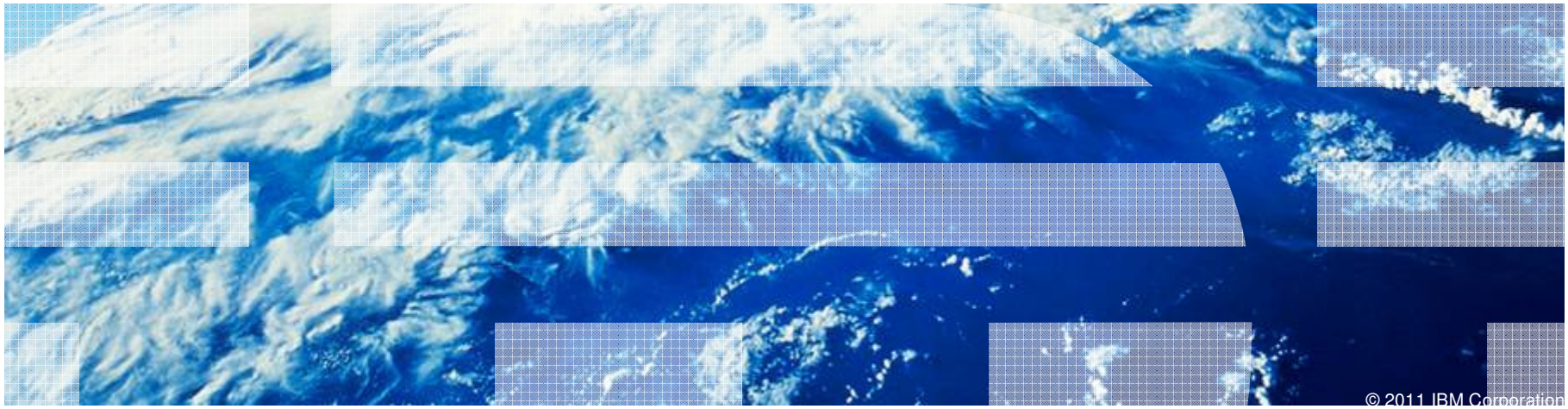
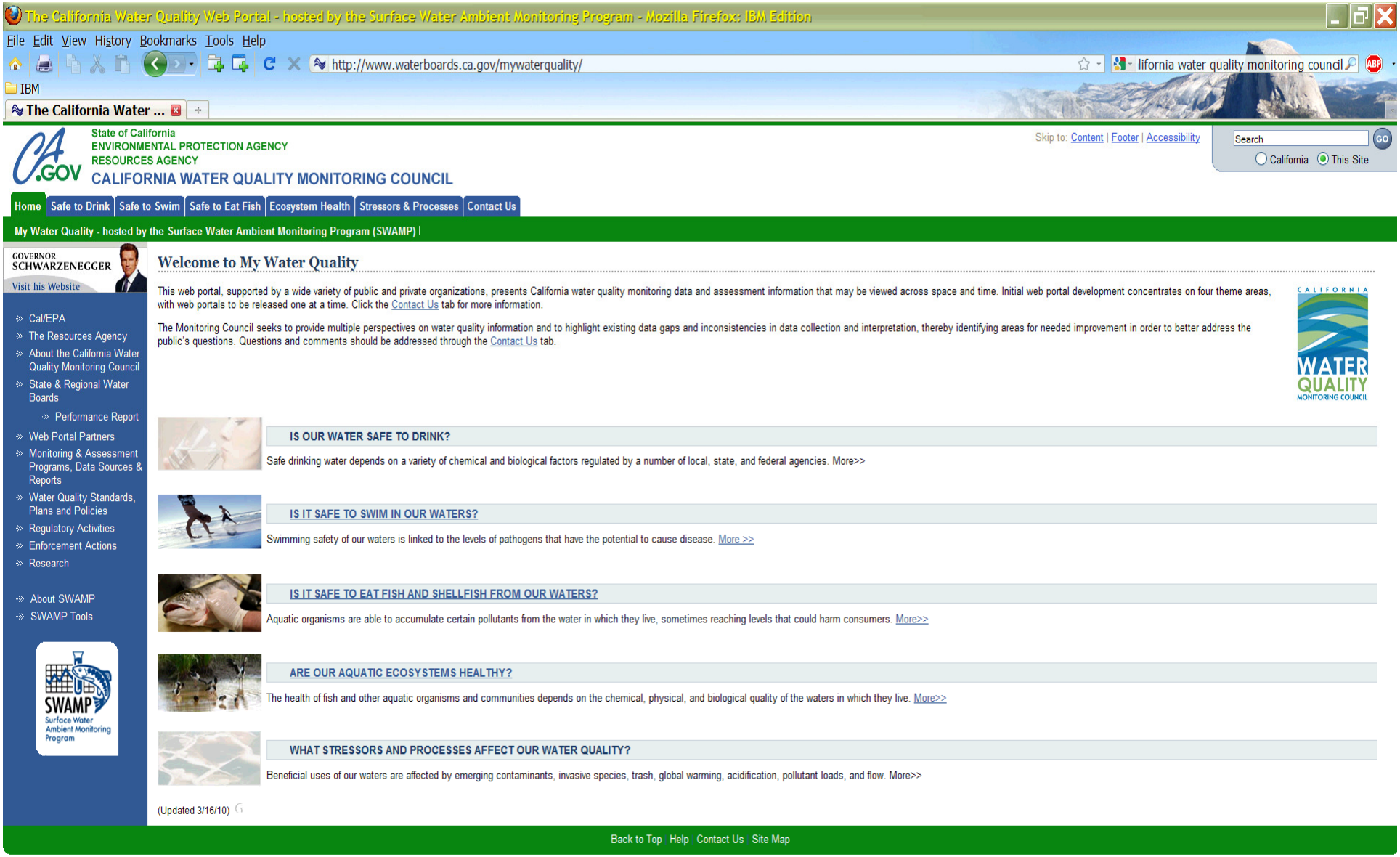
California Water Quality Monitoring Council – Data Management Workgroup:

Data Integration Approaches and Technologies

Peter Williams, CTO, "Big Green Innovations", IBM
Tuesday, November 22nd 2011



Begin with what we know: “My Water Quality” portal



Why “My Water Quality” gets it right…

- Uses **questions** as the expression of business need for data:
  - Captures that need from the user’s perspective
  - Can be broken down into subsidiary questions
- Uses questions to identify data that matters and prioritize data sets to work with:
  - Can be made to use results of previous questions - builds over time to a wider level of integration
  - If you can’t frame the question, you don’t know the need (or the value)
- Uses **questions** as “integration points” for multiple data sources:
  - Drives data architecture
  - Forces de facto standards around semantics and scale
  - (Can incorporate other standards)
  - Data aggregated for each question should then support additional related, but less structured queries too
  - Provides a focus to collaborate around
  - Allows value to be derived faster
- Avoids GADWITS - the “Great Amorphous Data Warehouse In The Sky”!!
- Enables “purpose-driven” data federation

Example: “Is it safe to swim in our waters?”

Top level question (exists)

Level 1 decomposition of questions (exists)

Examples of data required to answer questions

Other ad hoc uses of the same data

Additional structured question supported

Additional required data (to be integrated)

IS IT SAFE TO SWIM IN OUR WATERS?

Swimming safety of our waters is linked to the levels of pathogens that have the potential to cause disease. More >>

QUESTIONS ANSWERED

- Can I swim at my beach, lake, or stream?
- How clean was my beach, lake, or stream during the past week or month?
- What are the long-term trends at my beach, lake, or stream?
- Which beaches, lakes, and streams are currently closed by county health agencies?
- Which beaches, lakes, and streams are listed by the State as impaired?
- Are the problems getting better?

Geo-tags for every beach, lake or stream

Time-series data on [list of] pathogens – lab tests, sensed data, referenced to geo-tag

Time series data on [list of] contaminants - lab tests, sensed data, referenced to geo-tag

Data on County Health Agency or State actions, reasons

Public reports of illness, etc

(Locating and classifying water resources)

Support for other ad hoc water quality-related queries

Support for Epidemiological research

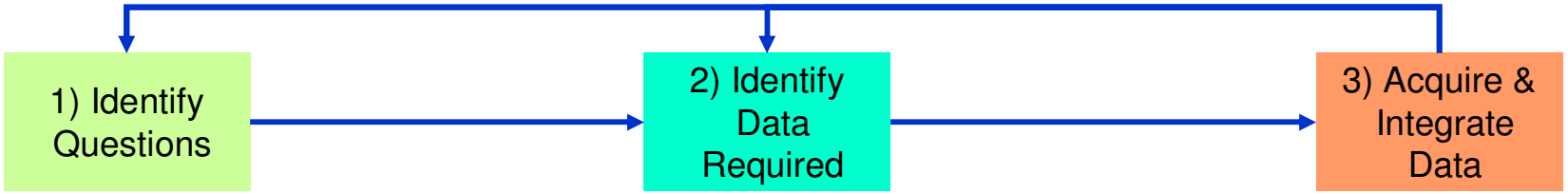
(Hypothetical example): What is the impact of known pig-slurry run-off sources on river and lake bathing water quality (or fresh water quality generally)?

Geo-tags, maybe photos for slurry source locations

Data on slurry run-off events (dates, amounts etc)

(etc)

Structured questions as the core of a data integration process?



- Identify key questions to be answered
  - Identify decomposition questions
- Test with users
  - Identify value of answers to questions (who benefits; how; $$$ if available)
- Prioritize key questions and decompositions
- Add questions to catalog

- Identify data required to answer key and decomposition questions
- Identify consuming applications, models, portals & needs arising
- Identify available data-sets & gaps
  - Search datasets already integrated
  - Identify proxy data if needed
- Confirm scale and semantic match with need
- Identify other possible ad-hoc uses and other questions from catalog that could be answered with this dataset
- Confirm cost and feasibility of acquiring and using each data-set
- Confirm question priorities in light of technical feasibility

- Acquire data
- Create adaptors etc necessary for semantic matches
- “Fill in” or “widen out” to address scale mismatches
- Test with consuming applications, models, portals –
- Confirm that question is “answered”/user need is met
- Create pub/sub or other interfaces to enable continuous acquisition
- Populate data warehouse/ data-set catalog as applicable (depends on degree of data federation)
- Repeat the above for additional uses

Core principles

- "Question-driven": data only integrated as needed to answer an unambiguously articulated questions – never "just because"
  - Can be adapted for historical perspective: for example, "how has water cleanliness in lakes and rivers changed since 1950?"
  - Can be adapted for causation: for example, "what has been the correlation between levels of contaminant x and high rainfall events?"
- Data warehouse, if used, is populated as data and datasets are assembled to answer questions – not before
- Each dataset is catalogued by ref to content, format, scale etc and also questions it can be used to answer
- Value of any given data set and integration activity =

$$\frac{\text{(Formal questions + ad hoc uses supported)}}{\text{(Technical risk x cost of integration)}}.$$

- Assembling data to answer each question builds on data sets integrated – cumulative coverage of the field, driven by value of integration activity

- So what technological trends and developments do we need to take into account as we frame integration goals?

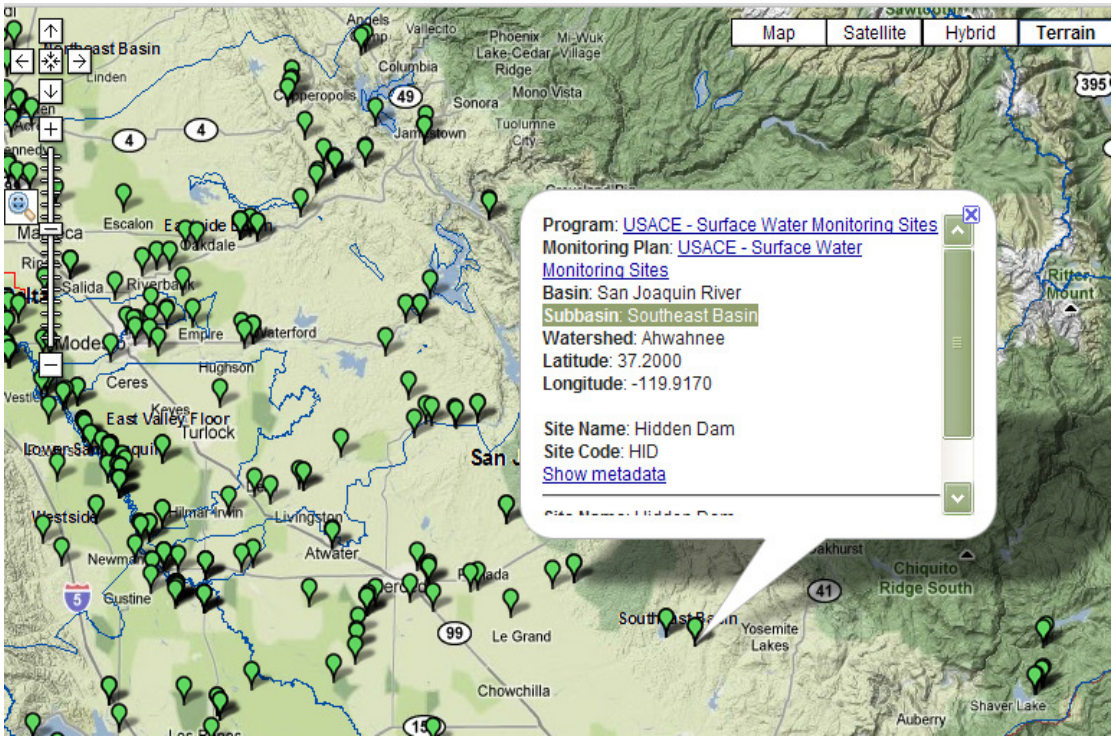
Data – in the past, often the elephant in the room…?

1. No data (rarer than one might think!)
2. Data is in the wrong scale (spatial or temporal) for the decision – too slow, late or infrequent, too scattered.
3. Data is fragmented between different stakeholders: different formats, scales, frequencies, standards; re-capture many times.
4. Too much data to use or analyze.
5. Incompatible or incomplete models mean that data is not leveraged, information is incomplete, or that solutions are partial or ineffective.
6. Poor visualization of information impedes effective decision-making: “so what’s this telling us?” syndrome.

7. Lack of awareness that 1-6 may be a problem. Possibly the most insidious problem of all!

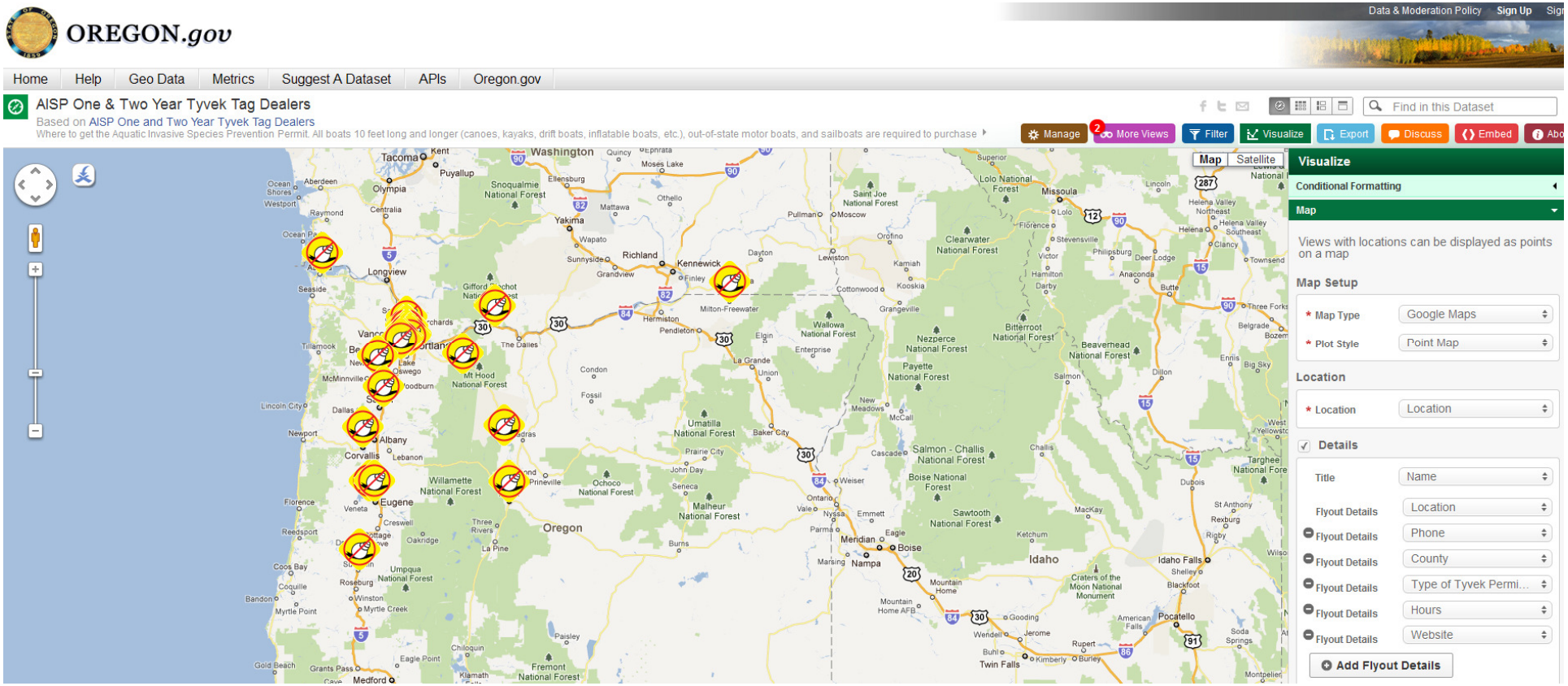
But data is becoming more available – even democratic

- Data is getting cheaper (Moore's law, more interest => more data and data sources)
- More and more data in the public domain – the open data movement.
- More and more ability to "mash it up" to create applications with it – often in the public domain.
- Socrata offers one of the most advance sets of open data/publishing tools around – starting to attract major attention.
  - San Francisco, Washington, Oregon (see over), Oklahoma, Chicago and others
- Especially useful in publishing data for consumption by other apps – eg smart phone apps



http://www.centralvalleymonitoring.org/

Socrata example: Oregon's open data portal



- Example shows location of invasive species tag dealers.
- Note user configurability – panel on the right.

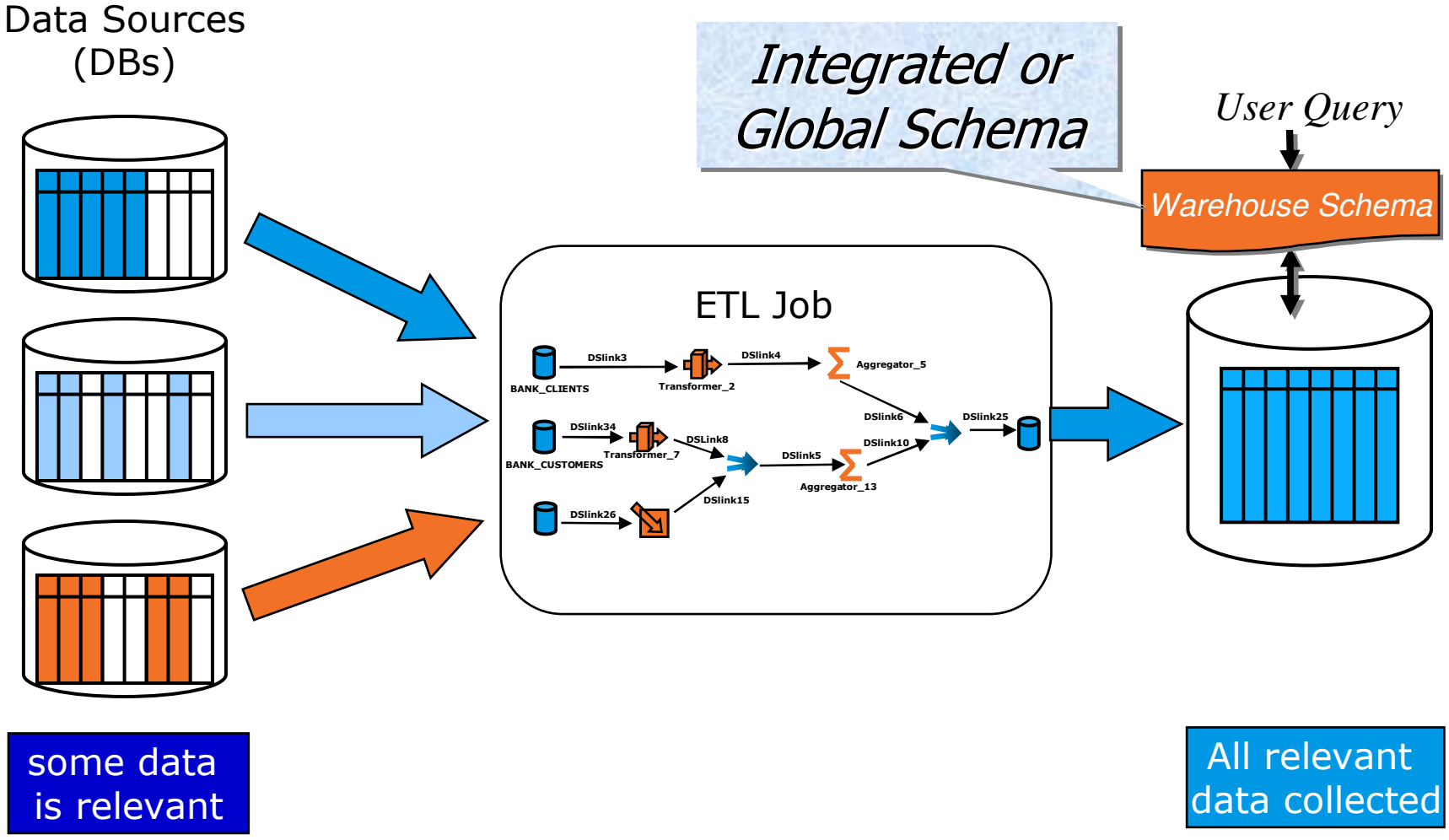
Socrata | The Open Data Company™

Data integration – a choice of approaches

- Materialization: create integrated data set via extract, transform, load (ETL) from multiple data sources or replication
- Enterprise Application Integration (EAI): writing a special application linked to workflow
- Federation: virtual representation of integrated data set, materializing only what is needed, when needed
- Indexing: single index, enables data or docs to be fetched dynamically at user request

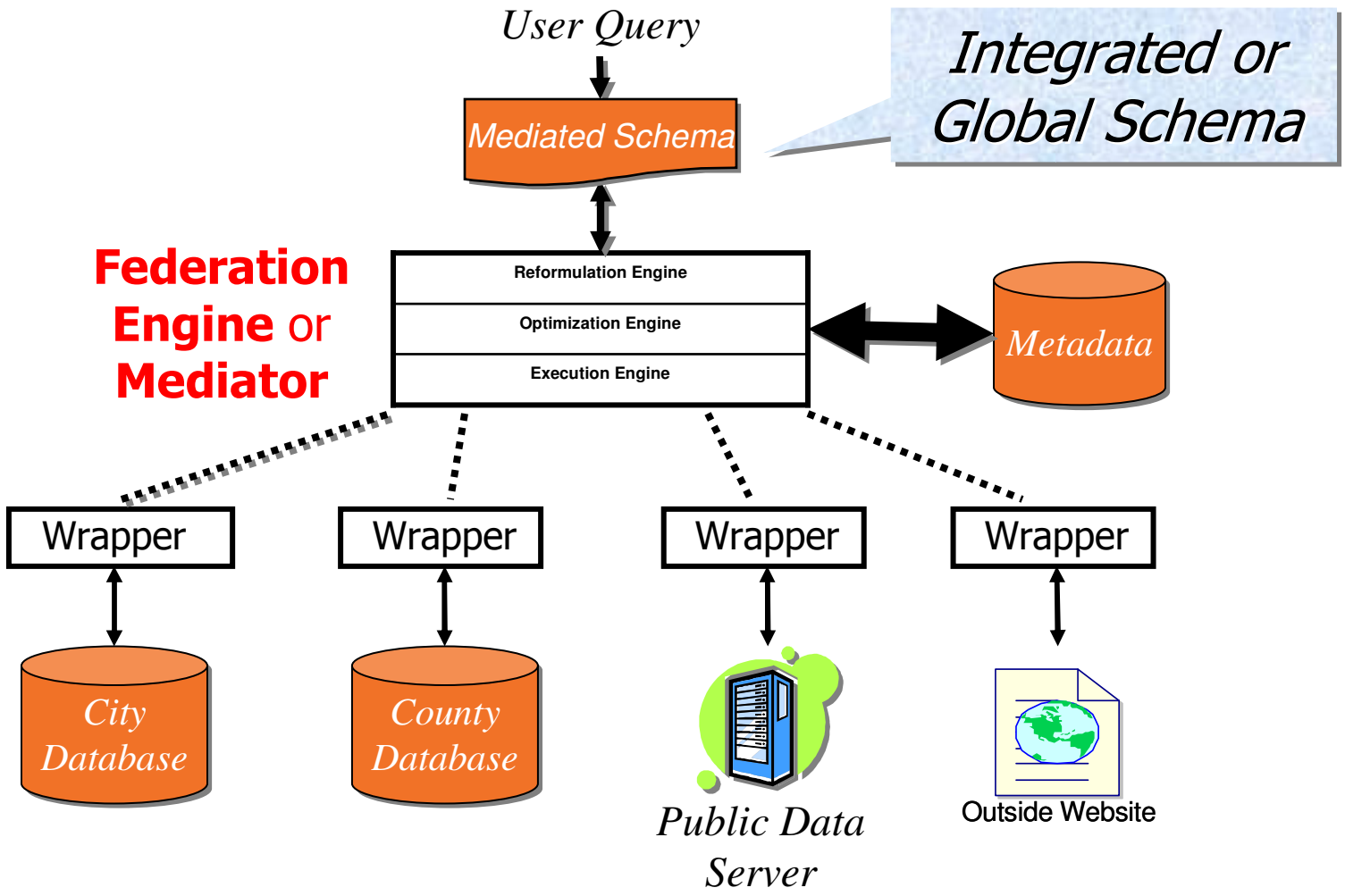
Data warehousing is the oldest form of information integration

Materialization, Data Exchange, ETL (Extract, Transform, Load), Integration



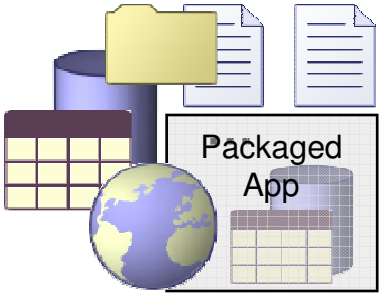
Data federation and indexing is the other extreme

Mediation, Virtual Integration, Data Integration, Lazy Integration



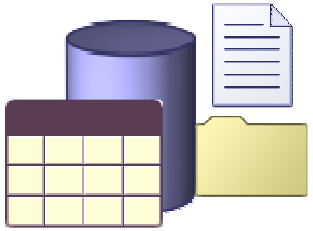
Pros and cons of integration approaches

Federation Services

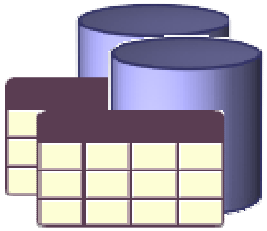


...

Consolidation (eg ETL) Services



...



Federation or index	Dimension	Consolidation (eg ETL)*
Latest, most current data – some network delay can be tolerated	**Key need**	Response time critical - rapid queries with no network delays
Small to medium result sets	**Data volumes**	Medium to very large result sets
Heterogenous - any	**Data types**	Largely relational
"Best endeavors"	**Availability**	High availability
Ad hoc	**Usage**	Predictable
Publishers'/originators' systems	**Commonest failure mode**	Internal errors
Medium – tend to be faster	**Implementation complexity**	High – long complex projects
Stays with the originator	**Data control/politics**	Moves to the aggregator

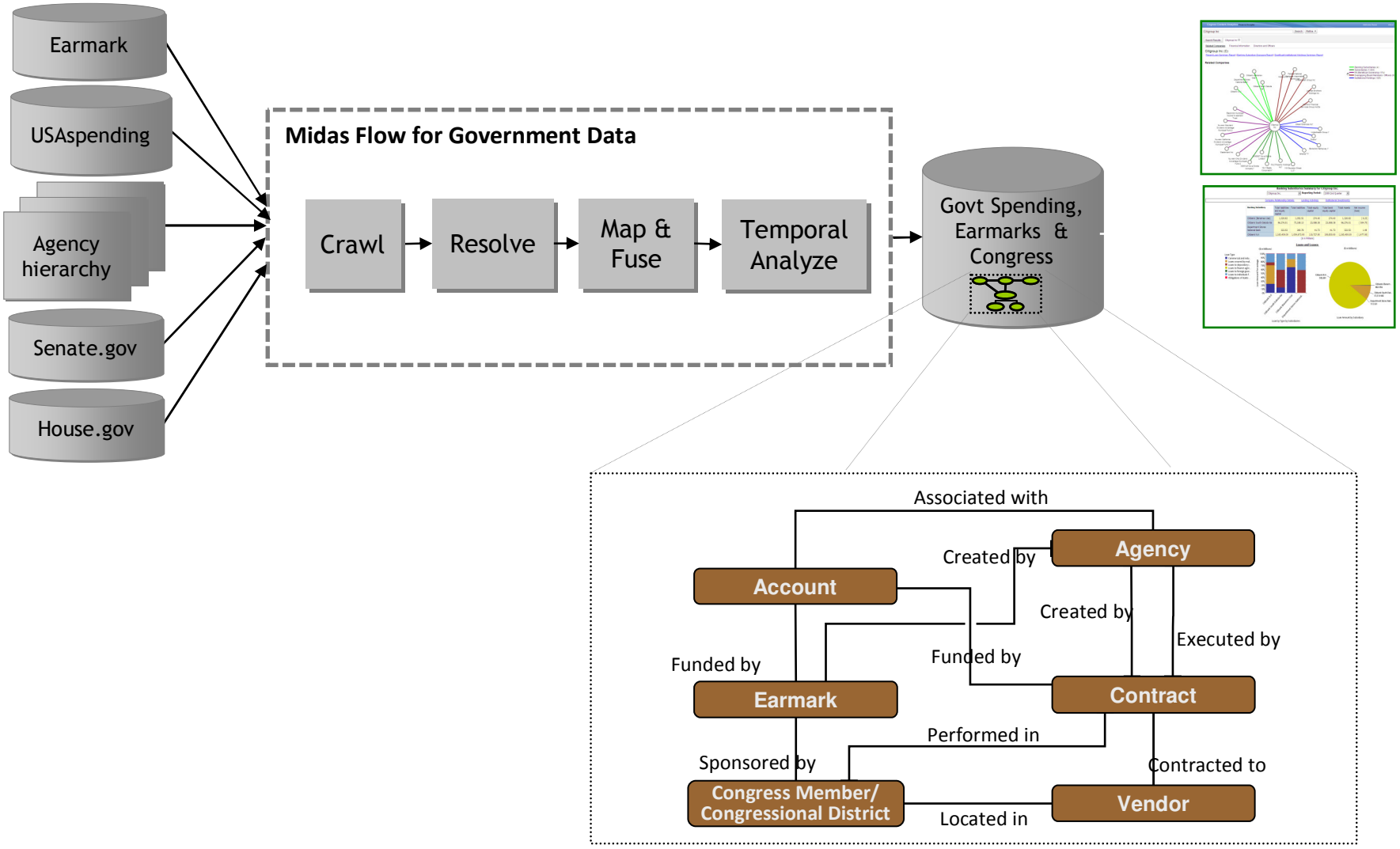
Complex integration targets - MIDAS

- MIDAS – IBM approach to enable queries over multiple, non-rationalized, data sources: integrates multiple partial or overlapping references into single complex entities.
  - May need to integrate multiple data types – text sensor data, numerical and so on
  - These data types may be both structured (tables) and unstructured (text, images, etc etc)
- Tested on work with financial services and government data, to answer queries like:
  - “How many and what value of earmarks in 2008 were solely sponsored by Republican or Democrat congress members?”
  - “Which public companies currently share one or more board member?”
- Significant open-sourced component – based on Apache Hadoop (data-intensive distributed application processing)

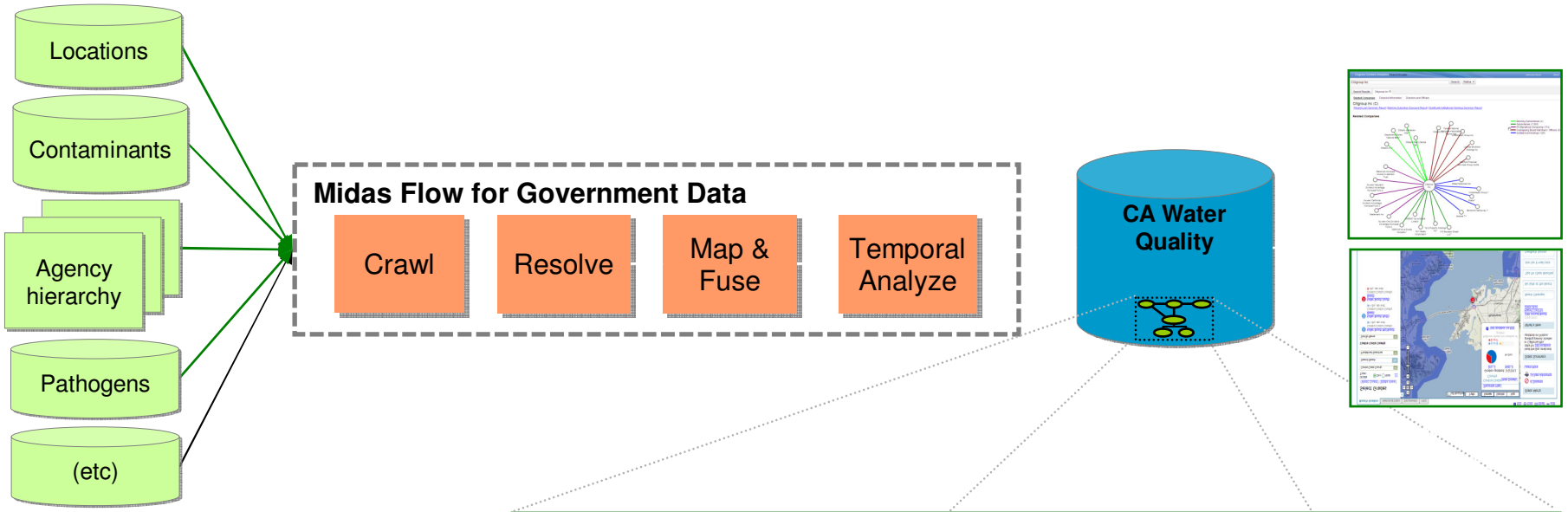
Technical challenges of complex entity creation

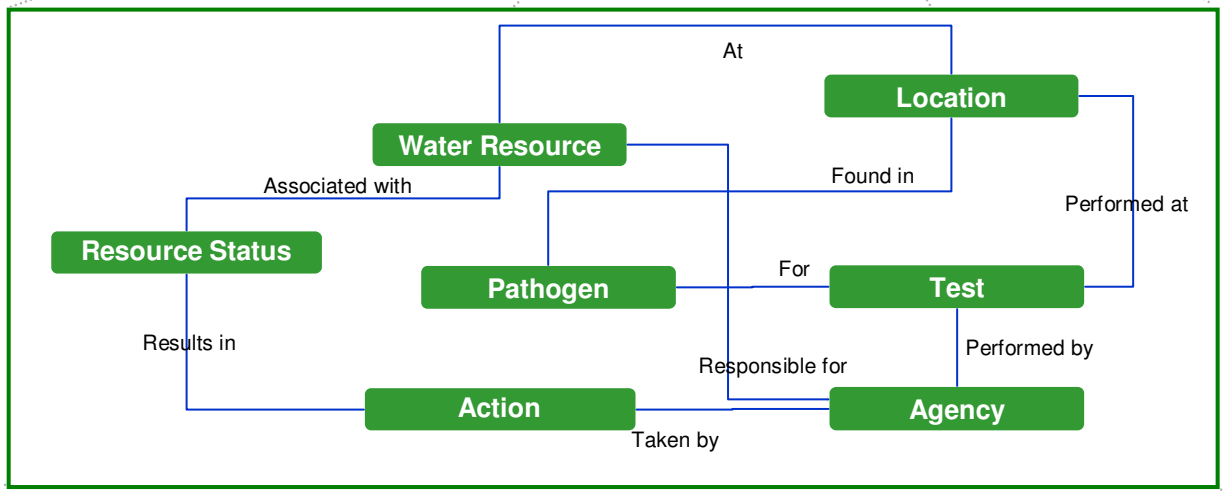
- Entity Extraction
  - Extracting structured data from unstructured texts about key entities, and relationships between key entities of the same type or different types
- Entity Resolution
  - Identifying that two instances (i.e., references) of the same entity type refer to the same real-world entity instance
  - Example: “Tahoe” vs “Lake Tahoe”, or whether “Tahoe” is a reference to the lake, the geological basin, or the region
- Data Mapping and Fusion
  - Map each extracted record (from its own format) to the target format
  - Fusion: aggregate multiple records of an entity into one (complex) object
- Temporal Analysis
  - Creation and management of timeline for relationships
- Scalable architecture
  - Complex analysis over millions of documents in a scalable manner

Example: MIDAS flow for government spending data



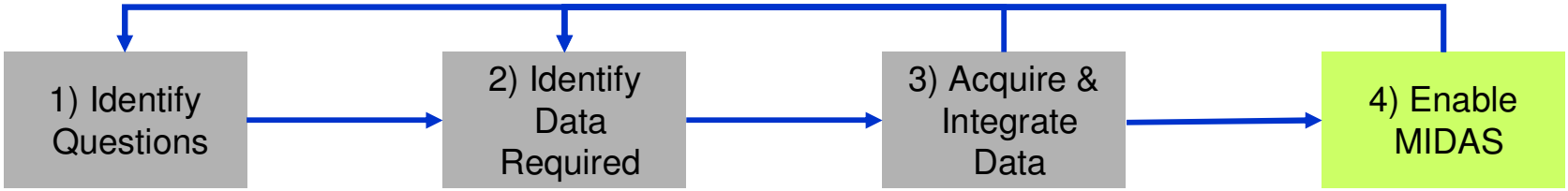
MIDAS for water management…?





Q: “How clean was my beach, lake or stream during the past week or month”?

Incorporation of MIDAS into the process…



- Identify key questions to be answered
  - Identify decomposition questions
- Test with users
  - Identify value of answers to questions (who benefits; how; $$$ if available)
- Prioritize key questions and decompositions
- Add questions to catalog

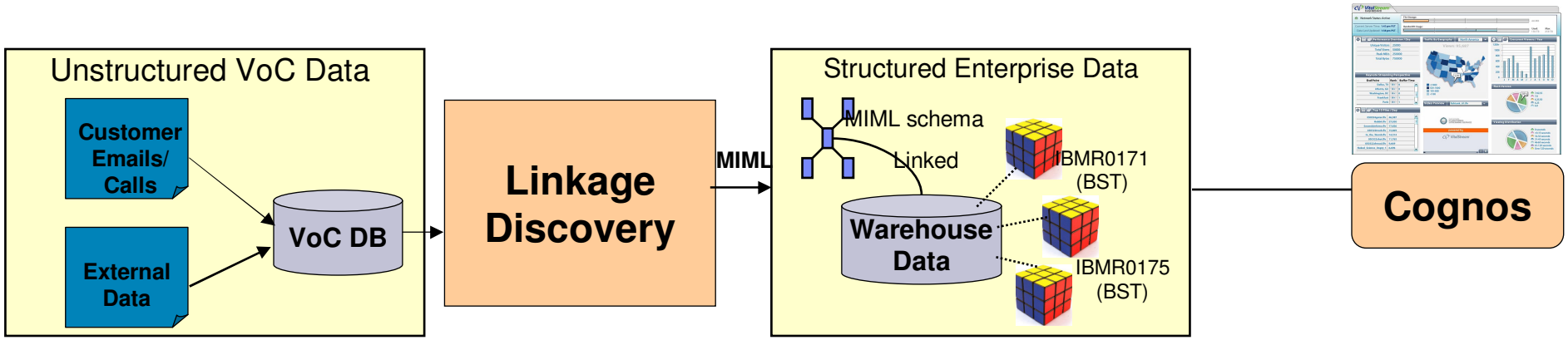
- Identify data required to answer key and decomposition questions
- Identify consuming applications, models, portals
- Identify available data-sets and gaps
  - Search datasets already integrated
  - If no direct data, identify proxies, if any
- Confirm scale and semantic match with need
- Identify other possible ad-hoc uses and other questions from catalog that could be answered with this dataset
- Confirm cost and feasibility of acquiring and using each data-set
- Confirm question priorities in light of technical feasibility

- Acquire data
- Create adaptors etc necessary for semantic matches
- "Fill in" or "widen out" as needed to address scale mis-matches
- Test with consuming applications, models, portals –
- Confirm that question is "answered"/user need is met
- Create pub/sub or other interfaces to enable continuous acquisition **(see right)**
- Populate data warehouse/ data-set catalog as applicable (depends on degree of data federation)
- Repeat the above for additional uses

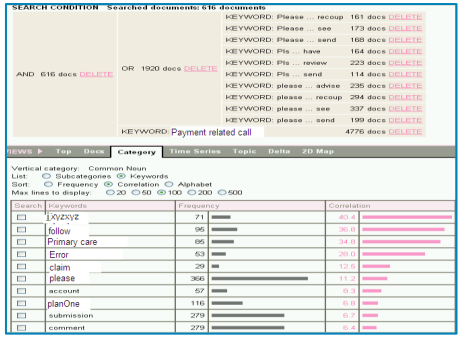
- Identify entities and sub entities
- Create nested entity relationships
- Aggregate records for each entity into complex objects and index these
  - By question supported
  - By other ad hoc use type supported
- Create search tool that enables queries
- If required, export to RDB to enable conventional BI tools to operate

Enhanced reporting using unstructured data, in Cognos

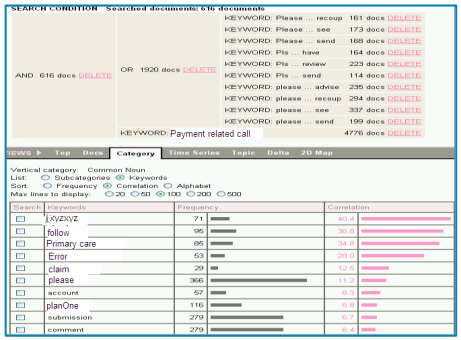
Example Reports generated from Structured Data	*Example Reports generated from Unstructured Data*
Location Profitability **- To identify the contribution to profit of geographic areas served by the financial institution.**	**Analyze complaint categories and their distribution with region and location.**
Channel Profitability **- To identify the contribution to profit of the Financial Institution's channel including branch networks, agencies, correspondents, and electronic channels**	**Distribution of complaints across service channels, user sentiments vs. service channels**
Customer Attrition Analysis **- To understand the reason and impact of customers ceasing to use the Financial Institution's products and services**	**Reporting on key complaints, opinion features (specific feature of the product/ product that customers have opinion on)**
Customer Complaint Analysis **- To understand the pattern of complaints and the effectiveness of the resolution process.**	**How are complaints distributed across customer segments? Is there a pattern to the type of problems faced by customers in a specific segment**

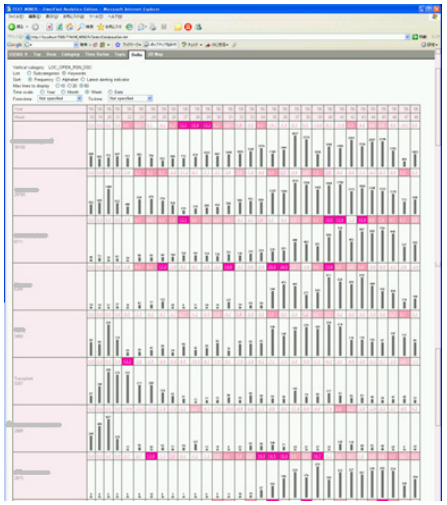


Example analysis scenario in a typical financial services company. What's the equivalent for water?

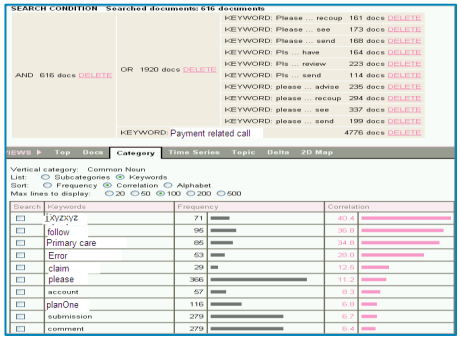


1) Find all customer contacts regarding cash advances on credit cards





3) Link and correlate information about account status against these complaints



2) Find the contacts where the customer complains about the service ("high rate", "overcharge" etc).

4) Identify higher correlation between customer complaints about cash advance rate and inactive accounts

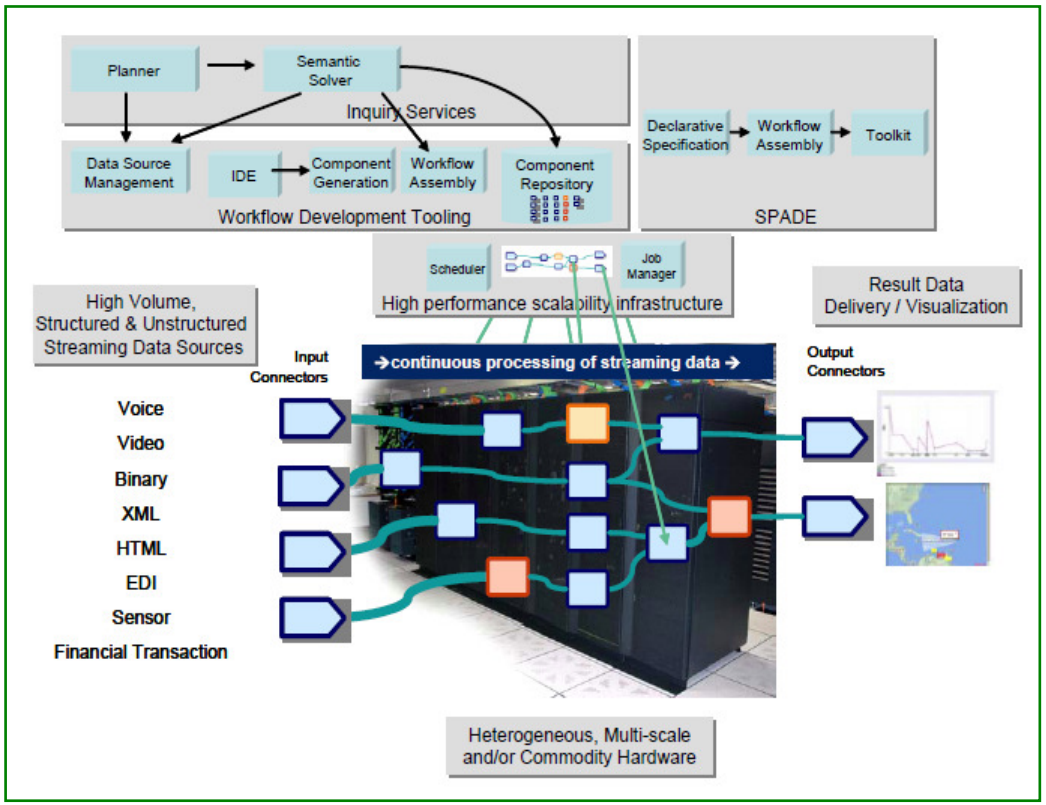
As a result of the analysis:

- Develop new rate structure and offer to premium account holders
- Resulting in reduced churn from this segment

- From integrating data to integrating models?

The role of models that consume data is changing

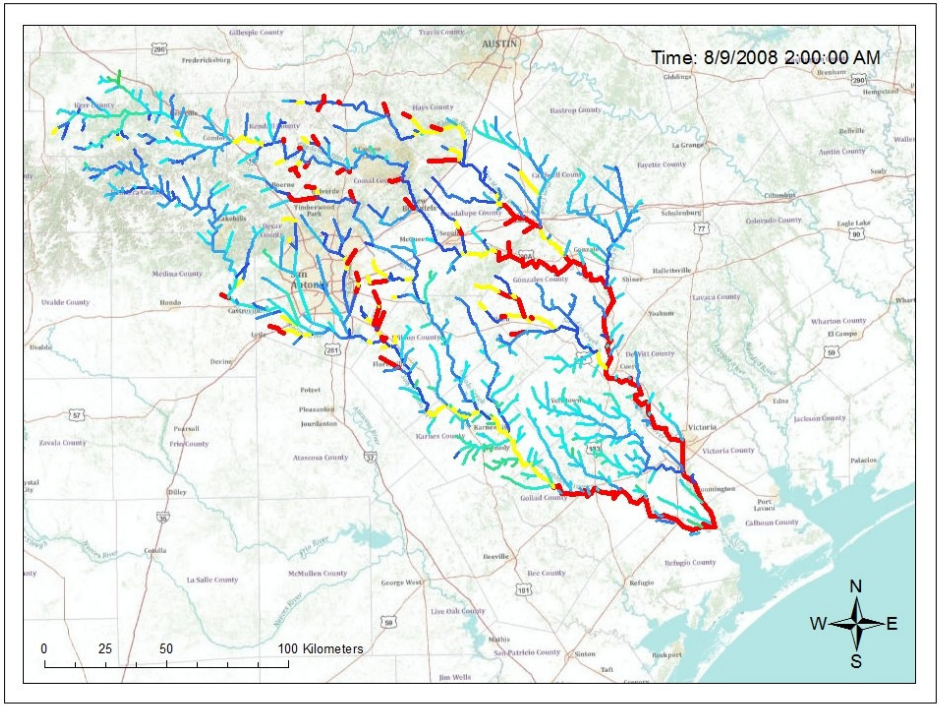
- Advent of the model as a management tool
  - Used to be an “off-line”, back room scenario generation or exploration tool
  - Now more likely to be “on-line”, part of the decision stream
- What’s driving this?
  - Moore’s law: models that took hours to run can now run in minutes (combining computing power with model tuning).
  - Extreme case – can run models in real time as part of event processing system



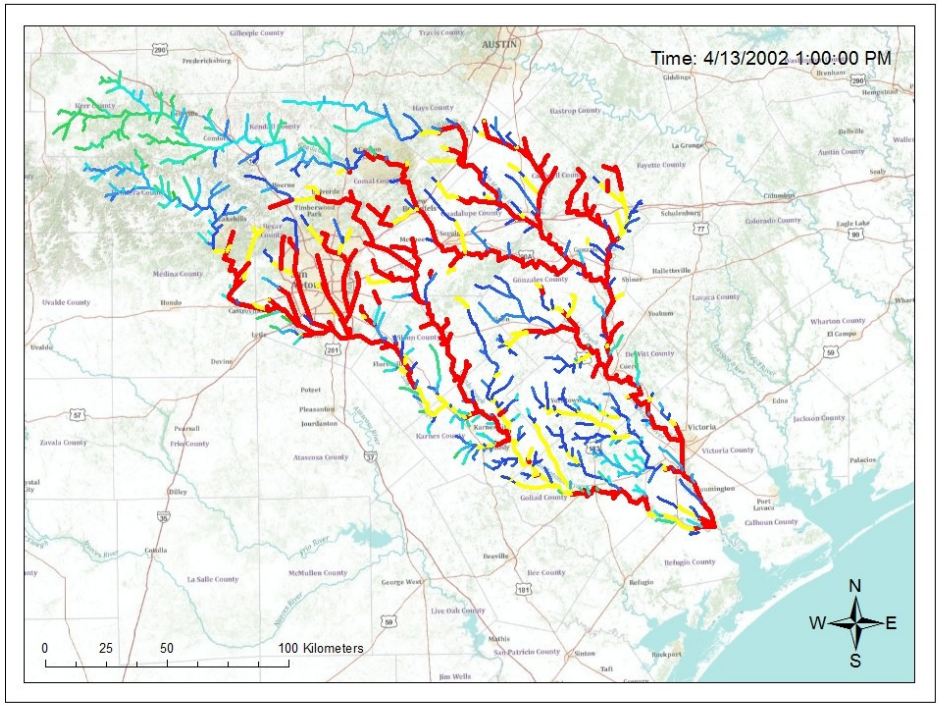
  - Those models increasingly have sufficiently granular data from sensors and other systems to produce “operational-grade” conclusions
  - Commercial modeling and visualization software increasingly focused on supporting operational decision-making. Examples: IBM/ILOG, Optimatics, GL, Bentley and others.

Model tuning example

- Guadalupe-San Antonio river basin in central Texas
- Approximately 3,500 reaches, 15,000 km total length
- Modeled by 110K segments
- Fully dynamic computation for a seven day event simulated in one hour on a one-node x86 computer.
- Adapts “netlist” techniques from semiconductor industry to allocate processing power to segments of the model highly efficiently



Snapshot of event 1



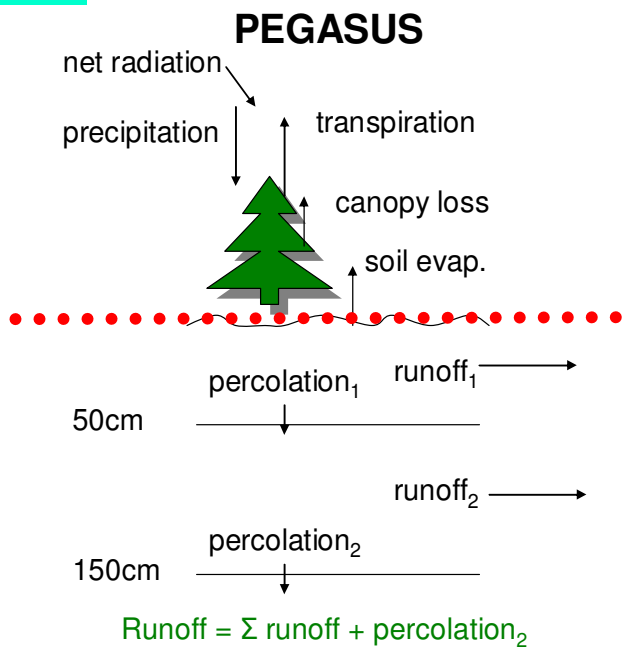
Snapshot of event 2

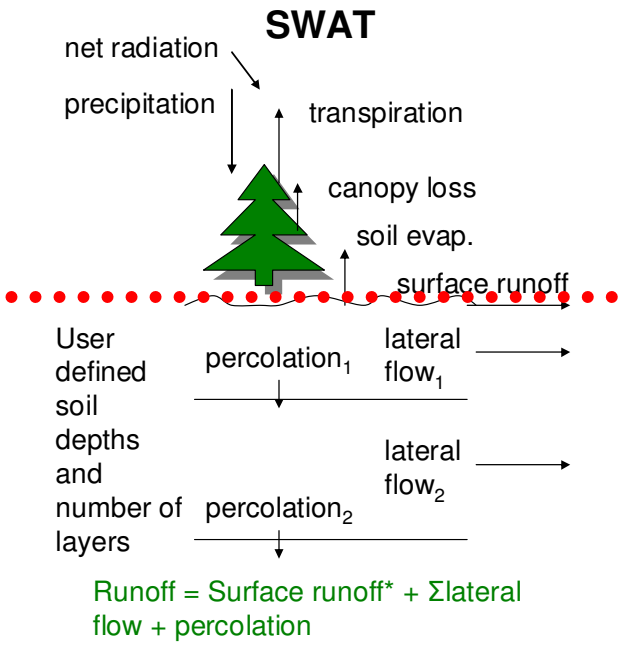
Mpeg files available

Integrative modeling frameworks*

*See final slide

- Model codes are typically monolithic: hard to integrate
- A more component based approach would allow components to be integrated into chains or compositions specific to the task at hand.
  - See right: the above-ground components of these two run-off models are identical, but the below-ground components are different
- Needed:
  - A way to break models into components, and compile and manage component sets and (larger) libraries
  - A way to allow only scientifically valid combinations of components - but include all such valid combinations
  - A consistent interface set and set of services to enable modules to interoperate
  - A concept of workflow to enable simulations using the chain of components
  - A rich semantic framework to express the above – captures context, not just checking inputs and outputs
  - Automatic co-calibration of model components
  - Ideally, ease of use for the non-scientist

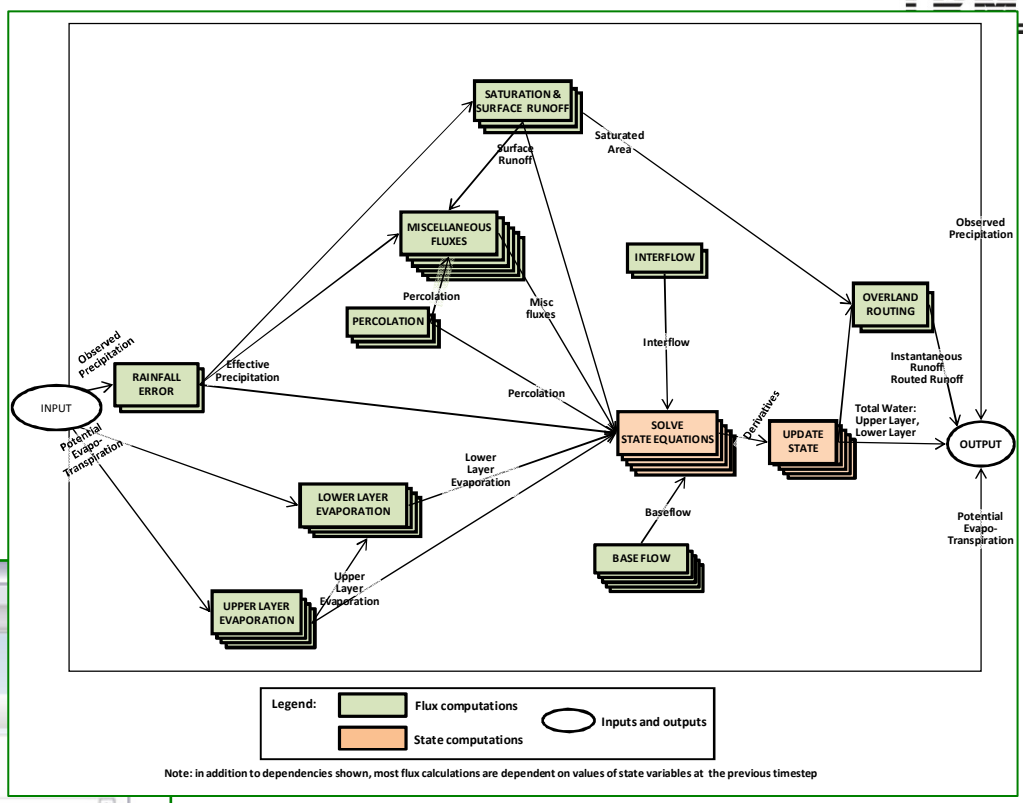


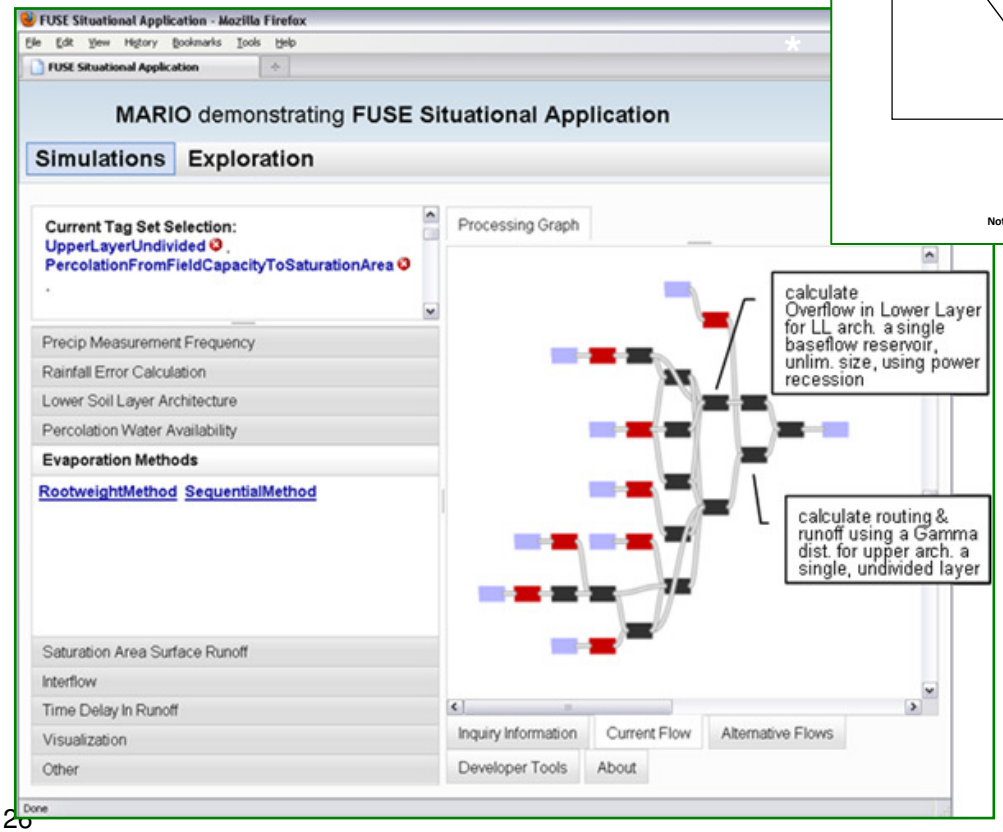


Mashup Automation with Runtime Invocation & Orchestration (MARIO)*

*See final slide

- The models on the previous page were integrated manually.
- Now demonstrated as an automated process using FUSE* framework for estimating stream discharge given observed precipitation:





  - 250,000 theoretical module combinations, of which just 316 are scientifically valid (for example, tying flux equations to appropriate soil layer architectures)
- MARIO enables selection of processing goals via a set of tags, then selects and ranks possible chains of model components that satisfy those goals according to a set of rules

- Visualization – the afterthought

Visualization

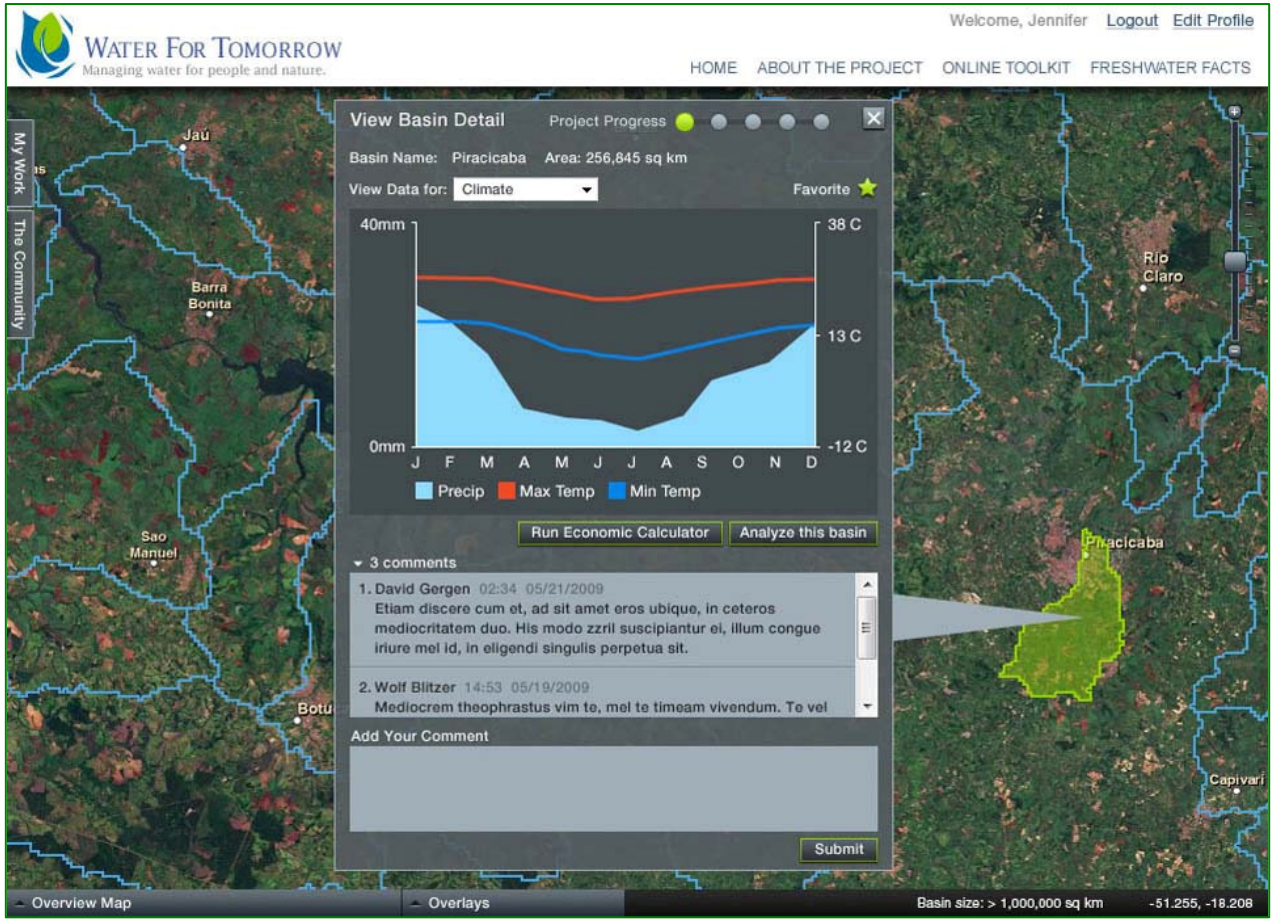
- **"A method of computing in which the enormous bandwidth and processing power of the human visual (eye-brain) system becomes an integral part of extracting knowledge from complex data",*** for example in:
  - Feature, trend or event identification
  - Comparison or fusion of data from multiple sources (visual fusion)
  - Decision support
  - Communication
- Visualization is too often an afterthought. It should be considered from the start when designing any application of environmental informatics – as distinct from "ok, here's the data, now how do we display it?"

*See final slide

Visualization as more than an afterthought: Water for Tomorrow*

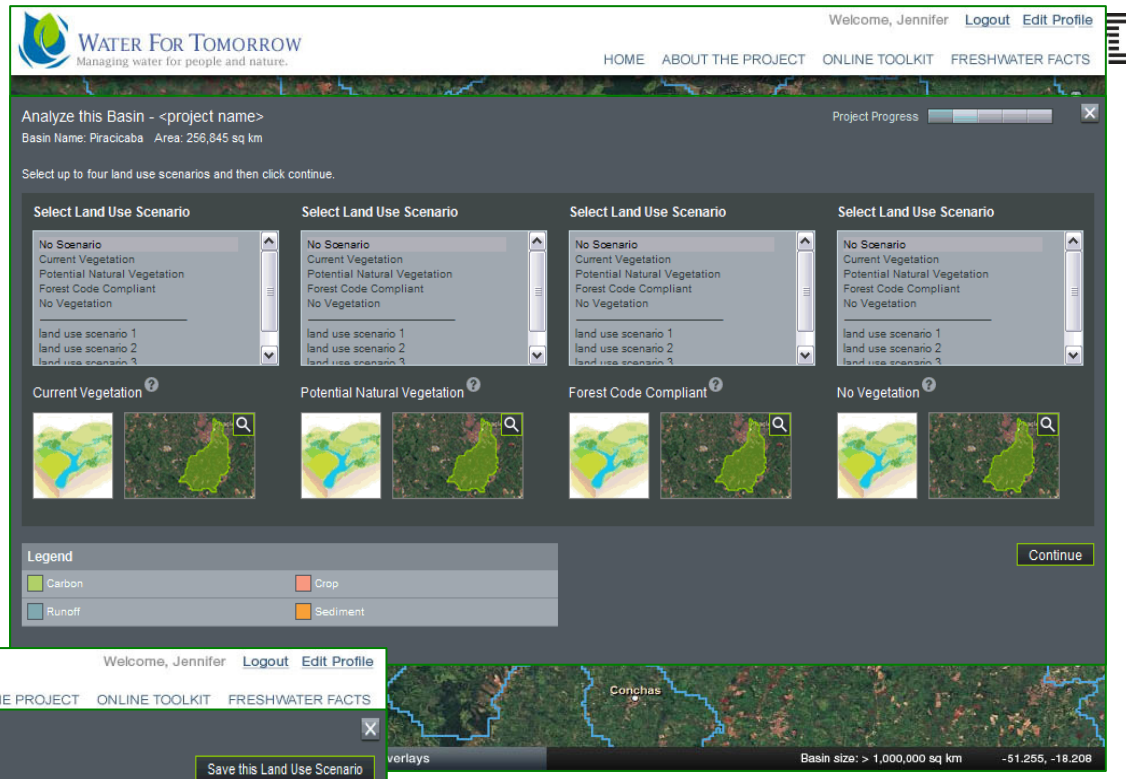
*See final slide

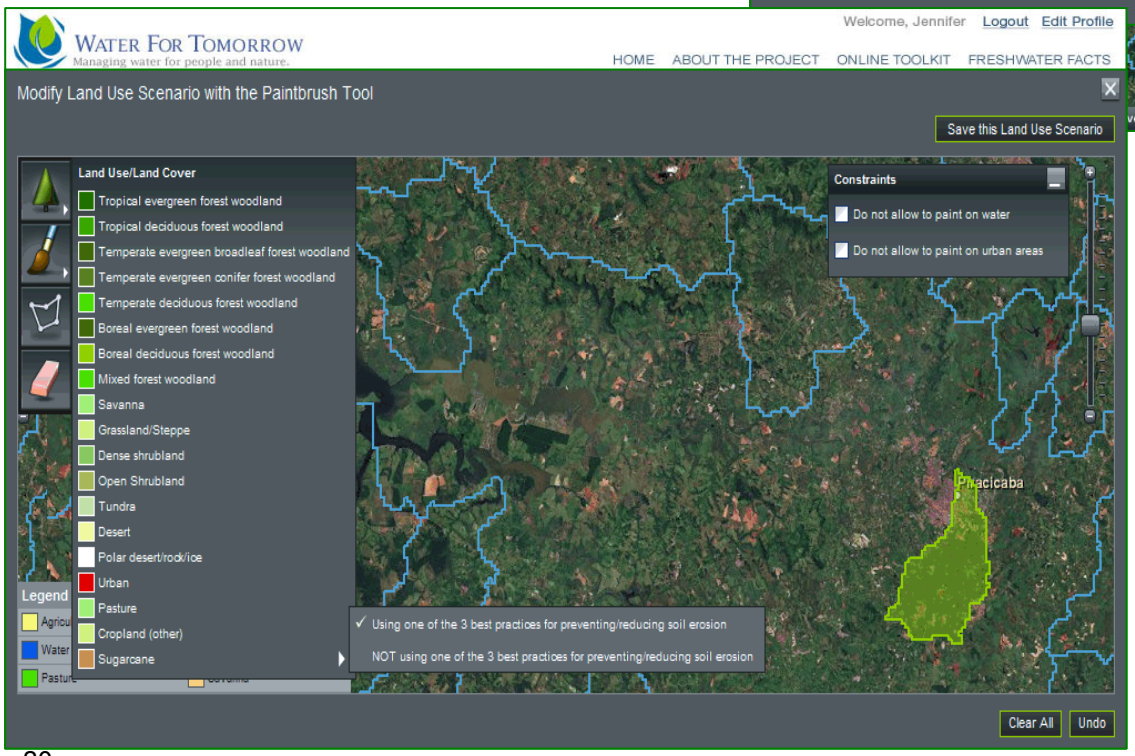
- Helps planners and scientists analyze river basins and visualize the effects of different management scenarios on basin health
  - In so doing, initiates collaboration to develop sustainable water management policies
- Marries state-of-the-art technology with collaboration tools to support conservation of freshwater ecosystems
  - Combines rich graphics and dynamic mapping capabilities
  - Based on the Pegasus and SWAT hydrological modules reviewed earlier
- Basins selected by delineating areas on the map, or uploading a basin file
- This basin is part of the Piracicaba, Capivari and Jundiaí rivers (PCJ) pilot site in Brazil



Scenario Selection and modification

- A land use scenario is a set of land use practices. Scenarios are used by the WFT hydrology model to compute crop production, water quality, water balance and other variables.
- Pre-canned scenarios are provided in the tool, or custom scenarios can be created or imported.

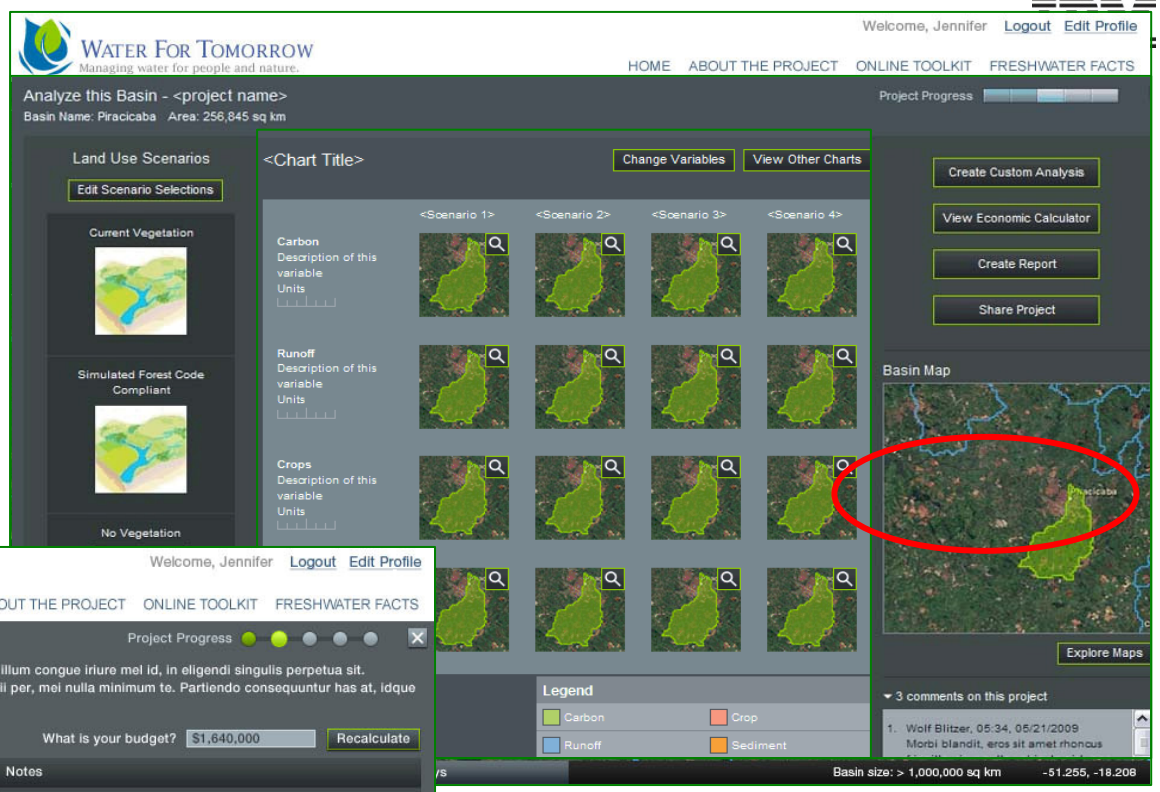


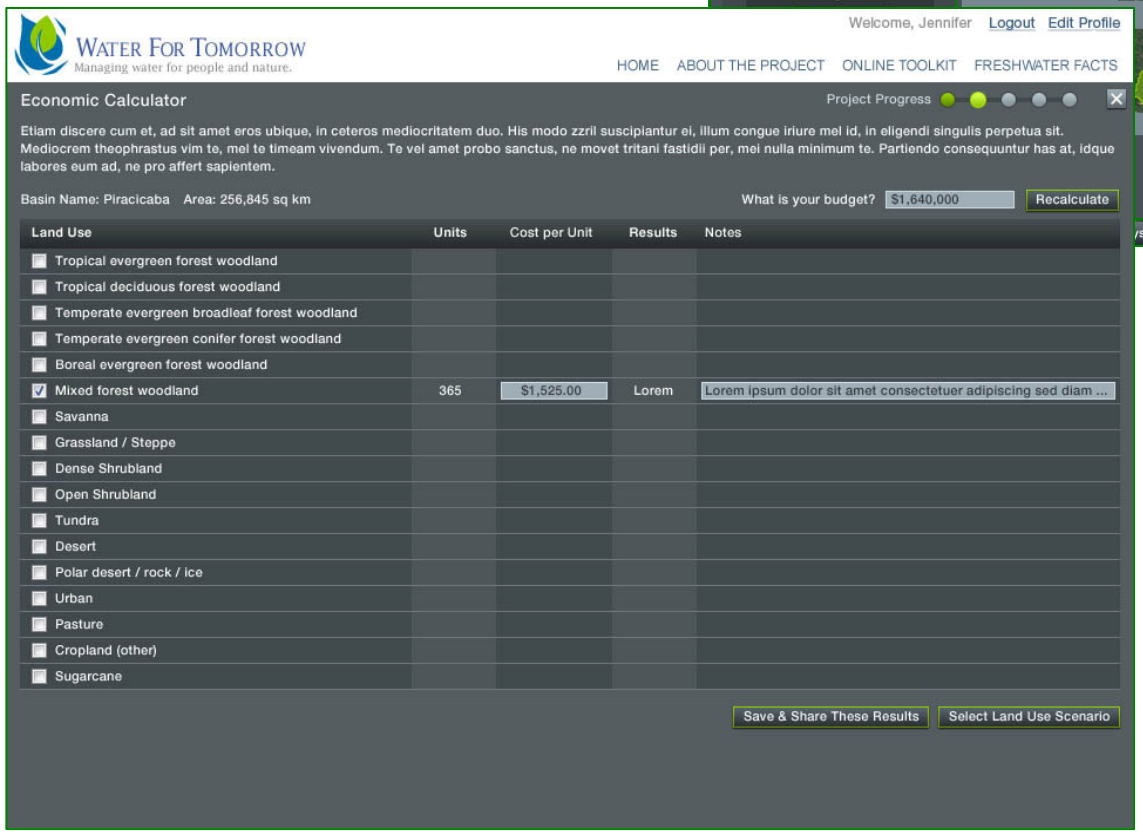


- Direct link between data and visualization: scenarios modified by using simple paintbrush tools
- Multiple visualization methods – charts, graphs, maps etc.
- Aim is to provide a “workbench” - analyses can be edited, saved, commented on and shared with the user community or pre-defined work groups

Interactive Visualizations

- Contains an ROI calculator that enables the economic value/impact of land use scenarios, and cost benefit of current policies vs ecosystem protection vs remediation to be calculated and weighed into the discussion.

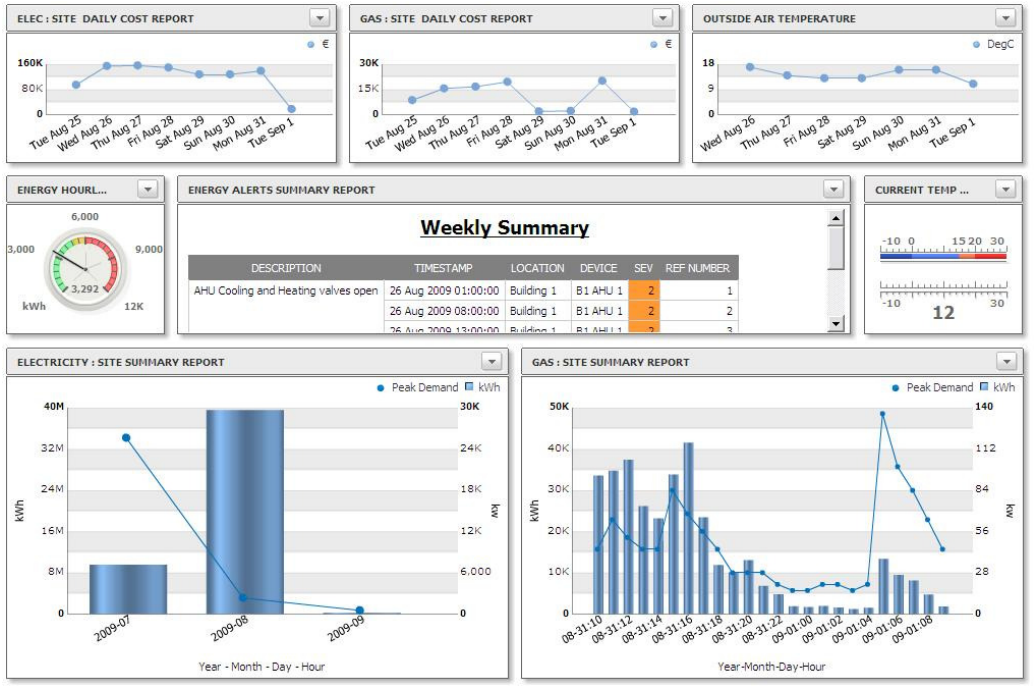




- Visualization as the integration of data, models and decision processes
- Visualization drives the definition of the system, function offered and model integration – by implication, also used to structure underlying data

Scorecards and dashboards

- Commonly used metaphors in business reporting, and for infrastructures. Excellent for depicting environmental performance of a business.
- But, they are under-investigated uses for environmental informatics:
  - They rely on key performance indicators: these could be defined for ecosystems, say, as easily as for a business.
  - Factors affecting performance against those indicators can be derived by systems dynamics modeling, or other methods.



- Can include considerable logic, up to and including models
  - Can operate in real time, with streaming data if needed – for process control, say
- Dashboards and scorecards are a beguilingly easy concept to grasp – in fact, unless all data is coming from an ERP system, they can require VERY significant integration work with systems generating core data
  - Direct linkages with sensors would in many cases be easier!!

References

- Slide 25:
  - "Toward an integrative software infrastructure for water management in the Smarter Planet" B Eckman, M Feblowitz, A Mayer, and A Riabov, IBM Systems Journal 2009.
  - Deconstruction of Pegasus and SWAT produced by Eckman and others in course of IBM/Nature Conservancy "Water For Tomorrow" project (see below)

- Slide 26:
  - Eckman, et al. The authors demonstrated the application of MARIO to automating the chaining of modules from the hydrological models incorporated in the Framework for Understanding Structural Errors (FUSE) (Clark et al, referenced in Eckman et al). In the graphic, stacks of green boxes are options for flux calculations; stacks of brown boxes are options for state equations.
  - Other model integration frameworks are starting to appear in hydrology and water management:
    - OpenMI: standard software component interface that allows water models to exchange data on a time-step-controlled basis – allows modeling of process interactions, irrespective of temporal and spatial resolutions of component models. See http://www.openmi.org/reloaded/about/what-is-openmi.php
    - The Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) has launched the a Community Hydrologic Modeling Platform (CHyMP). See http://www.cuahsi.org/chymp-20090331.html

- Slide 28:
  - "Visual Data Fusion for Applications of High Resolution Numerical Weather Prediction", L Treinish, IBM, date unknown. Available from: http://www.research.ibm.com/weather/vis/df.htm

- Slide 29:
  - Water for Tomorrow: http://www.nature.org/ourinitiatives/habitats/riverslakes/howwework/The-Partnership.xml and http://www.ibm.com/ibm/ideasfromibm/us/environment/100807/images/IFI_10082007.pdf