# Indexing and Selection of Data Items Using Tag Collections

**Sebastien Ponce**
CERN – LHCb Experiment
EPFL – Computer Science Dpt

**Pere Mato Vila**
CERN – LHCb Experiment

**Roger D. Hersch**
EPFL – Computer Science Dpt

March 27, 2002

Sebastien Ponce

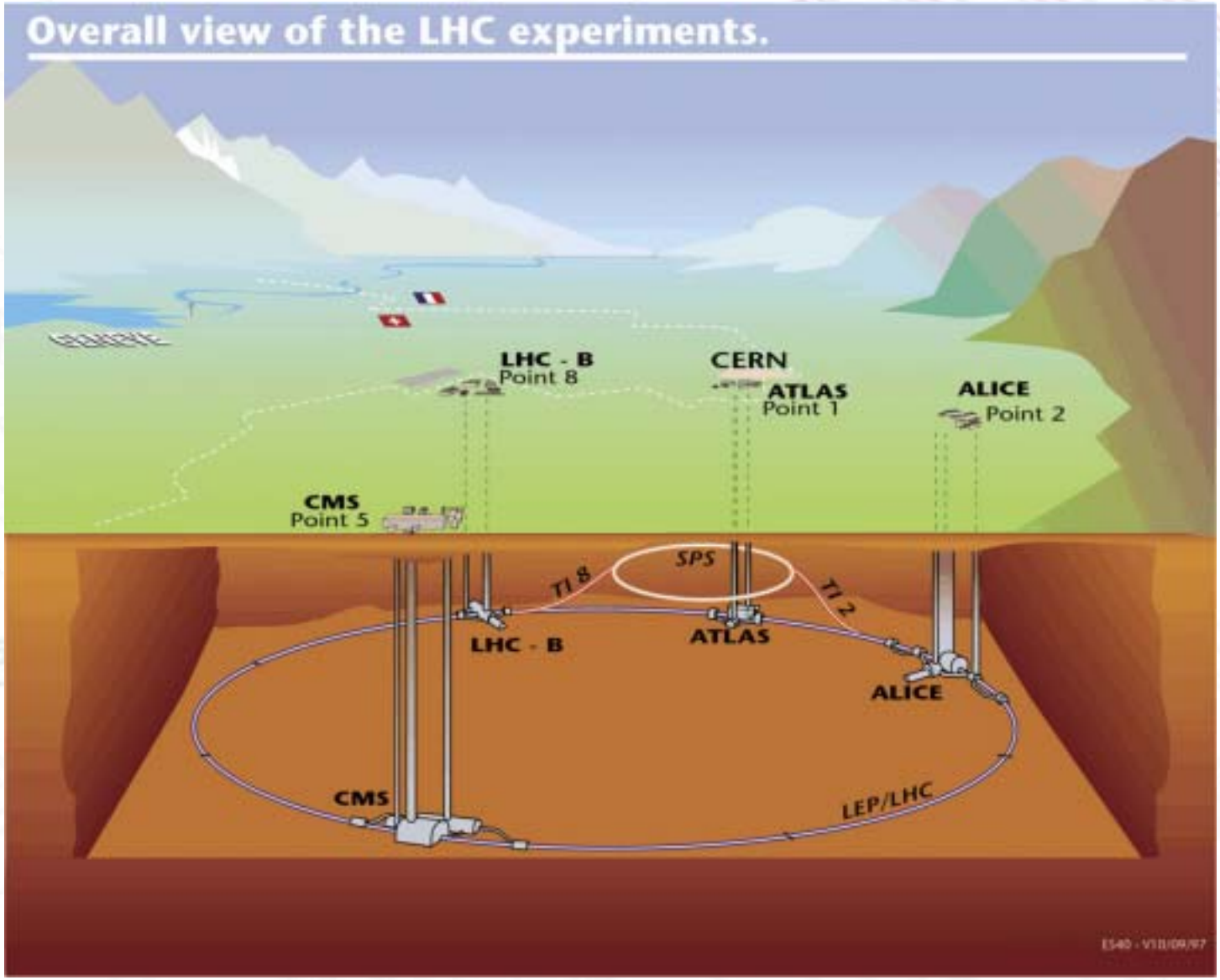
# Overview

- The context : LHCb problems
- A new indexing Schema
- Selection Process
- Theoretical Performance
- First measurements

# Context

- Work developed as part of the LHCb experiment at CERN (European Organization for Nuclear Research)
- Final aim is high energy particle physics, to study the behavior of the B-Meson and the CP-violation.
- Tool : the LHCb detector, being built on the future CERN accelerator : the LHC (Large Hadron Collider)
- The principle is to look at billions of particle collisions every second and understand what's happening

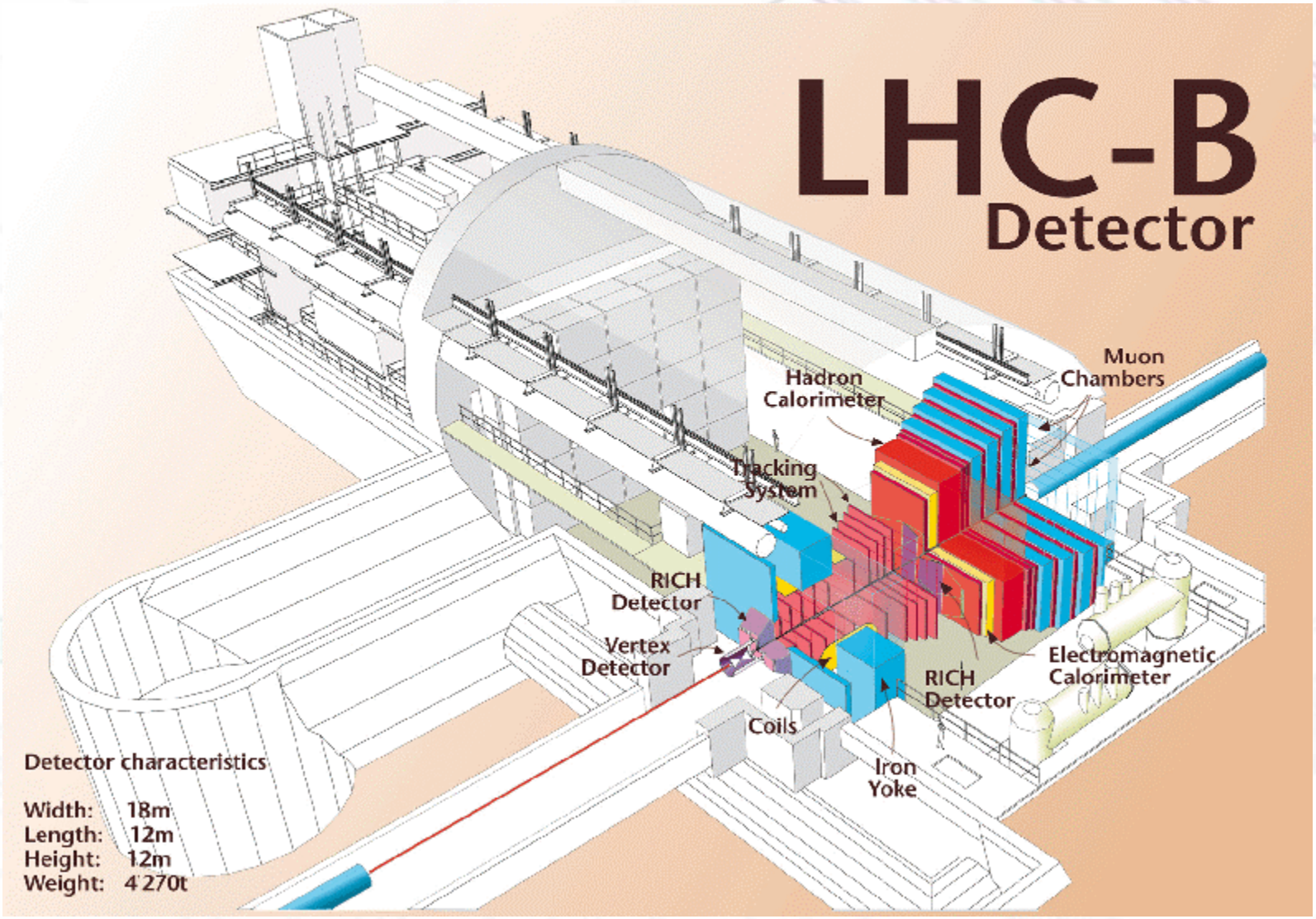
# The LHC




**Near Geneva**
**On the French-Swiss border**

**length : 27 km**
**depth : 50-150 m**

# LHCb




**Width : 18m**
**Length : 12m**
**Height : 12m**
**Weight : 4.3t**

# Some Figures

- Particle collision every 25 ns (40 millions per second).
- 950 000 channels → 1 MB of data for each collision
- Net result of **40 TB/s** of output data
- 24h a day, 6 months per year (15 millions seconds each year)
- + simulations & reconstructed data → * 3

**BUT**

- Interesting physics phenomena are really seldom
- A very efficient three levels trigger system removes 99.999% of the collisions (keeps 200 events per second)
- Only 100 KB are kept for each event
- "Only" **20 MB/s** or **~ .3 PB/year** are stored for real data
- Still **~ 1PB/year** in total

# Data Content

- The basic item is an event
- Events are independent one from the other
- A “per event” indexing is needed in order to make a selection among the $10^{10}$ events (real + simulated + reconstructed)
- The content of an event is a mix of booleans, strings, numbers
- Size and content of an event may vary

# Data Selection Needs

- **Typical physics analysis :**
  - selection of interesting events
  - download these events
  - compute some histogram
  - modify the criteria and restart
- **Selection is highly important**
- **Selection characteristics :**
  - many variables (up to 30, typically 10-15)
  - mixture of types (boolean, numbers, strings)
  - complicated rules, that may need a structured language

# Previous Solution

- Sequential scan of the whole database.
- Every item was converted to a C++ structure and the selection was carried out in the code
- Weber et al(1) demonstrated that this approach was the best one in high dimension

- The goal is to optimize this sequential scan

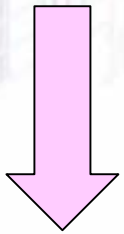
(1) R. Weber, H.-J. Schek, and S. Blott.
A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces.
VLDB'98

# Tags

- **A tag contains :**
  - a subset of the data item it represents
  - a "pointer" to this item
- **The subset of the item contains few values that will be available for fast selection criteria**
- **A tag is a small, well-structured entity that can be easily stored in a relational database**

**Event**

blablabla **Energy** blablabla blablabla
blablabla blablabla blablabla blablabla
blablabla blablabla **NbOfTracks** blabla
**InteractionType** blablabla blablabla
blablabla blablabla blablabla blablabla
blablabla blabla **MuonChamberDeposit**



**Tag**

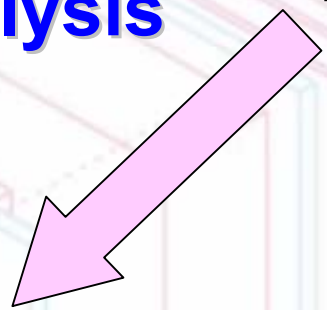
| | |
|---|---|
| float | Energy |
| int | NbOfTracks |
| int | InteractionType |
| float | MuonChamberDeposit |
| string | Pointer to Event |

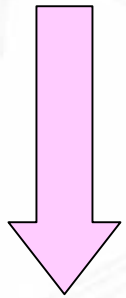
# Tag Types

- Several types of tags can be defined for a single event
- Their content depends on the type of analysis

**Event**

blablabla **Energy** blablabla **CaloEfficiency**
**CaloDeposit** blablabla blablabla blablabla
blablabla blablabla **NbOfTracks** blablabla
**InteractionType** blablabla blablabla blabla
blablabla blablabla **CaloNoiceLevel** blabla
blablabla blablabla **MuonChamberDeposit**





**Tag1**

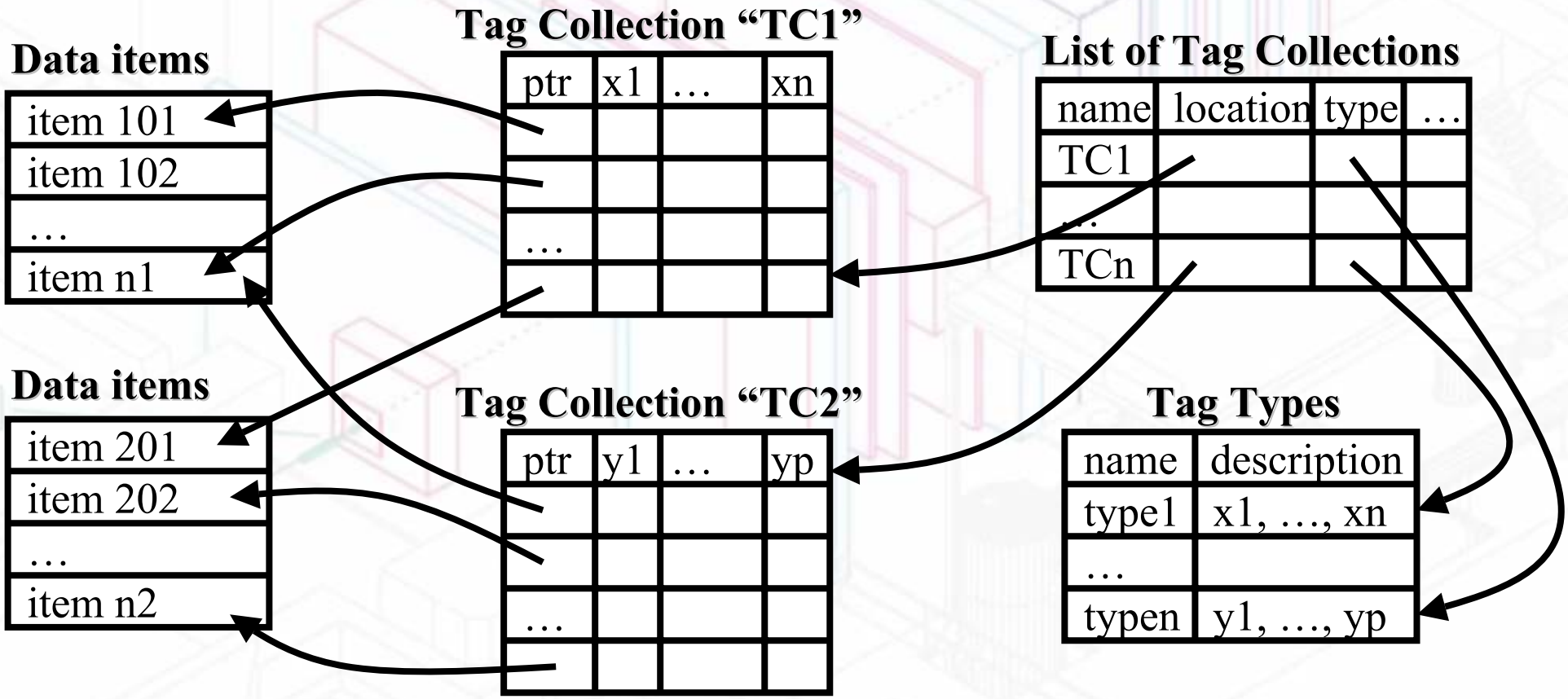
| | |
|---|---|
| float | Energy |
| float | CaloEfficiency |
| float | CaloDeposit |
| float | CaloNoiceLevel |
| string | Pointer to Event |

**Tag2**

| | |
|---|---|
| float | Energy |
| int | NbOfTracks |
| int | InteractionType |
| float | MuonChamberDeposit |
| string | Pointer to Event |

# Tag Collections

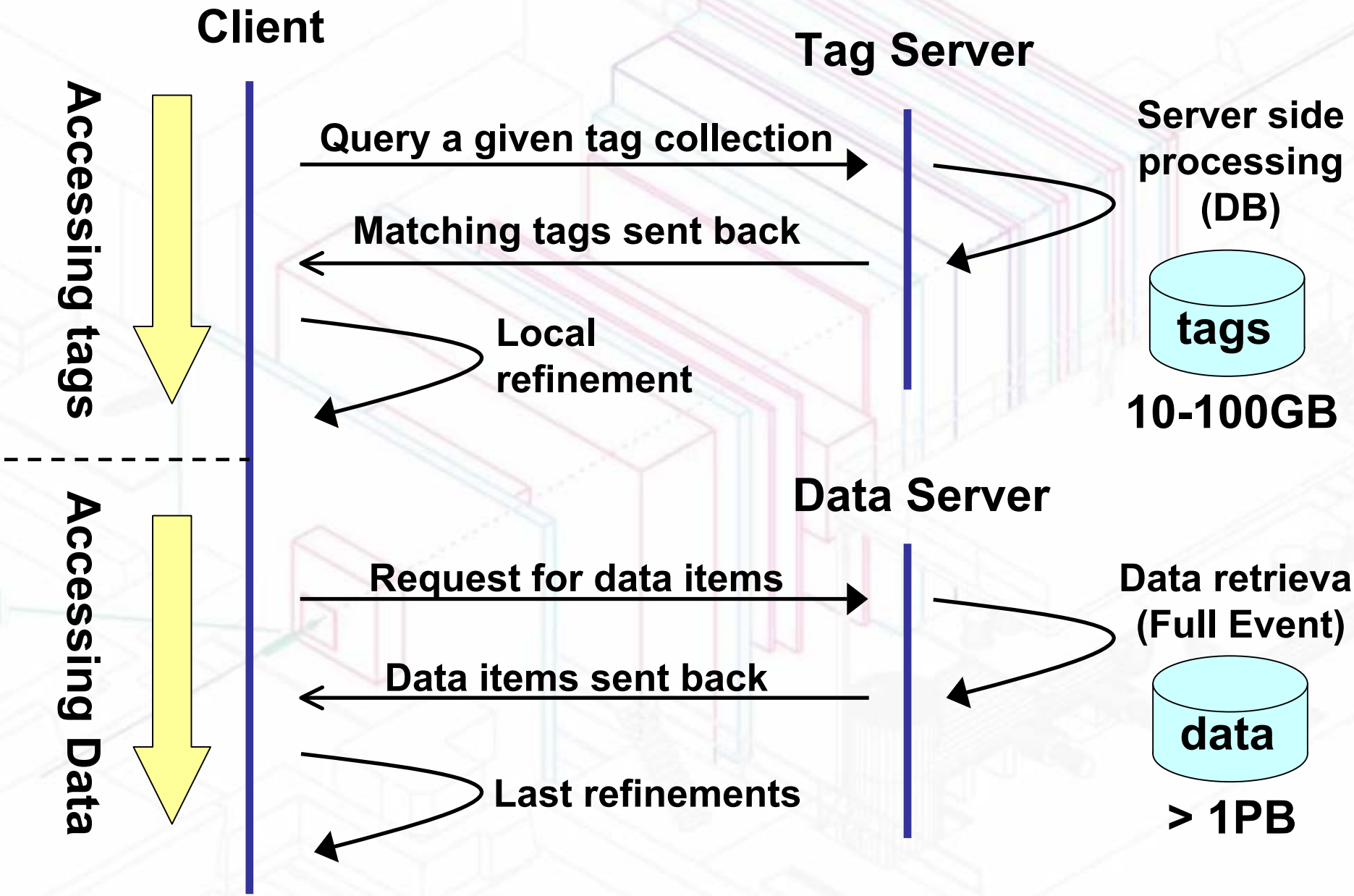
- A Tag Collection is a list of tags of the same type.
- There may be many collections with the same type.



# Selection Process

- **The selection process is very flexible :**
  - Selection of the tag collection implies a reduction of the number of data items of interest
  - Server-side preselection on tags using SQL-like criteria
  - Client-side refinement on tags using a high level programming language to maximize the preselection efficiency
  - Carry out the final refinement by reading selected full data items (high level programming language)

# Selection Process (2)

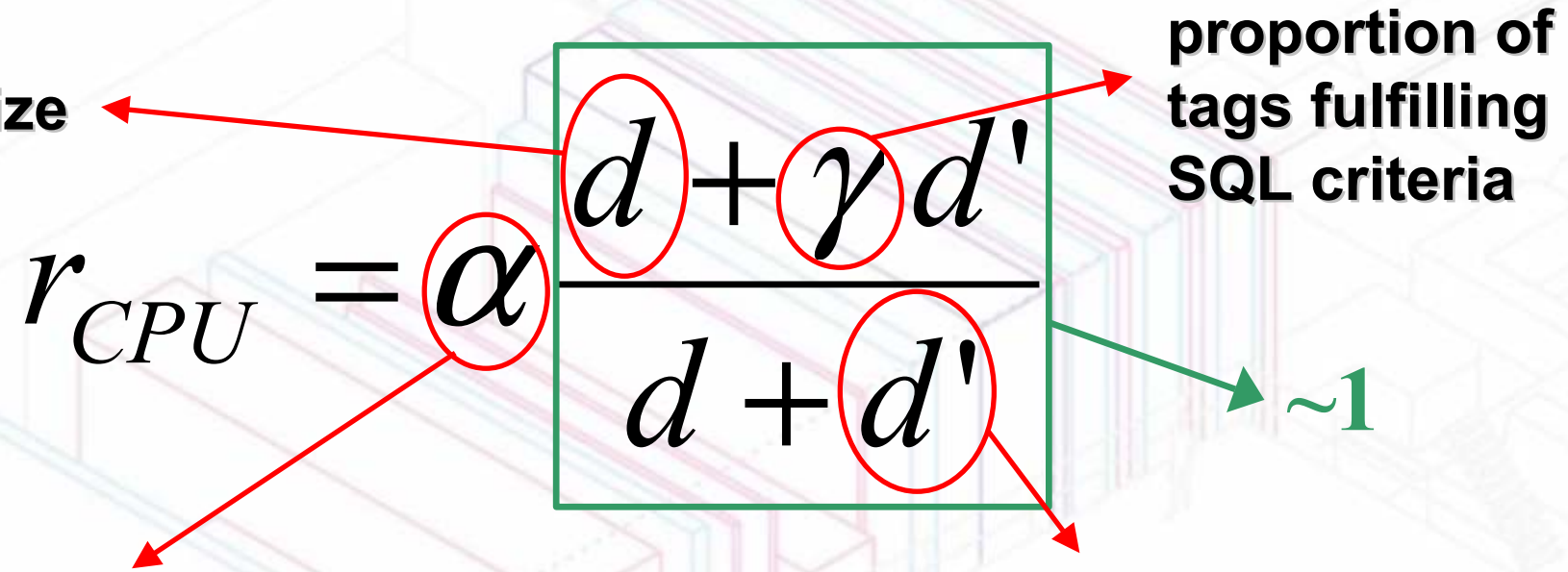


# Performance

- **The performance of the new retrieval schema can be evaluated by comparing it with a sequential scan**
- **Approximations :**
  - **data contains only integers**
  - **no optimizations at all (no pipelining, sequential scans...)**
  - **no local refinement step**
- **Performances are given under the form of ratios :**

$$ratio = \frac{selection\ with\ proposed\ indexing\ schema}{selection\ with\ sequential\ scan} < 1$$

# Processing Time Ratio




- Slightly better than α
- Main improvement : use of reduced size tag collection

# Network Load Ratio

$$r_{NET} \leq 2\,\alpha\,\gamma$$

**proportion of items present in tag collection.**

**proportion of tags fulfilling SQL criteria**

- $\alpha$ is due to the use of a tag collection (subset of events).
- $\gamma$ is the tag selection ratio
- 2 is a maximum. Depending on the latency, it can go down to 1 + $\beta$, $\beta$ being the tag size versus the data item size
- In practice, $\gamma << 1$ (~1% in LHCb) and $r_{NET} << \alpha$

# Retrieval Ratio (From Disk)

tag size versus
data item size

$$r_{DR} = \alpha(\beta + \gamma)$$

Reading
Data Items

proportion of items
present in tag collection.

Reading
Tags

proportion of tags
fulfilling SQL criteria

- $\beta$ is due to loading small tags instead of larger items
- $\gamma$ is the tag selection ratio
- $\alpha$ is due to the use of a tag collection (subset of events).
- usually $\beta << 1$ and $\gamma << 1$ ($10^{-4}$ and $10^{-2}$ in LHCb) thus $r_{DR} << \alpha$
- Tag size versus selection efficiency can be optimized

# Net Gain for LHCb

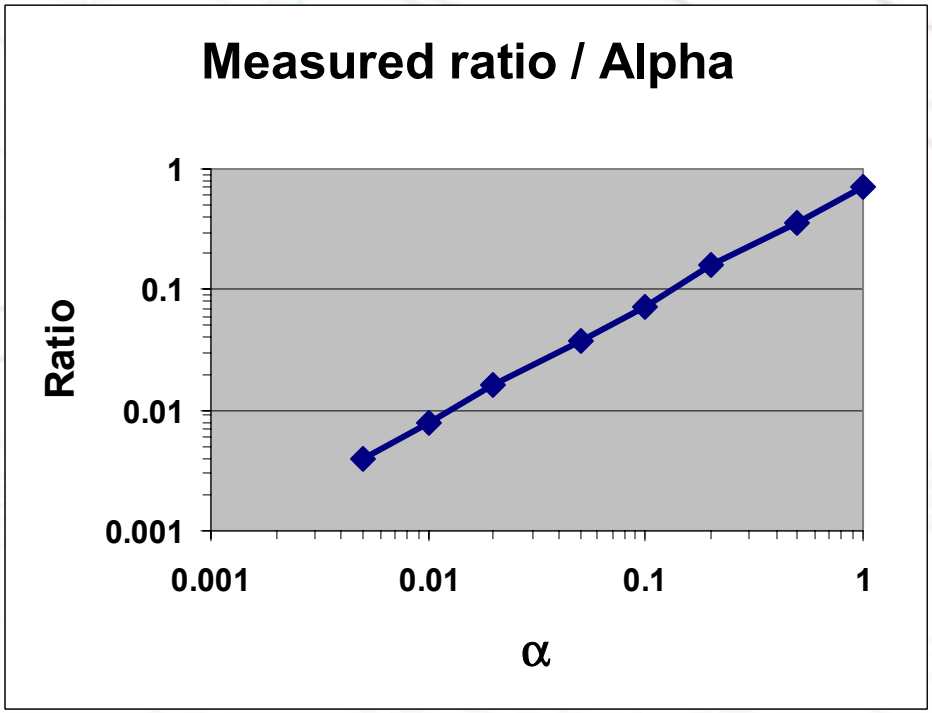
- **Typical values for are :**
  - Proportion of items in a collection : $\alpha \sim 10^{-4}$
  - Tag size versus item size : $\beta \sim 10^{-4}$
  - Proportion of tags fulfilling SQL criteria : $\gamma \sim 10^{-2}$
- **Typical gains are**
  - CPU time : $r_{CPU} \sim 10^{-4}$
  - Network load : $r_{NET} \sim 2.10^{-6}$
  - Retrieval time : $r_{DR} \sim 10^{-6}$

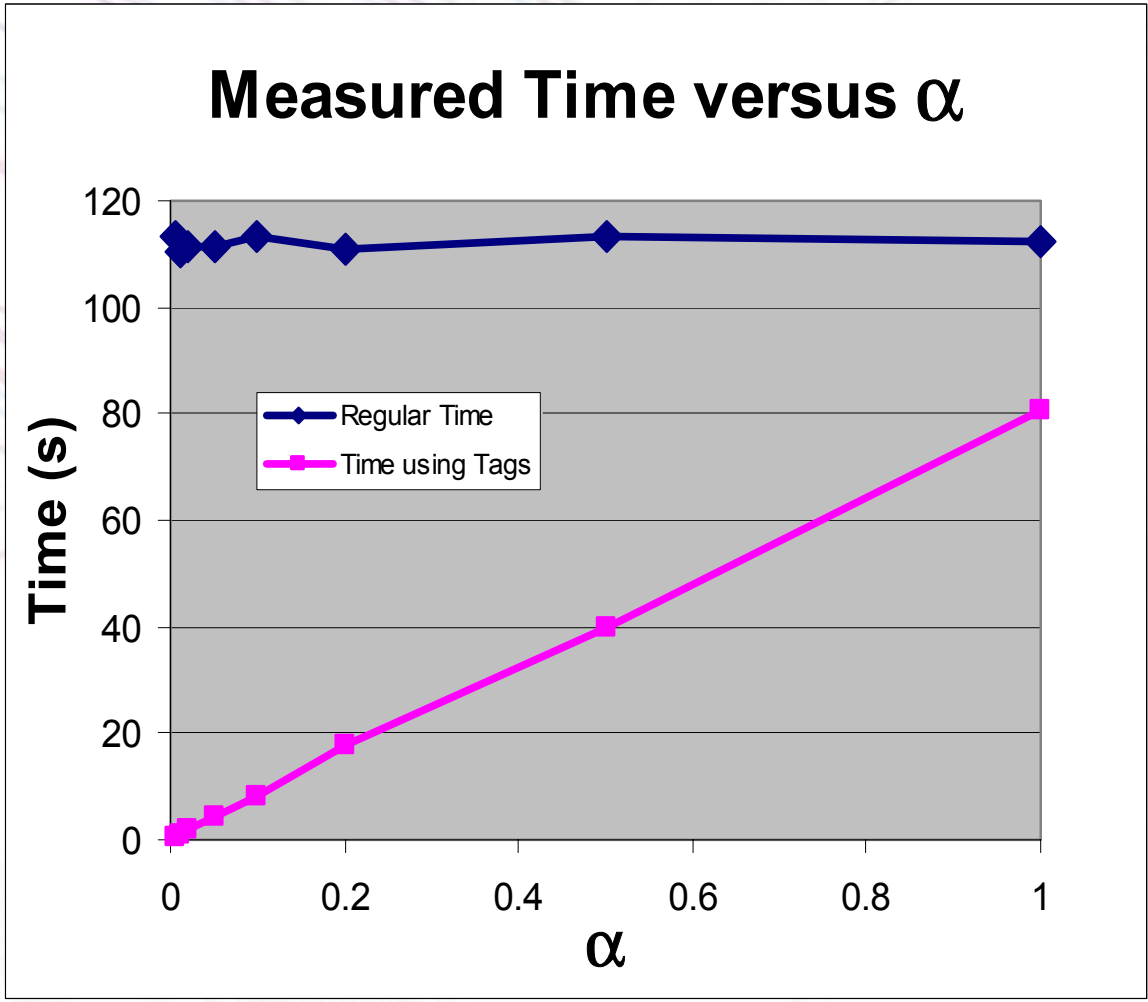
# First Measurements

- **The proposed schema is implemented within Gaudi (C++ LHCb event computation framework)**
- **Measurement conditions :**
  - **MySQL as a database.**
  - **Items of 160 KB, tags reduced to 15 B ($\beta \sim 10^{-4}$)**
  - **Only 5000 events in total (~800 MB)**
  - **No network, few CPU needed**
  - **Bottleneck is the retrieval from hard disk**
- **Overall ratio essentially equal to $\alpha\,\gamma$ :**
  - **$\alpha$ proportion of items present in tag collection**
  - **$\gamma$ proportion of tags fulfilling SQL criteria**

# Dependence on α
## (Tag Collection Size)

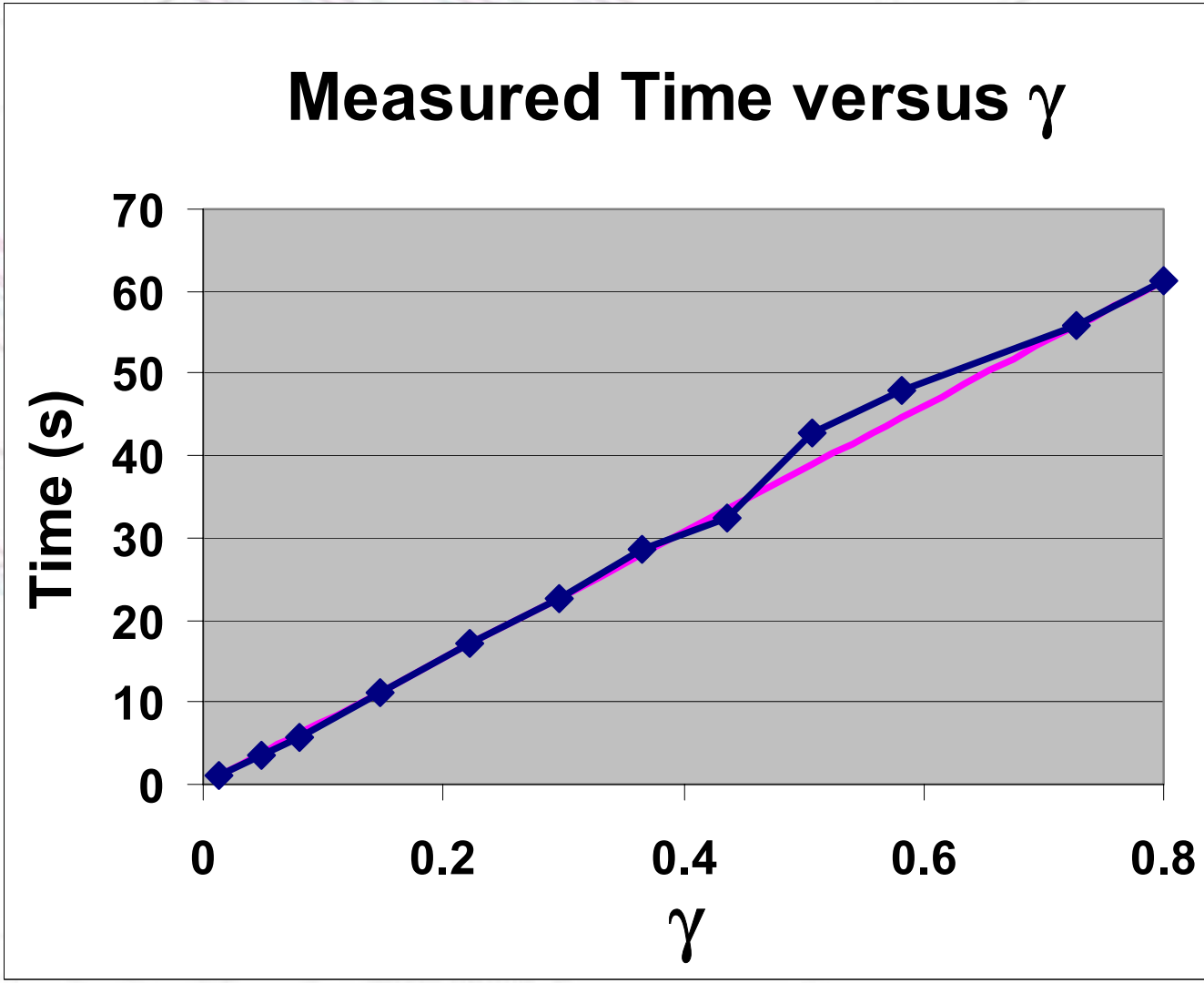
- N = 5000
- No SQL selection
- Measured dependency on α is linear as expected



# Dependence on γ
## (Selectivity of the SQL query)

- N = 5000
- Tag Collection containing all data items
- Dependency on γ is linear as expected



# Conclusions

- **Tag collections based indexing allows :**
  - various and powerful preselections (tag collection, SQL, high level programming language)
  - optimized network load (of the order of loading only matching items)
  - Large global gains (at least $10^4$ for LHCb)
- **Although developed as solution to a specific problem, the method is generic :**
  - adapted to data selection problems with highly selective multidimensional criteria, making use of a small subset of the data items
- **Tag collections may be accessed more efficiently by using existing indexing techniques on tags.**

# Future Work

- **The data selection schema will be parallelized :**
  - retrieving of tags/data items in parallel
  - carrying out I/O and local refinement as a pipeline
- **Interface with Grid software is foreseen :**
  - storage of data items in world-wide distributed databases
  - replication of the tag collections on different sites