1. Introduction

Web Data and the two cultures:

- WEB provides a simple and universal standard for the exchange of information.

  Information is decomposed into named units (URLs corresponding to a file) and transmitted.

  HTML is the language used to structure the test for visual presentation (able to describe intra-document structure - layout and inter-document structure - hyperlinks).

  HTTP is the protocol used by Web servers and Web Clients to exchange information (typically HTML documets).

- DATABASES provide mechanisms to represent and manipulate data using strict structure guidelines (schemas).

  Query languages are used to access the information in an ad-hoc manner.

- The need for a BRIDGE between the two cultures. Consider the following scenario:

  An organization publishes financial information on the Web using a relational database as the source. Web pages are generated on demand by executing queries in SQL and converting the results into HTML.

  A second organization wants to obtain financial analysis of this data, but has access only to the HTML pages.

One solution: second organization writes software that parses HTML data and converts into a suitable format for the data analysis software.

Problems with this approach:

(1) A minor change in the HTML formatting could break the parsing software

(2) The software may have to download a large part of the underlying database through repeated requests for HTML pages to compute some quantity such as an average of a single column! This could have been accomplished easily with an SQL query.

XML is the first step towards the convergence of the two views.

XML provides ability to structure data independent of the display format; Easily transmitted and exchanged.

A data model for semi-structured data (XML) with querying capability is a step towards the convergence of the two cultures.

Shift in paradigm:

traditional 2-tier (client/server) to multi-tier

several SERVERS/DATA SOURCES
several CLIENTS/CONSUMERS of the data
MIDDLE TIER responsible for transforming/integrating/adding value