

**Presentation for Macao for ICE-GOV Conference**

**JalaSRI Data Warehousing For  
District Jalgaon (INDIA)  
To Help Transition to  
Better Governance and Data Empowerment**

**Vinay Dharmadhikari , Sanjay Pawde,  
G.P Patil and Kishor Mahajan**

**Watershed Surveillance Research Institute (JalaSRI)  
Jalgaon, India**

1. Background
2. Motivation
3. Rationale
4. The OLAPs
5. Why Data warehousing ?
6. Strategy For Design & Development Of Data Warehouse
7. Data Warehouse Configuration
8. Data Warehouse Architecture Review & Design
9. Identified Data Sources In Context Of JalaSRI
10. Identified Data Requirements
11. Future Work
12. Acknowledgements
13. References

# 1. Background

□ JalaSRI has perceived a critical need for multi-disciplinary and multi-institutional R&D program which aims at developing methodologies for building and promoting the use of spatial and non-spatial data, its management and its analysis.

□ There is a need for making available to the decision makers, planners, researchers and public institutions, the necessary data and data handling tools and techniques, for accessing locale relevant data, such as on natural resources and human resources endowments, agriculture, climate, meteorology, etc for generating alternative scenarios for development, planning and interventions.

□ The ultimate goal is to develop an integrated system for data sharing, data accessing and data use for solving locale specific problems.

□ Geo-Informatics as conceived will address all the vital elements viz. geographic measurements, geo-accounting, spatial analysis and integrated spatio-temporal decision-making.

## 2. Motivation

2.1 The data management practices that exist at the district or lower levels of administration such as Taluka and village governance institution, is not yet fully geared to address the information needed for achieving integrated developmental planning, monitoring and follow up.

2.2 The conventional methods of data collection/ collation, storage are not amenable for quick updation, retrieval and holistic analysis, several problems are encountered:

- (i) Incompatible and non- standard datasets and databases, which inhibits easy integration
- (ii) Inefficient inter-sectoral data flow
- (iii) Problems related to quality, completeness and lineage of data

2.3 The process of integrated approach to planning requires a detailed knowledge of the interrelations and interdependencies between various sectors to resolve often-conflicting demands of the community.

## 2. Motivation (contd...)

2.4 This leads to a requirement for appropriate data management and analyzing tools and techniques and a large matrix of sectoral data, in digital format.

2.5 There is a critical need of enabling people, communities and Institutions of self-governance, with Spatial and non-spatial Data, along with technologies for data handling and data massaging for enabling informed Decision-making.

2.5 There is a critical need of enabling people, communities and Institutions of self-governance, with Spatial and non-spatial Data, along with technologies for data handling and data massaging for enabling informed Decision-making.

# 3. Rationale

3.1 The JalaSRI team in the Jalgaon District of India, is taking initiative for attempting transformation from:-

- 1 an average district governance system,
- 2 of rule bound implementation of plethora(2727) of schemes,in various silos,
- 3 through dozens of subject line-departments,
- 4 managed in a matrix fashion,
- 5 with Line bosses sitting hundreds of miles away in a State Secretariat,
- 6 functionally reporting to the local administrator, called District Magistrate.

3.2 The various working groups of JalaSRI will need relevant information/data, on which they are supposed to analyze and propose cost-effective alternative for achieving equitable growth and prosperity generation.

## 3. Rationale (Contd...)

3.3 In this initiative all such requirements will be progressively identified by the JalaSRI team along with its multidisciplinary working groups, working closely with the District and PRI officials throughout the District.

3.4 Based on the information and the data analysis from various JalaSRI working groups, appropriate recommendations will be developed and discussed with the concerned officials of the district administration for modification and implementation. Socio Economic impact, and on ground evaluation will be measured and reported as well as recorded into the data warehouse, for enabling future learning and sharing with other districts.

## 4. The On-line Application Processing Systems (OLAPS)

- ❑ The primary role of **DBMS** is to ensure adequate storage of data and efficient transaction-processing environment offering the user rapid access to data.
- ❑ The *database (DB) systems* were used (and still are used) mainly to ensure daily business operations, such as sales, production, etc. and they are called *operational system or operational databases*.
- ❑ The main concern in these systems is to ensure concurrent access and recovery techniques that guarantees data consistency.
- ❑ Some of these systems manage high number of concurrent transactions and give the user the possibility of on-line interactions with the database executing operations according to the business activities. These special kind of operational systems receive name *On-Line Application Processing Systems (OLAPS)*.

## The OLAPS (Contd...)

- ❑ The decision-making users need different views and manipulations of data; thus different kinds of systems called *Information Systems* (ISs) are required.
- ❑ Physically, the IS should be separate system with its own data , but in some situations the operational data are queried in order to give necessary information for decision-making users, thus playing the role of an IS.
- ❑ It leads to the evolution of *Decision Support System (DSS)*, which aims to improve the effectiveness of decision-making process and include sophisticated database management capabilities with the access to internal and external data

## 5. Why Data Warehousing ?

- ❑ Data Warehouses (DW) have become the core technology for DSSs offering the large amounts of *integrated, consistent, and historical* data for processing, in order to respond the decision-making user requirements and to meet the IS characteristics.
- ❑ In order to build the scientific databases with the help of modern geo-informatics tools and techniques and to be the unique 'reliable source' for such scientific data, *JalaSRI, Watershed Surveillance Research Institute, Jalgaon, India* has initiated the design and development of Data Warehouse as a DSS.
- ❑ A DW is a software infrastructure, which supports OLAP operations by providing a collection of tools to collect data from a set of distributed heterogeneous sources, *cleans and integrates* this data into uniform representation. It *aggregates and organizes* this data into multidimensional structure and refresh it periodically to maintain its currency and accuracy

## Why Data Warehousing ? (contd..)

- ❑ JalaSRI intends to build DW as an *overall strategy and continuous process*, for building decision support systems and a knowledge-based applications architecture and environment that supports both everyday tactical decision-making and long-term strategizing.
- ❑ The Data Warehouse environment will position JalaSRI to utilize an enterprise-wide data store to link information from diverse sources and make the information accessible for a variety of user purposes, most notably, strategic analysis.
- ❑ Analysts will be able to use the Data Warehouse for strategic purposes such as- *trend identification, forecasting, competitive analysis, and targeted market research*.

## 6. Strategy For Design & Development Of Data Warehouse

- ❑ Accurately identify the information that must be contained in the Data Warehouse.
- ❑ Identify , prioritize and manage the scope of the subject areas to be included in the Data Warehouse.
- ❑ Develop a scaleable architecture to serve as the Warehouse's technical and application foundation, and identify and select the hardware -software - middleware components to implement it.
- ❑ Extract, cleanse, aggregate, transform and validate the data to ensure accuracy and consistency.
- ❑ Defining the correct level of summarization to support decision-making.

## Strategy For Design & Development Of DW (Contd..)

- ❑ Define the correct level of summarization to support decision-making.
- ❑ Providing user-friendly, powerful tools at the desktop to access the data in the Data Warehouse.
- ❑ Educate the community about the realm of possibilities that are available to them through Data Warehousing.
- ❑ Establish the processes for maintaining, enhancing, and ensuring the ongoing evolution and applicability of the Warehouse.
- ❑ Establish a Data Warehouse Help Desk and training users to effectively utilize the desktop tools.

## Strategy For Design & Development Of DW (Contd..)

- ❑ Data Warehouse will be designed around the major **subject areas** of the JalaSRI, while, the operational environment will be designed around applications and functions.
- ❑ The data within the Data Warehouse will be **integrated**. This means that there will be consistency among naming conventions, measurements of variables, encoding structures, physical attributes, and other salient data characteristics.
- ❑ All data in Data Warehouse must be accurate as of some moment in time, providing an **historical perspective**.
- ❑ Data in the Warehouse will be static, not dynamic. The only operations that will occur in Data Warehouse applications will be the initial loading of data, access of data, and refresh of data

# 7. Data Warehouse Configuration

A Data Warehouse configuration, also known as the logical architecture, will include the following components:

- One *Enterprise Data Store (EDS)* - a central repository, which supplies atomic (detail level) integrated information to the whole organization.
- One *Operational Data Store* - a "snapshot" of a moment in time's enterprise-wide data
- One or more individual *Data Mart(s)* - summarized subset of the enterprise's data specific to a functional area or department, geographical region, or time period
- One or more *Metadata Store(s) or Repository (ies)* - catalog(s) of reference information about the primary data. Metadata is divided into two categories: information for technical use, and information for end-users.

## 8. DW Architecture Review & Design

- ❑ The **Architecture** is the logical and physical foundation on which the Data Warehouse will be built. The Architecture Review and Design stage, as the name implies, is both a requirements analysis and a gap analysis activity.
- ❑ The **logical architecture** is a configuration map of the necessary data stores that make up the Warehouse; it includes a central Enterprise Data Store, an optional Operational Data Store, one or more (optional) individual area Data Marts, and one or more Metadata stores.
- ❑ Once the logical configuration is defined, the Data, Application, Technical and Support Architectures are designed to physically implement it.

# DW Architecture Review & Design (contd..)

- ❑ *Gap analysis* is conducted to determine which components of each architecture already exist in the organization and can be reused, and which components must be developed (or purchased) and configured for the Data Warehouse.
- ❑ The *Data Architecture* organizes the sources and stores of information and defines the quality and management standards for data and metadata.
- ❑ The *Application Architecture* is the software framework that controls the movement of data from source to user, including the functions of data extraction, data cleansing, data transformation, data loading, data refresh, and data access (reporting, querying).

# DW Architecture Review & Design (contd..)

- ❑ The *Technical Architecture* provides the underlying computing infrastructure that enables the data and application architectures. It includes platform/server, network, communications and connectivity hardware/software/middleware, DBMS, client/server 2-tier vs.3-tier approach, and end-user workstation hardware/software.
- ❑ The *Support Architecture* includes the software components e.g., tools and structures for backup/recovery, disaster recovery, performance monitoring, reliability/stability compliance reporting, data archiving, and version control/configuration management
- ❑ *Architecture Review and Design* applies to the long-term strategy for development and refinement of the overall Data Warehouse, and is not conducted merely for a single iteration.
- ❑ Where Design tells what to do; Architecture Review and Design tells what pieces one needs in order to do it.

## 9. Identified Data Sources In Context Of JalaSRI

S1 -AISLUS- All India Soil And Land Use Survey, M/O Agriculture

S2 - NBSSLUP - National Bureau Of Soil Survey And Land Use

Planning, Indian Council of Agriculture Research

S3- NNRMS- National Natural Resources Management System,  
and SRSAC- State Remote Sensing Application Center, D/O Spac

S4- Population Census /NSS

S5- Agriculture Census

S6- Animal Husbandry Census

S7- BPL Census

S8- Land Records

S9- Net Area Sown , crop-wise

S10- Flood Control Agencies

S11- Meteorological Data

S12- Open domain Analysis of Paper on Watershed Surveillance and  
Inventions

# 10. Identified Data Sources In Context Of JalaSRI

S13- Agriculture Research Outputs

S14- Irrigation Department Records

S15- Fertilizer Company Soil Tests

S16- Seed Companies Data

S17- Employment Exchanges

S18- Industry Associations

S19- Vocational Education Surveys

S20- Department Of Forest Data

S21- Bio Diversity Registers From PRIs

S22- Ecological NGOs (WWF etc)

S23- Department Of CO-OP / Registrar

S24- Rural Banks/ Credit Societies

S25- TIFR , TISS,Data-bases

S26- Labor Bureau

S27- Factory Statistics from ASI

# Identified Data Requirements For JalaSRI

R1-Latest maps of various scale

R2-Satellite imaginaries of various seasonal times

R3-Land and soil use

R4-Employment seekers

R5- District irrigation and water needs

R6- Crop information, new varieties ,seed/fertilizers,pest management

R7- Land allotment to land less

R8- Credit delivery

R9- Disease Incidences

R10- Water quality

R11- Sanitation data

# Identified Data Requirements For JalaSRI

R12- No of wells with locations and water levels

R13- Fertilizers usage data

R14- Crop area sown

R15- Climate records

R16- Water table depths

R17- Acreage under HYV

R18- Child labour/bonded labour

R19- Species availability / abundance /scarcity

R20- Web link to relevant Data Warehouses like  
PASDA,EPA,USGS etc

# 11. Future Work

- ❑ The field of **spatio-temporal data warehouse** is new, it is still not very well exploited, and it needs to integrate the knowledge from three different research topics: data warehouses, spatial databases, and temporal databases.
- ❑ While JalaSRI researchers will have **free access** to all raw data and analyses, other collaborators will be given access, **only through a designated JalaSRI interface**, and a policy on attribution and feeding back use/publication of material using our data, will be enunciated. Others may be given access only to **public domain** material.
- ❑ Studies and comparisons will be undertaken on other existing data-warehouses, like **PASDA** of Penn state University and of **EPA** and **USGS** of government of USA, etc; and learnings will be made use of.

# Acknowledgements

Junior authors acknowledge the guidance of *Dr. G. P. Patil*, Distinguished Professor, Centre For Statistical Ecology And Environmental Statistics, Pennsylvania State University; University Park; PA 16802-2112, *Khandesh College Education society, Jalgaon* and *Researchers Team of JalaSRI*, M. J. College for their support.

## REFERENCES

1. Agrawal R., Gupta A., Sarawagi S. Modeling Multidimensional Data. In Proc. of 13<sup>th</sup> Int. Conf. on Data Engineering, ICDE, 1997.
2. Los Alamos National Laboratory. Earth & Environmental Sciences. GISLab. Spatial Data Warehouse.  
[http://www.gislab.lanl.gov/data\\_warehouse.html](http://www.gislab.lanl.gov/data_warehouse.html), 2003.
3. <http://www.esri.com/software/arcgis/arcinfo>.  
<http://www.esri.com/software/arcgis/arcinfo>, 2003.
4. Bauer A., Hümmer W., Lehner W. An Alternative Relational OLAP Modeling Approach. In Proc. of the 2nd Int. Conf. on Data Warehousing and Knowledge Discovery, DeWaK, 2000.
5. <http://www.isprs.org/commission4/proceedings/paper.html>, 2002.
6. Berson A., Smith S. Data Warehousing, Data Mining and OLAP. Mc Graw- Hill, 1997.
7. Borges K., Laender A. and Davis C. Spatial Data Integrity Constraints in Object-Oriented Geographic Data Modeling. In Proc. of the ACM Symposium on Advances in Geographic Information Systems, ACM GIS, 1999

Thanks and

Look forward to feedback/collaborations !