Lehrstuhl fuer Systeme und Informationsverwaltung

Prof.Dr.Ing. Klemens Boehm
Hoang Vu Nguyen
Marco Neumann

KIT

Karlsruher Institut für Technologie

Functional Specification

WHAT

WareHouse Analyzing Tool

November 19, 2012

Functional Specification	**Alexander Noe**
Design	**Jonathan Klawitter**
Implementation	**Anas Saber**
QA / Testing	**Niko Moraitakis**
Final	**Lukas Ehnle**

About this document

General Information

PSE is a mandatory course for students of Bachelor Informatik in the Karlsruhe Institute of Technology (KIT). Therefor groups of five to six students are formed to write programs of 'medium size' (about 5000 Lines of Code). The task given in this course consists of analyzing and visualizing the server log of a database.

This specification is the one of PSE Group #10 in the winter semester 2012/13.

What this specification does

The purpose of this document is a outline of the functional specifications and requirements for the WHAT application. Is tries to give a complete and exact model of the future system. In the course of this it wants to give the developers answers to all possible question concerning what should be implemented.

What this specification does not

This specification does not give any information on how the system should be implemented. Also it doesn't contain any project planning or deadlines.

A 'Living Document'

Software development is a dynamic process. So with proceeding development the content of this specification will change continuously.

Skyserver

The concept of the system described in this specification should work with any database storing it's query-log. The SkyServer description will function as example and testing reference for this system. SkyServer is one of the biggest databases for astronomical data.

About us

Because no one of us speaks English as his first language, we can't guarantee that our documentation doesn't contain simple or incorrectly used language. Our focus lies more on having easy-to-understand and correct documentation than on perfect English by a native speaker.

If you have any questions or comments regarding this document feel free to send us an E-Mail.

Contents

1 Goals

With this program the user should be put in the position to visualize the prepared data of queries, made against his data base.

To structure the criteria the system has to fulfill, it is divided into three parts for this and other sections.

- Web page

- Parser

- Analyzer

See 8.5 for a overview of this parts and their relationships. In the following their specific goals and criteria are described.

1.1 Web page

1.1.1 Core criteria

- The web page is the graphical interface for users and administrators.

- Users can choose the charts, they want to see and also the variables and scales for those.

- The web page provides help for how to use the diagrams and variables.

- A admin-login provides administrative rights, which are needed to use the parser.

1.1.2 Optional criteria

- The language of the website can be changed into different languages. (e.g. German)

- The administrator gets the ability to handle the data warehouse and load new log-files.

- A little history of the last requested charts is stored and viewable.

1.1.3 Exclusion criteria

- The web page does not allow normal users to load new data into the data warehouse.

- The web page does not provide statistic tabular.

1.2 Parser

1.2.1 Core criteria

- The Parser is able to perform the ETL-process on CSV-formatted log files (from Skyserver). This means he extracts the data he needs from the files, transforms them and loads them

into the data-warehouse. Dimensions and measures are specified in 8.6.

- The Parser will recognize invalid logs and won't add them to the data-warehouse. Every log with a mistake won't be accepted, because an error-message is not as bad as a corrupted data-warehouse.

- The Parser will be fed with log files from the administrator via web page.

1.2.2 Exclusion criteria

- This parser is only able to operate on CSV-formatted logs from SkyServer. It can neither read logs in another formats nor logs from another source.

- There is no way to avoid using this Parser when adding data to the warehouse. This can stop corrupting the warehouse to guarantee correct data in the warehouse.

- The Parser doesn't correct mistakes in the log file.

1.3 Analyzer

1.3.1 Core criteria

- The analyzer is the gate to the data warehouse. It extracts the specific data, needed for the diagrams, passing them to the java-script front-end.

- The analyzer can take filtered information via web page from the user to use only certain data for the charts.

- The charts that are supported are at least:
  - scatter plots
  - histograms
  - bubble charts

1.3.2 Optional criteria

- The analyzer will support more different chart-types, especially the combination of a histogram and a scatter plot.

- The analyzer does a little bit of data mining. This function will be called Niko's data mining.

2 Usage

2.1 Applications

Let's say you have some data in a database(who hasn't, right?). Let's say you don't feel like wasting your only life and what to see what parts of that data ara important, fast. Let's say you don't have the time, incination, or even (gasp!) technical background to do so yourself.

Fear not. WHAT is exactly what you are looking for, and more! With features such as drawing of scatterplots, histograms, bubble charts* and other obscure charts you have never heard of before, statistics about the importance of your tables*, automatic reports of* of interesting correlations*, what is the right tool for your job!

As a concrete implementation, all this will be provided for the SKYSERVER logs dataset, but other datasets may be used as well, with minor or not as minor modifications.

(* indicates that this is not a core feature, may be included later in the project, or not at all.)

2.2 Target groups / Audience

The target group of this application are people that want to analyse and visualize the amount of queries made against their database - or like in our case - against the SkyServer.

No matter if operating on a private or a public database, this implies

- the people that run the database wanting to optimize the access to their data,
- people that are interested from where the database is used,
- people knowing or using the database, wondering for what and when other people use the database,
- people new to the database, wanting to know, what may be of interest on the database.

To summarize briefly, WHAT will be of interest for many people, whether running the database, making market research or just all the folks loving statistics and diagrams.

This is a technical audience.

That said, this does not prevent the (web) user interface from being functional, usable and prettier than what you would expect a group of computer science students to design.

2.3 Operating conditions

The program is mainly used as a website, with the primary difference being that the server has to be started if the capacity for it to run all the time on a dedicated machine doesn't exist. The program needs a server to run.

If a dedicated server exists, the program can be used from anywhere with a decent network

connection with the server.

If not, the program can still be run on the same computer as the server (on localhost), but the server will have to be started first.

3 Operating environment

Whereas the programm and the data warehouse run on a sever, the access to it will be via web browser on a workstation computer. The web page making this access possible needs a hosting web server too.

3.1 Software

- The server hosting the data warehouse needs MySQL
- The server hosting the web page needs a recent version of th Java Runtime Environment / Java VM

This software is required on the workstation computer:

- Latest version of a modern browser
  - Chrome or Firefox will be supported
- JavaScript (enabled)

3.2 Hardware

The server has to be fast enough to support all clients. This depends on the expected number of clients. Most computations will be done on the server.

The client needs to be fast enough to visualize the data received from the server. The required hardware thus depends greatly on the amount of data that needs to be visualized. That said, any recent computer, with for example, an Intel® Core™2 Duo CPU E8400 @ 3.00GHz $\times$ 2 processor, 4GB(2x2GB) of 667Mhz DDR2 SDRAM, running Ubuntu Linux 12.10 Quantal Quetzal, as a point of reference, should be able to visualize about 50.000 data points on a scatterplot with ease.

3.3 Orgware

The server needs to be able to connect to the client with a reasonable latency. Also it has to be set up before the client will be able to connect to it.

To fulfill optional requirements, the programm musst be able to run queries against the origin database.

3.4 Product interfaces

To fulfill optional requirements .csv files have to be imported.

4 Functional requirements

4.1 Main functions

This functions are required to fulfill the core criteria.

General

 /F10/ Provide access via web page

 /F20/ Provide option for chart types

 /F30/ Show diagrams

 /F40/ Show histograms

 /F50/ Provide option for the two variables of the histogram

 /F60/ Provide option for the interval or selection of the x-axis variable

 /F70/ Show scattered plots

 /F80/ Provide option for the two variables of the scattered plot

 /F90/ Provide option for the intervals or selections of the two variables

/F100/ Show bubble charts

/F110/ Provide option for the three variables of the scattered plot

/F120/ Provide option for the intervals or selections of the three variables

/F130/ Show information about chart types

/F140/ Show information about selectable variables

Administrator specific

This functions are required for the administrative business.

/F150/ Provide access via web page with administrator rights

/F160/ Provide the opportunity to pass log-files to the parser

Parser specific

The following functions specify the parses functionality. Thereby /F180/ to /F200/ specify the function /F170/.

/F170/ Extract specific data from the log-files

/F180/ Extract access database, access time and user information (dimensions)

/F190/ Extract number of rows, elapsed time, busy time (measures)

/F200/ Extract type of data requested from the `where`-part

/F210/ Transform the data to fit into the data warehouse schema

/F220/ Load the data into the data warehouse

Analyzer specific

/F230/ Run specific queries against data warehouse

/F240/ Transform received data serving the web page

4.2 Extending functions

To fulfill the optional goals the following functions are required.

General

/F250/ Select language on web page

/F260/ Show bubble map

/F270/ Show combination of histogram and scattered plot

/F280/ Show other diagrams and charts

/F290/ Show history of the 5 last requested charts

Administrator specific

/F300/ Provide the opportunity to initialize the data warehouse

/F310/ Provide the opportunity to request new log-files for specific time intervals from the database

/F320/ Provide the opportunity to clean the data warehouse

5 Data

The product data are divided in two groups, static data, which are delivered with the program and in the data warehouse data, which come from the Skyserver server logs, are parsed and loaded into the Data-Warehouse. If optional criteria are fulfilled they can be replaced by clearing the data warehouse and loading new data via web page as administrator.

5.1 Static Data

/D10/ Language files

/D20/ Manual

/D30/ Source Code

/D40/ Documentation

/D50/ Graphics for GUI

/D60/ HTML backbone

/D70/ Javascript files

/D80/ Stylesheet files

5.2 Data-Warehouse data

/D90/ Database accessed by Users

/D100/ Server accessed by Users

/D110/ City of Users

/D120/ Country of Users

/D130/ Time of SkyServer access

/D140/ Number of rows accessed

/D150/ Elapsed time of access

/D160/ CPU busy time of access

6 Nonfunctional requirements

6.1 First Whatever

/NF10/ blabla

/NF20/ tadaa

6.2 Next Whatever

/NF30/ 30!

/NF40/ 40!

7 Global test cases

is a set of conditions and variables under which a test cases will determine whether an this Application is working correctly or not.

7.1 Test case

is the description of Workflow of this Application and depend on the Functional requirements. This Workflow must successful and correctly with alle these points.

 /T10/ Start internet browser (/F10/)

 /T20/ Type the web address - say pse10-analyzer.edu - in browser's address bar (/F10/)

 /T30/ Call the Home Page with a Table Bottun /F20/

 /T40/ Select a Type of chart /F30/

 /T50/ Call the diagram Page

 /T60/ Call histogram /F40/

 /T70/ Call the option for tow variables of the histrogram /F50/

 /T80/ Call the option for the interval or selection of the x-axis variable /F60/

 /T90/ Call scattered plots /F70/

/T100/ Call the option for tow variables of the histrogram /F80/

/T110/ Call the option for the interval or selection of the x-axis variable /F90/

/T120/ Call scattered plots /F70/

/T130/ Call the option for tow variables of the scattered plots /F80/

/T140/ Call the option for the interval or selection of tow variables /F90/

/T150/ Call bubble charts /F100/

/T160/ Call the option for three variables of the scattered plots /F110/

/T170/ Call the option for the interval or selection of tow variables /F120/

/T180/ Call the information about chart types (/F130/)

/T190/ Call information about selectable variables (/F140/)

/T200/ Leave the web page

8 Models

8.1 User stories

As a user, want to get some Information from skyserver to analyze the database of this server and want to know, how many user or client that used this server from wich Countries and in wich time (Month and Year) .

To analyze this Information in best way, i would like to get Visualizing and Statistically of these Information. That's mean, i can get some diagram with axises like histogram.

Steps to descrip how can a normal user use this Application:

1 Start internet browser

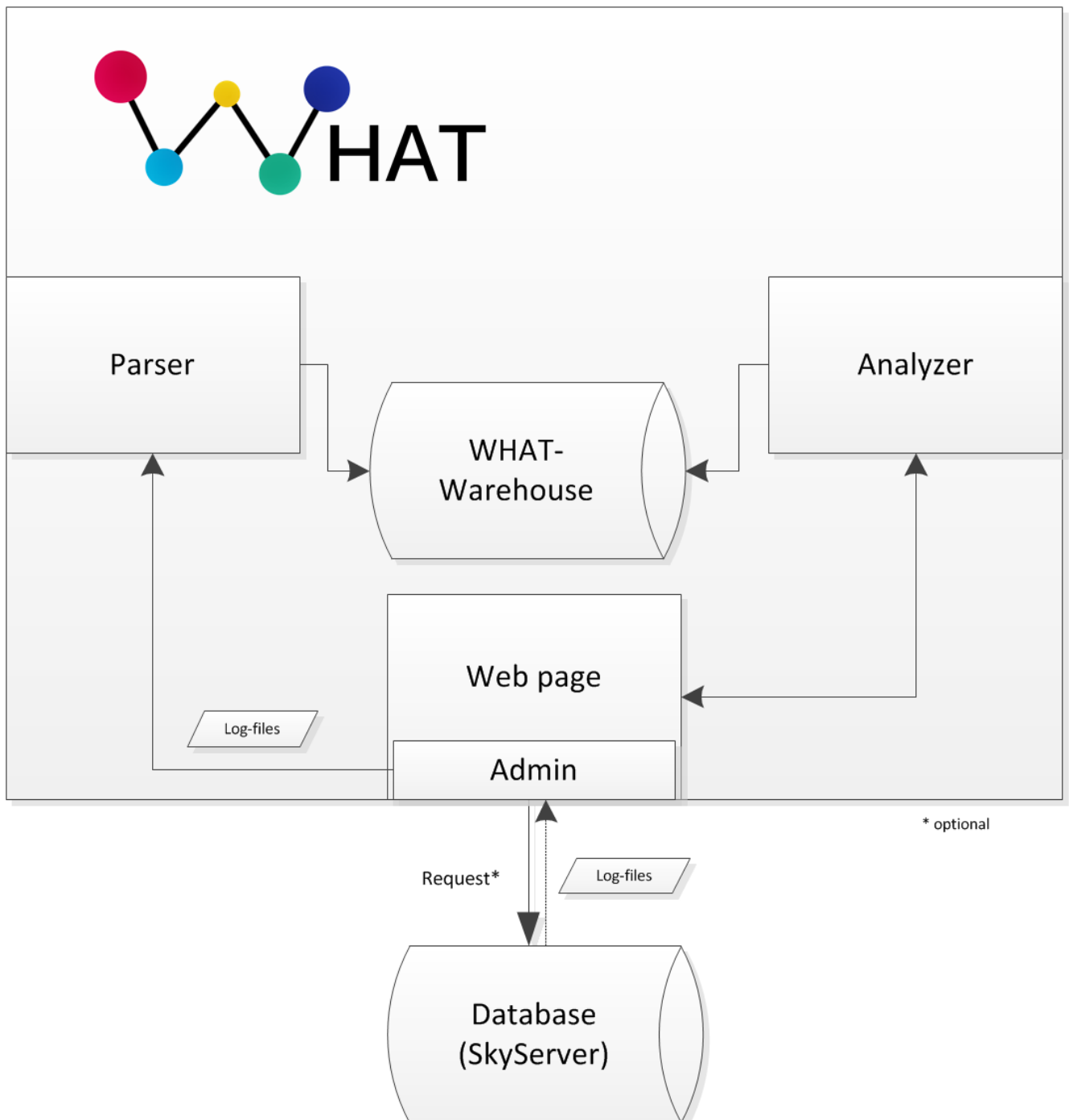2 Vist www.say pse10-analyzer.edu web page

3

8.2 Object models

8.3 Dynamic models

8.4 Web interfaces

8.5 Overview

This picture describes the generell concept of WHAT.
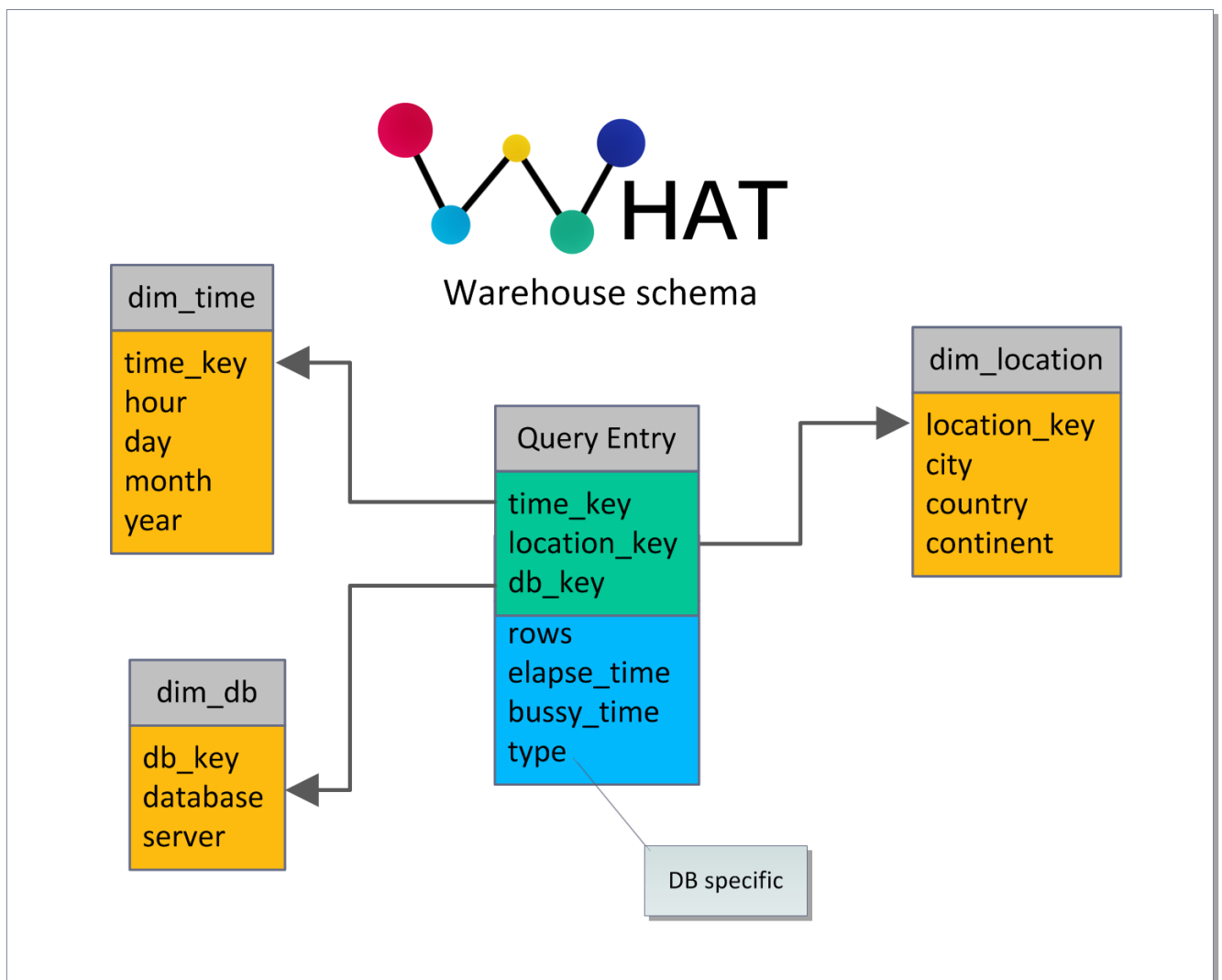
8.5.1 Parser

Parser blabla

8.5.2 Analyzer

8.5.3 Web page

8.6 Warehouse schema

The warehouse may use a star schema.

9 Development Environment

Operating System

- Windows 7
- Fedora 17
- Whatever Niko uses - EDIT THIS, NIKO!

Object Models

- Microsoft Visio

Version Control

- Git

Miscellaneous

- Github (Hosting of Repository, Issue Tracking)
- Travis (Integration)
- LaTeX (Documents)
- Gradle (Building)

10 Glossary

.. ..

KIT is a technological university in Karlsruhe, Baden-Württemberg, Germany (Karlsruhe Institute of Technology)

SkyServer is a huge database for astronomical data. It contains pictures and other information concerning astronomy and tries to form a 'map of the universe'. Web page ofSkyServer SkyServer.