

Databases 'R' Us: Databases at the Turning Point

Presentation to IU's Data and Search Institute

Antonio Badia
University of Louisville

October 7, 2008

Outline

The Current Situation

Diagnostic

The Information Architecture View

A new collar on an old dog?

My take: Some Future Projects

Databases as Social Resources

Database Workflow

Database Dialogs

Some (Preliminary) Conclusions

The Current Situation

Change is on the air: The Claremont Report

The Current Situation

Change is on the air: The Claremont Report

- ▶ Turning point for databases: "[these factors] signal an urgent, widespread need for new data management technologies."

The Current Situation

Change is on the air: The Claremont Report

- ▶ Turning point for databases: "[these factors] signal an urgent, widespread need for new data management technologies."
- ▶ Reading it, one has the feeling that there is worry in the air: "in recent years, our externally visible impact has not evolved sufficiently beyond traditional database systems and enterprise data management."

The Current Situation (Cont.)

More Clarendon Report

- ▶ New agenda: DB researchers should be "focusing outside the traditional RDBMS stack and its existing interfaces, emphasizing new data management systems for growth areas like e-science. In addition, database researchers should take data-centric ideas (declarative programming, query optimization) outside their original context in storage and retrieval, and attack new areas of computing where a data-centric mindset can have major impact."

The Current Situation (Cont.)

More Clarendon Report

- ▶ New agenda: DB researchers should be "focusing outside the traditional RDBMS stack and its existing interfaces, emphasizing new data management systems for growth areas like e-science. In addition, database researchers should take data-centric ideas (declarative programming, query optimization) outside their original context in storage and retrieval, and attack new areas of computing where a data-centric mindset can have major impact."
- ▶ "The time is ripe for various sub-communities to move out of the conceptual and algorithmic phase, and work together on comprehensive artifacts (systems, languages, services) that combine multiple techniques to solve complex user problems." Amen!

The Current Situation (Cont.)

Other voices:

- ▶ Pat Hanrahan (CS, Stanford): "When analyzing information, no single person knows it all. When you have a group look at data, you protect against bias. You get more perspectives, and this can lead to more reliable decisions."

The Current Situation (Cont.)

Other voices:

- ▶ Pat Hanrahan (CS, Stanford): "When analyzing information, no single person knows it all. When you have a group look at data, you protect against bias. You get more perspectives, and this can lead to more reliable decisions."
- ▶ Martin Wattenberg and Fernando B. Vidas (I.B.M. Research, Cambridge): "The conversation about the data is as important as the flow of data from the database."

The Current Situation (Cont.)

Other voices:

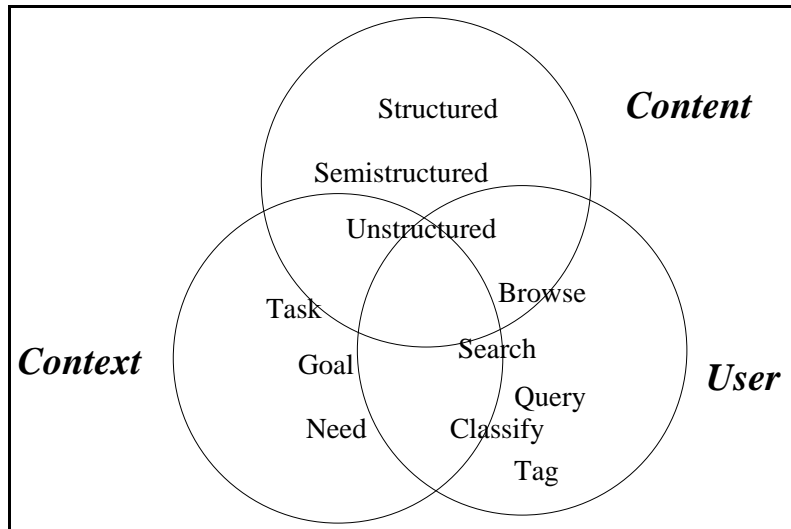
- ▶ Pat Hanrahan (CS, Stanford): "When analyzing information, no single person knows it all. When you have a group look at data, you protect against bias. You get more perspectives, and this can lead to more reliable decisions."
- ▶ Martin Wattenberg and Fernando B. Vidas (I.B.M. Research, Cambridge): "The conversation about the data is as important as the flow of data from the database."
- ▶ Ben Shneiderman (CS, U. Maryland-College Park): "The great fun of information visualization is that it gives you answers to questions you didn't know you had."

Diagnostic

To see the problem, I look at the *Information Architecture View*:
not just the database, but the environment it is in.

Diagnostic

To see the problem, I look at the *Information Architecture View*: not just the database, but the environment it is in.



Diagnostic (Cont.)

- ▶ Of all 3 factors in Information Architecture (Content, User, Context) only part of one is addressed.

Diagnostic (Cont.)

- ▶ Of all 3 factors in Information Architecture (Content, User, Context) only part of one is addressed.
- ▶ Of all user behavior (browse, search, query) only one thing (query) is supported.

Diagnostic (Cont.)

- ▶ Of all 3 factors in Information Architecture (Content, User, Context) only part of one is addressed.
- ▶ Of all user behavior (browse, search, query) only one thing (query) is supported.
- ▶ Of all data, only (semi)structured is supported.

Diagnostic (Cont.)

- ▶ Of all 3 factors in Information Architecture (Content, User, Context) only part of one is addressed.
- ▶ Of all user behavior (browse, search, query) only one thing (query) is supported.
- ▶ Of all data, only (semi)structured is supported.
- ▶ Of all organization (classification, tagging, indexing) only one is supported.

Diagnostic (Cont.)

- ▶ Of all 3 factors in Information Architecture (Content, User, Context) only part of one is addressed.
- ▶ Of all user behavior (browse, search, query) only one thing (query) is supported.
- ▶ Of all data, only (semi)structured is supported.
- ▶ Of all organization (classification, tagging, indexing) only one is supported.
- ▶ Information flow is severely restricted.

Diagnostic (Cont.)

- ▶ Of all 3 factors in Information Architecture (Content, User, Context) only part of one is addressed.
- ▶ Of all user behavior (browse, search, query) only one thing (query) is supported.
- ▶ Of all data, only (semi)structured is supported.
- ▶ Of all organization (classification, tagging, indexing) only one is supported.
- ▶ Information flow is severely restricted.
- ▶ And, on top of that, *dynamic dimensions* (anything that can change over time, including context, user profile, etc.) are usually ignored.

Diagnostic (Cont.)

Conclusion:

What Databases do is too narrow

Diagnostic (Cont.)

Conclusion:

What Databases do is too narrow

Narrow, technically well-defined, research. It produces results. But is it relevant?

Diagnostic (Cont.)

Conclusion:

What Databases do is too narrow

Narrow, technically well-defined, research. It produces results. But is it relevant?

Example: the problem of sharing data is much more than matching schemas. It involves:

- ▶ discovery of data to share;
- ▶ Access to the data;
- ▶ Understanding of the data;

(Smith, Seligman, and Swarup, IEEE Computer, Sept 2008)

Diagnostic (Cont.)

What are we to do?

Diagnostic (Cont.)

What are we to do?

Listen to Engelbert: "When adults accomplish something important, they almost always do it as a group activity. If computing is to amount to anything, it should become an amplifier of the collective intelligence of groups" (quoted by Alan Key, interview in CIO Insight, Feb 2007).

Diagnostic (Cont.)

What are we to do?

Listen to Engelbert: "When adults accomplish something important, they almost always do it as a group activity. If computing is to amount to anything, it should become an amplifier of the collective intelligence of groups" (quoted by Alan Key, interview in CIO Insight, Feb 2007).

The truth is: Users can not only create content, they can also create structure (perhaps implicitly, through use, or explicitly, through tags).

My Take

My Take

- ▶ We should take seriously the modeling and addition of Context, User.

My Take

- ▶ We should take seriously the modeling and addition of Context, User.
- ▶ Challenges:
 - ▶ if people (social, cognitive, psychological factors are the most critical resource, can technology do anything?
 - ▶ if context is vague, unbounded, can technology handle it?

My Take

- ▶ We should take seriously the modeling and addition of Context, User.
- ▶ Challenges:
 - ▶ if people (social, cognitive, psychological factors are the most critical resource, can technology do anything?
 - ▶ if context is vague, unbounded, can technology handle it?
- ▶ Be an *enabler*: understand (model) those resources, then *influence* (enhance, diminish) them.

My Take

- ▶ We should take seriously the modeling and addition of Context, User.
- ▶ Challenges:
 - ▶ if people (social, cognitive, psychological factors are the most critical resource, can technology do anything?
 - ▶ if context is vague, unbounded, can technology handle it?
- ▶ Be an *enabler*: understand (model) those resources, then *influence* (enhance, diminish) them.
- ▶ But biggest problem is old-mentality: we know what is best about data/information, so we decide. This is a top-down, hierarchical, taxonomical thinking in a bottom-up, tag-as-you-go, users create/decide world.

My take: Some Future Projects

So, what should we be doing?

My take: Some Future Projects

So, what should we be doing?

Basic idea #1: support *Interaction*

My take: Some Future Projects

So, what should we be doing?

Basic idea #1: support *Interaction*

Supporting Interaction

- ▶ Among users:
 - ▶ Collaboration
 - ▶ Sharing

My take: Some Future Projects

So, what should we be doing?

Basic idea #1: support *Interaction*

Supporting Interaction

- ▶ Among users:
 - ▶ Collaboration
 - ▶ Sharing
- ▶ Between humans and data:
 - ▶ Browsing, visualizing, exploring.
 - ▶ Several modalities of search.

My take: Some Future Projects

So, what should we be doing?

Basic idea #1: support *Interaction*

Supporting Interaction

- ▶ Among users:
 - ▶ Collaboration
 - ▶ Sharing
- ▶ Between humans and data:
 - ▶ Browsing, visualizing, exploring.
 - ▶ Several modalities of search.

Basic idea #2: take *mechanisms* (like queries, transaction, triggers) to the conceptual/semantic level.

My take: Some Future Projects

So, what should we be doing?

Basic idea #1: support *Interaction*

Supporting Interaction

- ▶ Among users:
 - ▶ Collaboration
 - ▶ Sharing
- ▶ Between humans and data:
 - ▶ Browsing, visualizing, exploring.
 - ▶ Several modalities of search.

Basic idea #2: take *mechanisms* (like queries, transaction, triggers) to the conceptual/semantic level.

- ▶ Query Languages.
- ▶ Workflows.
- ▶ Smart publish/subscribe.

A New Collar on an Old Dog?

Isn't this Knowledge Management again? Didn't KM attempt this (and fail)?

A New Collar on an Old Dog?

Isn't this Knowledge Management again? Didn't KM attempt this (and fail)?

KM did attempt something like this, but not quite.

A New Collar on an Old Dog?

Isn't this Knowledge Management again? Didn't KM attempt this (and fail)?

KM did attempt something like this, but not quite.

KM did fail because

- ▶ It treated tacit knowledge as something that could be made explicit and stored. For that to work, context has to be saved too.

A New Collar on an Old Dog?

Isn't this Knowledge Management again? Didn't KM attempt this (and fail)?

KM did attempt something like this, but not quite.

KM did fail because

- ▶ It treated tacit knowledge as something that could be made explicit and stored. For that to work, context has to be saved too.
- ▶ Workflows and hierarchies impose certain ways of working/thinking on people.

A New Collar on an Old Dog?

Isn't this Knowledge Management again? Didn't KM attempt this (and fail)?

KM did attempt something like this, but not quite.

KM did fail because

- ▶ It treated tacit knowledge as something that could be made explicit and stored. For that to work, context has to be saved too.
- ▶ Workflows and hierarchies impose certain ways of working/thinking on people.
- ▶ Not all knowledge can be neatly organized and classified (see John Seely Brown, or Elaine Svenonius).

A New Collar on an Old Dog?

Isn't this Knowledge Management again? Didn't KM attempt this (and fail)?

KM did attempt something like this, but not quite.

KM did fail because

- ▶ It treated tacit knowledge as something that could be made explicit and stored. For that to work, context has to be saved too.
- ▶ Workflows and hierarchies impose certain ways of working/thinking on people.
- ▶ Not all knowledge can be neatly organized and classified (see John Seely Brown, or Elaine Svenonius).

So, in learning lessons from the past, let us not make the same mistakes again!

A New Collar on an Old Dog?

Isn't this Knowledge Management again? Didn't KM attempt this (and fail)?

KM did attempt something like this, but not quite.

KM did fail because

- ▶ It treated tacit knowledge as something that could be made explicit and stored. For that to work, context has to be saved too.
- ▶ Workflows and hierarchies impose certain ways of working/thinking on people.
- ▶ Not all knowledge can be neatly organized and classified (see John Seely Brown, or Elaine Svenonius).

So, in learning lessons from the past, let us not make the same mistakes again!

More than a new collar on old dog, it's trying to teach new tricks to an old dog!

Future Projects 1

- ▶ Database as a shared, social resource: users can add tags, extend the database schema.

Future Projects 1

- ▶ Database as a shared, social resource: users can add tags, extend the database schema.
- ▶ In fact, maybe one schema does not fit all: multiple views of the same data should be supported. *The schema can be shaped by the users.*

Future Projects 1

- ▶ Database as a shared, social resource: users can add tags, extend the database schema.
- ▶ In fact, maybe one schema does not fit all: multiple views of the same data should be supported. *The schema can be shaped by the users.*
- ▶ What can we say about users? With user, communities (of interest, of expertise) are created. We can analyze them, use them for answers (like recommendation).

Future Projects 1

- ▶ Database as a shared, social resource: users can add tags, extend the database schema.
- ▶ In fact, maybe one schema does not fit all: multiple views of the same data should be supported. *The schema can be shaped by the users.*
- ▶ What can we say about users? With user, communities (of interest, of expertise) are created. We can analyze them, use them for answers (like recommendation).
- ▶ Tags can be used to further analyze data semantics.

Future Projects 2

- ▶ What do we do with the data? Manipulate it for some goal.

Future Projects 2

- ▶ What do we do with the data? Manipulate it for some goal.
- ▶ Workflows in the database: not as an add-on, but an integral part of the database engine.

Future Projects 2

- ▶ What do we do with the data? Manipulate it for some goal.
- ▶ Workflows in the database: not as an add-on, but an integral part of the database engine.
- ▶ Rationale: workflows where decisions are taken based on data, where many actions are information-flow actions (asking questions, gathering data) are better supported at the database level.

Future Projects 2

- ▶ What do we do with the data? Manipulate it for some goal.
- ▶ Workflows in the database: not as an add-on, but an integral part of the database engine.
- ▶ Rationale: workflows where decisions are taken based on data, where many actions are information-flow actions (asking questions, gathering data) are better supported at the database level.
- ▶ Note that this allows attacking a difficult problem: long duration transactions.

Future Projects 2

- ▶ What do we do with the data? Manipulate it for some goal.
- ▶ Workflows in the database: not as an add-on, but an integral part of the database engine.
- ▶ Rationale: workflows where decisions are taken based on data, where many actions are information-flow actions (asking questions, gathering data) are better supported at the database level.
- ▶ Note that this allows attacking a difficult problem: long duration transactions.
- ▶ At implementation level, this may mean reconsidering what we put into a log.

Future Projects 2 (Cont.)

- ▶ A similar process should be followed up to define events for monitoring and event-based processing.

Future Projects 2 (Cont.)

- ▶ A similar process should be followed up to define events for monitoring and event-based processing.
- ▶ This implies that publish/subscribe systems should be supported in the database. A smart publish/subscribe system forms an "alliance" with the client and agrees on ways to communicate back and forth, prompting a dialog.

Future Projects 2 (Cont.)

- ▶ A similar process should be followed up to define events for monitoring and event-based processing.
- ▶ This implies that publish/subscribe systems should be supported in the database. A smart publish/subscribe system forms an "alliance" with the client and agrees on ways to communicate back and forth, prompting a dialog.
- ▶ In the end, it's about taking transactions and monitoring up to the semantic level too.

Future Projects 3

- ▶ Database theory sees queries as mappings. This implies a one-shot process where the users know

Future Projects 3

- ▶ Database theory sees queries as mappings. This implies a one-shot process where the users know
 - ▶ exactly what they need,

Future Projects 3

- ▶ Database theory sees queries as mappings. This implies a one-shot process where the users know
 - ▶ exactly what they need,
 - ▶ what the database contains,

Future Projects 3

- ▶ Database theory sees queries as mappings. This implies a one-shot process where the users know
 - ▶ exactly what they need,
 - ▶ what the database contains,
 - ▶ and how to express it in a query language.

Future Projects 3

- ▶ Database theory sees queries as mappings. This implies a one-shot process where the users know
 - ▶ exactly what they need,
 - ▶ what the database contains,
 - ▶ and how to express it in a query language.
- ▶ Usually, these assumptions are (all) wrong.

Future Projects 3

- ▶ Database theory sees queries as mappings. This implies a one-shot process where the users know
 - ▶ exactly what they need,
 - ▶ what the database contains,
 - ▶ and how to express it in a query language.
- ▶ Usually, these assumptions are (all) wrong.
- ▶ Web searching shows that satisfying an informational need is more of an iterative and interactive process.

Future Projects 3

- ▶ Database theory sees queries as mappings. This implies a one-shot process where the users know
 - ▶ exactly what they need,
 - ▶ what the database contains,
 - ▶ and how to express it in a query language.
- ▶ Usually, these assumptions are (all) wrong.
- ▶ Web searching shows that satisfying an informational need is more of an iterative and interactive process.
- ▶ Databases should support a much wider view of query that incorporates *interactions with DB*, beyond querying: browsing (of data, which means showing related data, plus metadata), dialog support.

A Quick Note

Note that theory is still needed. In particular, a good theory of *information flow*.

A Quick Note

Note that theory is still needed. In particular, a good theory of *information flow*.

Less restrictive assumptions urgently needed.

A Quick Note

Note that theory is still needed. In particular, a good theory of *information flow*.

Less restrictive assumptions urgently needed.

There is nothing as useful as a good theory

Applications

If we do this, applications will just naturally fall: e-science, publishing,...

Applications

If we do this, applications will just naturally fall: e-science, publishing,...

My particular favorite: Intelligence.

- ▶ Data is partial.
- ▶ Data is unreliable.
- ▶ Data needs to be integrated (can be contradictory!).
- ▶ Data is ambiguous/can be seen in different ways.

The ultimate testbed for theories/systems!

Some (Preliminary) Conclusions

Change is indeed in the air.

Some (Preliminary) Conclusions

Change is indeed in the air.

While not forgetting/throwing away what we've learned, we clearly need to move beyond current boundaries.

Some (Preliminary) Conclusions

Change is indeed in the air.

While not forgetting/throwing away what we've learned, we clearly need to move beyond current boundaries.

New applications call for a new perspective on data: data is shared, not owned by DB -so others have a right to organize it differently.

Some (Preliminary) Conclusions

Change is indeed in the air.

While not forgetting/throwing away what we've learned, we clearly need to move beyond current boundaries.

New applications call for a new perspective on data: data is shared, not owned by DB -so others have a right to organize it differently.

New applications will only succeed if supported from this new mentality!

Thank you!

Any questions, comments, criticism, please send to

antoniobadia@gmail.com

I love to get feedback!!