DataMart (Data Warehouse) Tool: Mondrian + JRubik

Edwin Rojas (CIP)
ICIS workshop 2006, CIMMYT

May 2006

Data Warehouse Motivation and Examples

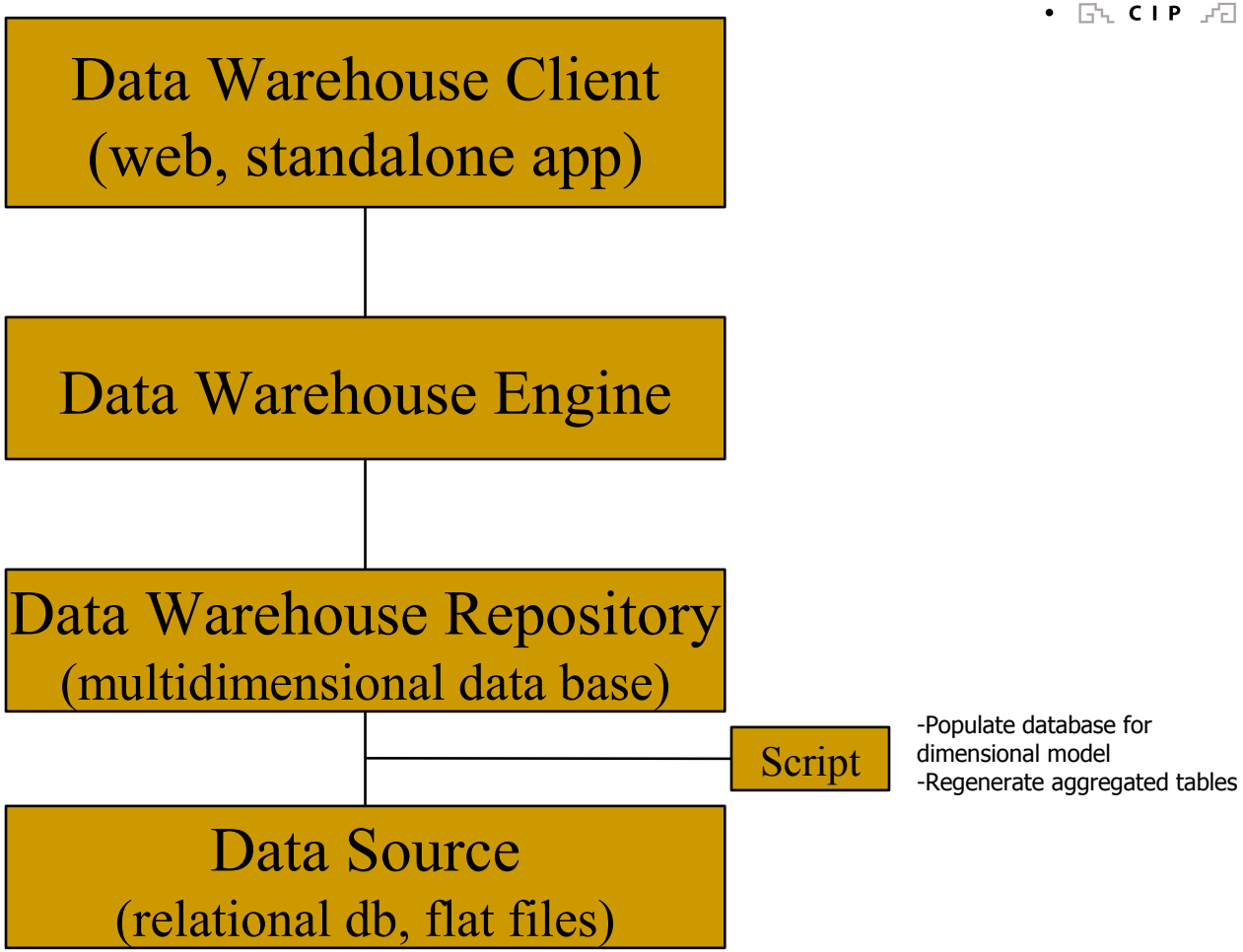
Data Warehouse Motivation

Huge amounts of data need to be summarized in various forms to enable data creators and data users to get quick overviews and dig into details as needed with high performance and flexibility

CIP Example Solutions

1. Holdings by Biological Status	Totals of accesions by biological status.
2. Holdings by Health Status	Totals of accesions by health status.
3. Holdings by Continent-Country	Totals of accesions by administrations (continent and country).
4. Distribution by Accession	Specific information about accessions by region and country of distributed materials from CIP-Lima.
5. Distribution by Institution Type	Distribution materials by accessions, crop, type institution (CGIAR centers, NARS, NGO, etc).
6. Distribution by Country	Number of consignments by country and crop distributed by CIP-Lima.
7. Distribution by Biological Status	Distribution materials by accessions, crop, type form and biological status.
8. Distribution in Invitro Form	**Invitro** distributions from internal (CIP), national and international distributed materials.
9. Distribution by Region	Distribution materials by region. It includes the number of approved requests by region and crop.
10. Morphology	Morphology of native potato with status active.

Data Warehouse Architectural



Data Warehouse Types – Part I

In the OLAP world, there are mainly two different types: Multidimensional OLAP (MOLAP) and Relational OLAP (ROLAP).
Hybrid OLAP (HOLAP) refers to technologies that combine MOLAP and ROLAP.
MOLAP, This is the more traditional way of OLAP analysis.

In MOLAP, data is stored in a multidimensional cube. The storage is not in the relational database, but in proprietary formats.

Advantages:

Excellent performance: MOLAP cubes are built for fast data retrieval, and is optimal for slicing and dicing operations.
Can perform complex calculations: All calculations have been pre-generated when the cube is created. Hence, complex calculations are not only doable, but they return quickly.

Disadvantages:

Limited in the amount of data it can handle: Because all calculations are performed when the cube is built, it is not possible to include a large amount of data
in the cube itself. This is not to say that the data in the cube cannot be derived from a large amount of data. Indeed, this is possible. But in this case, only
summary-level information will be included in the cube itself.
Requires additional investment: Cube technology are often proprietary and do not already exist in the organization. Therefore, to adopt MOLAP technology,
chances are additional investments in human and capital resources are needed.

Data Warehouse Types – Part II

ROLAP

This methodology relies on manipulating the data stored in the relational database to give the appearance of traditional OLAP's slicing and dicing functionality.
In essence, each action of slicing and dicing is equivalent to adding a "WHERE" clause in the SQL statement.

Advantages:
Can handle large amounts of data: The data size limitation of ROLAP technology is the limitation on data size of the underlying relational database. In other words,
ROLAP itself places no limitation on data amount.
Can leverage functionalities inherent in the relational database: Often, relational database already comes with a host of functionalities. ROLAP technologies, since
they sit on top of the relational database, can therefore leverage these functionalities.

Disadvantages:
Performance can be slow: Because each ROLAP report is essentially a SQL query (or multiple SQL queries) in the relational database, the query time can be long if
the underlying data size is large.
Limited by SQL functionalities: Because ROLAP technology mainly relies on generating SQL statements to query the relational database, and SQL statements do not
fit all needs (for example, it is difficult to perform complex calculations using SQL), ROLAP technologies are therefore traditionally limited by what SQL can do. ROLAP
vendors have mitigated this risk by building into the tool out-of-the-box complex functions as well as the ability to allow users to define their own functions.

HOLAP

HOLAP technologies attempt to combine the advantages of MOLAP and ROLAP. For summary-type information, HOLAP leverages cube technology for faster performance.
When detail information is needed, HOLAP can "drill through" from the cube into the underlying relational data.

Multidimensional Model Elements – Part I

Dimension
A category of information. For example, the taxonomy dimension.

Hierarchy Levels
The specification of levels that represents relationship between different attributes within a hierarchy. For example, one possible hierarchy in the Taxonomy dimension is Family --> Genus --> Series --> Species.

A fact table is a table that contains the **measures** of interest. For example, accessions count would be such a measure. This measure is stored in the fact table with the appropriate granularity.

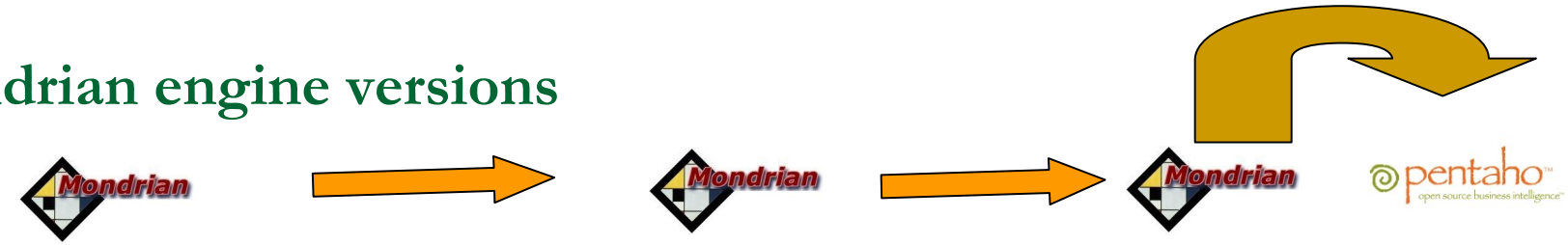
A dimensional model includes fact tables and lookup tables. Fact tables connect to one or more lookup tables, but fact tables do not have direct relationships to one another. Dimensions and hierarchies are represented by lookup tables. Attributes are the non-key columns in the lookup tables.

Data Warehouse Viewers

Mondrian = Data Warehouse Web Viewer + Data Warehouse Engine
http://mondrian.sourceforge.net/

JRubik = Data Warehouse Standalone Viewer (Java Swing)
http://rubik.sourceforge.net/

Mondrian engine versions



1.0 version
2004

Precomputed totals created when the first user run
Stored in temporary cache

2.0 version
August 2005

Precomputed totals created when the model db is created
Stored in tables db

2.1 version
March 2006

Mondrian as a component of business intelligent framework - BI

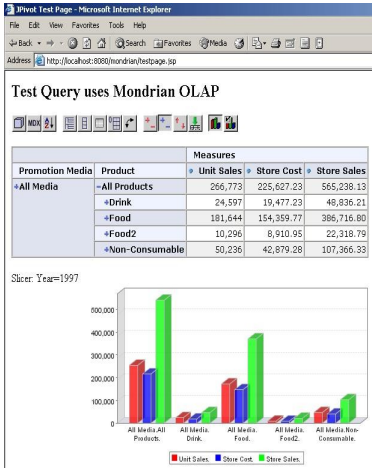
Open Source Data Warehouse Technology

MySQL
MS-Access
MS-SQL
PostgreSQL

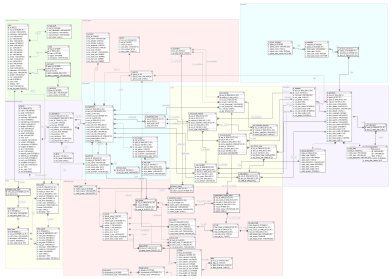
Job Script

MySQL
MS-Access
MS-SQL
PostgreSQL

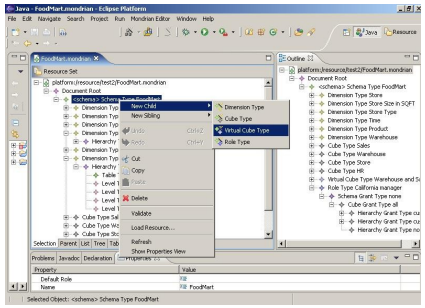
HTML
Pivot table and chart



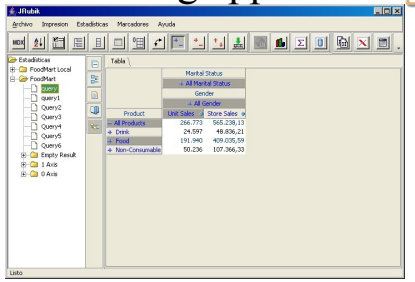
Relational Model



Multidimensional Model



Java/Swing application



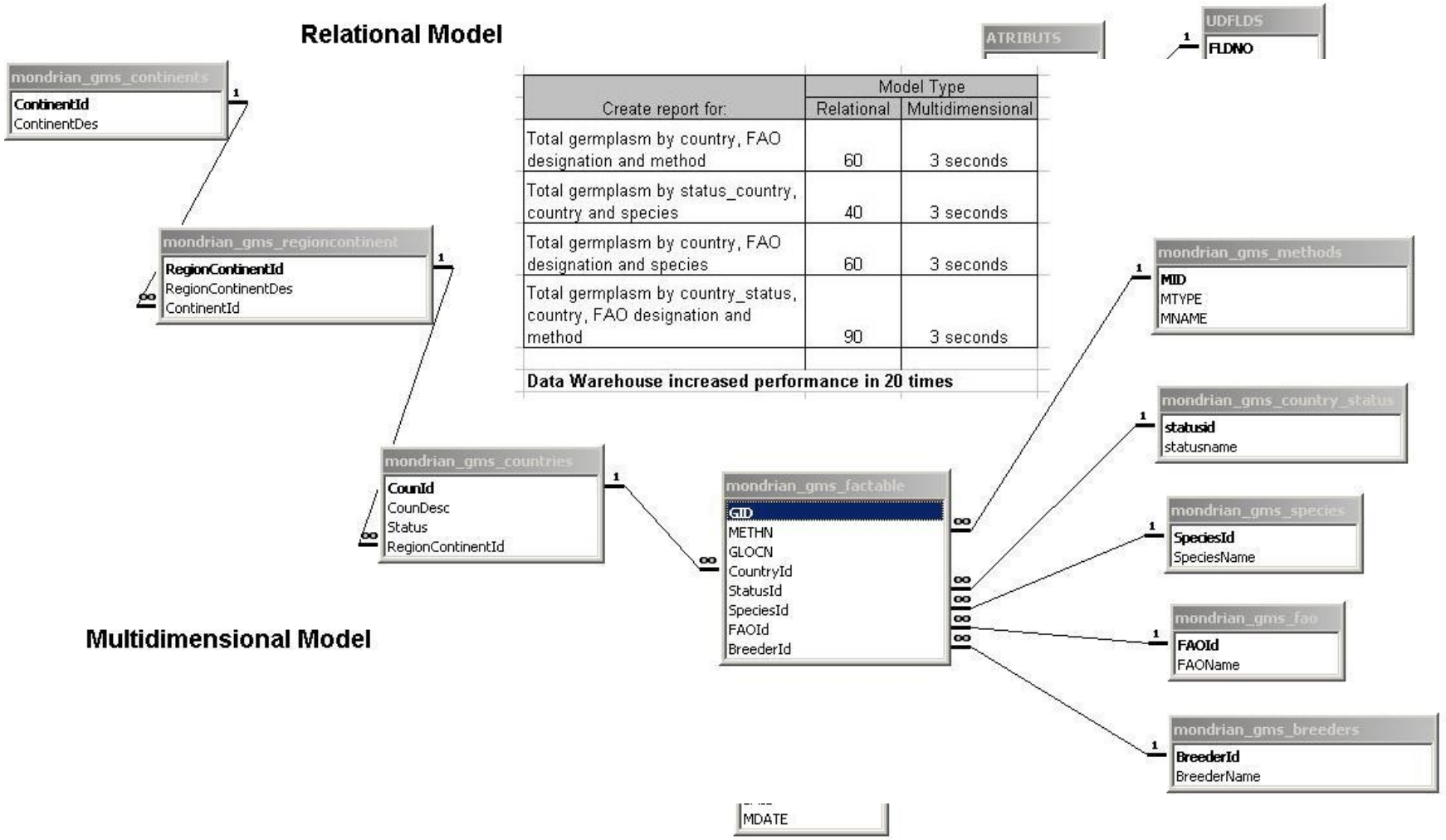
DBDesigner

Eclipse
Plug-in Mondrian

Case Study for ICIS Inventory (IMS) Database


```
<?xml version="1.0"?>
<Schema name="ICIS-IMS">

<Cube name="inventory">
  <Table name="mondrian_inventory_factable">

  </Table>


<Dimension name="WithSeeds" foreignKey="WithSeeds">
      <Hierarchy hasAll="true" allMemberName="All WithSeeds" primaryKey="WithSeeds">
            <Table name="mondrian_inventory_withseeds"/>
            <Level name="WithSeeds" column="WithSeeds" uniqueMembers="false"/>
      </Hierarchy>
</Dimension>


<Dimension name="Shelves" foreignKey="SNL2ID">
      <Hierarchy hasAll="true" allMemberName="All Shelves" primaryKey="LOCID">
            <Table name="mondrian_inventory_shelf"/>
            <Level name="Shelves" column="LNAMESHELF" uniqueMembers="false"/>
      </Hierarchy>
</Dimension>

<Dimension name="Boxs" foreignKey="SNL3ID">
      <Hierarchy hasAll="true" allMemberName="All Boxs" primaryKey="LOCID">
            <Table name="mondrian_inventory_box"/>
            <Level name="Boxs" column="LNAMEBOX" uniqueMembers="false"/>
      </Hierarchy>
</Dimension>


  <Measure name="Seeds_gr_" column="LOTQTY" aggregator="sum" formatString="#,###.00"/>
  <Measure name="Lots_Count" column="LOTID" aggregator="count" formatString="#,###"/>
  <Measure name="Germplasm" column="EID" aggregator="count" formatString="#,###"/>
  <Measure name="GermplasmDistinct" column="EID" aggregator="distinct count" formatString="#,###"/>
  <Measure name="StorageDaysMax" column="StorageDays" aggregator="max" formatString="#,###"/>
  <Measure name="StorageDaysAvg" column="StorageDays" aggregator="avg" formatString="#,###"/>
  <Measure name="StorageDaysMin" column="StorageDays" aggregator="min" formatString="#,###"/>

</Cube>

</Schema>
```

Case Study for ICIS Genealogy (GMS) Database



Create report for:	Model Type	
	Relational	Multidimensional
Total germplasm by country, FAO designation and method	60	3 seconds
Total germplasm by status_country, country and species	40	3 seconds
Total germplasm by country, FAO designation and species	60	3 seconds
Total germplasm by country_status, country, FAO designation and method	90	3 seconds

Data Warehouse increased performance in 20 times

Demos and Tutorial

For Standalone: Rubik viewer View Video: http://research.cip.cgiar.org/docs/mondrian/videos/general_rubik_summary/general_rubik_summary.html

For Web: Mondrian viewer View Video: http://research.cip.cgiar.org/docs/mondrian/videos/general_mondrian_summary/general_mondrian_summary.html

PFD Tutorial for Mondrian: http://research.cip.cgiar.org/docs/mondrian/Tutorial_Mondrian.pdf