Coping with Unstructured Dynamic Data Sets – NoSQL: a Buzzword or a Savior?

Dr. Johannes Scholz

Vienna University of Technology
Institute of Geoinformation and Cartography
scholz@geoinfo.tuwien.ac.at

TU WIEN

REAL CORP 2011

Geo info
Institute for Geoinformation and Cartography

Contents

Introduction I

- Contemporary Relational Database Management Systems (RDBMS)
  - follow the relational model (Codd 1980) – relational algebra
  - follow the table metaphor
- ACID properties (Atomitcity, **Consistency,** Isolation, Durability) (Gray 1983) are „holy grail“
  - Consistency: „*serious practical problem as more and more different types of data are integrated into common data banks*“ (Codd 1970, p. 387)
- Web 2.0 – challenges for RDBMS
  - Increasing number of users and large data volumes
    → Facebook, Amazon, Google
  - **NoSQL databases** emerged (Evans 2009)

Introduction II

- Research question:
  - Applicability of NoSQL databases – especially Document Stores– in cadastral document collection
- Why of interest?
  - Cadastral document collection – unstructured dataset
  - Cadastral document collection contains data necessary for spatial planning
  - Cadastral document collection shares commonalities with Web 2.0

Challenges for RDBMSs

- Relational concept has a certain „age“ – given the first publication (Codd 1970)

- Upcoming Web 2.0 challenges RDBMS (Agrawal, Ailamaki et al. 2008):
  - Frequent read/write with low data volumes vs. data intensive tasks
  - ACID vs. horizontal scaling over a great number of network nodes
  - Structured vs. unstructured data
    - Predefined data schema are problematic in Web 2.0
  - Static vs. mobile applications

NoSQL – Global Characteristics

- Edlich, Friedland et al. (2010) define NoSQL, as databases following (some) of the principles (left) and implementing key concepts (right):

NoSQL principles

- Non-relational data model
- Tailored towards distributed and horizontal scalability
- Open source
- Schema free or at least weak schema restrictions
- Support for a simple replication approach
- Simple application programming interface
- Other consistency approaches than ACID are used
  - *eventual consistency*
  - *basically available*, *soft state* *eventually consistent* (BASE)
  - **not ACID!**

NoSQL key concepts

- Map reduce
- Eventual Consistency
- Consistent hashing
- Multiversion Concurrency Control System
- Vector Clocks

NoSQL – Database Types

- Literature distinguishes four basic types:
  - Wide Column Stores
    - "*spars* … al. 2006)
    - Goog…
  - Docum…
    - Hand…
    - Datab…
  - Key/Va…
    - Stora…
  - Graph …
    - Follo…
      - Nodes
      - Edges …

```
"firstName": "Johannes",
"lastName": "Scholz",
"academicTitle": "DI(FH) Dr.",
"Position": "University Assistant",
"University": "Vienna University of Technology",
"Institute": "Department of Geoinformation and Cartography",
"Address":
{
    "streetAddress": "Gusshausstrasse 27-29 E127",
    "city": "Vienna",
    "ZIP": "1040",
    "country": "Austria"
},
"phoneNumber": "0043-1-58801-12721",
"faxNumber": "0043-1-58801-12799",
"Email": "scholz@geoinfo.tuwien.ac.at",
"Website": "www.geoinfo.tuwien.ac.at"
```

…hDB relax

NoSQL for Spatial Domain? Cadastral Document Collection

- Analyze the applicability of NoSQL Databases
  - cadastral document collection – in particular on the purchase of land contracts
    - → part of the document collection (Republik Österreich 2010; Kodek 2007; Feil, Marent et al. 2005)

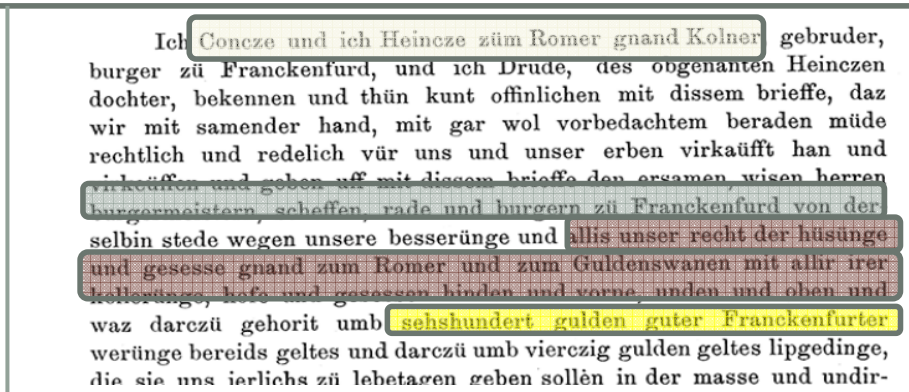
- Analysis of purchase of land contracts:
  - Contract items (Österreichische Notariatskammer 2011)
  - Relevance for spatial planning
  - Fixed/no fixed schema

contract item	relevant for spatial planning	fixed schema	no fixed schema
vendor		✔	
purchaser		✔	
property	✔		✔
legal situation	✔		✔
infrastructure	✔		✔
purchase price		✔	
payment modalities		✔	
defects liability	✔		✔
other registers			✔

NoSQL for Spatial Domain? Cadastral Document Collection

- Comparison of historical and contemporary documents in document collection

Purchase contract of "Römer" building (1405) – town hall of Frankfurt/Main



Ich Concze und ich Heincze züm Romer gnand Kolner gebruder, burger zü Franckenfurd, und ich Drude, des obgenanten Heinczen dochter, bekennen und thün kunt offinlichen mit dissem brieffe, daz wir mit samender hand, mit gar wol vorbedachtem beraden müde rechtlich und redelich vür uns und unser erben virkaüfft han und virkeüffen und geben uff mit dissem brieffe den ersamen, wisen herren burgermeistern, scheffen, rade und burgern zü Franckenfurd von der selbin stede wegen unsere besserünge und allis unser recht der hüsunge und gesesse gnand zum Romer und zum Guldenswanen mit allir irer hofereünge, hofe und gesessen hinden und vorne, unden und oben und waz darczü gehorit umb sehshundert gulden guter Franckenfurter werünge bereids geltes und darczü umb vierczig gulden geltes lipgedinge, die sie uns ierlichs zü lebetagen geben sollèn in der masse und undir-

Legend

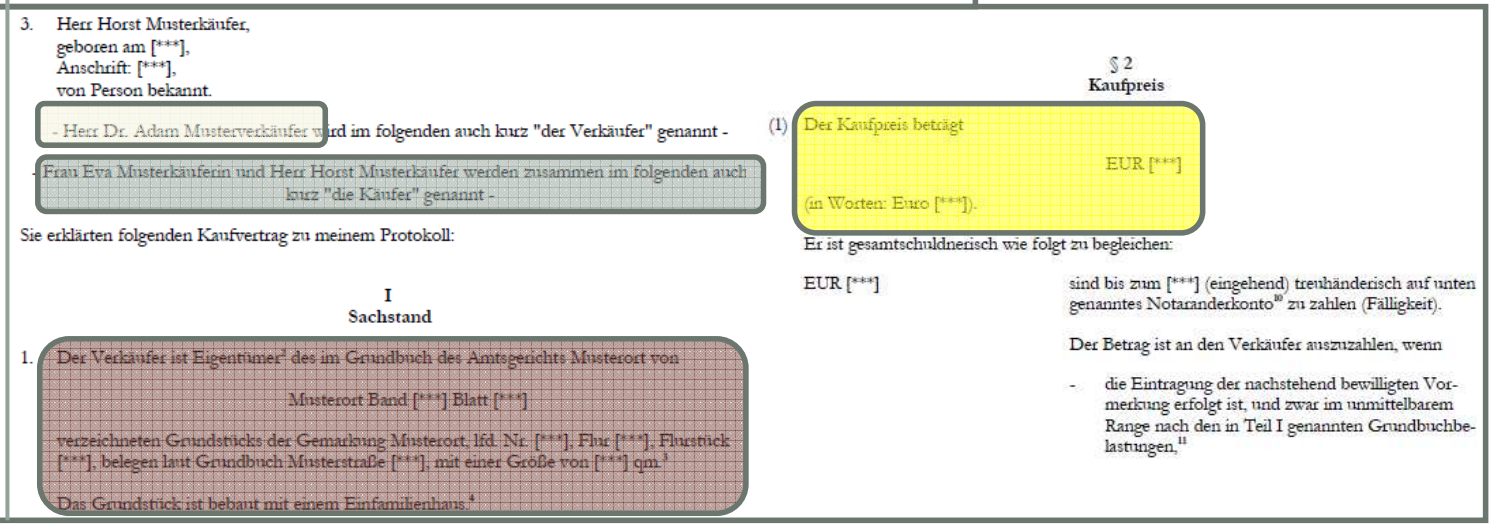
vendor

purchaser

property description

purchase price

Contemporary purchase of land contract example



3. Herr Horst Musterkäufer,
geboren am [***],
Anschrift: [***],
von Person bekannt.

- Herr Dr. Adam Musterverkäufer wird im folgenden auch kurz "der Verkäufer" genannt -

- Frau Eva Musterkäuferin und Herr Horst Musterkäufer werden zusammen im folgenden auch kurz "die Käufer" genannt -

Sie erklärten folgenden Kaufvertrag zu meinem Protokoll:

I
Sachstand

1. Der Verkäufer ist Eigentümer[2] des im Grundbuch des Amtsgerichts Musterort von Musterort Band [***] Blatt [***] verzeichneten Grundstücks der Gemarkung Musterort, lfd. Nr. [***], Flur [***], Flurstück [***], belegen laut Grundbuch Musterstraße [***], mit einer Größe von [***] qm.[3]

Das Grundstück ist bebaut mit einem Einfamilienhaus.[4]

§ 2
Kaufpreis

(1) Der Kaufpreis beträgt
EUR [***]
(in Worten: Euro [***]).

Er ist gesamtschuldnerisch wie folgt zu begleichen:

EUR [***] sind bis zum [***] (eingehend) treuhänderisch auf unten genanntes Notaranderkonto[10] zu zahlen (Fälligkeit).

Der Betrag ist an den Verkäufer auszuzahlen, wenn

- die Eintragung der nachstehend bewilligten Vormerkung erfolgt ist, und zwar im unmittelbarem Range nach den in Teil I genannten Grundbuchbelastungen,[11]

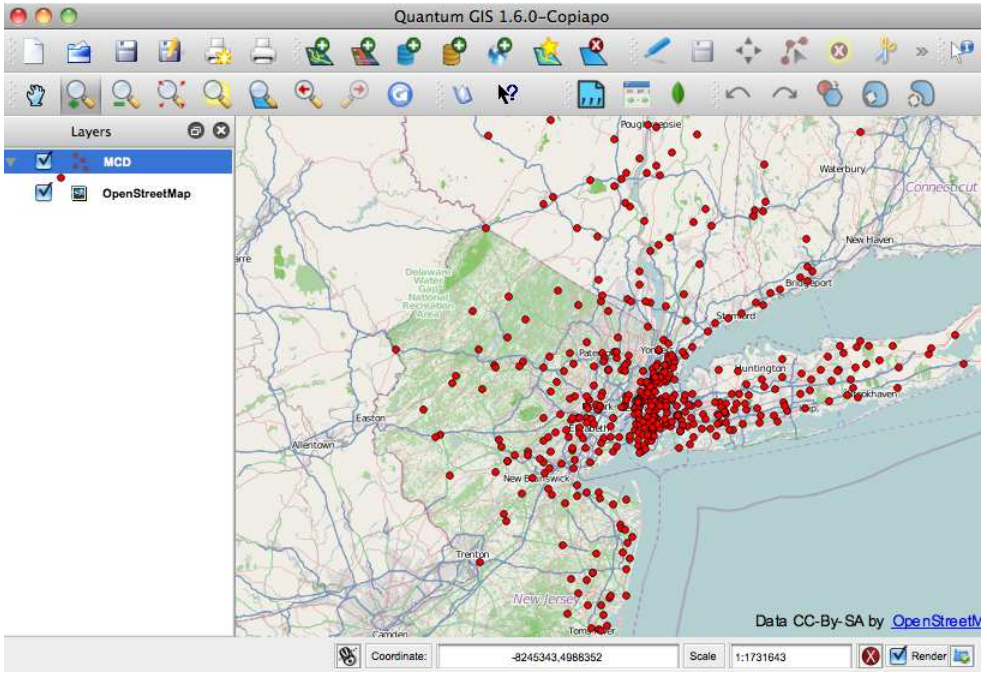
NoSQL for Spatial Domain? Cadastral Document Collection

- Why NoSQL can be important?
  - Historic documents show genesis of parcels
  - **Syntax changes** while **semantics** is **fixed** – thus of **unstructured nature**!
  - Cadastre is a “living” environment – constant change
    - purchase, subdivision, merge, creation or deletion processes
    - Austrian Land register: 684.000 new entries, 12 mio queries (for 2009)
  - Similarities between NoSQL document storages and document collection

	Document collection	Document Database (NoSQL)
purpose	storage and retrieval of documents	storage, retrieval and query mechanisms for documents
nature of data	predominantly unstructured	structured and unstructured (schema free)
dynamic behavior	yes (at least to a certain extent)	supports dynamic processes
multiple versions	may exist	multiple versions are supported (MVCC)
concurrent processes	may exist	handled by MVCC
data security	important	supports distributed data storage

„Spatial NoSQL“

- Currently the following NoSQL Projects support spatial data and spatial query mechanisms (at least rudimentary)
  - CochDB with GeoCouch (Mische 2010, 2011)
    - Document Store
  - mongoDB (10gen Inc. 2011)
    - Document Store
    - QGIS Plugin for mongoDB
  - Neo4j using Neo4j spatial (Neo Technology 2011)
    - Graph database



Source: http://geokoder.com/mongodb-plugin-for-quantum-gis

Conclusion

- General analysis of NoSQL Databases
- Properties of Cadastral Document Collection
  - Certain dynamic
  - Unstructured datasets – syntax changes whereas semantics remains invariant!
    - →Comparison: historic vs. contemporary purchase of land contracts
- NoSQL for document collection?
  - NoSQL (document stores) meets needs of document collection
  - NoSQL is not there to remove RDBMS – but to serve as an alternative for certain tasks and are offering new possibilities

References

- 10GEN INC.: mongoDB Manual: Geospatial Indexing, Url: http://www.mongodb.org/display/DOCS/Geospatial+Indexing, visited: 24-02-2011, 2011.
- ABADI, D.: Column-Stores For Wide and Sparse Data. In: Proceedings of the 3rd Biennial Conference on Innovative Data Systems Research (CIDR), Url: http://db.csail.mit.edu/projects/cstore/abadicidr07.pdf, visited 17-02-2011, 2007.
- AGRAWAL, R., AILMAKI, A., BERNSTEIN, P. A., BREWER, E. A., CAREY, M. J., CHAUDHURI, S., DOAN, A., FLORESCU, D., FRANKLIN, M. J., GARCIA-MOLINA, H., GEHRKE, J., GRUENWALD, L., HAAS, L. M., HALVEY, A. Y., HELLERSTEIN, J. M., IOANNIDIS, Y. E., KORTH, H. F., KOSSMANN, D., MADDEN, S., MAGOULAS, R., OOI, B. C., O'REILLY, T., RAMAKRISHNAN, R., SARAWAGI, S., STONEBAKER, M., SZALAY, A. S., WEIKUM, G.: The Claremont report on database research. ACM SIGMOD Record, Volume 37, Issue 3, 9-19, 2008.
- ANDERSON, J. C., LEHNARDT, J., SLATER. N: CouchDB: The Definitive Guide. O'Reilly Media, Inc., 2010.
- ANGLES, R., GUTIERREZ, C.: Survey of graph database models. ACM Computing Surveys, Vol. 40, Issue 1, pp. 1:1-1:39, 2008.
- CHANG, F., DEAN, J., GHEMAWAT, S., HSIEH, W. C., WALLACH, D. A., BURROWS, M., CHANDRA, T., FIKES, A., GRUBER, R. E.: Bigtable: A distributed storage system for structured data. In: Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation OSDI'06, pp. 205–218, 2006.
- CODD, E.F.: A Relational Model of Data for Large Shared Data Banks. In: Communications of the ACM, Vol. 13, Issue 6, pp. 377–387, 1970.
- DEAN, J. and GHEMAWAT, S.: MapReduce: Simplified Data Processing on Large Clusters. In: OSDI'04, 6th Symposium on Operating Systems Design and Implementation, Url: http://labs.google.com/papers/mapreduce-osdi04.pdf, visited: 17-02-2011, 2004.
- EDLICH, S., FRIEDLAND, A., HAMPE, J., BRAUER, B.: NoSQL: Einstieg in die Welt nichtrelationaler Web 2.0 Datenbanken. Hanser Fachbuchverlag, München, 2010.
- EVANS, E.: NOSQL 2009. Url: http://blog.sym-link.com/2009/05/12/nosql_2009.html, visited: 16-02-2011, 2009.
- FEIL, E., MARENT, K.-H., PREISL, G.: Grundbuchsrecht. Linde Verlag, Vienna, 2005.
- GRAY, J.: The Transaction Concept: Virtues and Limitations. In: Proceedings of the 7th International Conference on Very Large Databases, pp. 144–154, Cannes, France, 1981.
- KARGER, D., LEHMAN, E., LEIGHTON, T., PANIGRAHY, R., LEVINE, M., AND LEWIN, D.: Consistent hashing and random trees: distributed caching protocols for relieving hot spots on the World Wide Web. In: Proceedings of the Twenty-Ninth Annual ACM Symposium on theory of Computing. STOC '97. ACM Press, pp. 654-663, New York, 1997.
- KHOSHAFIAN, S., COPELAND, G., JAGODIS, T., BORAL H., VALDURIEZ, P.: A query processing strategy for the decomposed storage model. In: ICDE, pp. 636–643,1987.
- LAMPORT, L.: Time, Clocks, and the ordering of events in a distributed system. Communications of the ACM, Volume 21, Issue 7, pp. 558-565, 1978.
- MISCHE, V.: GeoCouch: A spatial index for CouchDB. Presentation at FOSS5G 2010, Url: http://2010.foss4g.org/presentations/3048.pdf, visited: 24-02-2011, 2010.
- NEO TECHNOLOGY INC.: Neo4j spatial wiki. Url: http://wiki.neo4j.org/content/Neo4j_Spatial, visited: 24-02-2011, 2011.
- ÖSTERREICHISCHE NOTARIATSKAMMER: Checkliste Kaufvertrag/Liegenschaft, Url: http://www.notar.at/notar/de/home/infoservice/checklisten/kaufvertrag__liegenschaft/, visited: 17-02-2011, 2011.
- ÖSTERREICHISCHES JUSITZMINISTERIM: IT-Anwendungen in der österreichischen Justiz. Url: http://www.justiz.gv.at/internet/file/8ab4ac8322985dd501229ce3fb1900b4.de.0/folder_justiz-online_0310_de.pdf, visited: 23-02-2011, 2010.

Coping with Unstructured Dynamic Data Sets – NoSQL: a Buzzword or a Savior?

Dr. Johannes Scholz

Vienna University of Technology
Institute of Geoinformation and Cartography
scholz@geoinfo.tuwien.ac.at

TU WIEN

REAL CORP 2011

Geo info
Institute for Geoinformation and Cartography