Matthew J. Graham (Caltech)

Alternative databases

Jet Propulsion Laboratory
INSTITUTE FOR SPACE STUDIES
Sponsored by the Jet Propulsion Laboratory and by the Keck Institute for Space Studies, California Institute of Technology
Caltech

beyond rdbms

- "Beyond 300TB is difficult" - Alex Szalay
- RDMBSs are tuned for small but frequent read/write transactions or large batch transactions with rare write accesses
- Scalable in terms of dataset size and read/write concurrancy
- Problems come with:
  - Too many reads, writes, joins
  - Server swamped with server-side computations

types of data store

- QServ: MySQL + Xrootd, shared-nothing (LSST solution)
- SciDB: Column-oriented db; arrays rather than tables; maintains ACID
- NoSQL: Largely optimized key-value stores; not ACID; web scale solutions
- NewSQL (H-Store, Google Spanner): Relational model + SQL; sharding middle layer

xml

- W3C standard markup language for structured data
- Hierarchical data model
- Supporting technologies:
  - -- xpath: standard for pointing to element, attributes and values
  - -- xslt: standard for converting XML to other formats
  - -- xquery: standard for querying XML documents
- Native XML dbs or RDBMSs with XML support

xml example

```
<resource xsi:type="vs:DataCollection" created="2000-01-01T09:00:00" status="active">
    <title> The Catalogue of Palomar-Quest Transient Sources </title>
    <shortName> pqtrans </shortName>
    <identifier> ivo://nvo.caltech/pqtrans </identifier>
    <curation>
        <creator> Matthew Graham </creator>
        <version> 1.0 </version>
        <contact> mjg@caltech.edu </contact>
    </curation>
    <content>
        <subject> variable stars </subject>
        <description> This contains the transient sources discovered in the Palomar-Quest survey </description>
        <referenceURL> http://nvo.caltech.edu/catalogs/pqtrans </referenceURL>
    </content>
    <format> text/xml+votable </format>
    <format> text/plain+csv </format>
    <coverage> optical </coverage>
    <catalog>
        <table><column><name> ID </name>
            <description> Source identifier </description>
            <unit/>
        </column></table>
    </catalog>
</resource>
```

flwor example

```
declare namespace cs = "http://www.ivoa.net/xml/ConeSearch/v1.0";
let $res := /resource[capability/@xsi:type = 'cs:ConeSearch']
for $cs in $res where contains($cs/description, 'quasar')
order by $cs/identifier
return
<conesearch>
      <title> {string($cs/title)} </title>
      <url> {string($cs/capability/interface/accessURL)} </url>
</conesearch>
```

rdf

- W3C standard for data interchange
- Associative data model: subject - predicate - object
- Supporting technologies:
  - sparql: standard for querying RDF data
  - rdfs: standard for modeling RDF data
  - skos: standard for representing controlled vocabularies
  - owl: standard for representing concept schemes
- Triple stores or RDBMS with SPARQL interfaces

sparql example

```
PREFIX abc: <http://example.com/exampleOntology#>
SELECT ?capital ?country
WHERE {
  ?x abc:cityname ?capital ;
     abc:isCapitalOf ?y .
  ?y abc:countryname ?country ;
     abc:isInContinent abc:Europe .
}
```

```
Sample data:
Berlin abc:isCapitalOf Germany
Germany abs:isInContinent abc:Europe
Chile abs:isInContinent abc:SouthAmerica
Shanghai abc:isCityIn China
```

ontology-driven databases

- Data modeled at both syntactic and conceptual level
- Employs formally-expressed domain knowledge:
  - Consistency checks
  - Logical inferencing
- Facilitates smart applications