

Facilitating Knowledge Discovery in Large Archives of Astronomical Spectra using Distributed Cloud-based Engine

Petr Škoda

Astronomical Institute, Czech Academy of Sciences Ondřejov

Jakub Koza, Andrej Palička, Jiří Nádvorník
and Tomáš Peterka

Faculty of Informatics, Czech Technical University, Prague

Supported by grant GAČR 13-08195S

Astroinformatics 2015
Dubrovnik, Croatia, 5th October 2015

Concept of scientific „CLOUD“

ITERATIVE REPEATING of SAME computation (workflow)

Global non-linear optimization (spectra disentangling)

Synthetic spectra (various elements, wavelength-ranges)

Machine Learning (almost all methods)

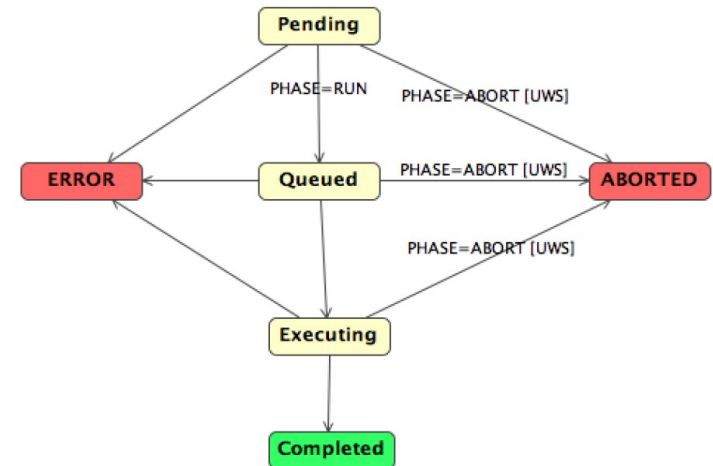
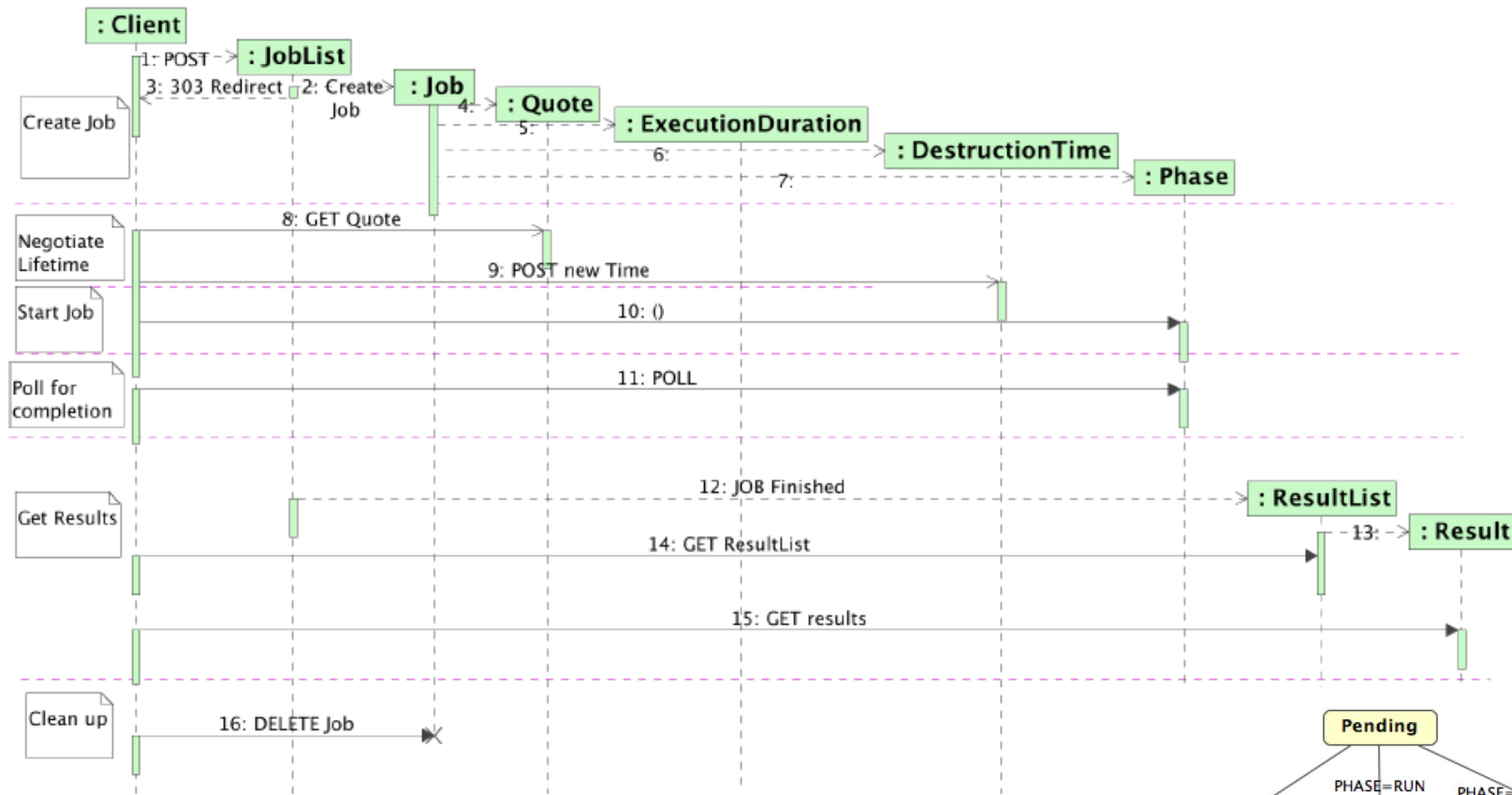
LARGE stable INPUT data + small changing PARAMS

Many runs on SAME data (tuning required)

Graphics visualization from postprocessed output (text) files

Using WWW browser - supercomputing in PDA/mobil

IVOA Universal Worker Service (UWS)



VO-CLOUD Architecture

VO-CLOUD (former VO-KOREL)

Distributed engine

MASTER (frontend)

Database of users and their experiments

Visualization

Scheduling

Load balancing

WORKERS (backend)

Computation [+ output for visualization]

CZECH TECHNICAL UNIVERSITY IN PRAGUE
FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF SOFTWARE ENGINEERING



Bachelor's thesis

**VO-KOREL, server for astronomical cloud
computing**

Lumír Mrkva

Supervisor: RNDr. Petr Škoda, CSc.

18th May 2012

CZECH TECHNICAL UNIVERSITY IN PRAGUE
FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF SOFTWARE ENGINEERING



Bachelor's thesis

**Design and implementation of a
distributed platform for data mining of big
astronomical spectra archives**

Jakub Koza

Supervisor: RNDr. Petr Škoda, CSc.

12th May 2015

CZECH TECHNICAL UNIVERSITY IN PRAGUE
FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE



Bachelor's thesis

Application of Random Decision Forests in Astroinformatics

Andrej Palička

Supervisor: RNDr. Petr Škoda, CSc.

12th May 2014

CZECH TECHNICAL UNIVERSITY IN PRAGUE
FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF THEORETICAL COMPUTER SCIENCE



Bachelor's thesis

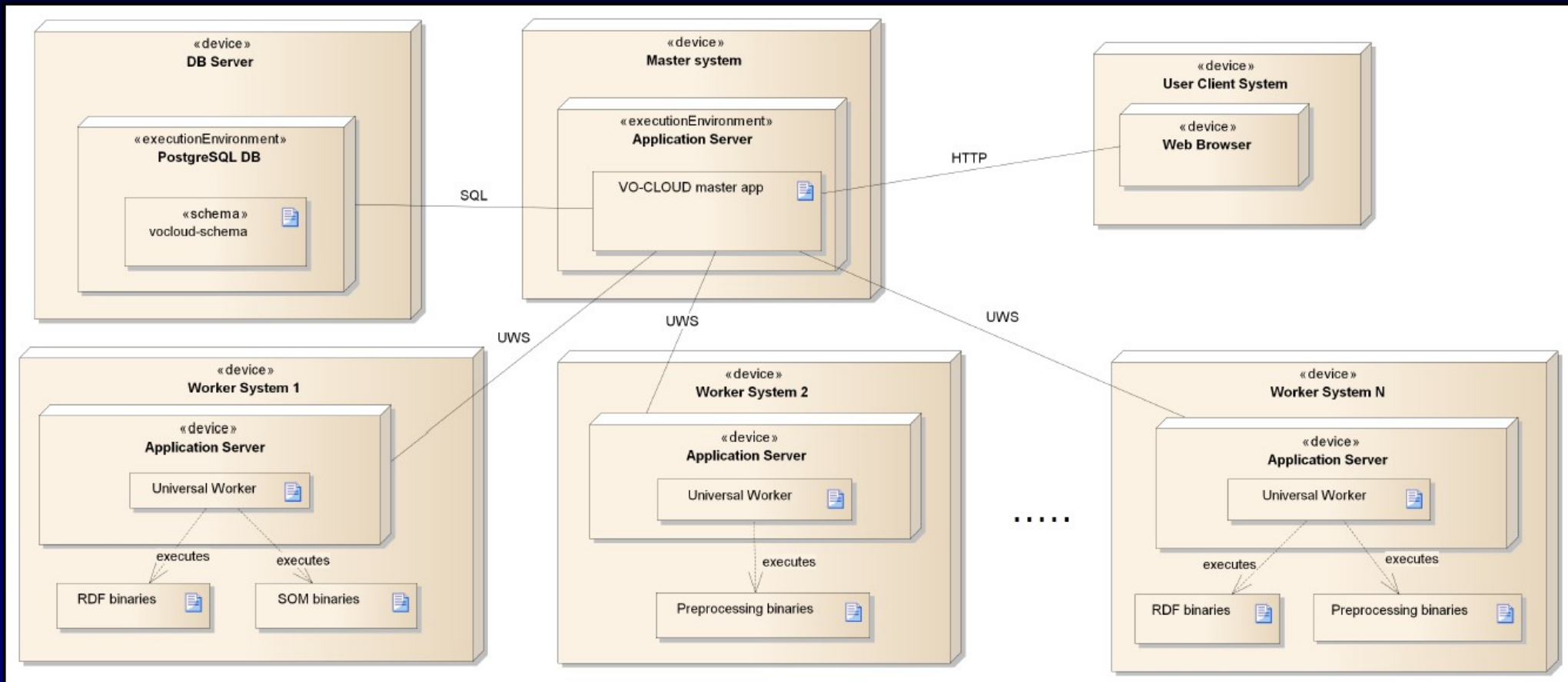
Application of Self-Organizing Maps in Astroinformatics

Lopatovský Lukáš

Supervisor: RNDr. Petr Škoda, Csc.

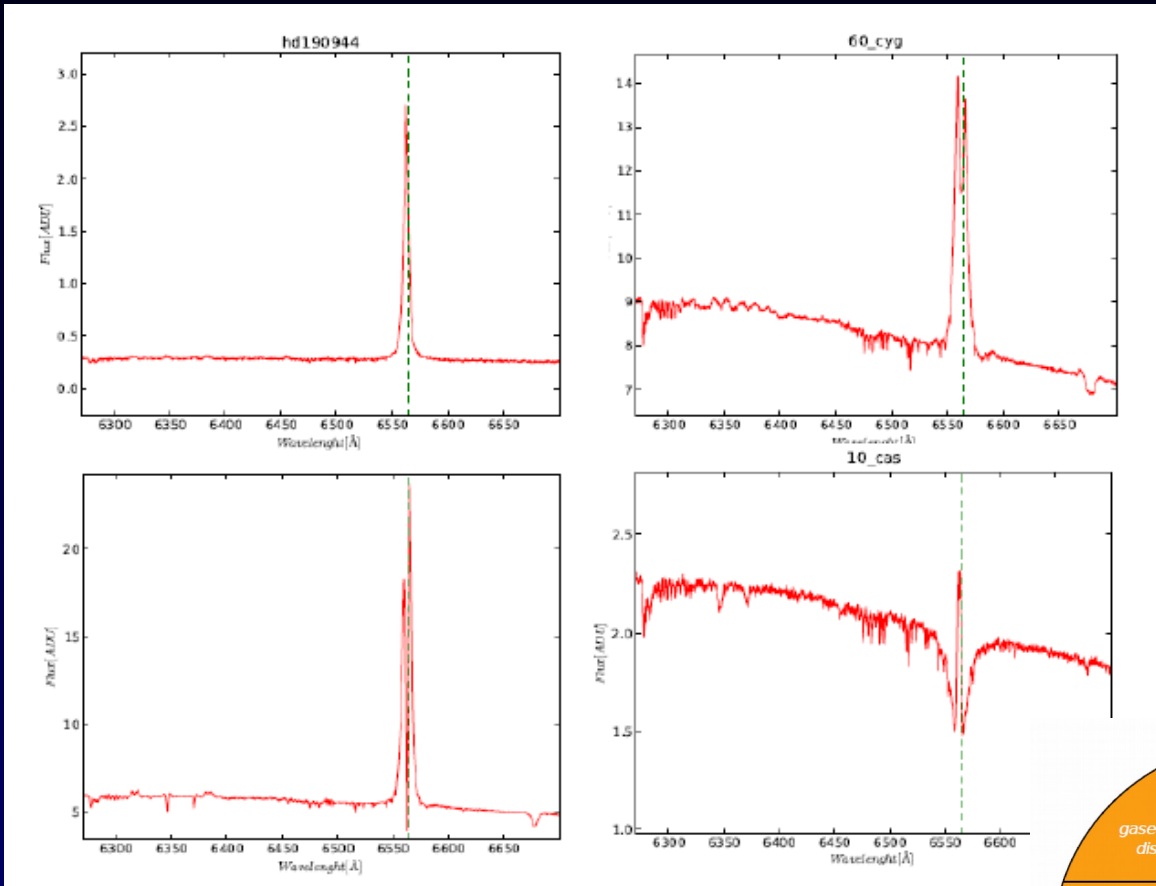
14th May 2014

VO-CLOUD Deployment Schema



Machine Learning of Spectra

Use case: ML of spectra profile of H α line (Be stars)

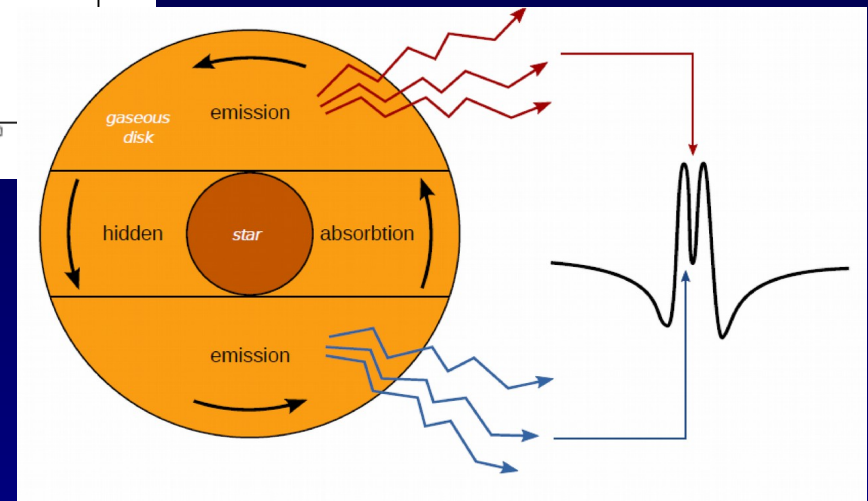


Be stars

Disk or envelope

Rotates, Hot

Origin ?????



Machine Learning of Spectra

Science case

Ondřejov 2m Perek Telescope – 1700/10 000 spectra

PRE-PROCESSING

Normalization to continuum, Cutout (SSAP+DL)

Rebinning (same wavelegth points) + Renormalization [-1,+1]

(Reduction of dimensionality (wavelets, PCA, LLE...))

Produces **feature vectors** in CSV (same length, dimensions)

MACHINE-LEARNING

Unified wrapper running multiple applications - same call

Name-of-wrapper + parameters (json) – method as param

VISUALIZATION

JavaScript (dygraph, HighCharts)

Sources of Spectra

Getting spectra + store

(restricted access – big files)

Files

UPLOAD from given local directory (recursive)

DOWNLOAD by http + index, FTP (recursive)

VOTable

UPLOAD VOTable (e.g. prepared in TOPCAT - meta)

REMOTE VOTable

SSAP query + Accref

+ DataLink (PUBDID + mime)

SAMP control - send to SPLAT

Machine Learning of BIG Archive?

Idea – 4.1 mil of LAMOST spectra (4 mil. in SDSS)

NOT Upload data by user (VO compatible archive)

Driven by SPECTRA LIST (votable obtained by TAP ?)

Workers on same **hi-speed network** as archive

Calling SSAP + DL always (client on GRID worker ?)

Pre-cache ?

Compute feature vectors – store for whole experiment ?

PERSISTENT STORAGE - network FS ?

Visualisation - needs input data (spectrum), lists from class

CZECH TECHNICAL UNIVERSITY IN PRAGUE
FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF THEORETICAL COMPUTER SCIENCE



Master's thesis

**Machine Learning in Astroinformatics
Using Massively Parallel Data Processing**

Bc. Tomáš Peterka

Supervisor: RNDr. Petr Škoda, CSc.

14th May 2015

Deep Learning

Caffe + Big Data Layer

GPU /CPU switch

Will be part of VO-CLOUD
soon

Source Code

<https://github.com/vodev/vocloud>

<https://github.com/vodev/vocloud-preprocessing>

<https://github.com/vodev/vocloud-som>

<https://github.com/vodev/vocloud-RDF>

<https://github.com/vodev/vocloud-deeplearning>

DEMO

<http://vocloud-dev.asu.cas.cz/vocloud2>