Do you have dependability related data? Share it!

Marco Vieira
mvieira@dei.uc.pt

University of Coimbra - Portugal

Winter 2009 Meeting of IFIP WG 10.4
Cortina d'Ampezzo, Italy, January 31st, 2009

The AMBER Repository

- Worldwide repository for dependability related data
- Key objectives:
  - Provide state-of-the-art data analysis
  - Allow data comparison and cross-exploitation
  - Facilitate worldwide data sharing and dissemination
- Potential tool to increase the impact of research

Motivation

- **Analyzing** large amounts of raw data produced in dependability evaluation is difficult
- **Comparing** results from different experiments or results of similar experiments across different systems is complex
  - Different and incompatible tools, data formats, and setup details…
- **Sharing** raw experimental results among research teams is hard

Current situation

- The situation today is not good!!!
- Spreadsheets and other specific tools to analyze results
  - Not standard and difficult to build
- Difficult to compare data and generalize conclusions
- Researchers share the final results and the conclusions
  - Papers, mainly
  - Raw data is not shared

"One of us is in serious trouble!"

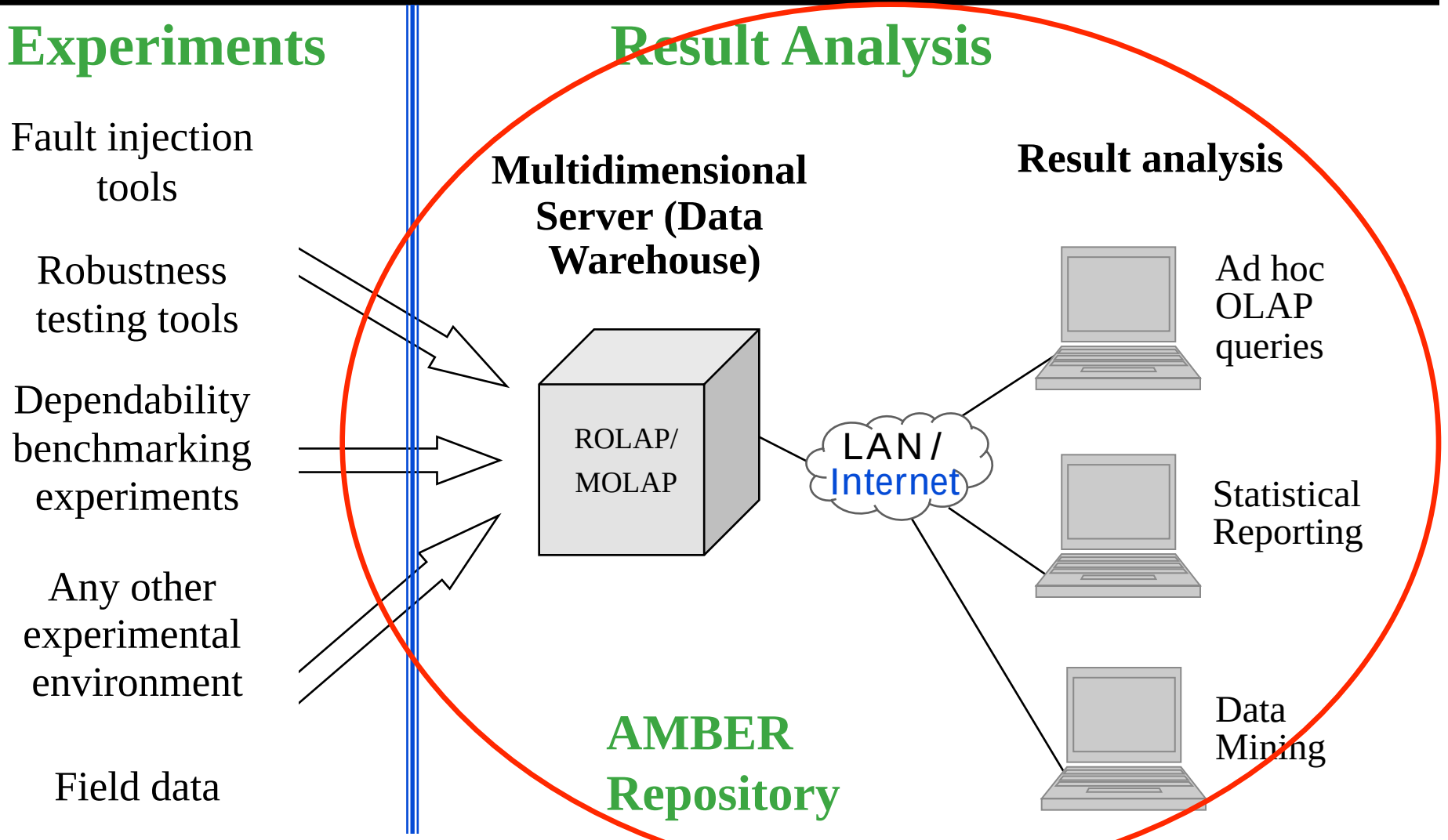
Potential use of the ADR

- Research team level
  - Perform the analysis of data in an efficient way
  - Efficient dissemination of the results of the team
- Project level
  - Sharing and cross-exploitation of results from different project teams
- World wide
  - Common repository to store and share data
  - Many teams are performing dependability evaluation but there are no results available at the web

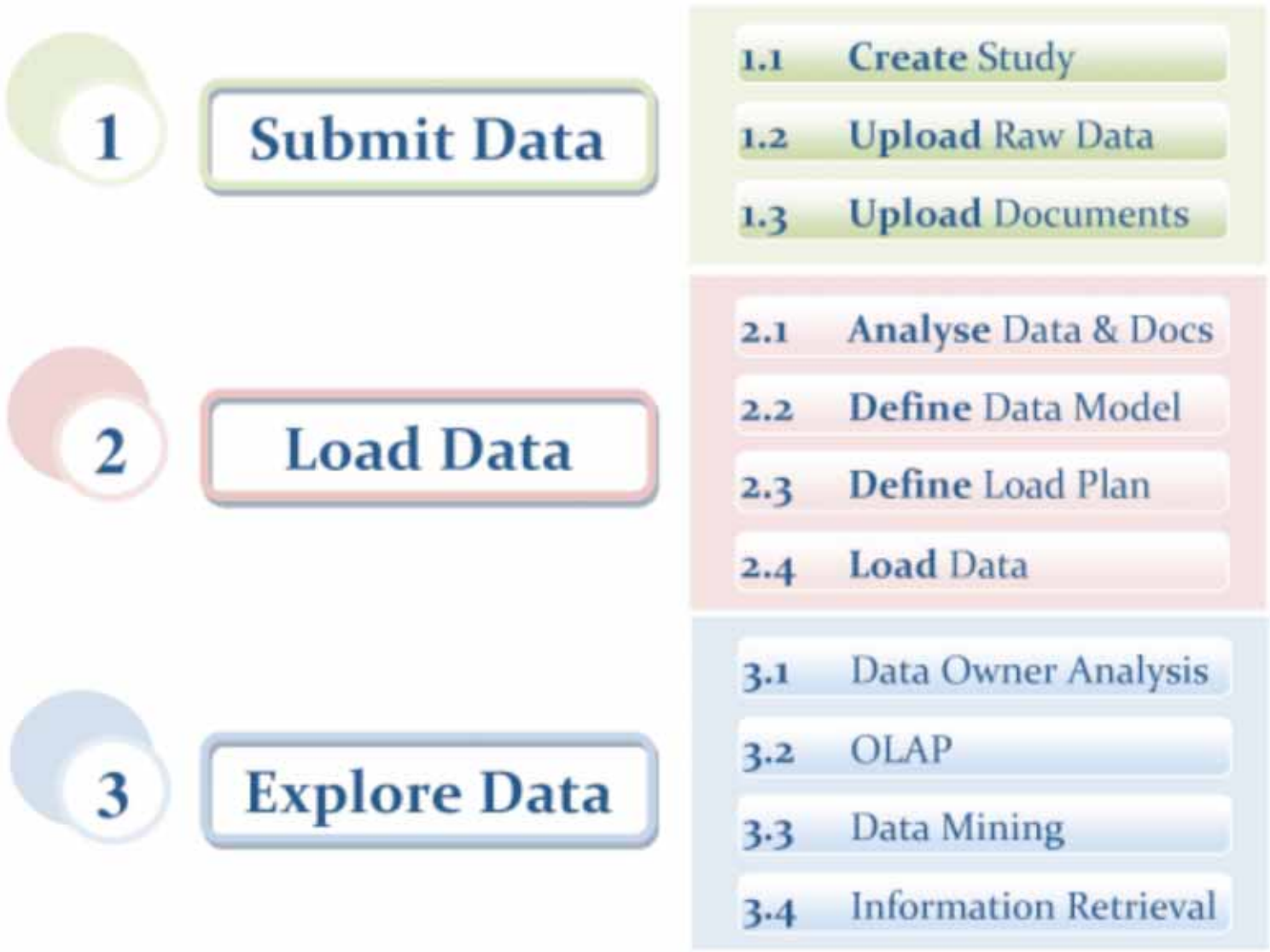
Data analysis approach

- Repository to analyze, compare, and share results
- Use a business intelligence approach:
  - Data warehouse to store data
  - On-Line Analytical Processing (OLAP) to analyze data
  - Data mining algorithms to identify (unknown) phenomena in the data
  - Information retrieval to access data in textual formats
- Adopt the same life cycle of BI data
- Use technology already available for BI

AMBER Repository

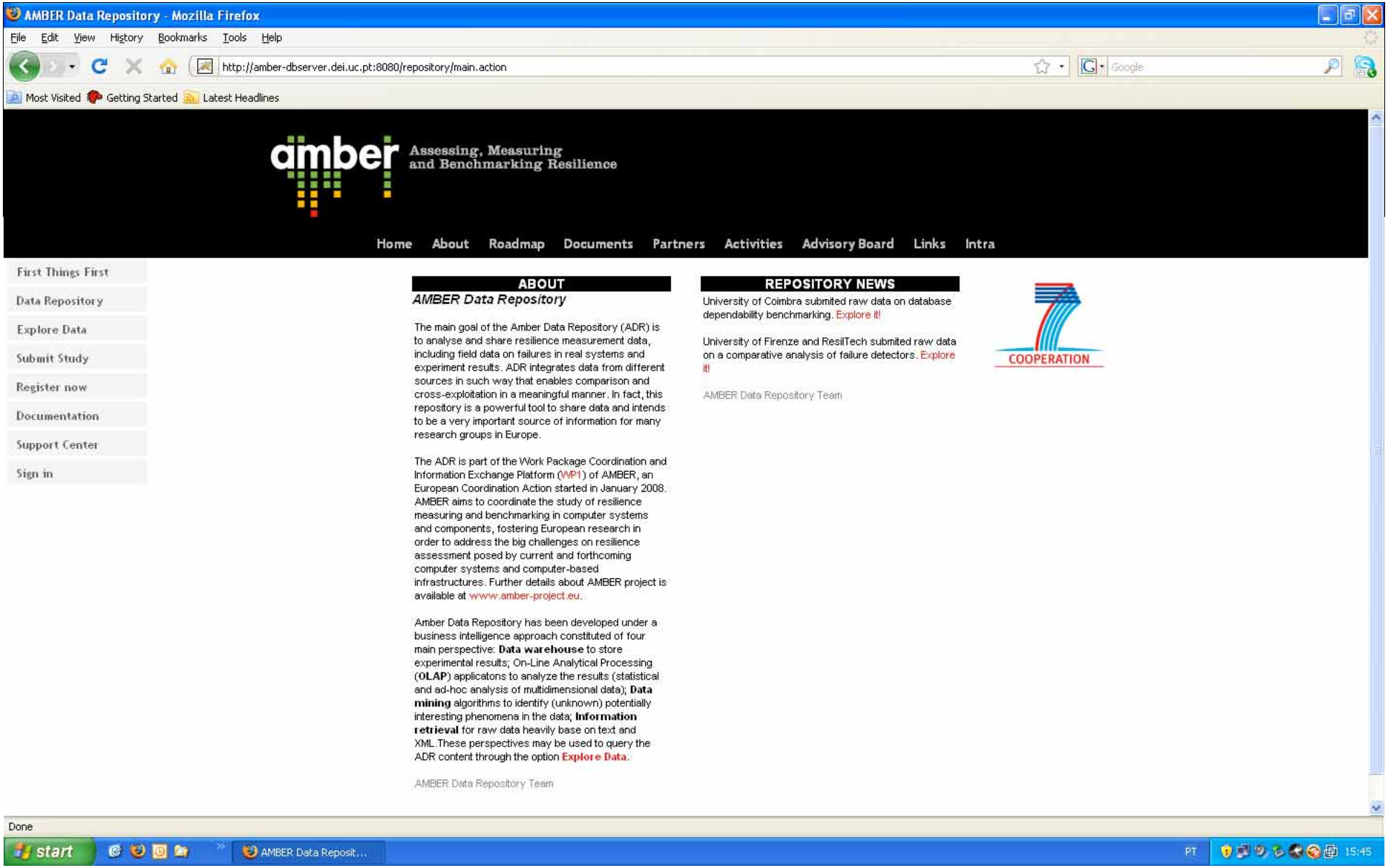


How to use it?


1
Submit Data
1.1 Create Study
1.2 Upload Raw Data
1.3 Upload Documents
2
Load Data
2.1 Analyse Data & Docs
2.2 Define Data Model
2.3 Define Load Plan
2.4 Load Data
3
Explore Data
3.1 Data Owner Analysis
3.2 OLAP
3.3 Data Mining
3.4 Information Retrieval

It does exist! ☺

- http://www.amber-project.eu


AMBER Data Repository - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://amber-dbserver.dei.uc.pt:8080/repository/main.action
Most Visited Getting Started Latest Headlines
amber
Assessing, Measuring and Benchmarking Resilience
Home About Roadmap Documents Partners Activities Advisory Board Links Intra
First Things First
Data Repository
Explore Data
Submit Study
Register now
Documentation
Support Center
Sign in
ABOUT
AMBER Data Repository
The main goal of the Amber Data Repository (ADR) is to analyse and share resilience measurement data, including field data on failures in real systems and experiment results. ADR integrates data from different sources in such way that enables comparison and cross-exploitation in a meaningful manner. In fact, this repository is a powerful tool to share data and intends to be a very important source of information for many research groups in Europe.
The ADR is part of the Work Package Coordination and Information Exchange Platform (WP1) of AMBER, an European Coordination Action started in January 2008. AMBER aims to coordinate the study of resilience measuring and benchmarking in computer systems and components, fostering European research in order to address the big challenges on resilience assessment posed by current and forthcoming computer systems and computer-based infrastructures. Further details about AMBER project is available at www.amber-project.eu.
Amber Data Repository has been developed under a business intelligence approach constituted of four main perspective: Data warehouse to store experimental results; On-Line Analytical Processing (OLAP) applicatons to analyze the results (statistical and ad-hoc analysis of multidimensional data); Data mining algorithms to identify (unknown) potentially interesting phenomena in the data; Information retrieval for raw data heavily base on text and XML.These perspectives may be used to query the ADR content through the option Explore Data.
AMBER Data Repository Team
REPOSITORY NEWS
University of Coimbra submited raw data on database dependability benchmarking. Explore it!
University of Firenze and ResilTech submited raw data on a comparative analysis of failure detectors. Explore it!
AMBER Data Repository Team
COOPERATION
Done
start
AMBER Data Reposit...
PT
15:45

Conclusions

- Powerful tool to disseminate research results
  - Simplicity
  - Support for older data
  - Well-proven analysis techniques and technologies
  - Automated data discovery facilities
  - Cross exploitation
  - Dissemination
- Do you have data? Try it!!!
  - We will help ☺

Questions (for you ☺)

- Is it useful?
- How can we convince you to try it?
- How can we convince companies to provide data?
- Should we have a business model?


amber

How to use it? (1)

1. User registration
   - Provide identification information that is verified by the ADR support team
     - To filter malicious users
2. Multidimensional analysis
   - Design an adequate multidimensional data model
3. Definition of the loading plans
   - Data extraction, transformation, and loading

How to use it? (2)

4. Load the data
   - Executing the loading plans created before
   - If new data becomes available we just need to rerun the plans
     - e.g., if new systems are evaluated
5. Definition of data ownership policies
   - Private, proprietary, collaborative
6. Analysis of the data

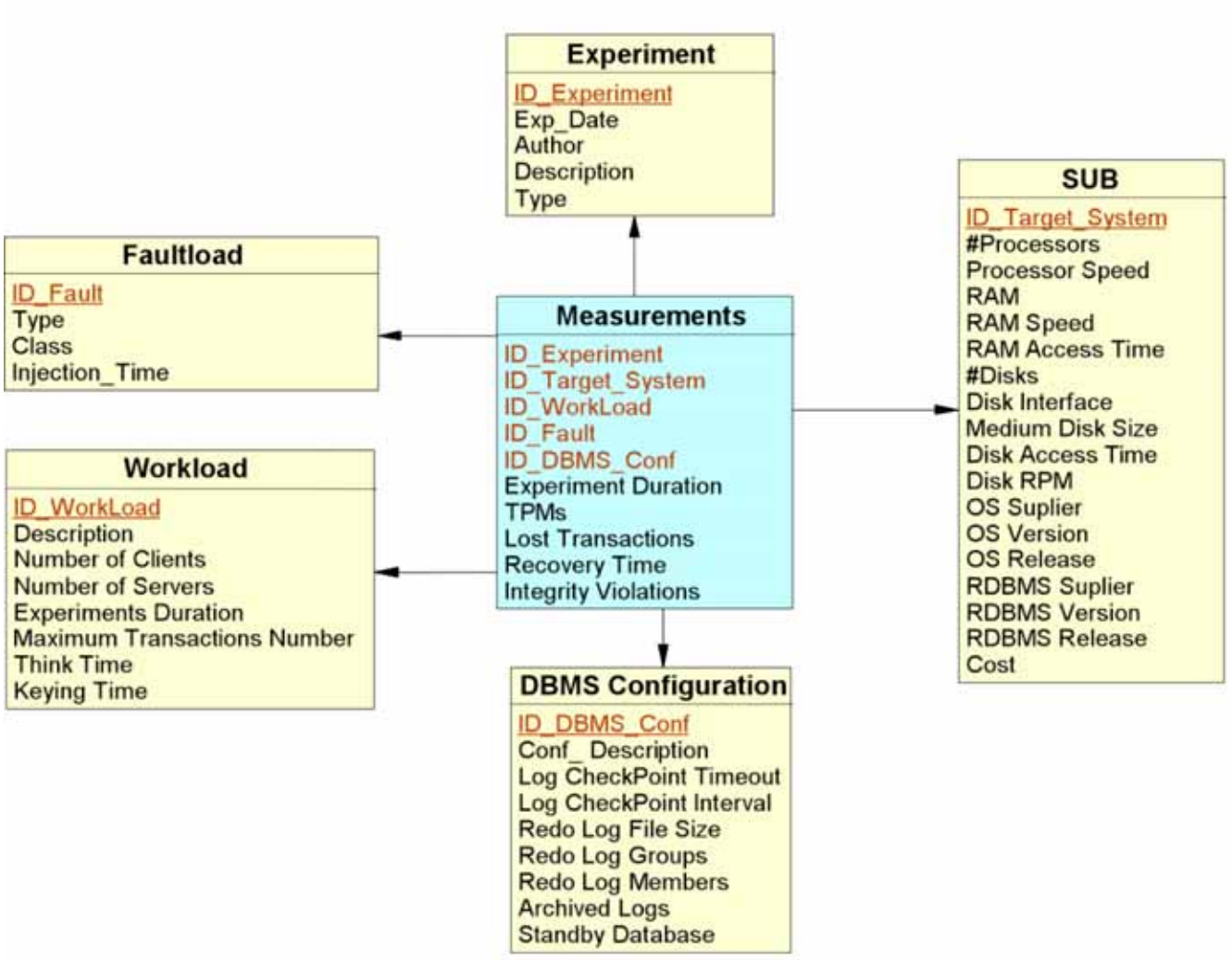
Analysis of the data

- On-line Analytical Processing (OLAP) tools
  - Support the analysis in a very flexible way
  - Provide high query performance and easy, intuitive data navigation
- Data mining
  - Automatic discover of correlations in the data
- Statistical analysis
- Information retrieval
  - For textual data

Data model (1)

- Key steps:
  - Identification of the facts that characterize the problem under analysis
  - Identification of the dimensions that may influence the facts
  - Definition of the granularity of the data stored in the star schema

Data model (2)


Experiment
ID_Experiment
Exp_Date
Author
Description
Type
SUB
ID_Target_System
#Processors
Processor Speed
RAM
RAM Speed
RAM Access Time
#Disks
Disk Interface
Medium Disk Size
Disk Access Time
Disk RPM
OS Suplier
OS Version
OS Release
RDBMS Suplier
RDBMS Version
RDBMS Release
Cost
Faultload
ID_Fault
Type
Class
Injection_Time
Measurements
ID_Experiment
ID_Target_System
ID_WorkLoad
ID_Fault
ID_DBMS_Conf
Experiment Duration
TPMs
Lost Transactions
Recovery Time
Integrity Violations
Workload
ID_WorkLoad
Description
Number of Clients
Number of Servers
Experiments Duration
Maximum Transactions Number
Think Time
Keying Time
DBMS Configuration
ID_DBMS_Conf
Conf_ Description
Log CheckPoint Timeout
Log CheckPoint Interval
Redo Log File Size
Redo Log Groups
Redo Log Members
Archived Logs
Standby Database