DESWeb 2014

March 31st, 2014 – Chicago, IL, USA

LODHub - A Platform for Sharing and Integrated Processing of Linked Open Data

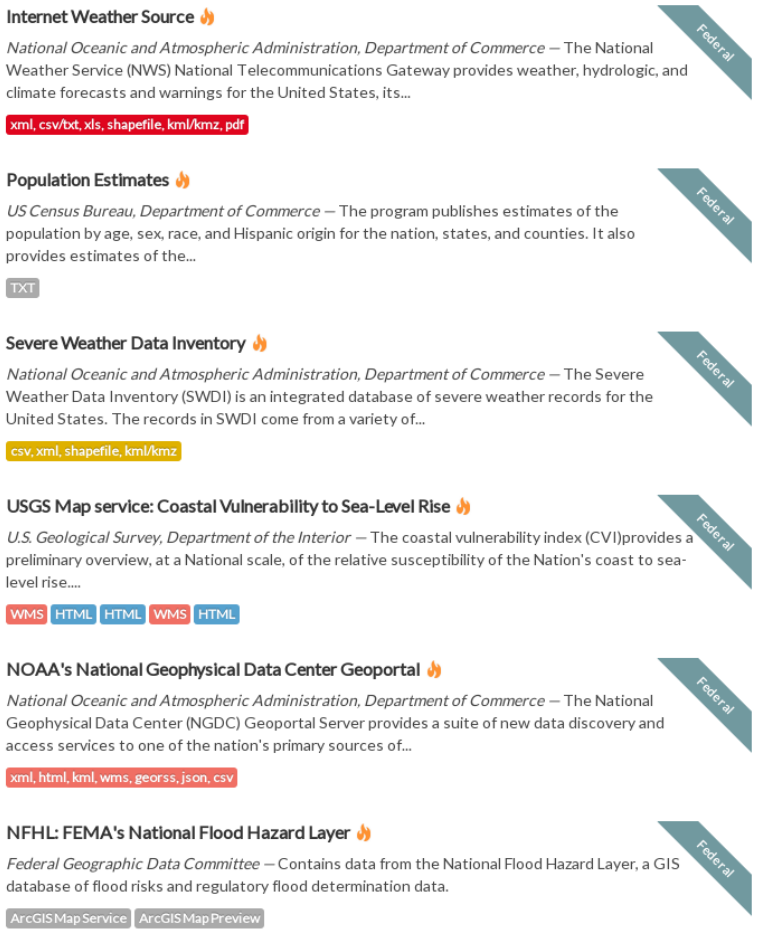
Technische Universität Ilmenau
Databases and Information Systems Group
Stefan Hagedorn and Kai-Uwe Sattler

Outline

Introduction

Current situation

- Open Data platforms
  - semi-structured file format
  - only publish/download
    $\rightarrow$ no direct querying
- SPARQL endpoints
  - not possible to use own datasets
  - reliability and availability too low
- plenty of frameworks and platforms – no infrastructure



Internet Weather Source

National Oceanic and Atmospheric Administration, Department of Commerce — The National Weather Service (NWS) National Telecommunications Gateway provides weather, hydrologic, and climate forecasts and warnings for the United States, its...

xml, csv/txt, xls, shapefile, kml/kmz, pdf

Federal

Population Estimates

US Census Bureau, Department of Commerce — The program publishes estimates of the population by age, sex, race, and Hispanic origin for the nation, states, and counties. It also provides estimates of the...

TXT

Federal

Severe Weather Data Inventory

National Oceanic and Atmospheric Administration, Department of Commerce — The Severe Weather Data Inventory (SWDI) is an integrated database of severe weather records for the United States. The records in SWDI come from a variety of...

csv, xml, shapefile, kml/kmz

Federal

USGS Map service: Coastal Vulnerability to Sea-Level Rise

U.S. Geological Survey, Department of the Interior — The coastal vulnerability index (CVI)provides a preliminary overview, at a National scale, of the relative susceptibility of the Nation's coast to sea-level rise....

WMS HTML HTML WMS HTML

Federal

NOAA's National Geophysical Data Center Geoportal

National Oceanic and Atmospheric Administration, Department of Commerce — The National Geophysical Data Center (NGDC) Geoportal Server provides a suite of new data discovery and access services to one of the nation's primary sources of...

xml, html, kml, wms, georss, json, csv

Federal

NFHL: FEMA's National Flood Hazard Layer

Federal Geographic Data Committee — Contains data from the National Flood Hazard Layer, a GIS database of flood risks and regulatory flood determination data.

ArcGIS Map Service ArcGIS Map Preview

Federal

Screenshot from data.gov

Introduction

LODHub

- provide infrastructure for storage & query processing
- share datasets among users
- link datasets to discover new knowledge
- support data analysis tasks

LODHub

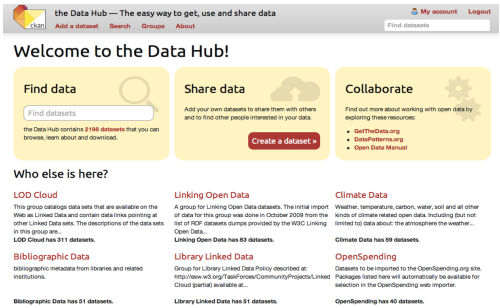
Features

Multi-organization capabilities

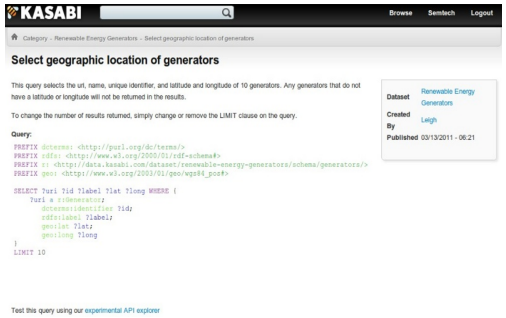
- Aggregate user accounts to teams
- Share data within team

Versioning & provenance

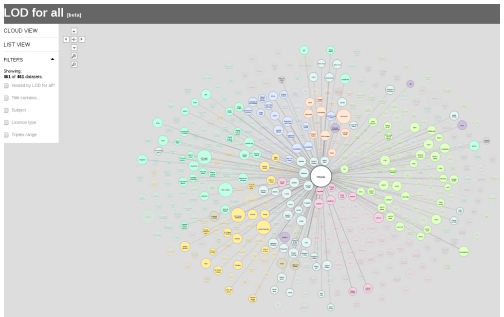
- keep history of changes
- track where data comes from



http://datahub.io/



Kasabi (offline)



http://lod4all.net/

LODHub

Features

Elastic server infrastructure

- keep only optimal number of machines active
- scale with size of data sets and active queries
- distribute data and queries transparent to the users

$\Rightarrow$ Amazon, Google, IBM, ...

Business model

- host datasets for free (size limit)
- pay per query
  - no. of involved hosts
  - execution time

LODHub

Features

Explorative queries

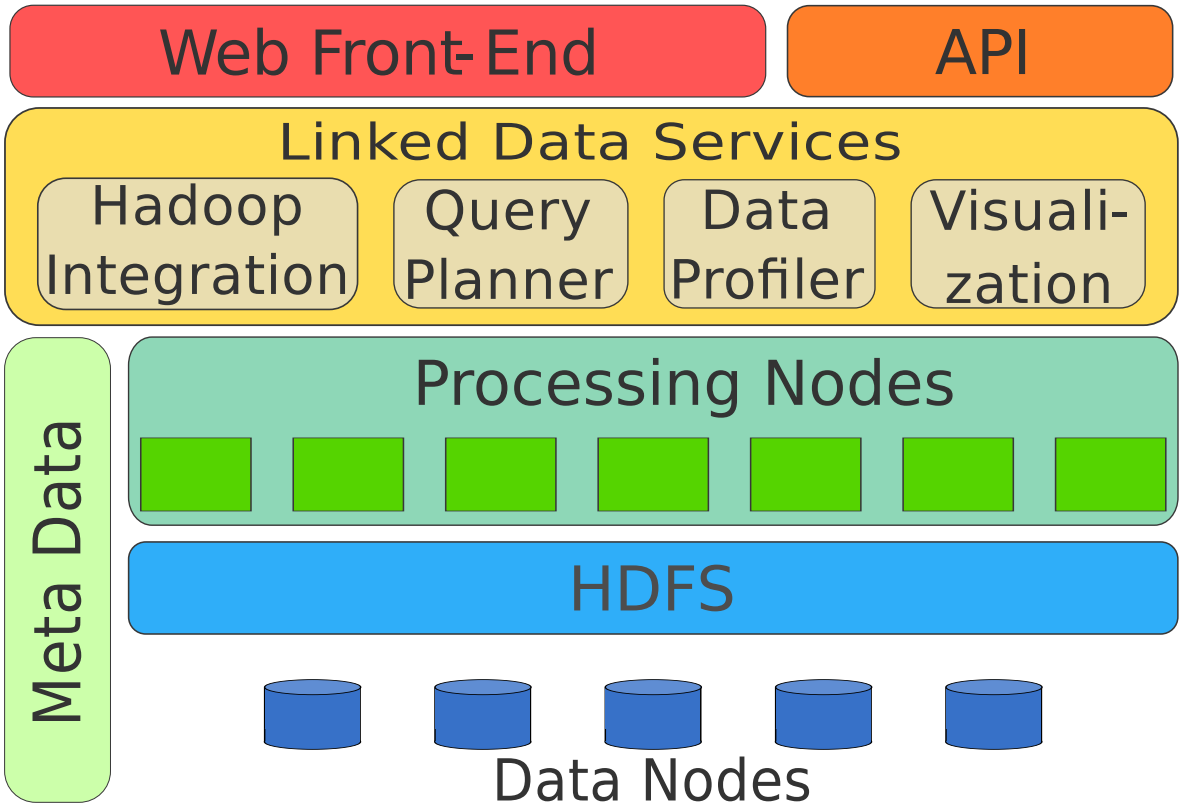
- provide algorithms to explore dataset
- discover links and facts
- enhance original query result

Tools & APIs

- Use standard APIs and formats: SPARQL & RDF
- conversion between formats
- include more tools & languages (Hadoop, R, Python, ...)
- data visualization

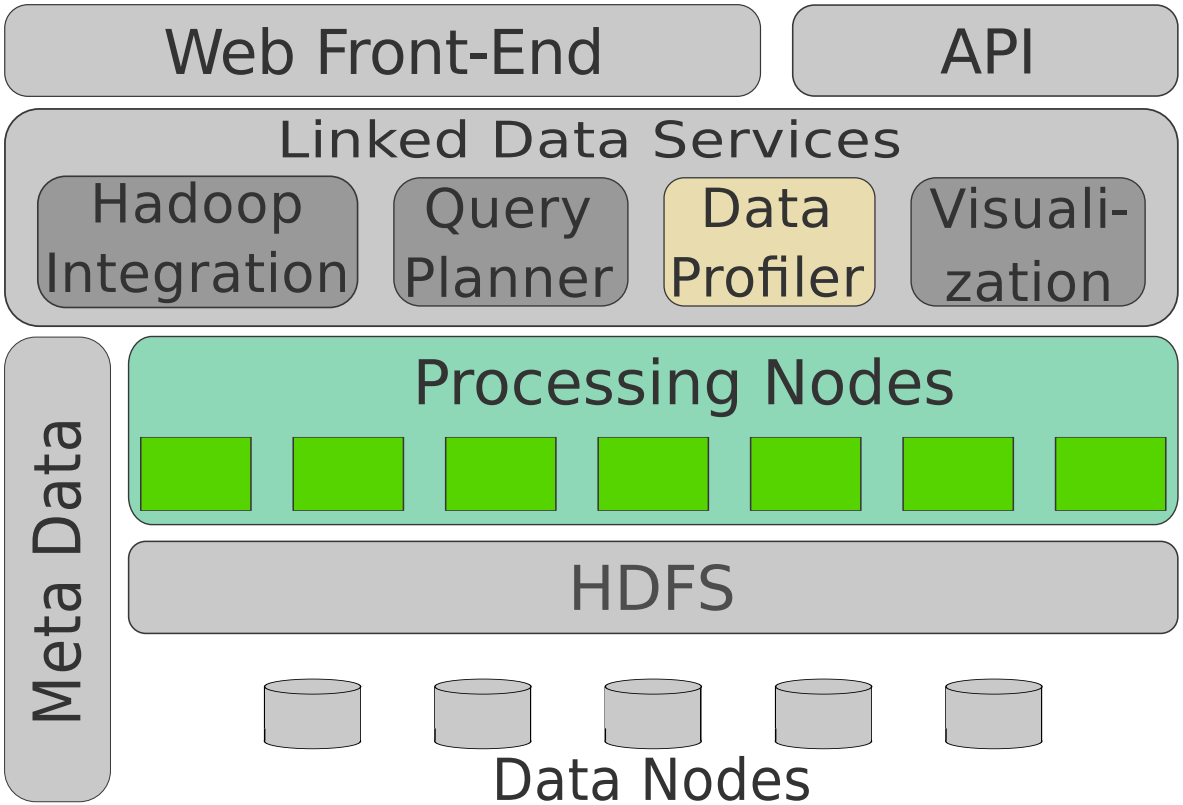
LODHub

Architecture



LODHub

Architecture

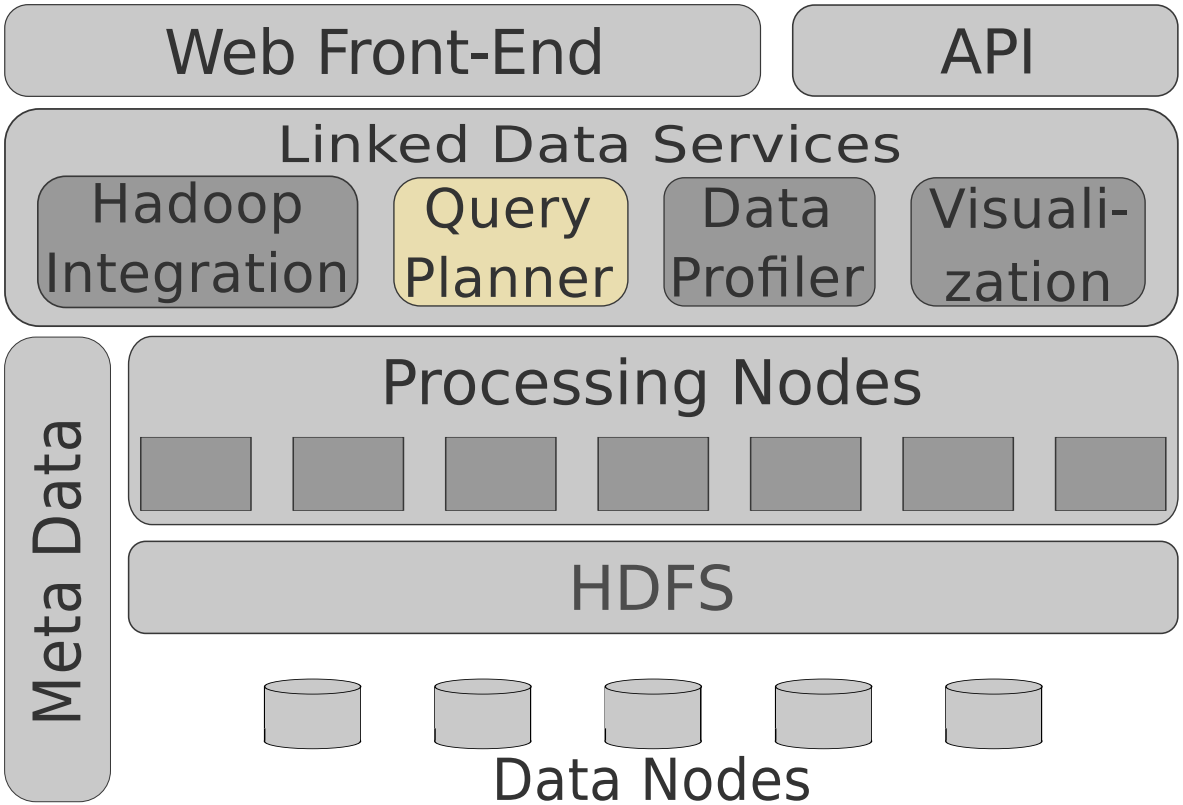


Stefan Hagedorn and Kai-Uwe Sattler,

Efficient Parallel Processing of Analytical Queries on Linked Data, OTM ODBASE 2013

LODHub

Architecture

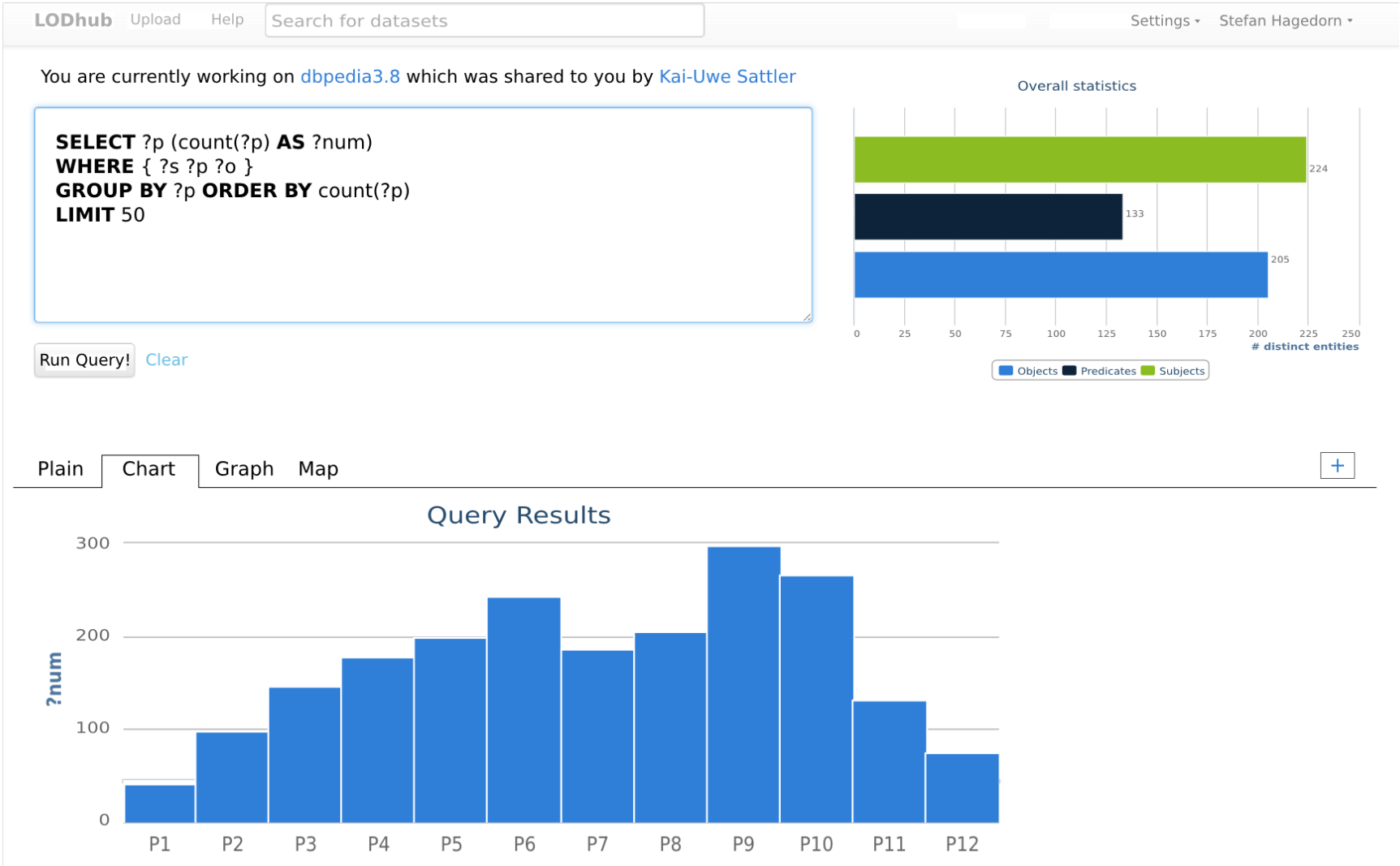


Hose, Umbrich, Sattler, Hagedorn

Multi-processor SPARQL execution, Work in Progress

LODHub

Preview



Challenges

- partitioning of datasets
- distribute queries according to data distribution
  - query multiple datasets/partitions
- predict and react on workload (SLA)
- users with different backgrounds
  - different tasks need different tools

Thank you!