Topics

- Project History & Status
- Downloading the code
- Current 11.06 release
- Coming 11.12 release
- HDF5 and FITS loaders
- Global EPICS data archiver project
- Community feedback & requests
Project History

• Kicked-off October 2007 at XLDB1
  - First demo’d at VLDB and XLDB-3 in August ‘09

• New code base starting January 2010

• Paradigm4 is the sponsor
  - 15 employees are designing, implementing, testing, documenting

• Ideas & contributions from academics and gov’t labs

• Just getting project organization and infrastructure in place for wider community involvement
  - Forums, processes, project management committee
Integrated data management & advanced analytics platform

KEY FEATURES

- Array Oriented Data Model
- Data is updated, not overwritten
- Support for versions, provenance, time travel
- Massively scalable computations
- Scalable on commodity HW grid or cloud
- Extensible with UDTs, UDFs
- Native support for uncertainty
- Multiple flavors of ‘null’
- Two APIs: SQL-like (AQL) and functional (AFL)

Partner scientists in
Astronomy
High Energy Physics
Computational Biology
Oceanography
Earth Science
... and more
Project Status

• 3 releases this year (2011)
  - R0.75 in January
  - R11.06 in June
  - R11.12 coming in December

• Still not quite a “R1.0”
  - Expect that in Spring ’12

• Support for Ubuntu & RedHat

• Pilot projects underway
  - EPICs data archiver (SciDB)
  - Computational Genomics (P4)
  - Insurance Telematics (P4)
## Downloads

<table>
<thead>
<tr>
<th>Type</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Govt Lab</td>
<td>47</td>
<td>9%</td>
</tr>
<tr>
<td>Commercial</td>
<td>96</td>
<td>18%</td>
</tr>
<tr>
<td>Personal</td>
<td>138</td>
<td>26%</td>
</tr>
<tr>
<td>Academic</td>
<td>248</td>
<td>47%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>529</strong></td>
<td></td>
</tr>
</tbody>
</table>

- Register at forum
  - [www.scidb.org/forum](http://www.scidb.org/forum)
- Forum moderated daily
Available now: R11.06

• Data transformation
  - ‘redimension’ loads csv formatted data into multi-dimensional arrays

• *iquery* command line processor

• python connector

• Core set of AQL and AFL operations
  - AQL: SELECT FROM WHERE GROUP BY JOIN aggregates et al
  - AFL: subsample, regrid, lookup, project, explain, et al

• Updates and versioning

• User-defined types, aggregates, operators

• Unlimited ‘null’ or ‘missing’ codes
Available now: P4 Math Add-ons

- Scalable math
  - Matrix multiply, covariance
  - Cumulative sum & product, quantiles, rank
  - Distance metrics (euclidean, manhattan)
  - Correlation metrics (Pearson, Kendall-tau, Spearman)
Coming in December: R11.12

- Failover and recovery
  - Transactional updates (ACID)
  - K-replication, failover, automatic node restart
- Unified storage model for sparse & dense data
- Executor vectorization
- Window aggregates – OLAP windowing in more than one dim
- Parallel loading
December P4 Math Add-ons

• Linear & Logistic regression
• Inverse
• SVD
• Statistical tests
  - e.g. Students T
• Distributions functions
  - Gaussian, Poisson, Geometric, ....
• R connector
June ‘12 release

• Preliminary provenance
• Improved cluster management
  - monitoring, status reporting system admin
• Stability and performance
HDFS & FITS loaders

• HDF5: Daniel Wang, SLAC
danielw@slac.stanford.edu

• FITS: Miguel Branco, EPFL
École Polytechnique Fédérale de Lausanne
miguel.branco@epfl.ch
Motivation for loader

People want to try SciDB on their big data, but...

**Existing method = barrier**
- built-in `load()` uses an textual format—but data is binary
- original → text → scidb native = inefficient
- manual schema conversion = tedious

**Loaders are better**
- Faster: binary to binary import
- Easier: Automatic schema conversion

* Not the whole story: “in-situ” backends would skip loading.
HDF5 loader

Features

- Fast: I/O bound
- Loads n-dimensional arrays of primitives and compounds of primitives
- Collapses some sub-arrays: 1-D of 2-D into 3-D

Usage

- load_library('loadhdf');
- loadhdf('image','/data/file.h5','/run0/ccd/image');
- show(image);

Available now: https://github.com/wangd/SciDB-HDF5
FITS Loader: Current Status

- SciDB operator for FITS binary tables
- Includes two user-defined operators
  - fits_input(), similar to the built-in input() operator
  - fits_show(), used to show the schema of a FITS file (i.e. size and data type of the array)
FITS Loader: Sample Usage

AFL% fits_show('rosat_pspc_rdf2_3_bk2.fits');
    [(true,"float","512,512")]

AQL% create array fits_test<v: float NOT NULL>
    [d0=0:511,512,0, d1=0:511,512,0];

AFL% fits_input(fits_test,
    'rosat_pspc_rdf2_3_bk2.fits');
FITS Loader: Features

- Written from scratch, no dependencies on external libs such as CFITSIO
- Supports exclusively FITS binary tables
- Most common data types are already supported
  - Remaining data types easy to add!
- Happy to support and extended it as needed!
FITS Loader: Next Steps

• Merge into trunk, as an example of a UDO loader:
  - Probably a good idea to agree on conventions regarding UDO names, arguments, etc:
    ▪ e.g. `<format>_show()`, `<format>_input()`

• **EPFL/DIAS** group has plans to develop an in-situ FITS storage for SciDB
  - (As a proof-of-concept for an in-situ DB applied to an array data model)
    ▪ We’re waiting the outcome of an ESA proposal to get started!
  - Comments welcome: workloads, benchmarks, FITS vs HDF-5, ...
EPICS Data Archiver

• Nikolay Maltisky, Brookhaven National Labs
  - malitsky@bnl.gov

• A global monitoring and analytic infrastructure for a wide variety of "big science" projects
  - http://www.aps.anl.gov/epics/

• Loading million data points a second
Community Input

• Questions?
• Feedback?
• Requests?