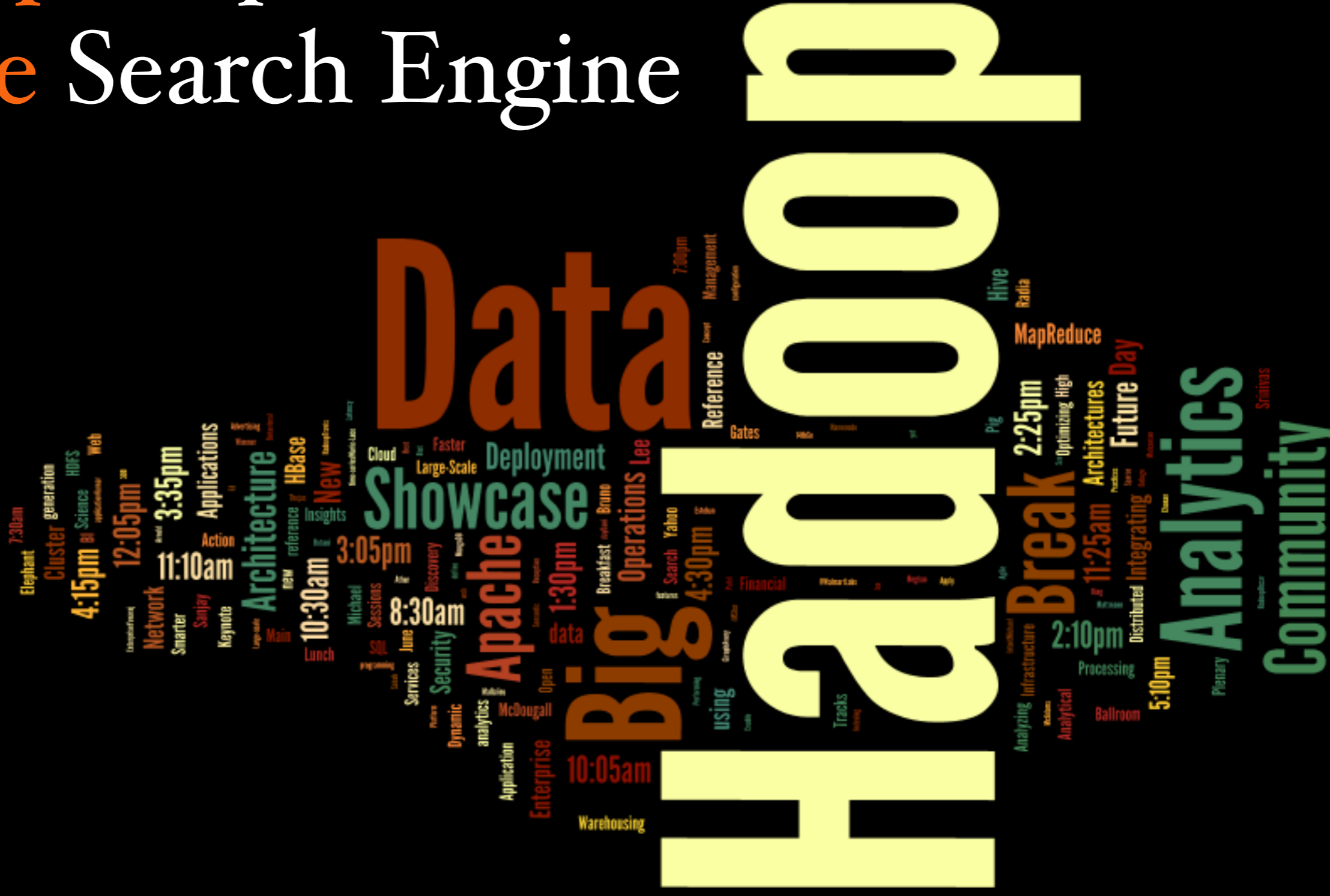


# Distributed Calculus with Hadoop MapReduce inside Orange Search Engine



# What is Big Data ?

\$ 5 billions (2012)

to

\$ 50 billions

(by 2017)

Forbes

*«Big Data is the new definitive  
source of competitive  
advantage across all  
industries»*

Jeff Kelly

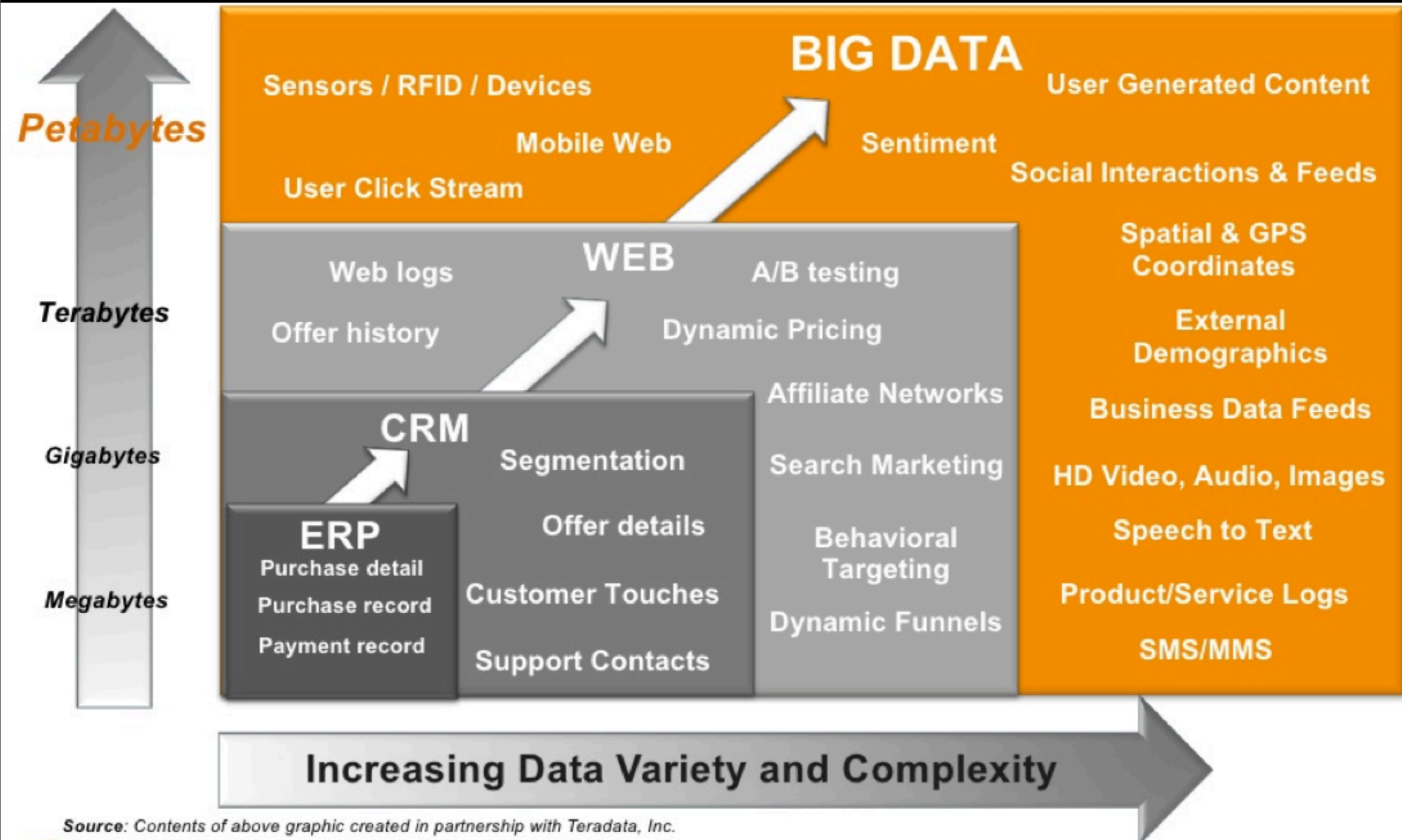
# Product success

«The days are over when you build a product once and it just works.

You have to take ideas, test them, iterate them, use data and analytics to understand what works and what doesn't in order to be successful»

LivingSocial @PCWorld 2012

# Big Data



Source: Contents of above graphic created in partnership with Teradata, Inc.

# What about Hadoop ?

# Beliefs

« We believe that by 2015, more than half the world's data will be processed by Apache Hadoop »



HortonWorks @HadoopSummit 2012





# Actors



# Hadoop eco-system



# Context at Orange

(more than 2 years ago)

# Orange Search Engine

<http://www.orange.fr>

orange un check-up pour votre forfait? toutes vos recherches sur le web et orange.fr autonomie | page d'accueil pro | page d'accueil personnalisée

rechercher

actualités | sports | finance | femmes | tv/vidéo | musique | jeux | photo | annuaire | auto | météo | cinéma | voyages | t'Chat | horoscope | tous les services

mercredi 20 juin - St Silvere

Prise d'otages à Toulouse actualité

En bordure de l'océan Indien Entourée par le Kenya et l'Ouganda, laissez-vous éblouir par la Tanzanie | voyages

Adieu Berthe cinéma

la bourse en direct CAC40 : 3126,40 +0,27%

Appareil photo shopping Découvrez notre sélection pour partir en vacances.

Du 1er coup code de la route Mettez toutes les chances de votre côté pour l'obtenir.

<http://www.lemoteur.fr>

voila Mercredi 20 Juin 2012

johnny depp Sur le Web Rechercher

M'identifier | Me créer un compte | Personnaliser ma page | Voila en page d'accueil Tous les services de Voila

Recherche

639 693 réponses Pages Web : johnny depp Web francophone | Web mondial

107 000 000 réponses Images : johnny depp

22 réponses Vidéos : johnny depp

Paradis en enfer, Cinéma - Dark Shadow..., Sweeney Todd - Le Di...

<http://www.voila.fr>

orange le moteur

web vidéos thématiques plus de rubriques

clara morgane rechercher

139 688 réponses

sur le web francophone mondial favoris

vidéos images

Recherches associées clara morgane films clara morgane photos clara morgane biographie clara morgane age

Clara Morgane Nom de naissance : Emmanuelle Munos Age : 31 ans Profession : Chanteuse, Animatrice à la télévision, Entrepreneur en lingerie Trouver : photos, vidéos, actualités

wikipedia.orange.fr/wiki/Clara\_Morgane

Clara Morgane . Le site officiel . Accueil Bienvenue sur le site officiel de Clara Morgane, son actu, ses albums, ses clips, et des centaines de photos et vidéos exclusives... www.claramorgane.com/ - ajouter à mes favoris

Blog de clara-diary . Clara morgane ... . Skyrock.com Clara morgane .... Clara Morgane la star de se sky allé vos commis ^^ ils sont rendu alors j'attend .... Infos. Liens Skyrock lorsque je voi flou ... a mes seins je me voue. Commenter. N... clara-diary.skyrock.com/ - ajouter à mes favoris

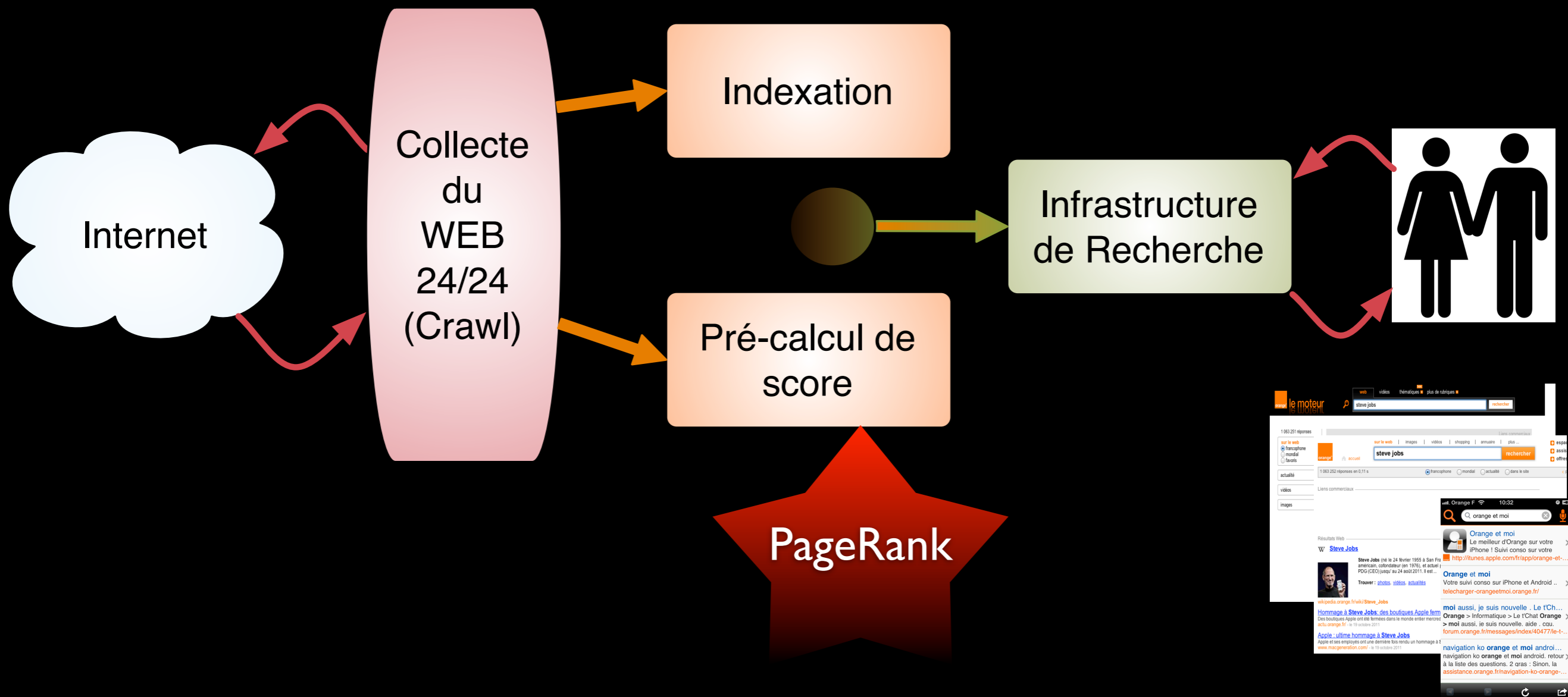
Clara Morgane à Narbonne . Fil info . Pro D2 . Rugby . Sport 24 Le coup d'envoi de la rencontre de Pro D2 entre Narbonne et Aix-en-Provence sera donné par Clara Morgane samedi. www.sport24.com/rugby/pro-d2/fil-info/clara-morgan...narbonne-443536/(language)/fre-FR - ajouter à mes favoris

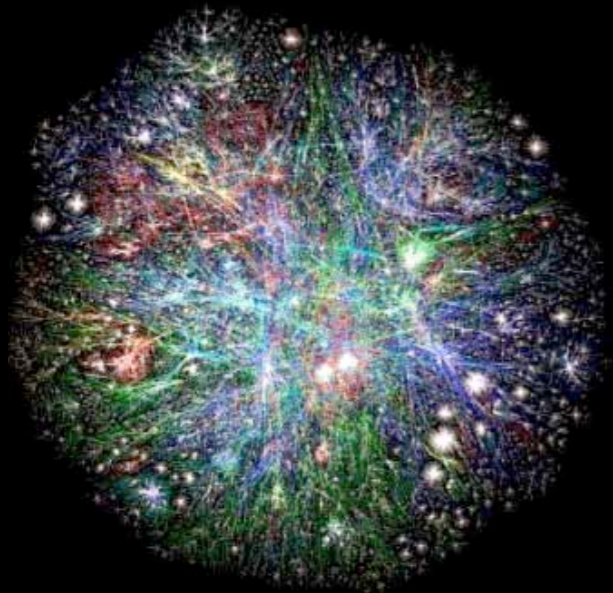
[olivier.varene@orange.com](mailto:olivier.varene@orange.com)

SophiaConf2012

mardi 3 juillet 12

# Search Engine Architecture





# Main issue

- PageRank calculus on billions nodes and 10s billions edges
- regularly failed ! (hardware ...)
- 4 to 8 weeks calculus
- unscalable
- failure rate around 80%
- One person full time to supervise



# Answer



# PageRank portable to Hadoop / MapReduce ?

- Simple programming model:

Map(in\_k,in\_v) => list(out\_k,intermed\_v)

Reduce(out\_k,intermed\_v) => list(out\_v)

- Scalable
- Batch Processing
- **YES !**

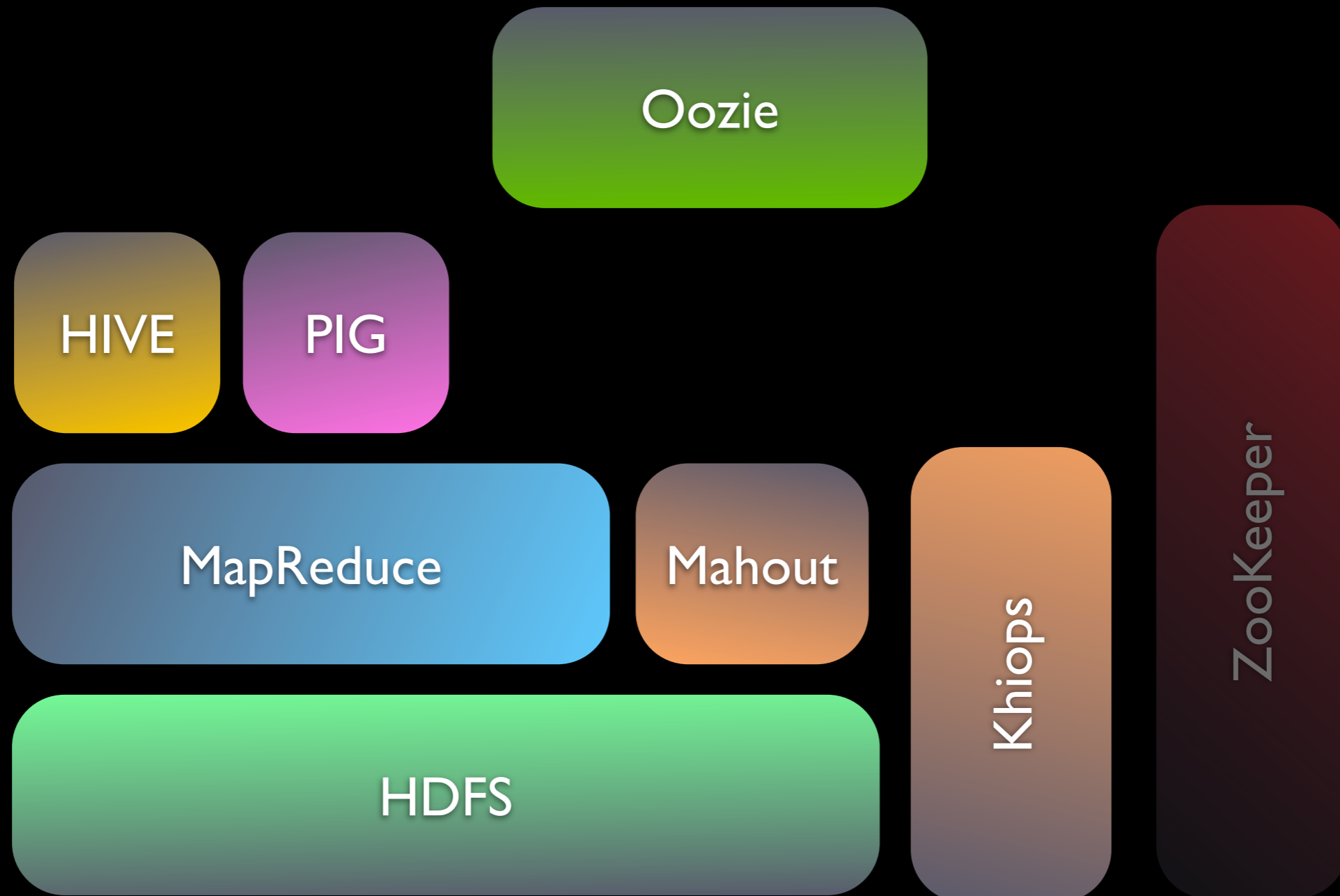


# Hadoop Axioms

- System shall manage and heal himself
- Performance shall scale linearly
- Compute shall move to data
- Modular and extensible

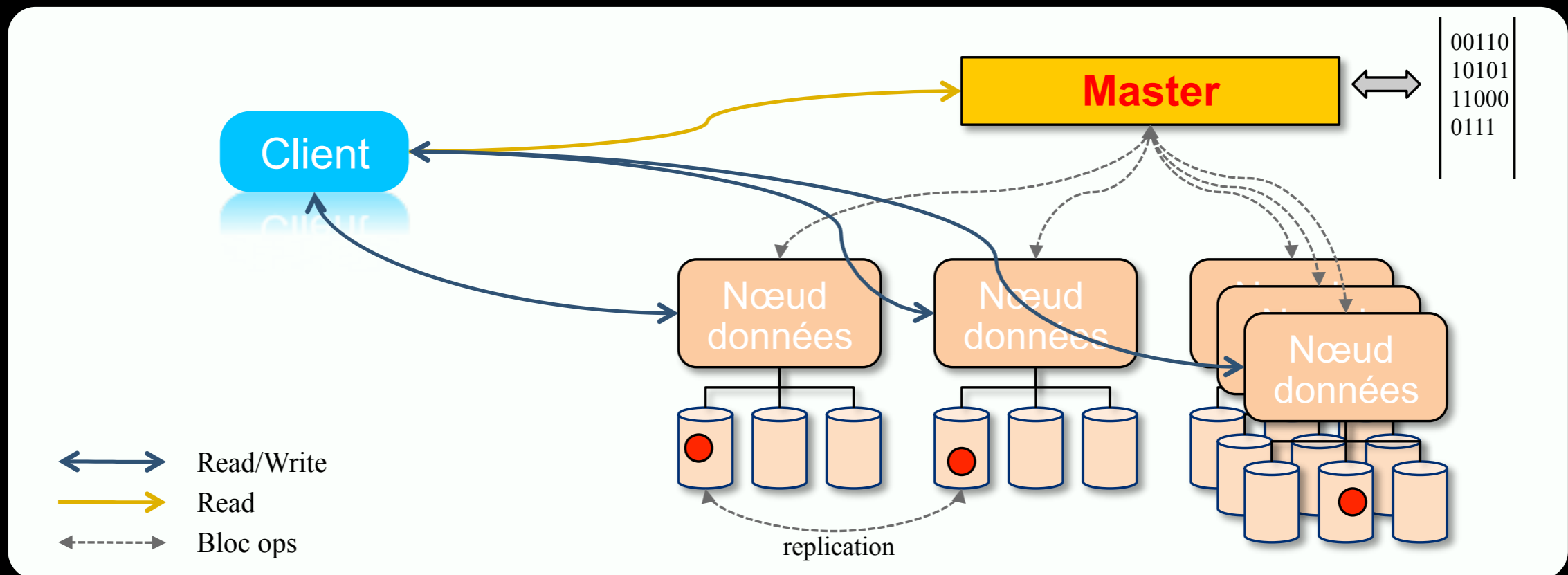
# Our install ?

# Our install



# HDFS ?

# Hadoop - HDFS



COTS - replication - big blocks  
maximize throughput - Metadata in RAM

# Map Reduce ?

# MapReduce

cat | «your map» | sort -u | «your reduce»



Programming paradigm

# MapReduce

cat | «your map» | sort -u | «your reduce»



**FrameWork**



# MapReduce

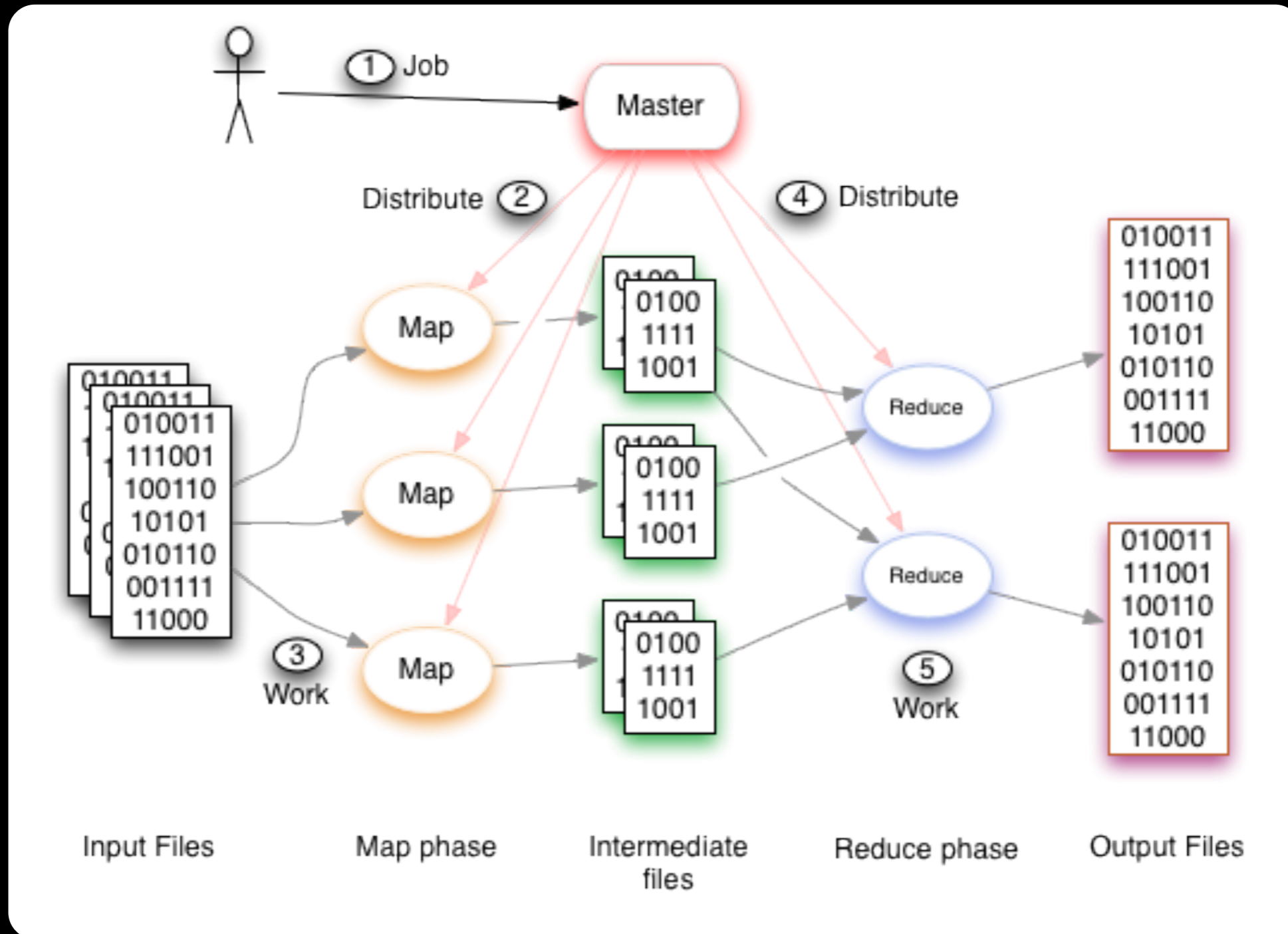
cat | «your map» | sort -u | «your reduce»



your Job

The diagram illustrates a MapReduce pipeline. It consists of three main stages: 'cat', '«your map»', and '«your reduce»'. The 'cat' command is followed by a pipe to '«your map»', which is then piped to 'sort -u', which is finally piped to '«your reduce»'. Two orange double-headed arrows are positioned below the pipes, one between 'cat' and '«your map»', and another between '«your map»' and '«your reduce»'. Below these arrows, the text 'your Job' is written in orange, indicating that the entire pipeline represents a single MapReduce job.

# MapReduce Insides



# Interfaces

- Java API
- Pipes
- Streaming (python, perl, C/C++, ...)

# PIG

- High level data analysis script language
- extensible via UDF
- Structure of a Pig script
  - load
  - filter
  - foreach | group by | join | your functions
  - order
  - store

# HIVE

- High level SQL-like query and analysis language
- extensible via UDF
- Structure of a Hive script
  - create table
  - load data
  - select ... from ...
  - insert | group by | join

# Application domains ?

# Projects

- Scoring
- User profiling
- Log analysis and statistics
- ... and many others to come



ROI ?



# ROI

- Lines Of Code

**10X** gain

- Development Time

**2X** gain

- IT cost

**4X** gain

less bug, automatic, scalable ...

# Perfect World ?

## ★ YES

- Run cost
- Development cost
- Scalable
- Stable
- Heterogeneity

## ★ NO

- SPOF (almost solved)
- Fastidious debugging
- Locally non optimum
- mono-site

# Thank you

Olivier Varene - Orange

[olivier.varene@orange.com](mailto:olivier.varene@orange.com)

@VareneO



# thanks to

- Hadoop - Apache (<http://hadoop.apache.org/>)
- Khiops - Orange (<http://www.khiops.com>)
- Shawn Connolly - HortonWorks  
(<http://youtu.be/yPfysFAGv8s>)
- Forbes - article  
(<http://www.forbes.com/sites/siliconangle/2012/02/17/big-data-is-big-market-big-business/>)
- Living Social (sentence)
- Terradata (Volumetry Graph)
- <http://www.wordle.net/> (Words Cloud)