongoing work in astronomy and remote sensing of affiliated projects.

- WP 2 YP 2012

4. SPAM detection

   A special case of geo-temporal event sequences comes from large compute clusters, e.g. the 3000 machines managed by Hyves. Within this context, malicious user behavior is one of the prime challenges to be recognized quickly. This involves detection within seconds an increased workload on some of the servers and/or flooding the system with malicious content. The Time-vault reference platform is used to mine for patterns quickly and to intercept such behavior using emergeny management techniques developed in an accompanying project.

   - WP 8 YP 2012

Software

1. MonetDB Data mining toolkit

   The data-mining activities are centered around finding optimal partial encodings for the time series. By repeating the process, taking into account noise and errors, both a highly compressed representation of the database emerges, but also the code tables. The latter embody the rules hidden in the data. The final goal of this project is a toolkit for data mining, tightly integrated with the Time-vault software. A key challenge is to scale out from data mining a few thousand records sofar and the GBs provided by the application owners.

   - WP 1 YP 2014

2. MonetDB Data Vaults

   Description Many sciences have developed file-based repositories for archiving observations. The key challenge is to unlock these treasure chests through modern database technology. The focus is seismic data, stored as the Mseed standard for data interchange, a format for highly compressed partial time series. The seismology community shares this data for obvious reasons. The first step in concerns direct support through database technology, we develop a software module that provides transparent access to mseed data for querying with SQL. In combination with a prototype query optimizer, it would allow for content-based search over data hitherto only accessible by laborious file-based low-level programming. The next phase consists of development of an array-based query processing system, which would solve the hitherto problem of blending relational and arraybased paradigms. It allows the scientist to combine semantic annotations with mathematical rigorous processing.

   - WP 7 YP 2013

3. Public trajectory warehouse

   Description The current proprietary file format for route information has reached its end-oflife.
   It becomes increasingly difficult to extend it with properties that characterise the services of the user and the dynamic properties of the routes. Instead, a migration to a database enabled dynamic route information system is urgently needed. Within COMMIT, the MonetDB database technology is exercised to stress the limits of this approach and to derive a

prototype implementation for route mining. A public benchmark is developed based on the TomTom use case to assess 1) loading speed 2) high volume updates 3) geospatial reporting, and 4) data mining. The first target is to define the benchmark and establish the experimentation platform on MonetDB.
- WP 3 YP 2012

4. Time-Trail Vault Reference Architecture
   Description The KNMI and TomTom application settings both call for innovative techniques to handle both bulk loads, enable short circuit responses, and integrate seamlessly with large trajectory repositories. Novel data exploration techniques have to be developed to meet the special characteristics of data that is "only passing by".
   The challenges range from fast outlier detection, on-line clustering and aggregation of data streams to continuous analysis of the data stream(s) to derive information "on-the-fly" and support instant decisions.
   The design and realisation will be aligned with requirements stemming from both remote sensing and emergency management use cases.
   WP WP 4 YP 2012

5. MonetDB DataCyclotron
   Distributed processing on the SciLens platform, a 300-node database system, will be developed as a module for the MonetDB system. This software code base will provide the necessary performance improvement for real-time data mining the large datawarehouses.
   - WP 8 YP 2015

User studies

1. GeoPortal evaluation
   Consensus of a community on a development project in public space involves several interest groups. The consensus process is based on discussing on various design alternatives in a spatial representation (e.g., a map). To keep the spatial representation readable, aggregations of contributions have to be provided according to topic, space and time dimension. The GeoPortal will be evaluated over the course of the project.
   - WP 5 YP 2014

2. GeoTagging evaluation
   Description The Eurocottage portal will be used to gather user study data by keeping track of the web interaction.
   - WP 6 YP 2014

Other results

1. Scilens demonstrators

The results of the database technology in the context of various science disciplines are elicited through the SciLens portal. This portal is maintained and extended to cover as broad as possible exposure of the results. In year one, we plan to include the TomTom and KNMI cases, but also some initial results from remote sensing applications derived from the TELEIOS project. It will be subsequently extended with use-cases from within the project.
- WP? YR?

References

[1] S.Idreos, M.L. Kersten, S. Manegold, Updating a cracked database, In *Proceedings 2007 ACM SIGMOD International Conference on Management of Data*, 2007

[2] S. Idreos, M. L. Kersten, S. Manegold. Database Cracking. In *Proceedings of the Biennial Conference on Innovative Data Systems Research (CIDR),* Asilomar, CA, USA, 2007.

[3] J. Vreeken, M. van Leeuwen, A. Siebes: Characterising the difference. KDD 2007: 765-774

[4] M. van Leeuwen, A. Siebes: StreamKrimp: Detecting Change in Data Streams. ECML/PKDD (1) 2008: 672-687

[5] A. Y. Halevy, M. J. Franklin, D. Maier: Principles of Dataspace Systems. In *Proceedings of PODS 2006*, Chicago, IL, USA, 2006.

[6] M. Goodchild. Citizens as sensors: the world of volunteered geography. In *GeoJournal*, 69(4): 211-221, 2007.

[7] A. Dilo, R. A. de By, and A. Stein. A spatial algebra for vague objects. In *International Journal of Geographical Information Science*, 21(4): 397-426, 2007.

[8] R. Lemmens, C. Granell, M. Gould, A. Wytzisk, R. de By, and P. van Oosterom. Integrating Semantic and Syntactic Descriptions to Chain Geographic Services. In *IEEE Internet Computing*, 10(5): 42-52, 2006.

[9] P. DeRose, W. Shen, F. Chen, A. Doan, and R. Ramakrishnan. Building structured web community portals: A top-down, compositional, and incremental approach. In *Proceedings of the 33rd Int'l Conf. on Very Large Data Bases (VLDB), Vienna, Austria, September 23-27, 2007*, pages 399-410. ACM, 2007. ISBN 978-1-59593-649-3. 6

[10] M. van Keulen, A. de Keijzer, and W. Alink. A probabilistic XML approach to data integration. In *Proceedings of the 21st Int\u2019l Conf. on Data Engineering (ICDE2005), 5-8 April 2005, Tokyo, Japan*, pages 459-470. IEEE Computer Society, 2005. ISBN 0-7695-2285-8.

[11] van Keulen, M. and de Keijzer, A., "Qualitative Effects of Knowledge Rules and User Feedback in Probabilistic Data Integration". Accepted for The VLDB Journal. 2009. Springer, ISSN 1066-8888. DOI 10.1007/s00778-009-0156-z.

[12] Pavel Serdyukov and Djoerd Hiemstra. Modeling documents as mixtures of persons for expert finding. In *Proceedings of the 30th European Conference on IR Research (ECIR2008), Glasgow, UK, March 30-April 3, 2008*, volume 4956 of LNCS, pages 309-320. Springer, April 2008. ISBN 978-3-540-78645-0.

[13] Peter Boncz, Torsten Grust, Maurice van Keulen, Stefan Manegold, Jan Rittinger, and Jens Teubner. MonetDB/XQuery: a fast xquery processor powered by a relational engine. In *SIGMOD '06: Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 479-490, New York, NY, USA, 2006. ACM. ISBN 1-59593-434-0.

[14] A.P.J.M. Siebes, J. Vreeken & M. van Leeuwen. Item Sets that Compress. In J. Ghosh, D. Lambert, D.B. Skillicorn & J. Srivastava (Eds.), SIAM Conference on Data Mining. SIAM 2006.

[15] F. Groffen, M. L. *Kersten, S. Manegold. Armada: a Reference Model for an Evolving Database System. In Proceed*ings of Datenbanksysteme in Business, Technologie und Web, Aachen, Germany, March 2007.

[16] Duckham M, Mason K, Stell J, and Worboys M (2001) A formal approach to imperfection in geographic information. *Computers, Environment and Urban Systems* 25**,** pp 89–103.

[17] C. Batini, M. Scannapieco, "*Data Quality---Concepts, Methodologies and Techniques*". Data-centric Systems and Applications Series, M.J. Carey & S. Ceri (eds.), Springer, 2006.

[18] Peter A. Boncz, Stefan Manegold, Martin L. Kersten: Database Architecture Evolution: Mammals Flourished long before Dinosaurs became Extinct. PVLDB 2(2): 1648-1653 (2009).

[19] Milena Ivanova, Martin L. Kersten, Niels J. Nes, Romulo Goncalves: An architecture for recycling intermediates in a column-store. ACM Trans. Database Syst. 35(4): 24 (2010)

[20] R. Goncalves, M. Kersten: The Datacyclotron Architecture, EDBT 2010.