OCTO Technology
There is a better way

Anticiper et prédire les sinistres avec une approche Big Data

Julien Cabot
Directeur Big Data Analytics
OCTO
jcabot@octo.com
@julien_cabot

50, avenue des Champs-Elysées
75008 Paris - FRANCE

Tél : +33 (0)1 58 56 10 00
Fax : +33 (0)1 58 56 10 01
www.octo.com

Internet comme source de données…

Bloomberg Our Company | Professional | Anywhere

HOME QUICK NEWS OPINION MARKET DATA PERSONAL FINANCE TECH POLITICS SUSTAINAE

Hedge Fund Will Track Twitter to Predict Stock Moves

By Jack Jordan - Dec 22, 2010 4:20 PM GMT+0100

0 COMMENTS QUEUE

Derwent Capital Markets, a family- owned hedge fund, will offer investors the chance to use Twitter Inc. posts to gauge the mood of the stockmarket, said co-owner Paul Hawtin.

The Derwent Absolute Return Fund Ltd., set to start trading in February with an initial 25 million pounds ($39 million) under management, will follow posts on the social-networking website. A trading model will highlight when the number of times words on Twitter such as "calm" rise above or below average.

Enlarge image

A paper published in October said the number of emotional words on Twitter

A paper by the University of Manchester and Indiana University published in October said the number of emotional words on Twitter could be used to predict daily moves in the Dow Jones Industrial Average. A change in emotions expressed online would be followed between two and six days later by a move in the index, the researchers said, and this information let them predict its movements with 87.6 percent accuracy.

The New York Times Internet

WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY SCIENCE HEALTH

Search Technology Go

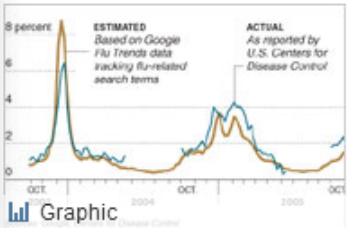
Inside Technology Internet | Start-Ups | Business Computing

Google Uses Searches to Track Flu's Spread

By MIGUEL HELFT
Published: November 11, 2008

SAN FRANCISCO — There is a new common symptom of the flu, in addition to the usual aches, coughs, fevers and sore throats. Turns out a lot of ailing Americans enter phrases like "flu symptoms" into Google and other search engines before they call their doctors.

Multimedia



That simple act, multiplied across millions of keyboards in homes around the country, has given rise to a new early warning system for fast-spreading flu outbreaks, called Google Flu Trends.

Tests of the new Web tool from

Internet, comme la voix des « mass markets »

Les sources d'information utilisées en matière de santé

Question : De manière générale, quelles sont parmi les suivantes toutes les sources d'information que vous utilisez lorsque vous cherchez des informations en matière de santé ?

Base : A tous

Résultats supérieurs à 100, plusieurs réponses possibles

**Item non suggéré*

Ipsos Public Affairs 4

Conseil National de l'Ordre des Médecins – Les conséquences des usages d'internet sur les relations patients – médecins
Avril 2010

71% à propos de
- Maladie
- Symptome
- Médicament
- Avis / opinion

Les principales sources sont les forums, et non les réseaux sociaux

Quels bénéfices pour les Assureurs?

- **Comprendre** les centres d'intérêt des communautés patients
- **Anticiper** l'effet psycho-social d'Internet
- **Prédire** les tendances d'évolution en suivant l'évolution de signaux faibles

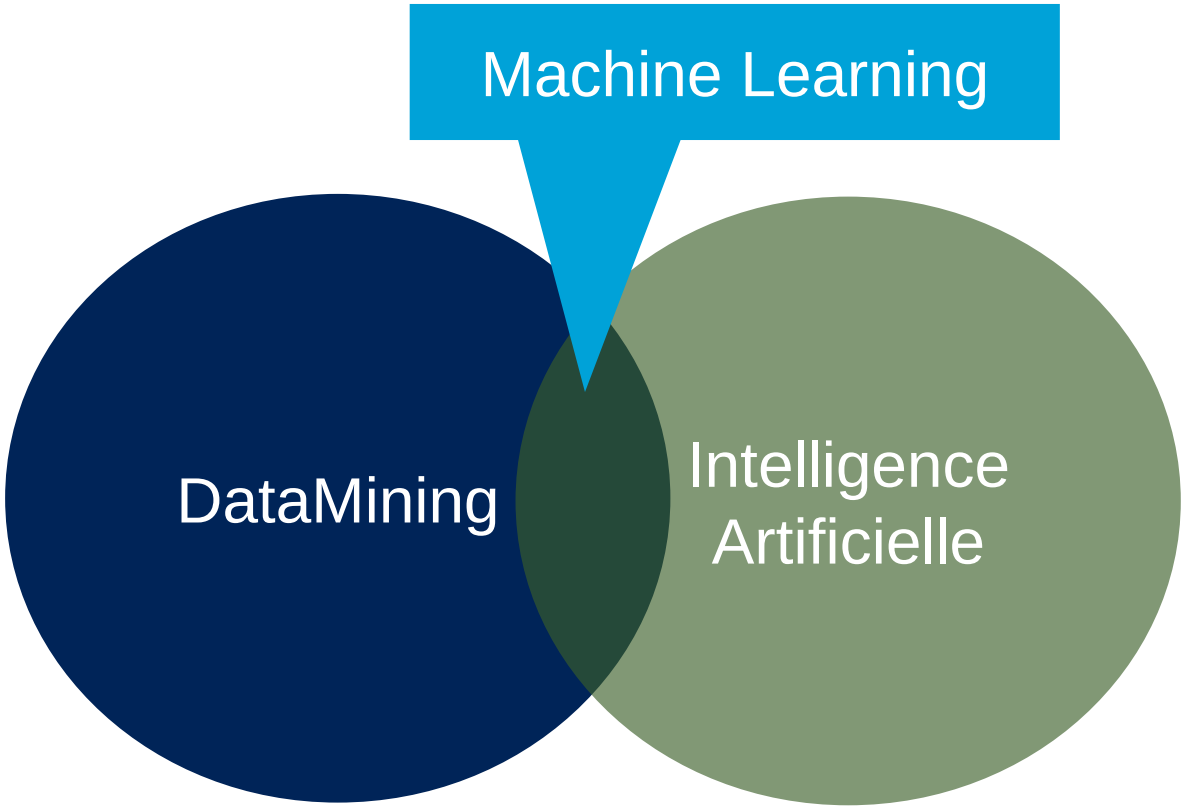
Comment procéder?

Le problème à résoudre

- Comprendre le **champ sémantique** de la Santé … utilisé sur Internet
- Identifier les corrélations entre l'évolution des sinistres et des … **millions de variables externes** non-identifiées
- Trouver des variables corrélées … **anticipant les sinistres**

Dépasser les limites humaines

Traiter des millions de corrélations n'est plus possible à l'échelle humaine



Trouver des corrélations anticipatives

Forum

Analyse du langage naturel des messages par date et **interprétation sémantique**

Google Trends LABS

Volume de recherche Google sur des mots clefs de symptomes et de médicaments

data.gouv.fr BETA
INNOVATION TRANSPARENCE . OUVERTURE

Evolutions des données socio-économiques issues de l'Open Data

Evolution de termes de Santé sur les forums

Evolution des recherches de termes de Santé

Evolution du contexte socio-économique

Sinistres Santé par typologies

Matrice de corrélations "lagged"

Construire une base de données sémantique pour la santé

Forum → Analyse des messages par date → Text Mining → Tendances d'évolution de mots clefs santé

Comment tagger les mots clefs du domaine de la santé?

1-Construire manuellement une liste de mots clefs

Google AdWords

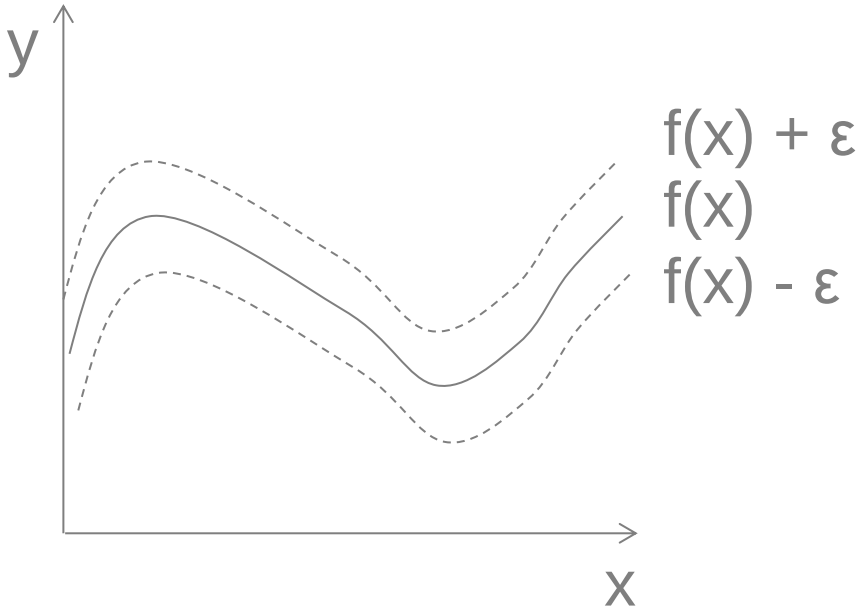
2-Enrichir la liste avec les mots clefs les plus cherchés

3-Apprendre automatiquement des catégories Santé

Base de données de mots clefs santé

Comment trouver des correlations dans des séries de temps?

- Comparer l'évolution de variables exogènes et des sinistres dans le temps
- Trouver des régressions non linéaires et identifier fonction prédictive polymorphique f(x) à partir du dataset avec les Support Vector Regressions (SVR)



Problème à résoudre

$$\min_{w} \frac{1}{2} w^T . w$$

$$\begin{cases} y_i - (w^T \cdot \phi(x) + b) \leq \varepsilon \\ (w^T \cdot \phi(x) + b) - y_i \leq \varepsilon \end{cases}$$

Résolution

- Descente de gradient stochastique
- Test de la réponse au travers du coef. de determination R^2

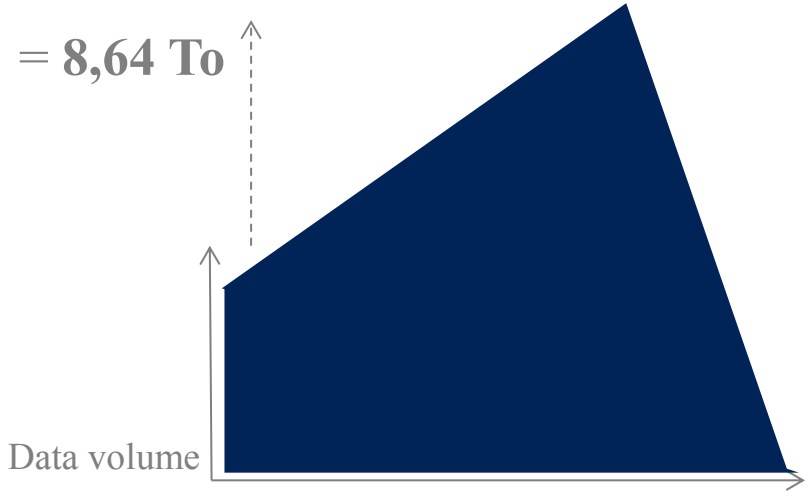
Les librairies de Machine Learning sont précieuses!

Volume de données à traiter

- **Le volume de données à analyser n'est pas si important (~ 10 Go)**
- **Agrégation de données**
  Ex. Select … Group By Date

- **Recherche corrélation**
  Ex. SVR computing

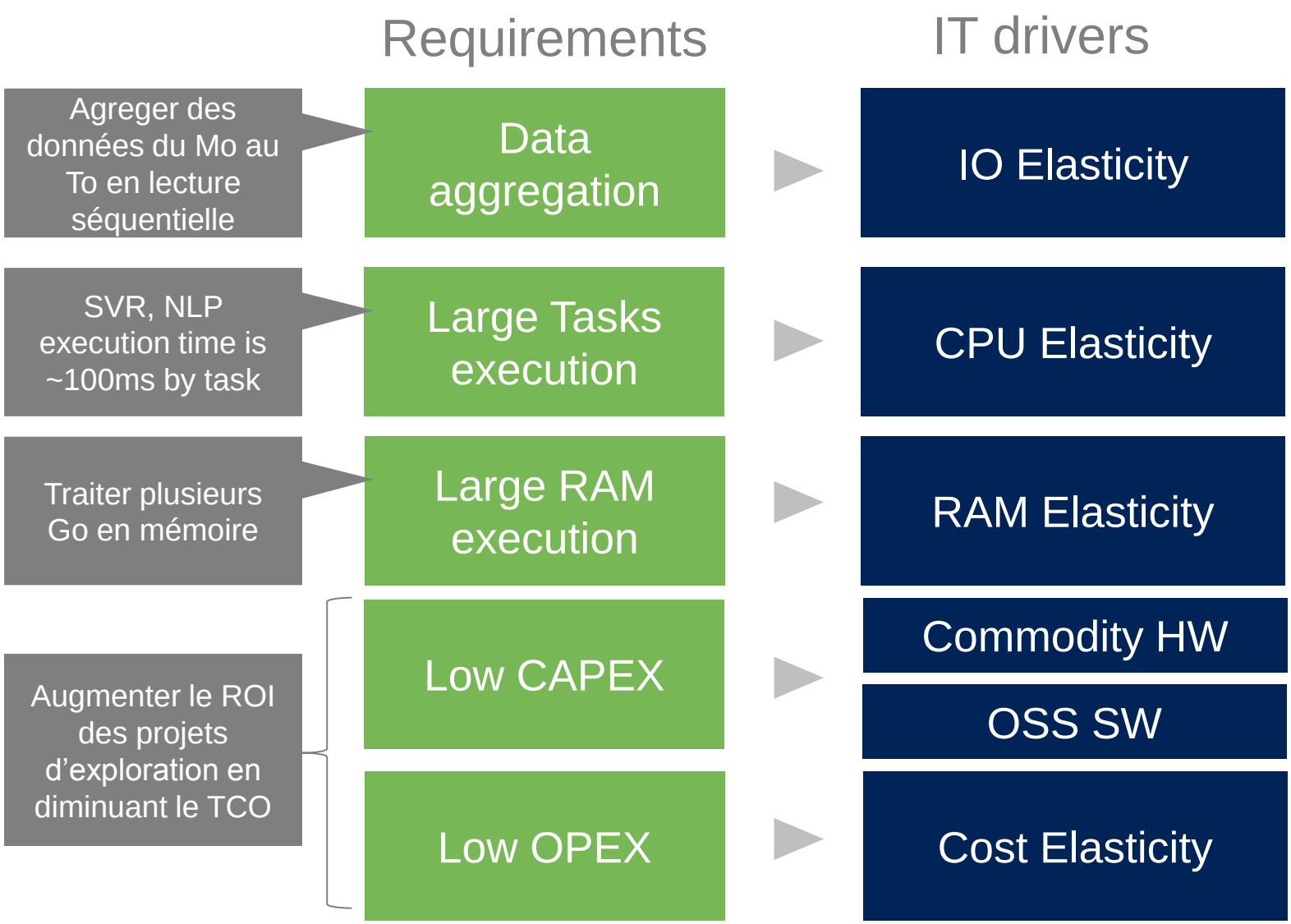
$\sim 5\text{Go} \cdot 12^3 = \mathbf{8{,}64\ To}$



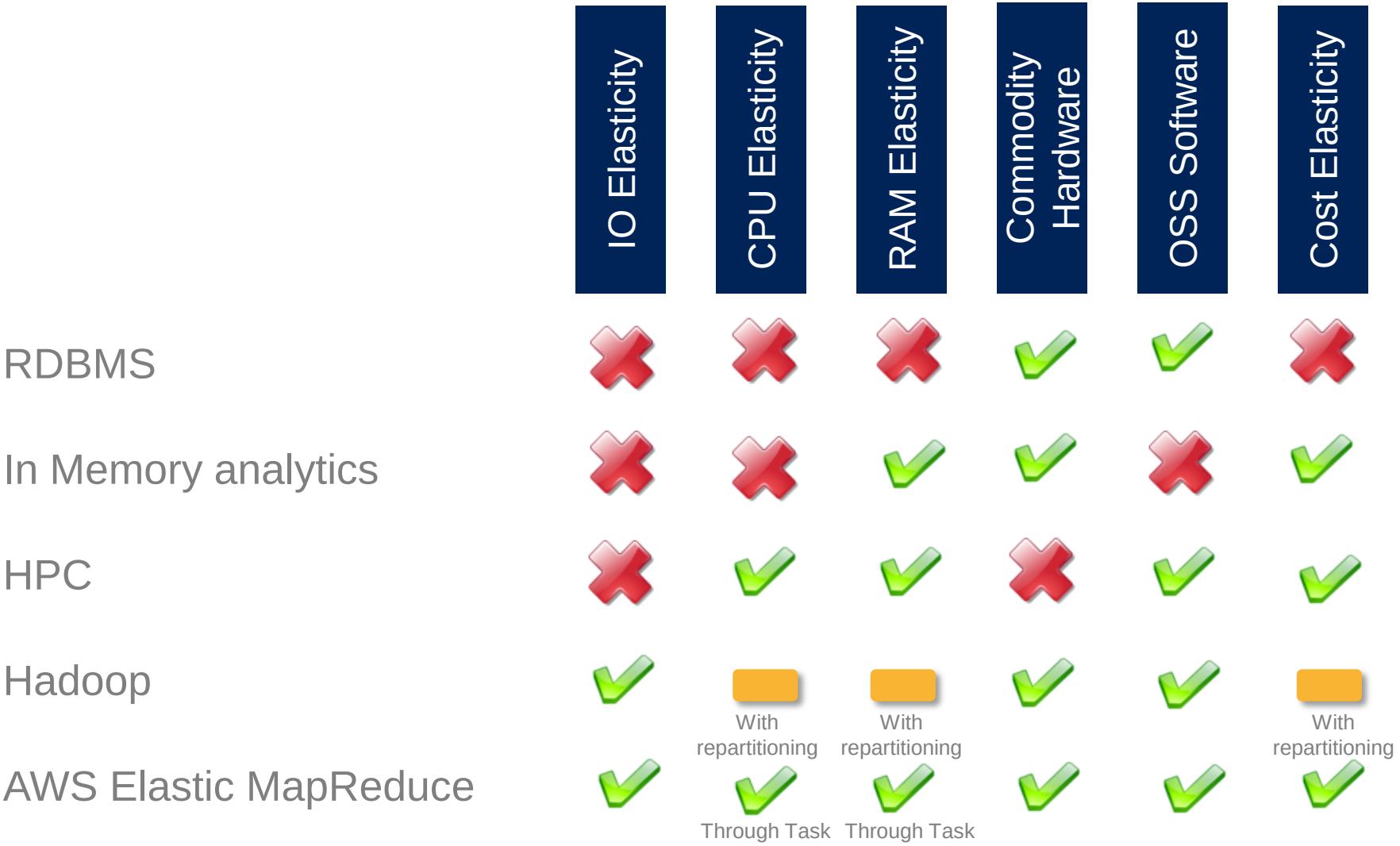
Le Parallel Computing permet de diviser le temps de traitement et le Cloud Computing les coûts !

Avec quelle plateforme?

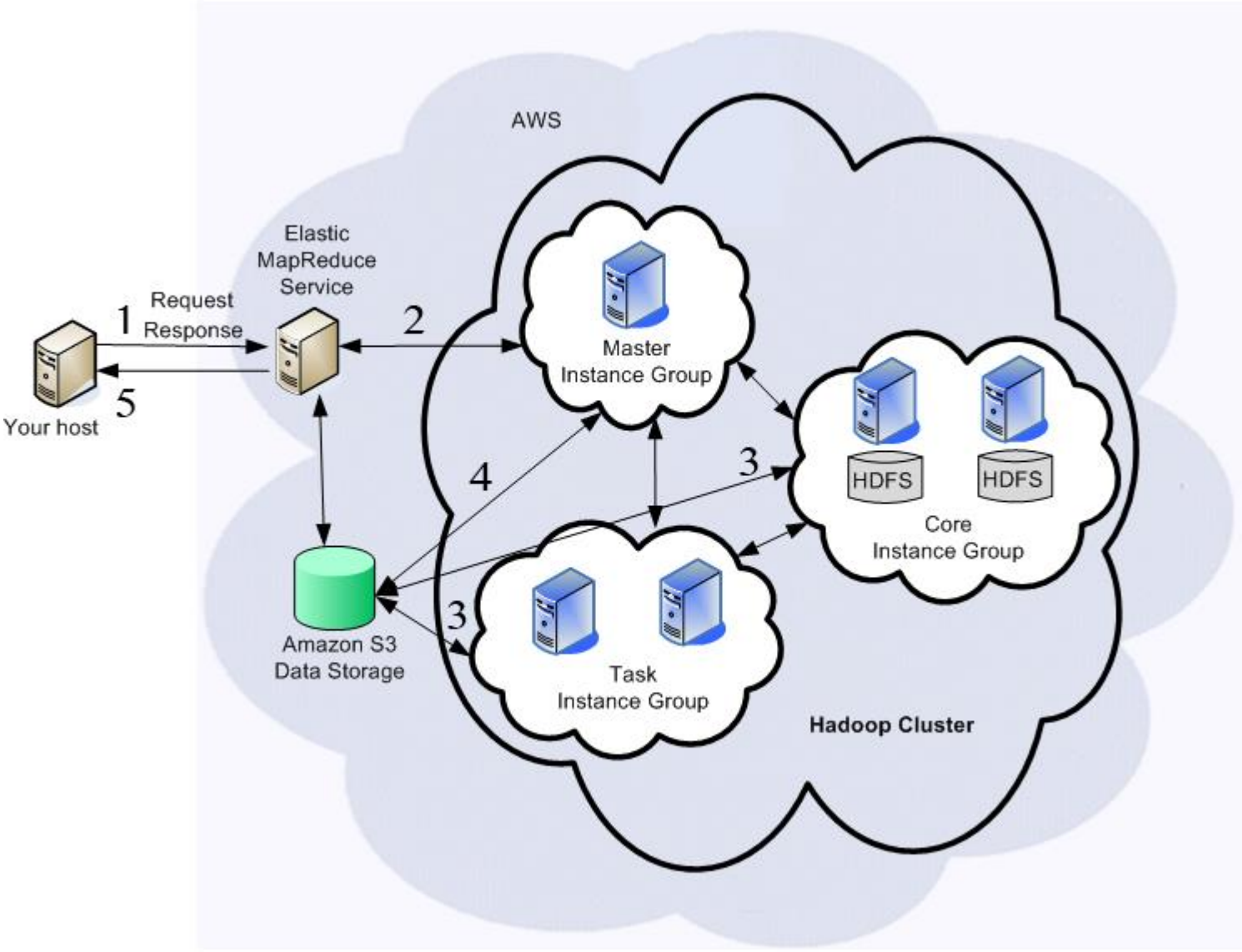
IT drivers


Requirements
IT drivers
Agreger des données du Mo au To en lecture séquentielle
Data aggregation
IO Elasticity
SVR, NLP execution time is ~100ms by task
Large Tasks execution
CPU Elasticity
Traiter plusieurs Go en mémoire
Large RAM execution
RAM Elasticity
Augmenter le ROI des projets d'exploration en diminuant le TCO
Low CAPEX
Commodity HW
OSS SW
Low OPEX
Cost Elasticity

Solutions pour adresser le problème



AWS Elastic MapReduce Architecture



Source: AWS

Custom App
Python, Java, C#, …

Dataming tools
R, SAS

BI tools
Tableau, Pentaho, …

Pig
Flow processing

Streaming
MR scripting

Hive
SQL-like querying

Oozie
MR workflow

Mahout
Machine Learning

Hama
Bulk synchronous processing

MapReduce
Parallel processing framework

HBase
NoSQL on HDFS

HDFS
Distributed file storage

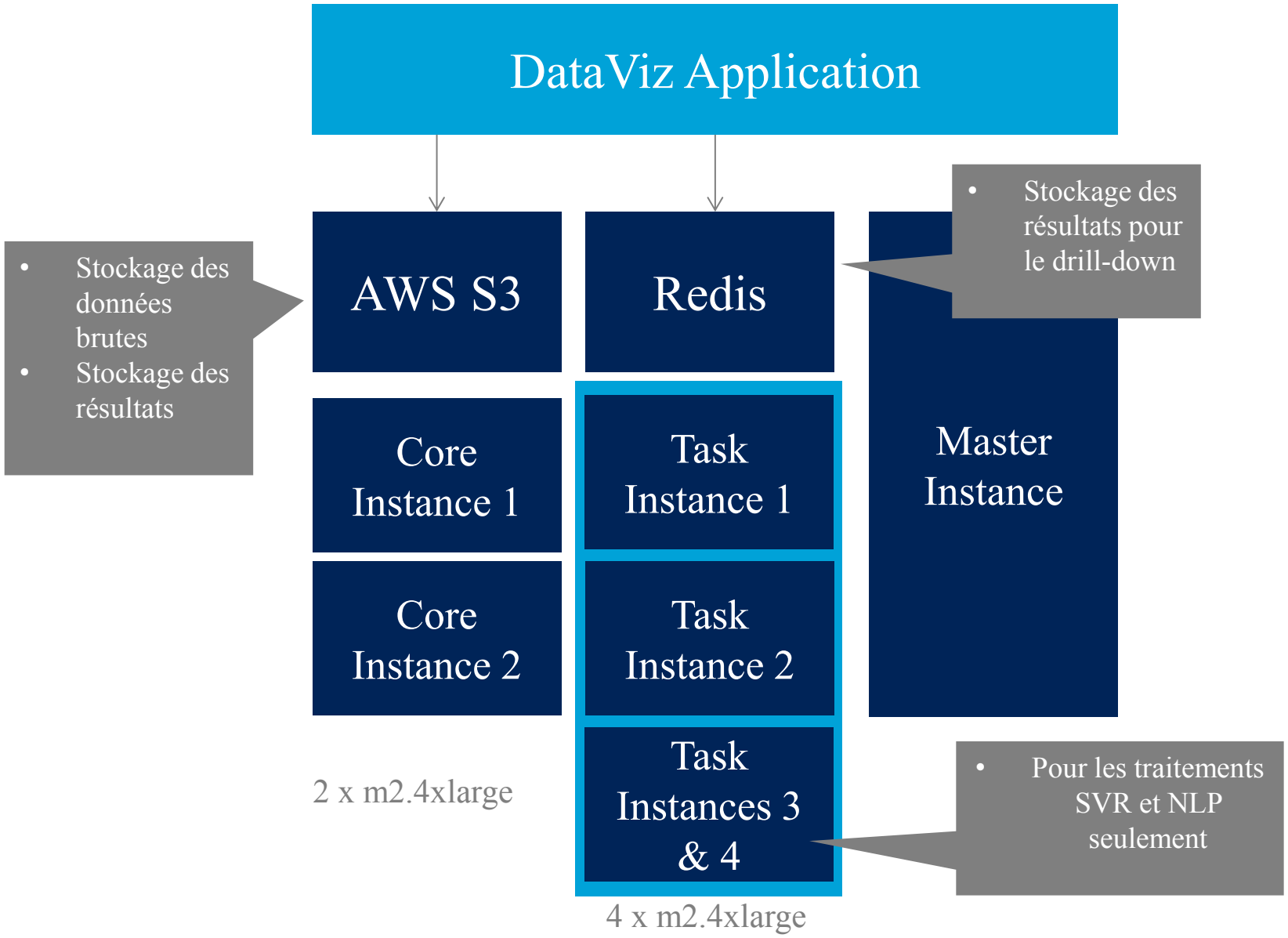
Zookeeper
Coordination service

Sqoop
RDBMS integration

Flume
Data stream integration

Grid of commodity hardware – storage and processing

Architecture générale de la plateforme



Num_of_messages_by_date.pig

```
records = LOAD '/input/forums/messages.txt'
AS (str_date:chararray, message:chararray,
url:chararray);

date_grouped = GROUP records BY str_date

results = FOREACH date_grouped GENERATE
group, COUNT(records);

DUMP results;
```

- Hadoop streaming exécute des jobs map/reduce ecrits en scripts au travers des standard I/O

- Cela resemble à cela (sur un cluster) :
  - cat input.txt | map.py | sort | reduce.py

- Pourquoi Hadoop streaming?
  - Usage intensif de NLTK pour le Natural Language Processing
  - Usage intensif de NumPy et Scikit-learn pour le Machine Learning

Stemmed word distribution with Hadoop streaming, mapper.py

Stem_distribution_by_date/mapper.py

```
import sys
import nltk
from nltk.tokenize import regexp_tokenize
from nltk.stem.snowball import FrenchStemmer

# input comes from STDIN (standard input)
for line in sys.stdin:
    line = line.strip()
    str_date, message, url =  line.split(";")

    stemmer = FrenchStemmer("french")
    tokens = regexp_tokenize(message, pattern='\w+')
    for token in tokens:
        word = stemmer.stem(token)
        if len(word) >= 3:
            print '%s;%s' % (word, str_date)
```

Stem_distribution_by_date/reducer.py

```
import sys
import json
from itertools import groupby
from operator import itemgetter
from nltk.probability import FreqDist

def read(f):
    for line in f:
        line = line.strip()
        yield line.split(';')

data = read(sys.stdin)

for current_stem, group in groupby(data, itemgetter(0)):
    values = [item[1] for item in group]
    freq_dist = FreqDist()

    print "%s;%s" % (current_stem, json.dumps(freq_dist))
```

Conclusions

Conclusions

- La recherche de corrélations identifie
  - **42 variables externes corrélées** à plus de 70% sur une période supérieure à 24 mois sur 2,7 millions de variables
  - un facteur d'**anticipation supérieur à 4 mois**

- L'analyse sémantique permet de répondre à des questions :
  - Quels sont les centres d'intérêt Santé à la date T?
  - Quelle est l'évolution d'un centre d'intérêt dans le temps?
  - Où les internautes se retrouvent-ils pour échanger sur un sujet donné?

Merci de votre attention!

Julien Cabot
jcabot@octo.com
@julien_cabot