



Wendelin Exanalytics 去“IOE”的警察大数据

2014-06-11 – 北京



MariaDB



www.wendelin.io



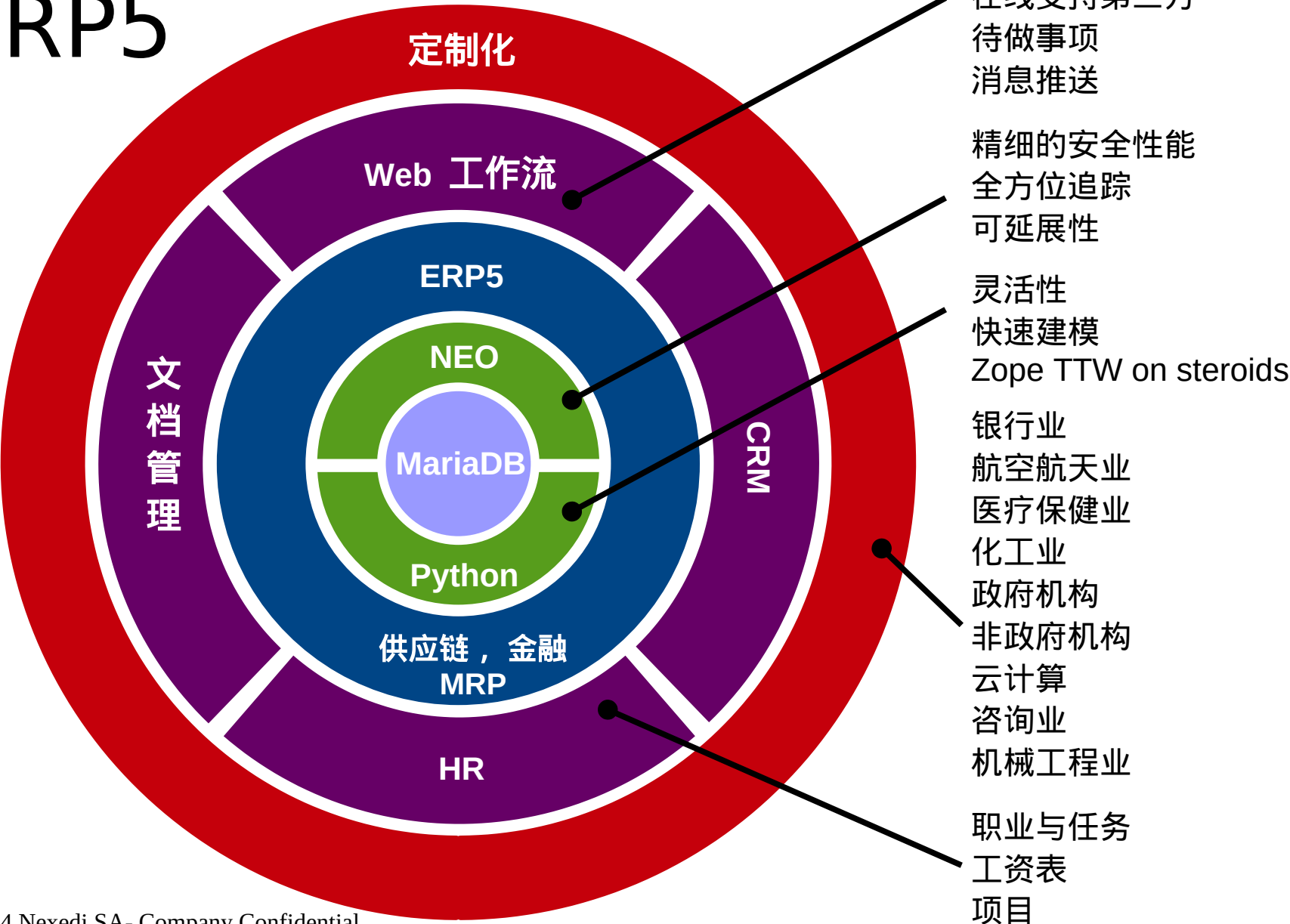
日程表



我们的背景：ERP5 与 SlapOS

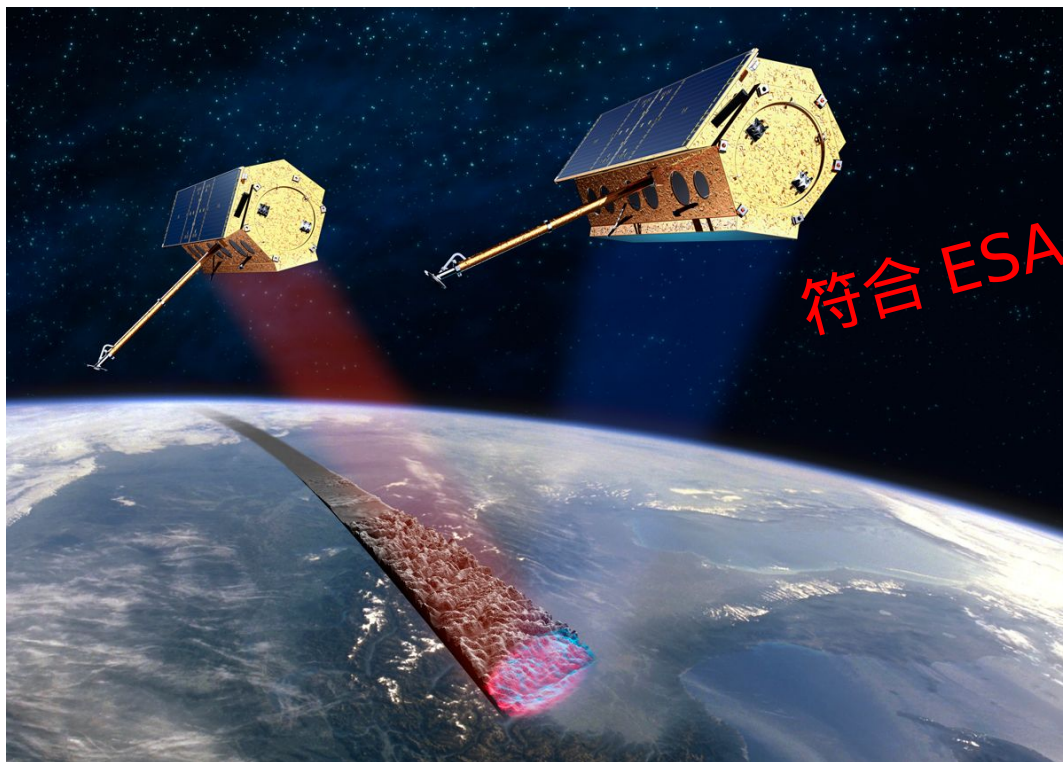
- **我们的未来：Wendelin Exanalytics**
- **我们的挑战：out-of-core**
- **实地应用：Wangrin 自动化**

ERP5



Terra-SAR X 卫星

卫星图像的销售管理及生产



符合 ESA 标准 (ECSS)



可访问空客 (Airbus)
合作伙伴及分销商
与 DLR 对接
(德国航空局)

« With ERP5, our partners all over the world can access our infrastructure and order online with complete security(通过使用ERP5, 我们全球的合作伙伴都可以访问我们的基础结构并在完全安全的环境下进行在线订购)» Ralf Duering

象牙海岸警务云

可恢复性 IPv6 云



通用中国产硬件
无 IBM 无 EMC 无 Oracle
100% 开源软件



IPv6 Mesh 网络
可恢复性数据存储
分布式基础架构

“Resiliency and auditability are required for government Cloud. We chose SlapOS to gain auditability through open source software and resiliency through decentralized architecture (可恢复性及可审核性是政府云所需要的。我们选择SlapOS 通过开源软件来获得可审核性，通过分布式基础架构获得可恢复性)” - Stéphane Konan – DITT MEMI

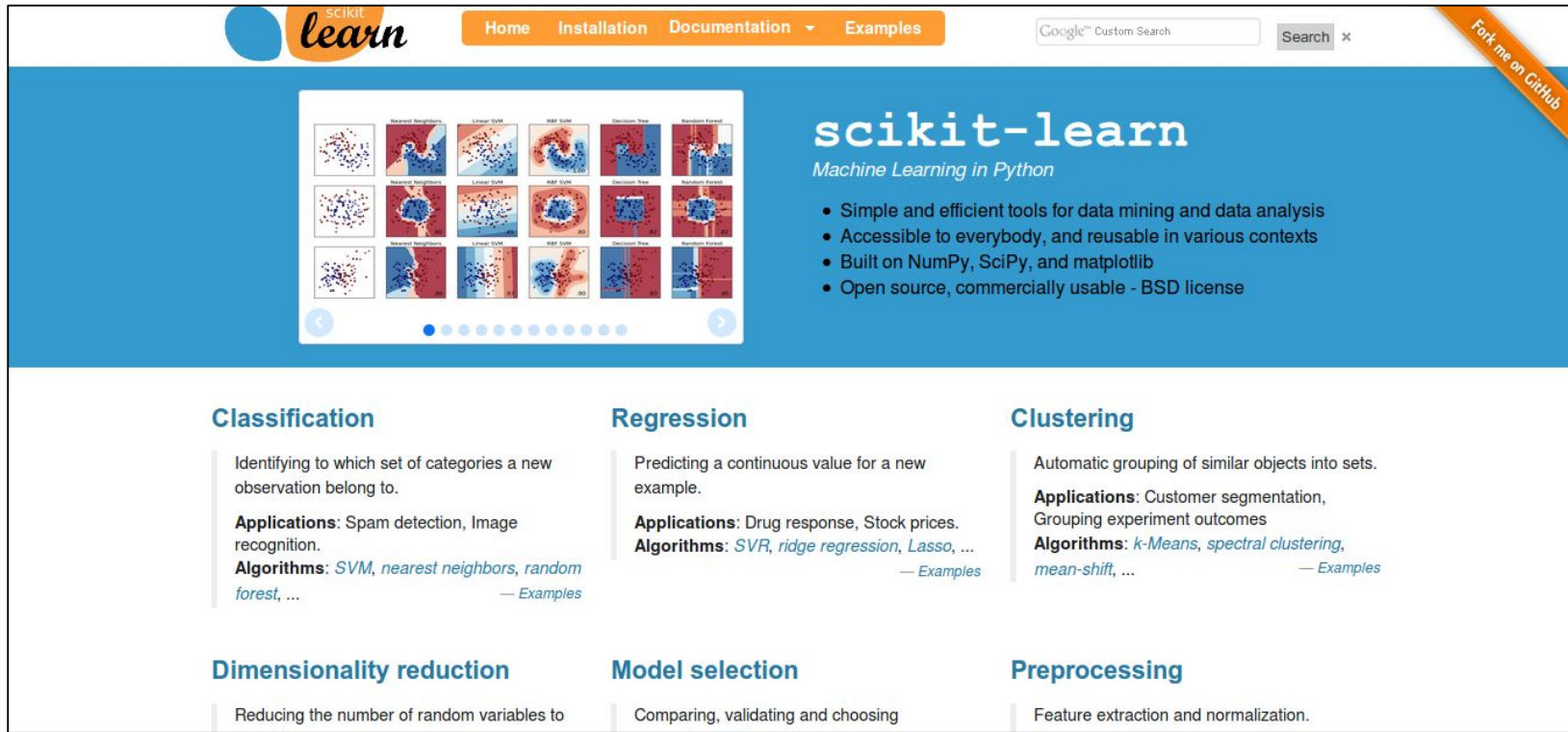
日程表



- 我们的背景：ERP5 与 SlapOS
- 我们的未来：Wendelin Exanalytics
- 我们的挑战：out-of-core
- 实地应用：Wangrin 自动化

运用最完美的解析学

scikit-learn.org



The screenshot shows the scikit-learn website homepage. At the top, there is a navigation bar with links for Home, Installation, Documentation, and Examples. A search bar is also present. The main header features the scikit-learn logo and the tagline "Machine Learning in Python". Below this, a grid of 12 small plots illustrates various machine learning models. To the right of the grid, a list of bullet points highlights the library's features: simple and efficient tools for data mining and data analysis, accessibility to everybody, built on NumPy, SciPy, and matplotlib, and being open source with a BSD license. A "Fork me on GitHub" button is located in the top right corner. The main content area is divided into six sections: Classification, Regression, Clustering, Dimensionality reduction, Model selection, and Preprocessing. Each section provides a brief description, applications, and algorithms.

scikit-learn
Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification
Identifying to which set of categories a new observation belong to.
Applications: Spam detection, Image recognition.
Algorithms: SVM, nearest neighbors, random forest, ... — Examples

Regression
Predicting a continuous value for a new example.
Applications: Drug response, Stock prices.
Algorithms: SVR, ridge regression, Lasso, ... — Examples

Clustering
Automatic grouping of similar objects into sets.
Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, ... — Examples

Dimensionality reduction
Reducing the number of random variables to

Model selection
Comparing, validating and choosing

Preprocessing
Feature extraction and normalization.

Inria
INVENTORS FOR THE DIGITAL WORLD

TELECOM
ParisTech

AWeber
COMMUNICATIONS

EVERNOTE



Spotify

Google

cloudera

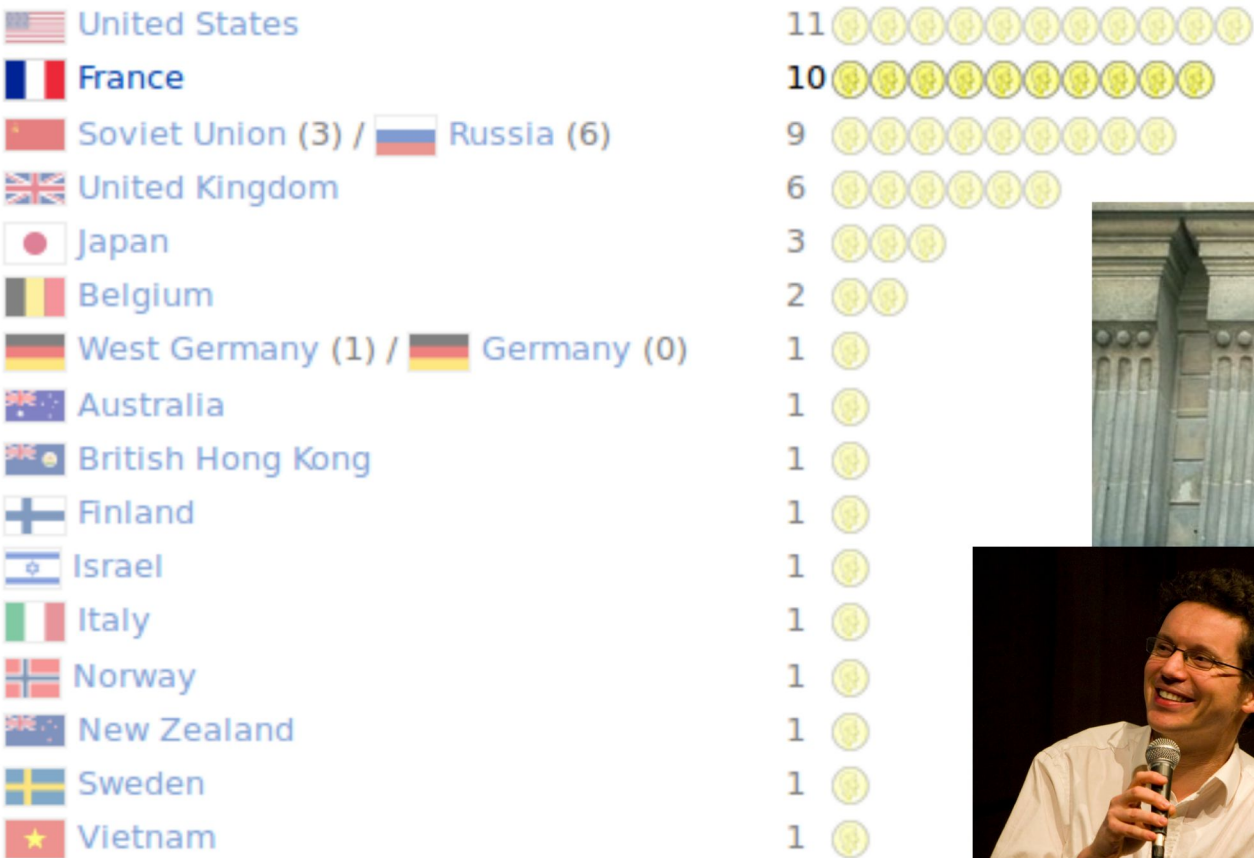


WENDELIN

由最优秀的数学家们设计

http://en.wikipedia.org/wiki/Fields_Medal

Number of Fields Medallists by country [\[edit\]](#)



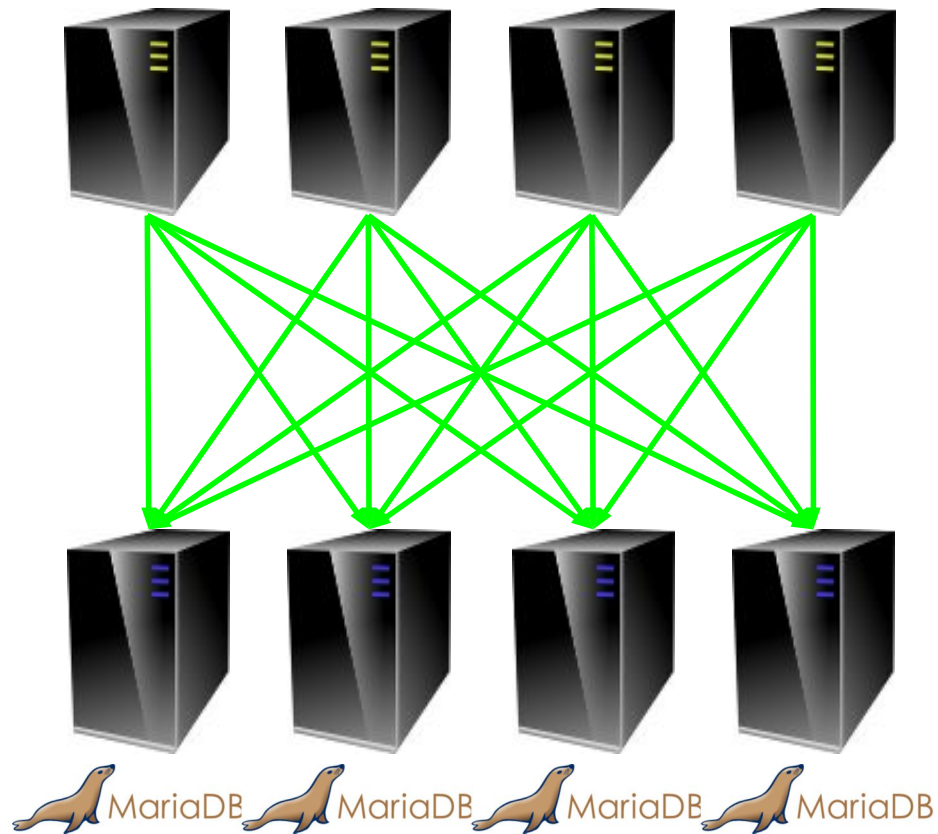
Wendelin Werner

加入分布式存储

neoppod.org



NEO



加入弹性 PaaS erp5.com

```
# Initialize data
data_size = 1000000
server_count = 1000
chunk_size = data_size / server_count
data = array(data_size)

# Process data in parallel on each server (Map Reduce, Batch, etc.)
for server in server_count:
    data.activate().process(server*chunk_size, chunk_size)
```



加入多重云部署 slapos.org

```
0 [buildout]
1
2 extends =
3 # "slapos" stack describes basic things needed for 99.9% of SlapOS Software
4 # Releases
5 ../../stack/slapos.cfg
6 # Extend here component profiles, like openssl, apache, mariadb, curl...
7 # Or/and extend a stack (lamp, tomcat) that does most of the work for you
8 # In this example we only need the dash binary to run a simple "hello world"
9 # shell script.
10 ../../component/dash/buildout.cfg
11
12 parts =
13 # Call installation of slapos.cookbook egg defined in stack/slapos.cfg (needed
14 # in 99.9% of Slapos Software Releases)
15 slapos-cookbook
16 # Call creation of instance.cfg file that will be called for deployment of
17 # instance
18 template
19
20 # Download instance.cfg.in (buildout profile used to deployment of instance),
21 # replace all ${foo:bar} parameters by real values, and change ${foo:bar} to
22 # ${foo:bar}
23 [template]
24 recipe = slapos.recipe.template
25 url = ${:_profile_base_location_}/instance.cfg.in
26 output = ${buildout:directory}/instance.cfg
27 # MD5 checksum can be skipped for development (easier to develop), but must be filled for production
28 md5sum = 1fc461c00e86485bee77a942f39e3c43
29 mode = 0644
30
```

Save



MMC Rus



L'Education change le monde



Wendelin Exanalytics 核心 100% 开源

100% Python

Scikit Learn

数据分析

NEO

分布式存储



ERP5

弹性 PaaS

SlapOS

多重云部署

多重数据中心



Wendelin 的选择

100% 开源

100% Python	NLTK	Natural Language Toolkit U. Texas / Chalmers
	Blaze	完整的 out-of-core 数组 Continuum / DARPA
	Numba / Parakeet	即时编译 / 类型推论 Continuum / DARPA
	Pandas	时间序列处理 DataPad / JP Morgan
	Scikit Learn	
	NEO	
	Fluentd	实时日志收集 Treasure Data / Amazon

Wendelin 用户界面

renderjs.org

Search

Shop:Order 2341
2013/04/02 · 49.02€ · France · Archive

Shop:Order 23412
2013/05/13 · 55.02€ · France · Paymen...

Sven Franck
Client · Last Order: 2013/04/02

Sven Franck

Available Options

Order Inbox
Review and Manage your orders. 12

Clients
Managing and marketing users.

Messages
Messages, special offers, news. 8

Shops
Setup and manage your shops.

Products
Upload and manage products.

Settings

Help

Settings

Report a problem

Sign out

Nexedi © 2013

Menu

Some

20+ Tasks

11 Messages

Home

Search

List

Tasks: 1-7 of 7 records

Criteria

	Company	Last Trade	Trade Time	Change	Prev Close	Open	Stuff		1y Target Est
							Bid	Ask	
<input type="checkbox"/>	GOOG Google Inc.	597.74	12:12PM	14.81 (2.54%)	582.93	597.95	597.73 x 100	597.91 x 300	731.10
<input type="checkbox"/>	AAPL Apple Inc.	378.94	12:22PM	5.74 (1.54%)	373.20	381.02	378.92 x 300	378.99 x 100	505.94
<input type="checkbox"/>	AMZN Amazon.com Inc.	191.55	12:23PM	3.16 (1.68%)	188.39	194.99	191.52 x 300	191.58 x 100	240.32
<input type="checkbox"/>	ORCL Oracle Corporation	31.15	12:44PM	1.41 (4.72%)	29.74	30.67	31.14 x 6500	31.15 x 3200	36.11
<input type="checkbox"/>	MSFT Microsoft Corporation	25.50	12:27PM	0.66 (2.67%)	24.84	25.37	25.50 x 71100	25.51 x 17800	31.50
<input type="checkbox"/>	CSCO Cisco Systems, Inc.	18.65	12:45PM	0.97 (5.49%)	17.68	18.23	18.65 x 10300	18.66 x 24000	21.12
<input type="checkbox"/>	YHOO Yahoo! Inc.	15.81	12:25PM	0.11 (0.67%)	15.70	15.94	15.79 x 6100	15.80 x 17000	18.16

<<

<

>

>>

Outlier detection

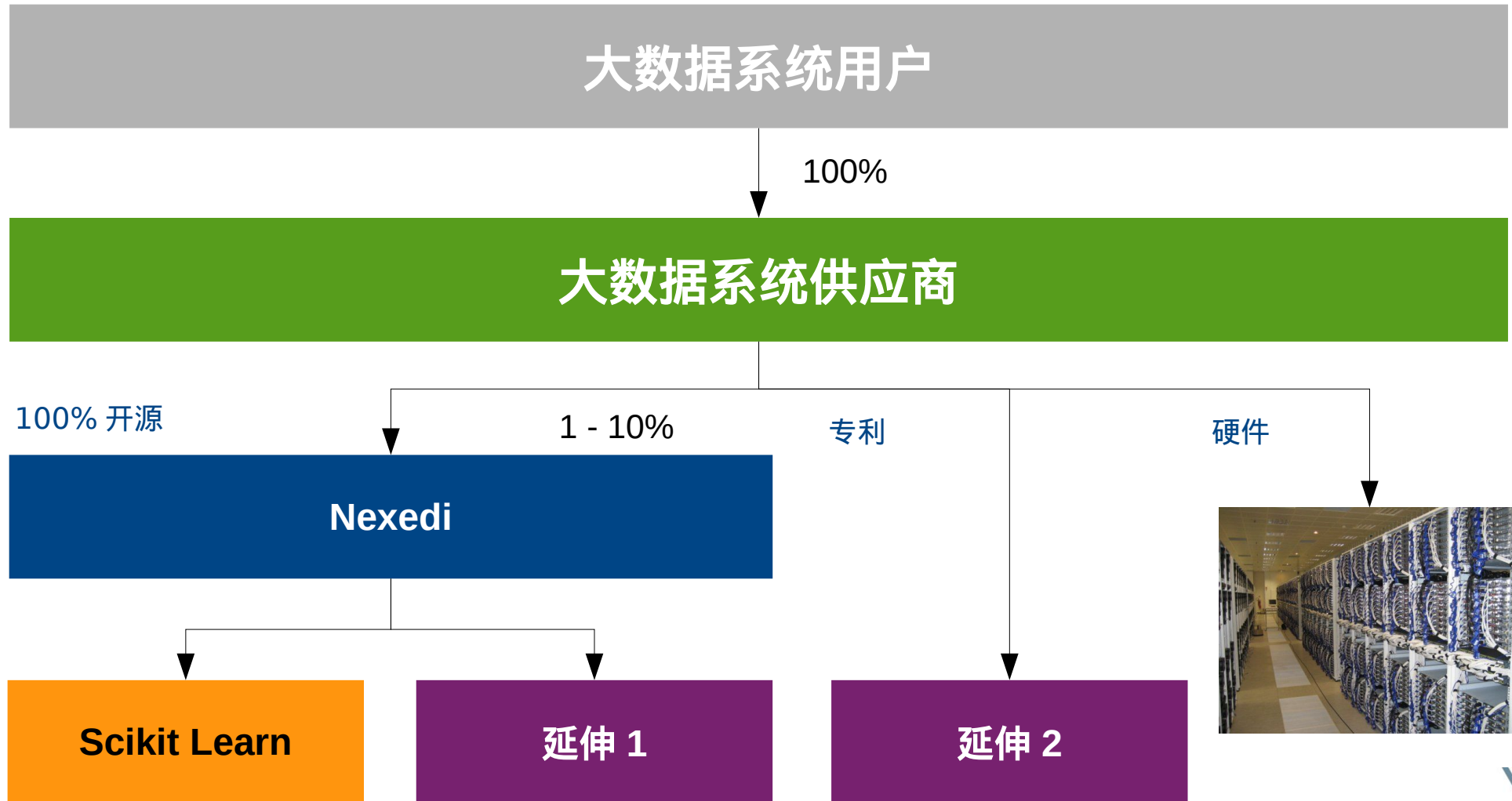
1. One-Class SVM (errors: 8)

LASSO Path

Wendelin 应用

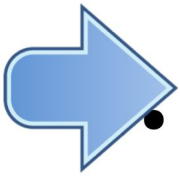
- 干扰侦查
- 欺诈侦查
- 商业及经济预测
- 市场分析
- 媒介分析
- 公共安全
- 脑机接口
- 物联网

商务模式：德国风格 No VC



日程表

- 我们的背景：ERP5 与 SlapOS
- 我们的未来：Wendelin Exanalytics
- 我们的挑战：out-of-core
- 实地应用：Wangrin 自动化



Out-of-core 数组

```
# Numpy
```

```
np.ndarray(shape=(2,2), dtype=float, order='F')
```

```
# Out-of-core data
```

```
np.ndarray(shape=(1e18,2), dtype=float, order='F') 1 Exabyte
```

```
# Full out-of-core
```

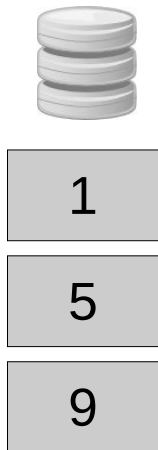
```
np.ndarray(shape=(1e9,2e9), dtype=float, order='F') 1 Exabyte
```

最好的 out-of-core 拓扑学需要依赖于运算法则和方阵
几何学

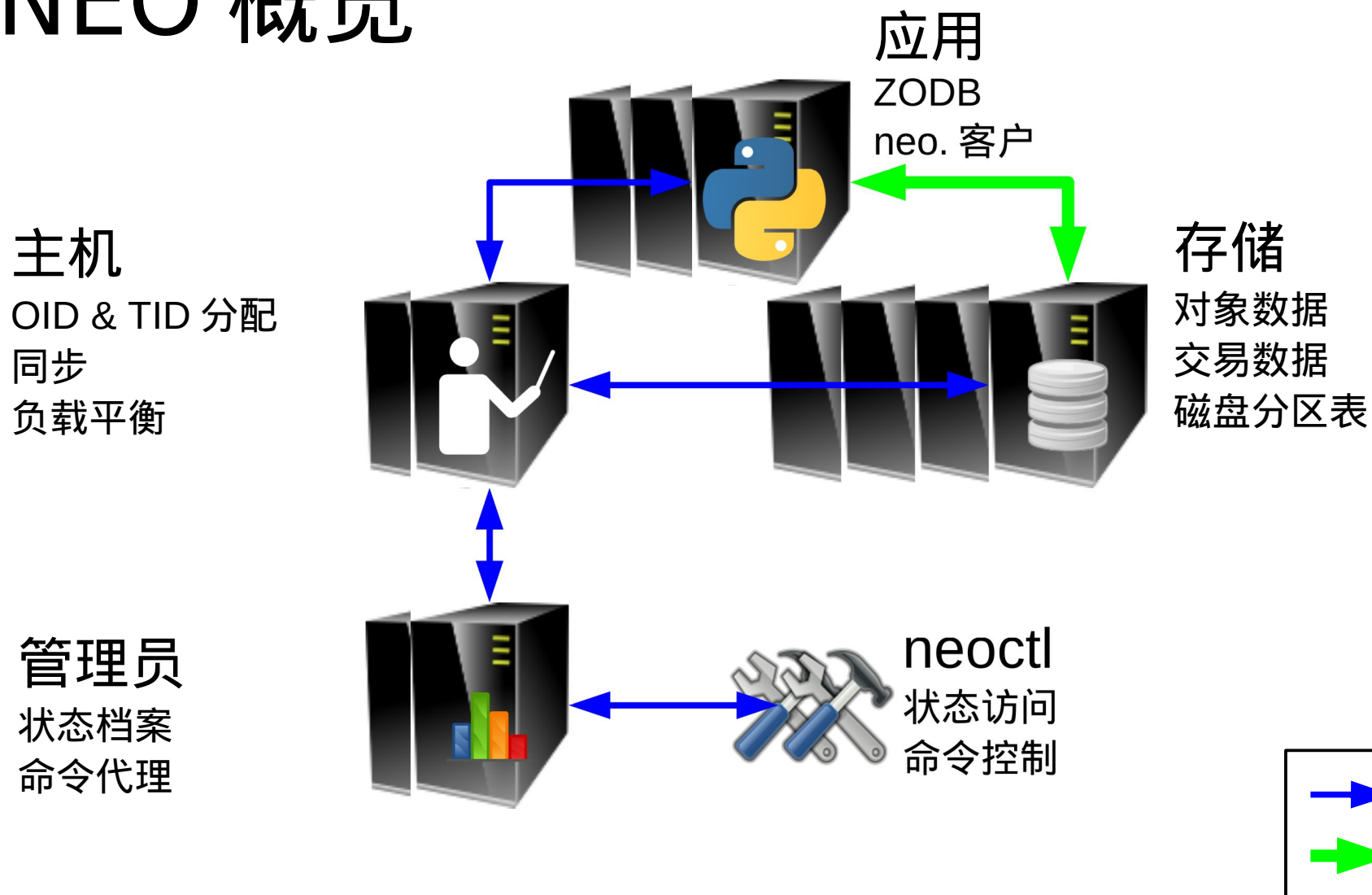
neo.ndarray out-of-core 数据

neo.ndarray

1	2	3	4	5	6	7	8	9	10	11	12
---	---	---	---	---	---	---	---	---	----	----	----



NEO 概览

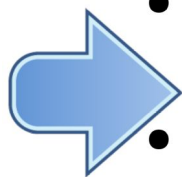


路径

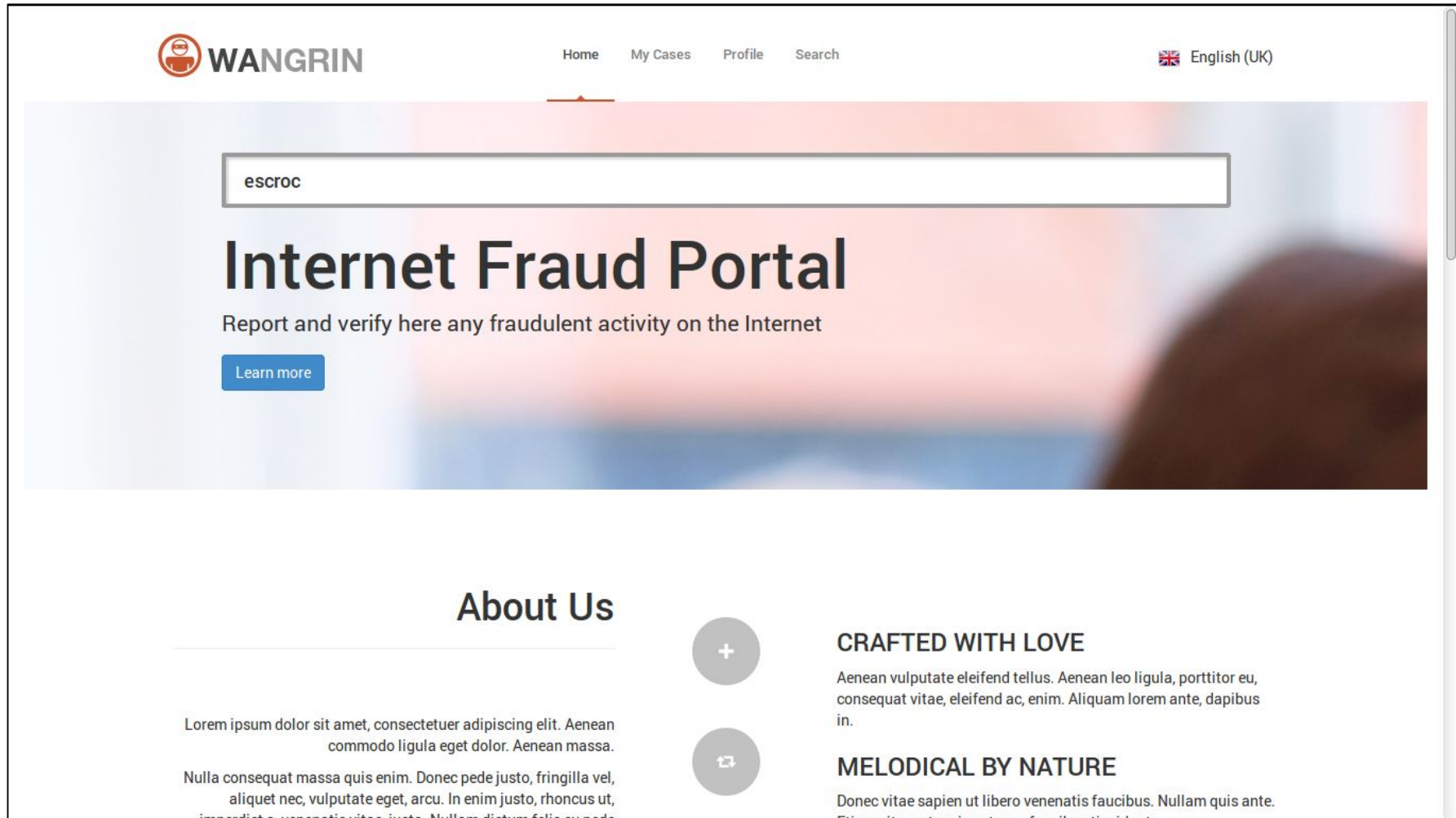
- Q3 2014: 开发者发布 Wendelin
- Q4 2014: 简单的优化 neo.ndarray
- Q3 2015: 完整的优化 neo.ndarray

日程表

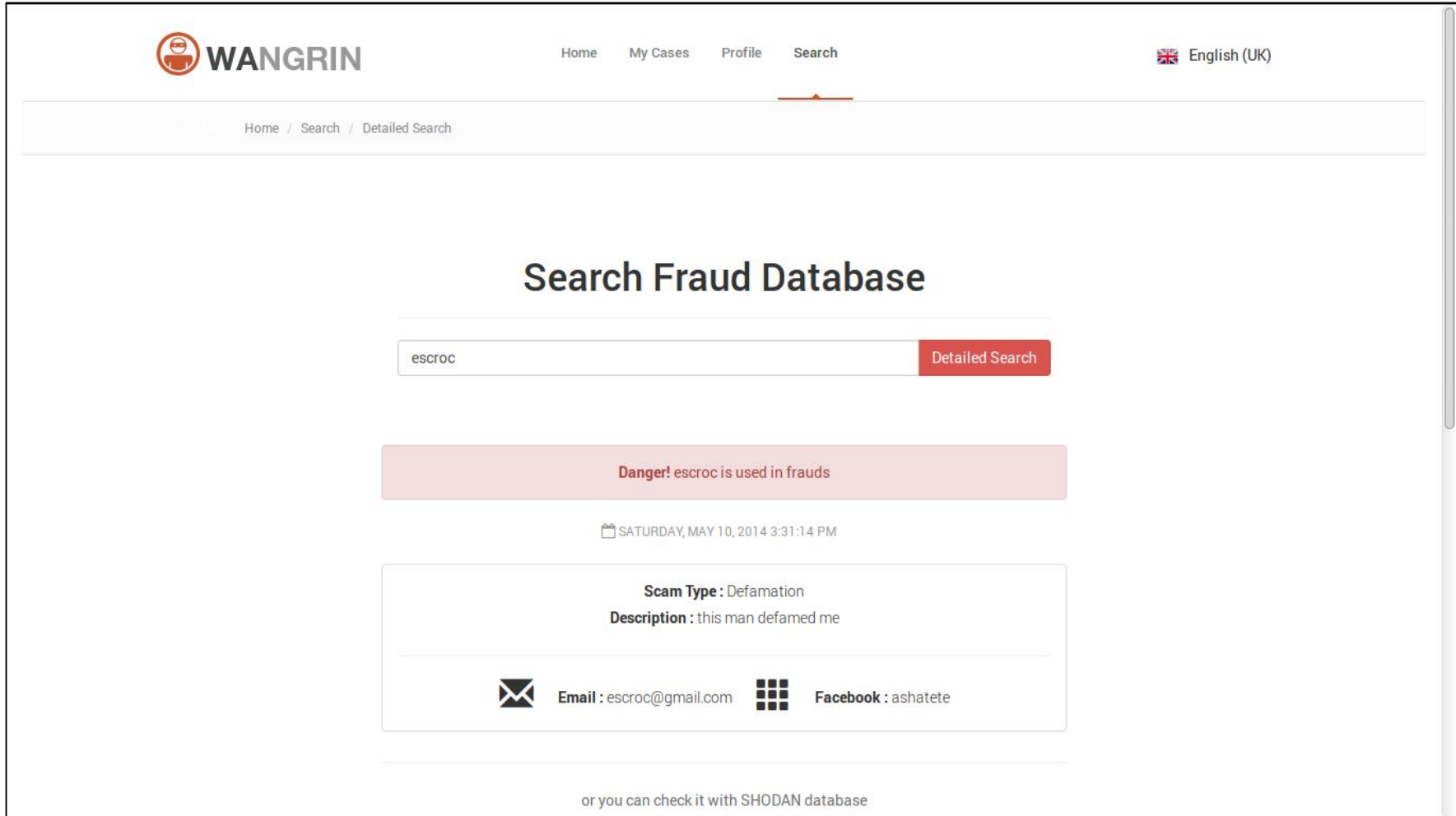
- 我们的背景 : ERP5 与 SlapOS
- 我们的未来 : Wendelin Exanalytics
- 我们的挑战 : out-of-core
- 实地应用 : Wangrin 自动化



wangrin.net – 互联网欺诈门户网



wangrin.net – 互联网欺诈门户网



The screenshot displays the Wangrin website interface. At the top, the logo 'WANGRIN' is on the left, and navigation links 'Home', 'My Cases', 'Profile', and 'Search' are in the center. A language selector 'English (UK)' is on the right. Below the navigation bar, a breadcrumb trail reads 'Home / Search / Detailed Search'. The main heading is 'Search Fraud Database'. A search input field contains the text 'escroc', and a red button labeled 'Detailed Search' is to its right. Below the search bar, a red warning box states 'Danger! escroc is used in frauds'. A timestamp 'SATURDAY, MAY 10, 2014 3:31:14 PM' is displayed. The search results section shows 'Scam Type : Defamation' and 'Description : this man defamed me'. At the bottom of the results, there are icons for email and Facebook, with the text 'Email : escroc@gmail.com' and 'Facebook : ashatete'. A footer note at the bottom of the page says 'or you can check it with SHODAN database'.

WANGRIN

Home My Cases Profile Search

English (UK)

Home / Search / Detailed Search

Search Fraud Database

escroc Detailed Search

Danger! escroc is used in frauds

SATURDAY, MAY 10, 2014 3:31:14 PM

Scam Type : Defamation
Description : this man defamed me

Email : escroc@gmail.com Facebook : ashatete

or you can check it with SHODAN database

案例提交流程

escroc@gmail.com

询问了我的中国工商银行密码
然后我就在 2 天后被偷扣了
100,000 元人民币

第一步



身份和犯罪
信息核实

第二步



案件已核实



案件已驳回

workflow automation

escroc@gmail.com
询问了我的中国工商银行密码
然后我就在 2 天后被偷扣了
100,000 元人民币

第一步



身份和犯罪
信息核实

第二步



第三步 : 1%

案件已核实



案件已驳回





Wendelin Exanalytics 去“IOE”的警察大数据

2014-06-11 – 北京





Wendelin Exanalytics 去“IOE”的警察大数据

2014-06-11 – 北京



MariaDB



© 2014 Nexedi SA- Company Confidential

www.wendelin.io



Good afternoon.

I am Jean-Paul Smets, CEO of Nexedi corporation. I will introduce today a new big data technology, that is actually only a few weeks old: Wendelin.

Through this presentation, you will learn how to implement Police Big Data without IBM hardware, without Oracle software and without EMC private cloud. In short, without IOE, only using open source technologies created by a worldwide community.

下午好，

我是 Jean-Paul Smets, Nexedi 公司的 CEO。我今天将介绍一个新的大数据技术，这个叫做 Wendelin 的技术刚刚在几周前被创建出来。

通过这个介绍，您将会学到如何在不用 IBM 硬件，Oracle 软件及 EMC 私有云的情况下实施警察大数据。简单说来，无需 IOE，只使用一个由世界级社区创建的开源技术。

日程表



我们的背景：ERP5 与 SlapOS

- 我们的未来：Wendelin Exanalytics
- 我们的挑战：out-of-core
- 实地应用：Wangrin 自动化

© 2014 Nexedi SA- Company Confidential



My presentation has four parts.

In the first part, I will introduce our company background: ERP5 ERP and SlapOS Cloud.

Then I will explain what is Wendelin technology, how it was built step by step.

In the third part I will explain what are the technical challenge to solve with big data.

I will finish with an example of application that we are currently building and that – maybe – could also be implemented in China in the future.

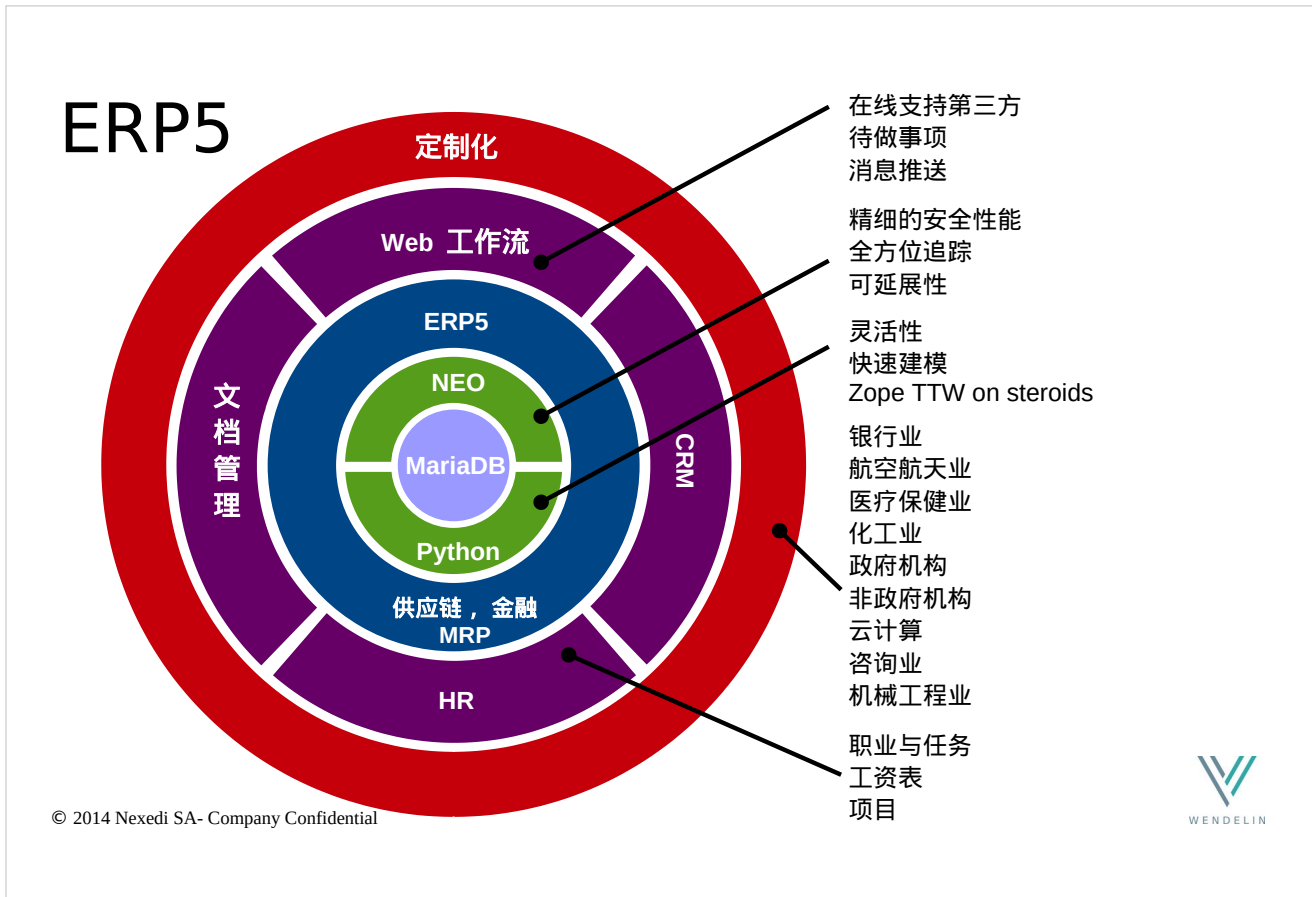
我的演讲分为四个部分。

首先，我会介绍一下我们公司的背景情况：ERP5，ERP 及 SlapOS 云。

然后，我会解释一下什么是 Wendelin 技术并介绍它是如何一步步建造起来的。

第三部分是我将解释大数据的技术挑战。

最后我会用一个我们正在搭建的实际应用案例作为结束，也许这个应用也可以在中国实施。



Nexedi is an open source software publisher that was created in 2001 in France.

We have offices in Germany, Japan, West Africa and Brazil. We opened our subsidiary in China a few months ago in Shanghai Free Trade Zone.

Our primary product is ERP5.

ERP5 is an open source framework based on python language and MariaDB database. It includes a NoSQL database engine for big data called "NEO", that is itself based on MariaDB. ERP5 core consists of generic applications such as ERP, CRM, HR, Document Management and Workflow. ERP5 is a "No IOE" ERP for about any large, complex, mission critical applications.

ERP5 has been customized to manage mission critical applications in many industries: banking, aerospace, chemical, government, cloud computing, etc. ERP5 framework is also the core of SlapOS decentralized Cloud Computing system and Wendelin Big Data.

Nexedi 是一个于 2001 年创立的开源软件发行方。

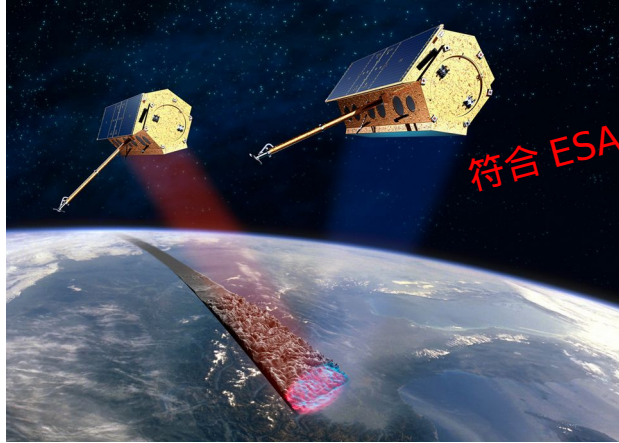
我们在德国，日本，西非和巴西都有办公处。几个月前，我们刚在上海自由贸易区设立了中国子公司。

我们的第一个产品是 ERP5。ERP5 是一个基于 Python 语言和 MariaDB 数据库的开源框架。它包含了专为大数据服务的 NoSQL 数据库引擎，叫做 "NEO"，它自己本身就是基于 MariaDB 的。ERP5 核心包括了一些通用的应用例如：ERP，CRM，人事管理，文档管理和工作流。ERP5 是一个去 "IOE" 的，适用于任何大型，复杂，关键任务应用的 ERP。

定制化的 ERP5 已被用于管理银行，航空航天，化工，政府及云计算等各行各业的关键任务应用。ERP5 框架同时还是 SlapOS 分布式云计算系统和 Wendelin 大数据的核心。

Terra-SAR X 卫星

卫星图像的销售管理及生产



可访问空客 (Airbus)
合作伙伴及分销商
与 DLR 对接
(德国航空局)

« With ERP5, our partners all over the world can access our infrastructure and order online with complete security(通过使用 ERP5, 我们全球的合作伙伴都可以访问我们的基础结构并在完全安全的环境下进行在线订购)» Ralf Duering

© 2014 Nexedi SA - Company Confidential



One typical customer of Nexedi that uses ERP5 is Airbus Defence & Aerospace.

Airbus Defence & Aerospace uses ERP5 to manage the orders and deliveries of satellite images produced by the Terra-SAR X missions. The ERP5 application is accessible worldwide to Airbus partners with strong access rules and security.

Nexedi 的一个 ERP5 典型客户就是空客国防及航空航天。

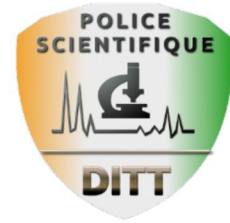
空客国防用 ERP5 管理 Terra-SAR X 任务的订单及传送。ERP5 应用是可以实现在严格的访问规则和安全控制下访问全球的空客合作伙伴，以确保比如说某些国家的某些地区无法被其他与之有领土冲突的国家查看。

象牙海岸警务云

可恢复性 IPv6 云



通用中国产硬件
无 IBM 无 EMC 无 Oracle
100% 开源软件



IPv6 Mesh 网络
可恢复性数据存储
分布式基础架构

"Resiliency and auditability are required for government Cloud. We chose SlapOS to gain auditability through open source software and resiliency through decentralized architecture (可恢复性及可审核性是政府云所需要的。我们选择 SlapOS 通过开源软件来获得可审核性，通过分布式基础架构获得可恢复性)" - Stéphane Konan - DITT MEMI

© 2014 Nexedi SA- Company Confidential



Another typical customer of Nexedi is the Ministry of Interior of Ivory Coast.

The Ministry of Interior of Ivory Coast deployed SlapOS cloud computing system to store all sensitive data of police as well as different databases. This cloud computing system only uses open source software and generic hardware. There is no high end hardware or proprietary software. No storage area network. No expensive router. No proprietary database. It is a "No IOE" Cloud.

Policemen in Ivory Coast have been trained to install the system by themselves. They are now able to reinstall their cloud without Nexedi. SlapOS is not only open source and "No IOE" but also probably the only cloud computing system that is so simple that it can be installed by non specialists.

Another features of this cloud is resiliency. By using modern networking technologies such as IPv6 and mesh topologies, it is possible to ensure continuous service even in case of destruction of data center or in case of human error.

Nexedi 的另外一个典型客户就是象牙海岸内政部。

象牙海岸内政部部署了 SlapOS 云计算系统来储存所有敏感的警务数据及不同的数据库。

这个云计算系统仅仅使用开源软件及通用硬件。没有高端硬件或是私有软件。没有存储区域网络。没有昂贵的路由器。没有私有数据库。它是一个去"IOE"的云。

象牙海岸的警察接受了培训并已经可以自己安装这个系统。警察现在已经可以自己重装云而不需要 Nexedi 来安装了。SlapOS 不仅是开源和去"IOE"的，也是唯一足够简单到可以由非专业人士安装的云计算系统了。

这个云的另一个优势就是可恢复性。通过使用像 IPv6 和 mesh 这样的现代网络技术，可以实现当数据中心被破坏或是人为错误出现时保证持续的服务。

日程表

- 我们的背景 : ERP5 与 SlapOS
-  • 我们的未来 : Wendelin Exanalytics
- 我们的挑战 : out-of-core
- 实地应用 : Wangrin 自动化

© 2014 Nexedi SA- Company Confidential



We believe in Nexedi that it is important to embed Big Data into ERP5 because future business applications will rely on the processing of large amounts of data generated by the Web or by Internet of Things. An ERP without native Big Data has no future in our opinion. This is why we created Wendelin.

With Big Data technology, ERP5 will be able natively to predict inventories or reply automatically to clients asking a question to after-sales support. Many other applications are possible.

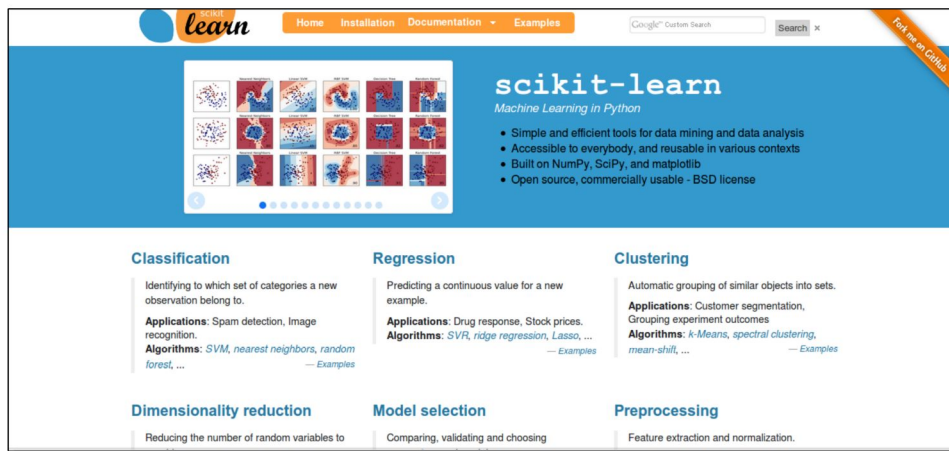
I will now explain how we are building Wendelin based on existing open source technologies.

在 Nexedi 我们相信在 ERP5 中嵌入大数据是非常重要的，因为未来的商务应用将依赖于处理由网络或物联网产生的大量数据。在我们看来，一个没有自带大数据的 ERP 是没有未来的。这也就是为什么我们开发了 Wendelin.

通过大数据技术，ERP5 将可以自行预测 ERP5 贸易库存，自动在 ERP5 CRM 里回复客户，许多其他的应用都可能实现。

我现在来解释一下我们是如何在现有的开源技术上建造 Wendelin 的。

运用最完美的解析学 scikit-learn.org



© 2014 Nexedi SA- Company Confidential

首先，什么是大数据？

对于 IT 企业来说，大数据就是 IT 公司继“云计算”之后用来向客户收取更多费用的另一标语。

但是事实上，大数据是一个叫做“机器学习”的统计科学应用的新名字用来应对逐渐增长的政府及企业数据。对比所有的机器学习软件，最好的之一或着可以称作最好的就是“scikit-learn”。它是由法国国家信息与自动化研究所创立，现在由巴黎高等电信学校和一个包括德国波恩大学在内的大型社区进行开发。它是一款免费的开源产品。Scikit-learn 已被 Evernote, Spotify, 谷歌，Cloudera 等公司使用。

Scikit-learn 的主要用途就是在数据中预测，猜测或者侦查对象。

例如，Scikit-learn 能够被用作猜测人的想法。让我来解释一下它是如何工作的。首先连接一个心电图（ECG）到 100 个不同的人，ECG 是一个可以连接到大脑并捕捉大脑活动从而形成一系列电子信号的设备。要每一个人想一个罗马字母（A，B，C 等）。记录这一百个人的心电图和所想的字母。总共就是 2600 个记录。

然后再另外找一个人，为他连上心电图，让他随机想一个字母并不说出来，记录心电图的信号。运行 Scikit-learn 来比较这个信号和之前的 2600 个信号记录。

Scikit-learn 将可以猜出这个人想得是哪一個字母。如果他想的是“P”，那么“p”就会在屏幕上显示出来。没有魔术，这就是一个统计学应用。

First, what is Big Data?

For many IT companies, Big Data is just the next catchword after “cloud computing” to collect more money from customers.

But in reality, Big Data is the new name for the application of a statistical science called “machine learning” to the ever increasing amount of government and business data. Among all machine learning software, one of the best if not the best is called “scikit-learn”. It was created by INRIA in France and is now developed by Telecom Paristech and a large community that includes for example the University of Bonn in Germany. It is open source. It is free. And it is now used by Evernote, Spotify, Google, Cloudera.

The main purpose of scikit-learn is to predict, guess or detect certain properties in data.

For example, scikit-learn can be used to guess human thoughts. Let me explain how this works. First connect an ECG to 100 different people. An ECG is a kind of device that can be connected to your head and that captures the activity of your brain in the form of many electrical signals. Ask each person to think about each letter of the roman alphabet (A, B, C, etc.). Record the signal of the ECG of each of the 100 person and each of the 26 letters. That is a total of 2600 recordings.

Then take one more person. Connect an ECG to that person. Ask that person to think about a random letter without telling which letter. Record the ECG signal. Run scikit-learn to compare that signal to the previous 2600 recorded signals. Scikit-learn will then be able to guess which letter the person was thinking in his mind. If he or she was thinking about letter “P”, then “P” will appear on the display. There is no magic. It is only an application of statistics.

由最优秀的数学家们设计

http://en.wikipedia.org/wiki/Fields_Medal

Number of Fields Medallists by country [\[edit\]](#)



Wendelin Werner



© 2014 Nexedi SA- Company Confidential

The reason why scikit-learn is better than other libraries is because it was created by better mathematicians. It also explains why scikit-learn was created in France.

If we consider the ranking of “Fields Medals” worldwide – the equivalent of Nobel prize for mathematics – USA is number one with 11, France number two with 10 and Russia number 3 with 9. A city like Paris counts nowadays as many “Fields Medalists” as the whole United States. 90% of them come from same institute: “Ecole Normale Supérieure”.

This situation creates a unique social network that connects mathematicians and computer scientists who graduated from that same institute. Gael Varroquaux – the creator of scikit-learn – graduated from that institute. Another famous alumni of this institute is “Wendelin Werner”. He is the first mathematician in “probability science” to be awarded a “Fields Medal”.

We eventually decided to call our Big Data engine “Wendelin” because we wanted to highlight that probabilities are the foundations of statistics which are themselves the primary science of “Big Data”.

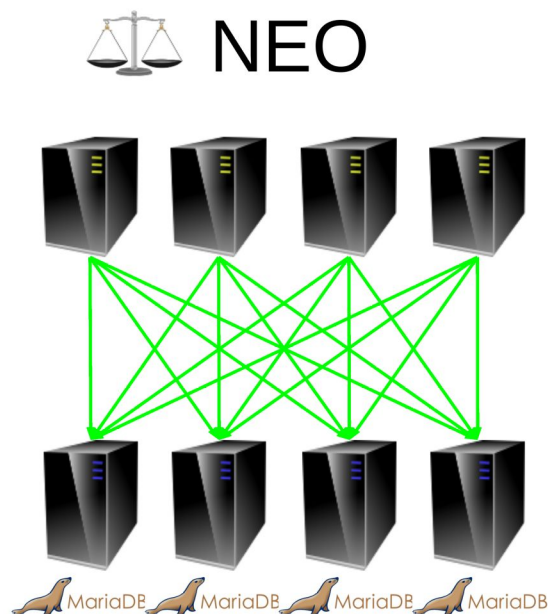
Scikit-learn 之所以比其他库要更好的原因就是它是由更优秀的数学家们设计的。这也解释了为什么 Scikit-learn 创造于法国。

如果我们来看一下全球“菲尔兹”奖——相当于数学界的诺贝尔奖的获得者，美国以 11 座奖牌占据第一位，法国以 10 座奖牌占据第二，俄罗斯以 9 座奖牌位居第三。像巴黎这座城市就有着和整个美国差不多数量的“菲尔兹奖牌”。因为 90% 的奖牌获得者都来自于同一所学校：“巴黎高等师范学院。”

这个现状给毕业于这所学校的数学家和计算机科学家们创造了一个独特的社会网络关系。Gael Varroquaux - Scikit-learn 的创造者 - 正是毕业于这所学院。这所学院的另一位著名校友就是“Wendelin Werner”。他是第一位“概率科学”界获得“菲尔兹奖”的数学家。

由于这个原因，我们决定将我们的大数据引擎命名为“Wendelin”，因为概率学就是统计学的基础，也就是“大数据”的原理。

加入分布式存储 neoppod.org



© 2014 Nexedi SA- Company Confidential



What was missing until now to scikit-learn was the ability to store and access large amounts of data transparently. Happily, Nexedi has developed for 6 years a Big Data database called “NEO” that uses the same programming language as scikit-learn: the “python” language.

By combining NEO with scikit-learn, scikit-learn can access very large amounts of data that can span across thousands of computers in a datacenter. This requires no change in the source code for developers, which is a huge advantage compared to other approaches.

在那个时候 Scikit-learn 还欠缺的一个能力就是存储和访问大量的数据。

非常幸运的是，Nexedi 已经开发了 6 年的一个叫做 “NEO” 的大数据数据库与 Scikit-learn 使用的是同一种编程语言：Python。

通过结合 NEO 和 Scikit-learn, Scikit-learn 能够访问跨越一个数据中心中的上千台电脑的大量数据。

开发人员无需修改任何的源代码，这相比其他方案来说是一个巨大的优势。

加入弹性 PaaS erp5.com

```
# Initialize data
data_size = 1000000
server_count = 1000
chunk_size = data_size / server_count
data = array(data_size)

# Process data in parallel on each server (Map Reduce, Batch, etc.)
for server in server_count:
    data.activate().process(server*chunk_size, chunk_size)
```



© 2014 Nexedi SA- Company Confidential

We took from ERP5 its distributed processing component to achieve elastic computing in Wendelin.

ERP5 Platform as a Service (PaaS) can actually use each computer of a datacenter and process bits of data in parallel to much faster result than with a single computer. Programmers can leverage this elastic feature very easily through a Web based, online text editor.

我们同时也添加了 ERP5 的分布式处理组件以实现 Wendelin 里的弹性计算。

ERP5 平台即服务 (PaaS) 事实上能够使用一个数据中心的每台电脑来各自以更快的速度，平行处理一小部分数据，最终和用一个独立的电脑来处理获得相同的结果。编程人员能够非常容易的通过一个基于网络的在线文字编辑器来实现这个弹性功能。

加入多重云部署 slapos.org



© 2014 Nexedi SA- Company Confidential



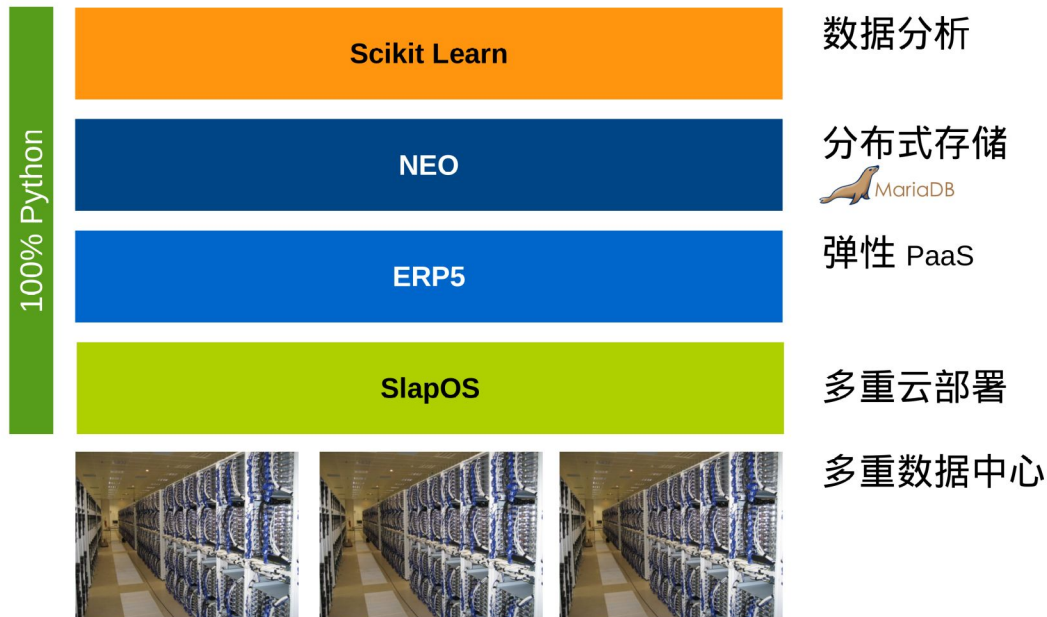
We added SlapOS mutli-cloud system to automate deployment in one or multiple datacenters.

This is the same technology as the one used by Ivory Coast police, Mitsubishi in Russia or SANEF in France. Wwith minimal training, system administrators can setup a private cloud based on SlapOS that can be operated automatically and that can deploy Wendelin in one or in many datacenters.

我们也添加了 SlapOS 多重云系统在一个或多个数据中心里进行自动化部署。

这和象牙海岸警察局，俄罗斯三菱以及法国 SANEF 使用的是同样的技术。这说明通过最少的培训，系统管理员就能够在 SlapOS 的基础上建立一个私有云，这个 SlapOS 除了可以自动化运行，还可以在一个或多个数据中心里自动化部署 Wendelin。

Wendelin Exanalytics 核心 100% 开源



© 2014 Nexedi SA- Company Confidential



Overall, we have created a complete Big Data stack that includes multi-cloud deployment, elastic Platform as a Service, distributed storage and machine learning. This stack is 100% open source and 100% based on the same language: python.

By using NEO, we no longer need disk bays from “IBM”, “EMC” or “Netapps” or database from “Oracle” (No “I”).

By using python, we no longer need “Java” from “Oracle” (No “O”).

By using SlapOS, we no longer need “VMWare” software from “EMC” (No “E”).

Wendelin thus demonstrates the possibility to deploy immediately a “No IOE” Big Data system in China..

总的来说，我们已经创造了一个包含了多重云部署，弹性平台即服务，分布式存储和机器学习的完整大数据堆栈。这个堆栈是百分百的开源并且百分百的基于同一种语言：Python.

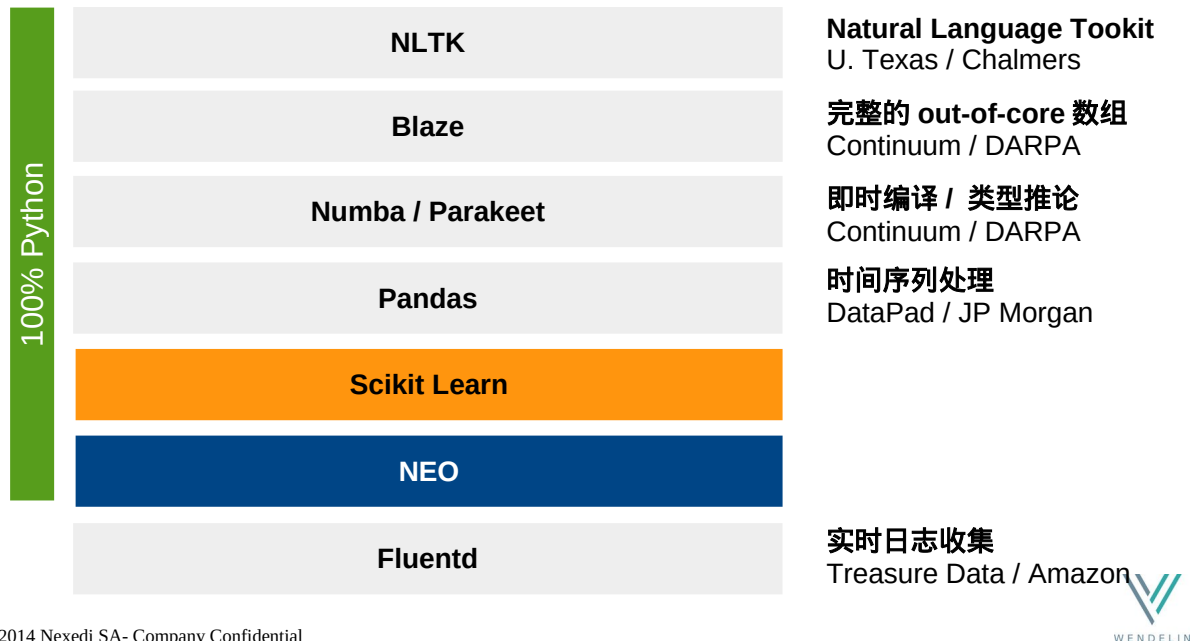
通过使用 NEO，我们不再需要来自“EMC”，“IBM”或者“Netapps”的**磁盘分区**或者是来自“Oracle”的数据库。（去“I”）

通过使用 Python，我们不再需要来自 Oracle 的 Java。（去“O”）

通过使用 SlapOS, 我们不再需要来自“EMC”的“WMware”软件。（去“E”）

Wendelin 证明了在中国立即部署一个“ No IOE” 大数据系统的可能性。

Wendelin 的选择 100% 开源



China is not the only country to consider “No IOE” approaches for Big Data.

中国不是唯一一个考虑去“IOE”方法大数据的国家。

The US defense department itself and Wallstreet banks have contributed many open source components that can be used as options for Wendelin. The US defense department has contributed a compiler that can accelerate Wendelin in some cases by a magnitude of 10 to 100 by using assembly language. New York University has contributed a component that can accelerate Wendelin in some cases by a magnitude of 1000 by using GPUs. JP Morgan has sponsored a component that provides traditional statistics to Wendelin in addition to scikit-learn. Chalmers University has contributed native language processing tools.

美国国防部门自己和华尔街银行已经贡献了许多开源组件作为 Wendelin 的选择。美国国防部门贡献了一个通过组合语言在某些情况下加速 Wendelin 10 到 100 倍的程序编译器。纽约大学贡献了一个通过使用 GPUs 加速 Wendelin 1000 倍的组件。JP Morgan 资助了一个除 Scikit-learn 之外另一个为 Wendelin 提供传统统计学的组件。查尔摩斯工学院贡献了本地语言处理工具。

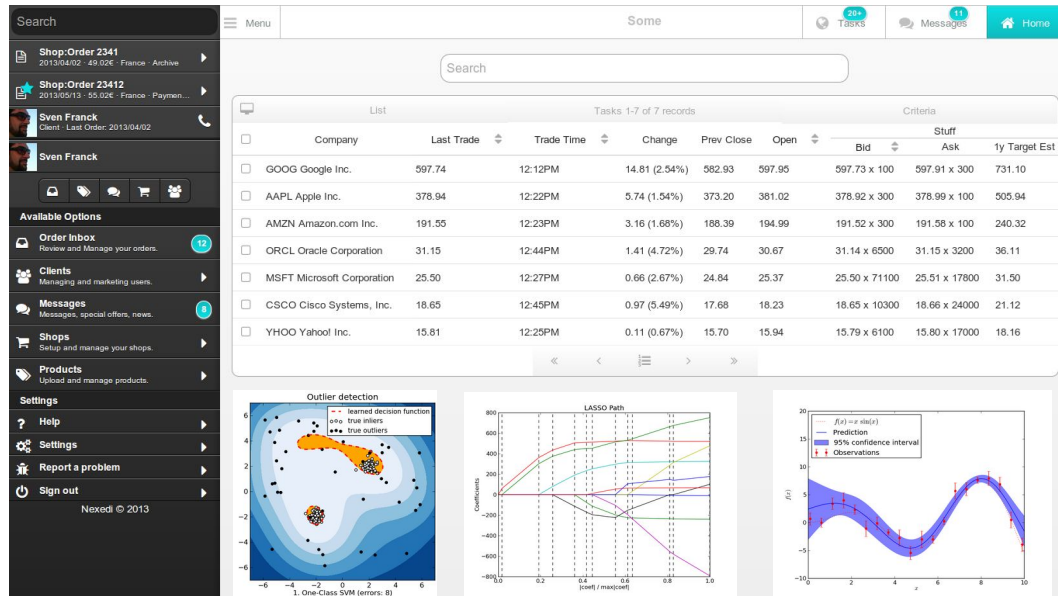
“No IOE” in Wendelin therefore means more contributions from a wider international community to make a better open source tool. We expect that many open source components will soon be contributed to Wendelin from China in the same spirit as what Taobao did recently with MariaDB.

Wendelin 的“无 IOE”说明了来自于更广泛的国际社区的贡献成就了一个更优质的开源工具。我们希望中国能够本着同样的精神尽快的开发出更多的开源组件，就像淘宝最近为 MariaDB 做的一样。

Andrew Ng – godfather of machine learning – inspired scikit-learn lectures and recently moved to Baidu. Wei Li - a young graduate of Tsinghua who lives now in United States – made 108 contributions to scikit-learn. This is only the beginning.

吴恩达 — 机器学习之父 — 启发了 Scikit-learn 课程并于近期转移就任于百度。李 Wei — 一个年轻的目前居住在美国的清华毕业生 — 为 Scikit-learn 做出了 108 项贡献。这还仅仅是个开始。

Wendelin 用户界面 renderjs.org



© 2014 Nexedi SA- Company Confidential



A short word now on our user interface: it is fully HTML5 based and responsive. This means that it can run on smartphone, tablet or desktop PC. It also means that it does not depend on Microsoft Windows and can run on any other operating system. If there was an “M” in “No IOE” then we could say here “No M”. 现在简单介绍一下我们的用户界面：完全的 HTML5 并且反应迅速。这说明它可以在智能手机，平板电脑或着是台式机上运行。这也说明它不需要依赖于微软 Windows 并且可以在任何的操作系统上运行。如果在“区 IOE” 里还有一个“M”，我们在这里就可以加上“去 M”。

Wendelin 应用

- 干扰侦查
- 欺诈侦查
- 商业及经济预测
- 市场分析
- 媒介分析
- 公共安全
- 脑机接口
- 物联网

© 2014 Nexedi SA- Company Confidential



Here are some possible applications of the Wendelin platform.

Intrusion detection, for example in the control system of a nuclear powerplant

Fraud detection in banking transactions

Sales prevision of a company

Customer segmentation for marketing

Automated analysis of recorded videos and incident detection

Automated detection of copyright infringement in Internet

Brain computer interface to control computer from human thoughts

Automated maintenance of device that are connected to the Internet of Things

I am sure you can find many other applications

这里是一个 Wendelin 平台的可行应用的列表。

干扰侦查，例如装在一个核动力装置的控制系统里。

在银行交易中的欺诈侦查。

一个公司的销售预测。

市场的客户分类。

录制视频及事故侦查的自动分析。

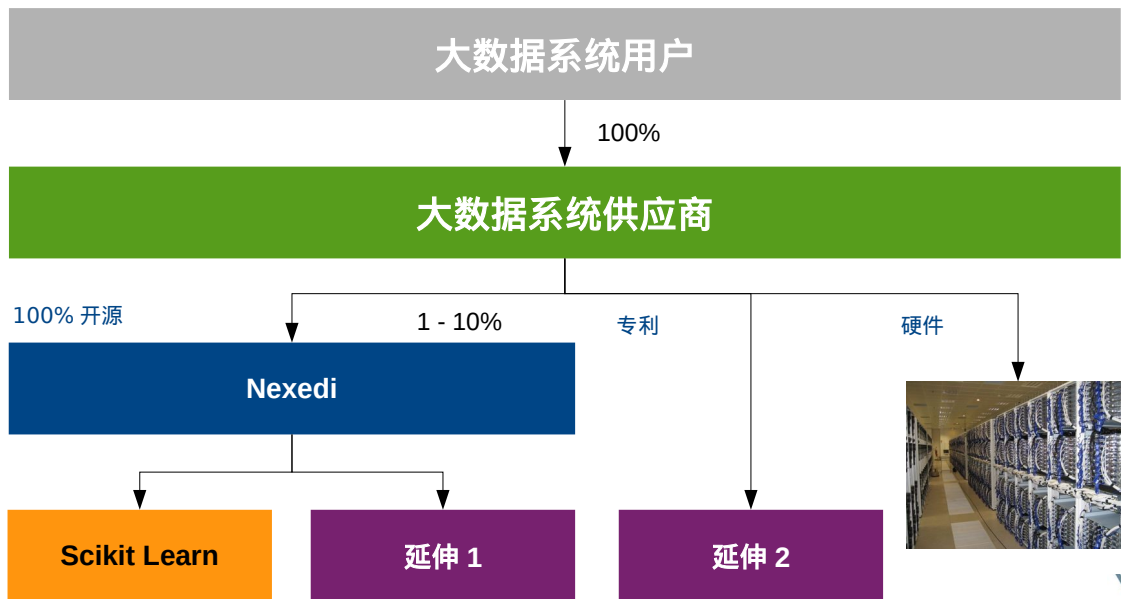
互联网版权侵犯的自动侦查。

用人的思维通过脑机界面控制电脑。

连接到物联网的设备自动化维护。

我保证您还可以很发现其他的应用。

商务模式：德国风格 No VC



© 2014 Nexedi SA- Company Confidential



Wendelin follows vertical business model which is typical of open source and German style business.

Wendelin developers team with a supplier of Big Data vertical solutions. This supplier can be for example a company that provides Big Data systems for police applications.

Nexedi provides maintenance and support contract on the Wendelin core and initial software setup.

The Wendelin core itself is open source. It can be combined with proprietary code made by the supplier of Big Data vertical solution and deployed on any kind of hardware.

Nexedi receives a small share of the total revenue that the supplier of Big Data vertical solution receives from users. This small share is used to maintain Wendelin core, extend scikit-learn or extend other open source components.

Through this business model, the supplier of Big Data vertical solution controls both business and technology, yet benefits from a stable R&D team in Europe where turnover of engineers is much lower than in China for example.

This business model also helps sharing R&D among different vertical markets.

Wendelin 使用的是垂直商务模型，是一个典型的德国式开源商务。

Wendelin 开发团队加上一个大数据解决方案供应商。这种供应商可以是一个提供警务应用大数据系统的公司。

Nexedi 提供了维护及支持 Wendelin 核心和原始软件建立的合同。

Wendelin 核心是开源的。它可以被合并到大数据解决方案供应商的私有代码中并被部署到任何类型的硬件上。

Nexedi 收取一小部分来自于大数据解决方案供应商的收益并用来维护 Wendelin 核心，扩充 scikit-learn 或者其他开源组件。

通过这个商务模式，大数据解决方案供应商控制了商务及技术，并得益于来自欧洲的稳定的研究开发组，其人员流动率远低于中国。

这种商务模式也同时帮助不同的垂直市场之间共享研究成果。

日程表

- 我们的背景 : ERP5 与 SlapOS
- 我们的未来 : Wendelin Exanalytics
-  • 我们的挑战 : out-of-core
- 实地应用 : Wangrin 自动化

Wendelin is still far from perfect and we still have some challenges to solve.

Wendelin 还远远不够完美，我们还面临着许多的挑战。

Out-of-core 数组

```
# Numpy
np.ndarray(shape=(2,2), dtype=float, order='F')

# Out-of-core data
np.ndarray(shape=(1e18,2), dtype=float, order='F')    1 Exabyte

# Full out-of-core
np.ndarray(shape=(1e9,2e9), dtype=float, order='F')    1 Exabyte
```

最好的 out-of-core 拓扑学需要依赖于运算法则和方阵
几何学

© 2014 Nexedi SA- Company Confidential



The goal of Wendelin is to process exabytes of data within 10 years

One exabyte is 1 billion of billions of bytes.

No system – even the Large Hadron Collider in Geneva – can yet process this in its core.

We believe that a first simple case of exabyte can be processed soon by Wendelin. We call it “out-of-core data”. This is typically the case of video or log processing to build machine learning models.

Another case will take more time. We call it “full out-of-core”. It is the case of the analysis of a full social graph of a billion people.

Wendelin 的目标就是在 10 年之内处理艾字节的数据。

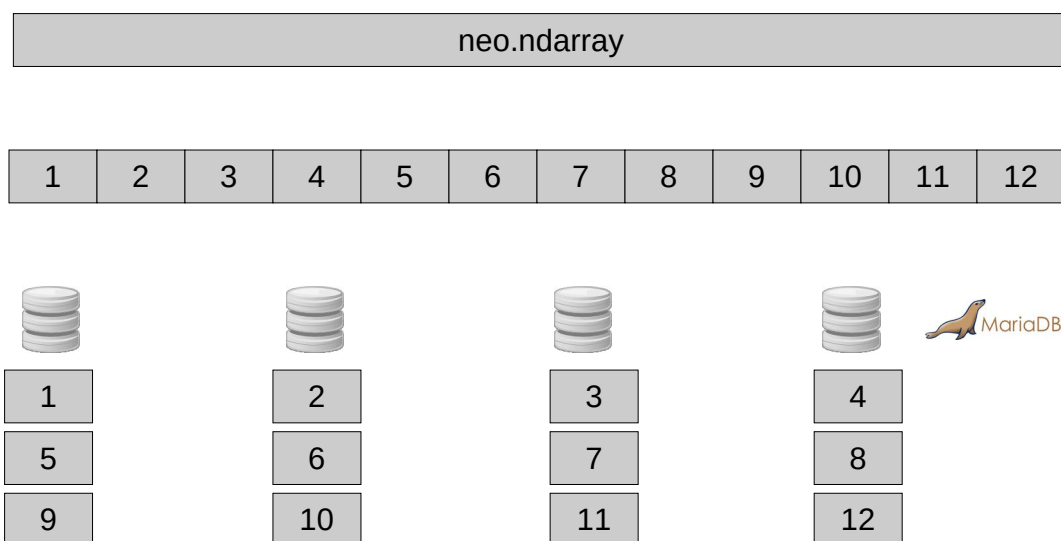
一个艾字节就是上亿的数量。

无系统 — 就算是日内瓦的大型核子对撞机 — 都还没能力在它的核心中处理这个数量的数据。

有一个简单的艾字节的案例，我们相信可以很快对它进行处理。我们把它叫做“Out-of-core 数据”。这是典型的视频或日志处理案例。

另一个案例将需要更久的时间，我们把它叫做“完全 out-of-core”。它是十亿人的全部社会曲线图。

neo.ndarray out-of-core 数据



© 2014 Nexedi SA- Company Confidential

The base idea to process large data sets is to split the data into small parts.

处理大量数据的基本原理就是将数据分成许多小份。

Here, we split one big block of data into 12 small blocks

这里我们将一个大码块的数据分成 12 个小码块

We have four storage nodes.

我们有四个存储节点。

Block 1 goes to storage 1

第一码块到第一个存储中，

Block 2 goes to storage 2

第二码块到第二个存储中，

Block 3 goes to storage 3

第三码块到第三个存储中，

Block 4 goes to storage 4

第四码块到第四个存储中，

Block 5 goes to storage 1

第五码块到第一个存储中，

Block 6 goes to storage 2

第六码块到第二个存储中，

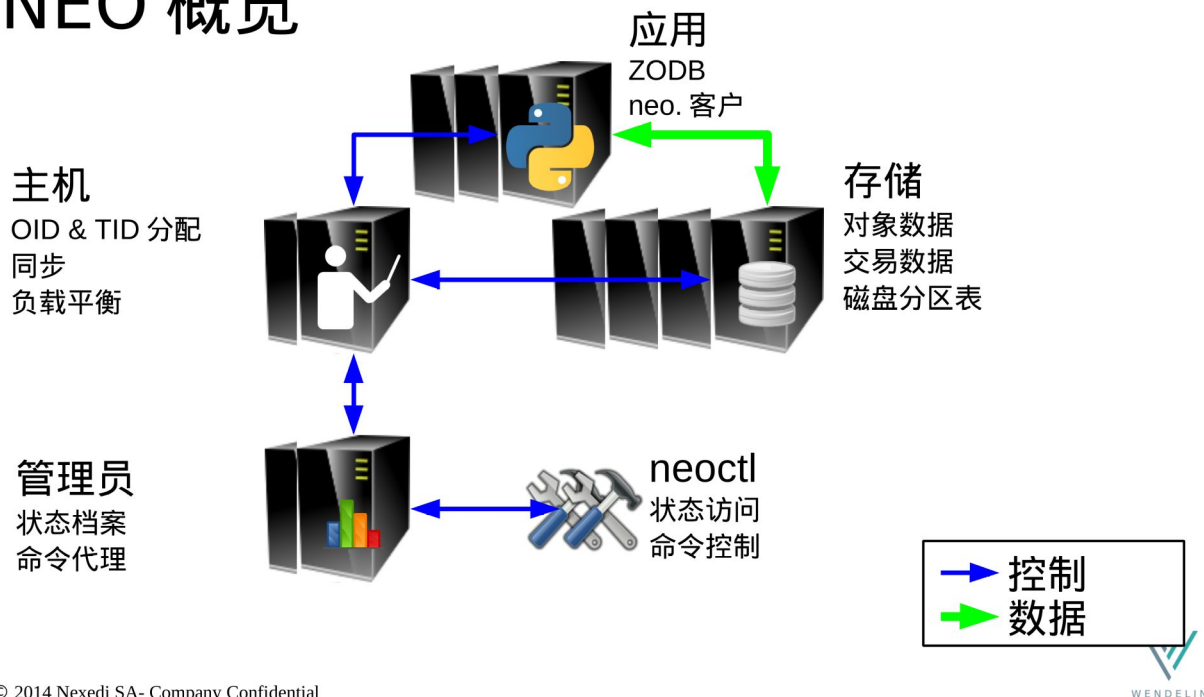
Etc.

等等

By adding more storage nodes, we can increase the size of the big block of data that can be stored and processed.

通过添加更多的存储节点，我们可以增加需要储存和处理的大码块数据的数量

NEO 概览



NEO is the technical component that splits big blocks of data into smaller blocks that are stored in multiple storage nodes. NEO is open source. Some people call it “NoSQL” database for python objects.

What is more important is that developers who use NEO do not need to change their code. The process of splitting big data into smaller one is completely transparent.

NEO has 5 components: storage, processing, master, administration and control.

Storage components are used to store small data blocks. A typical NEO setup consists of a few to thousands storage components.

Processing components access data stored in storage components and process it.

Master components makes sure that data remains consistent if two process components try to change the same data.

The administration component monitors the sanity of the NEO database and triggers decision to replace one component by another to achieve high availability.

The NEO control component is used by the system administrator to access the administration component and control low level aspects of the system.

The most innovative aspect of this architecture is that it achieves at the same time scalability and transactions with no single point of failure and no central index.

NEO 是一个将大码块数据分离成小码块并储存在多个小心存储节点的技术组件。 NEO 是开源的。一些人也叫它“ NoSQL”Python 对象数据库。

更重要的是使用 NEO 的开发人员无需改变他们的编码。分离大数据到小块的过程是完全透明的。

它有 5 个组件：存储，处理，主机，管理员和控制。

存储组件被用作储存小的数据码块。一个典型的 NEO 设置包含了一小部分到上千的存储组件。

处理组件访问存储组件中的数据并进行处理。

如果两个处理组件想要改变同样的数据，主机组件保证了数据的一致性。

管理员组件用来监督 NEO 数据库的健康状况，并作出由一个组件取代另一组件的决定，以此来达到更高的可用性。

Neo 控制组件被系统管理员用来访问管理员组件并控制系统中低级的部分。

这个架构最有创意的部分就是它能够同时满足延展性及交易无单独端点故障及无中央索引。

路径

- **Q3 2014: 开发者发布 Wendelin**
- **Q4 2014: 简单的优化 neo.ndarray**
- **Q3 2015: 完整的优化 neo.ndarray**

© 2014 Nexedi SA- Company Confidential



The first release of Wendelin is planned for Q3 2014, in a few months.

An optimized version will be released at the end of the year

A fully optimized version is expected in 2015.

Although it is still being developed, Wendelin was already sold to a company in Germany for a vertical solution related to the “Internet of Things”.

**Wendelin 的第一次发布预计安排在 2014 年第三季度，就在几个月内。
年末将会发布一个优化的版本。**

2015 年将会有一个完整的优化版本。

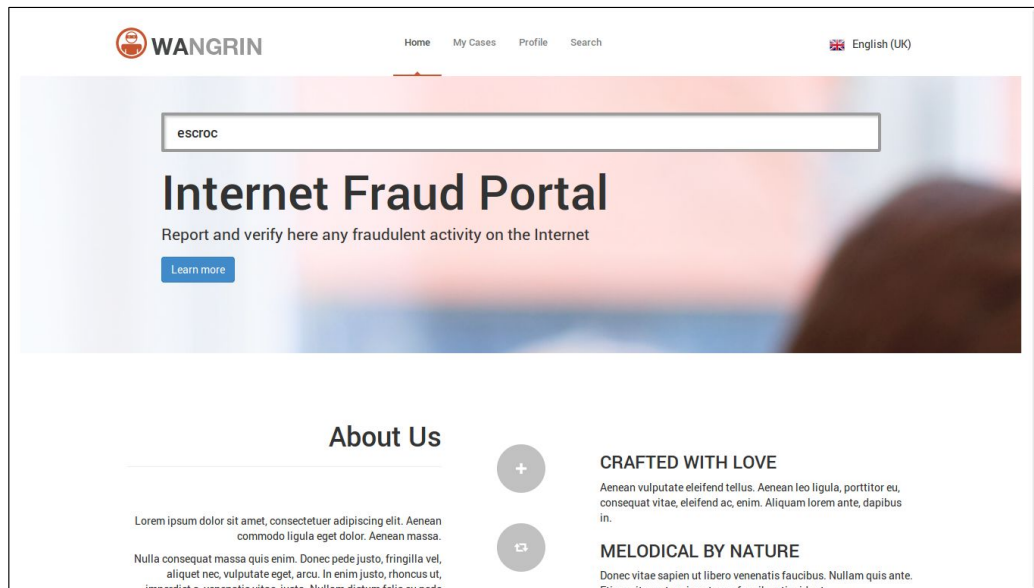
虽然在处于开发阶段，Wendelin 已经被一家德国公司购买用作“物联网”的垂直解决方案。

日程表

- 我们的背景 : ERP5 与 SlapOS
- 我们的未来 : Wendelin Exanalytics
- 我们的挑战 : out-of-core
-  • 实地应用 : Wangrin 自动化

To conclude, I would like to mention an application in the field of police and Big Data
作为总结，我想讲一个关于警务领域和大数据的应用。

wangrin.net – 互联网欺诈门户网



WAGRIN is a startup company created by Stéphane Konan to develop an “Internet Fraud Portal”

Its concept is simple.

Suppose that a person writes to you an email and asks you to change your bank account password.

Can you trust him ?

With WAGRIN, users can enter the email address of that person (ex. “escroc@gmail.com”) and requests a search of the fraud database.

WAGRIN 是由 Stéphane Konan 创立的初创公司开发的“互联网欺诈门户”的项目。

概念非常简单。

假设有一人给你写信请你换一个银行账户密码。

你要相信他吗？

通过 WagrIn, 用户可以输入这个人的邮件地址（例如：“escroc@gmail.com”）并要求在欺诈数据库中进行搜索。

wangrin.net – 互联网欺诈门户网

The screenshot displays the Wangrin Search Fraud Database interface. At the top, there is a navigation bar with links for Home, My Cases, Profile, and Search, along with a language selector for English (UK). Below the navigation bar, the page title "Search Fraud Database" is centered. A search input field contains the text "escroc", and a red button labeled "Detailed Search" is positioned to its right. Below the search field, a red warning box states "Danger! escroc is used in frauds". A timestamp indicates the search was performed on Saturday, May 10, 2014, at 3:31:14 PM. The search results section shows a "Scam Type" of "Defamation" and a "Description" of "this man defamed me". Below this, contact information is listed: an email address "escroc@gmail.com" and a Facebook profile "ashatete". At the bottom of the results section, a note suggests checking the SHODAN database.

© 2014 Nexedi SA- Company Confidential



The fraud database then shows that a complain case has already been filed for that person.

Knowing this, you should be careful before changing your bank password.

数据库然后会显示一个有关这个人的举报案例。

知道了这一点，在改变你的银行账户密码前，你会更加小心。

案例提交流程



© 2014 Nexedi SA- Company Confidential



The way WANGRIN works is simple.

Victims of cybercrime go to the www.wangrin.net portal. They enter their name, upload a copy of their ID card and explain what happened to them. They provide some details like phone number, email address, ID of the criminal in skype or in a social network, etc.

Then they submit their case.

Policemen or employee of Wagrín – depending on the existence of agreement with governments – review the case. They either approve it or reject it.

The more WANGRIN is used, the more policemen are required. This is where Wendelin helps.

WANGRIN 工作的方式很简单。

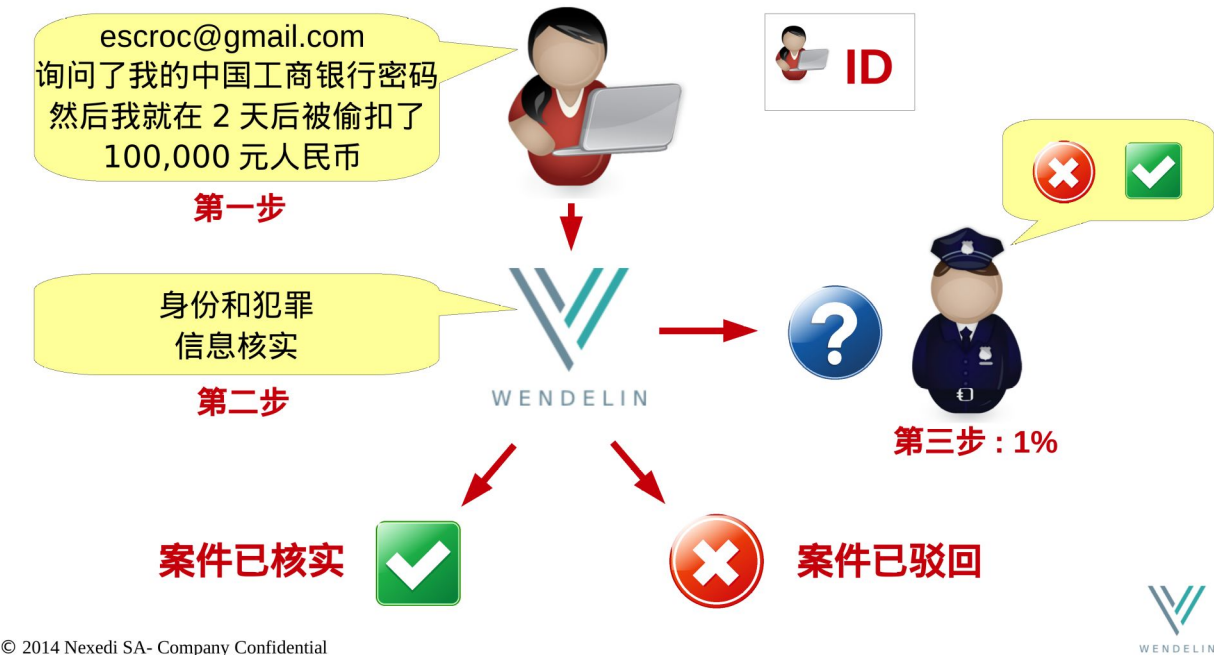
网络犯罪的受害者到 www.wangrin.net 门户网。输入他们的名字，上传一张他们的身份证并解释发生了什么。他们提供一些像电话号码，犯罪者在社交网络中的 ID 等细节信息。

然后提交他们的案例。

警察或者 Wagrín 的员工 — 基于与政府的协议情况 — 检测这个案例，可以作出核实或是驳回的决定。

Wagrín 用的越多，需要的警察数量也越多。这就是 Wendelin 该协助的地方。

workflow 自动化



Wendelin can be used to automate the review of criminal case submission.

By using appropriate algorithms in scikit-learn, it is possible to automatically accept or reject 99% of submissions. This works in the same way – yet more sophisticated – as anti-spam systems can recognize if email sent to you is worth reading or just advertising.

For the 1% remaining cases, policemen still have to process data.

This approach works best with a lot of data. The more data from the larger country with the more people using the same language, the most accurate the results of Wendelin.

Wendelin 将被用作对提交的案件进行自动审查。

通过使用 Scikit-learn 中合适的运算法则，可以自动化接受或驳回 99% 的提交案件。这个和反垃圾邮件系统一个道理 — 当然比它更加的复杂 — 因为反垃圾邮件可以分辨收到的邮件值得阅读还是只是广告。

对于那留下的 1% 的案件，还是需要警察来处理这些数据。

这个方法在许多数据上使用得都非常的好。如果有越多的数据来自于有大量人使用统一语言的大型国家，这个结果就越精确。



Wendelin Exanalytics 去“IOE”的警察大数据

2014-06-11 – 北京



MariaDB



© 2014 Nexedi SA- Company Confidential

www.wendelin.io



This simple application of Wendelin to police Big Data concludes my talk.

Thank you for your patience.

I will be happy to reply any question now or later.

这个简单的 Wendelin 警察大数据应用总结了我的演讲。

感谢关注，从现在开始，我将非常高兴回答任何的问题。