Distributed Knowledge Management

IndexMed GRAIL days, CESAB, Aix-en-provence Marseille, 2016

Víctor Méndez Muñoz vmendez@uab.es
Agenda

- A model of Distributed Knowledge Management (DKM)
- DIRAC Data Management System (DMS)
- A New DKM based in DIRAC DMS?
- Building Knowledge Methodologies and Technologies
- Conclusions
A model of Distributed Knowledge Management (DKM)

User Frontend

DB DB DB
A model of Distributed Knowledge Management (DKM)

The data path
A model of Distributed Knowledge Management (DKM)

User Frontend

Cataloguing System

Metadata Polling System

DB

DB

DB
A model of Distributed Knowledge Management (DKM)

Metadata Polling System

- Gets data from DB
- Filters the significance
- Integrates heterogeneous metadata sources
  - DB tech
  - Data schemas
A model of Distributed Knowledge Management (DKM)

Cataloguing System

• Provides categories of the whole domain
  • Logical identifications of physical data
  • Data tree structure
  • Provenance
  • Metadata
  • ACLs
• Integrates heterogenous data sources
  • DB tech
  • Data schemas
A model of Distributed Knowledge Management (DKM)

The process path
A model of Distributed Knowledge Management (DKM)

User Frontend

Transaction & Processing System

DB

DB

DB
A model of Distributed Knowledge Management (DKM)

Transaction/Processing System

- Data mining transactions over a set of DBs
- Writing ?
- Large processing ?
  - So needing additional computing power ?
A model of Distributed Knowledge Management (DKM)

The overall picture
A model of Distributed Knowledge Management (DKM)

User Frontend

Cataloguing System

Metadata Polling System

Transaction & Processing System

DB

DB

DB
A model of Distributed Knowledge Management (DKM)

User Frontend

Cataloguing System
Metadata Polling System
Transaction & Processing System

DB

Velocity?

Variety

Volume

Data

Knowledge

Information
DIRAC provides all the necessary components to build ad-hoc grid infrastructures **interconnecting** computing resources of different types, allowing **interoperability** and simplifying **interfaces**. This allows to speak about the DIRAC **interware**.
DMS can work without the need of installing all the DIRAC system
DIRAC DMS a file oriented system:

- File replicas are distributed over Storage Elements world wide
- Files are registered in a single name space
  - Metadata
  - ACL, UNIX system like
- For most of applications the file access is as simple as in a File System
  - Additional upload/download if necessary
DIRAC Data Management System (DMS)

**DMS components:**
- Files
- Replicas
- Datasets
- Storage Elements
- Transfer Services
- Catalogs
- File Systems

**GraphDB components:**
- DB objects
- Replicas
- Datasets
- Storage Elements
- Transfer Services
- Catalogs
- DBs
Storage Element is an abstraction
- SRM
- XROOTD
- POSIX FS, others
- Uses DIRAC auth
- Secure transfers
DIRAC Data Management System (DMS)

• DIRAC File Catalogue
• DIRAC File Catalogue
• Datasets are the result of a metadata query

• Datasets can be stored in the Catalogue as another object

• Datasets can be frozen or/and refresh

• Datasets can be operands of the data operations
DIRAC Data Management System (DMS)

Data Centric Example

[Diagram showing the components of the EISCAT system, including WebAppDIRAC, Configuration System, File Catalog, Workload Management System, File Server, and cloud resources.]
A New DKM based in DIRAC DMS?

User Frontend

Cataloguing System

Metadata Polling System

Transaction & Processing System

DB

DB

DB
A New DKM based in DIRAC DMS?

User Frontend

Cataloguing System

Metadata Polling System

Transaction & Processing System

DISET Secured connection

SE DB

SE DB

SE DB

Knowledge

Information

Data
A New DKM based in DIRAC DMS?

WebAppDIRAC Framework

USER Frontend

DIRAC Catalogue

Metadata Polling System

Transaction & Processing System

SE DB

SE DB

SE DB

Informed Metadata

Polling System

Cataloguing System

Transaction & Processing System

Knowledge

Information

Data
A New DKM based in DIRAC DMS?

User Frontend

Metadata Polling System

Transaction & Processing System
• Methods
  • Non-formal methods
    • For example, general graph theory
    • Variable association (e-commerce)
    • Data mining (predictive behaviour)
    • Clustering (marketing segmentation)

• Formal methods (formal validation)
  • For example, membrane computing
Building Knowledge Methodologies and Technologies

- How to build a **knowledge frontend**?
  - Having expert with a knowledge model
  - Moving expertise to computing model
- Following a suitable top-down developing methodology, for example:
  - ISO-9241-210:2010 standard, Human-centred design for interactive systems
Building Knowledge Methodologies and Technologies

- How to build the **information systems**?
- Having expertise in
  - Distributed systems
  - Complex data management
- Using existing framework solutions
- Following a suitable developing methodology, for example:
  - Rapid Application Development (RAD)
Building Knowledge Methodologies and Technologies

- Graph technologies:

  - Graph databases (OLTP): These technologies are used primarily for transactional online graph persistence, typically accessed directly in real time from an application.

  - Graph compute engines (OLAP): in the same category of bulk data analysis (large scale, data mining, MapReduce). These technologies are used primarily for offline graph analytics, typically performed as a series of batch steps.
Building Knowledge Methodologies and Technologies

- NOSQL Data Bases

Voldemort, Amazon's Dynamo

Apache Cassandra

Google BigTable, Apache Hbase

MongoDB

from “Graph Databases” by Ian Robinson, Jim Webber, and Emil Eifrem
Building Knowledge Methodologies and Technologies

from “Graph Databases” by Ian Robinson, Jim Webber, and Emil Eifrem
Building Knowledge Methodologies and Technologies

• Time Series Databases

  • Optimizes for time series analysis
  • Efficient data storage, analysis and retrieval
  • Easy to maintain
  • Scale horizontally
  • Easy to create complex reports (third party dashboard systems)
Building Knowledge Methodologies and Technologies

- Time Series Databases
  - InfluxDB distributed no external dependencies.
  - OpenTSDB is a distributed time series database in Hbase (Apache Hadoop ecosystem).
  - Elasticsearch is a distributed search and analytics.

- Performance evaluation use case

Information System → Publisher States Monitoring → RabbitMQ → Time series databases
ActiveMQ
Kafka
Data visualization dashboard examples are:

- Grafana is a metric dashboard and graph editor for InfluxDB, Graphite and OpenTSDB, which is based in Google plotting engine.
- Kibana is flexible analytic and visualization framework to create dashboards from generic DB.

DataBase → RabbitMQ → Data visualization dashboard
   ActiveMQ
   Kafka
<< Any organization that designs a system (defined broadly) will produce a design whose structure is a copy of the organization's communication structure. -- Melvyn Conway, 1967 >>
• Today complex problems require the integration of varied volumes of data, supported by different methodologies and technologies
• Graph technologies open the possibility of an storage model close to the conceptual representation, overcoming in some aspects to the conventional approaches
• Distributed data sources can be aggregated by a coherent information system
• DIRAC can help in the data cataloguing
Thank you!

victor.mendez@uab.es