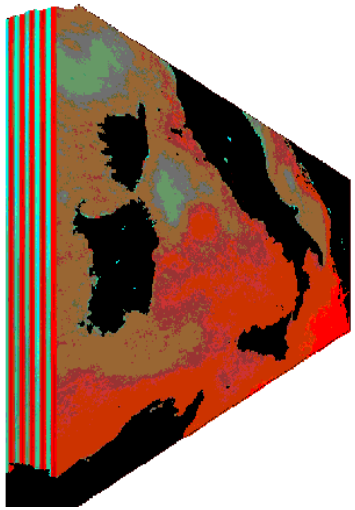
Supported by EU FP7 SCIDIP-ES, EU FP7 EarthServer



A Big Picture for Big Data

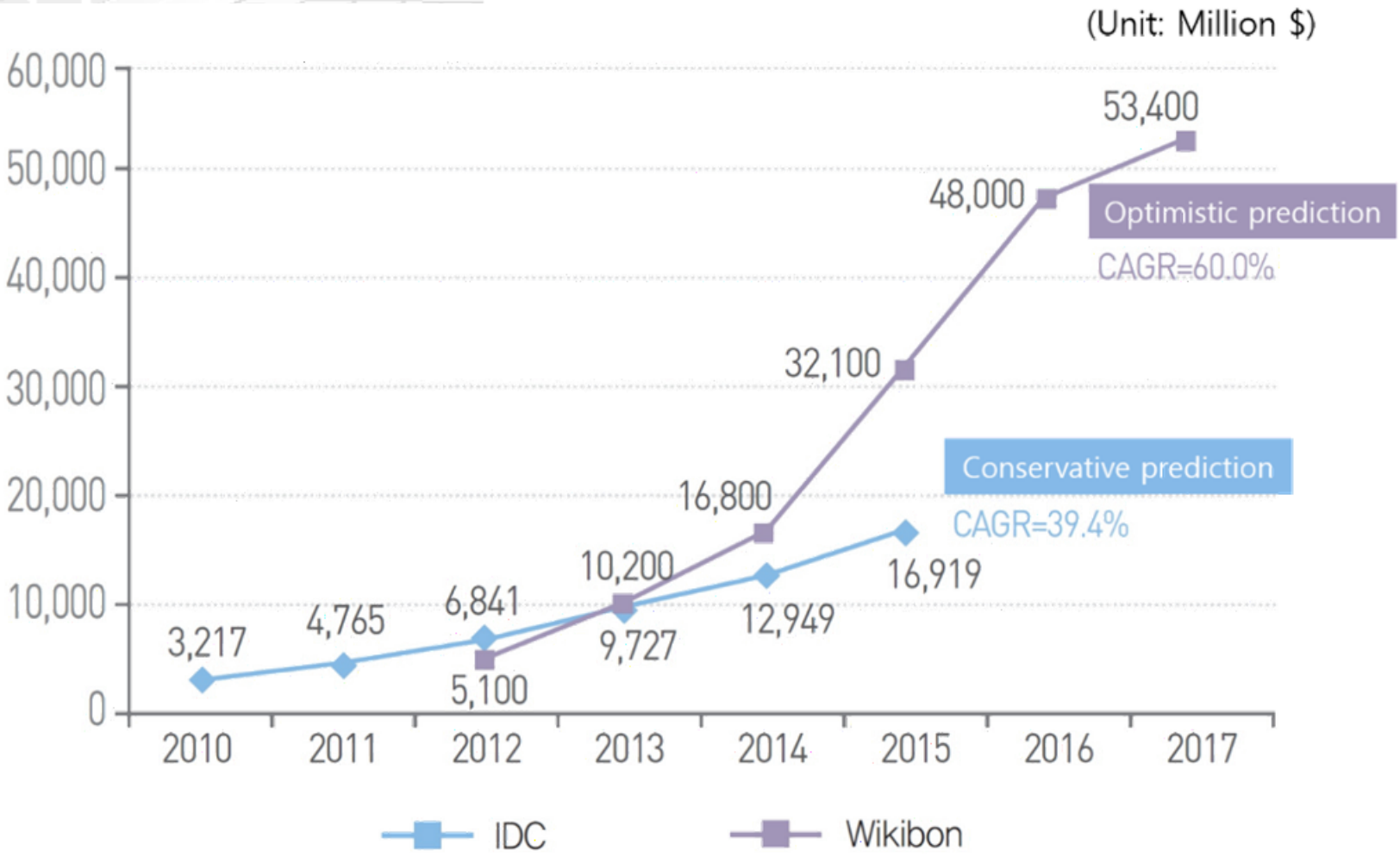
FOSS4G-Europe, Bremen, 2014-07-15

Peter Baumann
Jacobs University | rasdaman GmbH
p.baumann@jacobs-university.de

Our Stds Involvement

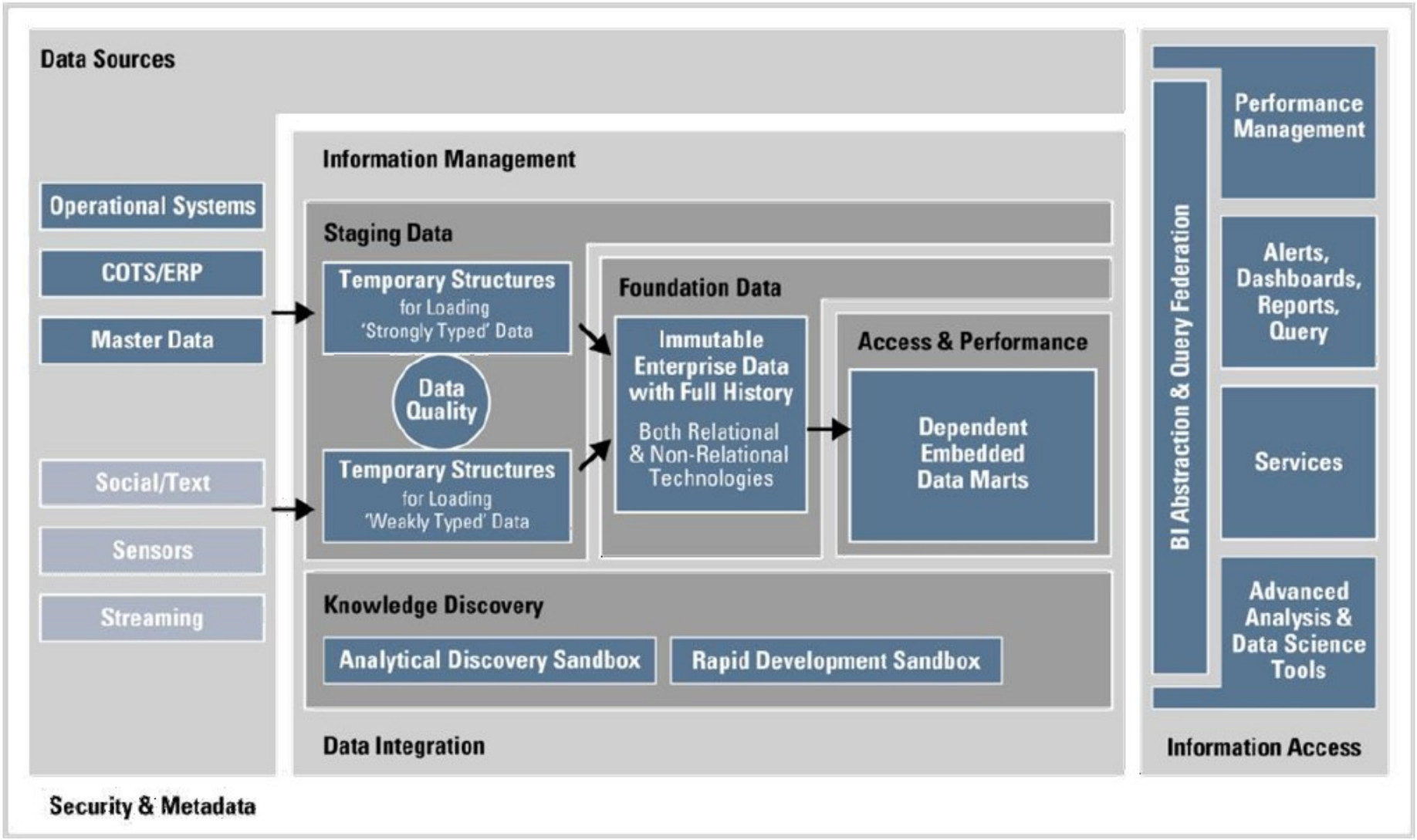
- ISO: member, SC32 / WG3 SQL; SC32 Big Data Study Group; OGC liaison, TC211
- Open Geospatial Consortium: co-chair, BigData.DWG, WCS.SWG, Coverages.DWG; co-founder, Temporal.DWG
- Charter Member, OSGeo
- Research Data Alliance: co-chair, Big Data Interest Group and Geospatial Interest Group
- member, ERCIM Expert Group Big Data
- member, Belmont Forum, WP 3 Harmonization of global environmental data infrastructure
- council member, CGI / IUGS
- founding member and secretary, CODATA Germany
- ...

Big Data Market Forecast



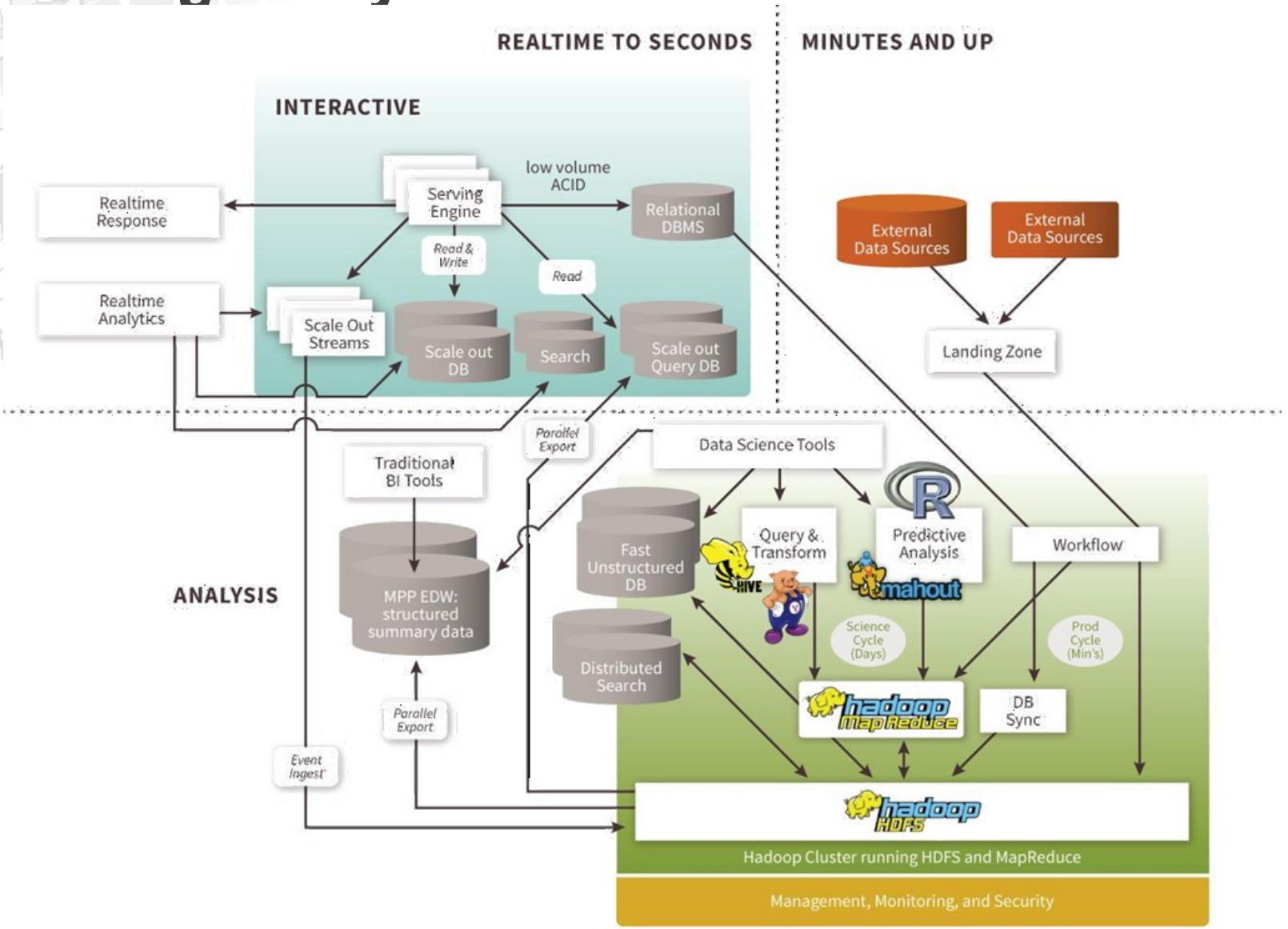
[ISO/IEC JTC1 SC32 N0065]

Oracle Reference Architecture



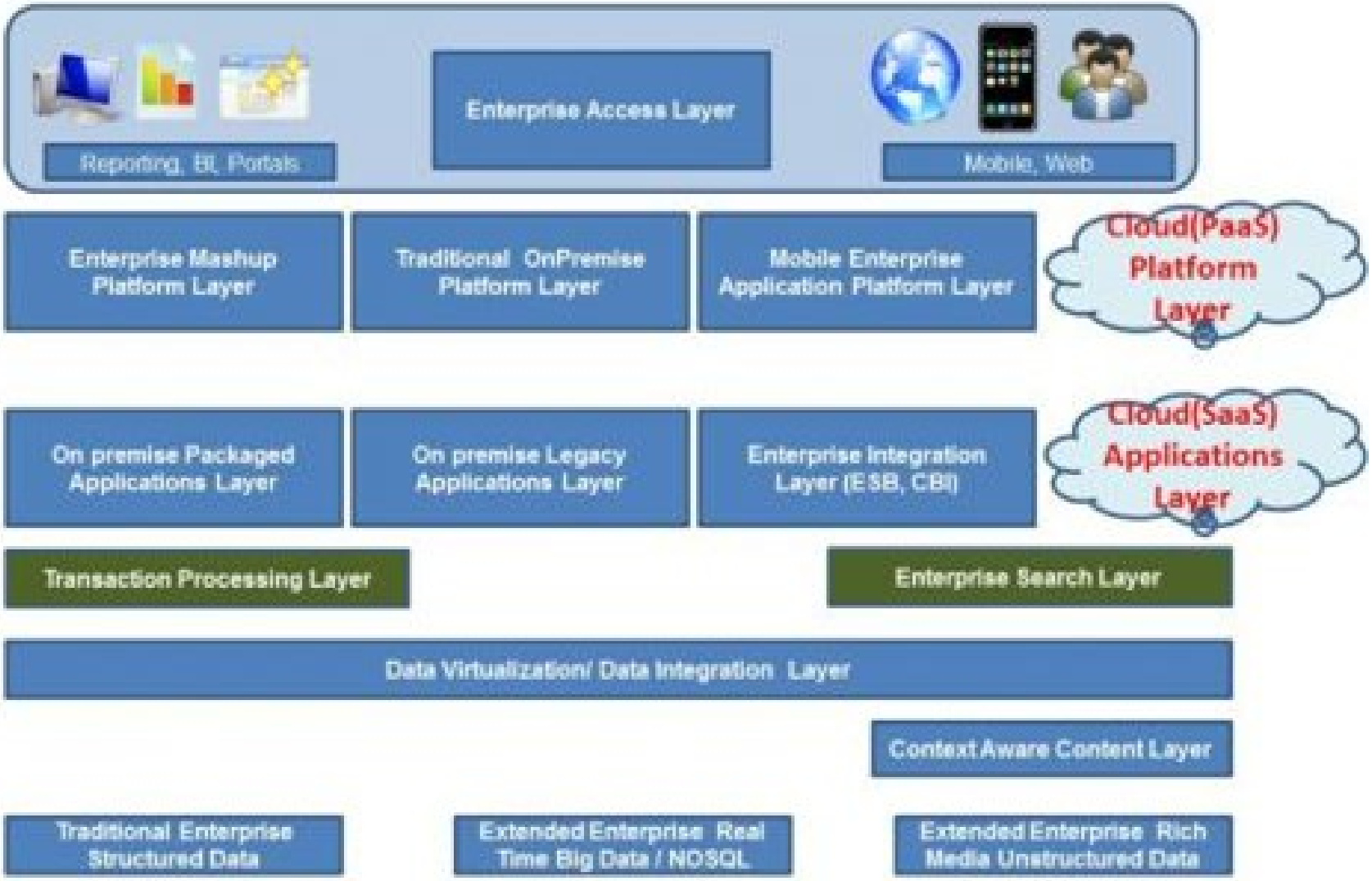
[oracle.com]

Think Big Analytics RA



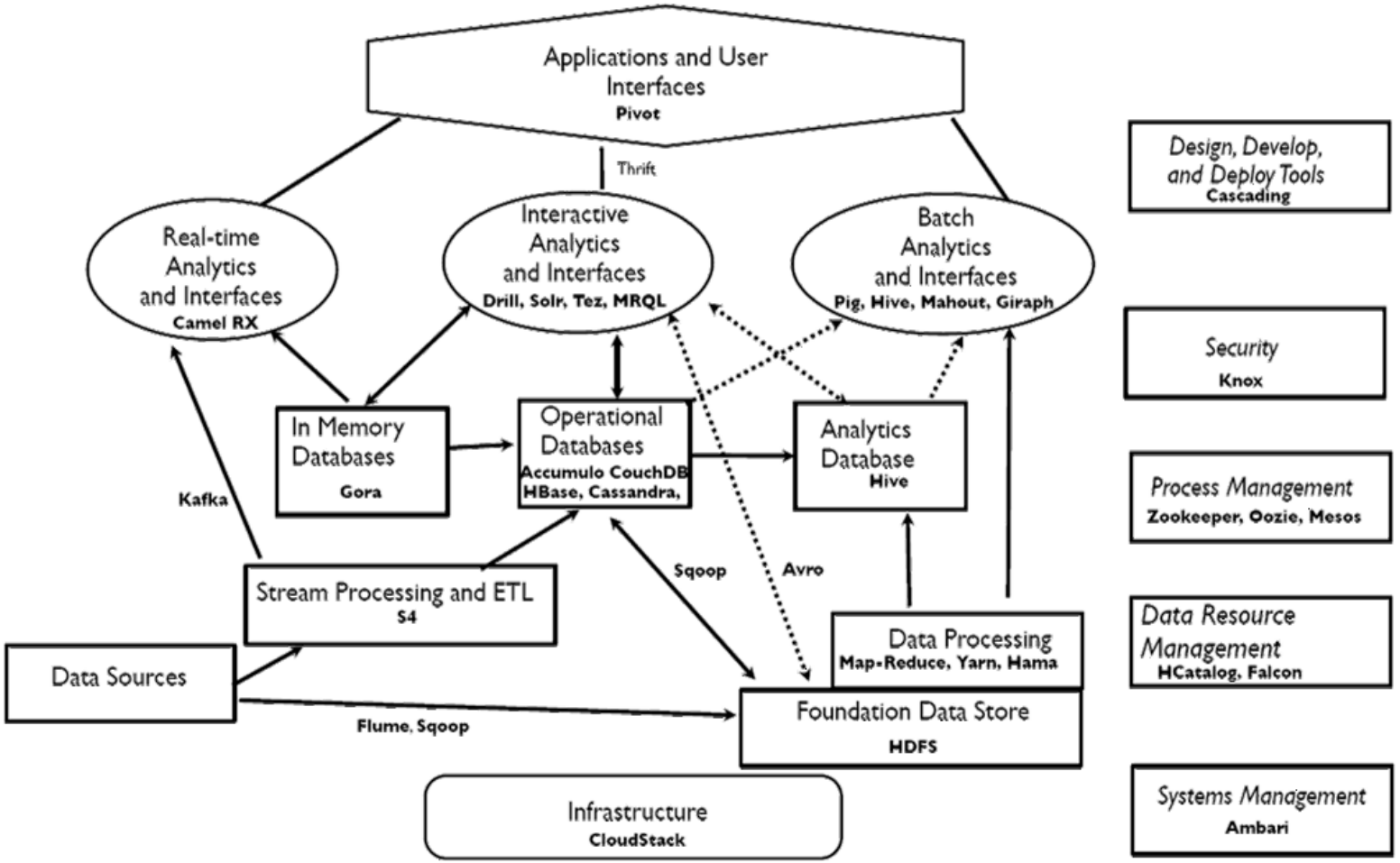
[thinkbiganalytics.com]

Ulitzer New Enterprise Reference Architecture



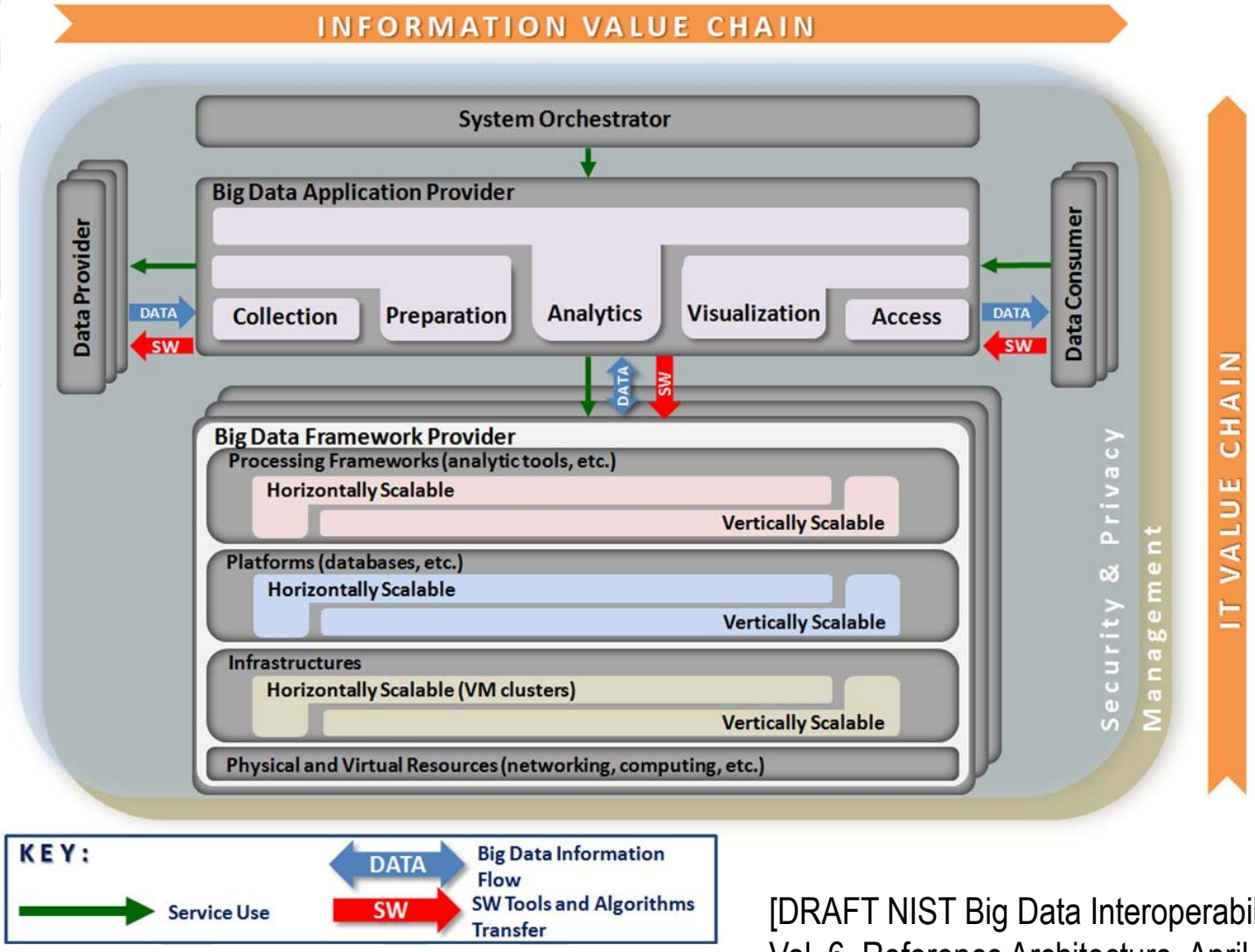
[ulitzer.com]

Apache Big Data Framework



[http://semanticommunity.info]

NIST Big Data Reference Architecture*



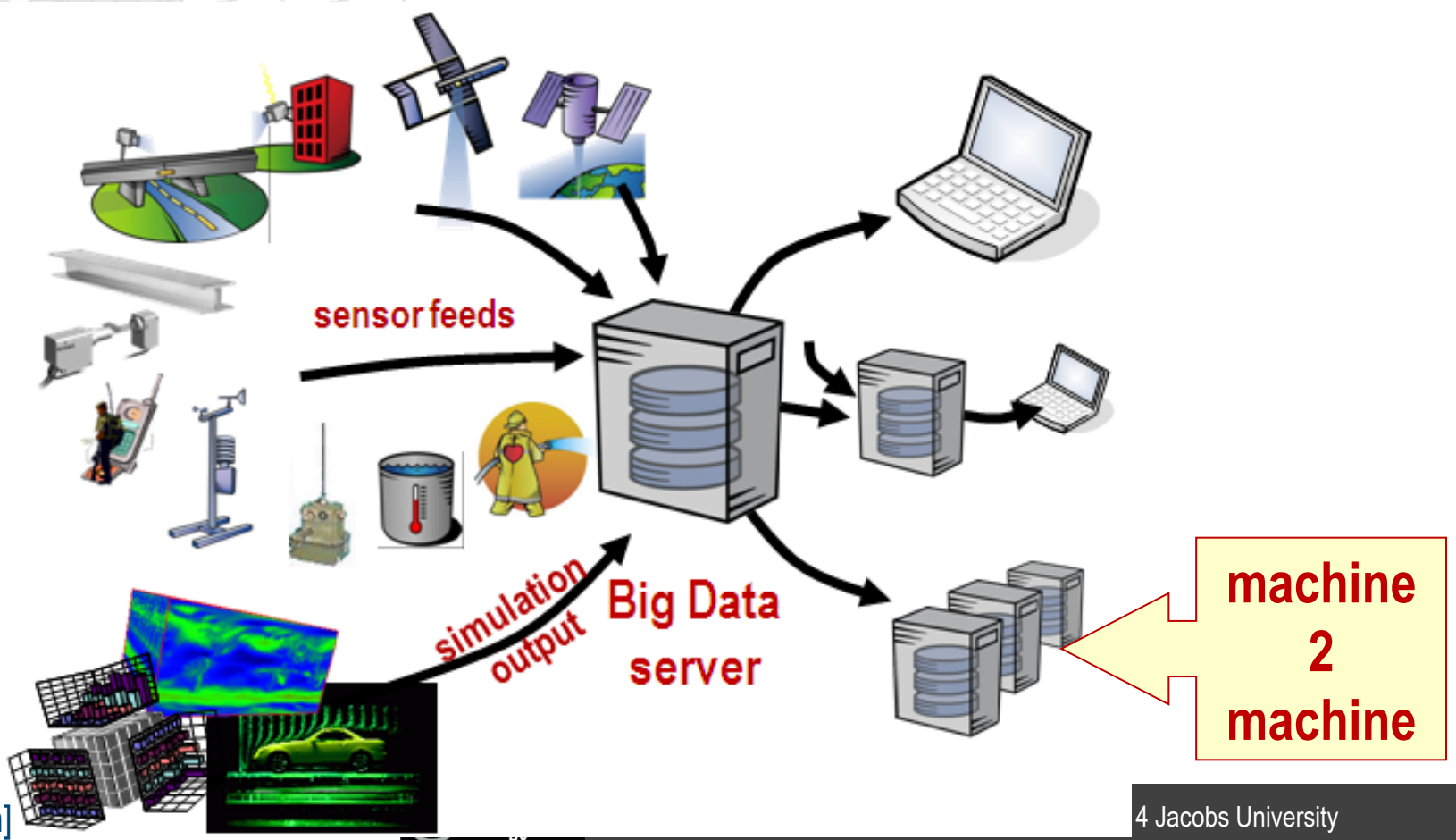
[DRAFT NIST Big Data Interoperability Framework: Vol. 6, Reference Architecture, April 23, 2014]

Reference Architectures?

- Static
- Mixing concept, architecture, interfaces, tools
- often 3-tier variants
  - what's new?

My Reference Architecture ;-)

- ...for Sensor, Image, Simulation, Statistics Data
  = spatio-temporal data cubes



[src: OGC & own]

A Data Structure Centric View

- Stock trading: 1-D sequences (i.e., arrays)
- Social networks: large, homogeneous graphs
- Ontologies: small, heterogeneous graphs
- Climate modelling: 4D/5D arrays
- Satellite imagery: 2D/3D arrays (+irregularity)
- Genome: long string arrays
- Particle physics: sets of events
- Bio taxonomies: hierarchies (such as XML)
- Documents: key/value stores: sets of unique identifiers + whatever
- etc.

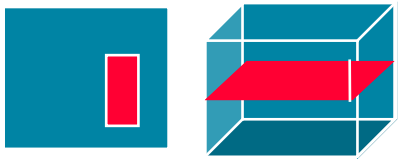
analyzing **domain standards** (like OGC WCS) can help

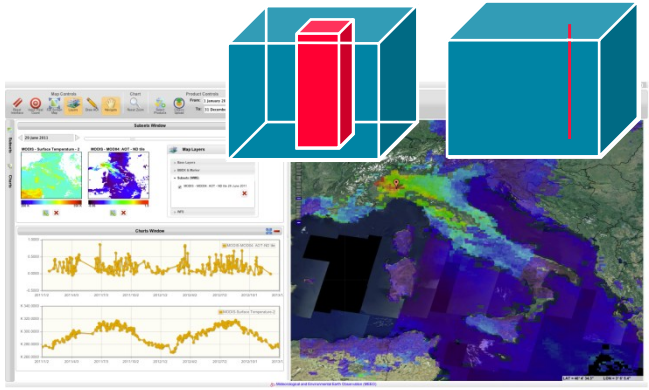
Embracing Variety

- Observation: (Big) Data means many things, at least:
  - Heterogeneous data types
  - Heterogeneous ways to access (download, extract, analyze, mix, ...)

- Proposition: Query languages as common c/s interfaces
  - Sets: SQL
  - Arrays: SQL/MDA
  - Graphs: SparQL, graph QLs
  - Hierarchies: XQuery/XPath
  - + *glue*

Sample Domain-Specific Queries





- Arrays: data cube paradigm

```
SELECT la.scene.red – la.scene.nir
FROM LandsatArchive as la
WHERE avg_cells( la.green > 17 )
```

- Hierarchies: XPath

```
/book/chapter[ @title = „Introduction“ ]
```

- Graphs:
  - Graph QLS (CYPHER):

```
MATCH (n:Crew)-[r:KNOWS*]-m
WHERE n.name='Neo'
RETURN n AS Neo,r,m
```

  - SPARQL:

```
SELECT ?name ?email
WHERE {
  ?person rdf:type foaf:Person.
  ?person foaf:name ?name.
  ?person foaf:mbox ?email.
}
```

Sample Integrated Queries

- Arrays + sets (SQL)

```
select encode( scene.nir, „image/tiff“ )
from LandsatScenes
where acquired between „1990-06-01“ and „1990-06-30“
```

- Arrays + hierarchies (XQuery)

```
for $c in doc(“WCPS”)//coverage/[ some( $c.nir > $c.red )]
return
   <id> { $c/@id } </id>
   <area> { $c/boundedBy } </area>
```

Major Consortia for Big Data Standardization

- ISO/IEC JTC 1/SC 32: Data management & interchange
  - SC32 Big Data Analytics Study Group
- ITU-T SG13: Cloud computing for Big Data
- W3C: Web & Semantic related standards for markup, structure, query, semantics, & interchange
- Open Geospatial Consortium (OGC): Geospatial related standards for the specification, structure, query, and processing of location related data
  - BigData.DWG, http://external.opengeospatial.org/twiki_public/BigDataDwg/
- Organization for the Advancement of Structured Information Standards (OASIS): Information access and exchange.
  - structured data content; security, Cloud computing, SOA, Web services, the Smart Grid, electronic publishing, emergency management, ...

Major Consortia for Big Data Standardization /2

- Research Data Alliance (RDA)
  - Big Data Analytics Interest Group: use case based
  - Geospatial Interest Group
- Transaction Processing Performance Council
  - Benchmarks for Big Data Systems; Specification of TPC Express, BenchmarkTM for Hadoop
- TM Forum
  - Enable enterprises, service providers and suppliers continuously transform to succeed in the digital economy
  - Share experiences to solve critical business challenges including IT transformation, business process optimization, big data analytics, cloud management, and cyber security.

Potential Gaps in Standardization

- Big Data use cases, definitions, vocabulary, reference architectures
- Specifications & standardization of metadata including provenance
- Application models (batch, streaming, etc.)
- Query languages for diverse data types (XML, RDF, JSON, multimedia, etc.) and Big Data operations (e.g. matrix operations)
- Domain-specific languages
- Semantics of eventual consistency
- efficient network protocols
- ontologies & taxonomies
- Security, privacy access controls
- Analytics, resource discovery, mining
- Data sharing & exchange
- Data storage
- human consumption (visualization!)
- Energy measurement for Big Data

[ISO/IEC JTC1 SC32 N0065]

Conclusion

- Query language concept successful...
  - ...if & when data model adequate
- Query language as general c/s paradigm
  - Client can hide behind visual interface
  - Server can optimize
  - Agnostic of low-level protocol: KVP, SOAP, REST, ...
- More options for future
  - Security, uncertainty, etc.
- Open research: next-gen mediators as „glue“
  - Issues: model integration, cross-model optimization, ...