



**BIG DATA PARIS • 4<sup>e</sup> Edition**  
Congrès & Exposition  
10 et 11 mars 2015 au CNIT • Paris La Défense

**ATELIERS PRODUITS**

**SALLE B**

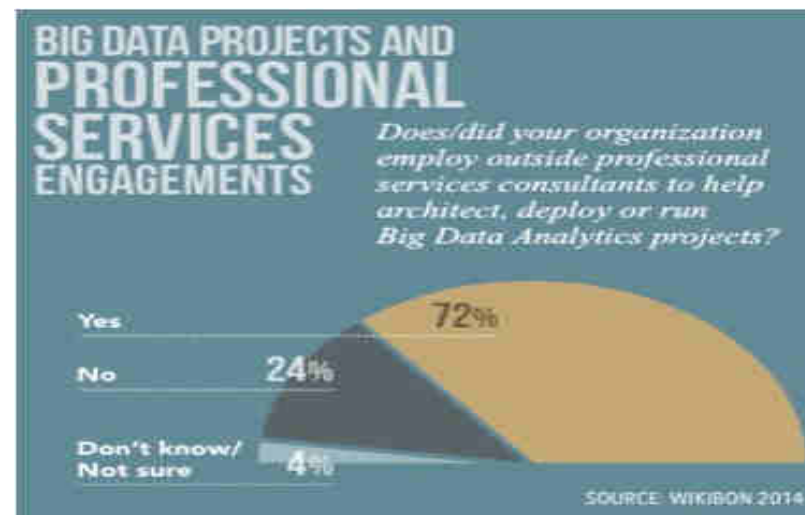
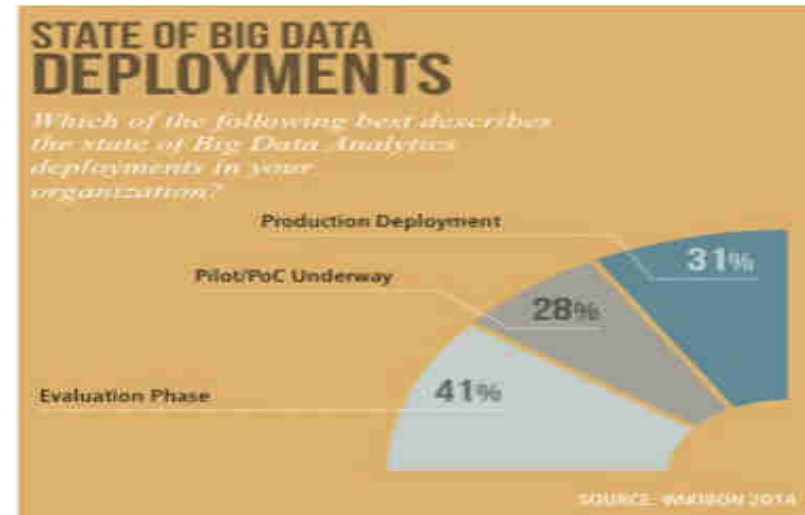
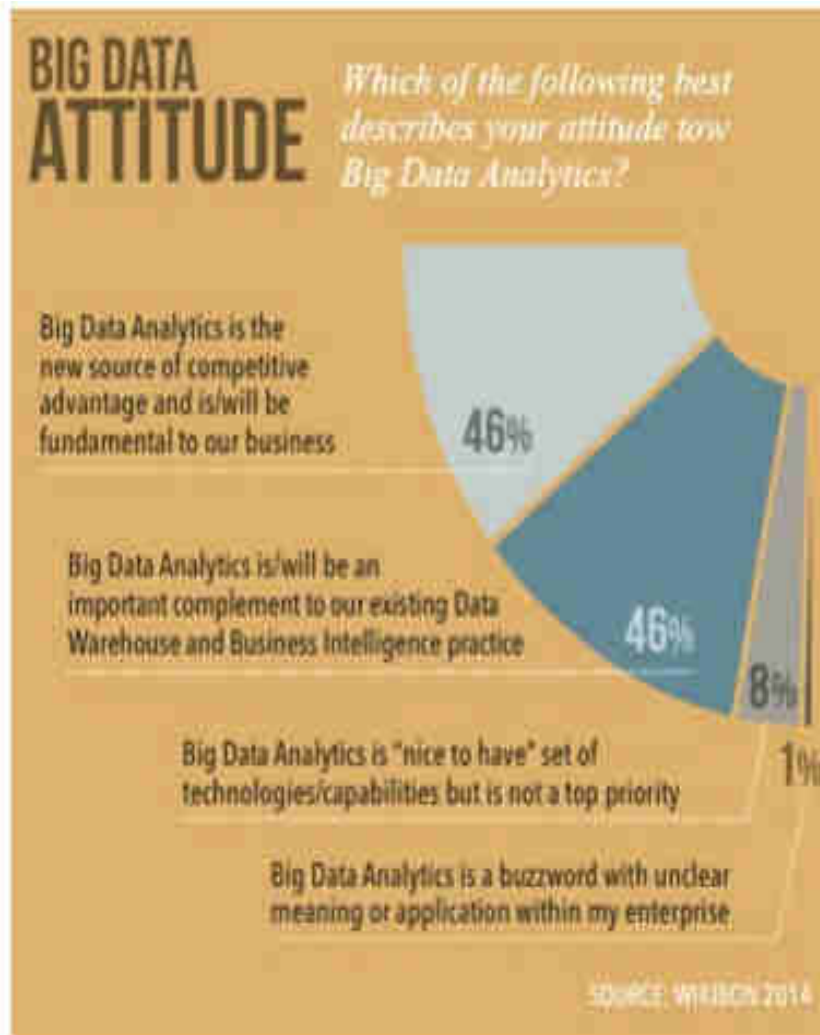
**10 Mars 17h30**

● ● ● ● ● ● ● ● ●  
**informatica**

**Facilitez l'intégration du Big Data  
à votre architecture**

Mathieu Lagrange,  
Product Specialist EMEA  
[mlagrange@informatica.com](mailto:mlagrange@informatica.com)

# Les tendances du Big Data Aujourd'hui



**HDFS**

**Chukwa**

**Mahout**

**Pig**

**Impala**

**Yarn**

**Jaql**

**Oozie**

**Scribe**

**Hive**

**MapReduce**

**Spark**

**Shark**

**Flink**

**Zookeeper**

**Kafka**

**Kylin**

**Flume**

**Myriad**

**Tachyon**

**Presto**

# 80% d'un projet big data concerne l'intégration et la qualité des données

“80% du travail d'un projet  
Big Data est le nettoyage  
des données”

“70% de ma valeur est  
d'ingérer des données, 20%  
d'utiliser mes compétences  
'data-scientiste'...”

“Je passe plus de la moitié  
de mon temps à intégrer,  
nettoyer et transformer les  
données sans faire la  
moindre analyse”

**RethinkDB**

**HBase**

**Weka**

**MongoDB**

**Cassandra**

**BayesDB**

**Neo4j**

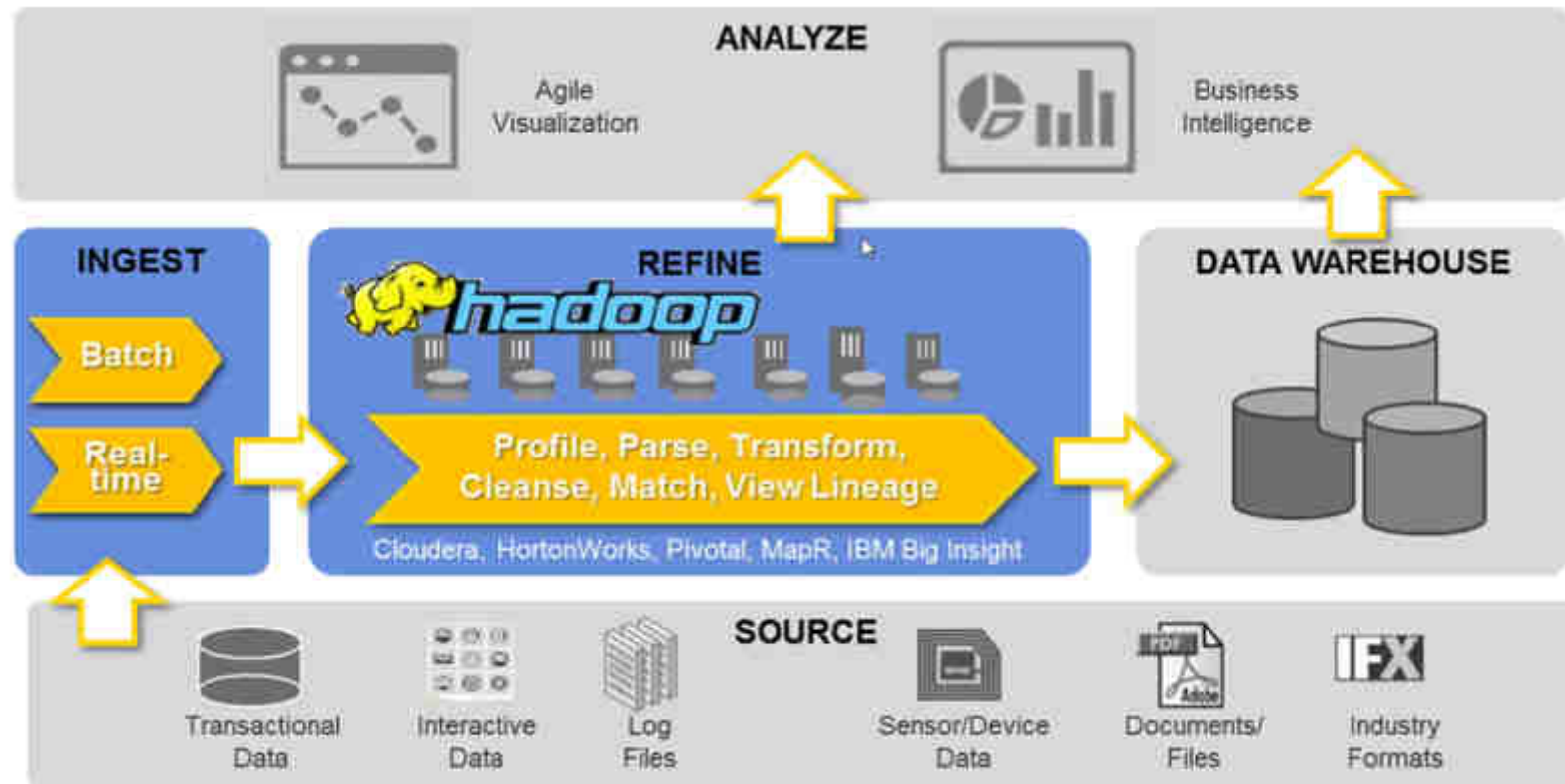
**Twitter Elephant Bird**

**HIHO**



# Les Solutions Informatica pour le Big Data:

Une architecture simple, vos projets accélérés



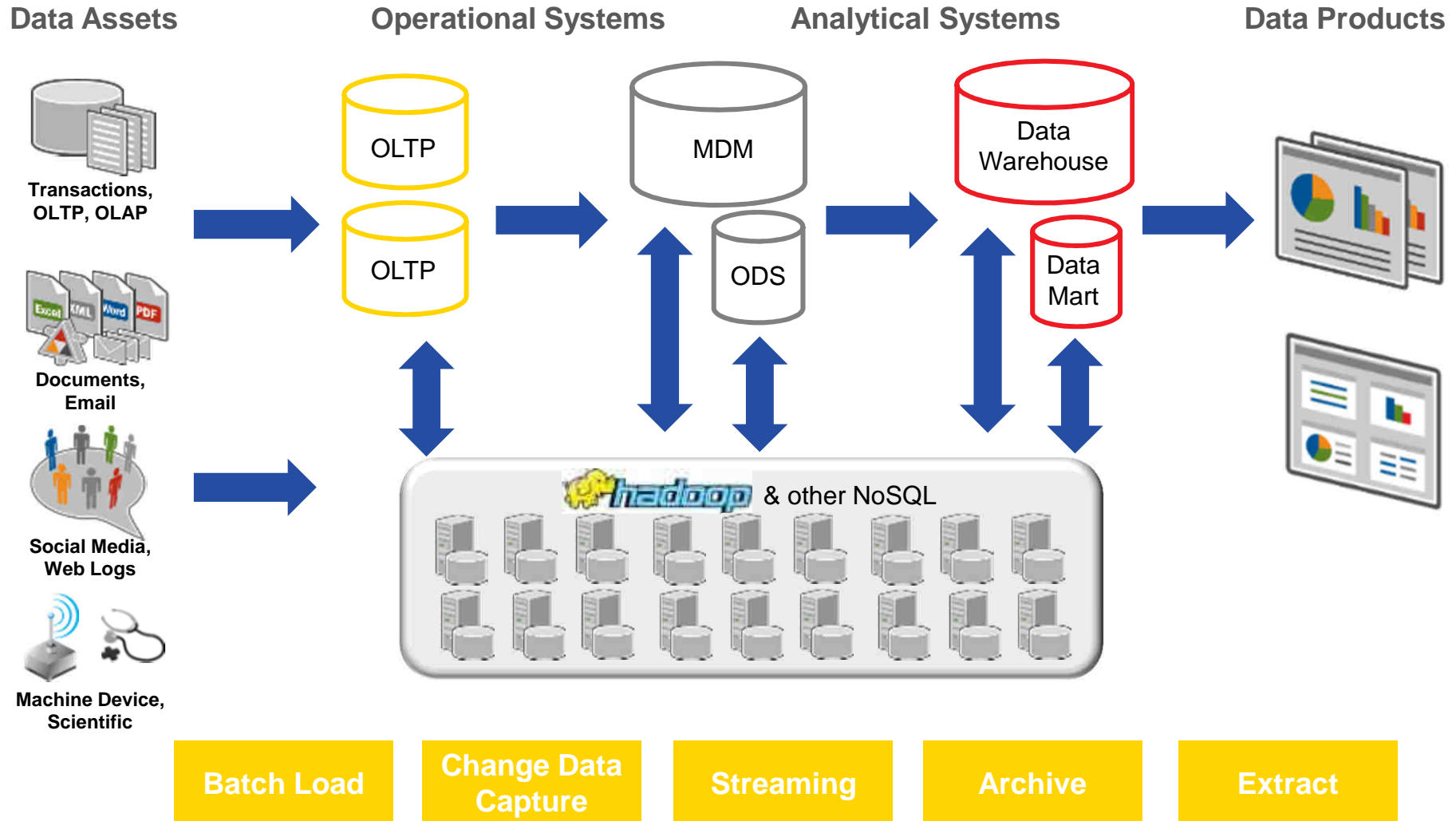
Collecte  
+Rapide

Intégration  
+Simple

Gouvernance  
des données

# Simplifiez votre architecture Big Data

Collecte  
+Rapide



# Connectivité Universelle

Collecte  
+Rapide

Messaging,  
and Web Services



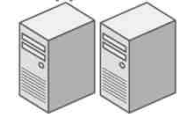
WebSphere MQ  
JMS  
MSMQ  
SAP NetWeaver XI

Web Services  
TIBCO  
webMethods

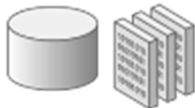
JD Edwards  
Lotus Notes  
Oracle E-Business  
PeopleSoft

SAP NetWeaver  
SAP NetWeaver BI  
SAS  
Siebel

Packaged  
Applications



Relational and  
Flat Files



Oracle  
DB2 UDB  
DB2/400  
SQL Server  
Sybase

Informix  
Teradata  
Nettezza  
ODBC  
JDBC

Salesforce CRM  
Force.com  
RightNow  
NetSuite

ADP  
Hewitt  
SAP By Design  
Oracle OnDemand

SaaS/BPO

salesforce.com  
experience success™



Industry  
Standards

Mainframe  
and Midrange



ADABAS  
Datacom  
DB2  
IDMS  
IMS

VSAM  
C-ISAM  
Binary Flat Files  
Tape Formats...

EDI-X12  
EDI-Fact  
RosettaNet  
HL7  
HIPAA

AST  
FIX  
SWIFT  
Cargo IMP  
MVR

Unstructured  
Data and Files



Word, Excel  
PDF  
StarOffice  
WordPerfect  
Email (POP, IMPA)  
HTTP

Flat files  
ASCII reports  
HTML  
RPG  
ANSI  
LDAP

XML  
LegalXML  
IFX  
cXML

ebXML  
HL7 v3.0  
ACORD (AL3, XML)

XML Standards



MPP Appliances



Pivotal  
Vertica  
Nettezza

Teradata  
Aster

Facebook  
Twitter  
LinkedIn

Kapow  
Datasift



Social Media

informatica



HBASE



# Connectivité Réseaux Sociaux

Collecte  
+Rapide



**Search-Post**

| Name             | Type   | Precision | Scale | Description |
|------------------|--------|-----------|-------|-------------|
| application      | STRING | 512       | 0     | applicati   |
| application_id   | STRING | 300       | 0     | applicati   |
| application_name | STRING | 520       | 0     | applicati   |
| caption          | STRING | 15680     | 0     | caption     |
| created_time     | STRING | 480       | 0     | created_t   |
| description      | STRING | 5940      | 0     | descripti   |
| from             | STRING | 512       | 0     | from        |
| from_category    | STRING | 420       | 0     | from_cat    |
| from_id          | STRING | 300       | 0     | from_id     |
| from_name        | STRING | 600       | 0     | from_nar    |
| icon             | STRING | 1180      | 0     | icon        |
| id               | STRING | 620       | 0     | id          |
| likes            | STRING | 512       | 0     | likes       |
| likes_count      | STRING | 20        | 0     | likes_cou   |

**Output**

| Name             | Type   | Precision | Scale |
|------------------|--------|-----------|-------|
| application      | string | 512       | 0     |
| application_id   | string | 300       | 0     |
| application_name | string | 520       | 0     |
| caption          | string | 15680     | 0     |
| created_time     | string | 480       | 0     |
| description      | string | 5940      | 0     |
| from             | string | 512       | 0     |
| from_category    | string | 420       | 0     |
| from_id          | string | 300       | 0     |
| from_name        | string | 600       | 0     |
| icon             | string | 1180      | 0     |
| id               | string | 620       | 0     |
| likes            | string | 512       | 0     |
| likes_count      | string | 20        | 0     |
| likes_data       | string | 1040      | 0     |

Overview Application\_Data\_Object\_Operation x

Properties Data Viewer x Tags

Configuration: (Default Settings) Run

**Output**

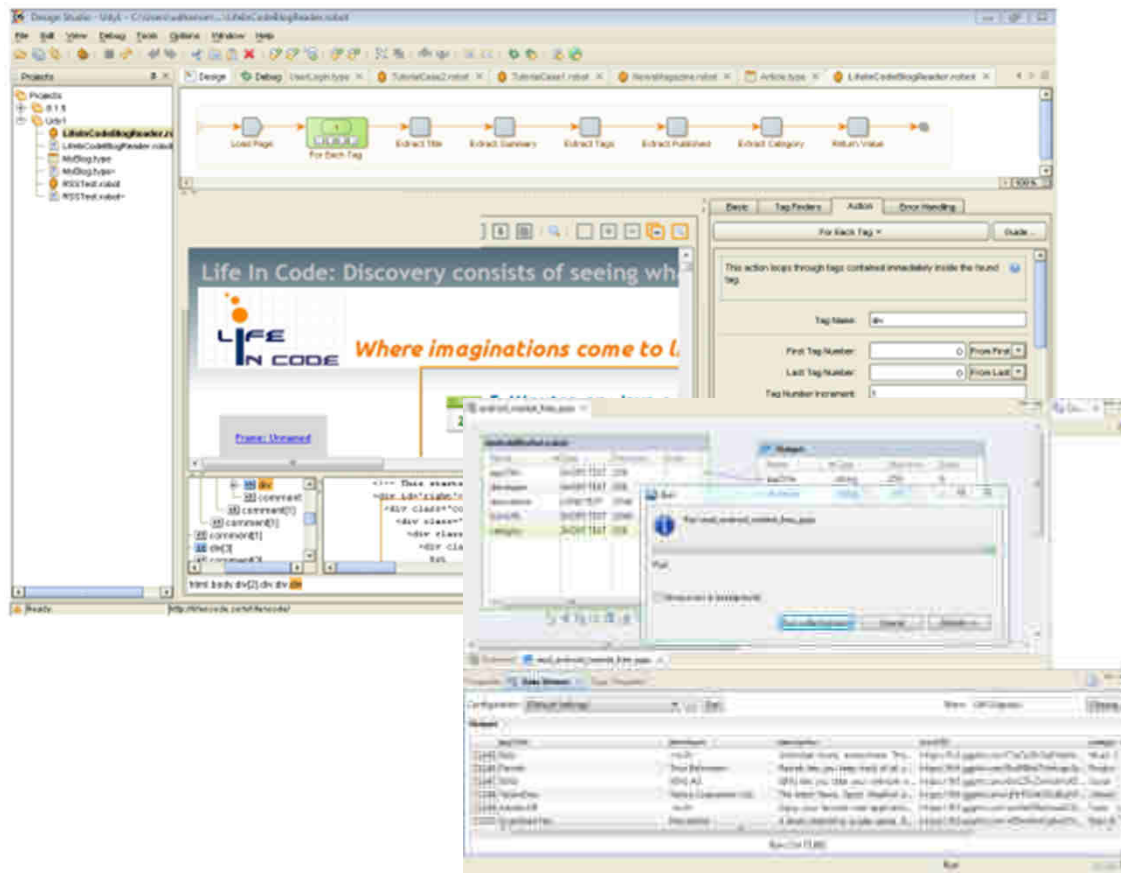
| id               | likes | likes_count | likes_data | link             | message   |
|------------------|-------|-------------|------------|------------------|---|
| 1 10000219970... |       |             |            | http://www.fa... | Proud to be a Kentucky Fan...CAL is a great coach with a strong team...was great listenin                 |
| 2 845634239_1... |       |             |            |                  | "The start is what stops most people." Don Shula Former NFL Coach   |
| 3 517350798_1... |       |             |            |                  | SHUT UP COACH MACGREGOR. IMMA PUNCH YOU. SMASH IS A PATHETIC SHOWBOATING ASS. IM...                       |
| 4 798644247_1... |       |             |            |                  | "The start is what stops most people." Don Shula Former NFL Coach   |
| 5 10000202373... |       |             |            |                  | Bigger-Better-Bolder GOING GOING GONE rules www.fashionandyou.com at 10am. Upto 90% OFF on...             |
| 6 1618964438...  |       |             |            |                  | WISDOM REFLECTIONS... A Devotional For Deep Minds - Monday 26th March '12 - Extending The Fro...          |
| 7 1570361212...  |       |             |            | http://www.fa... | AMEN Tha's my motto...JUST DO IT and make sure your heart is in it...The path less traveled that is wh... |

Row 1 to 100

| Connexion | Fonctionnalités   |
|-----------|---|
| Twitter   | <ul style="list-style-type: none"> <li>Recherche de tweets</li> <li>Extraction des données de profil utilisateur</li> </ul>                 |
| LinkedIn  | <ul style="list-style-type: none"> <li>Extraction des données de profil utilisateur</li> <li>Extractions des dates de connexions</li> </ul> |
| Facebook  | <ul style="list-style-type: none"> <li>Extraction des statuts Facebook + critères de recherche</li> </ul>                                   |

# Accéder aux données du Web

Collecte  
+Rapide



kapow  
SOFTWARE

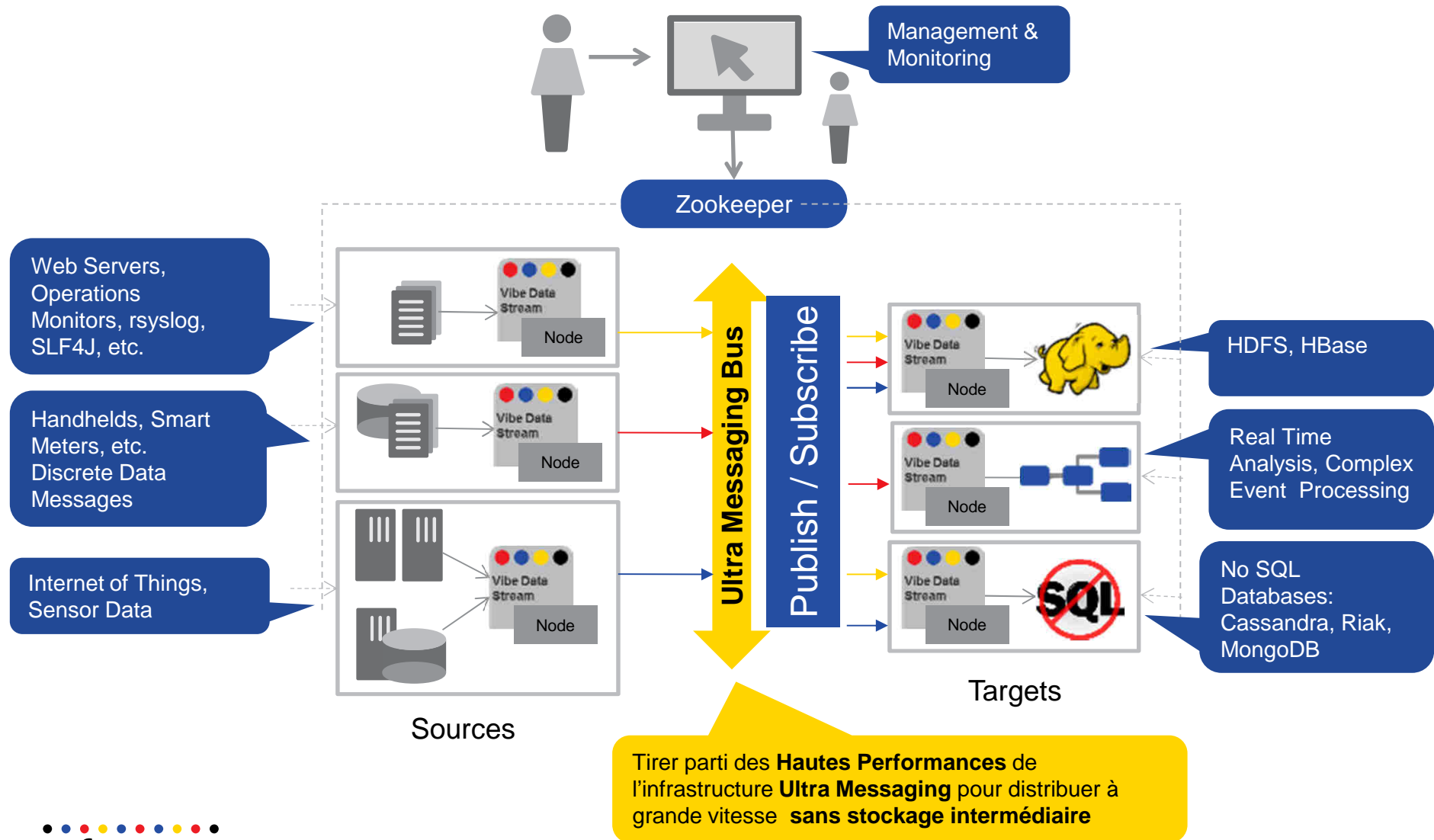
- **Simuler une navigation Web** pour extraire les bonnes informations de manière graphique et rapide
- Extractions **de tout type d'application Web**
- Outiller sa cellule de **veille technologique**
- **Connectivité native Informatica**



# Big Data Streaming

Vibe Data Stream- performance

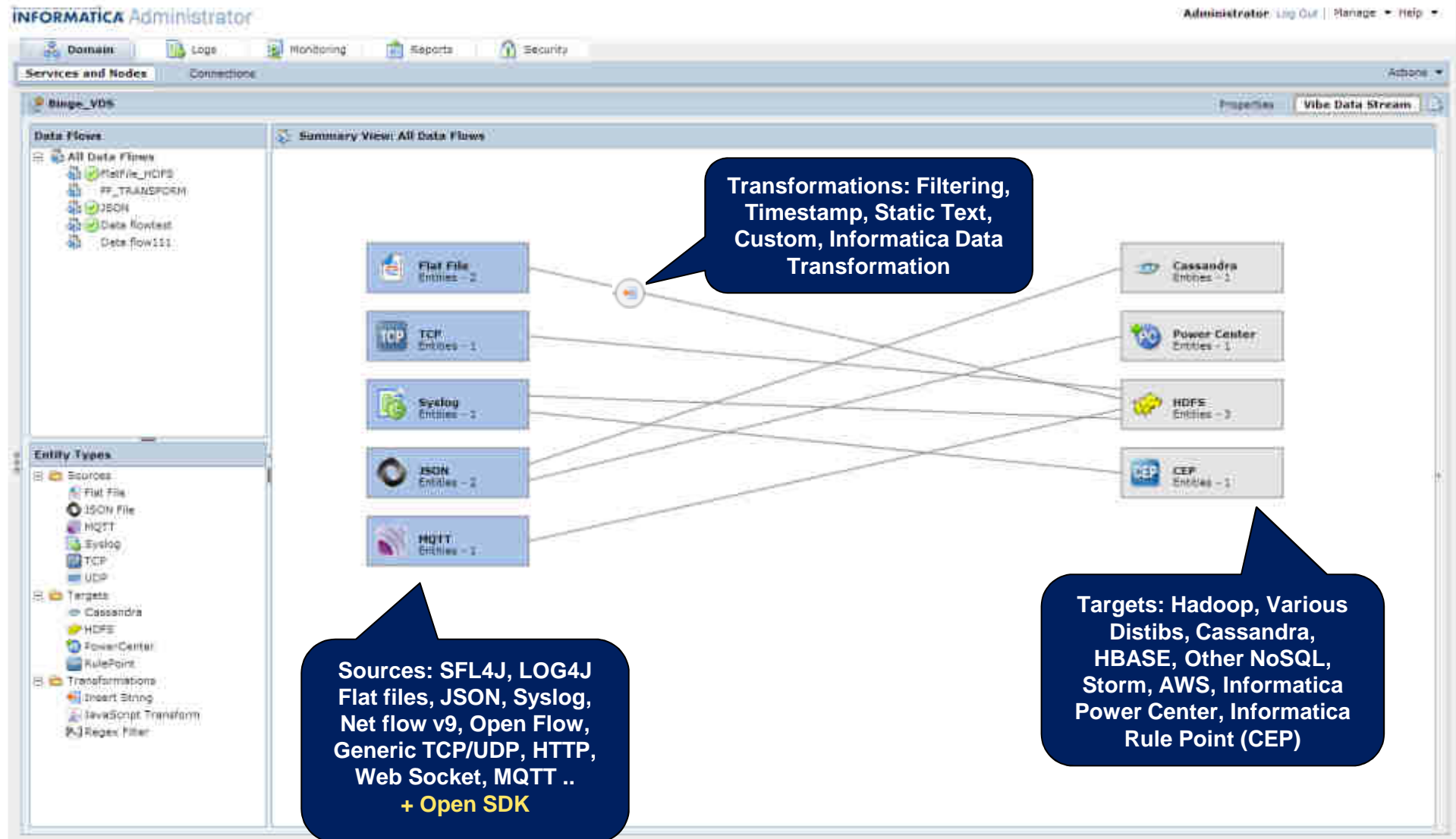
Collecte  
+Rapide



# Big Data Streaming

## Vibe Data Stream - Interface Graphique

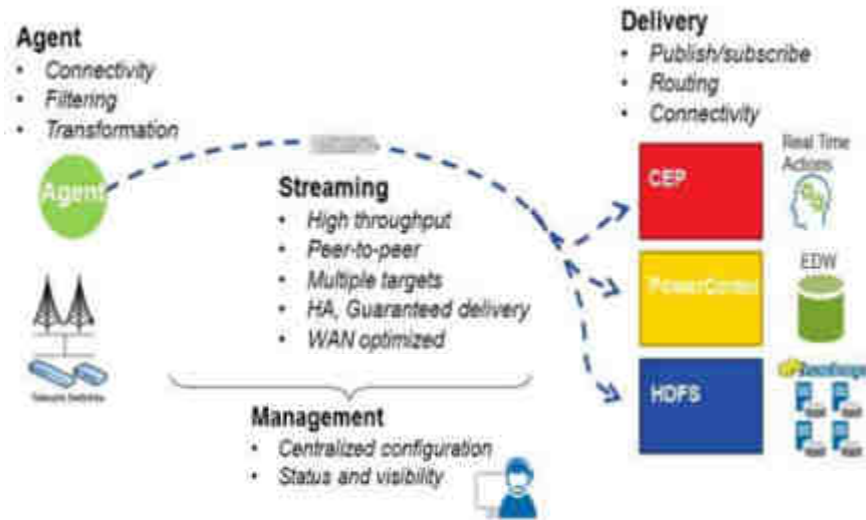
Collecte  
+Rapide



# Big Data Streaming

## Vibe Data Stream – Complex Event Processing

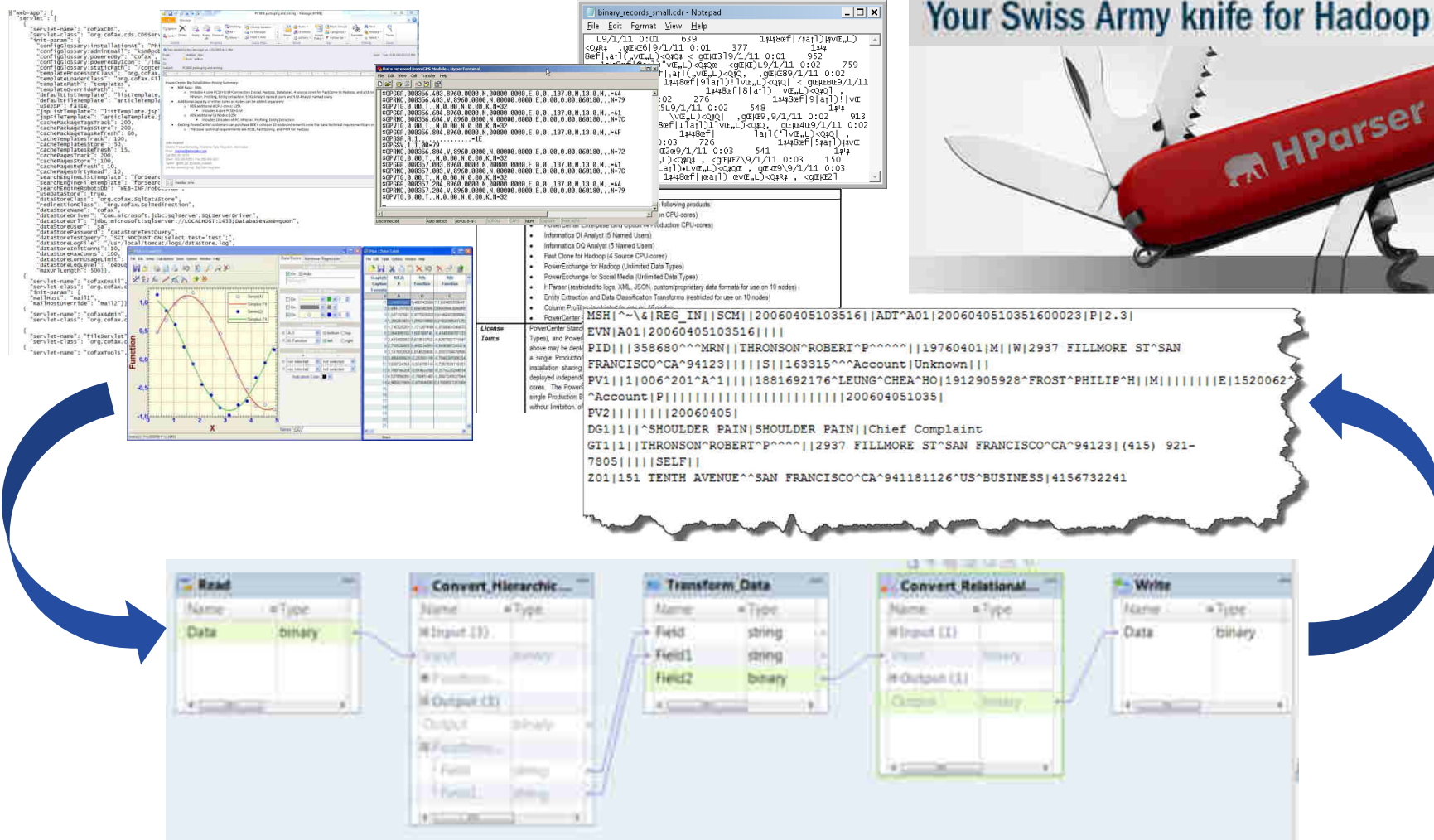
Collecte  
+Rapide



- **Collecte Haute Performance** des données streaming LAN/WAN **sans 'Router'** intermédiaire
- **Interface graphique & simple** pour configurer, (re)déployer, monitorer
- **Ingestion** continue en **temps-réel** des données logs, machines et autres (sdk)
- **Interactions temps-réel** (Complex Event Processing)
- **Déploiement immédiat** directement aux cibles multiple (batch/stream)
- **Haute disponibilité; efficacité; scalabilité**
- 'Light weights' agents source et cibles sur **écosystème étendu**

# Traiter les données Complexes

Intégration  
+Simple





# Un Exemple : librairie ASN.1

Intégration  
+Simple

1. Vue des données sources

2. Editeur interactif de la logique de parsing

3. Suivi d'exécution interactif

4. Résultats du parsing

```
<CALL-RECORD-COUNT>  
<CALL-RECORDS-NUM>00014</CALL-RECORDS-NUM>  
</CALL-RECORD-COUNT>  
<CALL-RECORD>  
<CALL-TIMESTAMP>9/1/11 0:01</CALL-TIMESTAMP>  
<CALL-DURATION>639</CALL-DURATION>  
<CALL-IMSI>  
<IMSI-ONE>310</IMSI-ONE>  
<IMSI-TWO>150</IMSI-TWO>  
<IMSI-THREE>389</IMSI-THREE>  
<IMSI-FOUR>667</IMSI-FOUR>  
<IMSI-FIVE>370</IMSI-FIVE>  
</CALL-IMSI>  
<CALL-ISDN>  
<ISDN-ONE>61186<  
<ISDN-TWO>29190<  
</CALL-ISDN>  
<CALL-IMEI>  
<IMEI-ONE>768</IMEI-ONE>  
<IMEI-TWO>844</IMEI-TWO>  
<IMEI-THREE>293</IMEI-THREE>  
<IMEI-FOUR>510</IMEI-FOUR>
```

# Traiter les données complexes Hadoop

Intégration  
+Simple

## The broadest coverage for Big Data

### Fichiers Plats & Documents

Name = Value  
^/>>Delimited<¥^



### XML

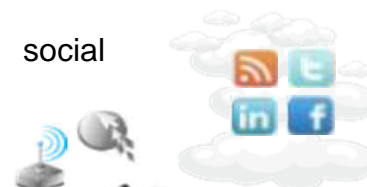


### Standards



### Interactions

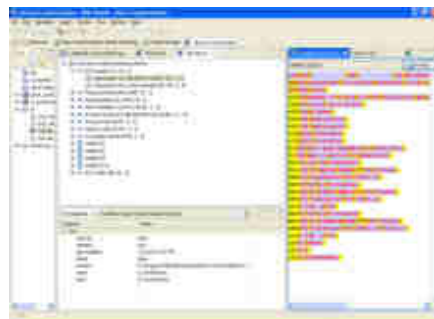
social



Device/sensor  
scientific

## Productivity

- Environnement Graphique
- Bibliothèques de transformation pré-packagées



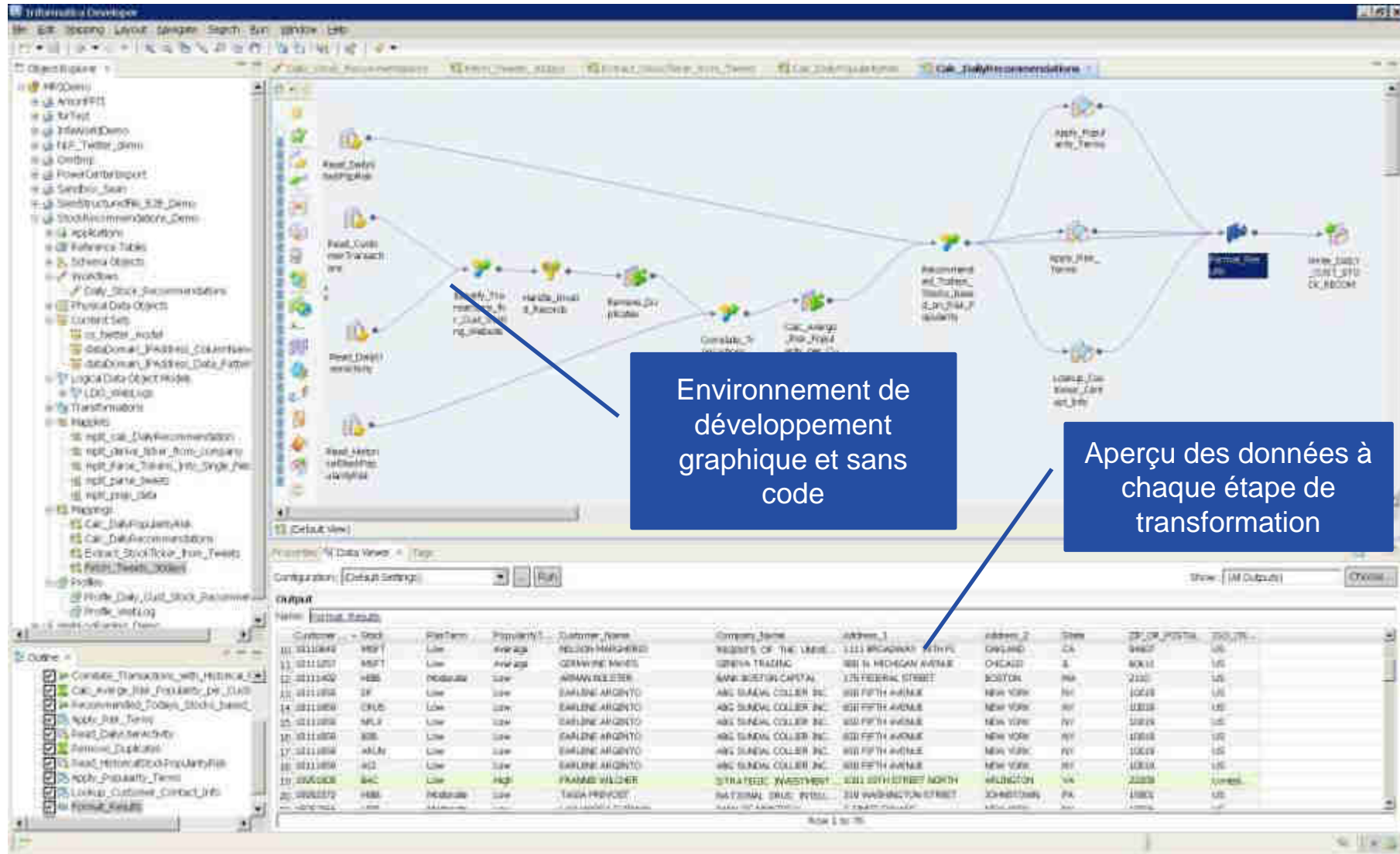
## Any DI/BI architecture



EDW  
MDM

# Intégrer les données sans codage

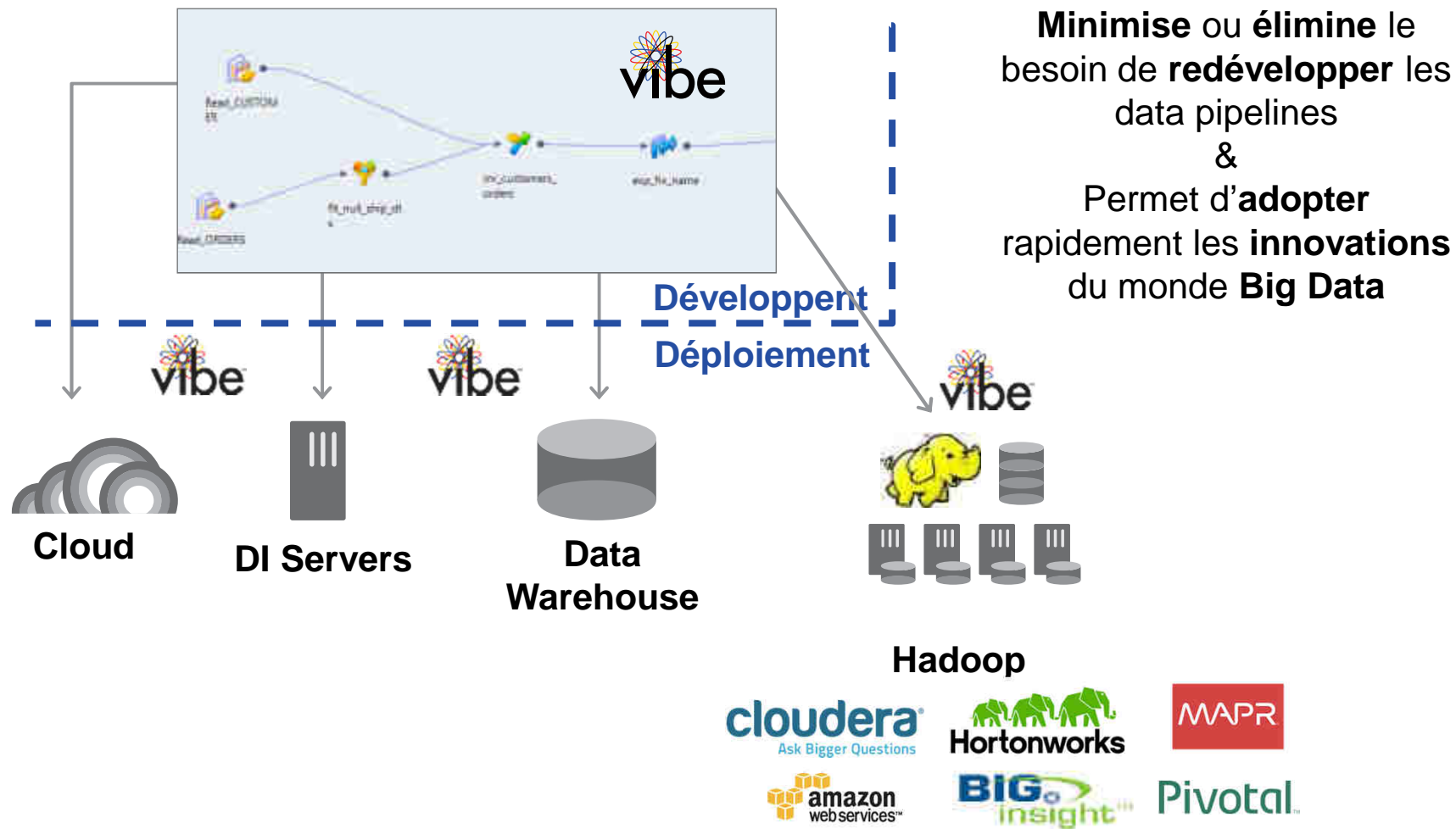
Intégration  
+Simple



# Changer de technologie sans risque

'Police d'assurance' aux changements Hadoop

Intégration  
+Simple





# Qualité de données sur Hadoop

## Nettoyer, corriger, dé-dupliquer les données

Intégration  
+Simple

| Address1       | Address2  | Address3 | Address4 | Address5 |
|----------------|-----------|----------|----------|----------|
| 7887 KATY FRWY | SUITE 333 | HOUSTEN  | TX       | 99999    |

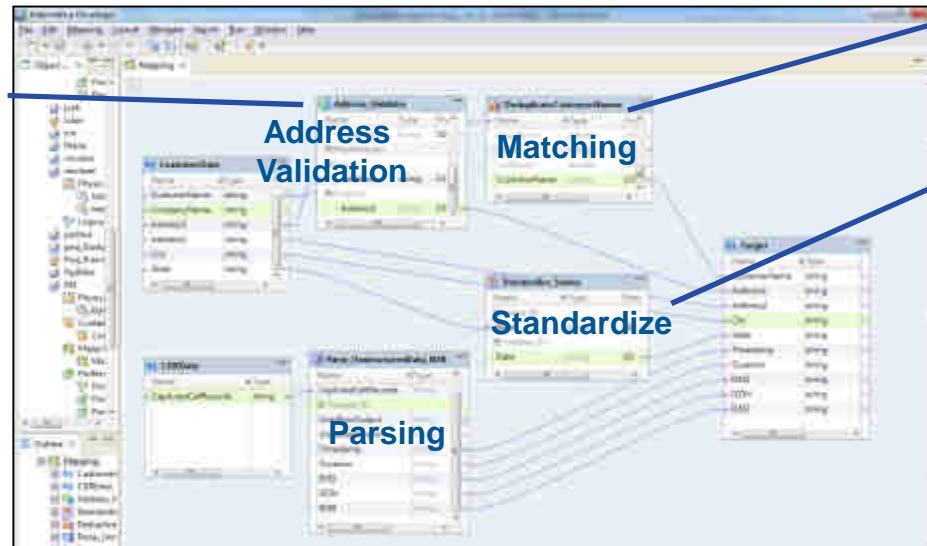
| Street                      | City    | County | StateCode | StateName | ZIP   | ZIP4 | Latitude  | Longitude |
|-----------------------------|---------|--------|-----------|-----------|-------|------|-----------|-----------|
| 7887 Katy Freeway Suite 333 | Houston | Harris | TX        | Texas     | 77024 | 2005 | 29.283427 | -95.46802 |

| Name                    | DOB       | Address                              | City          | State | Zip        |
|-------------------------|-----------|--------------------------------------|---------------|-------|------------|
| W. S. Harrison II PhD   | 1/33/1967 | Medical Center,117/2A #17497 Jackson | E. Hartford   | NY    | 16987      |
| William Stuart Harison  | 1/3/1967  | 117- 2a Jackson Rd.                  | Easthartford  | CT    | 06987      |
| William Stewart Harison | 9/9/99    | 117 Jackson Road, Suite 2A           | Hartford East | CT    | 06987      |
| Doctor Bill Harisen jr  | 1/13/1967 | 117 Jacson Room 2a                   | HartfordCT    |       | 6984       |
| Harrison William Doctor |           | 2a Jackson Rd #174978                | Hartford      | CT    | 06987-4573 |

Réconciliation (Matching)  
Probabiliste & Déterministe

Validation, enrichissement  
d'adresses & Géocodage sur  
260 pays

Parsage des données non  
structurées de tout types  
de données (client/  
produits/ réseaux sociaux /  
logs)



Standardisation & Reference  
Data Management

| Status/Tier | Status/Tier |
|-------------|-------------|
| Gold        | GOLD        |
| 1           | SILVER      |
| 3           | BRONZE      |
| Slvr        | unknown     |
| Brnz        |             |
| Gld         |             |
| TBD         |             |
| ??          |             |
| ...         |             |

Exécution native SUR Hadoop

| Product ID | Brand | Description                       |
|------------|-------|-----------------------------------|
| 90017      | iPod  | 4GB, Red iPod Nano //Special Edt. |

| Product_ID | Brand | Size | Color | Description                           |
|------------|-------|------|-------|---------------------------------------|
| 90017      | IPOD  | 4GB  | Red   | 4 Gigabyte Nano Special Edition (Red) |

Description / Unstructured Text

She's very upset. Need to send something to her ASAP: 'BROADCASTING HOUSE ATTN: HILARY THOMAS, ROOM G12 ,ACCOUNT SERVICES ENGINEERING'

| Location                    | Person        | Organization                    | Sentiment | Noise |
|-----------------------------|---------------|---------------------------------|-----------|-------|
| BROADCASTING HOUSE ROOM G12 | HILARY THOMAS | ACCOUNTING SERVICES ENGINEERING | Upset     | ATTN  |

# Natural Language Processing

## Extraction d'entités et classification

Intégration  
+Simple

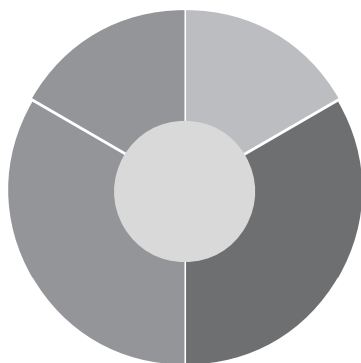
The screenshot displays the Informatica Developer interface. On the left, the 'Object Explorer' shows a project structure with folders like 'Applications', 'Reference Tables', and 'Workflows'. A workflow named 'pm\_twitter\_model' is selected. The main workspace shows a 'Probabilistic Mode data' view of a dataset. The data is presented in a table with columns for text, entity names (like AAPL, GOOG, MSFT), and other metadata. A blue arrow points from a text box to the 'Labels' column, which contains the output of the NLP model. The bottom of the interface shows a 'Data viewer' pane with a message: 'Data viewing is not supported for the current selection in the active editor.'

Apprentissage NLP pour trouver et classifier des entités dans une source de données non structurée

# Archivage de données

## Data Archive HDFS

Gouvernance  
des Données



Data Warehouse  
Production Data



Archiver les données dans un format optimisé pour réduire le stockage

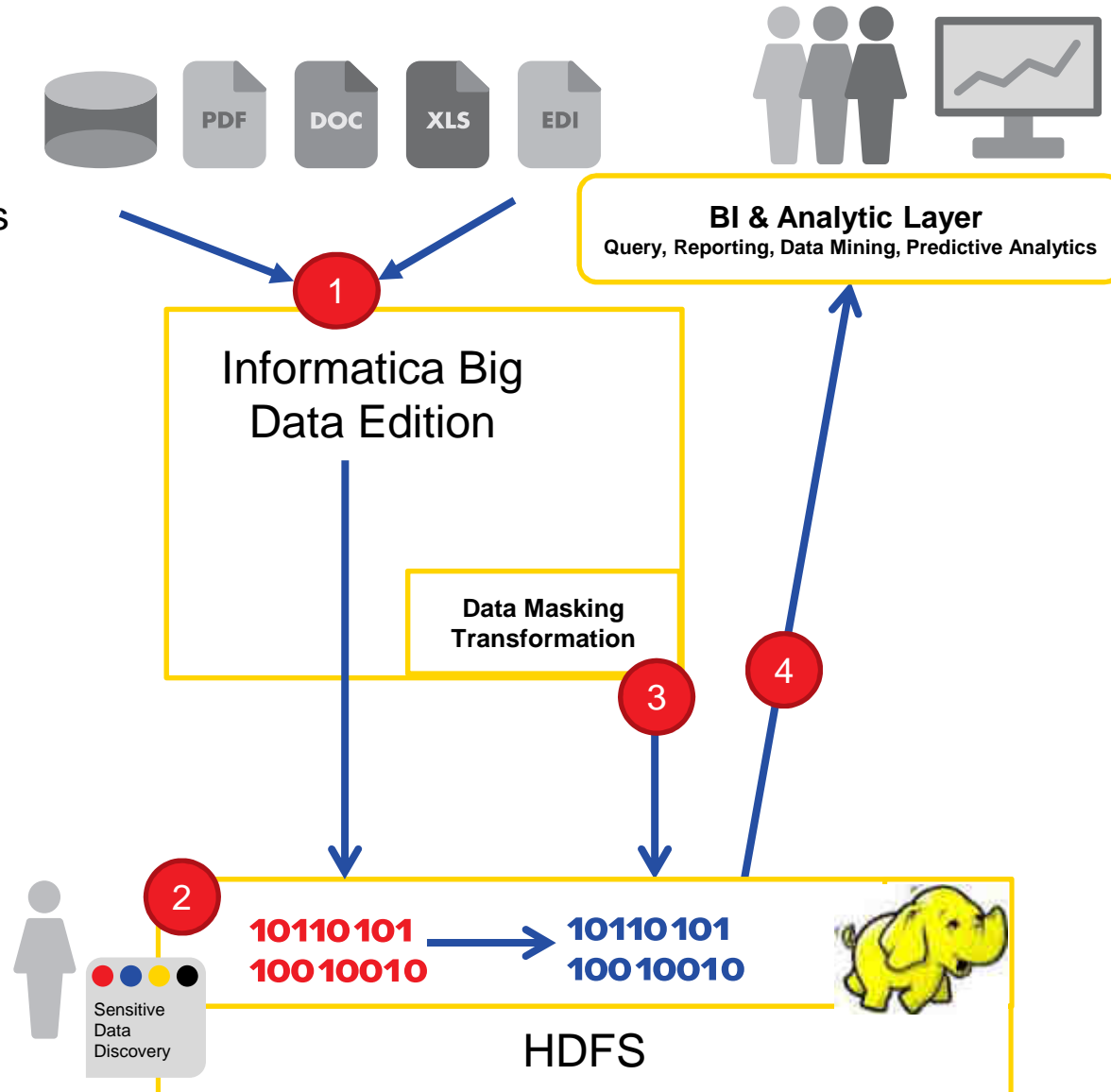
- Compressé (jusqu'à 98%)
- Immuable
- Accessible (SQL, ODBC, JDBC)

# Anonymisation persistante

Persistent Data Masking

Gouvernance  
des Données

1. **Accès universel** aux données
2. Scan et **découverte** des **données sensibles** ('Sensitive Data Discovery')
3. **Anonymisation** des données **dans Hadoop** (en amont ou en mode 'push down')
4. Les **analystes** métiers accèdent aux **données anonymisées**

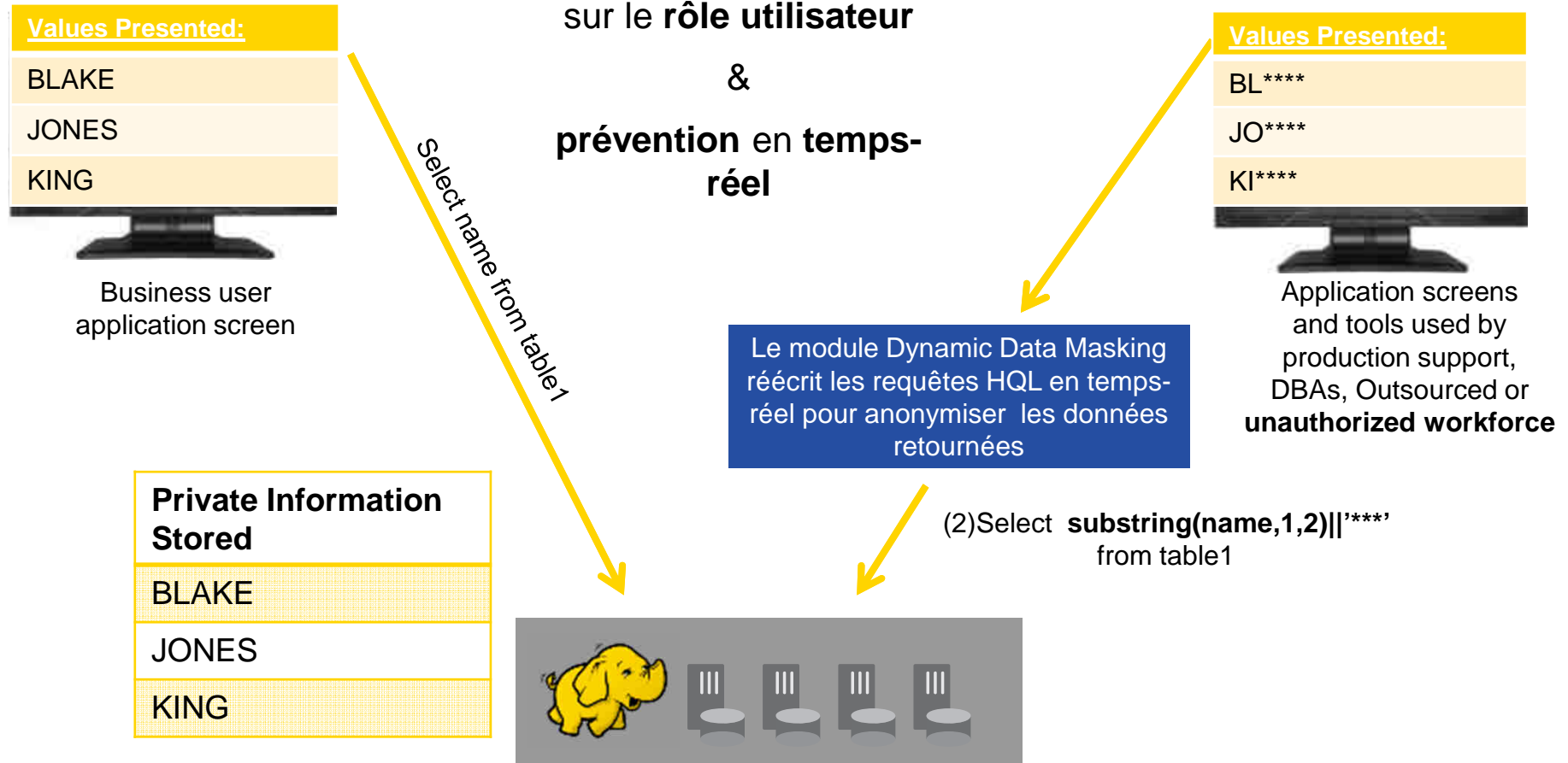




# Anonymisation dynamique

Dynamic Data Masking - Hive

Gouvernance  
des Données



# Profilage des données Hadoop

## Gouvernance des Données

Le résultat du Profilage des données Hadoop est accessible à tous via une interface Web

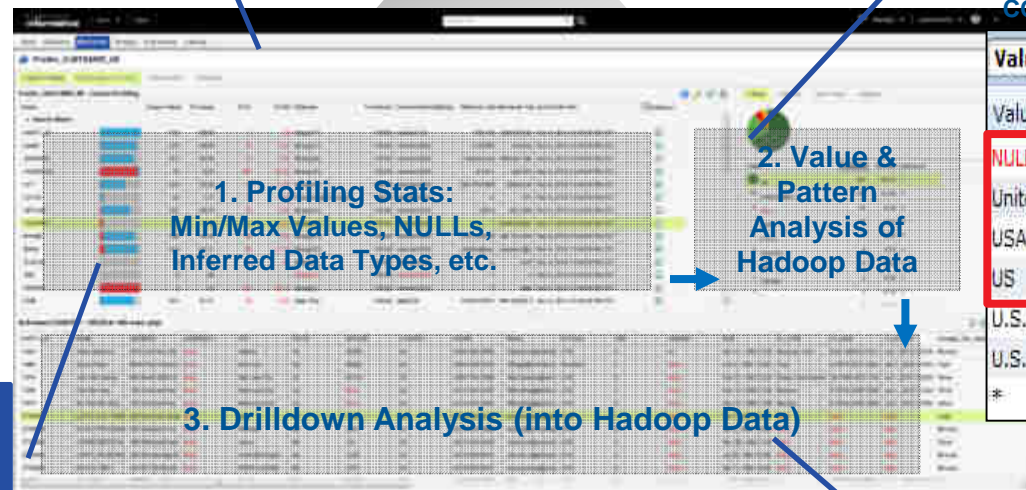


Fréquence des valeurs et des formats pour isoler les inconsistances ou les problèmes de qualité des données

### CUSTOMER\_ID example

| Statistic      | Value  |
|----------------|--|
| Maximum Length | 8  |
| Minimum Length | 6  |
| Bottom (5)     | 10110090<br>10110091<br>10110092<br>10110122<br>10110124 |
| Top (5)        | A5B334<br>A44563<br>A23456<br>19134136<br>19134134       |

Statistiques pour identifier les anomalies dans les données



### COUNTRY CODE example

| Values        | Fre... | Per... |
|---------------|--------|--------|
| Value         | Fre... | Per... |
| NULL          | 16     | 3.20   |
| United States | 2      | 0.40   |
| USA           | 8      | 1.60   |
| US            | 464    | 92.80  |
| U.S.A.        | 6      | 1.20   |
| U.S.          | 3      | 0.60   |
| *             | 1      | 0.20   |

'Drill down' dans les données pour inspecter le jeu de données en entier, y compris les doublons

| IN(GORDY SPAROW G) |                |                |               |          |          |               |            |
|--------------------|----------------|----------------|---------------|----------|----------|---------------|------------|
| CUSTOMER_ID        | CUSTOMER_NAME  | COMPANY_NAME   | ADDRESS1      | ADDRESS2 | ADDRESS3 | ZIP_OR_POSTAL | ISO_CTRY_C |
| 10110239           | GORDIE SPARROW | MCGREGOR GROUP | 1740 BROADWAY | NEW YORK | NY       | 10019         | US         |
| 10116657           | GORDON SPARROW | MCGREGOR GRP   | 1740 BRDWW    | NY       | NY       | 10019         | US         |
| 10178890           | GORDY SPAROW   | UNKNOWN        | BROADWAY      | NEW YORK | NY       | 10019         | USA        |

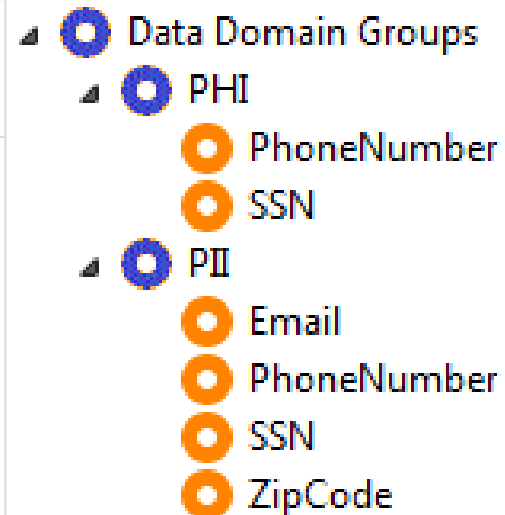
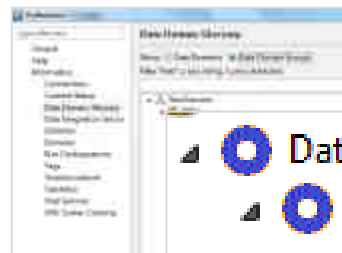
# Recherche de Patterns dans Hadoop

## Sensitive Domain Discovery

Gouvernance  
des Données

Utilisation de règles / mapplets Informatica pour identifier la signification des données Hadoop  
Extraction des données sensibles  
(Carte de crédit, Sécurité Sociale, etc.)

Discovery logic  
automatically pushed down  
and run on Hadoop



PHI: Protected Health Information  
PII: Personally Identifiable Information  
Scalable to look for/discover ANY Domain type



| Columns and Domains  |                   |                   |                    |                     |
|--|-------------------|-------------------|--------------------|---------------------|
| Show: <input type="radio"/> Data Domain <input checked="" type="radio"/> Columns |                   |                   |                    |                     |
| Name   | % Data Conforming | Column Name Match | Data Domain Groups | Documented DataType |
| zip5   |                   |                   |                    |                     |
| ZipCode  | 97.00%            | Yes               | PII                | decimal(5)          |
| phone  |                   |                   |                    |                     |
| PhoneNumber  | 76.00%            | Yes               | PHI, PII           | string(14)          |
| email  |                   |                   |                    |                     |
| Email  | 100.00%           | Yes               | PII                | string(17)          |
| credit_card_num  |                   |                   |                    |                     |
| CreditCardNumber   | 100.00%           | Yes               | PCI                | string(19)          |
| cust_no  |                   |                   |                    |                     |
| SSN  | 100.00%           | No                | PHI, PII           | string(11)          |

Vision/partage des rapports sur les domaines/données sensibles contenues dans Hadoop. Capacité de 'drill down' pour voir les données suspectes.

# Administration, Monitoring unifiés

Gouvernance  
des Données

The screenshot displays the Informatica Monitoring console. On the left, a tree view shows the hierarchy of workflows, with 'Workflow' selected. The main panel shows a table of workflows with columns: Name, Type, State, Start Time, Updated at, Stopped Time, and Started By. The table lists various workflows, including 'Daily\_Stock\_Recommendation', 'Recommend\_Stocks', 'Calc\_DailyRecommendations', and 'Fetch\_Latest\_Tweets'. The 'State' column indicates the status of each workflow, such as 'Running' or 'Completed'. A blue callout box points to the 'Recommend\_Stocks' workflow, stating: 'Traçabilité de bout en bout De l'exécution native Hadoop des jobs MapReduce'. Below the table, a 'Viewing 3 results' section shows the details of a selected workflow, including the Hive script used for execution. The script is a HiveQL query that creates a table and inserts data from a Hadoop MapReduce job. A blue callout box points to the script, stating: 'Vue des scripts Hive générés (~spécification fonctionnelle)'. The Informatica logo is visible in the bottom left corner.

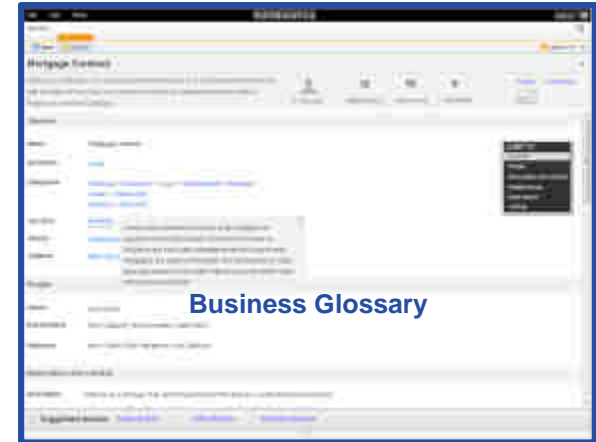
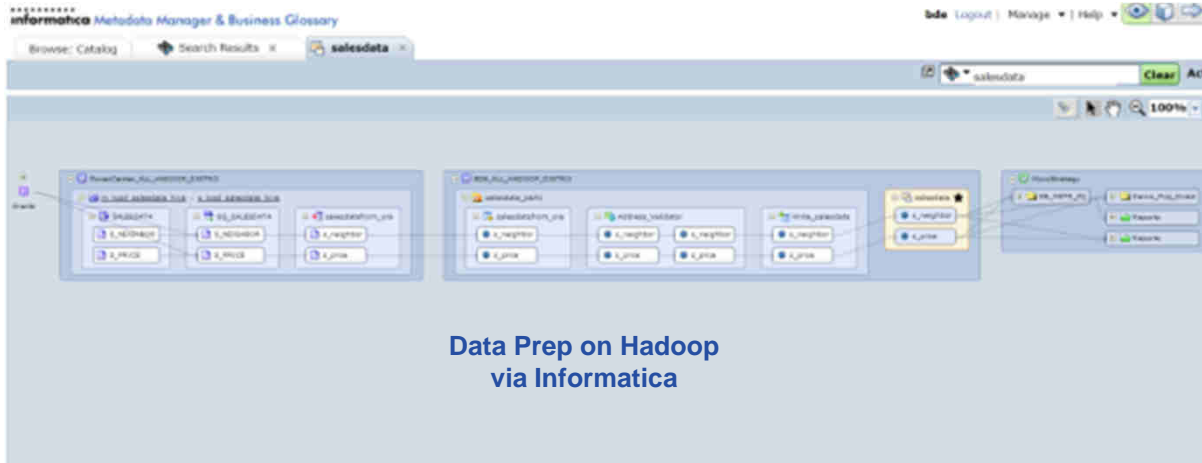
Traçabilité de bout en bout  
De l'exécution native Hadoop  
des jobs MapReduce

Vue des scripts Hive  
générés  
(~spécification  
fonctionnelle)



## Vue de bout en bout des flux de données

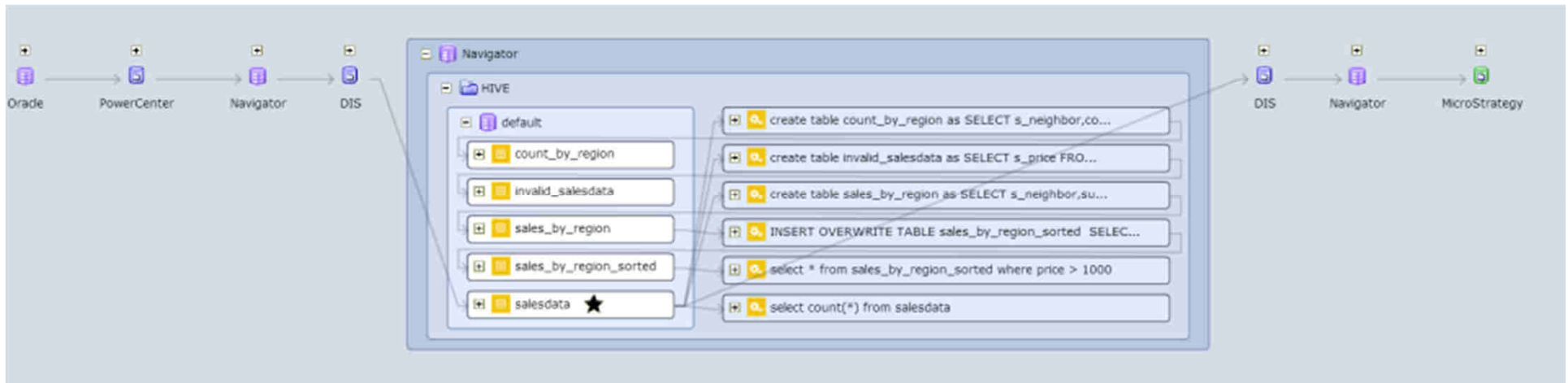
# Gouvernance des Données



## Data Source

## Hive HQL

## Target BI/Analytic App

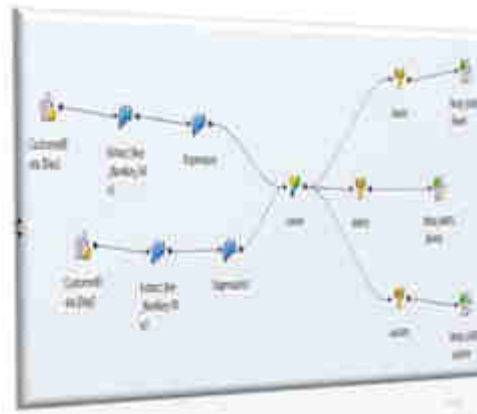


# Une architecture simple, vos projets accélérés

- 200+ connecteurs certifiés & performants
- 100+ parseurs pré-packagés
- Complex Data Parsing
- Ingestion multiple (batch, CDC)
- Data Streaming

- Développement simple et graphique
- Exécution native Hadoop
- Réutilisation des ressources
- Simple à exécuter et monitorer
- Anonymisation simple et intégrée

- Qualité de donnée étendue
- Profilage et découverte de patterns
- Gestion avancée des métadonnées
- Glossaire métier
- Data linéage 'end to end'



# Les Solutions Informatica Big Data:

## Une architecture simple, vos projets accélérés

### Des Ressources disponibles

**Careerbuilder.com** montre dans une étude qu'il y a **27000** demandes d'experts Hadoop et seulement **3000** CV avec des compétences **Hadoop**  
– Il y a +**100000** développeurs **Informatica** expérimentés mondialement



### Une Meilleure productivité

Avec son interface simple et graphique, robuste et connue, les Développeurs Informatica sont **5x plus productifs**  
- basé sur nos POC clients



**Hand coding**  
**4 semaines**

**Informatica**  
**4 jours**



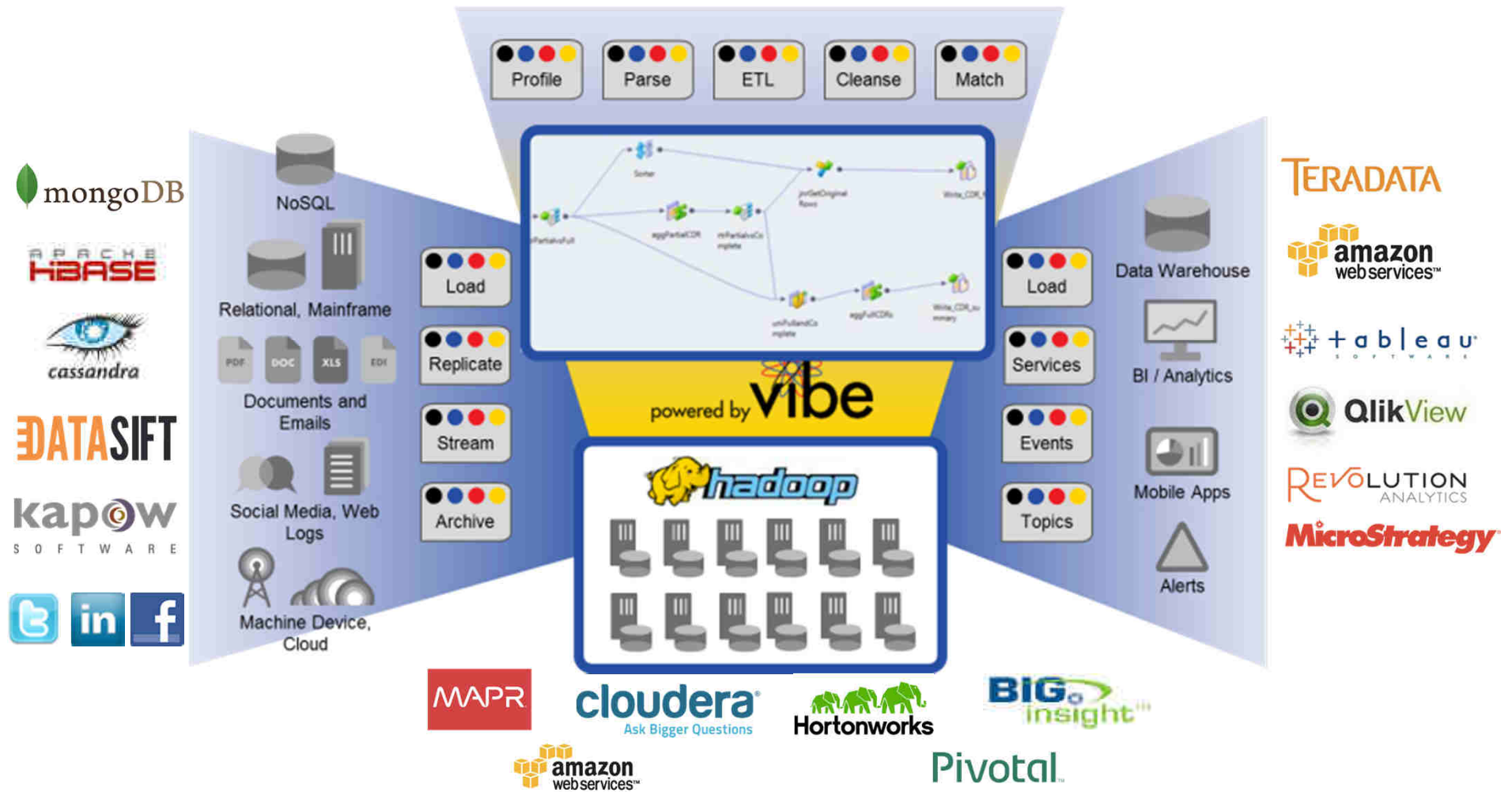
### Un déploiement accéléré

Avec Informatica tout ce que vous définissez en Développement peut immédiatement être déployé en Production.  
- "Notre POC Big Data POC était si concluant que le métier nous a demandé de le garder et de l'exécuter dans le système de production"



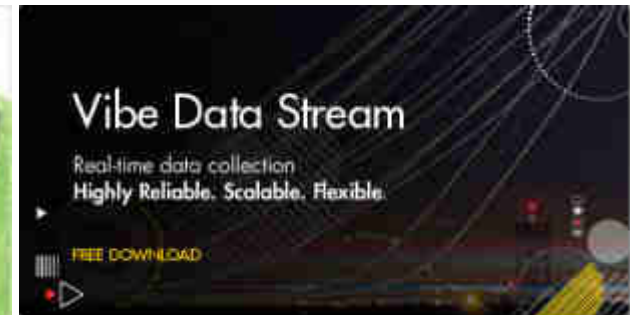
# Ecosystème Informatica & Big Data

Tirez le meilleur parti de l'open source, sans codage



# Essayez Informatica Big Data Edition

[marketplace.informatica.com/bigdata](http://marketplace.informatica.com/bigdata)



TRIAL DOWNLOADS

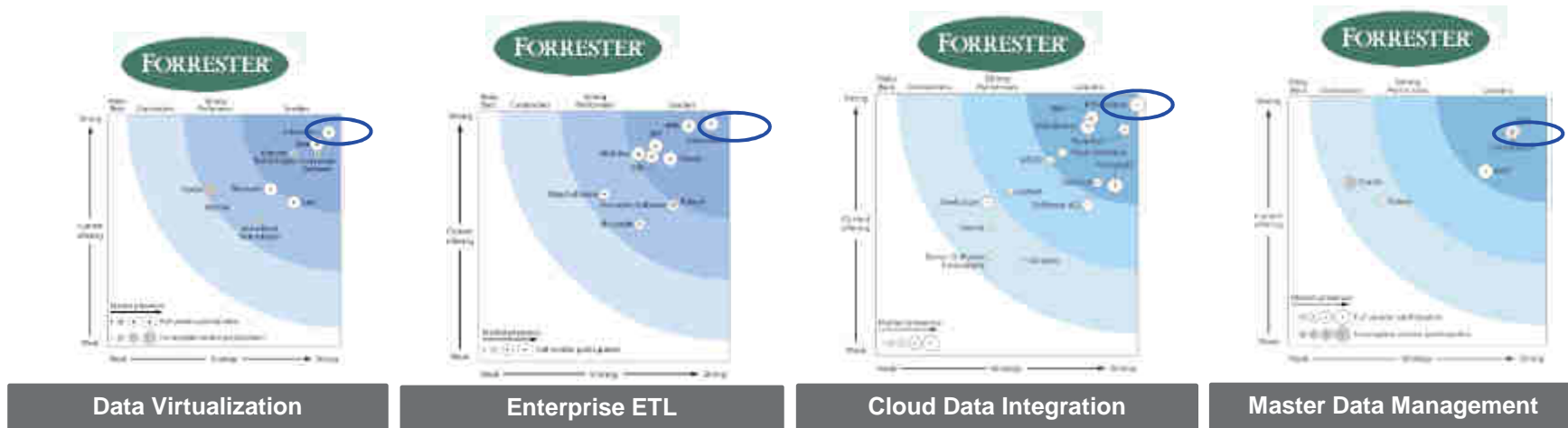
REFERENCE ARCHITECTURES

TRAINING & WEBINARS





# Informatica : une technologie éprouvée



## Informatica Platform

