TECHNISCHE UNIVERSITÄT DRESDEN

Dresden Database Systems Group

Data Analytics in Databases

Pro- und Hauptseminar am Lehrstuhl für Datenbanken

11.04.2017 - Lars.Kegel@tu-dresden.de; Johannes.Luong@tu-dresden.de

Einordnung

Dresden Database Systems Group

Proseminar

- **Informatik/Medieninformatik Bachelor**: B-510, B-520, B-530, B-540, B-610
- **Informatik/Medieninformatik Diplom**: Wahlpflichtfach im Vertiefungsgebiet Datenbanken oder im Fachgebiet Architektur verteilter Systeme ab dem fünften Fachsemester
- **Wirtschaftsinformatik Bachelor**: WI-BA-08
- (0 V / 2 Ü / 0 P)

Hauptseminar

- **Informatik/Medieninformatik Bachelor:** INF-AQUA
- **Informatik/Medieninformatik Master**: VERT-4, INF-D-940, INF-AQUA
- **Informatik/Medieninformatik Diplom (2010):** VERT-4, INF-D-940
- **Informatik/Medieninformatik Diplom (2004):** INF-04-HS
- (0 V / 2 Ü / 0 P)

Organisatorisches

Ablauf

Dresden Database Systems Group

Auswahl des Themas

- Heute (bzw. durch Anfrage)
- Papiere zum Thema auf der Webseite des Lehrstuhls:
  - Hauptseminar: https://wwwdb.inf.tu-dresden.de/study/teaching/ss_17/hauptseminar-datenbanken/
  - Proseminar: https://wwwdb.inf.tu-dresden.de/study/teaching/ss_17/proseminar-datenbanken/

Einarbeitung in Thema

- Bis Ende April

Ausarbeitung eines Papiers

- Bis 11.06.2017

Verteidigungsvortrag

- Ende Juni

Regularien

Dresden Database Systems Group

Einarbeitung

- Vorgegebene(s) Papier(e) lesen
- Darüber hinausschauen und einordnen
- Literatur Recherche (aus Uni-Netz)
  - http://dblp.uni-trier.de/
  - http://dl.acm.org/
  - http://ieeexplore.ieee.org/Xplore/dynhome.jsp?tag=1
  - http://scholar.google.de/

Ausarbeitung

- Schriftliche Aufarbeitung des Themas (in deutscher oder englischer Sprache)
- Vorlage: http://www.acm.org/sigs/publications/proceedings-templates
- Template "ACM proceedings template (standard)", sample-sigconf.tex bzw. ACM_SigConf.docx
- 8-10 Seiten (Hauptseminar); 6-8 Seiten (Proseminar)
- Abgabe: bis 11.06.2017 per PDF an Betreuer UND lars.kegel@tu-dresden.de
- Vorab mindestens eine Iteration mit Betreuer (Ihr müsst euch melden!)

Regularien

Dresden Database Systems Group

Verteidigungsvortrag: Ende Juni

- Formelle Präsentation des Themas mit Folienprojektion
- Dauer: 20 Minuten
- Folienvorlage nach eurer Wahl (Vorlage des Lehrstuhls auf Anfrage)
- Folien sind in englischer Sprache zu verfassen
- Reihenfolge der Vorträge wird noch mitgeteilt

Fachliche Betreuung

- Individuell
- Betreuer werden zeitnah zugeteilt und bekannt gegeben
- Betreuer beantwortet Fragen zum Thema und gibt Hinweise zu Qualität von Ausarbeitung und Vortrag

Bewertung

- Verständnis, Ausarbeitung, Vortrag, Selbstständigkeit

Inhaltliches

Analytics for Better Insights

Dresden Database Systems Group

Business Analytics

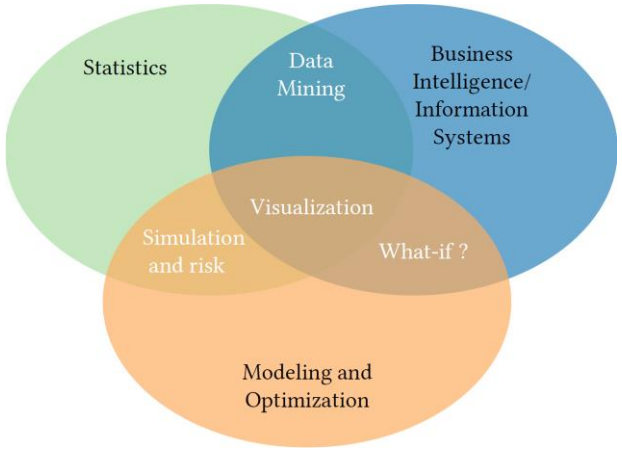
- Use data, information technology, and statistical analysis
- Make better, fact-based decisions

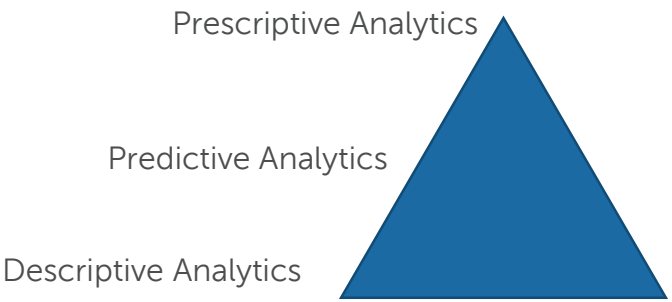
High Value for Investments [Evans, 2012]

- $10.66 for every $1 invested in analytics
- Companies create analytics departments
- Increase of professional with analytics expertise

Increase Analytical Capacity of Systems

- Analyzing past an current data …
- …taking into account further data…
- …suggesting actions for a given business goal

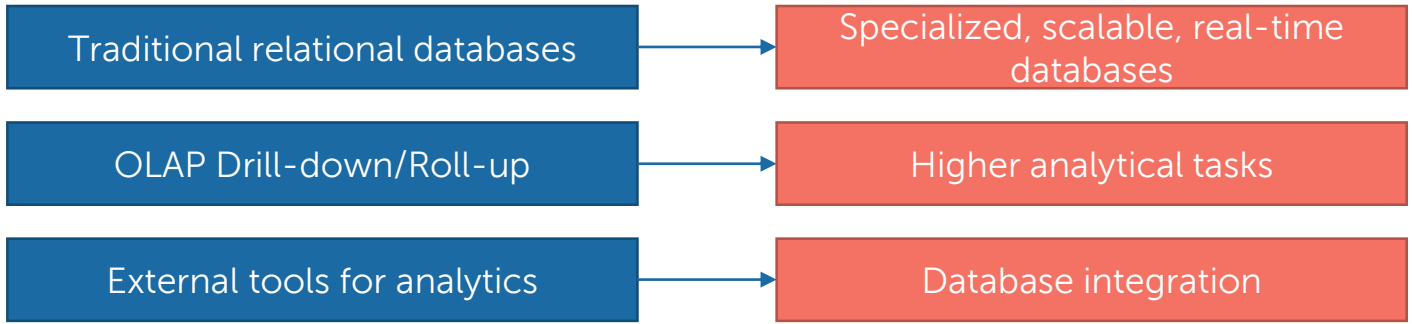




Effects on DBMS

Dresden Database Systems Group

Shift of DBMS Usage



Advantages

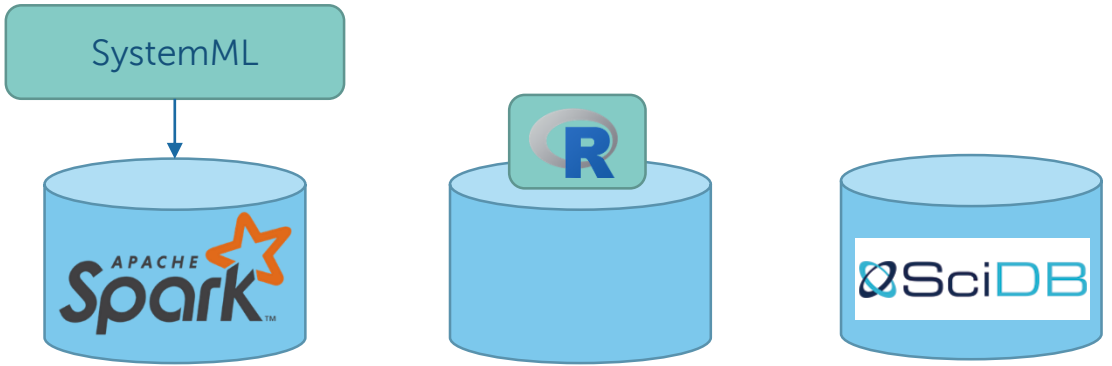
- No export of data
- Usage of database-specific optimizations
- Creation, Usage and Updates of statistical models within the database

Challenges for Databases

Dresden Database Systems Group

Architectural Integration

- No integration
- Partial integration
  - SQL and UDF usage
  - Proprietary Language
  - Bi-directional communiction
- Full integration



No Integration | Partial Integration | Full Integration

Analytical Methods

- Statistics (moments, interpolation, prinipal component analysis (PCA))
- Data Mining and Machine Learning (forecast, clustering, classification, association rule mining, and regression)
- Optimization problems

Goals of These Seminars

Dresden Database Systems Group

Architectural Integration

- Present the architecture and concepts of your seminar paper!
- How do the authors of your paper integrate analytical methods?
- Which concepts of databases do they exploit?
- What are advantages and disadvantages

Analytical Tasks

- Which analytical tasks are focussed?
- Is this a narrow approach or a flexible approach?

Evaluation

- Present the authors' evaluation (if available)!

What is your opinion about the system?

Proseminar

Statistical Model Computation with UDFs

Dresden Database Systems Group

Architectural Integration	Partial Inte-gration (SQL)
Application in Analytics	Statistics

Overview

- Partial integration of analysis by exploiting UDFs for computing statistical models inside a DBMS
- Application on statistical models (linear regression, PCA, etc.)
- Comparison of run-rime against a C++ implementation

Please explain

- Overview over technique and examples
- How are models represented in the database?
- Advantages, disadvantages and limits of using UDFs for analysis algorithms
- Presentation of evaluation

Statistical Model Computation with UDFs

Carlos Ordonez

SolveDB

Dresden Database Systems Group

Architectural Integration	Partial Integration (SQL)
Application in Analytics	Optimization problems

Overview

- DBMS with full integration for solving optimization problems
- Integration of different solvers, an SQL-based syntax for optimization problems, and extension of query optimization

Please explain

- Presentation of the database stack in SolveDB
- Advantages, Disadvantages and Limitation of this integration scheme
- Presenation of evaluation result and comparison against other systems

SolveDB: Integrating Optimization Problem Solvers Into SQL Databases

Laurynas Šikšnys, Torben Bach Pedersen

Ricardo: Integrating R and Hadoop

Dresden Database Systems Group

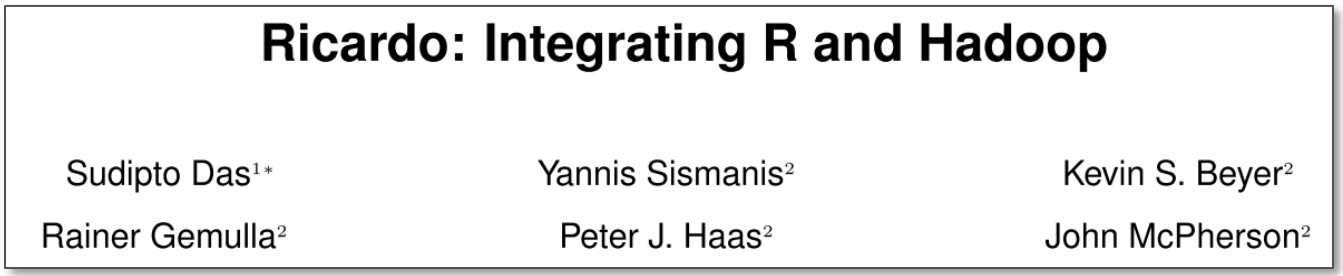
Architectural Integration	Partial Integration (Bi-directional)
Application in Analytics	Statistics

Overview

- Bi-directional communication
  - Hadoop as large-scale data management systems
  - R for complex analysis tasks

Please explain

- Classification of communication scheme
- Presentation of algorithm decomposition
- For which algorithms is this useful, for which not?
- Presentation of evalation



Efficient and Scaleable Dataflows

Dresden Database Systems Group

Architectural Integration	Processing API for Analytics
Application in Analytics	Processing Engine

Overview

- An efficient processing engine for deep learning

Please explain

- What is the difference between efficient and scaleable?
- How is the design of Project Adam specific to machine learning?
- Is the system design future proof?

Project Adam: Building an Efficient and Scalable Deep Learning Training System

Trishul Chilimbi Yutaka Suzue Johnson Apacible Karthik Kalyanaraman

Microsoft Research

Optimization for Machine Learning

Dresden Database Systems Group

Architectural Integration	No integration
Application in Analytics	Machine Learning

Overview

- Optimization of linear algebra expressions
- Similar to relational optimization in DBMS

Please explain

- What are the important optimization vectors?
- Why do these optimizations require a compilation step?

SPOOF: Sum-Product Optimization and Operator Fusion for Large-Scale Machine Learning

Tarek Elgamal[2]; Shangyu Luo[3]; Matthias Boehm[1], Alexandre V. Evfimievski[1], Shirish Tatikonda[4]‡ Berthold Reinwald[1], Prithviraj Sen[1]

Machine Learning on Distributed Systems

Dresden Database Systems Group

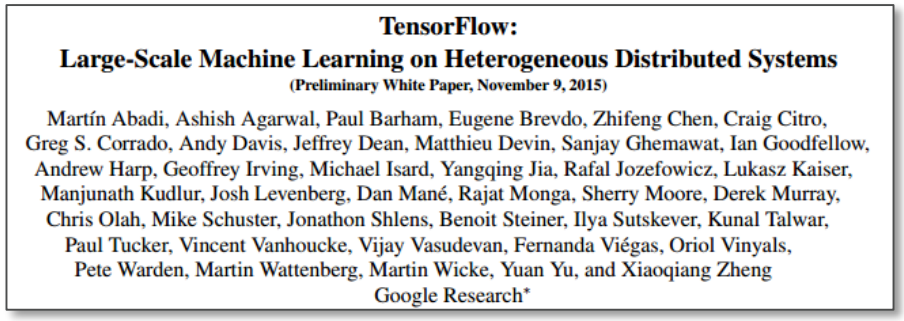
Architectural Integration	Full Integration
Application in Analytics	Machine Learning

Overview

- Dataflow engine for machine learning
- Scales to Google sized workloads

Please explain

- How does TensorFlow differ from general purpose dataflow engines?
- What programming model(s) does TensorFlow provide?
- Can TensorFlow optimize its workloads?



TensorFlow:
Large-Scale Machine Learning on Heterogeneous Distributed Systems
(Preliminary White Paper, November 9, 2015)

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng
Google Research*

COLUMBUS

Dresden Database Systems Group

Architectural Integration	No Integration
Application in Analytics	Feature Selection

Overview

- Feature selection arises in many analytical workflows
- COLUMBUS presents optimization techniques for feature selection workloads
  - Description of a feature-selection language
  - Materializations and caching results respecting human-in-the-loop process
  - Optimization techniques specific for feature selection with no correspondence to database optimizations

Please explain

- The architecture of COLUMBUS and the common workflow of feature selection
- Techniques of the COLUMBUS optimizer
- Their evaluation and performance compared to R and commercial RDBMS

Dresden Database
Systems Group

Hauptseminar

Declarative Machine Learning

Dresden Database Systems Group

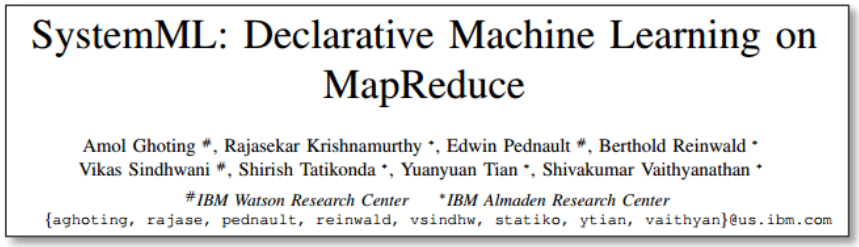
Architectural Integration	No integration
Application in Analytics	Machine Learning

Overview

- Machine learning on top of MapReduce/Spark
- Declarative programming environment

Please explain

- What does SystemML add to MapReduce?
- Why does it have to be declarative?
- How does SystemML support iterative algorithms?

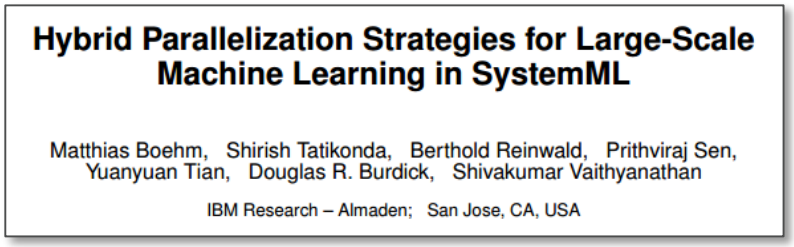


SystemML: Declarative Machine Learning on MapReduce

Amol Ghoting #, Rajasekar Krishnamurthy *, Edwin Pednault #, Berthold Reinwald *
Vikas Sindhwani #, Shirish Tatikonda *, Yuanyuan Tian *, Shivakumar Vaithyanathan *

#IBM Watson Research Center *IBM Almaden Research Center

{aghoting, rajase, pednault, reinwald, vsindhw, statiko, ytian, vaithyan}@us.ibm.com



Hybrid Parallelization Strategies for Large-Scale Machine Learning in SystemML

Matthias Boehm, Shirish Tatikonda, Berthold Reinwald, Prithviraj Sen,
Yuanyuan Tian, Douglas R. Burdick, Shivakumar Vaithyanathan

IBM Research – Almaden; San Jose, CA, USA

MLbase: A Distributed Machine-learning System

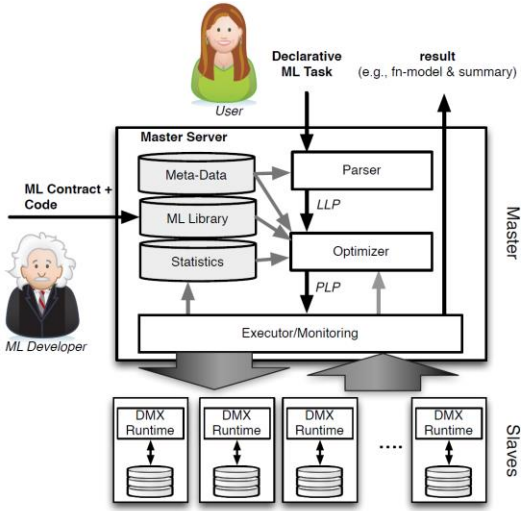
Dresden Database Systems Group

Overview

- Efficient and easy-to-use system for calculating ML tasks on distributed systems
- Declarative description of ML tasks
- Description of usual ML-related data structures (MLI)
- Optimization and model selection algorithms
- Covers a variety of application scenarios (Classification, Recommandation, Graph Analysis)

Please explain

- Present the system and its extensions
- Explain the workflow (declarative task, input structures, optimization and model selection, run-time evaluation) for a given an application scenario



Architectural Integration	No integration
Application in Analytics	Machine Learning

MAD – Magnetic, Agile, Deep

Dresden Database Systems Group

Architectural Integration	Partial Inte-gration (SQL)
Application in Analytics	Statistics

Overview

- Authors gathered requirements and reflexions of database analytics
- MADlib explores SQL extensions for calculating statistical methods (Regression, Classification, Clustering)
- Focus lies on strategies how to implement algorithms that read the data single pass, multi-pass (iterative) and state-full multi-pass

Please explain

- Overview over requirements for analytics in databases
- Present concepts how to translate algorithms in SQL and extensions
- How is MADlib (2012) used now?

BISMARCK: Towards a unified architecture for in-RDBMS analytics

Dresden Database Systems Group

Architectural Integration	Partial Integration (Low-Level)
Application in Analytics	Convex programming

Overview

- Database implementation of convex programming problems (SVM, regression, and others)
- Architecture makes use of user-defined aggregates, special data ordering and parallelization
- Prototype BISMARCK implemented in two commerical DBMSes and PostgreSQL

Please explain

- Presentation of database technologies BISMARCK makes use of
- Description of convex programming problems and applications
- Presenation of evaluation result and comparison against other systems

Array Oriented Database

Dresden Database Systems Group

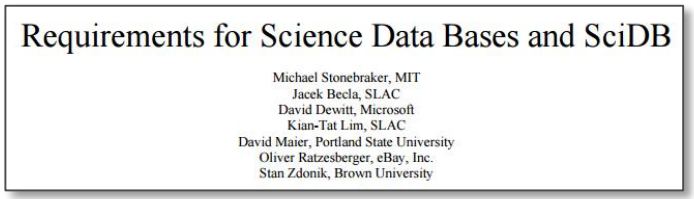
Architectural Integration	Full Integration
Application in Analytics	Fast Numeric Processing

Overview

- DBMS that stores arrays instead of relations
- Matrix and vector oriented query interface

Please explain

- Why do we need a special kind of databse?
- How do SciDB queries compare to SQL?
- Does SciDB scale to „Internet Size" workloads?

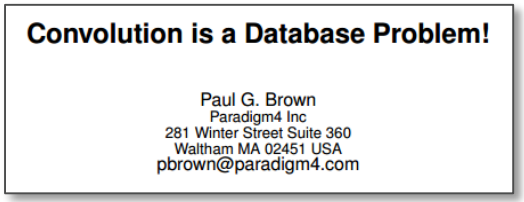

Requirements for Science Data Bases and SciDB

Michael Stonebraker, MIT
Jacek Becla, SLAC
David Dewitt, Microsoft
Kian-Tat Lim, SLAC
David Maier, Portland State University
Oliver Ratzesberger, eBay, Inc.
Stan Zdonik, Brown University


Overview of SciDB

Large Scale Array Storage, Processing and Analysis

The SciDB Development Team http://www.scidb.org


Convolution is a Database Problem!

Paul G. Brown
Paradigm4 Inc
281 Winter Street Suite 360
Waltham MA 02451 USA
pbrown@paradigm4.com

Iterative and Incremental Dataflows

Dresden Database Systems Group

Architectural Integration	Processing API for Analytics
Application in Analytics	Processing Engine

Overview

- Dataflow engines with support for iterative algorithms

Please explain

- Why do we need special support for iterative algorithms?
- How do Ciel and Naiad differ with regards to iteration?
- Which one is the better engine?

CIEL: a universal execution engine for distributed data-flow computing

Derek G. Murray Malte Schwarzkopf Christopher Smowton
Steven Smith Anil Madhavapeddy Steven Hand
University of Cambridge Computer Laboratory

Naiad: A Timely Dataflow System

Derek G. Murray Frank McSherry Rebecca Isaacs
Michael Isard Paul Barham Martín Abadi
Microsoft Research Silicon Valley
{derekmur,mcsherry,risaacs,misard,pbar,abadi}@microsoft.com

Comparison of Time Series Databases

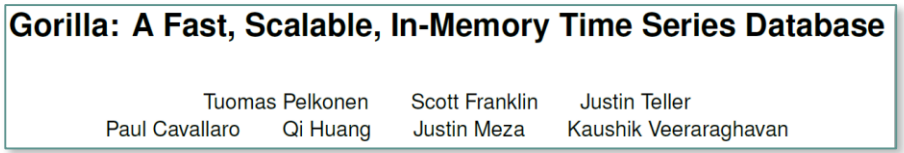
Dresden Database Systems Group

Overview

- Time Series Databases (TSDBs) are designed for
  - Efficiently storing time series and metadata
  - Supporting analytical tasks time series analysis
- Open/Closed-source TSDBs are
  - OpenTSDB
  - InfluxDB
  - Gorilla

Please explain

- The architecture of these TSDBs
- Their application domains
- Comparison of their concepts


Gorilla: A Fast, Scalable, In-Memory Time Series Database

Tuomas Pelkonen Scott Franklin Justin Teller
Paul Cavallaro Qi Huang Justin Meza Kaushik Veeraraghavan

A Sequence Database System

Dresden Database Systems Group

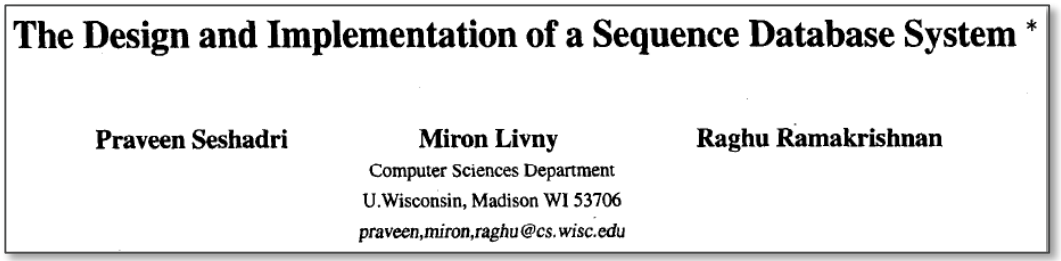
Architectural Integration	Full Integration
Application in Analytics	Statistics

Overview

- Design of a database with support for sequence data
- Development of the sequency query language SEQUIN
- Implementation of query optimization techniques that are specific for sequences

Task

- Presentation of …
  - Requirements for a sequence database
  - The query language, the design of the database and the optimization techniques
  - Evalation
- What has been done after this work?



The Design and Implementation of a Sequence Database System *

Praveen Seshadri **Miron Livny** **Raghu Ramakrishnan**
Computer Sciences Department
U.Wisconsin, Madison WI 53706
praveen,miron,raghu@cs.wisc.edu