Patrick Beaucamp

Founder of the Vanilla Project

Mail : Patrick.beaucamp@bpm-conseil.com

Open Source Platforms to deploy Search and Maps Visualization on top of a Big Data Database

II-SDV, Nice 15th April 2013

II-SDV

BPM-Conseil

the Company behind the Vanilla project

Presentation Agenda

Big Data Use case

Various Application

Big Data Landscape

Data Storage and Database Type
PlayerMap - Hadoop
Search & DataMining Tools
Data Visualisation

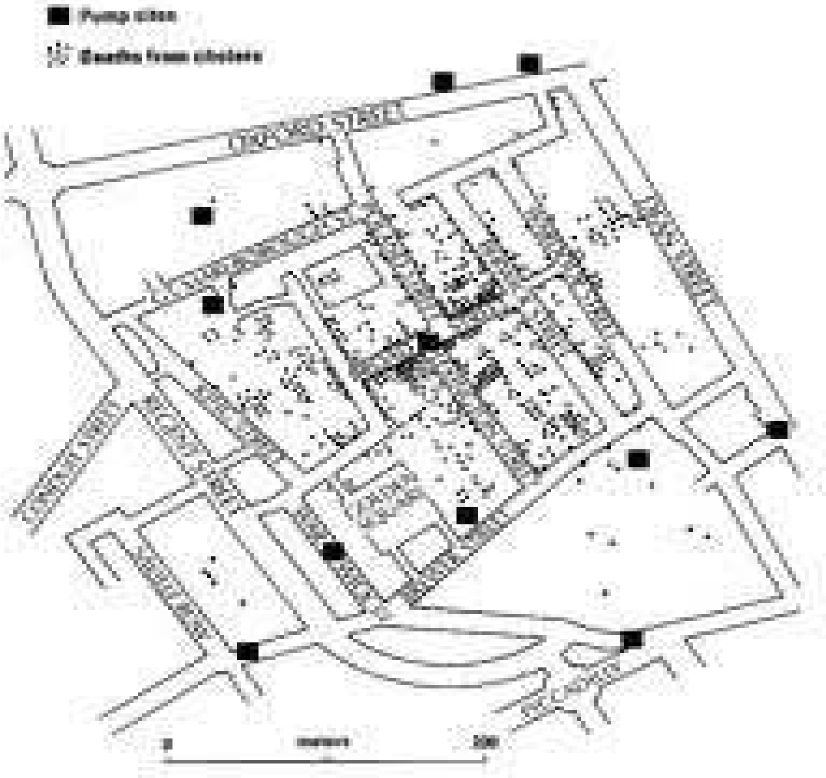
Big Data Use case

Social Network
Open Data

Various Applications

Discovery of Cholera – 1854 (John Snow)

http://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

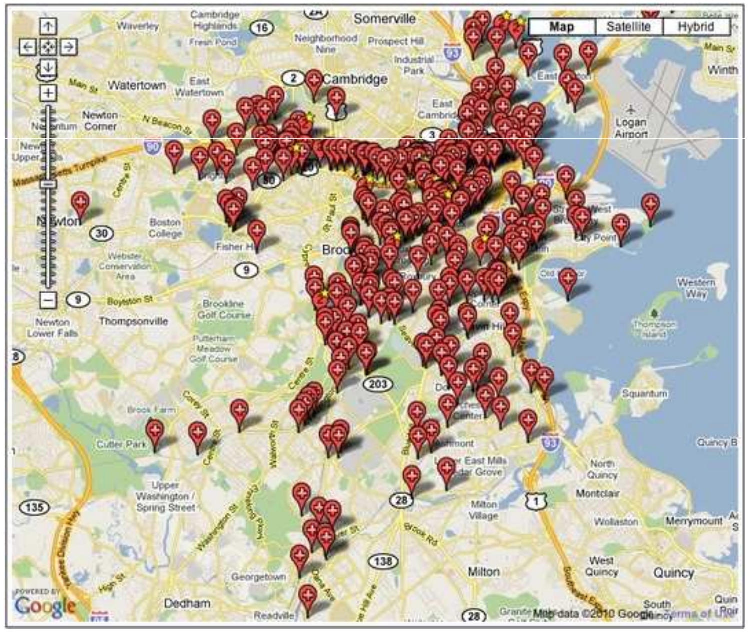


Various Applications

Bicycle Accident in Street : who is taking care of trafic management

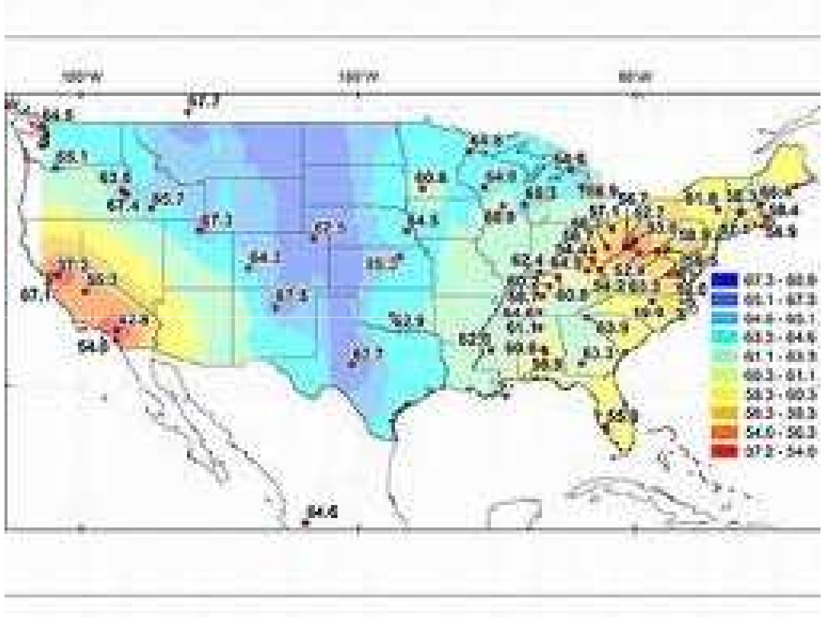
Example in Boston :

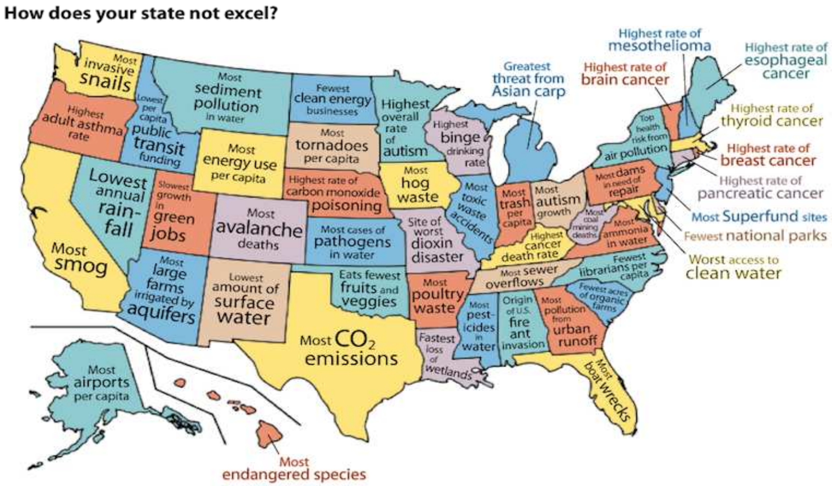
http://www.boston.com/bostonglobe/editorial_opinion/blogs/the_angle/2010/12/bike_crash_map.html



Various Applications

Pollution data with electoral data : who is taking care of pollution



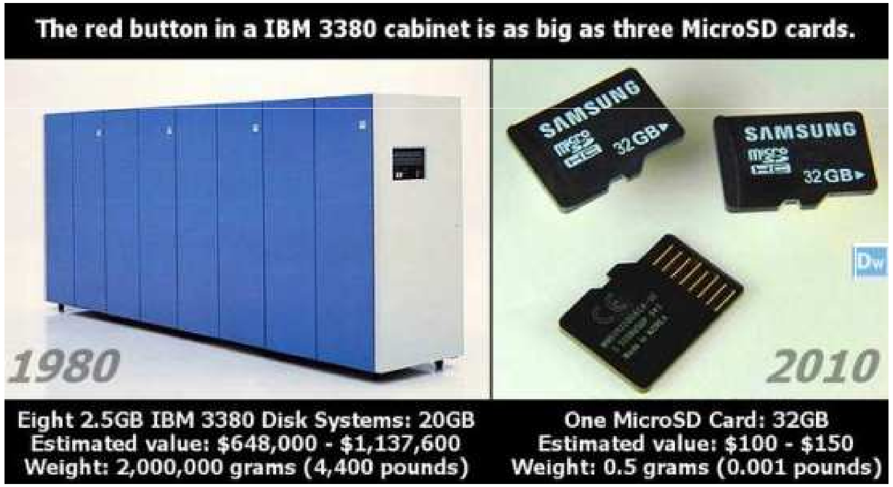

How does your state not excel?
Most invasive snails
Highest adult asthma rate
Lowest per capita public transit funding
Most sediment pollution in water
Fewest clean energy businesses
Highest overall rate of autism
Highest binge drinking rate
Greatest threat from Asian carp
Highest rate of mesothelioma
Highest rate of brain cancer
Highest rate of esophageal cancer
Highest rate of thyroid cancer
Top health risk from air pollution
Highest rate of breast cancer
Highest rate of pancreatic cancer
Most Superfund sites
Fewest national parks
Worst access to clean water
Most tornadoes per capita
Most energy use per capita
Lowest annual rain-fall
Slowest growth in green jobs
Highest rate of carbon monoxide poisoning
Most hog waste
Most toxic waste accidents
Most trash per capita
Most autism growth
Most dams in need of repair
Most coal mining deaths
Most ammonia in water
Most avalanche deaths
Most cases of pathogens in water
Site of worst dioxin disaster
Highest cancer death rate
Most sewer overflows
Fewest librarians per capita
Most smog
Most large farms irrigated by aquifers
Lowest amount of surface water
Eats fewest fruits and veggies
Most poultry waste
Most pest-icides in water
Origin of U.S. fire ant invasion
Most pollution from urban runoff
Fewest acres of organic farms
Most CO₂ emissions
Fastest loss of wetlands
Most boat wrecks
Most airports per capita
Most endangered species

BPM-Conseil

the Company behind the Vanilla project

1 - Big Data Landscape - Storage

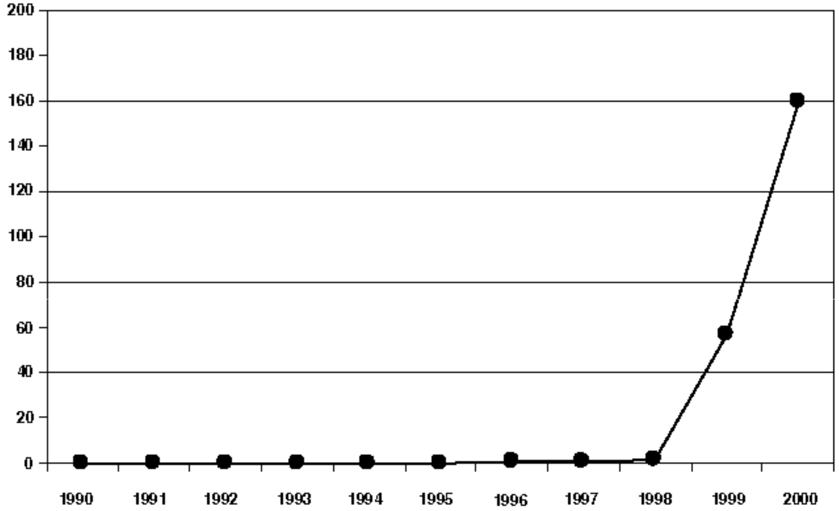
Storage … what for ?



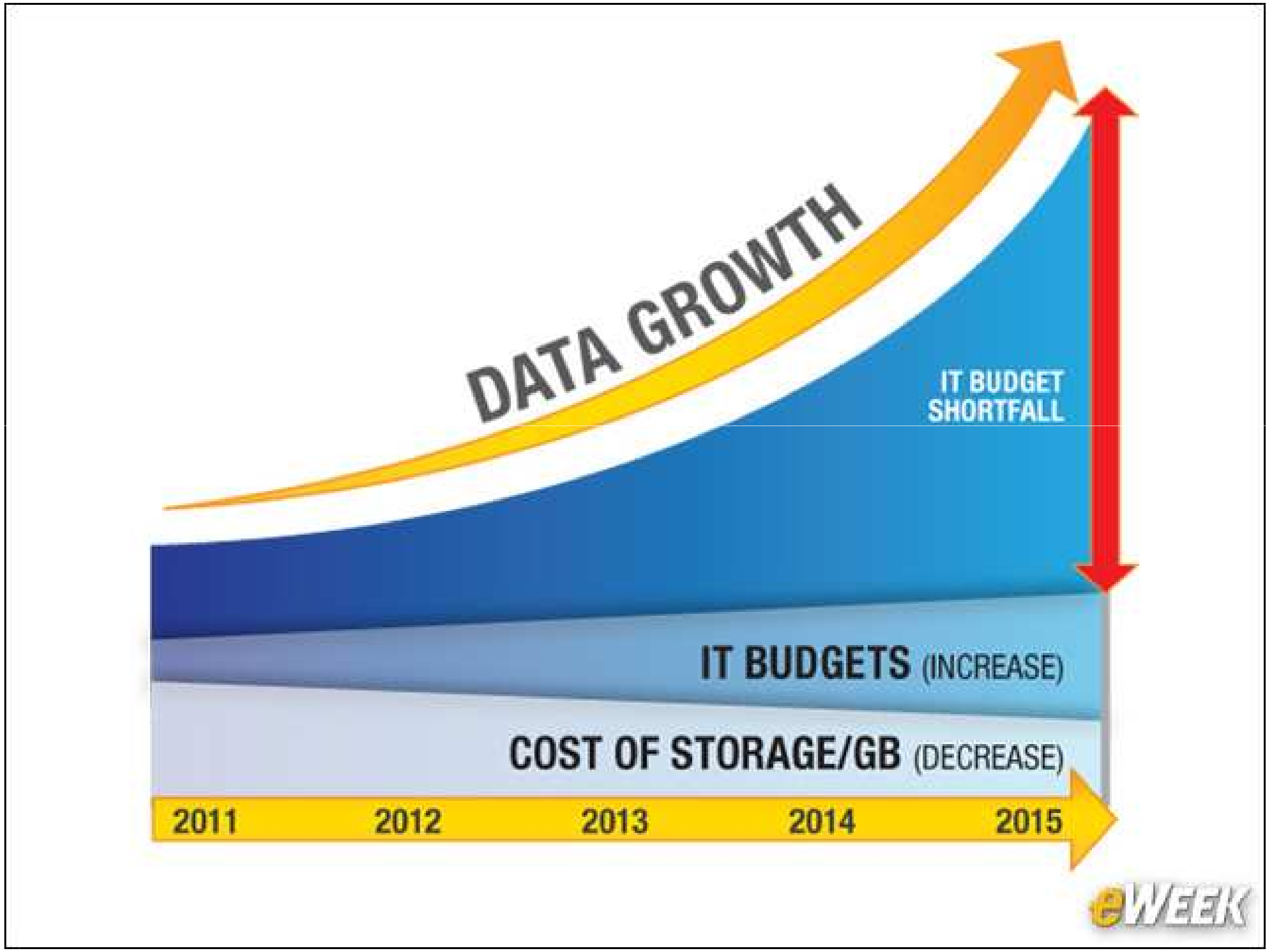
The red button in a IBM 3380 cabinet is as big as three MicroSD cards.

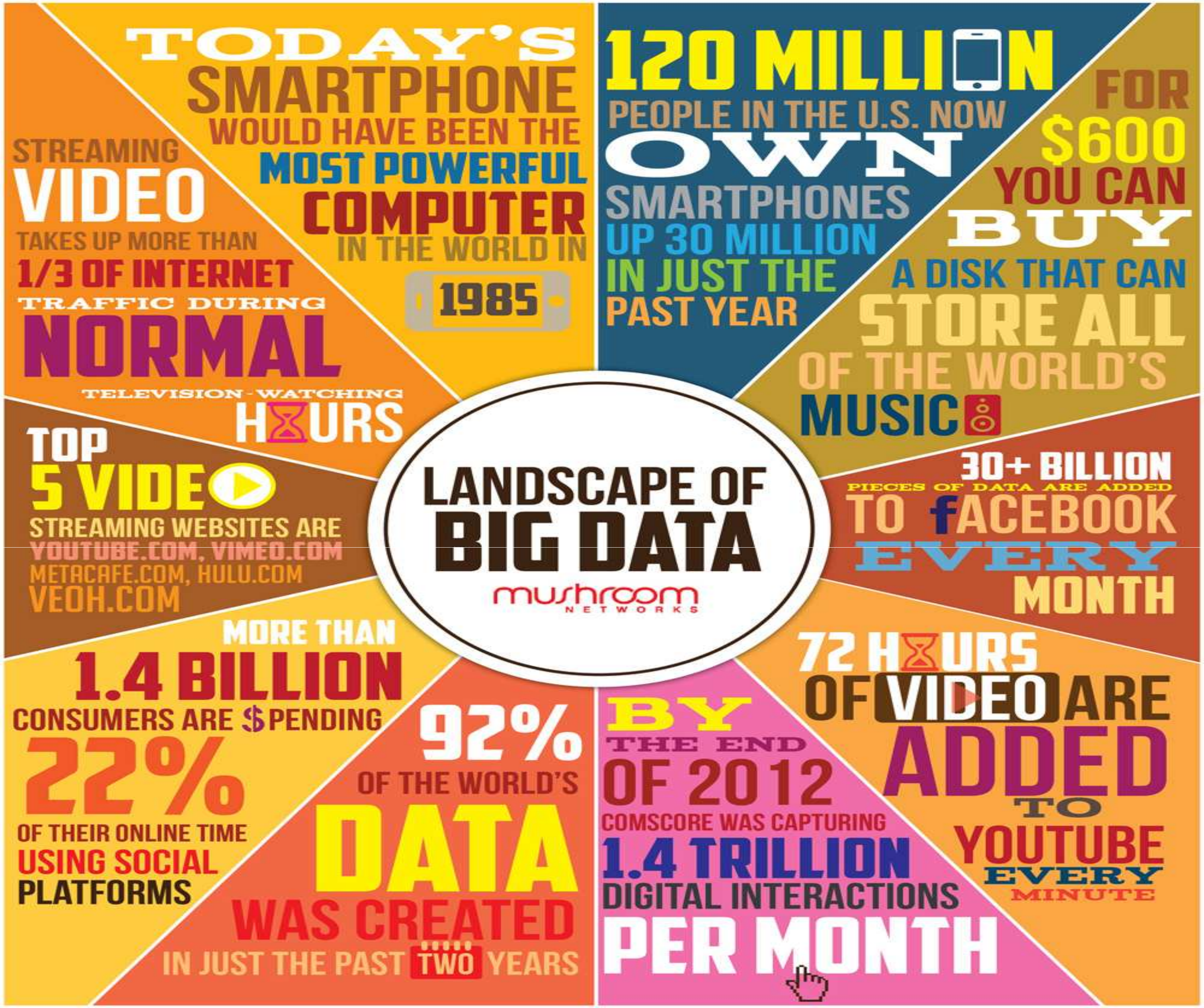
Eight 2.5GB IBM 3380 Disk Systems: 20GB
Estimated value: $648,000 - $1,137,600
Weight: 2,000,000 grams (4,400 pounds)

One MicroSD Card: 32GB
Estimated value: $100 - $150
Weight: 0.5 grams (0.001 pounds)



1 - Big Data Landscape - Storage





LANDSCAPE OF BIG DATA

mushroom NETWORKS

TODAY'S SMARTPHONE WOULD HAVE BEEN THE MOST POWERFUL COMPUTER IN THE WORLD IN 1985

120 MILLION PEOPLE IN THE U.S. NOW OWN SMARTPHONES UP 30 MILLION IN JUST THE PAST YEAR

FOR $600 YOU CAN BUY A DISK THAT CAN STORE ALL OF THE WORLD'S MUSIC

30+ BILLION PIECES OF DATA ARE ADDED TO FACEBOOK EVERY MONTH

72 HOURS OF VIDEO ARE ADDED TO YOUTUBE EVERY MINUTE

BY THE END OF 2012 COMSCORE WAS CAPTURING 1.4 TRILLION DIGITAL INTERACTIONS PER MONTH

92% OF THE WORLD'S DATA WAS CREATED IN JUST THE PAST TWO YEARS

MORE THAN 1.4 BILLION CONSUMERS ARE $PENDING 22% OF THEIR ONLINE TIME USING SOCIAL PLATFORMS

TOP 5 VIDEO STREAMING WEBSITES ARE YOUTUBE.COM, VIMEO.COM METACAFE.COM, HULU.COM VEOH.COM

STREAMING VIDEO TAKES UP MORE THAN 1/3 OF INTERNET TRAFFIC DURING NORMAL TELEVISION-WATCHING HOURS

BPM-Conseil
the Company behind the Vanilla project

1 - State of art of different database technologies (1/6)

Sql Database

Database that complies with SQL language and its sub-definition like pl/sql, transact-sql …, where data are structured in tables and columns

Usage : Erp, transactional application (Oltp), any kind of application with structured datas

Famous one : Oracle, Sybase, SqlServer, PostgreSql, Mysql

1 - State of art of different database technologies (2/6)

Column Database (vertical database)

Database that complies with SQL language where index strategy along with data compression are the key subjects

Usage : Bi Applications – Log Analysis application

Famous one : SybaseIQ, Vertica, Infobright, VectorWise, Exadata

INFOBRIGHT
Open Source Data Warehousing

VERTICA

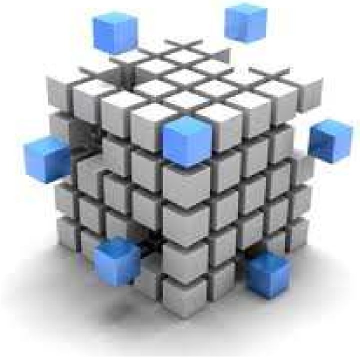
1 - State of art of different database technologies (3/6)

Olap Database (Cubes)

Database that complies with MDX language (and its Xmla implementation) where data is structured with dimensions and measures

Usage : Bi Applications (Olap)

Famous one : Hyperion Essbase, Cognos PowerPlay, Sap Bw



1 - State of art of different database technologies (4/6)

NoSql Database (Hadoop world, BigData)

Database that provides support for any kind of datastorage for structured data, unstructured data and any kind of document (image, message). Development language specific to read/write data (Api)

Usage : Large unstructured Web application (Twitter, Facebook : could be view like table with millions of columns)

Famous one : Cassandra, Hbase, MongoDb, Google BigTable

1 - State of art of different database technologies (5/6)

Application Database (Hidden structure)

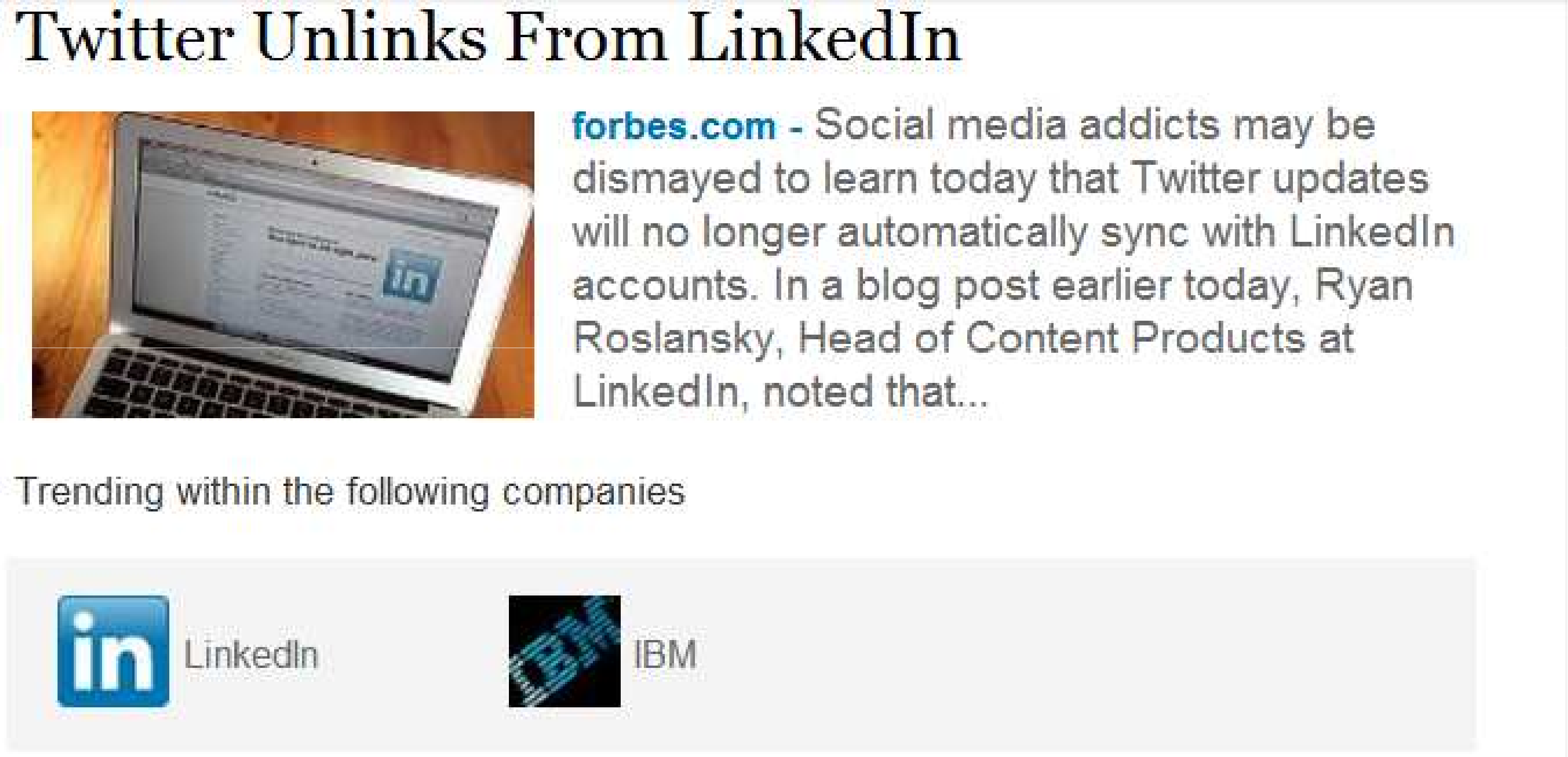
Database that provides access & exchange to their data through public Api, keeping hidden or too-complex the database structure.

Usage : Any kind of application, usually large Erp and large Web based cloud application

Warning : Api licence & usage (twitter example)

Famous one : GoogleAnalytics, Sap-Bapi (even Sql behind), twitter (Database & Application)

1 - State of art of different database technologies (6/6)



Twitter Unlinks From LinkedIn

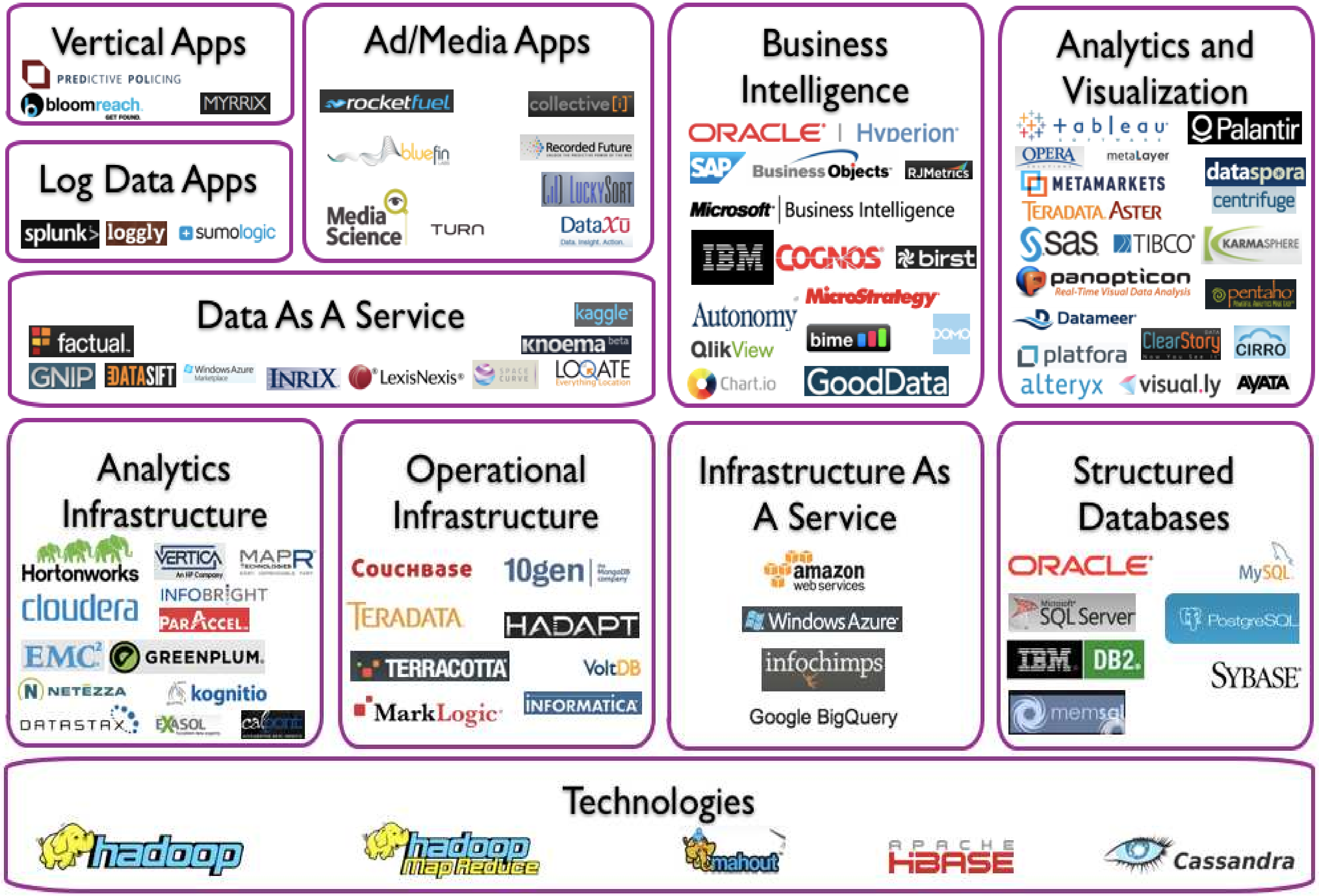
forbes.com - Social media addicts may be dismayed to learn today that Twitter updates will no longer automatically sync with LinkedIn accounts. In a blog post earlier today, Ryan Roslansky, Head of Content Products at LinkedIn, noted that...

Trending within the following companies

LinkedIn

IBM

Big Data Landscape



Copyright © 2012 Dave Feinleib

dave@vcdave.com

blogs.forbes.com/davefeinleib

BPM-Conseil

the Company behind the Vanilla project

2 - Database proposal in big data : nosql+column database, Hadoop (1/6)

Hadoop is the natural platform for BigData, providing the fundation and the Framework.

Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using a simple programming model

2 - Database proposal in big data : nosql+column database, Hadoop (2/6)

Column Database position is difficult, as Hadoop Hbase provides column indexation services. Column database editor trying to push « BigData offer » with some Hadoop components

3 majors databases for NoSql : Cassandra, Hbase and MongoDb, 2 of the 3 (Cassandra, Hbase) under the Apache fundation umbrella

2 - Database proposal in big data : nosql+column database, Hadoop (3/6)

Why Hadoop is unique

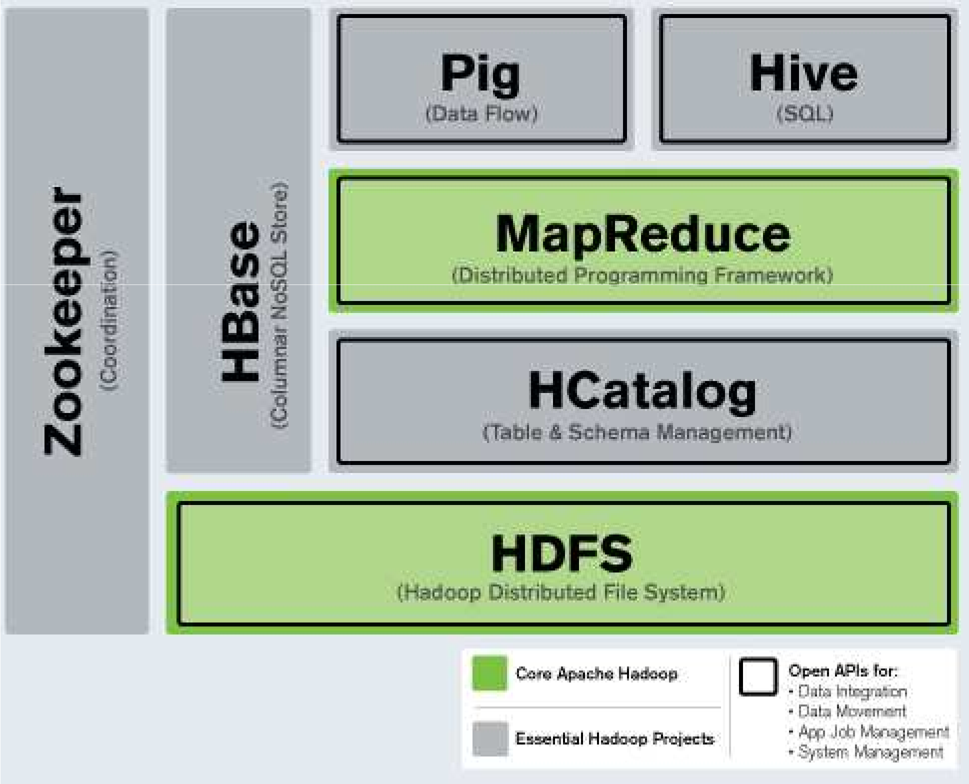
It's an Open Source Framework, without any commercial competitor at that level

New companies endorse Hadoop to build stable distribution with major Hadoop components : HortonWorks, MapR, Cloudera

Hadoop architecture target no single point of failure (multi master) , with usefull component (Zookeeper, MapReduce)

2 - Database proposal in big data : nosql+column database, Hadoop (4/6)

Hadoop Framework to manage big data set on the Cloud



2 - Database proposal in big data : nosql+column database, Hadoop (5/6)

Hadoop Components

It was inevitable that Hadoop would soon evolve to adopt some of the characteristics of a database

HBase, stepped in to partially fill the gap. It implements a column-oriented data store modeled on Google's Bigtable on top of Hadoop and HDFS

Pig, originally developed at Yahoo Research, is a high-level language for building MapReduce programs for Hadoop, thus simplifying the use of MapReduce. It is a data flow language that provides high-level commands.

2 - Database proposal in big data : nosql+column database, Hadoop (6/6)

Some Hadoop Components

Hive, initially a sub-project of Hadoop, evolved to provide a formal query capability. In effect, Hive turns Hadoop into a data warehouse-like system, allowing data summarization, ad hoc queries and the analysis of data stored by Hadoop

Hive holds metadata describing the contents of files and allows queries in HiveQL, a SQL-like language. It also allows MapReduce programmers to get around the limitations of HiveQL by plugging in MapReduce routines.

3 - Big Data, Search & Datamining

-Work on your data with Dataming component such as

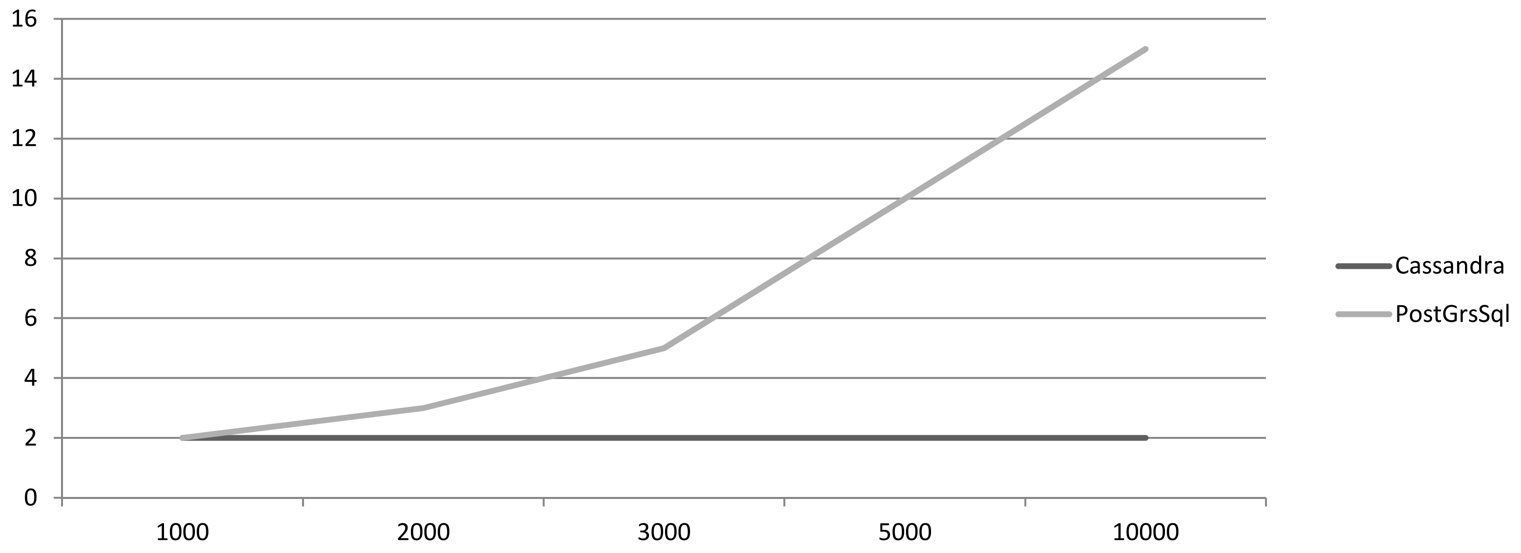
-R : http://www.revolutionanalytics.com/
- Weka : http://www.cs.waikato.ac.nz/ml/weka/
- Apache Mahout : http://mahout.apache.org/

- Manage different kind of data, using Hadoop Cassandra database

-Use existing Search Infrastructure like Solr/Lucene (Vanilla certified)
http://www.lucidworks.com/

3 - Big Data, Search & Vanilla

Response Time is stable with Cassandra – Impressive !!!



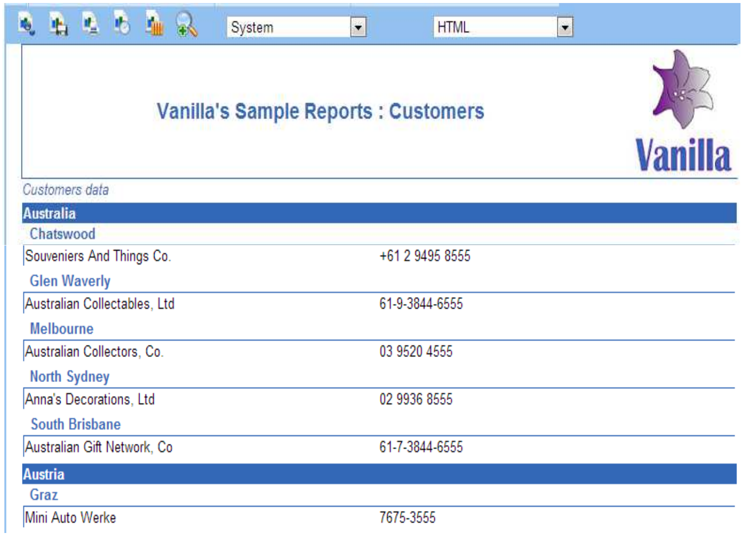
Cassandra

Vanilla

BPM-Conseil
the Company behind the Vanilla project

4 - Big Data & Data Visualisation

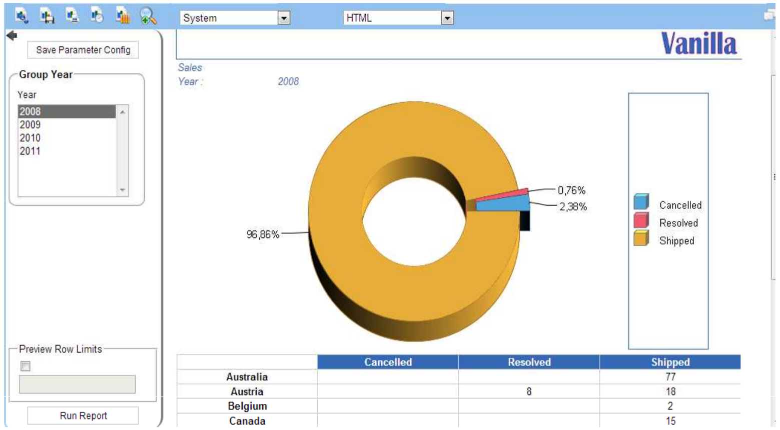
History of data visualisation



Vanilla's Sample Reports : Customers

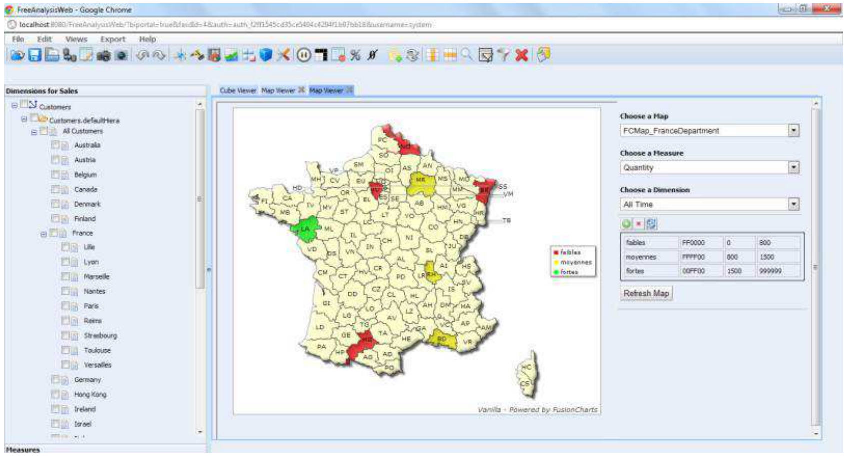
Customers data

Australia	
Chatswood	
Souveniers And Things Co.	+61 2 9495 8555
Glen Waverly	
Australian Collectables, Ltd	61-9-3844-6555
Melbourne	
Australian Collectors, Co.	03 9520 4555
North Sydney	
Anna's Decorations, Ltd	02 9936 8555
South Brisbane	
Australian Gift Network, Co	61-7-3844-6555
Austria	
Graz	
Mini Auto Werke	7675-3555



Sales
Year : 2008

	Cancelled	Resolved	Shipped
Australia			77
Austria		8	18
Belgium			2
Canada			15



Choose a Map: FCMap_FranceDepartment

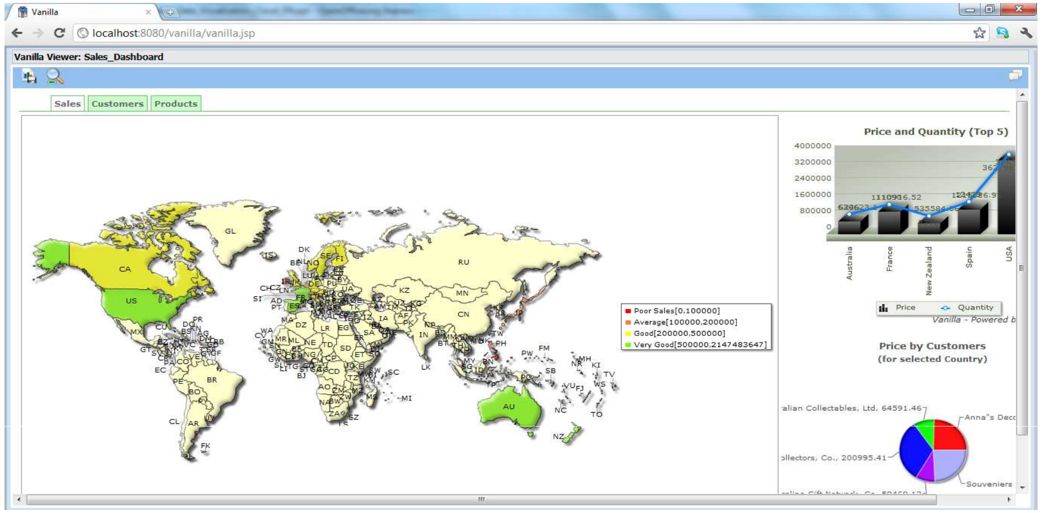
Choose a Measure: Quantity

Choose a Dimension: All Time

faibles	FF0000	0	800
moyennes	FFFF00	800	1500
fortes	00FF00	1500	999999

BPM-Conseil
the Company behind the Vanilla project

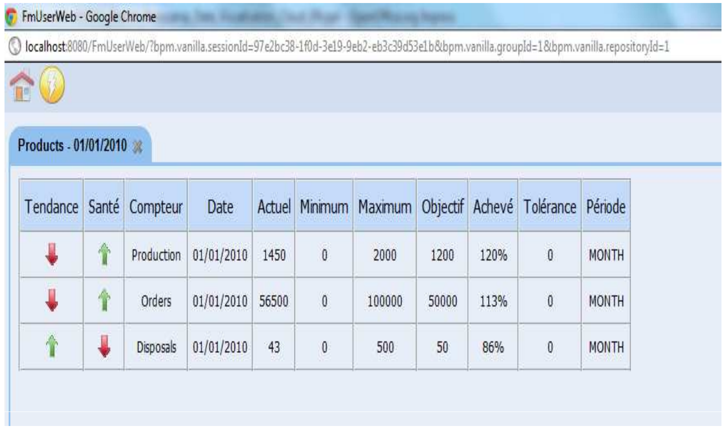
4 - Big Data & Data Visualisation



Vanilla

localhost:8080/vanilla/vanilla.jsp

Vanilla Viewer: Sales_Dashboard

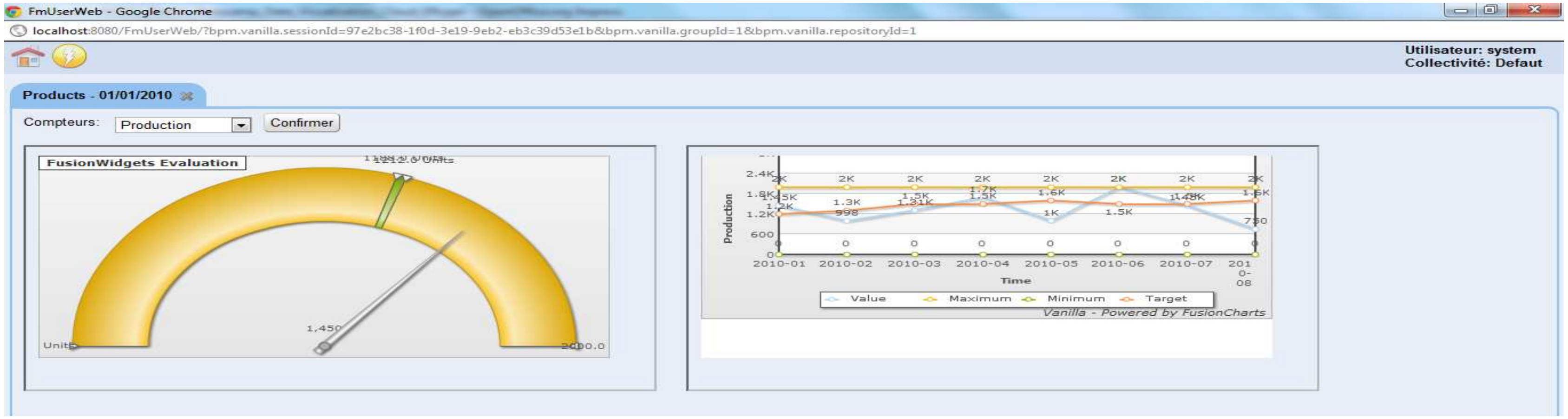


FmUserWeb - Google Chrome

localhost:8080/FmUserWeb/?bpm.vanilla.sessionId=97e2bc38-1f0d-3e19-9eb2-eb3c39d53e1b&bpm.vanilla.groupId=1&bpm.vanilla.repositoryId=1

Products - 01/01/2010

Tendance	Santé	Compteur	Date	Actuel	Minimum	Maximum	Objectif	Achevé	Tolérance	Période
↓	↑	Production	01/01/2010	1450	0	2000	1200	120%	0	MONTH
↓	↑	Orders	01/01/2010	56500	0	100000	50000	113%	0	MONTH
↑	↓	Disposals	01/01/2010	43	0	500	50	86%	0	MONTH



FmUserWeb - Google Chrome

localhost:8080/FmUserWeb/?bpm.vanilla.sessionId=97e2bc38-1f0d-3e19-9eb2-eb3c39d53e1b&bpm.vanilla.groupId=1&bpm.vanilla.repositoryId=1

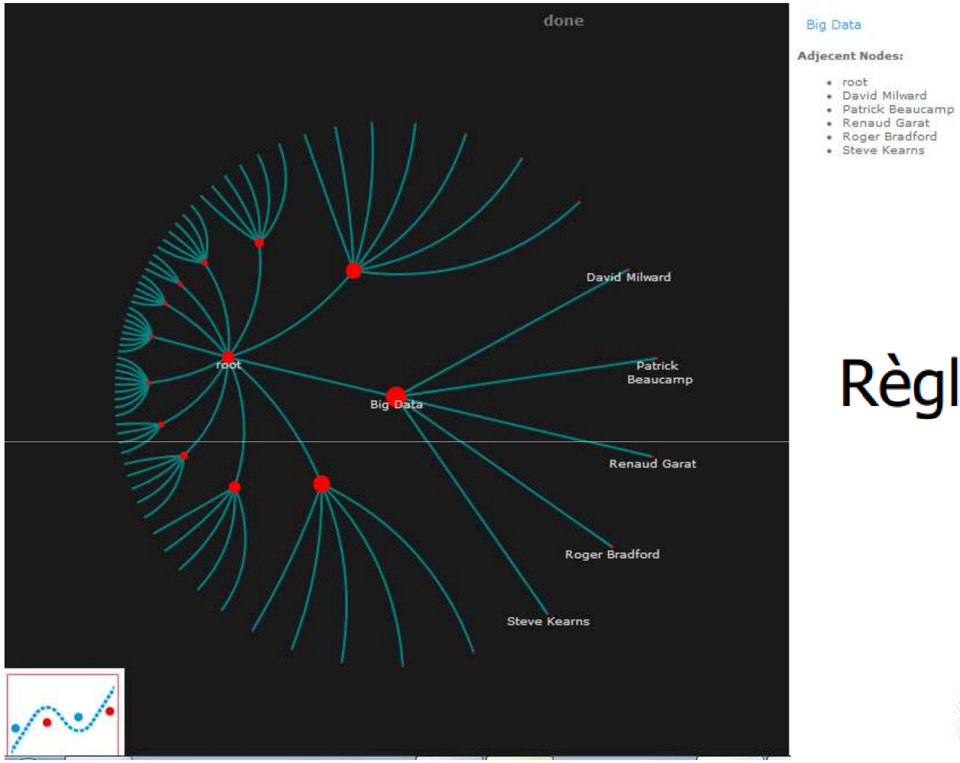
Utilisateur: system
Collectivité: Defaut

Products - 01/01/2010

Compteurs: Production

BPM-Conseil
the Company behind the Vanilla project

4 – Clustering & Data Mining



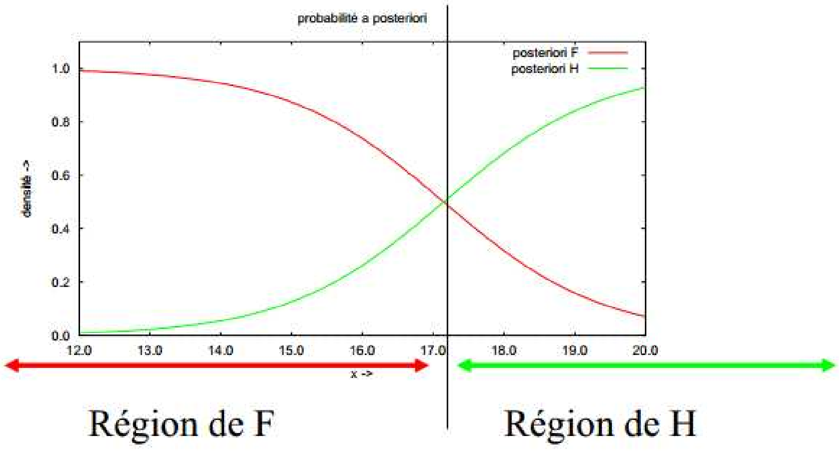
Big Data

Adjecent Nodes:

- root
- David Milward
- Patrick Beaucamp
- Renaud Garat
- Roger Bradford
- Steve Kearns

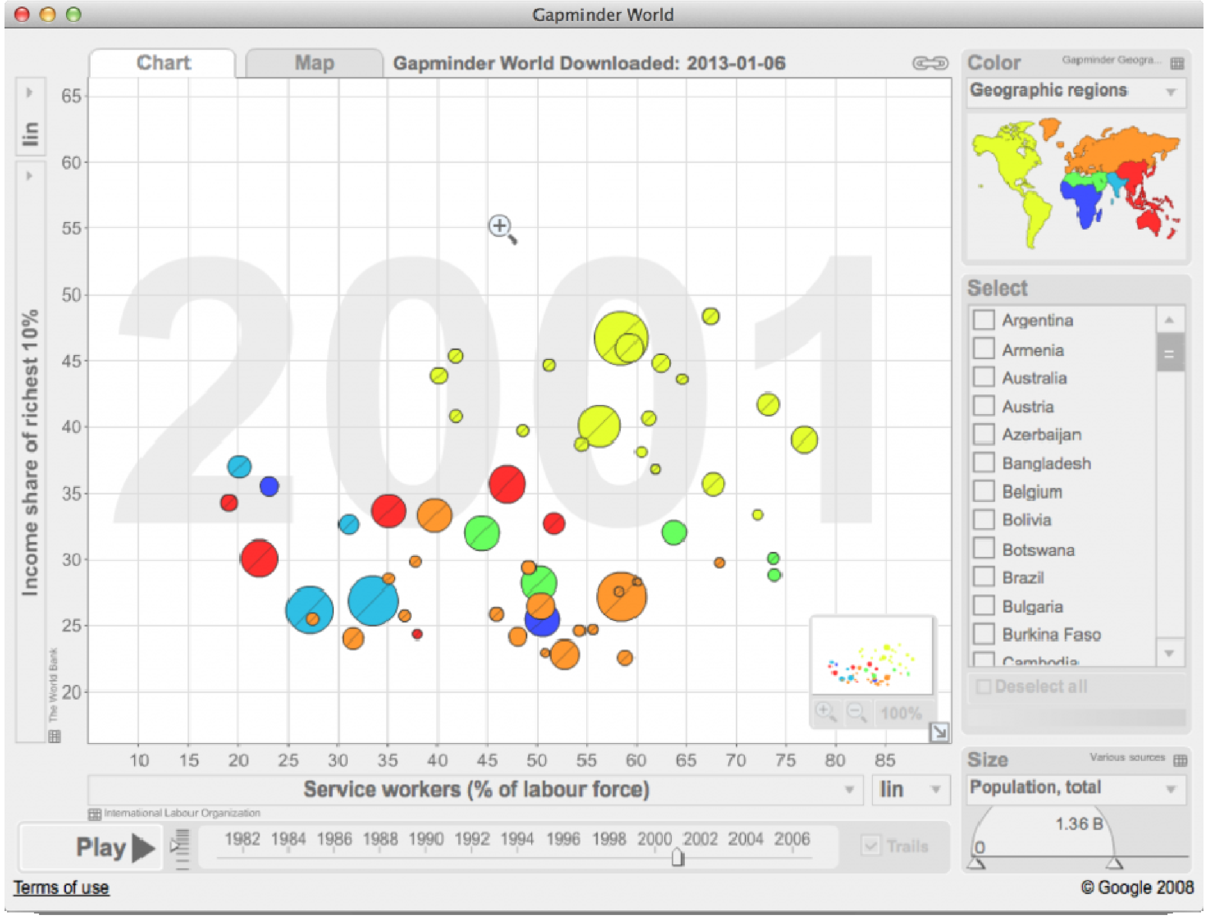
Règle de Bayes

$$\Pr(k / x) = \frac{\pi_k L_k(x)}{L(x)}$$



4 - Big Data & Data Visualisation

http://www.gapminder.org



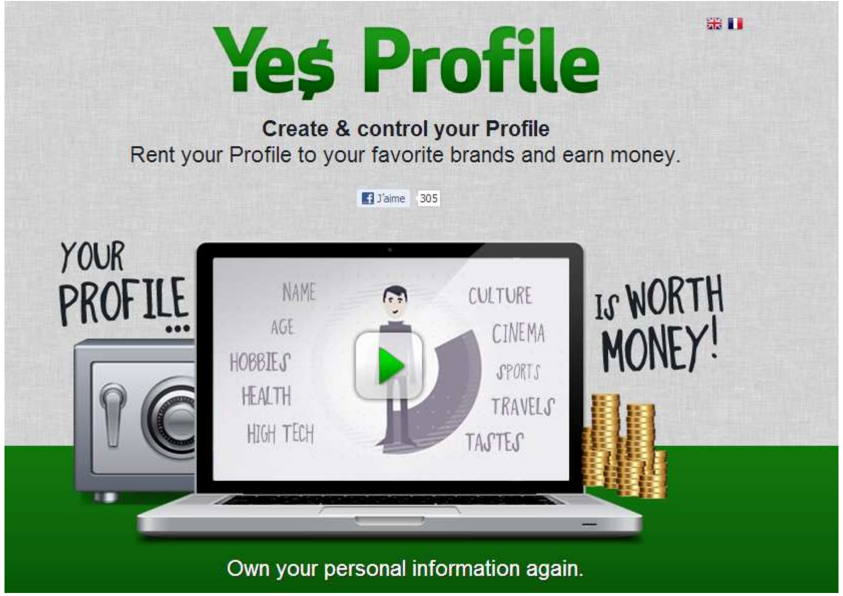
1 - Social Network

- Facebook – Linked – Twitter
  Your Private Life – Your Professional Life – What you think

  Intensive usage of BigData database (Cassandra) and Platform (Hadoop)
  Unstructured Data (chat) – Image and Video (face recognition)

  Just rethink what they do with your data

  http://www.yesprofile.com



Yes Profile

Create & control your Profile
Rent your Profile to your favorite brands and earn money.

Own your personal information again.

2 - Open Data Global Site

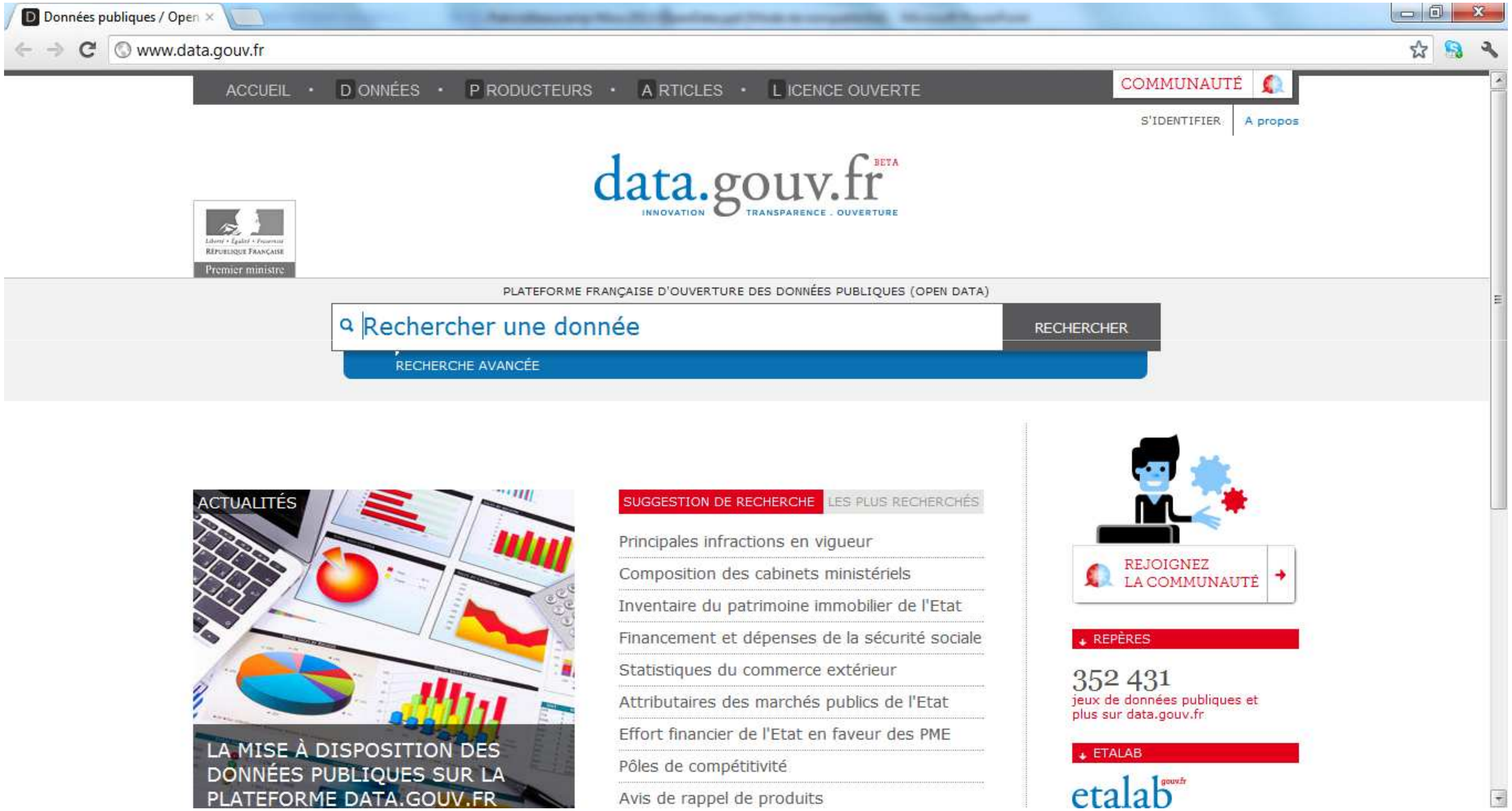
http://www.opengovdata.org/

The 8 Principles of Open Government Data

1. Data Must Be Complete
2. Data Must Be Primary
3. Data Must Be Timely
4. Data Must Be Accessible
5. Data Must Be Machine processable
6. Access Must Be Non-Discriminatory
7. Data Formats Must Be Non-Proprietary
8. Data Must Be License-free

2 - Open Data Government Site (1/4)

French Portal : http://www.data.gouv.fr/ - 352’431 dataset



Données publiques / Open

www.data.gouv.fr

ACCUEIL · DONNÉES · PRODUCTEURS · ARTICLES · LICENCE OUVERTE

COMMUNAUTÉ

S'IDENTIFIER | A propos

data.gouv.fr BETA

INNOVATION TRANSPARENCE . OUVERTURE

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
Premier ministre

PLATEFORME FRANÇAISE D'OUVERTURE DES DONNÉES PUBLIQUES (OPEN DATA)

Rechercher une donnée | RECHERCHER

RECHERCHE AVANCÉE

ACTUALITÉS

LA MISE À DISPOSITION DES DONNÉES PUBLIQUES SUR LA PLATEFORME DATA.GOUV.FR

SUGGESTION DE RECHERCHE | LES PLUS RECHERCHÉS

- Principales infractions en vigueur
- Composition des cabinets ministériels
- Inventaire du patrimoine immobilier de l'Etat
- Financement et dépenses de la sécurité sociale
- Statistiques du commerce extérieur
- Attributaires des marchés publics de l'Etat
- Effort financier de l'Etat en faveur des PME
- Pôles de compétitivité
- Avis de rappel de produits

REJOIGNEZ LA COMMUNAUTÉ

REPÈRES

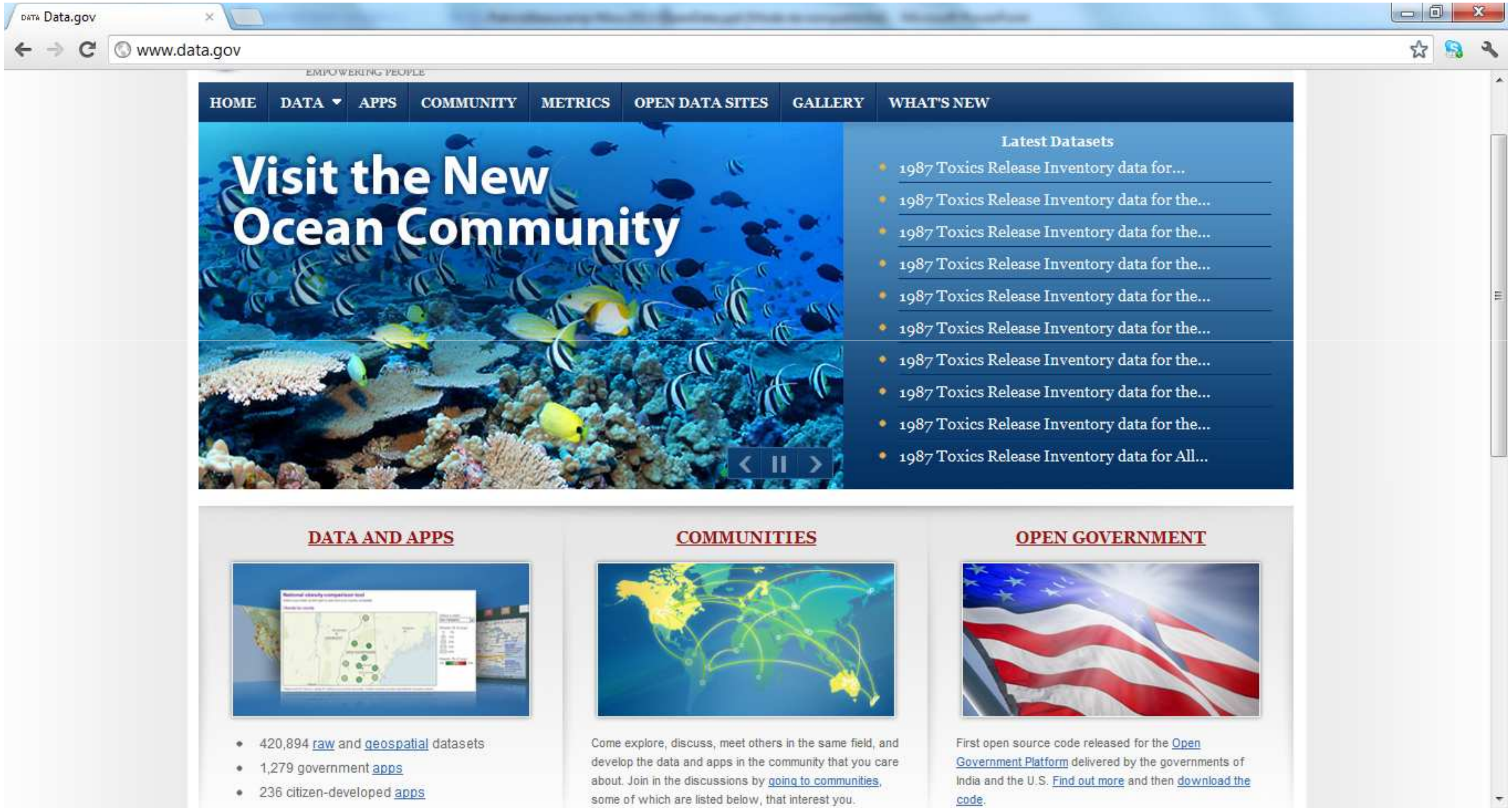
352 431
jeux de données publiques et plus sur data.gouv.fr

ETALAB

etalab gouv.fr

2 - Open Data Government Site (2/4)

US Portal : http://www.data.gov/ - 420’894 dataset



Latest Datasets

- 1987 Toxics Release Inventory data for...
- 1987 Toxics Release Inventory data for the...
- 1987 Toxics Release Inventory data for the...
- 1987 Toxics Release Inventory data for the...
- 1987 Toxics Release Inventory data for the...
- 1987 Toxics Release Inventory data for the...
- 1987 Toxics Release Inventory data for the...
- 1987 Toxics Release Inventory data for the...
- 1987 Toxics Release Inventory data for the...
- 1987 Toxics Release Inventory data for All...

DATA AND APPS

- 420,894 raw and geospatial datasets
- 1,279 government apps
- 236 citizen-developed apps

COMMUNITIES

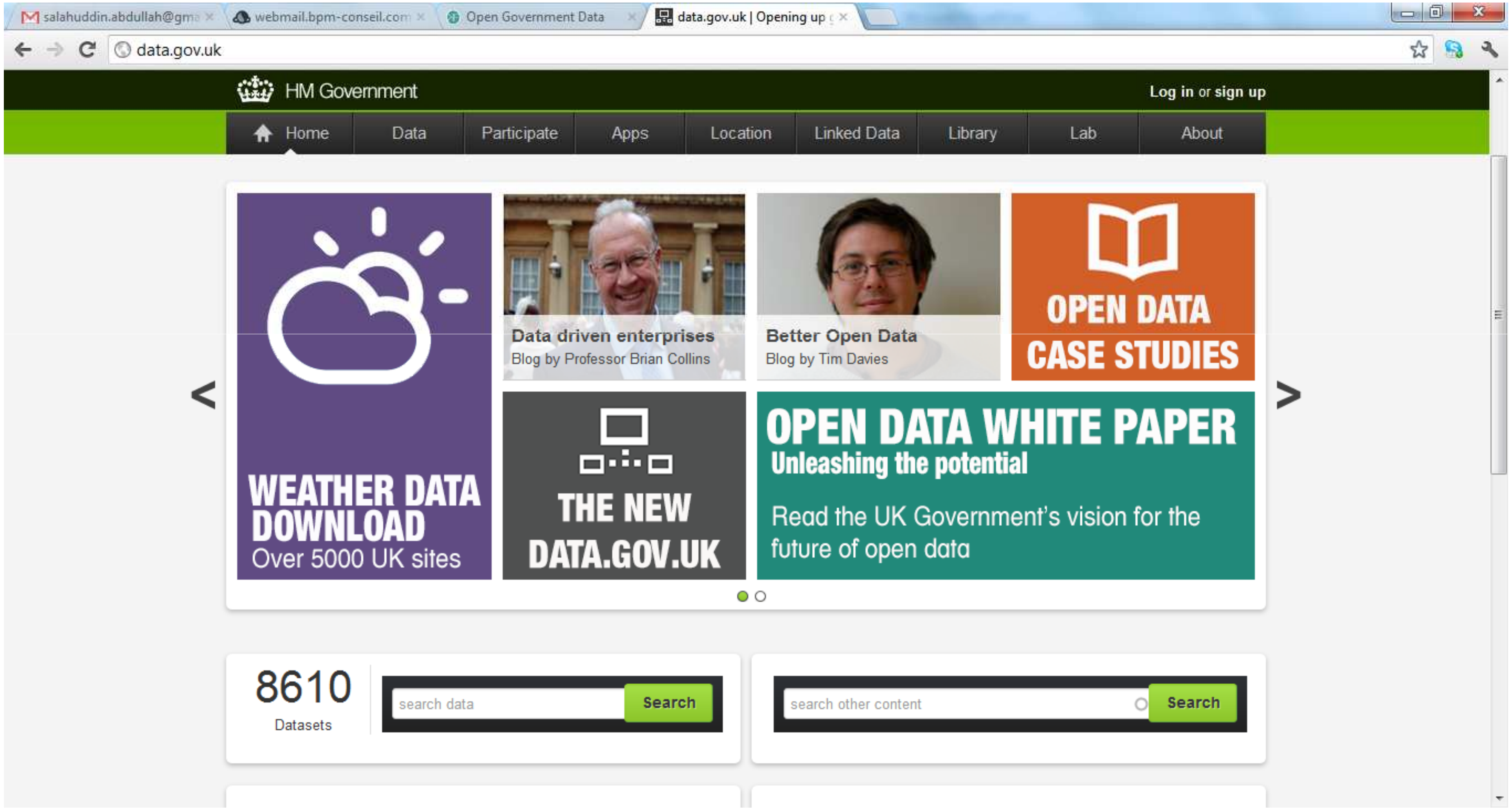
Come explore, discuss, meet others in the same field, and develop the data and apps in the community that you care about. Join in the discussions by going to communities, some of which are listed below, that interest you.

OPEN GOVERNMENT

First open source code released for the Open Government Platform delivered by the governments of India and the U.S. Find out more and then download the code.

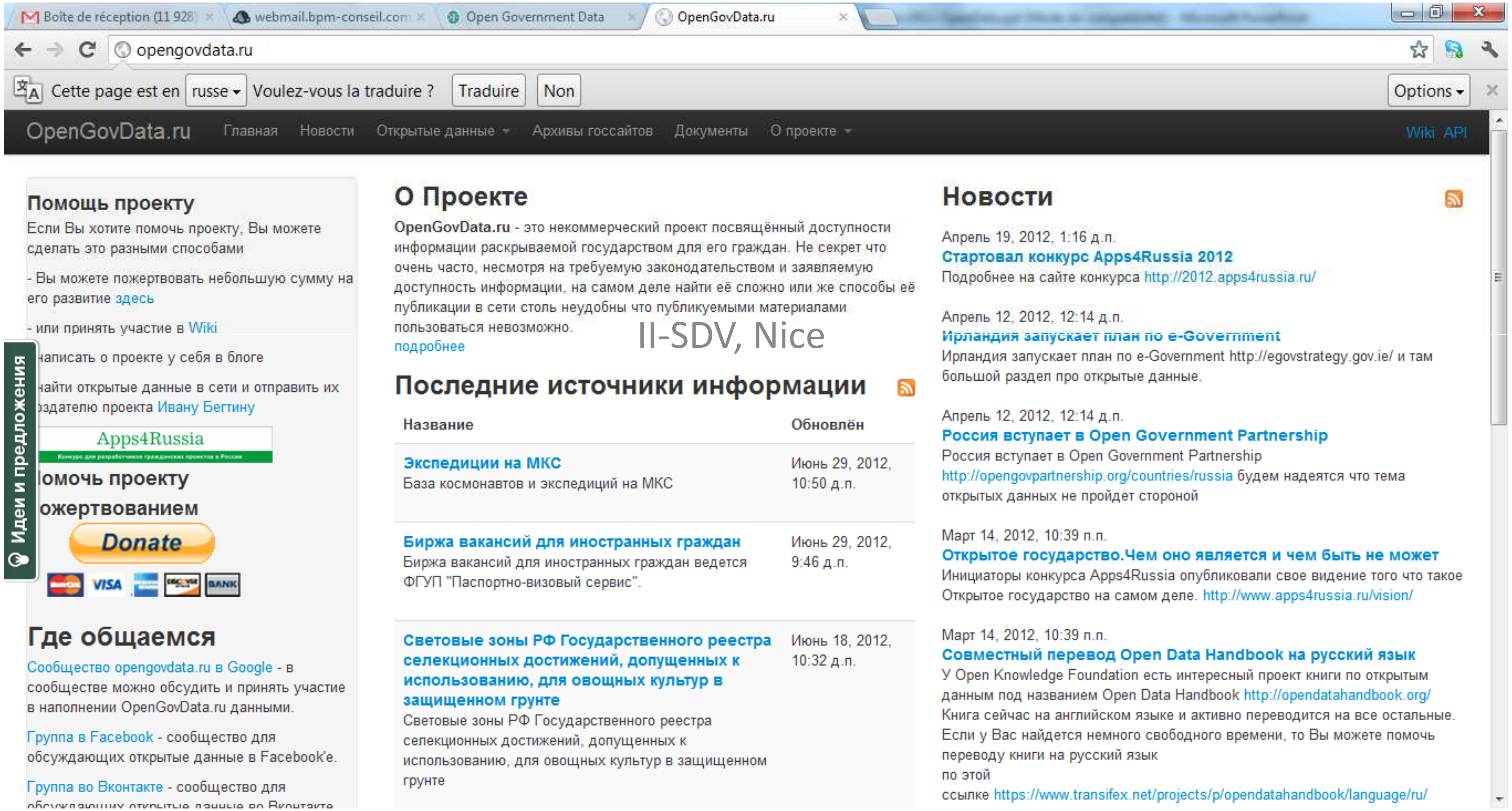
2 - Open Data Government Site (3/4)

UK Portal : http://data.gov.uk/ - 8’610 dataset



2 - Open Data Government Site (4/4)

Russian Portal : http://opengovdata.ru/ - ?? dataset



Помощь проекту

Если Вы хотите помочь проекту, Вы можете сделать это разными способами

- Вы можете пожертвовать небольшую сумму на его развитие здесь

- или принять участие в Wiki

написать о проекте у себя в блоге

найти открытые данные в сети и отправить их оздателю проекта Ивану Бегтину

омочь проекту ожертвованием

Где общаемся

Сообщество opengovdata.ru в Google - в сообществе можно обсудить и принять участие в наполнении OpenGovData.ru данными.

Группа в Facebook - сообщество для обсуждающих открытые данные в Facebook'e.

Группа во Вконтакте - сообщество для обсуждающих открытые данные во Вконтакте

О Проекте

OpenGovData.ru - это некоммерческий проект посвящённый доступности информации раскрываемой государством для его граждан. Не секрет что очень часто, несмотря на требуемую законодательством и заявляемую доступность информации, на самом деле найти её сложно или же способы её публикации в сети столь неудобны что публикуемыми материалами пользоваться невозможно.

подробнее

II-SDV, Nice

Последние источники информации

Название	Обновлён
Экспедиции на МКС База космонавтов и экспедиций на МКС	Июнь 29, 2012, 10:50 д.п.
Биржа вакансий для иностранных граждан Биржа вакансий для иностранных граждан ведется ФГУП "Паспортно-визовый сервис".	Июнь 29, 2012, 9:46 д.п.
Световые зоны РФ Государственного реестра селекционных достижений, допущенных к использованию, для овощных культур в защищенном грунте Световые зоны РФ Государственного реестра селекционных достижений, допущенных к использованию, для овощных культур в защищенном грунте	Июнь 18, 2012, 10:32 д.п.

Новости

Апрель 19, 2012, 1:16 д.п.
Стартовал конкурс Apps4Russia 2012
Подробнее на сайте конкурса http://2012.apps4russia.ru/

Апрель 12, 2012, 12:14 д.п.
Ирландия запускает план по e-Government
Ирландия запускает план по e-Government http://egovstrategy.gov.ie/ и там большой раздел про открытые данные.

Апрель 12, 2012, 12:14 д.п.
Россия вступает в Open Government Partnership
Россия вступает в Open Government Partnership http://opengovpartnership.org/countries/russia будем надеятся что тема открытых данных не пройдет стороной

Март 14, 2012, 10:39 п.п.
Открытое государство.Чем оно является и чем быть не может
Инициаторы конкурса Apps4Russia опубликовали свое видение того что такое Открытое государство на самом деле. http://www.apps4russia.ru/vision/

Март 14, 2012, 10:39 п.п.
Совместный перевод Open Data Handbook на русский язык
У Open Knowledge Foundation есть интересный проект книги по открытым данным под названием Open Data Handbook http://opendatahandbook.org/ Книга сейчас на английском языке и активно переводится на все остальные. Если у Вас найдется немного свободного времени, то Вы можете помочь переводу книги на русский язык
по этой
ссылке https://www.transifex.net/projects/p/opendatahandbook/language/ru/

2 - Open Data Government - Bonus

European Decision on Open Data for Government (April 10)

- https://ec.europa.eu/digital-agenda/en/public-sector-information-raw-data-new-services-and-products

- http://gigaom.com/2013/04/10/european-governments-agree-to-open-up-public-data/

II-SDV, Nice

European Decision on Mine (Diamond, Gold, Uranium), Gaz, Oil (April 13)

- http://eurodad.org/wp-content/uploads/2011/11/CBC-report_french-21.pdf

Thank You!