UNIVERSITÉ
FRANÇOIS - RABELAIS
TOURS

Leveraging query logs for user-centric OLAP

Habilitation à Diriger les Recherches

Patrick Marcel

November 13, 2012

Agenda

1. OLAP not dead!
2. What's in a log?
3. Where do we go from here?
4. What's hidden in a log?
5. Long live user-centric OLAP!
6. The holo-deck and beyond

Standard Reports Are Enough. A common lesson is that standard reports, containing pre-defined text or graphics layouts and produced at regular intervals, is by far the most used way of interacting with the BI system. Even relatively simple analytical functionality like the one offered by OLAP systems seem to be used at lot less. Even though this is probably related to the "BI maturity" of the organization, the trend will likely remain.

T.B. Pedersen, "How is BI used in industry", DaWaK 2004

OLAP NOT DEAD!

Buzzwords?

- Long ago: BI

Buzzwords?

- Long ago: BI

One of my first papers: “A rule based data manipulation language for OLAP systems”, DOOD 1997

Buzzwords?

- Long ago: BI
- 2007: BI 2.0

Gartner: It's business intelligence 2.0 time

Summary: *Analyst group claims that the next version of business intelligence is about fewer vendors and better data management*

By Colin Barker | January 30, 2007 -- 17:18 GMT (09:18 PST)

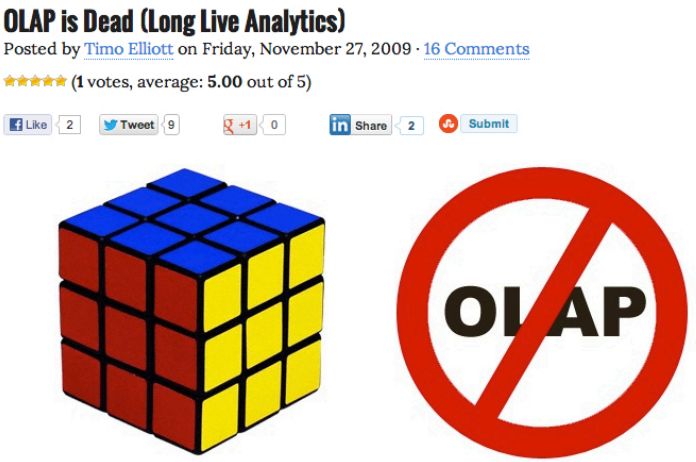
Buzzwords?

- Long ago: BI
- 2007: BI 2.0
- 2009 : Analytics

Gartner: It's business intelligence 2.0 time

Summary: *Analyst group claims that the next version of business intelligence is about fewer vendors and better data management*

By Colin Barker | January 30, 2007 -- 17:18 GMT (09:18 PST)



OLAP is Dead (Long Live Analytics)

Posted by Timo Elliott on Friday, November 27, 2009 · 16 Comments

(1 votes, average: 5.00 out of 5)

Like 2 Tweet 9 +1 0 Share 2 Submit

Buzzwords?

- Long ago: BI
- 2007: BI 2.0
- 2009 : Analytics
- 2012 : Big Data


Gartner: It's business intelligence 2.0 time

Summary: *Analyst group claims that the next version of business intelligence is about fewer vendors and better data management*

By Colin Barker | January 30, 2007 -- 17:18 GMT (09:18 PST)

Enterprise Data Analysis and Visualization: An Interview Study

- by Kandel, Paepcke, Hellerstein and Heer, in IEEE TVCG

Analysts often interacted closely with IT staff to complete aspects of their job. We observed that the IT team regularly provides four primary

Data sets may contain a number of quality issues that affect the validity of results, such as missing, erroneous or extreme values. Many

Another difficulty, reported by 23 analysts, was integrating data from multiple sources. Identifiers useful for joining records across data sets

It is really hard to know where the data is. We have all the data, but there is no huge schema where we can say this data is here and this variable is there. It may be written

Most respondents (31/35) noted that existing analytic packages, tools or algorithms did not scale with the size of their data sets. The thresh-

Some analysts (16/35) noted difficulty performing ad hoc grouping of observations, as in path or funnel analysis [36]. One analyst at a web

A number of analysts (17/35) also complained that reports were too inflexible and did not allow interactive verification or sensitivity analysis. Often reporting and charting tools were used directly on the out-

Enterprise Data Analysis and Visualization: An Interview Study

- by Kandel, Paepcke, Hellerstein and Heer, in IEEE TVCG

Analysts often interacted closely with IT staff to complete aspects of their job. We observed that the IT team regularly provides four primary

Data sets may contain a number of quality issues that affect the validity of results, such as missing, erroneous or extreme values. Many

Another difficulty, reported by 23 analysts, was integrating data from multiple sources. Identifiers useful for joining records across data sets

It is really hard to know where the data is. We have all the data, but there is no huge schema where we can say this data is here and this variable is there. It may be written

Most respondents (31/35) noted that existing analytic packages, tools or algorithms did not scale with the size of their data sets. The thresh-

Some analysts (16/35) noted difficulty performing ad hoc grouping of observations, as in path or funnel analysis [36]. One analyst at a web

A number of analysts (17/35) also complained that reports were too inflexible and did not allow interactive verification or sensitivity analysis. Often reporting and charting tools were used directly on the out-

Enterprise Data Analysis and Visualization: An Interview Study

UNIVERSITÉ FRANÇOIS - RABELAIS TOURS

- by Kandel, Paepcke, Hellerstein and Heer, in IEEE TVCG

Analysts often interacted closely with IT staff to complete aspects of their job. We observed that the IT team regularly provides four primary

Data sets may contain a number of quality issues that affect the validity of results, such as missing, erroneous or extreme values. Many

Another difficulty, reported by 23 analysts, was integrating data from multiple sources. Identifiers useful for joining records across data sets

It is really hard to know where the data is. We have all the data, but there is no huge schema where we can say this data is here and this variable is there. It may be written

Most respondents (31/35) noted that existing analytic packages, tools or algorithms did not scale with the size of their data sets. The thresh-

Some analysts (16/35) noted difficulty performing ad hoc grouping of observations, as in path or funnel analysis [36]. One analyst at a web

A number of analysts (17/35) also complained that reports were too inflexible and did not allow interactive verification or sensitivity analysis. Often reporting and charting tools were used directly on the out-

Very early DW literature? This paper is from Oct. 2012 issue!

OLAP as an ideal use-case

- We need more user-friendly DBMSs
  - See e.g., [Jagadish & al., SIGMOD 2007], [Khoussainova & al., CIDR 2009], [Nandi & Jagadish, VLDB 2011]
  - But also: Skyline, Preference SQL, QueRIE, SnipSuggest, etc.
- and more user-friendly OLAP!
  - Specificities:
    - Well defined topology due to the multidimensional schema
    - Exploration by navigation, roll-up, drill-down, slice
    - Dedicated MDX language
    - Read mostly, non volatile, multi-user, etc.

User-centric approaches

- Formulation effort
- Proactiveness
  - Content based
  - Collaborative filtering
- Prescriptiveness
- Expressiveness

Low formulation effort, proactive, not too prescriptive, expressive enough… the best approach?

Idea: use the query log to reduce the formulation effort and enhance proactiveness

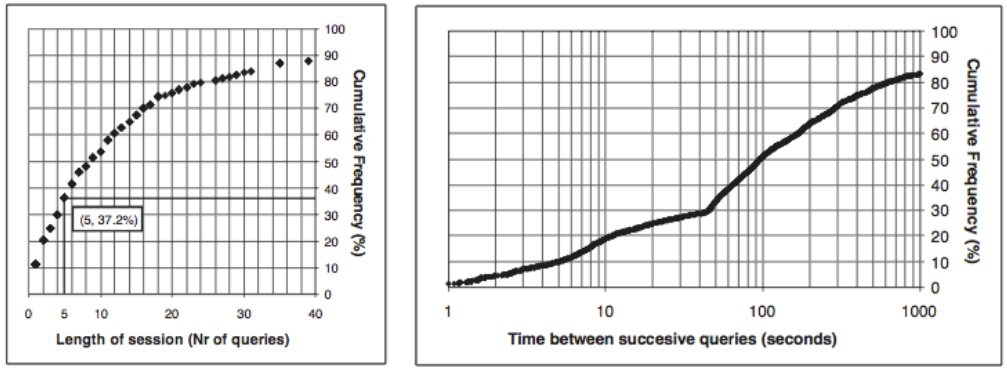


Fig. 2. Cumulative distributions of session length and consideration time

In OLAP systems, the *navigational nature* of the workload is guaranteed as long as the user interactively formulates his next request using the results of the previous request ([9]). We call such a sequence of navigational queries a session. The analysis showed that typical OLAP sessions have a considerable length and are thus suited for prediction approaches. The left hand side of Figure 2 shows the cumulative frequency distribution of the session length. It is obvious, that only 11% of the sessions consisted of executing a single query (simple reporting). On the other hand, some of the sessions contained more than 100 queries. If we assume that accurate prediction is possible for sessions with 5 or more consecutive queries, Figure 2 shows that 63,8% of the sessions fulfill this condition.

C. Sapia, "PROMISE: Predicting Query Behavior to Enable Predictive Caching Strategies for OLAP Systems", DaWaK 2000

Part 2: Modelling OLAP user activities

WHAT'S IN A LOG?

What is an OLAP session?

- No definition of session in the literature
- Though concept is used
  - e.g., in [Sapia, DaWak 2000], [Sarawagi, VLDB 2001], [Cariou & al., DaWaK 2008]
- Definitions exist in other domain
  - e.g., search session in the Web

Simple viewpoint: a sequence of queries, possibly separated by an OLAP operation

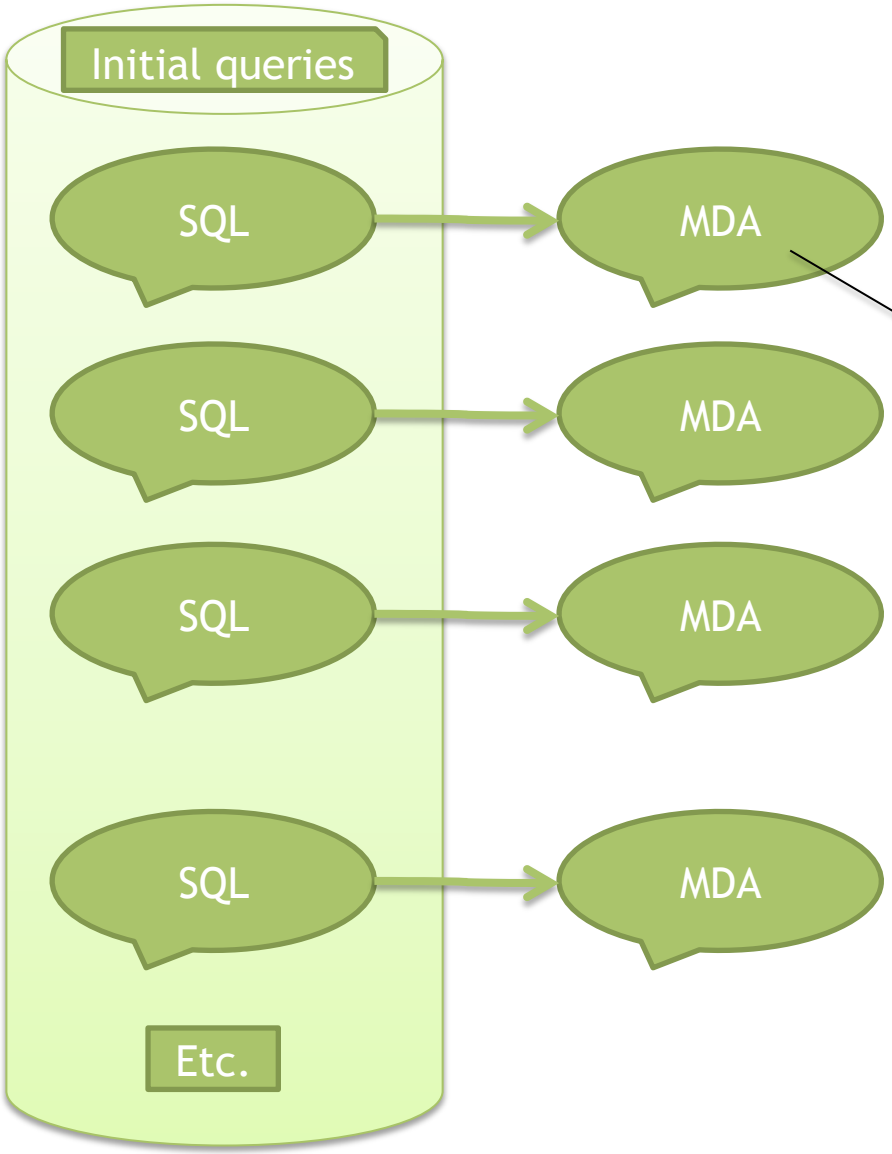
Detecting OLAP sessions

- Joint work with Univ. Polytecnica Catalogna and Ensma Poitiers
  - Master's thesis work of Jovan Varga

Idea: detect semantic connections between queries, where semantics is given by OLAP operations

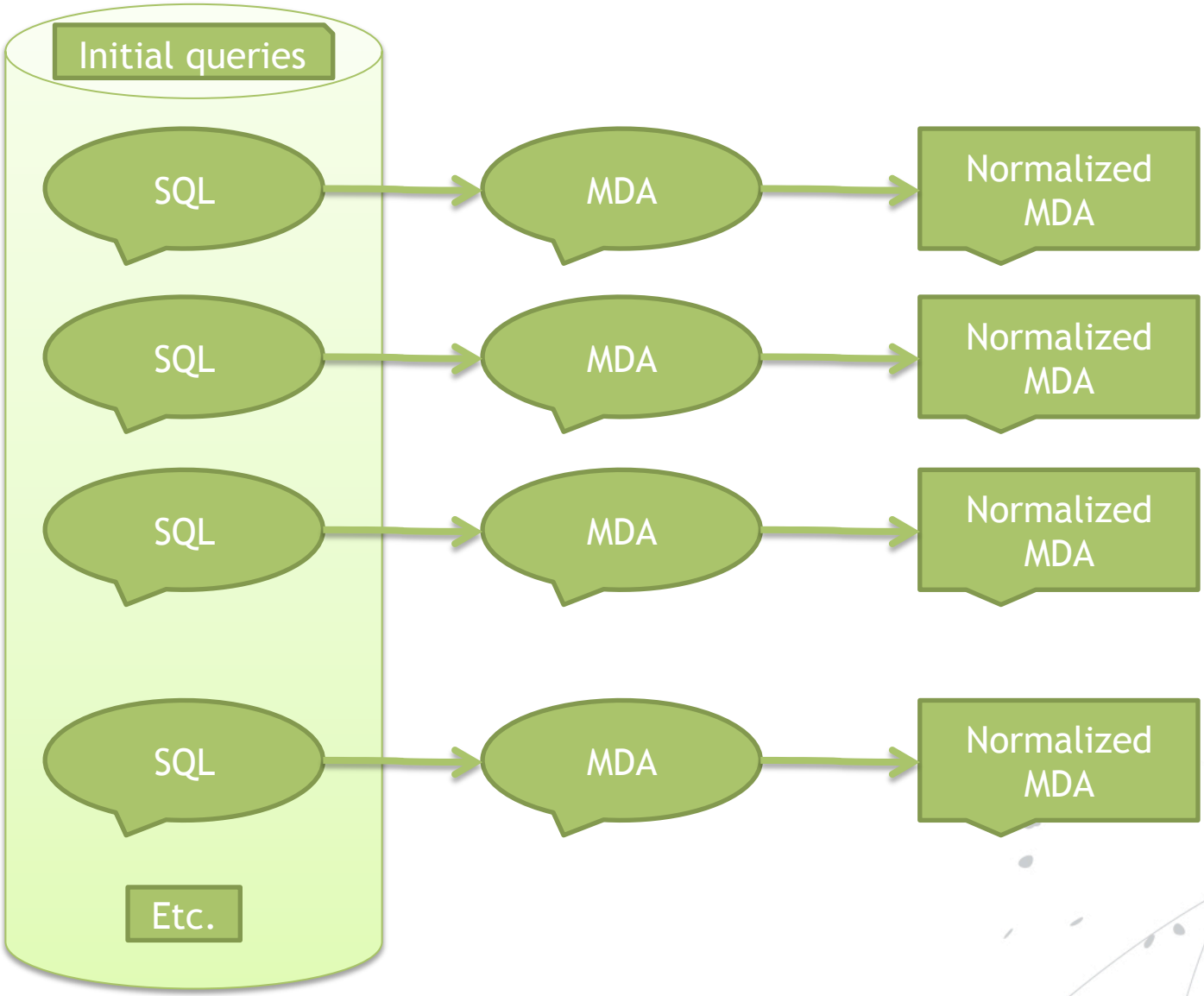
Detecting OLAP sessions

Detecting OLAP sessions

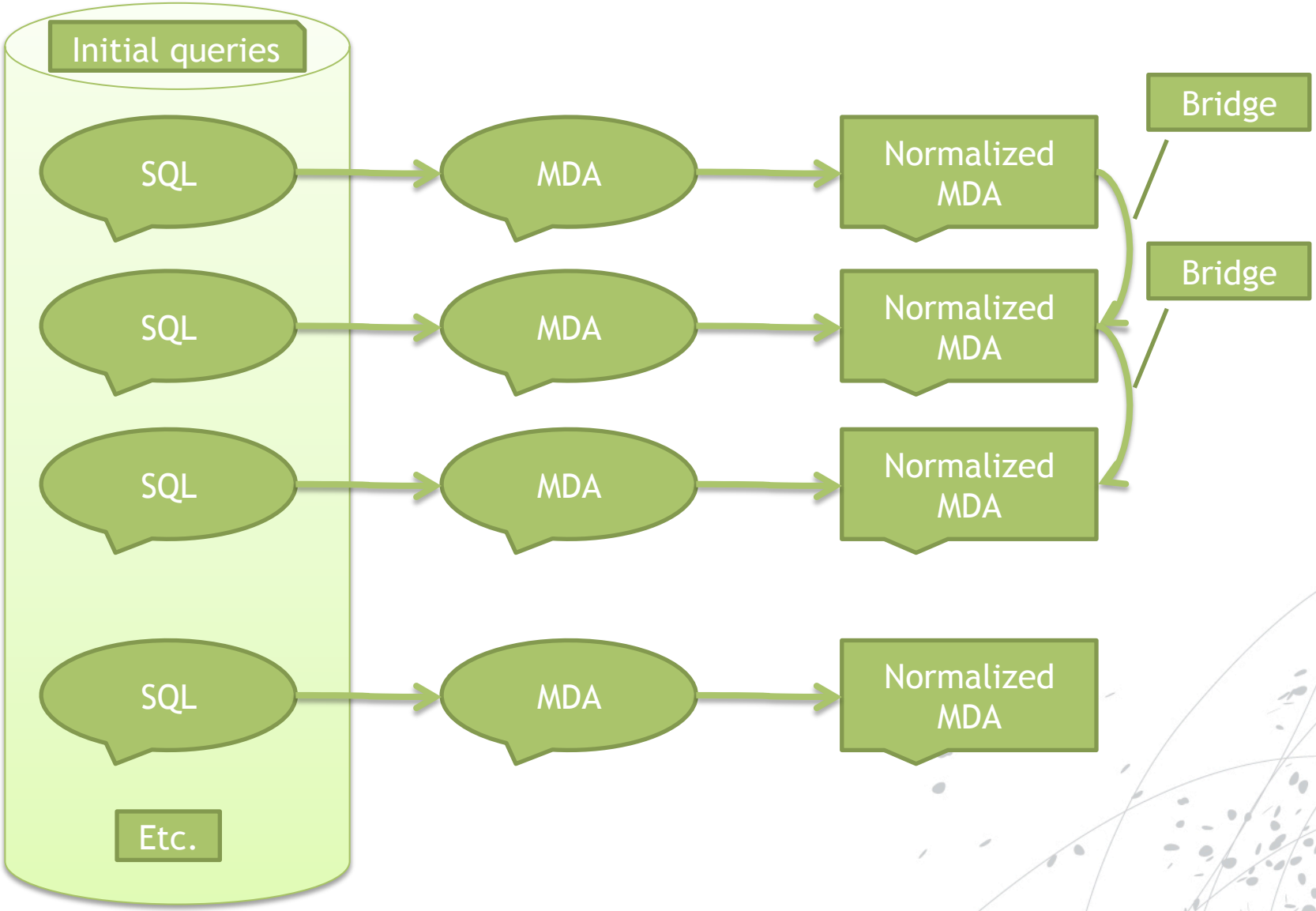


MultiDimensional Algebra [Abello & Romero, DaWak 2007]

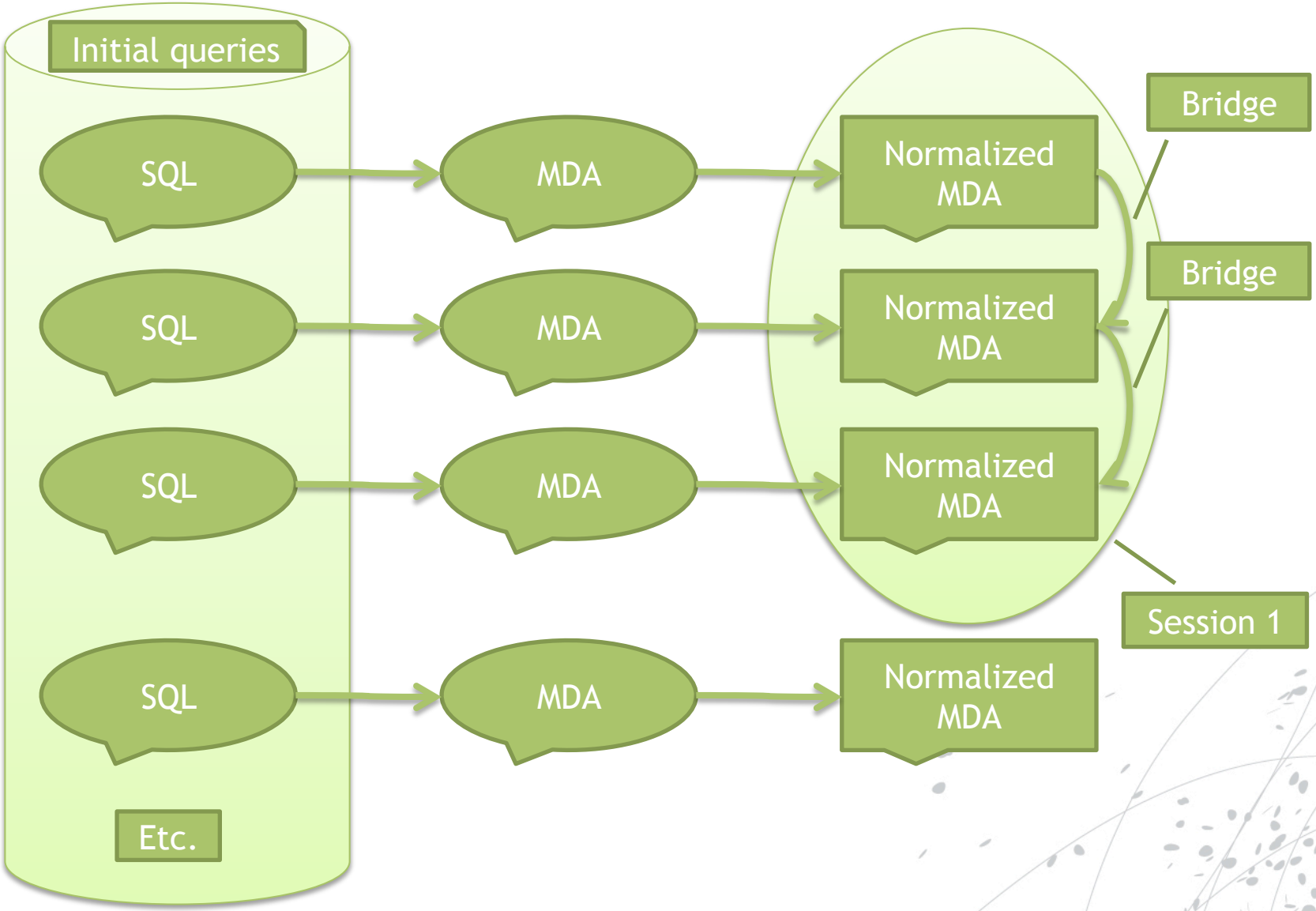
Detecting OLAP sessions



Detecting OLAP sessions



Detecting OLAP sessions



What is an OLAP query?

- A syntax (query intension)
  - Set of fragments
  - Slicers, group by set, measure set

```
SELECT CROSSJOIN({Paris,Bruxelles},
                 {2010,2011}) ON ROWS,
                 [Income Range].Members ON COLUMNS
FROM CENSUS
WHERE (Measure.[Elec. Consumption])
```

- A partially evaluated expression

Elec. Cons.		income<100	100<income<500	income>500
Paris	2010			
	2011			
Bruxelles	2010			
	2011			

- The retrieved answers

Elec. Cons.		income<100	100<income<500	income>500
Paris	2010	80	90	100
	2011	50	60	100
Bruxelles	2010	80	100	120
	2011	60	70	110

Effectiveness/efficiency trade-off
[Chatzopoulou & al., DE Bulletin 2011]

2.2 Search and Browse Interaction

One essential query management feature is the ability for users to search for and browse through past queries. We refer to this mode of interaction as the *Search and Browse Interaction Mode*.

Search. A *meta-query* is a query that searches for queries. Such queries enable users to locate past queries matching specific search conditions. The resulting queries can then be learned from, re-executed, or used as a starting point to compose a new query. A

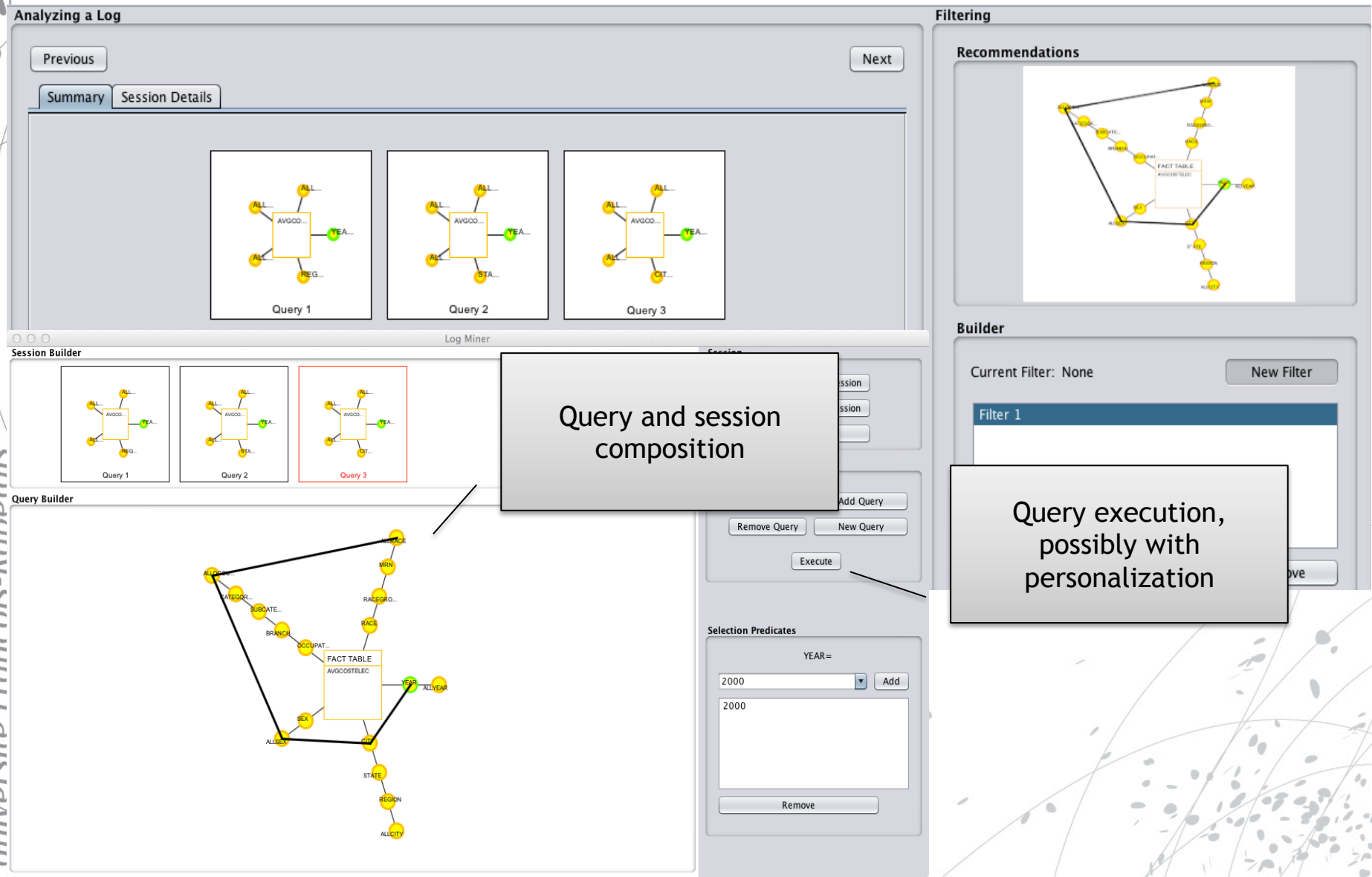
Browse. After finding the desired queries, the CQMS must allow the user to browse the results. Many systems that provide query logging [11, 15, 26, 32, 33] also allow the user to view the log in a table or a file. However, to make the query log suitable for browsing, the CQMS needs to present it in a comprehensible, summarized format. One possible method is to present *query sessions* instead of individual queries. A query session is a series of (often similar) queries with the same information goal in mind. Such

N. Khoussainova et al., "A Case for A Collaborative Query Management System", CIDR 2009

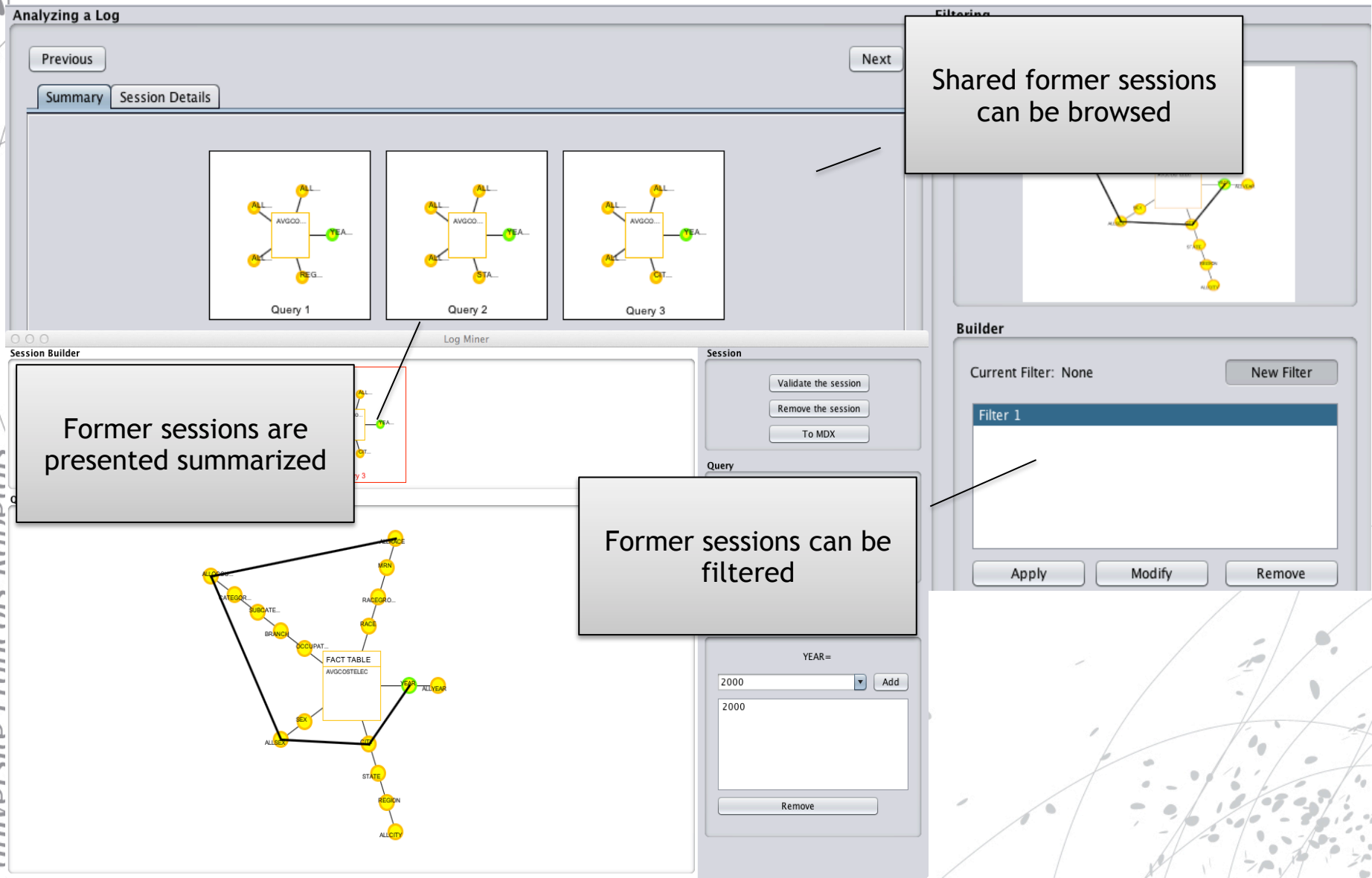
Part 3: An OLAP query management system

WHERE DO WE GO FROM HERE?

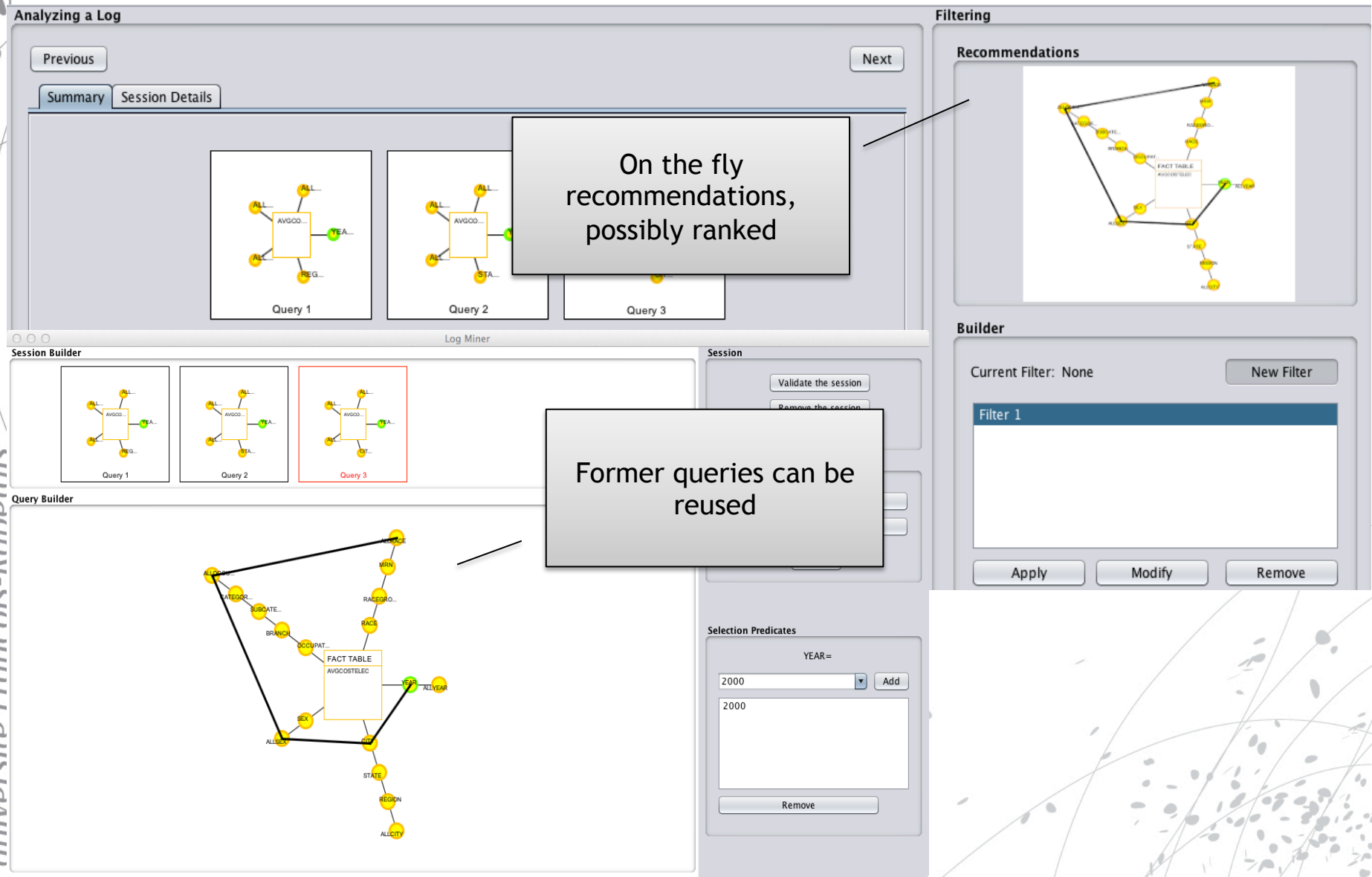
Envisioned CQMS for OLAP



Envisioned CQMS for OLAP

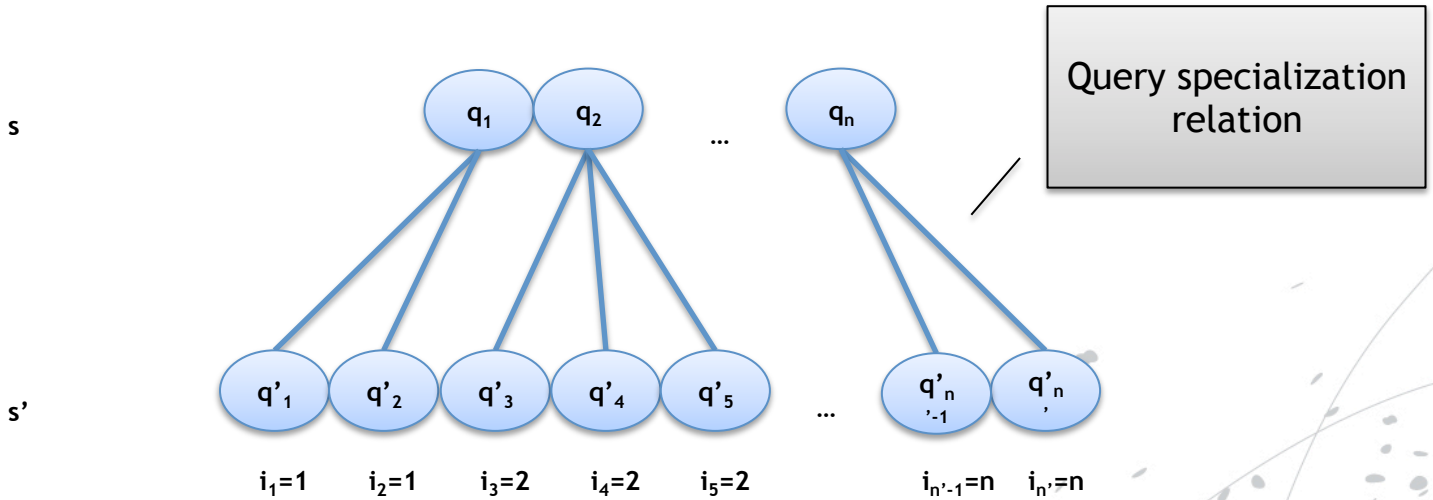


Envisioned CQMS for OLAP



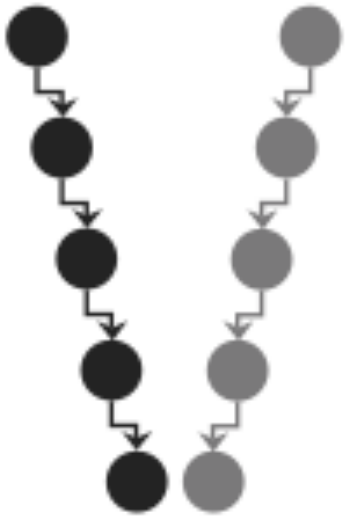
Session specialization

- PhD thesis of Julien Aligon
  - Specialization relations over queries and over sessions
  - For query intension or partially evaluated queries



Session comparison

- PhD thesis of Elsa Negre
  - Edit distance for sessions, Hausdorff distance for partially evaluated queries
- Joint work with Univ. Bologna
  - A dedicated similarity for query intensions
  - Sequence alignment, extensions of TF-IDF and Dice coefficient for sessions
  - Involving users to validate the approach



Towards a logical framework for log-based user-centric approaches

- Declarative expression
- Characterizing the expressiveness
- Derive and use logical properties, like e.g.,
  - Optimize user-centric techniques
  - Compare user-centric activities

"**All models are wrong**, but some are useful."

G. Box, Empirical model building and response surface, 1987.

Part 4: Extracting knowledge from the log

WHAT'S HIDDEN IN A LOG?

A log can tell about:

- Simple user preferences on multidimensional objects
- Navigational habits
- Analysis discoveries
- User expectations

The richer the query model, the better the user support

Simple user preferences

- PhD thesis of Hassina Mouloudi
- Weak orders over data, measures and dimensions
  - based on the frequency in the log
- Allowing to define orders over queries
  - E.g., Year 2008 > Year 2009, Location > Time,

Elec. Cons.		income<100	100<income<500	income>500
Paris	2008			
	2009			
	2010			
	2011			

>

Elec. Cons.		income<100	100<income<500	income>500
Paris	2010			
	2011			
Bruxelles	2010			
	2011			

Navigational habits

- Joint work with Univ. Bologna
- If-then like patterns over data and attributes

  - E.g., *if* query features Year 2008 and groups by Bank Names, *then* it (*often*) features Measure=Losses

Analysis discoveries

- PhD thesis of Elsa Negre
- Complex relations over query answers:
  - important difference of 2 measure values
  - the queries drilling down this difference

    - E.g., the *difference* in Profit between Years 2007 and 2008 was investigated by
      - *drilling down* to the Country level or
      - *drilling down* to the Customer Income level

User expectations

- Joint work with Univ. Bologna and Univ. Québec en Outaouais
- A model of the cube as expected by the user

		Redtab	Silvertab	Loose	Lowrise
Ontario	Jan.11	20	20	20	20
	Feb.11	20	20	20	20
NY	Jan.11	20	20		
	Feb.11	20	20		

What to expect here?
If total=320, 20 everywhere maximizes Entropy

What can we do about it: Talk to our inner Liberal Arts Major.

Acknowledge that most people hate numbers and math.

Recognize that they need answers to business questions that are much harder to formulate than one would think.

And still, they and nobody else should be in power.

Don't expect too much respect for structure from users. Give them the content they want.

Don't expect complete and clean thought models. Accept and accompany incremental, trial-and-error approaches.

Gently guide them to what makes common sense (which also means out of Excel).

Make BI pervasive and invisible.

SAP BusinessObjects

Yannick Cras, "Why simple BI questions are not that simple", eBISS 2011

Part 5: log-driven user-centric analysis

LONG LIVE USER-CENTRIC OLAP!

With a single-user log

- Personalizing queries for avoiding too large answers
- In a prescriptive or non prescriptive manner:
  - Inferring query fragments from simple preferences for expanding queries
  - Inferring preference constructs from navigational habits

With a multi-user log

- Proactive approaches
- Recommending queries to help users analyzing cubes, in a collaborative fashion:
  - Users who used this value/level/measure frequently used also that one
  - Users who launched session similar to yours also launched that query
  - Users who investigated this difference also launched that query

With a shared log

- For session browsing, searching and reuse
  - Cluster sessions
  - Summarize sessions
  - Filter, browse and drill through session summaries

Use case 1

- User browses queries that concern Year 2011 with measure Elec. Consumption
- Obtains two clusters summarized by:
  - C1:
    - Elec. Consumption
    - By Ownership
    - For Year 2011, France
  - C2:
    - Elec. Consumption
    - By Income range
    - For Year 2010, 2011
- Drills C1 down to a more precise session:
  - Q1:
    - Elec. Consumption
    - By Ownership
    - For Year 2011, France
  - Q2:
    - Elec. Consumption
    - By Ownership
    - For Year 2011, France, Germany
- User starts her session with Q2

Use case 2

- The current session:
  - Q1:
    - Elec. Consumption, Gas consumption
    - By Ownership
    - For Year 2011, France
  - Q2:
    - Elec. Consumption, Gas consumption
    - By Income range and Ownership
    - For Year 2011, France

Recommending

- This past session is very closed
  - Q'1:
    - Elec. Consumption, Gas consumption, Water consumption
    - By Occupation
    - For Year 2011, France
  - Q'2:
    - Elec. Consumption, Gas consumption, Water consumption
    - By Occupation and Ownership
    - For Year 2011, France, Germany
  - Q'3:
    - Elec. Consumption, Gas consumption, *Water consumption*
    - By Occupation, Ownership, *Region*
    - For Year 2011, France, *Germany*
- Q'3 is recommended

Personalizing

- Q'3's answer may be quite large
- User's navigational habits indicate that:
  - If the query contains France, then it often contains measure Elec. consumption
- Q'3 is personalized and becomes:
  - Elec. Consumption, Gas consumption, Water consumption
    - **Preferring** Elec. Consumption
  - **By** Occupation, Ownership, Region
  - **For Year** 2011, France, Germany

Use case 3

- The current session:
  - Q1:
    - Elec. Consumption
    - By Ownership
    - For Year 2010, 2011
  - Q2:
    - Elec. Consumption
    - By Income range
    - For Year 2010, 2011

- Q2 answer:

Elec. Cons.	Year	
Income range	2010	2011
income<100	80	50
100<income<500	90	60
income>500	100	100

Recommending

- Past sessions also investigated

Elec. Cons.	Year	
Income range	2010	2011
income<100	80	50
100<income<500	90	60
income>500	100	100

with

  - Q4:
    - Elec. Consumption
    - By Countries
    - For year 2010, 2011, 100<income<500
  - Q5:
    - Elec. Consumption
    - By Ownership
    - For Year 2010, 2011, income<100

- Recommend Q4 and Q5

Personalizing

- Simple preferences indicate that:
  - income<100 is preferred to 100<income<500
  - Dimension Income preferred to both dimensions Location and Ownership
- thus Q5 is preferred to Q4
- Recommend Q5 first

This paper is somewhere between a vision paper and a survey paper on interactive data warehouse systems. I very much like many ideas of the authors, such as the idea of a smart data warehouse system that recommends next operations, based on distributions in result cubes and the expected utility for the user. That may contain outlier identfification, cluster identfication, dependency estimation and finally visualization selection. This vision is somehow close to the holo-deck of a star wars episode.

Anonymous, review to our IJDWM paper entitled “an envisioned approach for modelling and supporting user-centric query activities on data warehouses”

Part 6: Perspectives

THE HOLO-DECK AND BEYOND

Chronology

- 2005
  - Collaboration with Ladjel Bellatreche (Poitiers)
  - First paper on OLAP and query personalization (Dolap)
- 2007
  - First paper on log browsing and searching for OLAP (EDA)
  - PhD defense on OLAP query personalization (Hassina Mouloudi)
- 2008
  - First paper on OLAP and query recommendation (Dolap)
- 2009
  - Various query recommendation approaches (DaWaK, Dolap)
  - PhD defense on OLAP query recommendation (Elsa Negre)
- 2010
  - Collaboration with Oscar Romero and Alberto Abello (Barcelona)
  - First paper on OLAP log summarization (EGC)
- 2011
  - Collaboration with Mateo Golfarelli and Stefano Rizzi (Bologna)
  - Navigational habits for proactive personalization (Adbis)
  - Detection of OLAP sessions (DaWaK)
- 2012
  - First logical framework for log manipulation (persDB)
  - Collaboration with Rokia Missaoui (Québec)
  - Various log-based summarization approaches (EDA, Dolap)
- 2013
  - The paper with the “holo-deck” review, written after the 2011 Dagstuhl seminar (IJDWM)

Short term perspectives

- Other recommendation and personalization approaches
  - navigational habits, expectations, etc.
  - Evaluated queries for current session and intensions from the log
- Making it all work together
  - Orchestrating personalization, recommendation, session reuse
  - To Browse or not to browse?
  - What about crowdsourcing?

Long term perspectives

- So far, research in DW & OALP has mostly tackled efficiency
- What about the effectiveness of analyses?
  - Quality of a query? Of an answer?
  - Quality of a session?
  - What about the user's understanding of data?
- Back to intensional answering?
  - Support queries like "what can you tell me about…"
  - Explain and motivate the answer

Thank you!

Tag cloud of my DBLP page