Schema Evolution Toolsuite:
integrating collection and analysis of schema evolution history

Technologies for Information Systems
A.Y. 2008/09

Professors: L.Tanca, F.A. Schreiber
In collaboration with: C.A. Curino

Student: Moroni Fabrizio – 706592

Previous work – Schema Evolution in Wikipedia

Schema Evolution Toolsuite

- Collected and dissected MediaWiki schema history (170+ schema versions in 4.5 years)
- Analysis of the data by hand, shell script (not portable)
- The interesting results of the work arose the need:
  - Extend the analysis to different systems
  - Automating the data collection

Our toolsuite...

Schema Evolution Toolsuite

Goals:

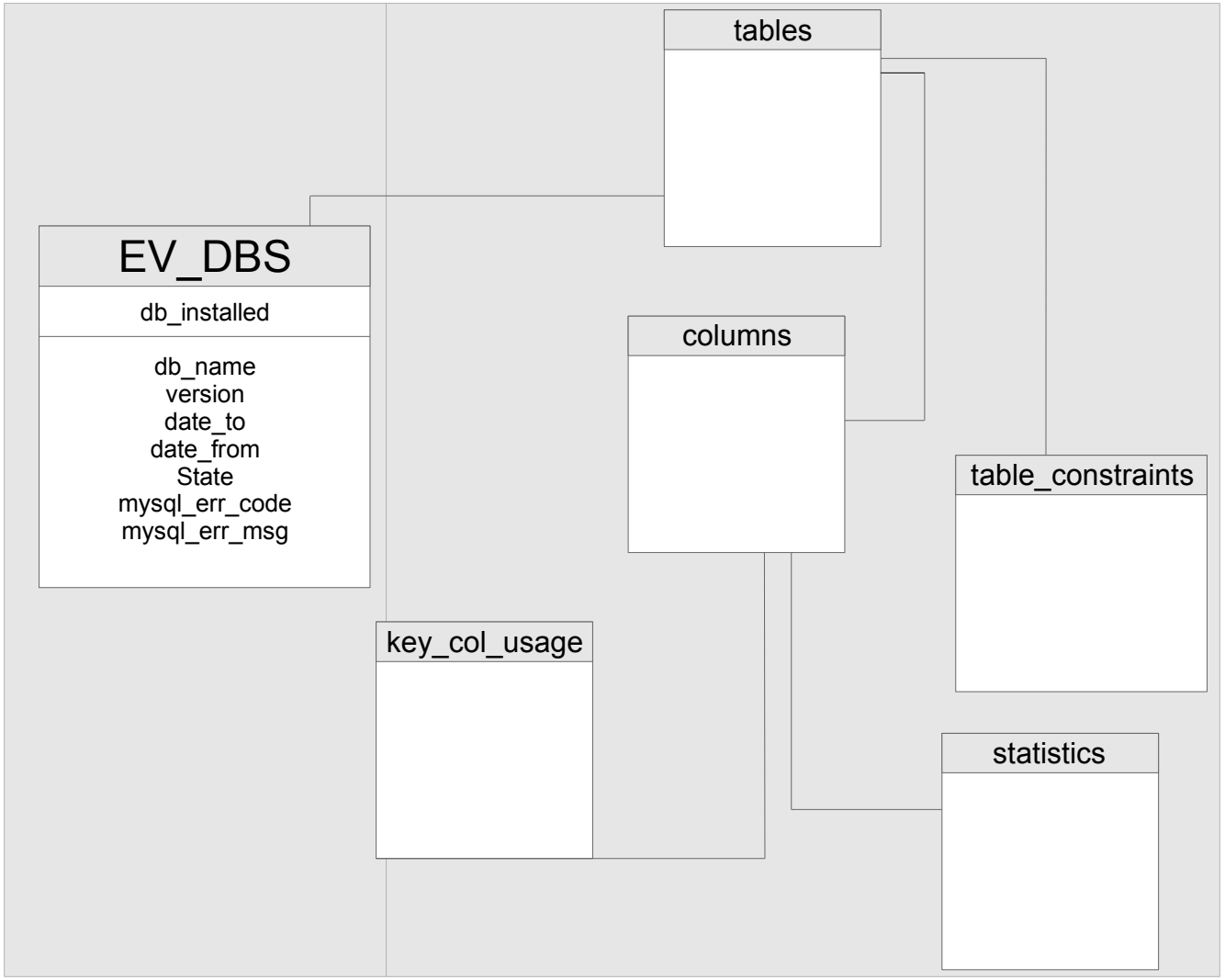
- Collect several schema evolution's history from repository
- Automate statistical analysis
- Understands database evolution rate

- It's a suite capable of downloading, install and collect data from several systems
- Multiple repository support CVS and SVN
- Provides an extended set of statistics performed over the desired system, giving the user/ researcher a quick understanding of the status of the system they're going to work with.

Evolution database's: make information_schema persistent

Schema Evolution Toolsuite

- Performance
- Portability
- Completeness



MySQL information schema tables

Data collection module

Schema Evolution Toolsuite

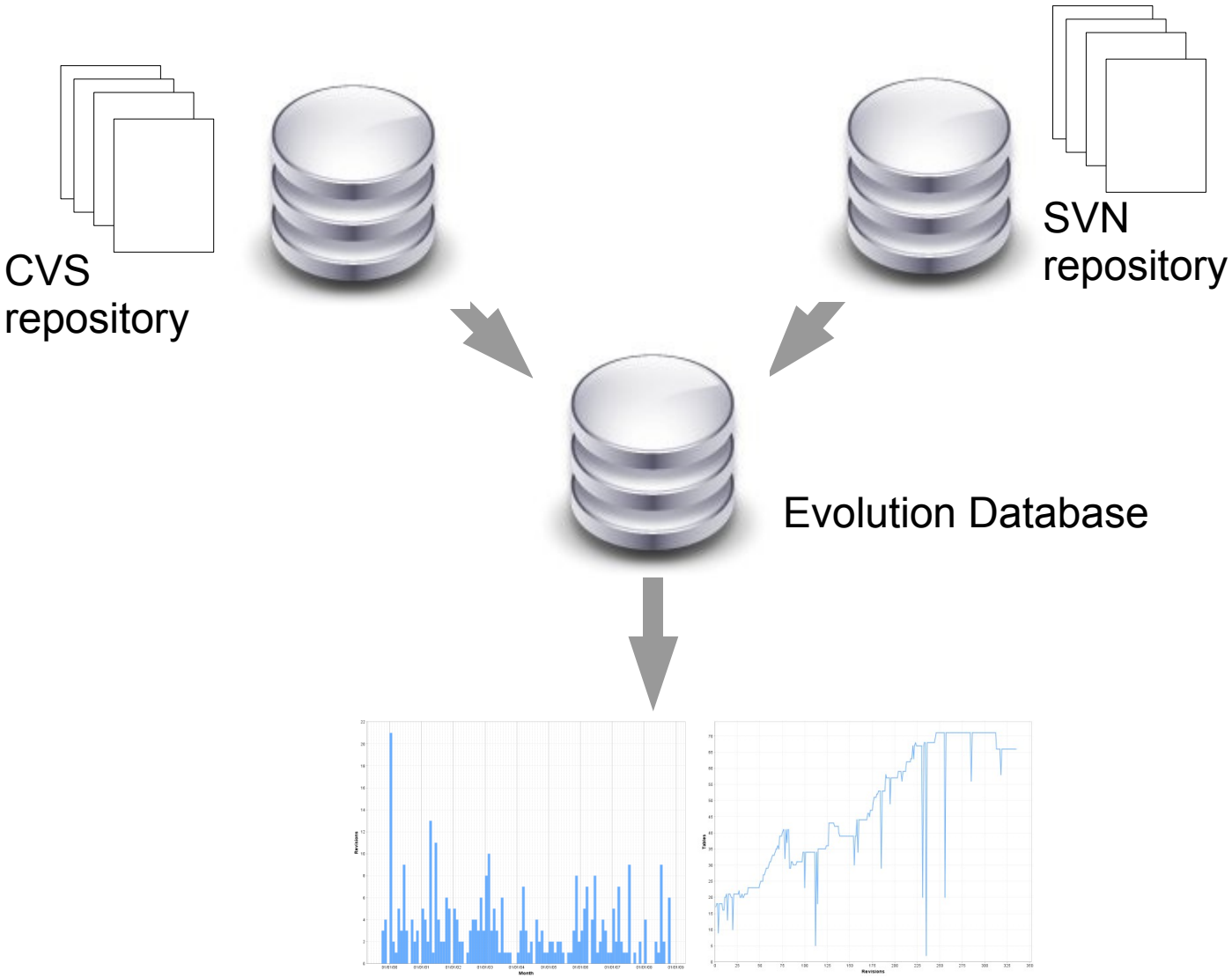
SCHEMA DOWNLOAD PHASE: the suite connects to the application's repository and retrieve the information about the database's revisions

SCHEMA INSTALLATION PHASE: the downloaded schema are installed inside the current MySQL instance

SCHEMA DATA COLLECTION: the suite reads the information schema's tables and collects the data relative
to the system installed

SCHEMA DROP: the database are deleted from the server in order to keep the server clean and efficient

Schema Evolution Toolsuite



Statistics module

Schema Evolution Toolsuite

The suite uses the data collected in the previos phase to generate statistics and display graphically the evolution of the system.

The statistics has been divided in five groups:

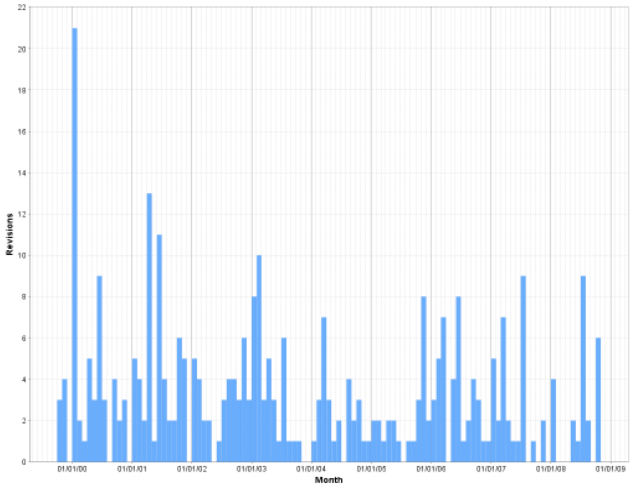
- Schema's Version
- Table
- Columns
- Constraints
- Index

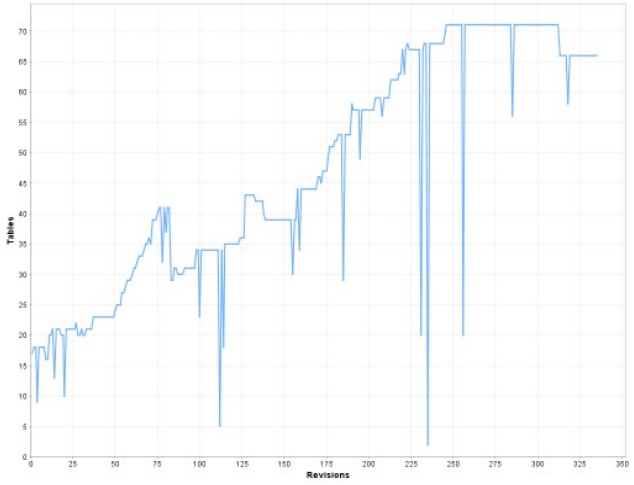
Statistics sample

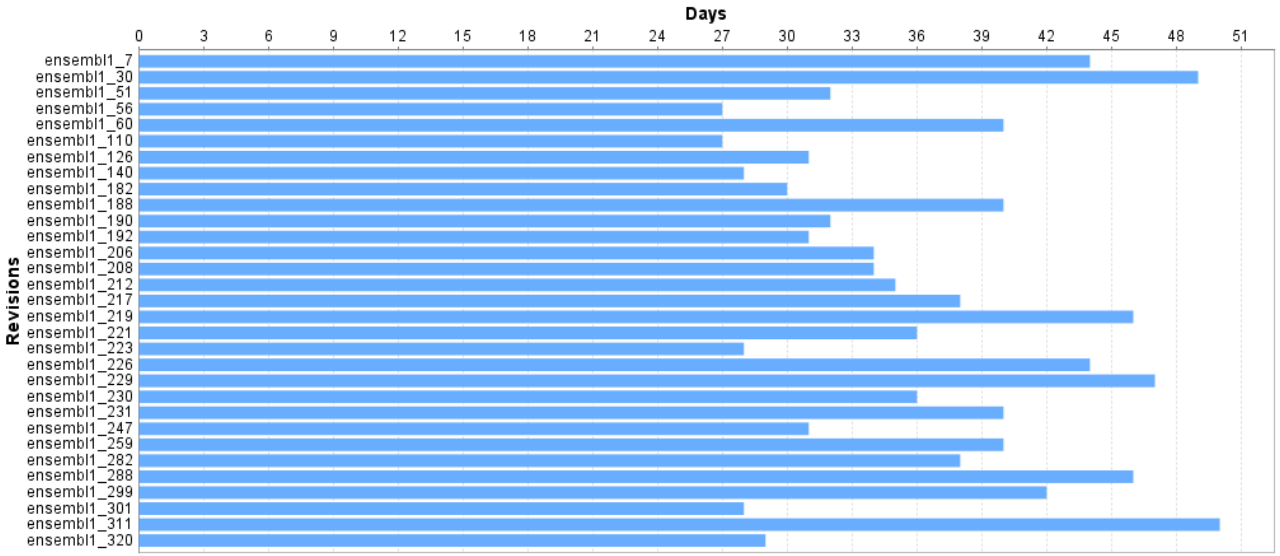
Schema Evolution Toolsuite

Basic set of statistic:
give an idea of the overall evolution of the schema

- Column's evolution
- Tables' evolution

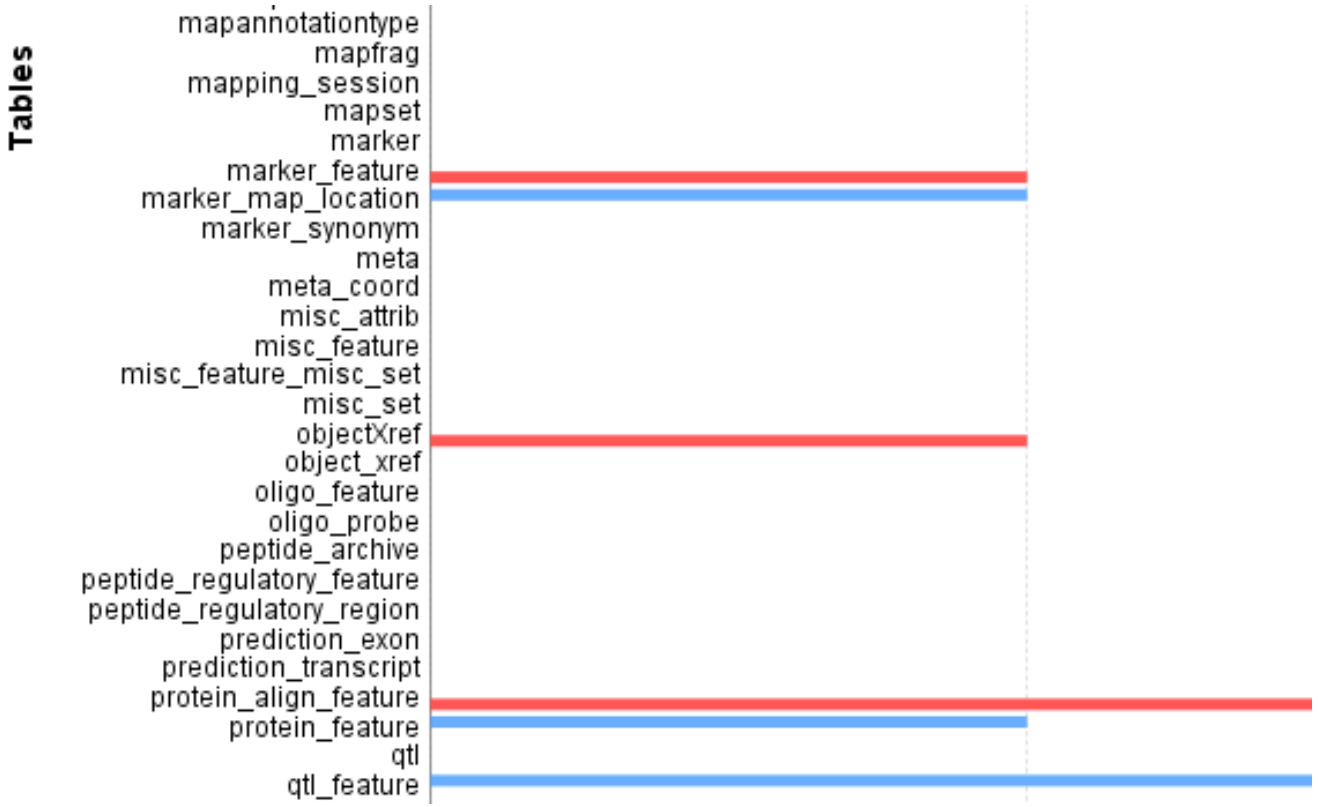






Statistics sample -2

Schema Evolution Toolsuite



Extended Set:
in depth analysis of the system

- Index evolution
- Constraints evolution

Definition of a new set of statistics
to support the schema analysis

- Rollback
- Name/number change

Statistic summary and evolution's rate

Schema Evolution Toolsuite

The suite complets the system's analysis defining a global (database) evolution rate

- Sum up the result of each statistic
- Define a table's evolution rate and visualize it with a red/green bar

Table Name	Colomns evolution	Colomns data type changes	Number of roll back	Primary key changes	Foreign key changes	Index changes	Table risk factor	Table Name
analysis	9	Show/Hide DataTypes...	0	0	0	0	5%	analysis
analysis_history	15		0	0	0	0	8%	analysis_history
clone	21		0	2	0	0	29%	clone
contig	36	Show/Hide DataTypes...	0	0	0	0	19%	contig
contig_equiv	13		0	0	0	0	6%	contig_equiv
dna	15	Show/Hide DataTypes...	0	0	0	0	8%	dna
exon	52	Show/Hide DataTypes...	0	4	0	0	65%	exon
exon_feature	15		0	0	0	0	8%	exon_feature
exon_transcript	10		0	0	0	0	5%	exon_transcript
feature	22	Show/Hide DataTypes...	0	0	0	0	12%	feature
fset	13	Show/Hide DataTypes...	0	0	0	0	7%	fset
fset_feature	5	Show/Hide DataTypes...	0	0	0	0	3%	fset_feature

Definition of a database evolution rate

- Arithmetic mean of the single table's rate

Table's risk factor computation

Schema Evolution Toolsuite

In order to compute analytically these value we followed these steps:

- execution of the SET statistic module over multiple sample systems
- collection of the results
- definition of the evolution rate associated to the table
- computation of the variables' coefficient via a statistical software

Table Name	Colomns count evolution	Data Type change	Number of roll back	Primary key changes	Foreign key changes	Index changes
analysis	9	3	0	0	0	0
analysis_history	15	0	0	0	0	0
clone	21	0	0	2	0	0
contig	36	3	0	0	0	0
contig_equiv	13	0	0	0	0	0

Table Name	Colomns count evo-lution	Data Type change	Number of roll back	Primary key changes	Foreign key changes	Index changes	Table evolu-tion rate
analysis	9	3	0	0	0	0	0,01
analysis_history	15	0	0	0	0	0	0,02
clone	21	0	0	2	0	0	0,1
contig	36	3	0	0	0	0	0,05
contig_equiv	13	0	0	0	0	0	0,02

Variables	coefficients	std. Error	sig.
Colomns count evolution	0,005	0,001	0,000
Datatype Changes	0,002	0,002	0,323
Number of roll back	0,082	0,004	0,000
Primary key changes	0,094	0,004	0,000
Index changes	0,020	0,003	0,000

Future developments

Schema Evolution Toolsuite

- Oracle support
- Statistic cross-history
- CVS/SVN spider
- SMO detection in schema evolution (possible thesis work)