Data Analysis Tools for Sensor-Based Science

Stuart Ozer, Jim Gray
Microsoft Research
San Francisco, CA
{struarto,Jim.Gray}
@microsoft.com

Alex Szalay, Andreas Terzis, Razvan Musaloiu-E.
Johns Hopkins University
Baltimore, MD
{szalay,terzis,razvanm,}@jhu.edu

Katalin Szlavecz, Randal Burns, Josh Cogan
Johns Hopkins University
Baltimore, MD
{szlavecz, randal, joshc}@jhu.edu

ABSTRACT
Science is increasingly driven by data collected automatically from arrays of inexpensive sensors. The collected data volumes require a different approach from the scientists' current Excel spreadsheet storage and analysis model. Spreadsheets work well for small data sets; but scientists want high level summaries of their data for various statistical analyses without sacrificing the ability to drill down to *every* bit of the raw data. This demonstration describes our prototype data analysis system that is suitable for browsing and visualization – like a spreadsheet – but scalable to much larger data sets

Categories and Subject Descriptors
H.3.3. **[Information Research and Retrieval]:** Information filtering.

General Terms
Management, Measurement, Experimentation, Human Factors.

Keywords: Sensor Networks, Data Cubes.

1. INTRODUCTION

While the proposed approach is applicable to any Wireless Sensor Network that generates large amounts of data, collected by large collections of sensors over long periods of time, we ground our design through a environmental monitoring application we developed and deployed during the fall of 2005 [3].

The purpose of our WSN is soil monitoring, in which motes periodically collect soil measurements including soil temperature and soil humidity, as well as ambient temperature and light. Measurements are stored locally on the motes' flash memory until they are retrieved by a network gateway using a reliable transfer protocol.

All collected measurements are subsequently inserted to a relational database. This database not only stores "raw" measurements, but also calibrated versions, calculated using stored procedures, and drives user interfaces including HTML and Web Services front-ends [2]. As a way to allow arbitrary analyses, the Web interface allows SQL queries to be sent directly to the database. This "guru" interface has already been very useful but at the same time domain scientists prefer to interact with visual and high-level summarization tools rather than having to use a different set of tools to analyze data extracted from the database.

The data analysis tools presented in the following section provide this service.

2. DATA CUBES FOR DATA ANALYSIS

The calibrated and interpolated data, available in the relational database, can answer a variety of scientific questions exploring both the time and spatial dimensions for small soil ecosystems. However, equally important to examining individual measurements and looking for unusual cases, ecologists want a high level view of the measured quantities. They want to analyze aggregations and functions of the sensor data, visualize trends, and cross-correlate them with other biological measurements at many different scales.

These requirements for slicing, aggregation and analysis can be summarized by general ad-hoc query requests such as: **(1)** Display the functions of measurements (*e.g.,* average, min, max, standard deviation) for a particular time or time interval, for one sensor, for a patch, for all sensors at a site, or for all sites. **(2)** Show the results as a function of depth, time, and category (land cover, age of vegetation, crop management type, upslope, downslope, etc.).

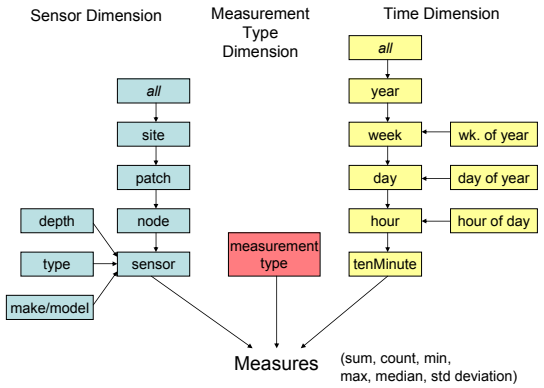

Figure 1. Sensor data cube dimensional model.

SenSys'06, November 1–3, 2006, Boulder, Colorado, USA.
ACM 1-59593-343-3/06/0011.

These later questions are ideally suited for a specialized database design typical of online analytical processing — a *data cube* that supports aggregation (rollup) and drill down operations across many dimensions [1]. Figure 1 illustrates an example of a data cube and unified dimension model derived from the relational database used in the soil monitoring experiment.

The cube provides access to all sensor measurements including air and soil temperature, soil water pressure and light flux averaged over 10-minute measurement intervals, in addition to daily averages, minima and maxima of weather data including precipitation, cloud cover and wind.

The cube also defines calculations of average, min, max, median and standard deviation that can be applied to any type of sensor measurement over any selected spatio-temporal range. Analysis tools querying the cube can display these aggregates easily and quickly, as well as apply richer computations such as correlations. Users can aggregate and pivot on a variety of attributes: position on the hillside, depth in the soil, under the shade vs. in the open, etc.

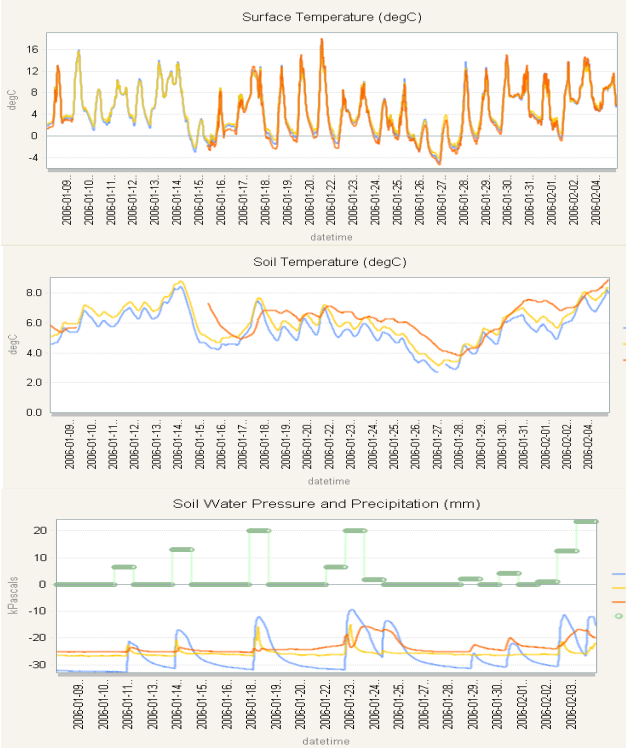

Figure 2. Temperature data recorded by three motes at soil surface (a) and at 10 cm depth (b) during January 2006. Soil moisture in January 2006 reported by three motes, superimposed with precipitation data (bars) (c).

The cube, implemented in SQL Server 2005 Analysis Services, organizes the measurements around three dimensions *when-where-what*: Time, Location-Sensor, and Measurement Type (see Figure 1.) Arrows connecting elements within the `Sensor` and `Time` dimensions document one-to-many relationships, and are essential to specify as *attribute relationships*.

The `Time` dimension includes a hierarchy providing natural aggregation levels for measurement data at the resolution of year, season, week, day, hour and minute (to the grain of 10-minute interval). Not only can data be summarized to any of these levels (*e.g.* average temperature by week), but this summarized data can then also be easily grouped by recurring cyclic attributes such as hour-of-day and week-of-year.

The `Sensor` dimension includes a geographic hierarchy permitting aggregation or slicing by site, patch, mote or individual sensor, as well as a variety of positional or device-specific attributes (patch coordinates, mote position, sensor manufacturer, etc.)

Data visualization, trending and correlation analysis is most effective when measurement data is available for uniform measurement points. While it is straightforward to handle large contiguous data gaps by eliminating a gap period from consideration, frequent gaps can interfere with calculations of daily or hourly averages. To avoid these problems, we interpolate small holes in the data prior to populating the cube.

3. RESULTS

For this demonstration we present the analysis tools for measurements collected by a network of 10 motes deployed into an urban forest environment [3]. A subset of the temperature and moisture data is shown on Figure 2. An interesting comparison can be made between air temperature at the soil surface and soil temperature at 10cm depth. While surface temperature dropped below 0ºC several times, the soil itself was never frozen. This might be due to the vicinity of the stream, the insulating effect of the occasional snow cover, and heat generated by soil metabolic processes. Several soil invertebrate species are still active even a few degrees above 0ºC and, thus, this information is helpful for the soil zoologist in designing a field sampling strategy.

4. REFERENCES

[1] J. Gray, A. Bosworth, A. Layman, and H. Pirahesh, "Data cube: A relational operator generalizing group-by, crosstab and sub-totals," *ICDE 1996*, pages 152–159, 1996.

[2] http://lifeunderyourfeet.org

[3] R. Musaloiu-E., A. Terzis , K. Szlavecz , A. Szalay, J. Cogan , J. Gray, "Life Under your Feet: A Wireless Soil Ecology Sensor Network." Proc. 3rd Workshop on Embedded Networked Sensors (EmNets 2006). May 2006, Cambridge MA.