*Prof. Dr.-Ing. Thomas Kudraß*
*HTWK Leipzig, F IMN*

# Themen im OS "Datenbanksysteme - Aktuelle Trends" SS2014

Nachfolgend finden Sie einige Hinweise und Fragestellungen zu den ausgegebenen Themen. Die mit * gekennzeichneten Materialien sind leihweise bei Prof. Kudraß erhältlich.

1. Big Data

In dem Überblicksvortrag ist das aktuelle Schlagwort "Big Data" näher zu beleuchten, also die Verarbeitung großer Datenmengen. Was kennzeichnet diese gegenüber herkömmlichen Datenbanktechnologien? Was treibt die Entwicklung voran? Folgende Aspekte sollte der Vortrag behandeln:

- Die vier V's, die drei F's
- Datenquellen und Kategorien der Verarbeitung von Big Data
- Anwendungsbereiche (einer davon etwas mehr im Detail)
- Überblick über Big-Data-Technologien
- Komponenten und Aufgaben einer Big-Data-Plattform
- Aspekte des Datenschutzes Abgrenzung: Geo-Informationssysteme (GIS)

*Quellen:*

- Informatik-Spektrum, Bd. 37, Heft 2, April 2014, Themenheft "Big Data" .

2. Smart Data / Data Streams

Der technologische Fortschritt im Bereich der Mikroelektronik und Kommunikationstechnik führt zunehmend zu stark vernetzten, mit Sensoren ausgestatteten verteilten Informationssystemen (Internet of Things). Die damit einhergehende steigende Anzahl an Sensorinformationen, deren Daten in Form von Datenströmen bereitgestellt werden, ermöglichen neue Anwendungsszenarien und treiben neue Verarbeitungstechniken. Smart Data können zur Verbesserung von Steuerungs- und Entscheidungsprozessen eingesetzt werden. Wesentliche Anwendungen hierfür sind Smart Cities oder das Smart Home.
Schwerpunktmäßig soll der Vortrag die Verarbeitung von Datenströmen (data streams) betrachten, bei dem kontinuierlich Anfragen an einen Strom von eingehenden Daten gestellt werden. Hierfür existiert auch der Begriff Complex Event Processing (CEP) kombiniert Ereignisse aus unterschiedlichen Quellen, um daraus bestimmte Muster abzuleiten, die auf ein relevantes Ereignis hindeuten, z.B. eine Bedrohungssituation, auf das umgehend reagiert werden muss.

Der Vortrag sollte auf folgende Aspekte eingehen:

- Anwendungsszenarien für CEP, / Data Streams insbesondere im Netzwerk- und Systemmanagement, Geschäftsprozessmanagement, Smart-City-Anwendungen und in der Finanzwirtschaft (Trading, Fraud Detection)

- Event Query Languages: Eventbegriff, Eventalgebra, Data Stream Query Language(CQL),
- Beziehung von CEP zu Zeitreihen-Datenbanken
- Forschungsprojekte zu CEP / Data Streams bzw. Smart*-Anwendungen

*Quellen:*

- K.P. Eckert, R. Popescu-Zeletin: Smart Data als Motor für Smart Cities, in: Informatik-Spektrum, Bd. 37, Heft 2, April 2014
- D. Luckham: The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems. Addison-Wesley Professional, 2002.
- M. Eckert, F. Bry: Complex Event Processing, in: Informatik-Spektrum: Bd. 32, Heft 2, 2009.

3. Cloud-Datenbanken (1 Vortrag)

Cloud Computing besitzt ein großes Potential für Unternehmen zur Reduktion ihrer Kosten sowie einer Verkürzung der Entwicklungszeiten für marktreife Produkte (Time-to-Market) durch Verschlankung notwendiger Hardware-Infrastruktur. Besonders betrachtet werden sollen Speicher- und Datenbank-Service, die von einer Cloud zur Verfügung gestellt werden können. Der Vortrag sollte auf folgende Aspekte eingehen:

- Einführung in das Cloud Computing: Klassifikation, Prinzipien und Vorteile
- Speicherkategorien in der Cloud: Blob Storage, Table Storage, Datenbankserver
- Überblick über Anbieter von Cloud-Datenbanken (insb. Amazon, Google, Microsoft)
- APIs, Datenmodelle und Speichermedien für Cloud-Datenbanken
- Bewertung der Speicherkategorien nach den Cloud-Computing-Kriterien: Elastizität hinsichtlich Datenvolumen, Ausfallsicherheit/Hochverfügbarkeit, Kosteneinsparung durch Elastizität, Administrationsaufwand
- Allgemeine Probleme von Cloud Storage: Partitionierung der Daten, Systemarchitektur (Konsistenzkontrolle), Skalierbarkeit, Performance, Migration,

*Quellen:*

- D. Kossmann, T. Kraska: Data Management in the Cloud: Promises, State-of-the-art, and Open Questions, in: Datenbank-Spektrum Bd. 10, Heft 3/Dezember 2010, Springer.*
- M.C. Jaeger, U. Hohenstein: Cloud Storage: Wieviel Cloud Computing steckt dahinter?, in: 14. Fachtagung Datenbanksysteme für Business, Technologie und Web (BTW), Kaiserslautern, 2011. *

4. NoSQL Datenbanken

NoSQL (zumeist interpretiert als "not only SQL") beschreibt ein breites Spektrum von Datenbankmanagementsystemen, die dadurch charakterisiert sind, dass sie nicht dem weitverbreiteten relationalen Datenmodell folgen. NoSQL Datenbanken operieren daher nicht primär auf

Tabellen und nutzen im Allgemeinen nicht SQL für den Datenzugriff. NoSQL-Datenbanksysteme sind oft optimiert für Anwendungen mit gleichzeitig hohen Datenanforderungen und häufigen Datenänderungen, wie sie im Web 2.0 auftreten. Sie zeichnen sich durch eine verbesserte (horizontale) Skalierbarkeit und Performance für bestimmte (nicht-relationale) Datenmodelle aus. Der Vortrag sollte auf folgende Aspekte eingehen:

- Historie von NoSQL, Grenzen bisheriger relationaler Systeme
- Motivation und Anwendungshintergrund (Web 2.0) für NoSQL-Datenbanken
- Theoretische Grundlagen, insbesondere Map/Reduce, CAP Theorem und Eventually Consistent als neuer Konsistenzbegriff, Consistent Hashing, Multiversion Concurrency Control, REST
- Kategorisierung von NoSQL-Systemen: Key/Value-Systeme, Column-Familiy-Systeme, Document Stores, Graphdatenbanken
- Vorstellung ausgewählter NoSQL-Systeme, zB. CouchDB, Neo4J ins. APIs

*Quellen:*

- S. Edlich, A. Friedland, J. Hampe, B. Brauer, M. Brückner: NoSQL : Einstieg in die Welt nichtrelationaler Web 2.0 Datenbanken. 2., aktualisierte und erweiterte Auflage. Hanser Verlag, München 2011.

5. Column Stores und Hauptspeicherdatenbanken

Traditionell werden Datenbankanwendungen in einem Unternehmen in OLTP (Online Trasactional Processing) und OLAP (Online Analytical Processing) unterteilt. OLTP- und OLAP-Systeme wurden in der Vergangenheit bereits sehr stark optimiert, die Leistung in entsprechenden Benchmarks bewertet. Dabei haben sich sowohl Hardware als auch Datenbanken weiter entwickelt. Einerseits gibt es DBMS, die Daten spaltenorientiert organisieren (Column Stores) und dabei ideal das Anforderungsprofil analytischer Anfragen abdecken. Andererseits steht heutzutage deutlich mehr Hauptspeicher zur Verfügung, der in Kombination mit der ebenfalls wesentlich gesteigerten Rechenleistung es erlaubt, komplette Datenbanken von Unternehmen komprimiert im Speicher vorzuhalten. Beide Entwicklungen ermöglichen die Bearbeitung komplexer analytischer Anfragen in Sekundenbruchteilen und ermöglichen so völlig neue Geschäftsanwendungen (z.B. im Bereich Decision Support). Der am Hasso-Plattner-Institut entwickelte Prototyp SanssouciDB vereinigt beide Konzepte und wurde bei SAP mittlerweile zur Produktreife unter dem Namen HANA geführt. Der Vortrag sollte auf folgende Aspekte eingehen:

- Hauptspeicherdatenbanken (In-Memory oder Main-Memory-Datenbanken)
- Spaltenorientierte Datenbanken (Column Stores)
- DBMS-Architektur am Beispiel von SanssouciDB
- Kompression in Datenbanken
- Insert-Only-Strategien
- Transaktionsmanagement

- Anfrageverarbeitung (Aggregation, Joins)
- Partitionierung und Replikation
- Anforderungen von Geschäftsanwendungen (z.B. Mahnungen, Available-to-Promise): Workload, Charakteristika von OLTP- und OLAP-Anwendungen

*Quellen*:

- J. Krueger, M. Grund, C. Tinnefeld, B. Eckart, A. Zeier, H. Plattner: Hauptspeicherdatenbanken für Unternehmensanwendungen - Datenmanagement für Unternehmensanwendungen im Kontext heutiger Anforderungen und Trends, in: Datenbank-Spektrum Bd. 10 Heft 3/Dez. 2010, Springer-Verlag. *
- H. Plattner: SanssouciDB: An In-Memory Database for Processing Enterprise Workloads, in: 14. Fachtagung Datenbanksysteme für Business, Technologie und Web (BTW), Kaiserslautern, 2011. *

6. Information Extraction

Information Extraction (IE) bezeichnet den Ansatz, strukturiertes Wissen aus unstrukturierten oder bestenfalls semi-strukturierten Daten (z.B. HTML- oder XML-Dokumente) zu gewinnen. Intelligente Informationsextraktionstechniken sind dabei die wichtigsten Bestandteile bei der Generierung und Repräsentation von Wissen für eine Vielzahl von Anwendungen, insbesondere bei der Auswertung des World Wide Web als weltgrößtem Informationsbestand.

Der Vortrag sollte folgende Schwerpunkte umfassen:

- Einordnung und Abgrenzung von IE gegenüber anderen Teilgebieten der Informatik: Natural Language Processing (NLP), Machine Learning, Text Mining, Information Retrieval
- Historie: Message Understanding Conferences (MUC)
- Anwendungen
- Extraktion von (named) Entities und Beziehungen, Attribute und Klassen von Entities
- Extraktionstechniken: Klassifikatoren, Sequenz-Modelle (Hidden Markov Modelle)
- hybride Techniken unter Einbeziehung von menschlicher Interaktion
- semantische Aspekte der Informationsextraktion
- Bewertungskriterien bei der Informationsextraktion
- Werkzeuge zur Informationsextraktion (z.B. Open-Source-Tool GATE, System T von IBM)

*Quellen:*

- W.-T. Balke: Introduction to Information Extraction: Basic Notions and Current Tremds, in: Datenbank-Spektrum Bd. 12 Heft 2, 2012. *
- P. Klügl, M. Toepfer: Informationsextraktion, in Informatik-Spektrum Bd. 37 Heft 2, 2014

- J. Piskorski, R. Yangarber: Information Extraction: past, Present and Future, in: Poibeau et. el. (eds.): Multisource, Multilingual Information Extraction and Summarization, Springer-Verlag, 2013, S. 23-48.

7. Datenbanken und Semantic Web am Beispiel von SPARQL

SPARQL ist eine graphbasierte Anfragesprache für Datenbanken, die es erlaubt, auf Daten lesend und schreibend zuzugreifen, die im Format des Resource Description Framework (RDF) gespeichert sind. SPARQL kann auch in andere Anfragesprachen wie SQL oder XQuery übersetzt werden. SPARQL-Anfragen werden an sogenannte SPARQL-Endpoints gesendet, die in der Lage sind, derartige Anfragen zu interpretieren und zu beantworten. Seit 2008 ist SPRQL auch eine offizielle W3C Recommendation und wird als eine der wichtigen Komponenten im Semantic Web der Zukunft beatrachtet. Der Vortrag sollte auf folgende Aspekte eingehen:

- Grundlagen von semantischen Daten: Begriff Semantik, Modellierung
- Resource Description Framework (RDF)
- Anfragetypen in SPARQL, Beispiel-Anfragen
- Semantik von SPARQL
- Erweiterungen von SPARQL
- Datenbanken zum Speichern und Anfragen von RDF-Daten
- Verfügbare SPARQL-Engines (Demo einer Open-Source-Implementierung wünschenswert)

*Quellen:*

- Datenbank-Spektrum, Bd. 13, Heft 2, Juli 2013, Schwerpunkt RDF Data Management, u.a. A. Brodt, B. Mitschang: Effiziente Verarbeitung allgemeiner Anfragen in RDF Trople Stores, in
- T. Segaran, C. Evans, J. Taylor: Programming the Semantic Web, O'Reilly, 2009.

8. Geodatenbanken (1 Vorträge)

Geodatenbanken sind ein wesentlicher Bestand von Geoinformationssystemen (GIS) und anderen Anwendungen, die räumliche Daten (Geodaten) verarbeiten. Sie dienen der Modellierung, der Speicherung und der Anfrage von Geodaten.

In einem Überblicksvortrag sollten folgende Aspekte behandelt werden:

- Einordung und Abgrenzung: Geo-Informationssysteme (GIS)
- Geodaten: Eigenschaften, Metadaten
- Standardisierung von Geodatenmodellen: Datenschemata
- Funktionalität von Geodatenbanksystemen
- Räumliche Datenbankanfragen

- Räumliche Indexe
- Geocoding
- Produkte (insbesondere Oracle Spatial), ausführliche Produktvorstellung als 2. Vortrag möglich

*Quellen:*

- T. Brinkhoff: Geodatenbanksysteme in Theorie und Praxis, Wichmann Verlag, 2005.
- T. Brinkhoff: Geodatenbanken, in: T. Kudraß (Hrsg.): Taschenbuch Datenbanken, Hanser-Verlag 2007.
- R. Kothuri, A. Godfrind, E. Beinat: Pro Oracle Spatial, Dokumentation Oracle Spatial Reference and User's Guide, Apress, 2004.*

9. Temporale Datenbanken (1 Vortrag)

Temporale Datenbanksysteme unterstützen die Verarbeitung und Speicherung von zeitbezogenen (temporalen) Daten über die zeitbezogene Datentypen hinaus. Derzeit existiert kein kommerzielles DBMS, das die Anforderungen der temporalen Datenhaltung vollständig abbildet. Allerdings gibt es Komponenten für bestimmte Arten von temporalen Daten, z.B. Oracle Time Series für die Verarbeitung von Zeitreihen auf der Basis von spezifischen Time-Series-Datentypen.

Der Vortrag sollte auf folgende Aspekte eingehen:

- Basiskonzepte temporaler Datenbanken: Gültigkeitszeit/Aufzeichnungszeit, temporale Datentypen, Historisierung, Kalender, Zeitstempel
- Integrität in temporalen Datenbanken
- Abbildung auf herkömmliche relationale Datenbanken
- Aktueller Stand der Standardisierung in SQL:2011, SQL-Erweiterungen
- Unterstützung in DBMS-Produkten, insbesondere IBM DB2 V10

*Quellen:*

- D. Petkovic; Was lange währt, wird endlich gut: Temporale Daten in SQL-Standard, in: Datenbank-Spektrum, Bd. 13, Heft 2, 2013.
- R.T. Snodgrass: The TSQL2 temporal query language, Springer-Verlag, Berlin 1995.
- R.T. Snodgrass, Michael Böhlen, Christian S. Jensen, Andreas Steiner: Transitioning Temporal Support in TSQL2 to SQL3, 1997, TIMECENTER Technical Report TR-8
- K. Kulkarni, J. Michels: Temporal Features in SQL:2011, SIGMOD Record Vol. 41, No. 3, 2012, S. 34-43.

10.Mobile Datenbanken und Informationssysteme

Die weite Verbreitung von mobilen Endgeräten wie Mobiltelefonen, Smartphones, Laptops und Tablet PCs in Verbindung mit ihrer stetig zunehmenden Leistungfähigkeit ermöglicht den Zugriff auf Informationen jederzeit von überall her. Dabei zeigt sich, daß die Portabilität mo-

biler Endgeräte und die Eigenarten von drahtlosen Netzwerken neue Fragestellungen aus der Sicht der Informationsverarbeitung aufwerfen. In einem Überblicksvortrag sind die wichtigsten Problemstellungen und Lösungsansätze auf dem Gebiet der mobilen Datenbanken und Informationssysteme darzustellen.

Als ein konkretes Produkt unterstützt Sybase mit seinem Produkt SQL Anywhere die Nutzung mobiler Datenbanken auf Laptops oder Smartphones und ist damit Marktführer bei mobilen Datenbanken.
Folgende Konzepte sollten enthalten sein:

- ortsabhängige Anfragen und Anfragen an bewegliche Objekte
- Replikation und Synchronisation
- mobile Transaktionen
- Architektur mobiler Datenbanksysteme
- Administration: System Management, User und Device Management
- Zuverlässigkeit / Skalierbarkeit
- Auffinden, Verwalten und Verbreiten von Informationen in mobilen, drahtlosen Umgebungen
- Systemüberblick Sybase SQL Anywhere

*Quellen*:

- H. Höpfner, C. Türker, B. König-Ries: Mobile Datenbanken und Informationssysteme, dpunkt Verlag, 2005.*
- SQL Anywhere Produktseite: http://www.sybase.com/products/databasemanagement/sqlanywhere

11.Objektdatenbanken am Beispiel db4o

Objektdatenbanken waren in den 1990-er Jahren ein großer Trend und beeinflusste die Weiterentwicklung relationaler Datenbanksysteme hin zu objektrelationalen Systemen. Heutzutage haben Objektdatenbanken als embedded Databases ein neues Anwendungsgebiet mit Wachstumspotential gefunden. Der Vortrag sollte auf folgende Aspekte eingehen:

- Basiskonzepte objektorientierter Datenbanken (auch in Abgrenzung zu objektrelationalen Datenbanken), insbesondere Persistenz
- Modellierung von Beziehungen in Objektdatenbanken
- API für einen Objektlebenszyklus (CRUD-Operationen) am Beispiel von db4o
- Anfrageschnittstellen: QBE (Query By Example), S.O.D.A. / Criteria Queris, Native Abfragen
- Transaktionen in db4o
- Client/Server-Modes in db4o

- weitere interessante Eigenschaften (Replikation, Callbacks, Ladeverhalten)

*Quellen*:

- I. Brenner: Datenbankentwicklung mit db4o - Einführung in eine objektorientierte Datenbank, online unter www.inabrenner.de
- http://odbms.org (Portal rund um das Thema Objektorientierte Datenbanken)

12.Wissenschaftliche Datenbanken (Scientific Data Management)

Herkömmliche Datenbanken konzentrieren sich auf Verwaltung und Analyse von geschäftsorientierten Daten. Dem steht aber eine noch größere Menge an wissenschaftlichen Daten gegenüber, die bisher bei DB-Forschung und Entwicklung nicht ausreichend berücksichtigt wurden. Die effiziente Verwaltung, Speicherung, Suche und Analyse wissenschaftlicher Daten stellt eine immense Herausforderung an diese und verwandte Bereiche der Naturwissenschaften dar. Wie kann man effektiv neues Wissen aus den Daten ableiten? Wo stoßen aktuelle Systeme an ihre Grenzen?
Durch den immensen Fortschritt bei der Instrumentierung von Experimenten, Simulation und Beobachtungen in allen Bereichen der Naturwissenschaften entstehen neue Herausforderungen für Datenbanktechnologien. Die bei Experimenten anfallenden Daten entstehen dabei oft schneller als sie verarbeitet werden können, was zu Bottlenecks führen kann. Wissenschaftliche Daten sind typischerweise sehr heterogen und komplex, erfordern neue Datenstrukturen und Zugriffsmuster. Dies bewirkt neue Aspekte der Zugriffsoptimierung und Datenintegration. Um rechenintensive und datenintensive Abläufe bei der Verarbeitung wissenschaftlicher Daten zu beschreiben, sind Scientific-Workflow-Technologien zu entwickeln, die sich von herkömmlichen business-orientierten Workflows unterscheiden. Dazu zählt insbesondere das Problem der Datenherkunft (Data Lineage). Der Vortrag sollte auf folgende Aspekte eingehen:

- Beispiele für naturwissenschaftliche Anwendungen: Biologie (Genetik, Moleekularbiologie), Astronomie, Meterologie,
- Datenbanksysteme vs. Dateisysteme für wissenschaftliche Anwendungen
- Bedeutung von Metadaten
- Data Lineage, Data Provenance
- Mengenorientierte Verarbeitung, Paraellelisierung, Map/Reduce
- Scientific Workflows: Modelle, Design, offene Fragen

*Quellen:*

- J. Gray, D. Liu, M. Nieto-Santisteban, A. S. Szalay, D. DeWitt, G. Heber: Scientific Data Management in the Coming Decade, SIGMOD Record, Vol. 34 No. 4, 2005.
- V. Cuevas-Vicenttin, S. Dey, S. Köhler, S. Riddle, B. Ludäscher: Scientific Workflows and Provenance: Introduction and Research Opportunities, Datenbank-Spektrum, Bd. 12, Heft 3, 2012.