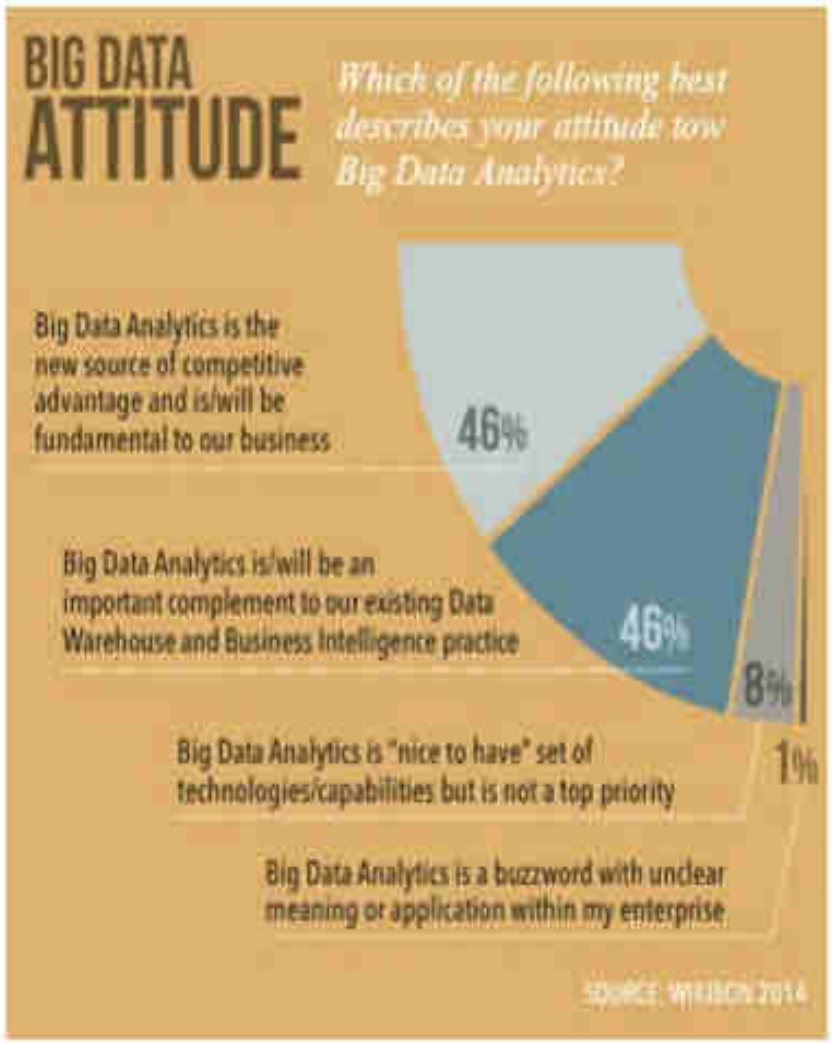
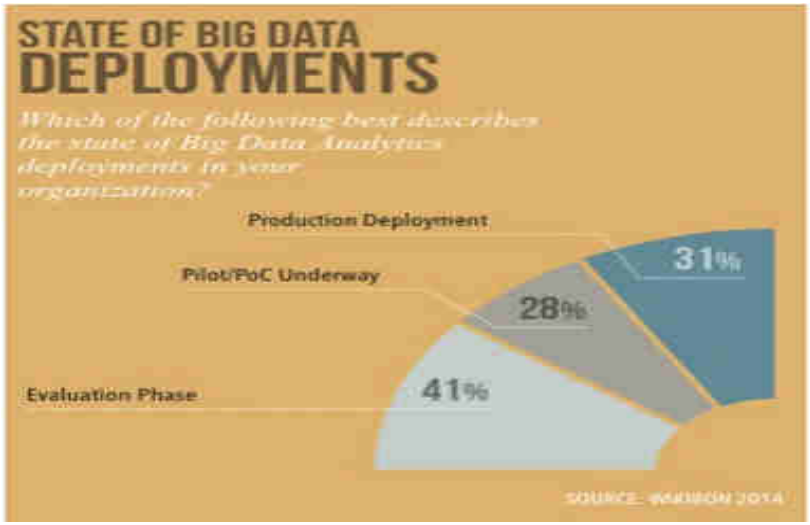
informatica

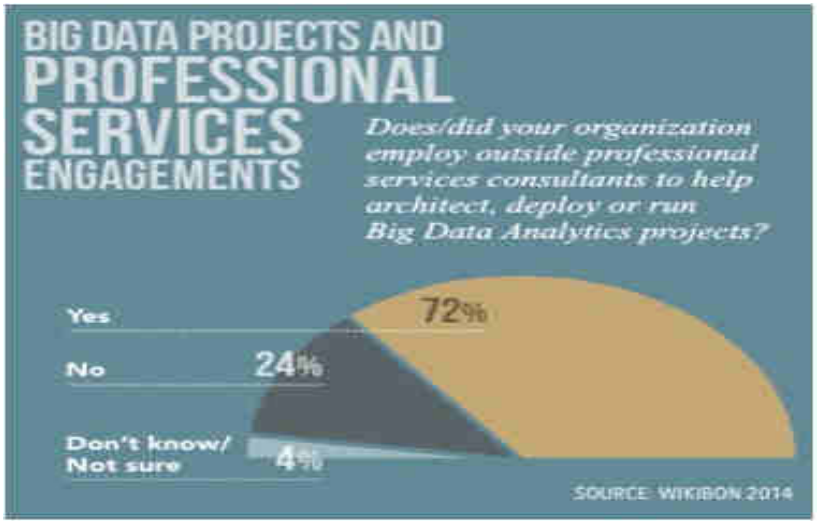
Facilitez l'intégration du Big Data à votre architecture

Mathieu Lagrange,
Product Specialist EMEA
mlagrange@informatica.com

Les tendances du Big Data Aujourd'hui





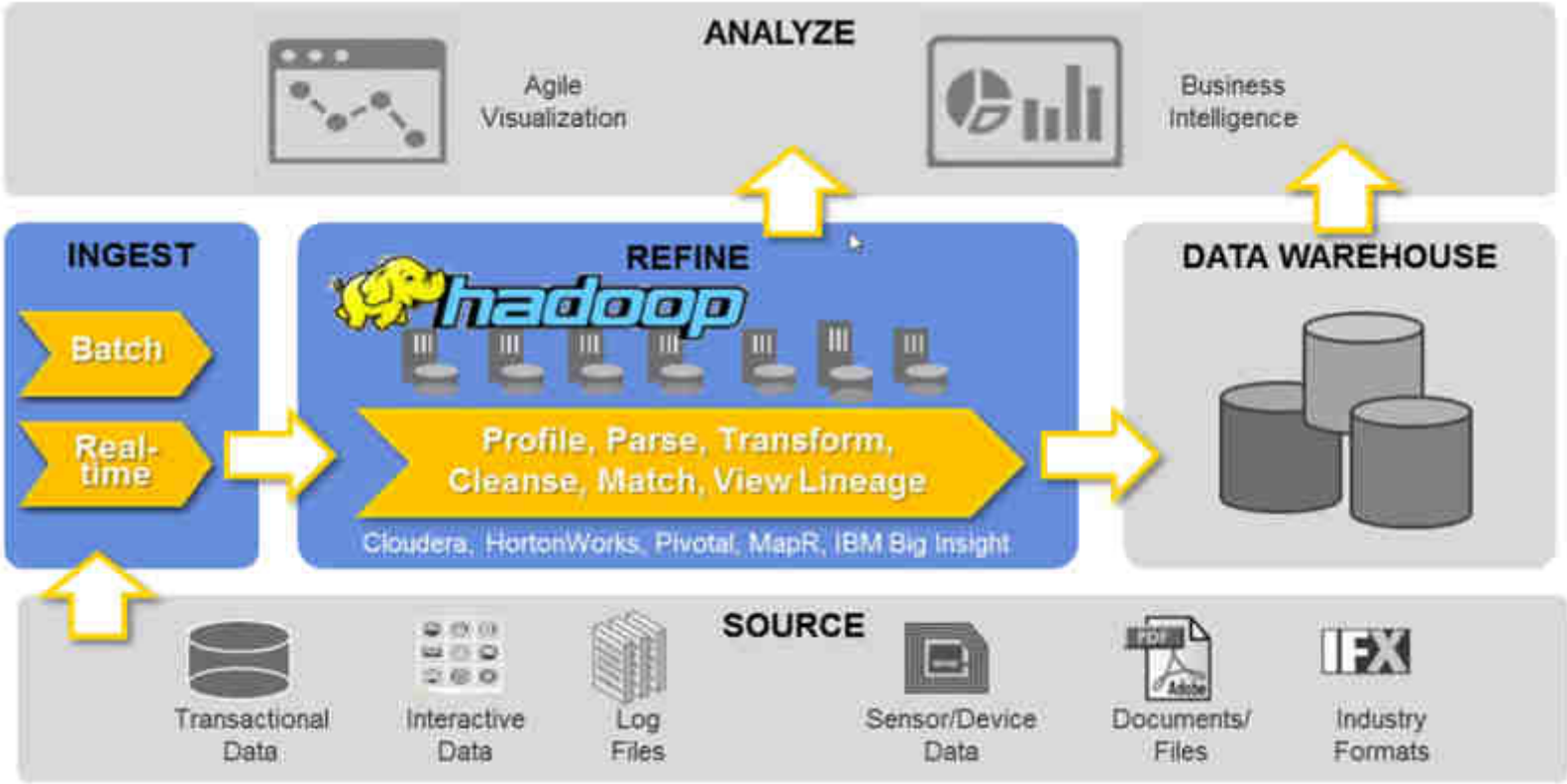


informatica


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

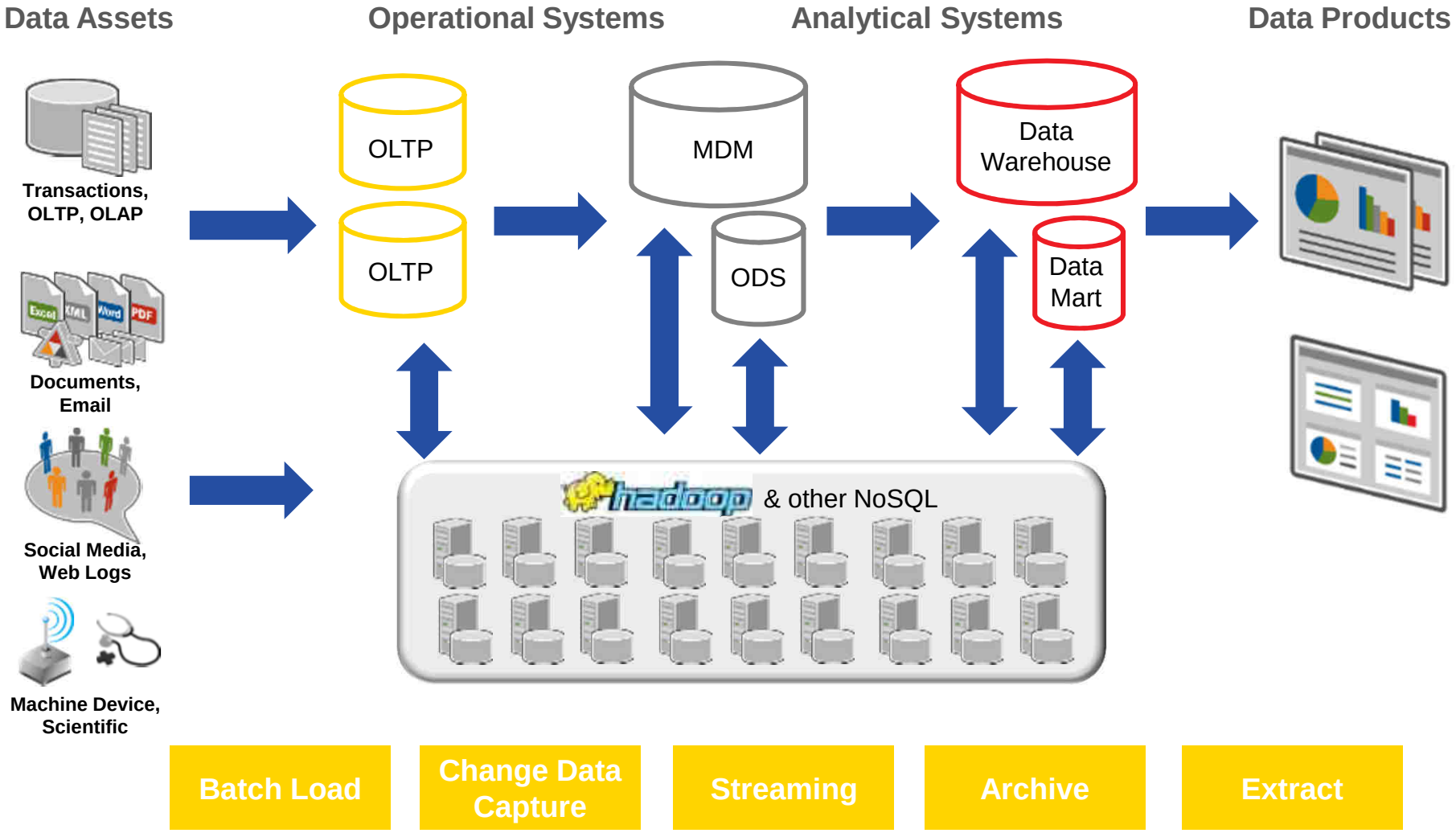
Intégration +Simple

Gouvernance des donnés

informatica

Simplifiez votre architecture Big Data

Collecte +Rapide



informatica

Connectivité Universelle

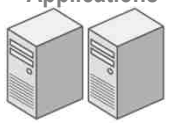
Collecte +Rapide

Messaging, and Web Services

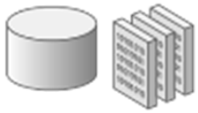
WebSphere MQ	Web Services
JMS	TIBCO
MSMQ	webMethods
SAP NetWeaver XI	

JD Edwards	SAP NetWeaver
Lotus Notes	SAP NetWeaver BI
Oracle E-Business	SAS
PeopleSoft	Siebel

Packaged Applications



Relational and Flat Files



Oracle	Informix
DB2 UDB	Teradata
DB2/400	Netezza
SQL Server	ODBC
Sybase	JDBC

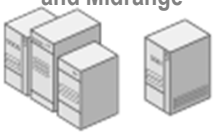
Salesforce CRM	ADP
Force.com	Hewitt
RightNow	SAP By Design
NetSuite	Oracle OnDemand

SaaS/BPO

salesforce.com

ADP

Mainframe and Midrange



ADABAS	VSAM
Datacom	C-ISAM
DB2	Binary Flat Files
IDMS	Tape Formats…
IMS	

EDI–X12	AST
EDI-Fact	FIX
RosettaNet	SWIFT
HL7	Cargo IMP
HIPAA	MVR

Industry Standards

HL7 PAA

Unstructured Data and Files

Word, Excel	Flat files
PDF	ASCII reports
StarOffice	HTML
WordPerfect	RPG
Email (POP, IMPA)	ANSI
HTTP	LDAP

XML	ebXML
LegalXML	HL7 v3.0
IFX	ACORD (AL3, XML)
cXML	

XML Standards

IFX ebXML

MPP Appliances

Pivotal	Teradata
Vertica	Aster
Netezza	

Facebook	Kapow
Twitter	Datasift
LinkedIn	

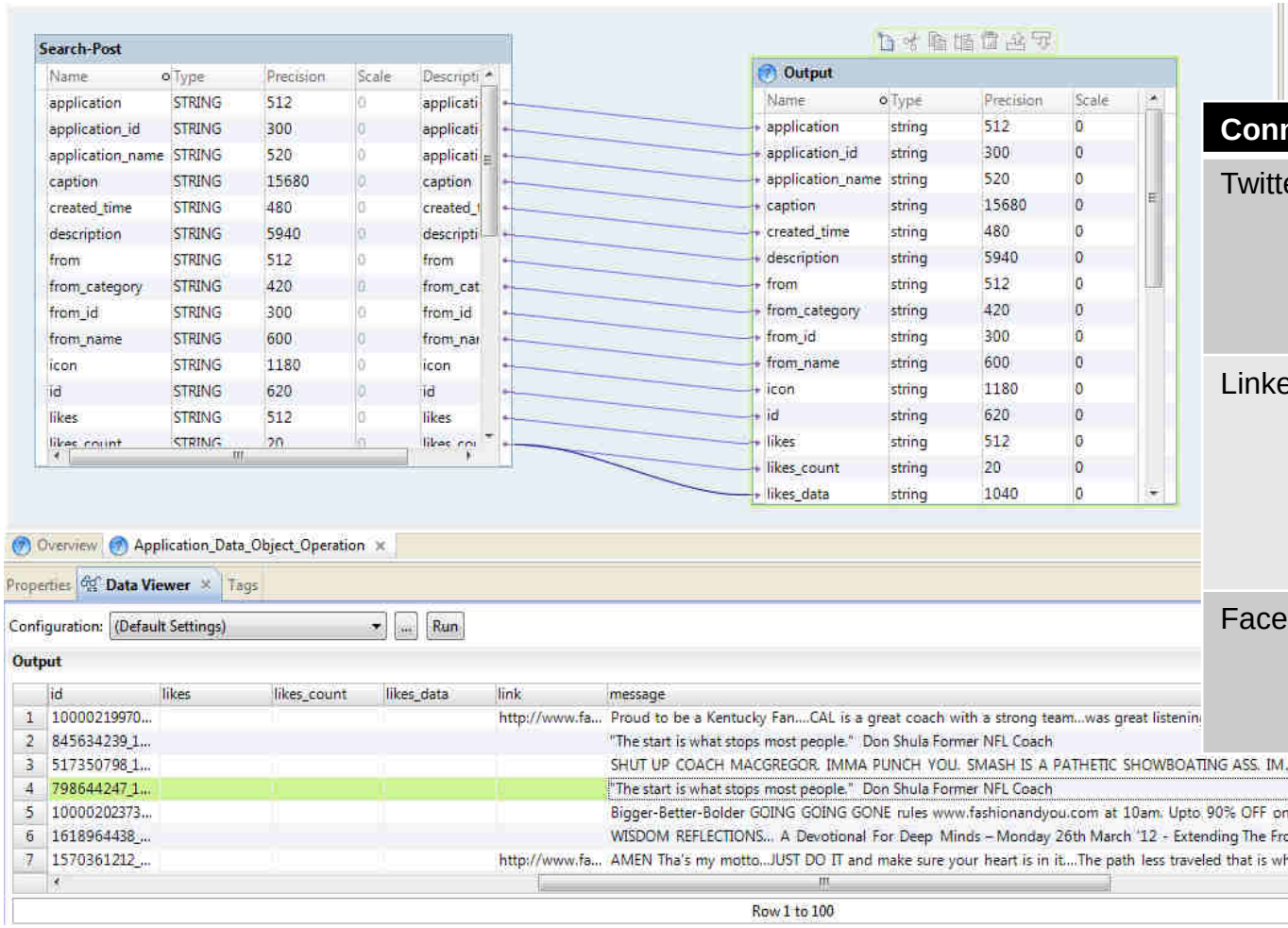
Social Media

informatica

hadoop mongoDB HBASE Cassandra

Connectivité Réseaux Sociaux

Collecte +Rapide

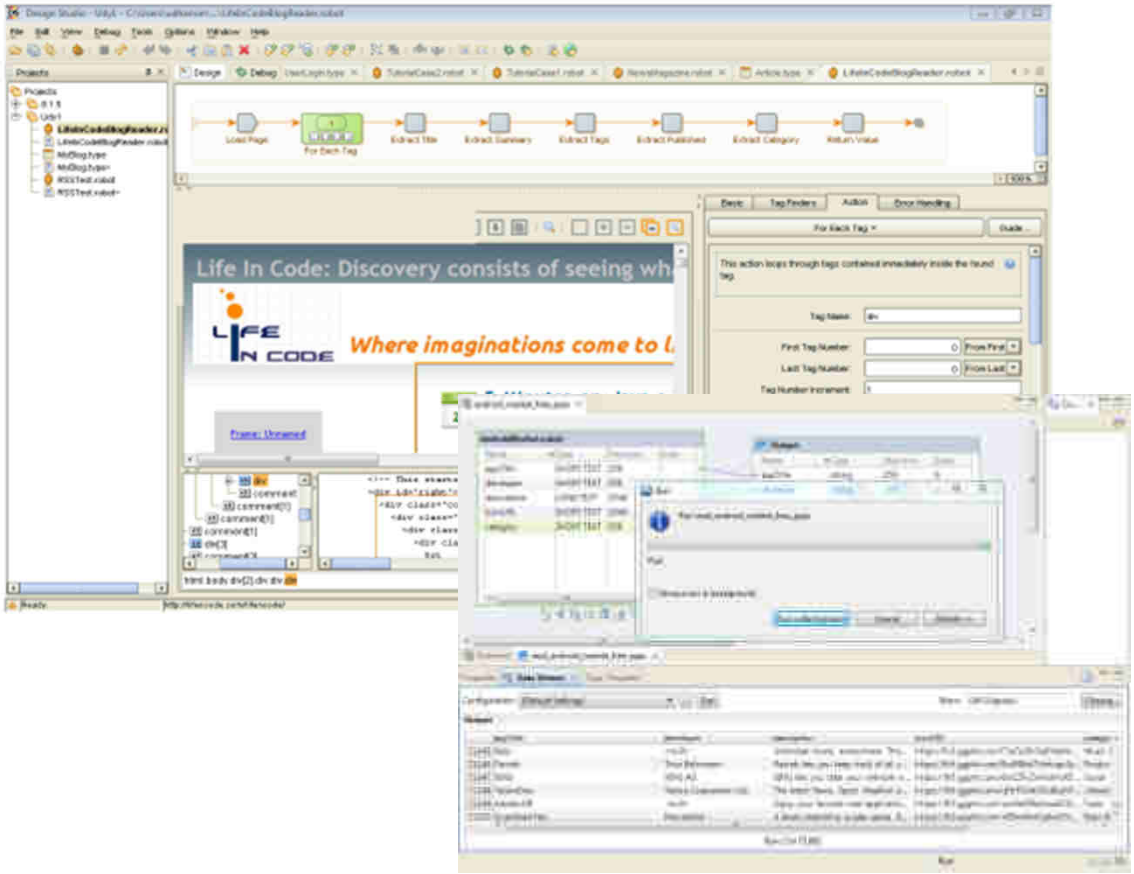


Connexion	Fonctionnalités
Twitter	• Recherche de tweets • Extraction des données de profil utilisateur
LinkedIn	• Extraction des données de profil utilisateur • Extractions des dates de connexions
Facebook	• Extraction des statuts Facebook + critères de recherche

informatica

Accéder aux données du Web

Collecte
+Rapide



kapow SOFTWARE

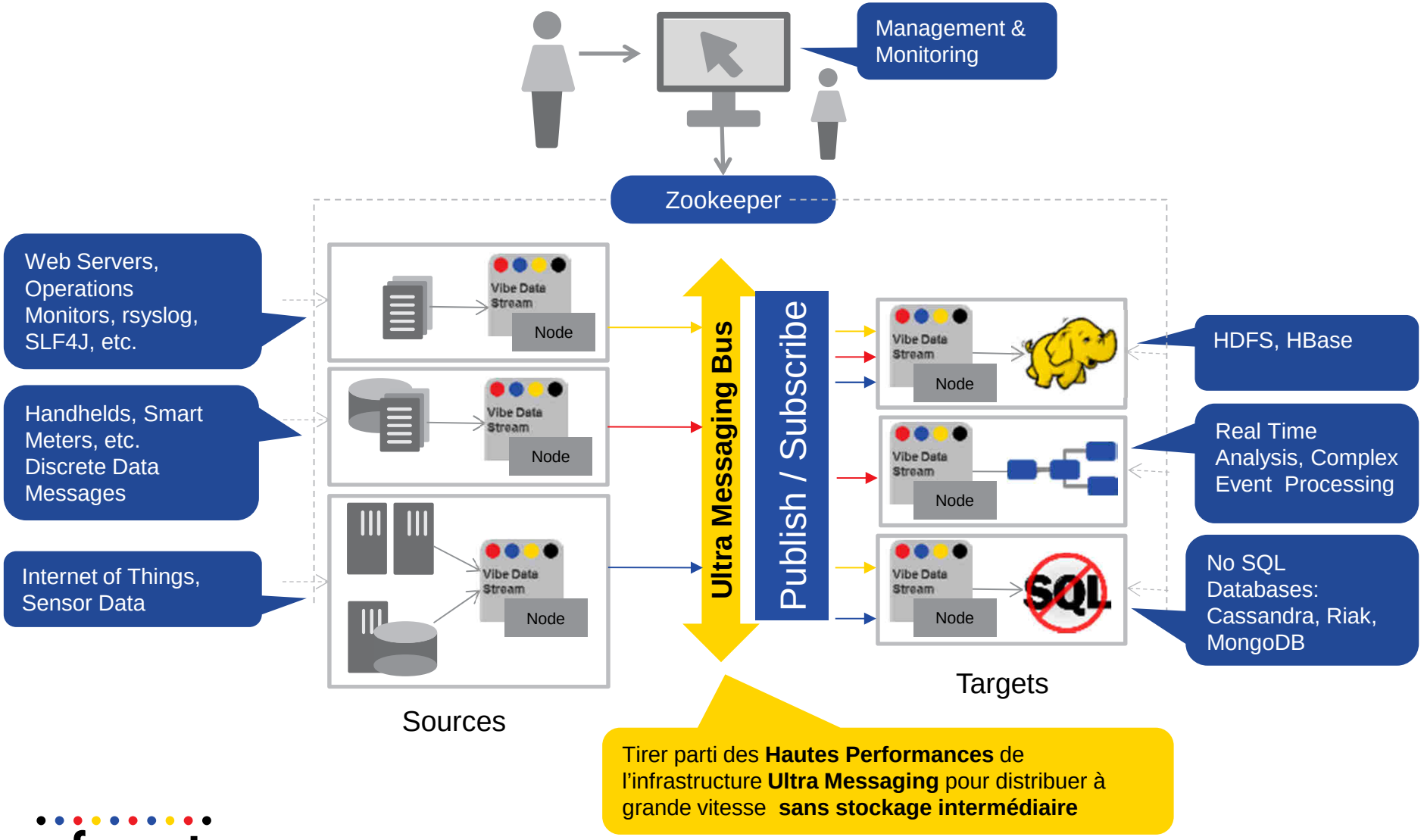
- **Simuler une navigation Web** pour extraire les bonnes informations de manière graphique et rapide
- Extractions **de tout type d'application** Web
- Outiller sa cellule de **veille technologique**
- **Connectivité** native **Informatica**

informatica

Big Data Streaming

Vibe Data Stream- performance

Collecte +Rapide

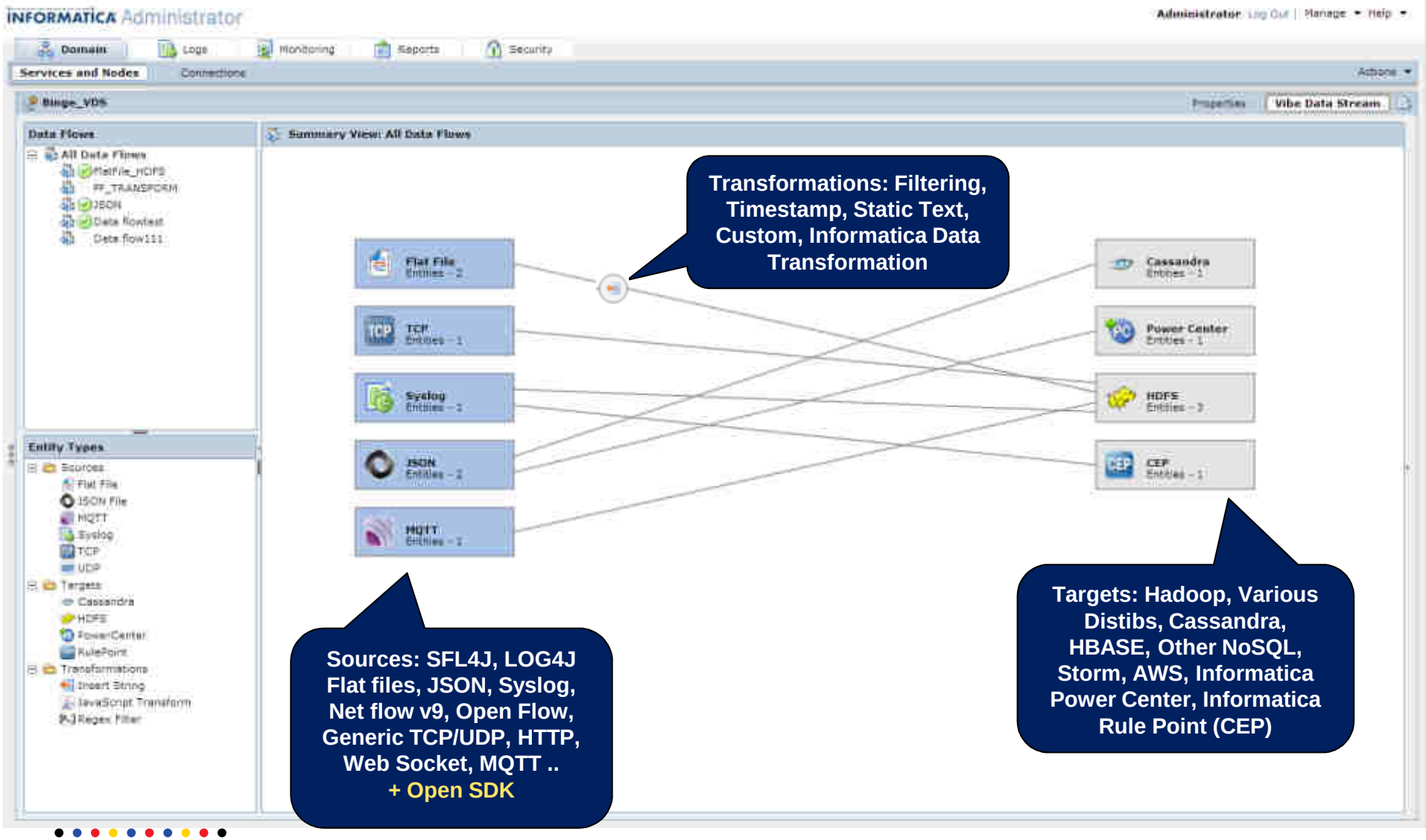


informatica

Big Data Streaming

Vibe Data Stream - Interface Graphique

Collecte +Rapide

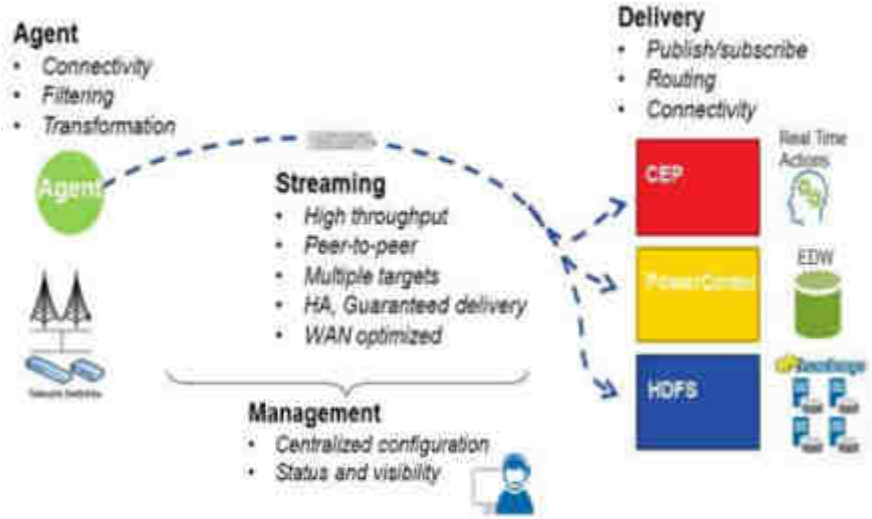


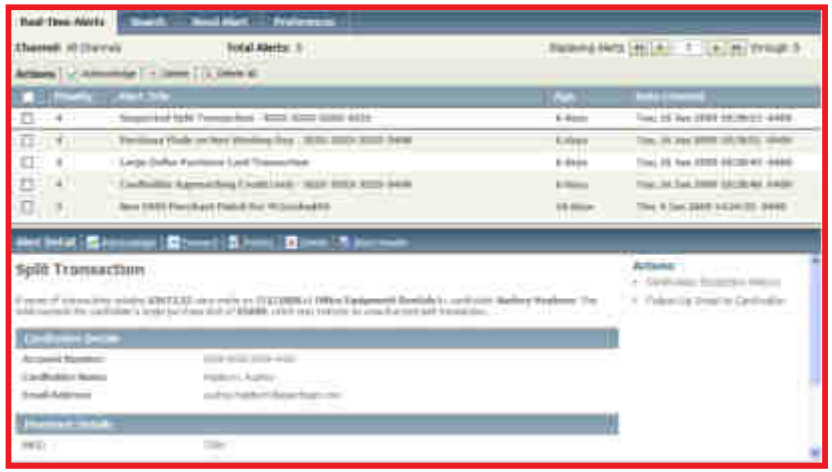
informatica

Big Data Streaming

Vibe Data Stream – Complex Event Processing

Collecte +Rapide



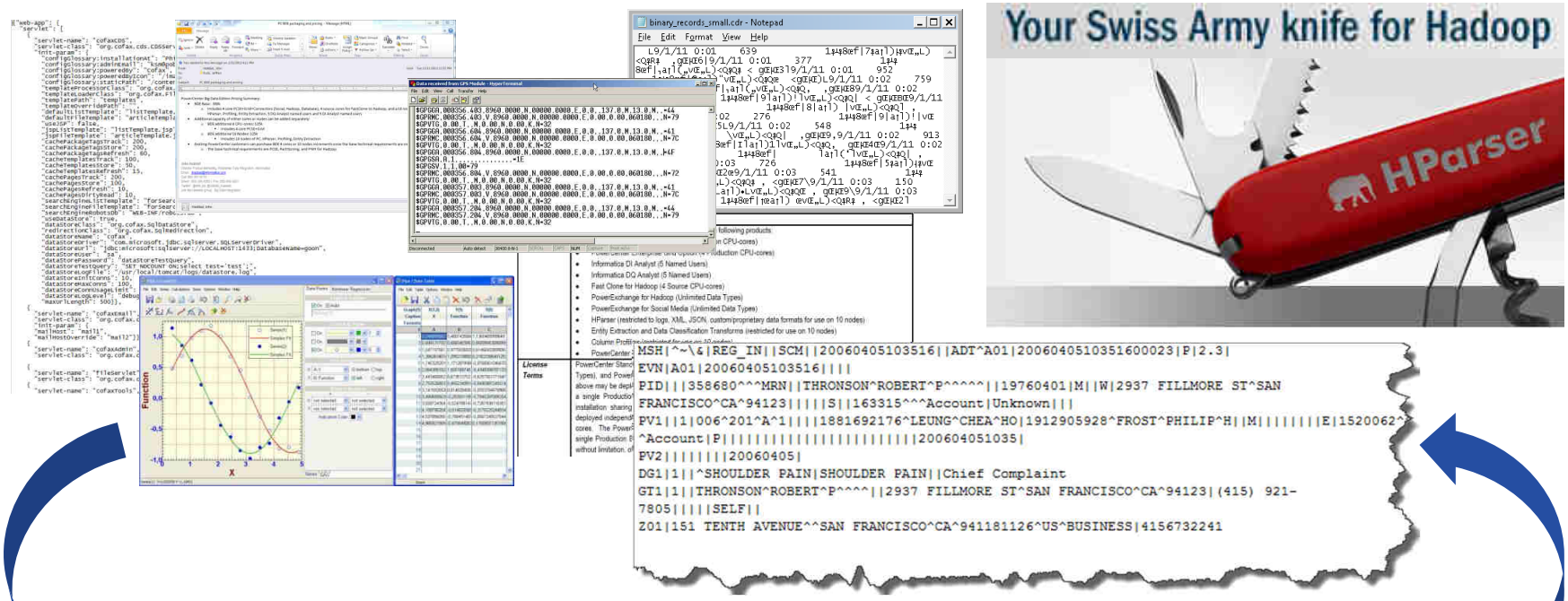


- **Collecte Haute Performance** des données streaming LAN/WAN **sans 'Router'** intermédiaire
- **Interface graphique** & **simple** pour configurer, (re)déployer, monitorer
- **Ingestion** continue en **temps-réel** des données logs, machines et autres (sdk)
- **Interactions temps-réel** (Complex Event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

informatica

Traiter les données Complexes

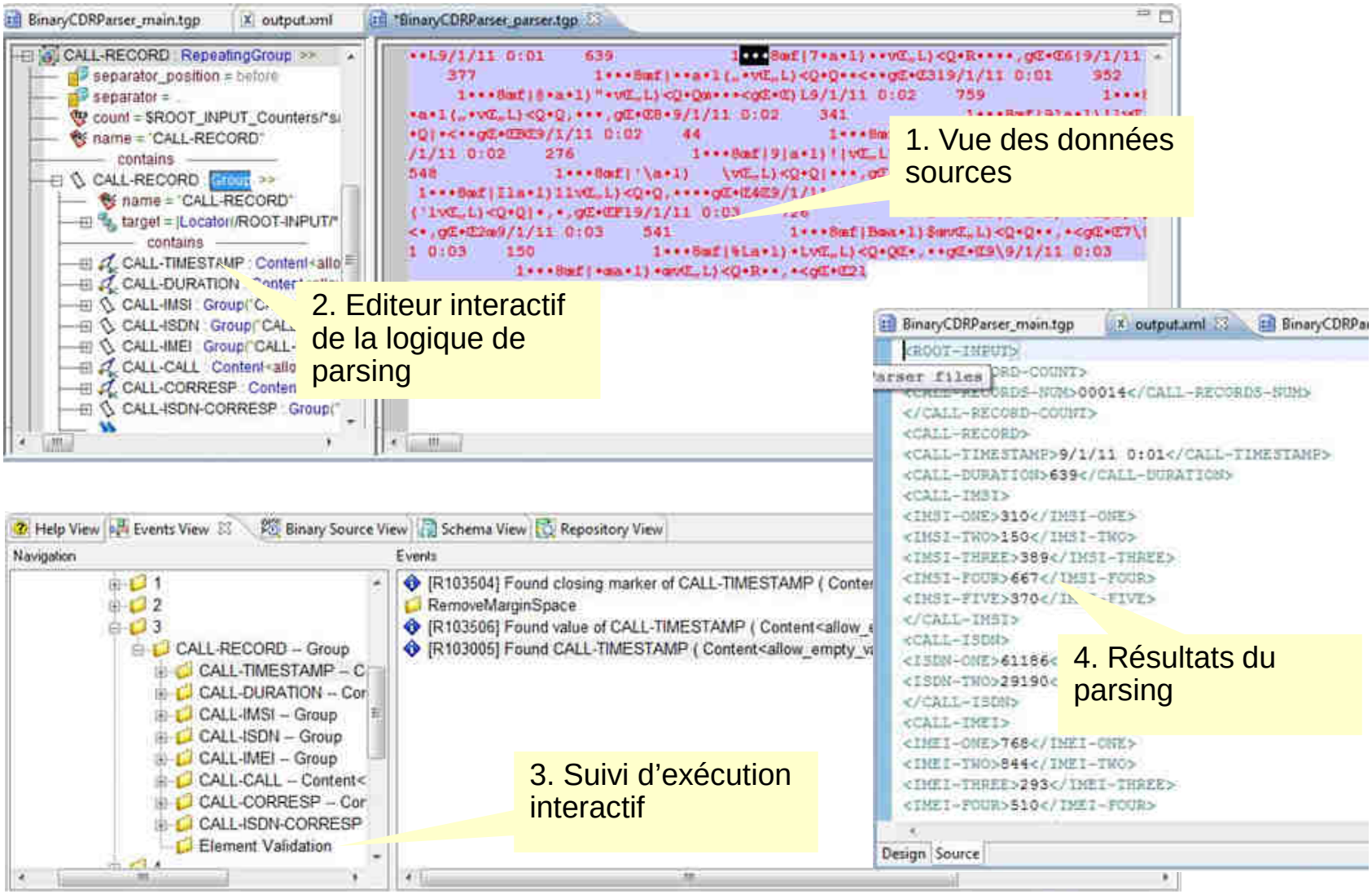
Intégration
+Simple



informatica

Un Exemple : librairie ASN.1

Intégration +Simple



informatica

Traiter les données complexes Hadoop

Intégration +Simple

The broadest coverage for Big Data

Fichiers Plats & Documents

Name = Value
^/>Delimited<¥^

XML

Standards

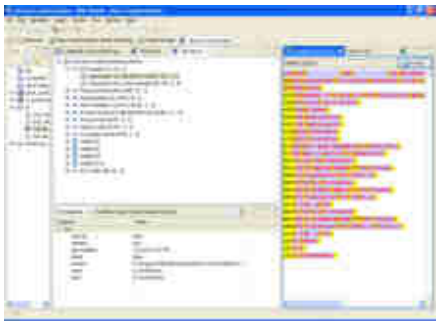
Interactions

social

Device/sensor
scientific

Productivity

- Environnement Graphique
- Bibliothèques de transformation pré-packagées



Any DI/BI architecture

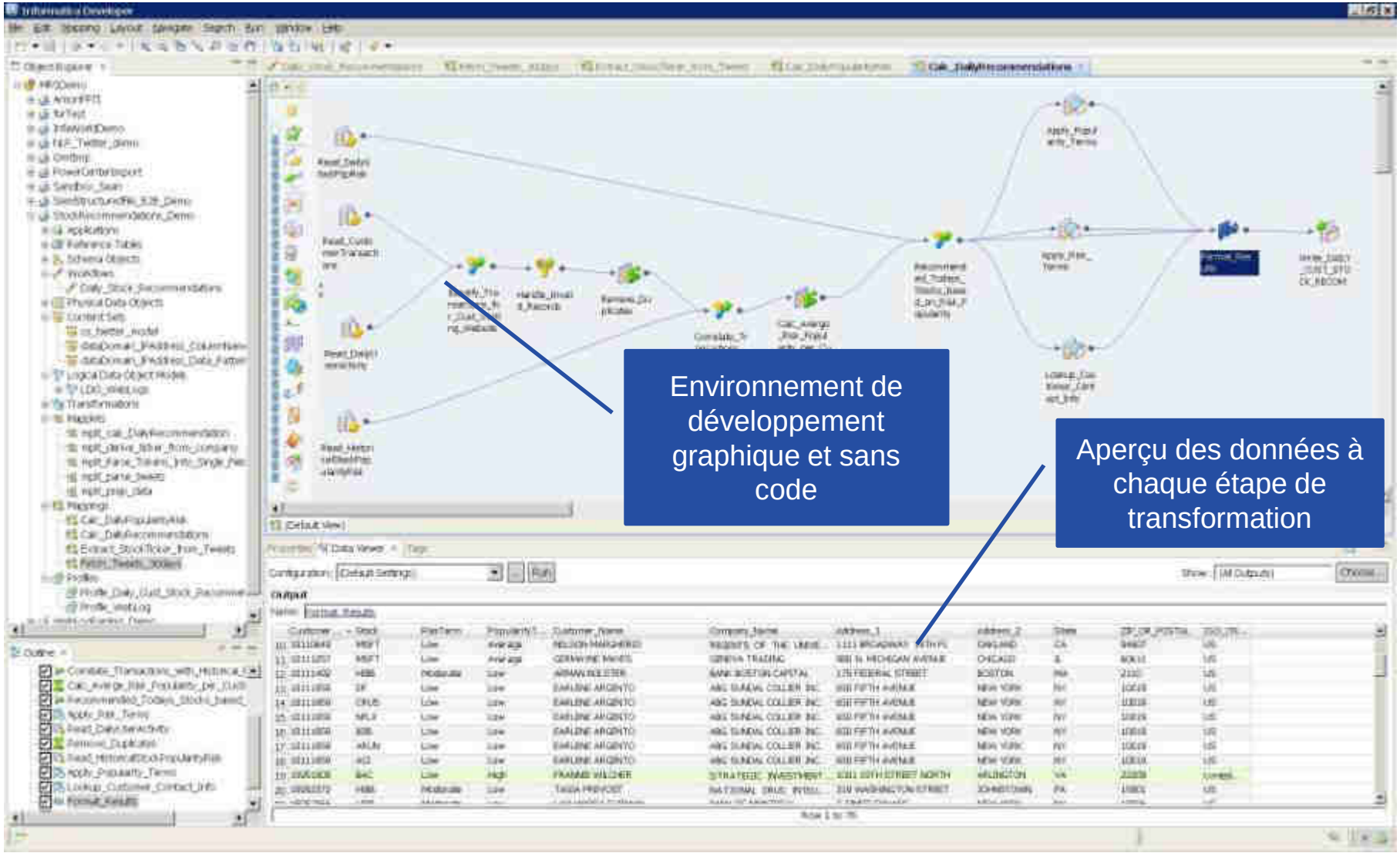
PIG

EDW
MDM

informatica

Intégrer les données sans codage

Intégration +Simple

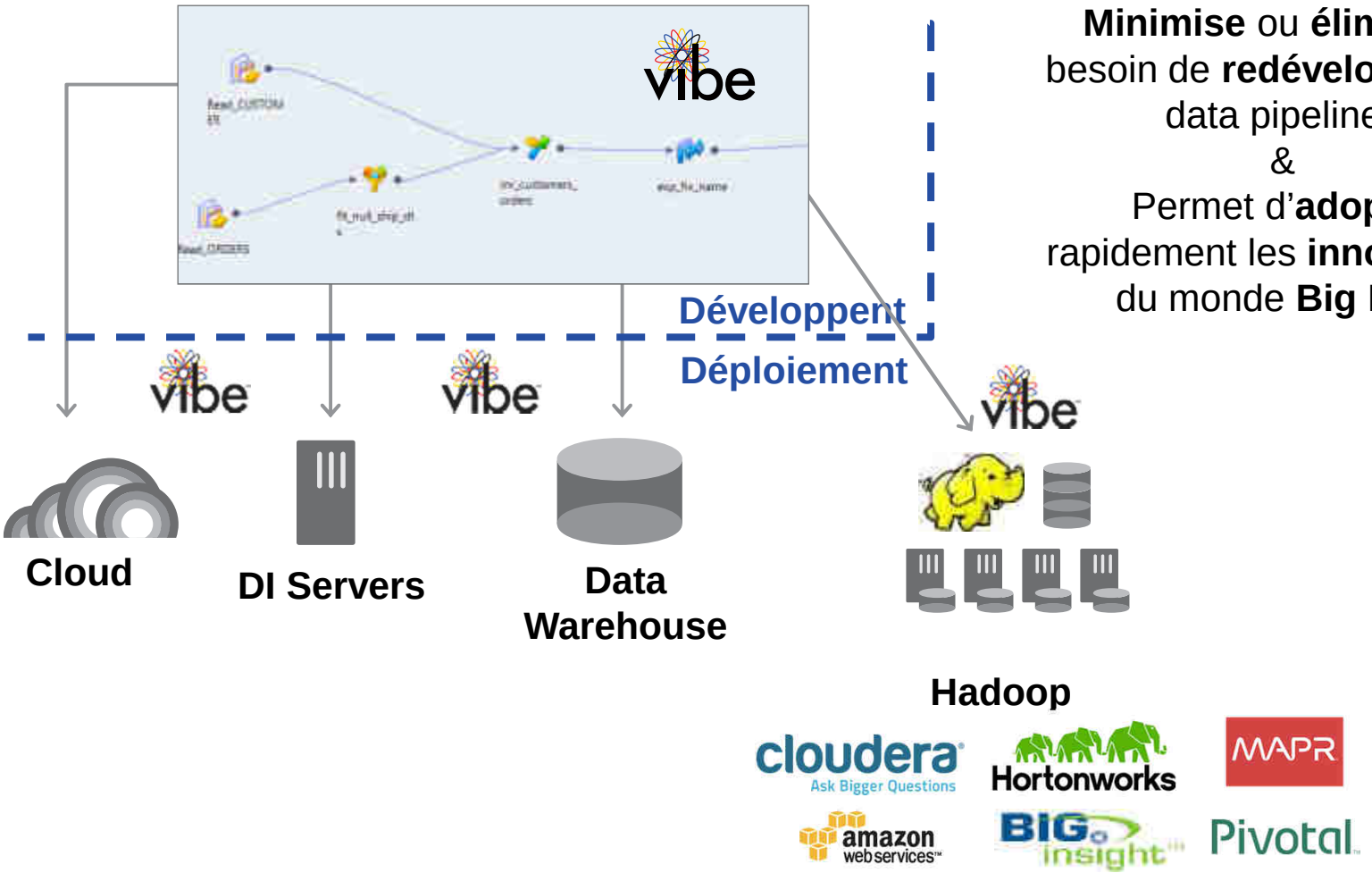


informatica

Changer de technologie sans risque

‘Police d’assurance’ aux changements Hadoop

Intégration +Simple



Minimise ou **élimine** le besoin de **redévelopper** les data pipelines
&
Permet d’**adopter** rapidement les **innovations** du monde **Big Data**

informatica

Qualité de données sur Hadoop

Nettoyer, corriger, dé-dupliquer les données

Intégration +Simple

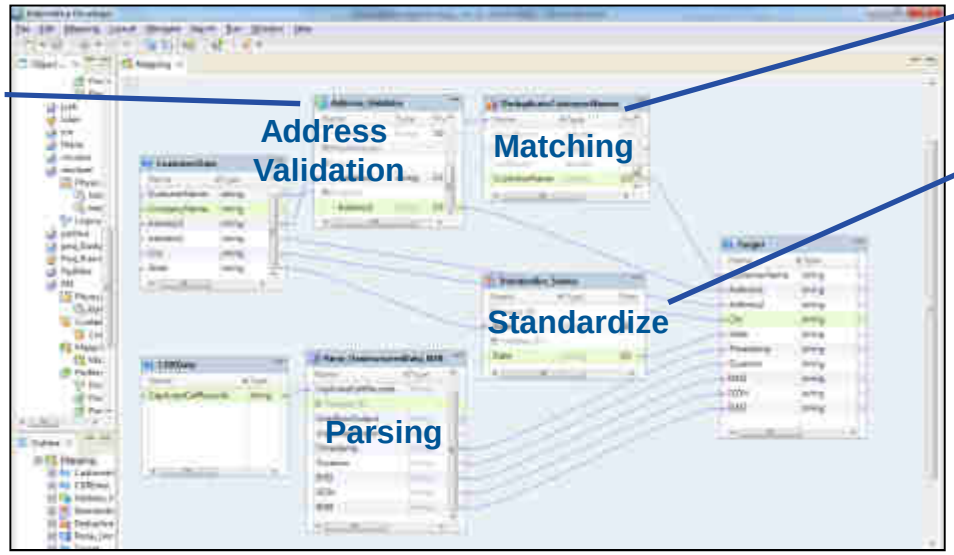
Name	DOB	Address	City	State	Zip
W. S. Harrison II PhD	1/33/1967	Medical Center,117/2A #17497 Jackson	E. Hartford	NY	16987
William Stuart Harison	1/3/1967	117-2a Jacksen Rd.	Easthartford	CT	06987
William Stewart Harison	9/9/99	117 Jackson Road, Suite 2A	Hartford East	CT	06987
Doctor Bill Harisen jr	1/13/1967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

Réconciliation (Matching) Probabiliste & Déterministe

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

Validation, enrichissement d'adresses & Géocodage sur 260 pays



Standardisation & Reference Data Management

Status/Tier	Status/Tier
Gold	GOLD
1	SILVER
3	BRONZE
Slvr	unknown
Brnz	
Gld	
TBD	
??	
…	

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

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)

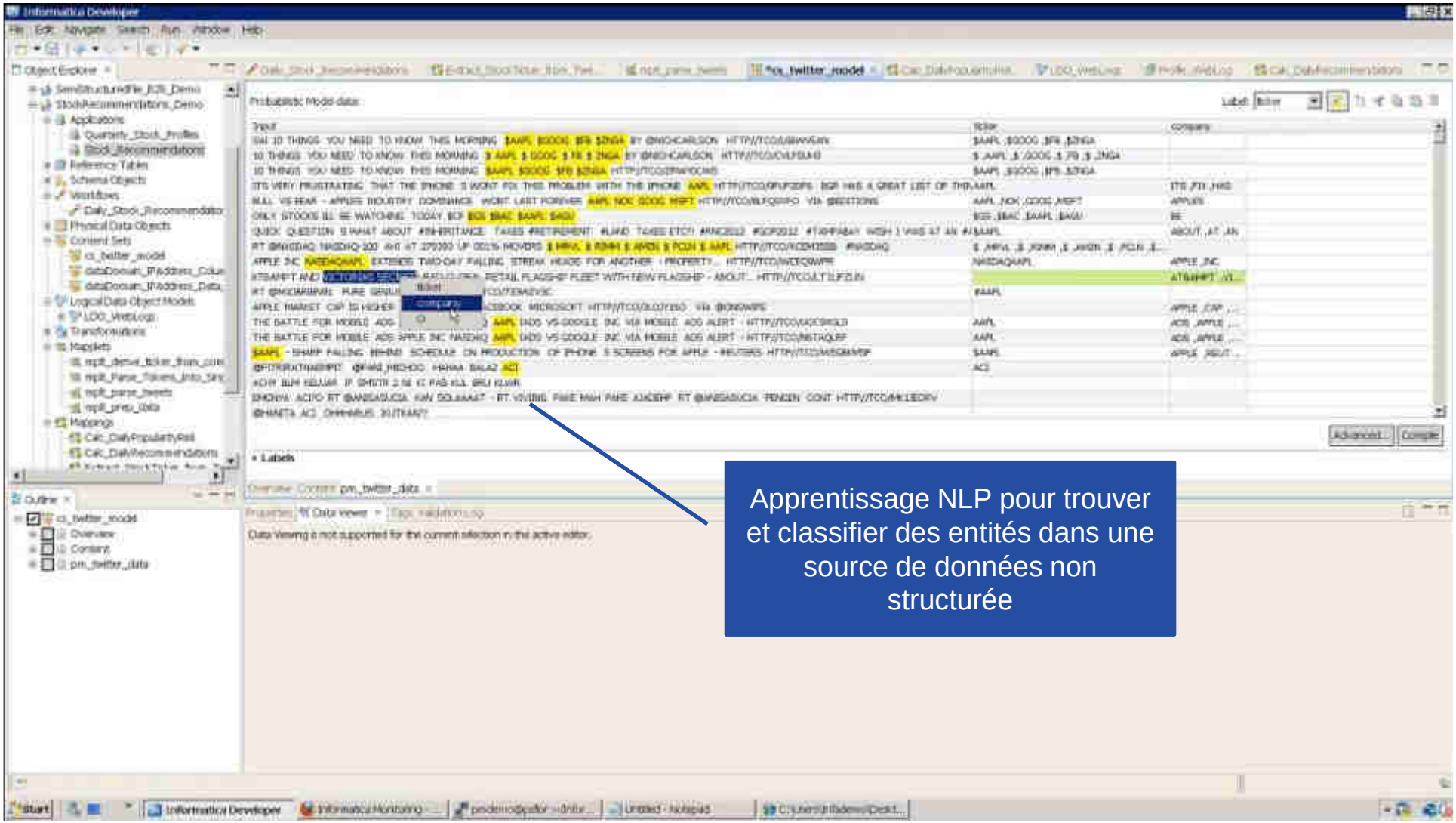
Description / Unstructured Text
She's very upset. Need to send something to her ASAP: 'BROADCASTING HOUSE ATTN: HILARY THOMAS, ROOM G12 ,ACCOUNT SERVICES ENGINEERING'

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

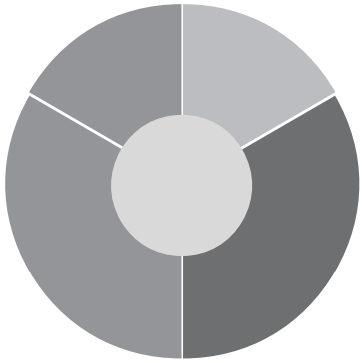


informatica

Archivage de données

Data Archive HDFS

Gouvernance des Données



Data Warehouse
Production Data

Optimized File Archive
Stored on Hadoop File System

Archiver les données dans un format optimisé pour réduire le stockage

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

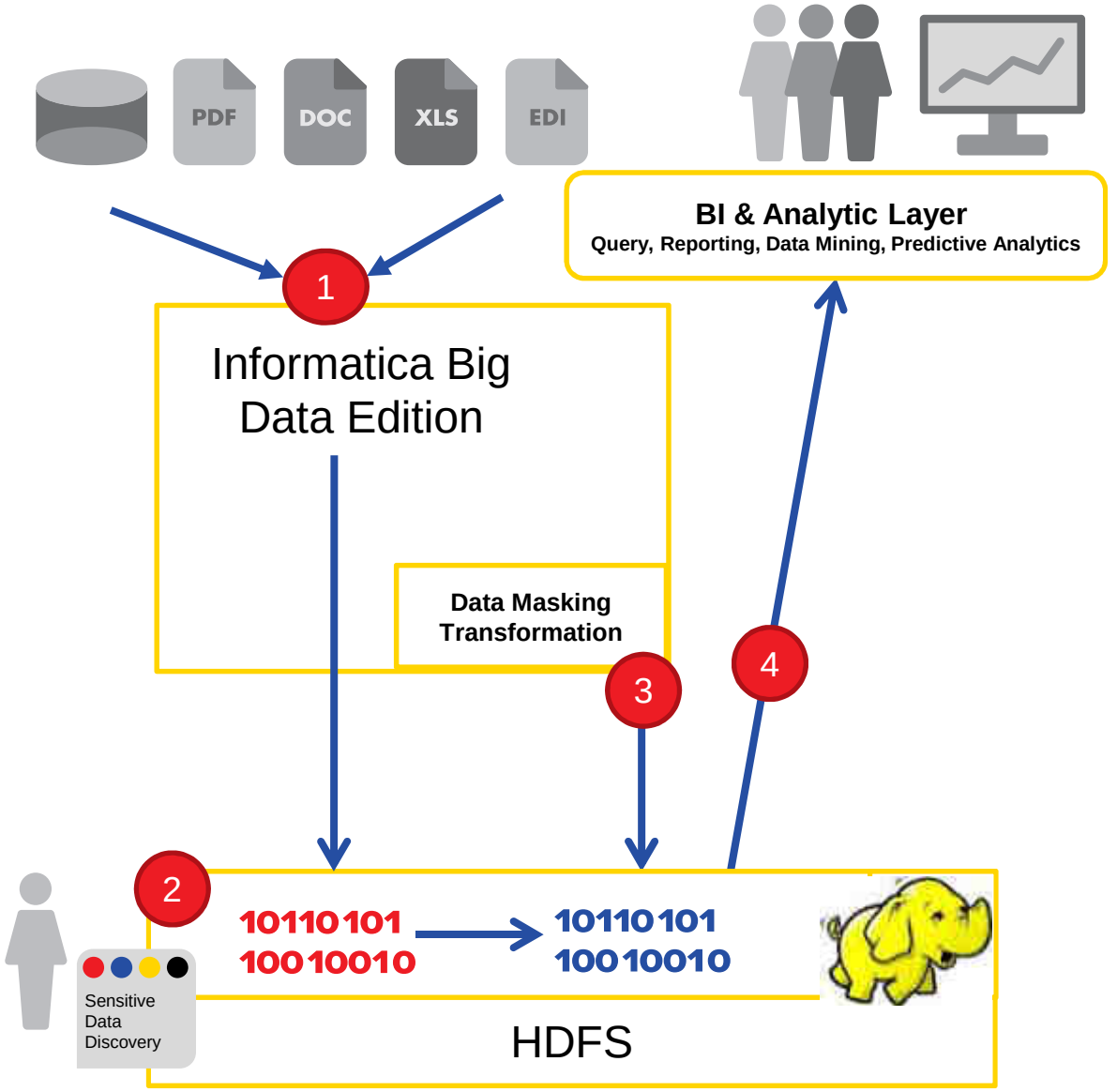
informatica

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**



informatica

Anonymisation dynamique

Dynamic Data Masking - Hive

Gouvernance des Données

anonymisation basée sur le **rôle utilisateur**

&

prévention en **temps-réel**

Values Presented:
BLAKE
JONES
KING

Business user application screen

Select name from table1

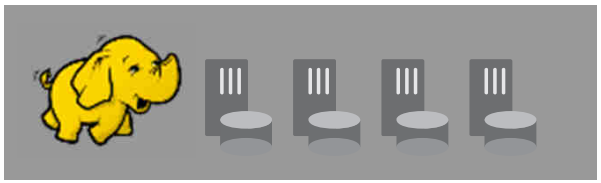
Values Presented:
BL****
JO****
KI****

Application screens and tools used by production support, DBAs, Outsourced or **unauthorized workforce**

Le module Dynamic Data Masking réécrit les requêtes HQL en temps-réel pour anonymiser les données retournées

(2)Select **substring(name,1,2)||'***'** from table1

Private Information Stored
BLAKE
JONES
KING



informatica

Profilage des données Hadoop

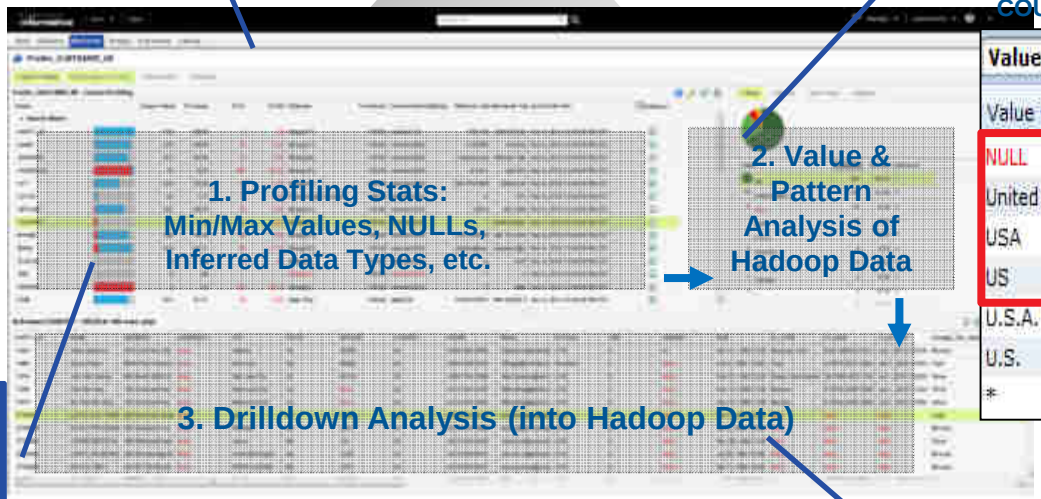
Gouvernance des Données

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
	10110091
	10110092
	10110122
	10110124
Top (5)	A5B334
	A44563
	A23456
	19134136
	19134134



COUNTRY CODE example

Values | Patterns | Statistics

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

Statistiques pour identifier les anomalies dans les données

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	MCcGREGOR GRP	1740 BRDWY	NY	NY	10019	US
10178890	GORDY SPAROW	UNKNOWN	BROADWAY	NEW YORK	NY	10019	USA

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

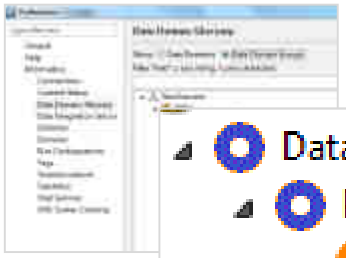
informatica

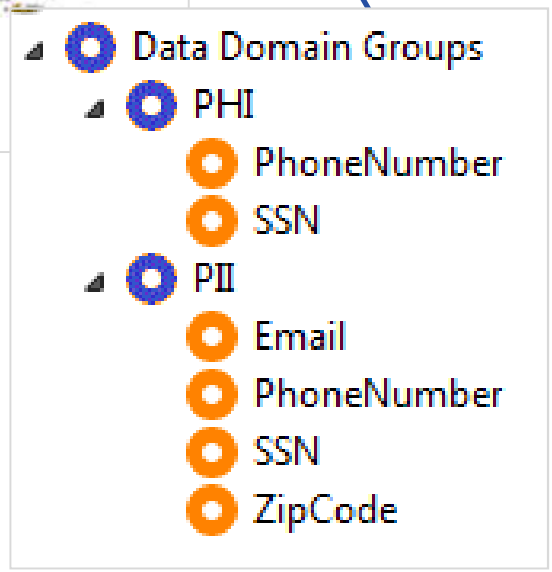
Recherche de Patterns dans Hadoop

Sensitive Domain Discovery

Gouvernance des Données

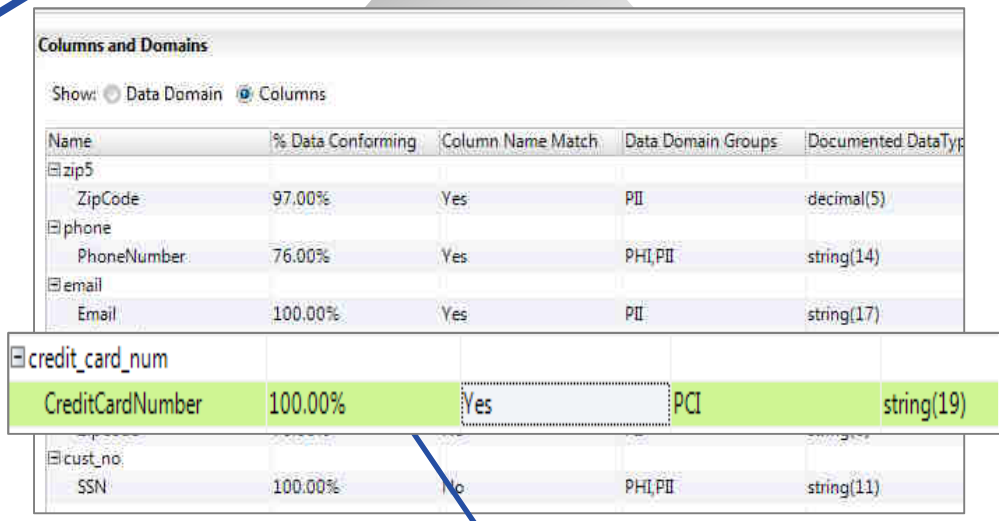
Utilisation de règles / mapplets Informatica pour identifier la signification des données Hadoop
Extraction des données sensibles
(Carte de crédit, Sécurité Sociale, etc.)





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

Discovery logic automatically pushed down and run on Hadoop

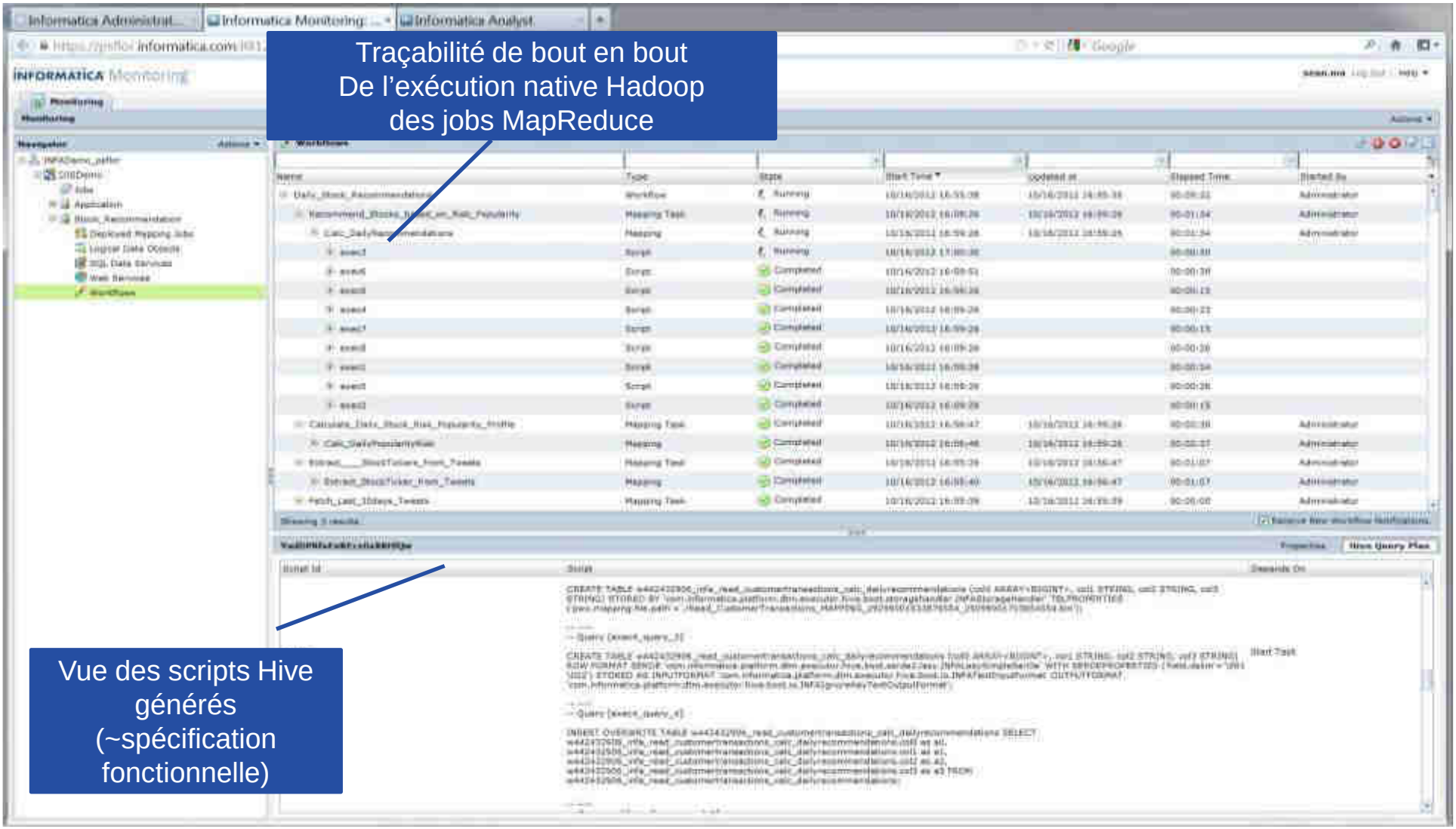


Vision/partage des rapport s sur les domaines/données sensibles contenues dans Hadoop. Capacité de 'drill down' pour voir les données suspectes.

informatica

Administration, Monitoring unifiés

Gouvernance des Données

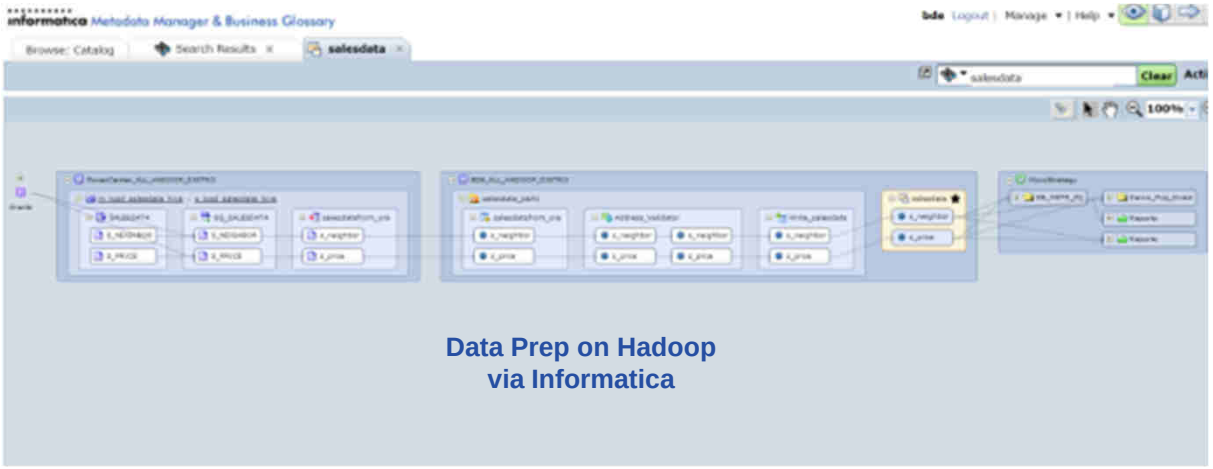


informatica

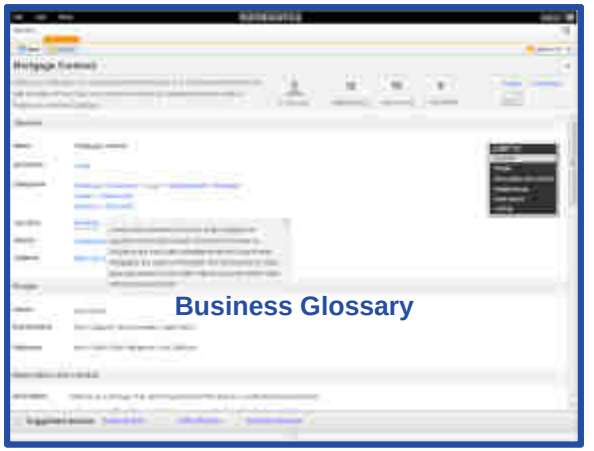
End-to-End Metadata Lineage

Vue de bout en bout des flux de données

Gouvernance des Données

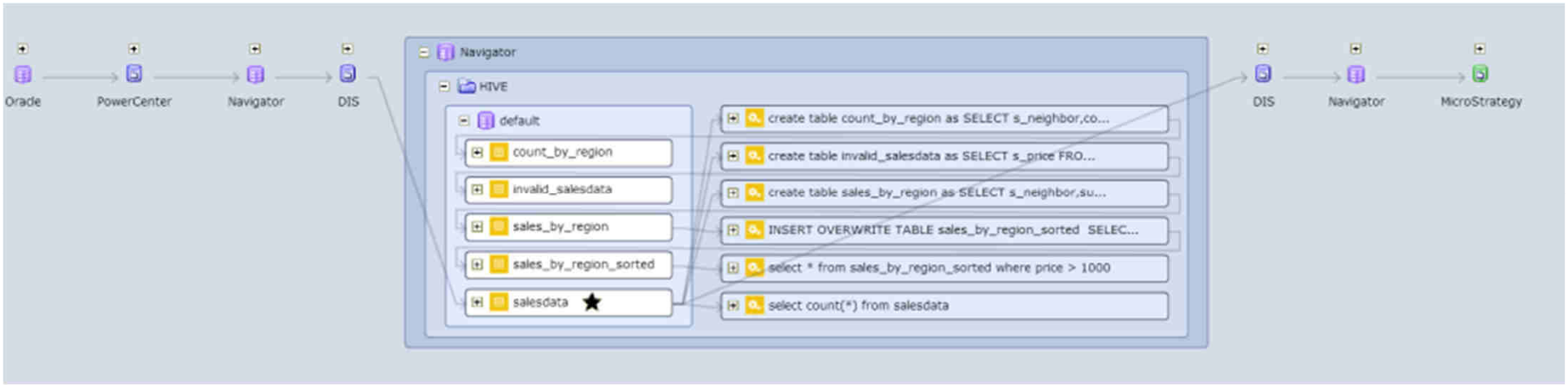


Data Prep on Hadoop via Informatica



Business Glossary

Data Source **Hive HQL** **Target BI/Analytic App**



informatica

Les Solutions Informatica Big Data:

Une architecture simple, vos projets accélérés

Collecte +Rapide

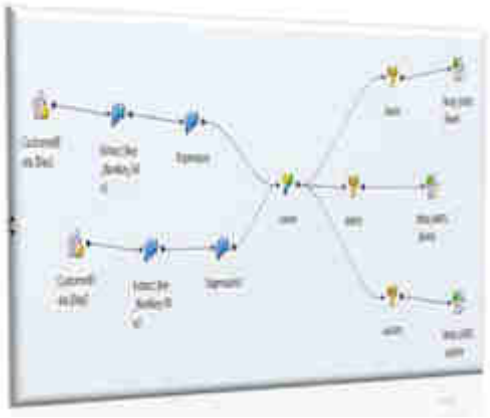
- 200+ connecteurs certifiés & performants
- 100+ parseurs pré-packagés
- Complex Data Parsing
- Ingestion multiple (batch, CDC)
- Data Streaming

Intégration +Simple

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

Gouvernance des donnés

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



informatica

Les Solutions Informatica Big Data:

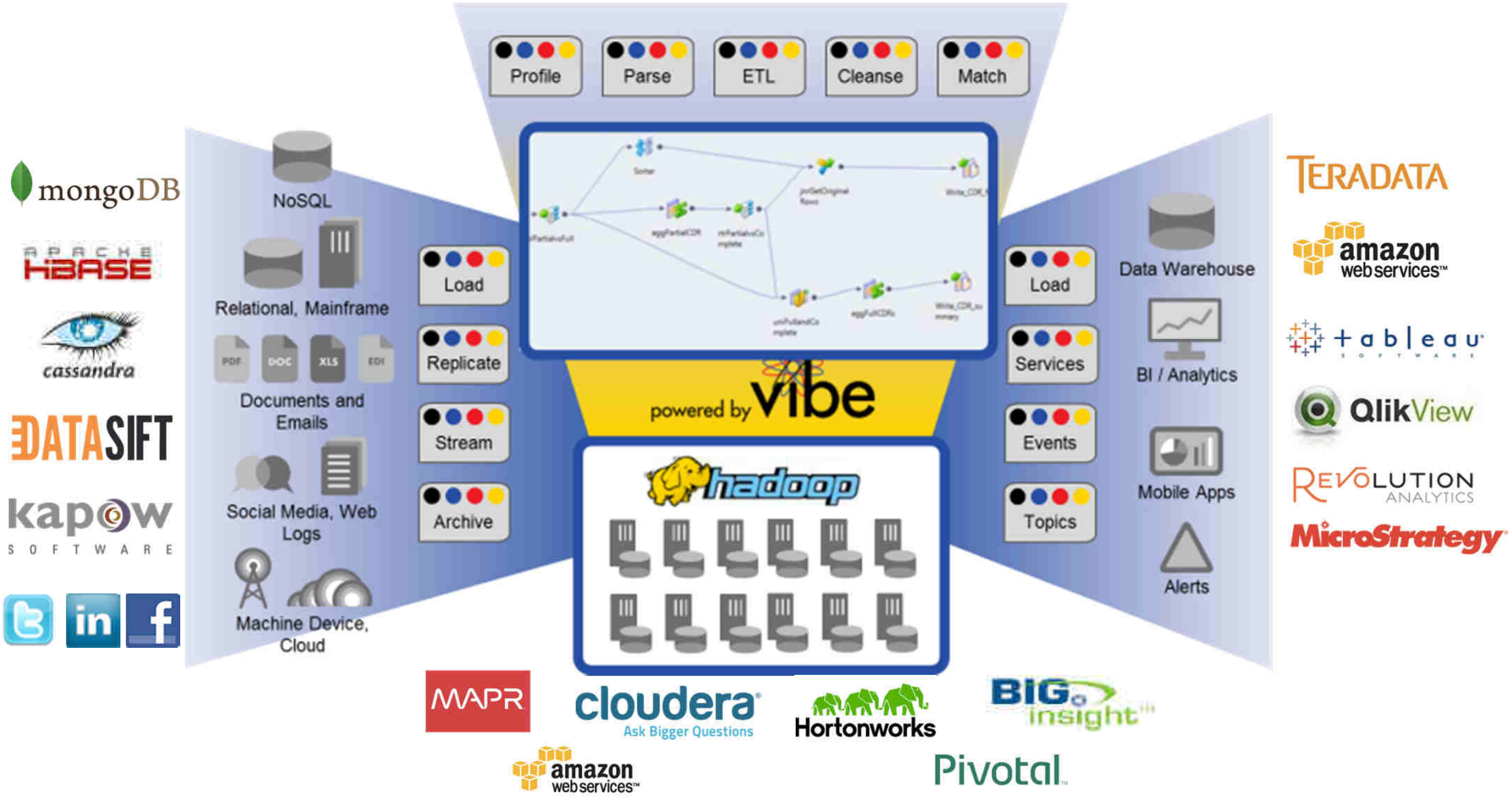
Une architecture simple, vos projets accélérés

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**"	

informatica

Ecosystème Informatica & Big Data

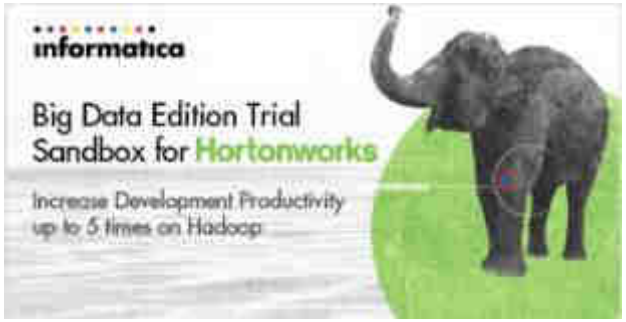
Tirez le meilleur parti de l'open source, sans codage

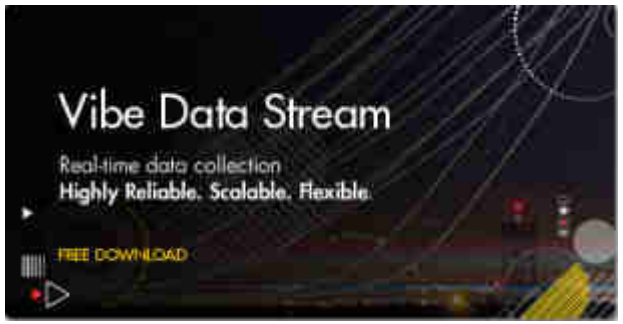


informatica

Essayez Informatica Big Data Edition

marketplace.informatica.com/bigdata





TRIAL DOWNLOADS

REFERENCE ARCHITECTURES

TRAINING & WEBINARS

informatica

Informatica : une technologie éprouvée

Data Virtualization

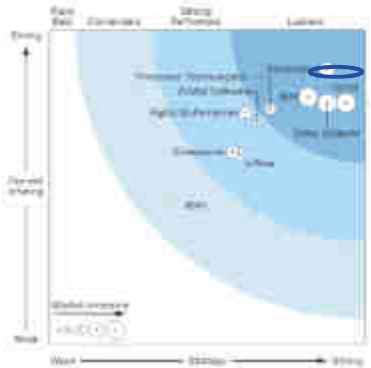
Enterprise ETL

Cloud Data Integration

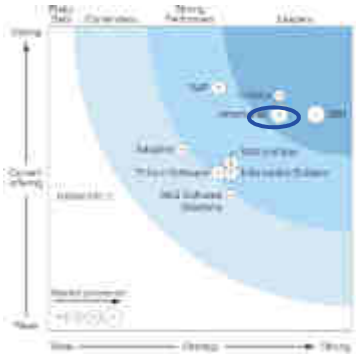
Master Data Management

Informatica Platform

Product Information Management



Data Governance Tools



Big Data Streaming Analytics Platforms

informatica