Triple store evaluation

Vasily Bunakov, STFC

EUDAT MetaData TaskForce meeting

Amsterdam, 13-15 January 2014

Test environment

- 3k records harvested from eudat-jmd.dkrz.de = about 100k RDF triples
- Scaled up to 600k EUDAT-like records = 20M RDF triples
- Uploaded in Jena TDB triple store:
  a part of an open source Java framework
  http://jena.apache.org

Ingest productivity

for 600K EUDAT-like records resulted in 20M RDF triples = 3Gb RDF graph

	Laptop 2 Gb	Desktop 2 Gb	Desktop 4Gb
Upload time for the whole set, sec	2018	2741	729
Upload rate, RDF triples / catalogue records per sec	1056 / 311	7403 / 229	27842 / 862

Laptop 2Gb = Ubuntu (64 bit) 2Gb VM on Intel Core i3 2.2GHz
Desktop 2 Gb = Ubuntu (64 bit) 2Gb VM on Intel Core i5 3.3GHz
Desktop 4 Gb = Ubuntu (64 bit) 4Gb VM on Intel Core i3 3.3GHz

Requests productivity

	Laptop 2 Gb	Desktop 2 Gb	Desktop 4Gb
Count languages ordered by their names, sec	48.5	2.5	2.5
Count languages ordered by their popularity, sec	47.9	2.6	2.4
(unordered) Retrieve first 20 records associated with a specific language, sec	0.1	0.05	0.05
(ordered by title) Retrieve first 20 records associated with a specific language, sec	42	2.6	2.3

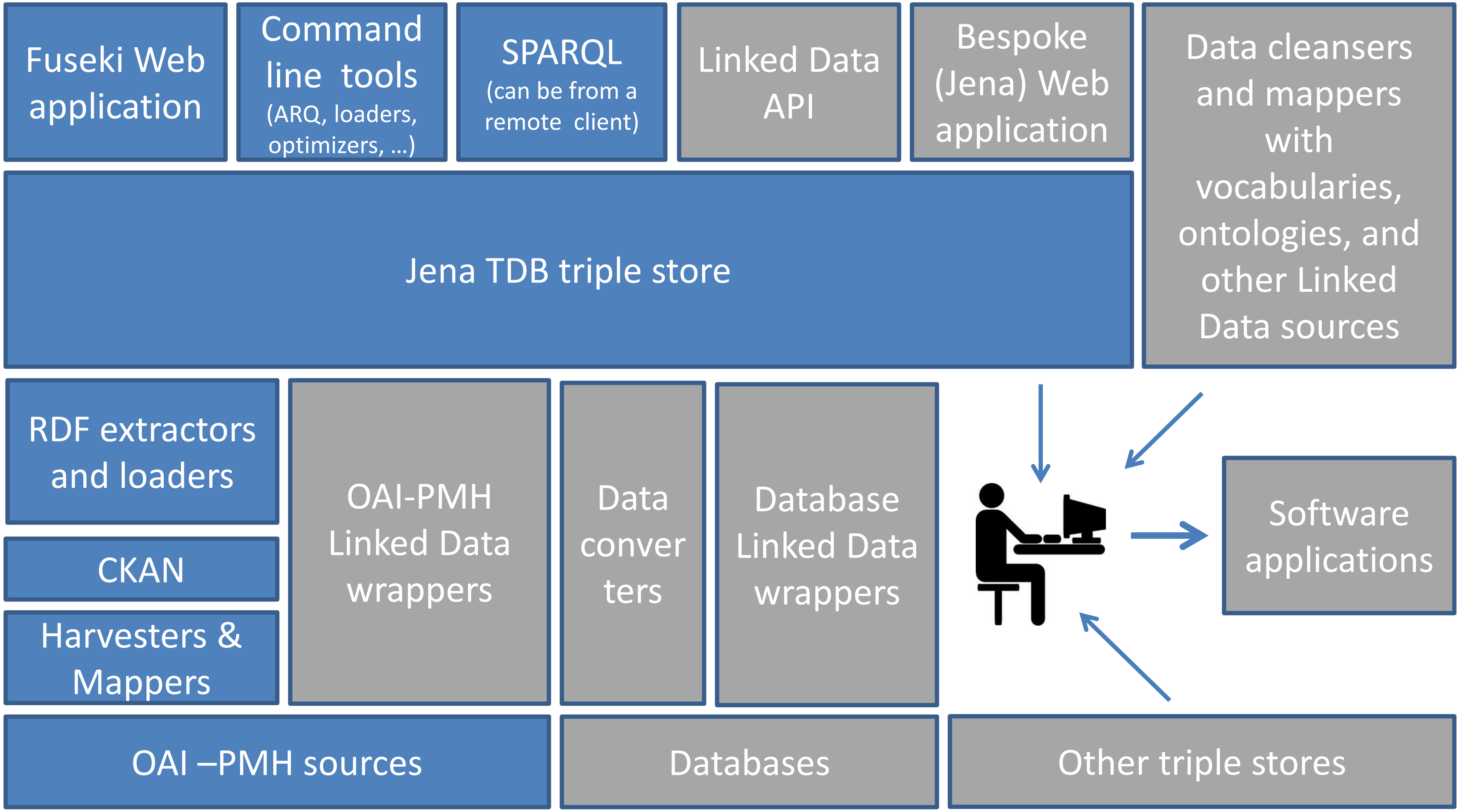
Effect of Jena TDB optimizer

	Laptop 2 Gb no optimizer	Laptop 2 Gb with optimizer	Desktop 2 Gb no optimizer
Count languages ordered by their names, sec	48.5	3.7	2.5
Count languages ordered by their popularity, sec	47.9	3.5	2.6
(unordered) Retrieve first 20 records associated with a specific language, sec	0.1	0.1	0.05
(ordered by title) Retrieve first 20 records associated with a specific language, sec	42	3.4	2.6

RDF advantages

- High data portability
- High interoperability (on data level)
- Potential for integration with various data and reference material
- Scalability on logical level
- Scalability on physical level

Possible technology stack



Fuseki Web application

Command line tools (ARQ, loaders, optimizers, …)

SPARQL (can be from a remote client)

Linked Data API

Bespoke (Jena) Web application

Data cleansers and mappers with vocabularies, ontologies, and other Linked Data sources

Jena TDB triple store

RDF extractors and loaders

CKAN

Harvesters & Mappers

OAI-PMH Linked Data wrappers

Data converters

Database Linked Data wrappers

Software applications

OAI –PMH sources

Databases

Other triple stores

Blue: tried out components Grey: to be considered

TDB comparison to other triple stores

(as per Berlin SPARQL Benchmark)

	100M	200M	1B
BigData	12512.278	10059.940	-
BigOwlim	14029.453	9170.083	1669.899
TDB	15381.857	10573.858	-
Virtuoso6	37678.319	32969.006	8984.789
Virtuoso7	47178.820	-	27933.682

Queries per hour; the larger number means better performance

Testing was done in April 2013 on the cluster of 8 machines as the following:
2 x Intel(R) Xeon(R) CPU E5-2650, 2.00GHz (8 cores & hyperthreading), memory 256GB

Suggestions

- Keep using CKAN as current MD catalogue and as a producer of RDF data
- As an experimental service, offer triple store and a few normalized vocabularies such as locations or languages (along with CKAN)
- Continue scalability experiments
- Develop basic GUI atop of triple store

Thank you!

data
research
LHC
e-Science
particle
petabytes
used
Tier-1
energy
collisions
experiments
STFC
physics
services
researchers
construction
experimental
year
users
beams
rates
computing
November
first
UK
number
years
R89
new
also
sites
experiment
Appleton
Laboratory
expected
storing
Rutherford
programme
project
started
recorded
science
received
initiatives
time
requirements
simulating
analysis
smoothly
plasma
activities
distributed
support
economic
room
move
start
network
Research
LHCb
SCARF
centre
world
Atlas
events
CMS
four
tier-1
universe
processing
store
CERN
machine
impact
March
handling
including
protons
high
wide
Two
TeV
last
main
Tier-2
collision
Central
Web
began
mid-December
discovery
increasing
explained
area
trillion
chemistry
academic
managing
computer
imbalance
predict
notable
bang
Source
PaNdata
days
exploitation
custodial
ensure
near
beam
higher
much
meet
peak
developed
benefit
current
future
collide
large
configuration
continues
computational
materials
simulation
projects
areas
particularly
background
sit
theories
long
review
enough
ALICE
twin
volts
opening
SKOS
vendors
consisting
big
light
Tevatron
advances
sustaining
simulate
make
ago