Single Point of Entry
Integrating relational and semi-structured data with PostgreSQL

Dr. Ernst-Georg Schmid / NoSQL, Cologne / April 2014
The Pink Pineapple Problem
The Pink Pineapple Problem

...AGCTTTGACCTAGAGATCCGT
CAATGAGCTAGGCTACATGAA
AGCTTTGACCTAGAGATCCGT
CAATGAGCTAGCTTATCT...

A  B  C
5.2345374 3.2543526 0.213541
sigma+1 alpha rho 2.2 1 0.75
The Pink Pineapple Problem
The Pink Pineapple Problem
The Pink Pineapple Problem

AGCTTTGACCTAGAGATCCG
CAATGAGCTAGGCCTACATGAA
AGCTCCTGACCTAGAGGTCCGA
CAATGAGCTAGCTTATCT...

A  B  C
5  4

A  B  C
3  2

A  B  C
1  0.2

2  2

sigma+1  alpha  rho
2.2  1  0.75

Ernst-Georg Schmid • NoSQL Matters, Cologne • April 2014
The Pink Pineapple Problem
Diverse data along the way

- Text
- Undirected cyclic graphs
- Uncompressed images
- Sets of mixed types
- Number arrays
- Trees

AGCTTTGACCTAGAGATCCGT
CAATGAGCTAGGCCTACATGAA
AGCTTTGACCTAGAGATCCGT
CAATGAGCTAGCTTATCT...

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2</td>
<td>3.2</td>
<td>0.2</td>
</tr>
<tr>
<td>sig 5.4 A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>2.2 sigma+1 alpha rho</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Ernst-Georg Schmid • NoSQL Matters, Cologne • April 2014

Bayer Business Services
Growing data along the way

600 GB per sequencer per week
+300TB per year

10 million +600000 per year

300TB +300TB per year

> 1 million nodes
Heterogenous storage

...AGCTTTGACCTAGAGATCCGT
CAATGAGCTAGGCCTACATGAA
AGCTTTGACCTAGAGATCCGT
CAATGAGCTAGCTTATCT...

5.2345374 3.2543526 0.213541
sigma+1  alpha  rho
2.2  1  0.75

Ernst-Georg Schmid • NoSQL Matters, Cologne • April 2014
Problem centric integration
Problem centric integration

Every organization that tried to integrate all of this data in a single database…

…has failed!
Problem centric integration

Still, some kind of integration is needed

- One system as Single Point of Entry
- SQL as lingua franca
- Broad driver support
- Permanently move data only when needed (e.g. for archival)
- „Projection first“ queries:

„Move computation, not data!“
PostgreSQL

Why PostgreSQL?

Because it’s the world’s most advanced open source database. ;-)

And here are some less bragging reasons…
XML

XML is a datatype to store XML data in a database column

- Checks for well-formedness
- Type-safe operations

XML Support is output-biased

- Relation to XML mapping functions with matching XML Schema generation
- Checks for `xml_is_well_formed`, `IS DOCUMENT`, `XMLEXISTS`
- XPath 1.0 queries
- Aggregation function

- Validation only with DTD, no XML Schema support
- XML cannot be transparently indexed
- No comparison operators
- Memory intensive
- Encoding sensitive
JSON

JSON is a datatype to store JSON data in a database column

- Checks for valid JSON
- Type-safe operations

http://www.postgresql.org/docs/9.3/static/datatype-json.html
JSON support is almost complete

- Relation to JSON and JSON to relation mapping functions
- JSON object manipulation functions
- JSON object access operators
- Aggregation function

- JSON cannot be transparently indexed
HSTORE is a self-contained key/value store in a database column

- Key and value are strings
- Keys are hashed
- Arbitrary number of k/v pairs per HSTORE
- Keys are unique per HSTORE
- Values can be NULL

http://www.postgresql.org/docs/9.3/static/hstore.html
HSTORE support is complete

- Relation to HSTORE and HSTORE to relation mapping functions
- HSTORE key, value manipulation functions
- HSTORE access operators
- HSTORE to JSON mapping
- HSTORE can be transparently indexed
FDW + NoSQL

Foreign Data Wrappers map external data sources transparently into PostgreSQL

- Supports SELECT, INSERT, UPDATE, DELETE
- Datatype conversion
- Table size estimation
- Provide ANALYZE statistics
- Provide EXPLAIN information
- Impose user/role/privilege security

http://www.postgresql.org/docs/9.3/static/sql-createforeigndatawrapper.html
http://www.postgresql.org/docs/9.3/static/fdwhandler.html
CSTORE provides a columnar data store

- Implemented as FDW
- Column compression
- Column projection
- Full PostgreSQL integration on datatype and optimizer level

http://www.citusdata.com/blog/76-postgresql-columnar-store-for-analytics
JSONB is the fusion of JSON and HSTORE

- Checks for valid JSON
- Type-safe operations
- Full transparent indexing
- Complete semi-structured document store in PostgreSQL
- Comes with PostgreSQL 9.4 (3rd Quarter 2014)

http://obartunov.livejournal.com/177247.html
Single Point of Entry
Integrating relational and semi-structured data with PostgreSQL

Dr. Ernst-Georg Schmid / NoSQL, Cologne / April 2014