Mise en œuvre d'une stratégie Big Data

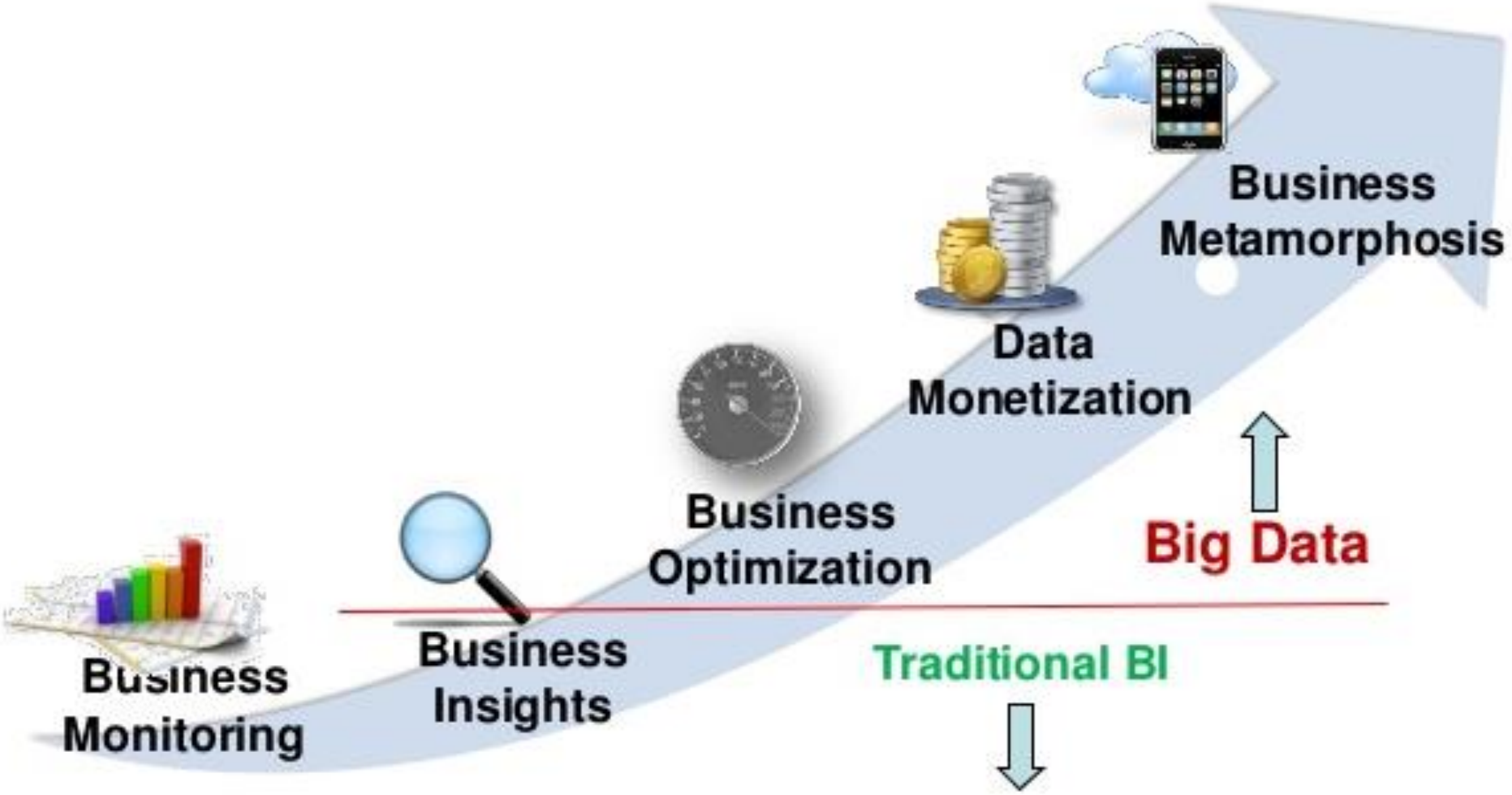
Les fondamentaux

Sommaire

- Stratégies dédiées au Big Data
- Evaluation des outils dédiés au Big Data
- Méthode analytique innovante
- Analyse statistique du Big Data

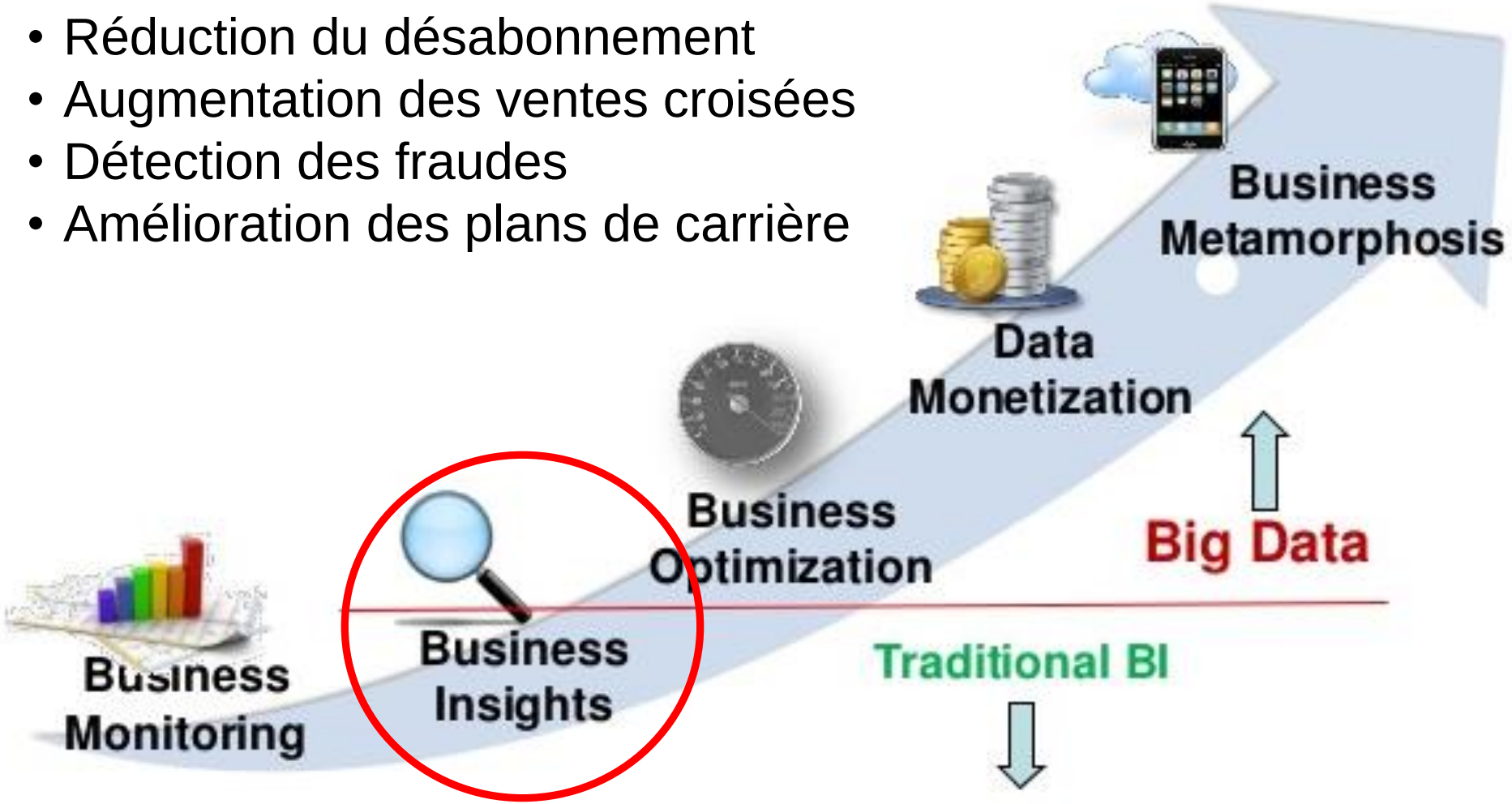
Stratégies dédiées au Big Data

Modèle de maturité en Big Data



Modèle de maturité en Big Data

- Amélioration de la relation clients
- Réduction du désabonnement
- Augmentation des ventes croisées
- Détection des fraudes
- Amélioration des plans de carrière



Comprendre et cibler les clients

Démarche:
Enrichir les données traditionnelles avec des données

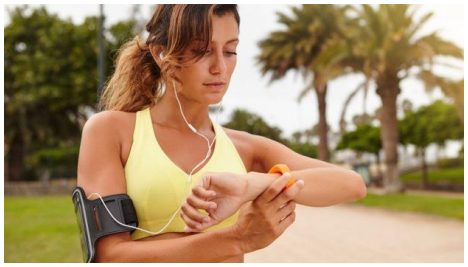
- de réseaux sociaux;
- des logs système
- des données de capteurs

pour mieux comprendre les clients, leurs comportements et leurs préférences

Objectif:
créer des modèles prédictifs pour:

- prédire le taux de désabonnement de la clientèle (ex: telco)
- prédire quels produits se vendent (ex: distributeurs),
- comprendre comment conduisent effectivement les gens (ex: assureurs).

Comprendre et améliorer nos performances



Démarche:
Bénéficier des données générées à partir d'appareils portables tels que les montres intelligentes ou bracelets à puce pour recueillir:

- des données sur notre consommation de calories,
- les niveaux d'activité,
- nos habitudes de sommeil

Objectif:
L'analyse de ces données agrégées collectivement apporte entièrement de nouvelles perspectives qui sont restituées à des utilisateurs individuels

Améliorer la santé

Démarche:
Enregistrer et analyser chaque battement cardiaque et la respiration modèle de chaque bébé d'une unité de bébés prématurés

Objectif:
L'unité a été en mesure de développer des algorithmes qui peuvent maintenant prédire les infections 24 heures avant que les symptômes physiques apparaissent.
De cette façon, l'équipe peut intervenir rapidement et sauver des bébés fragiles dans un environnement où chaque heure compte.

Modèle de maturité en Big Data

- Réduction des dépenses IT
- Optimisation des couts logistiques
- Optimisation de la maintenance
- Optimisation des pages webs
- Optimisation des centres d'appels
- Réduction des dépenses d'énergie

Comprendre et optimiser les itinéraires

Démarche:
Enrichir les données avec des informations provenant de capteurs de positionnement et d'identification par radiofréquence géographique pour suivre les marchandises ou les véhicules de livraison

Objectif:
optimiser les itinéraires en intégrant les données de trafic en temps réel

Maintenance predictive d'engins

Démarche:

Collecter des sources multiples de données provenant de capteurs

Analyser les données pour identifier les problèmes potentiels

Utiliser les outils de machine learning pour confirmer les problèmes et apporter des solutions

Objectifs:

Augmenter la disponibilité des engins

Espacer davantage dans le temps les révisions

Réduire les couts de maintenance

Optimiser les réseaux d'energie

Démarche:
Collecter les données de consommation des compteurs intelligents et enrichir avec les données météo

Objectif:
Optimiser la production et la distribution d'énergie

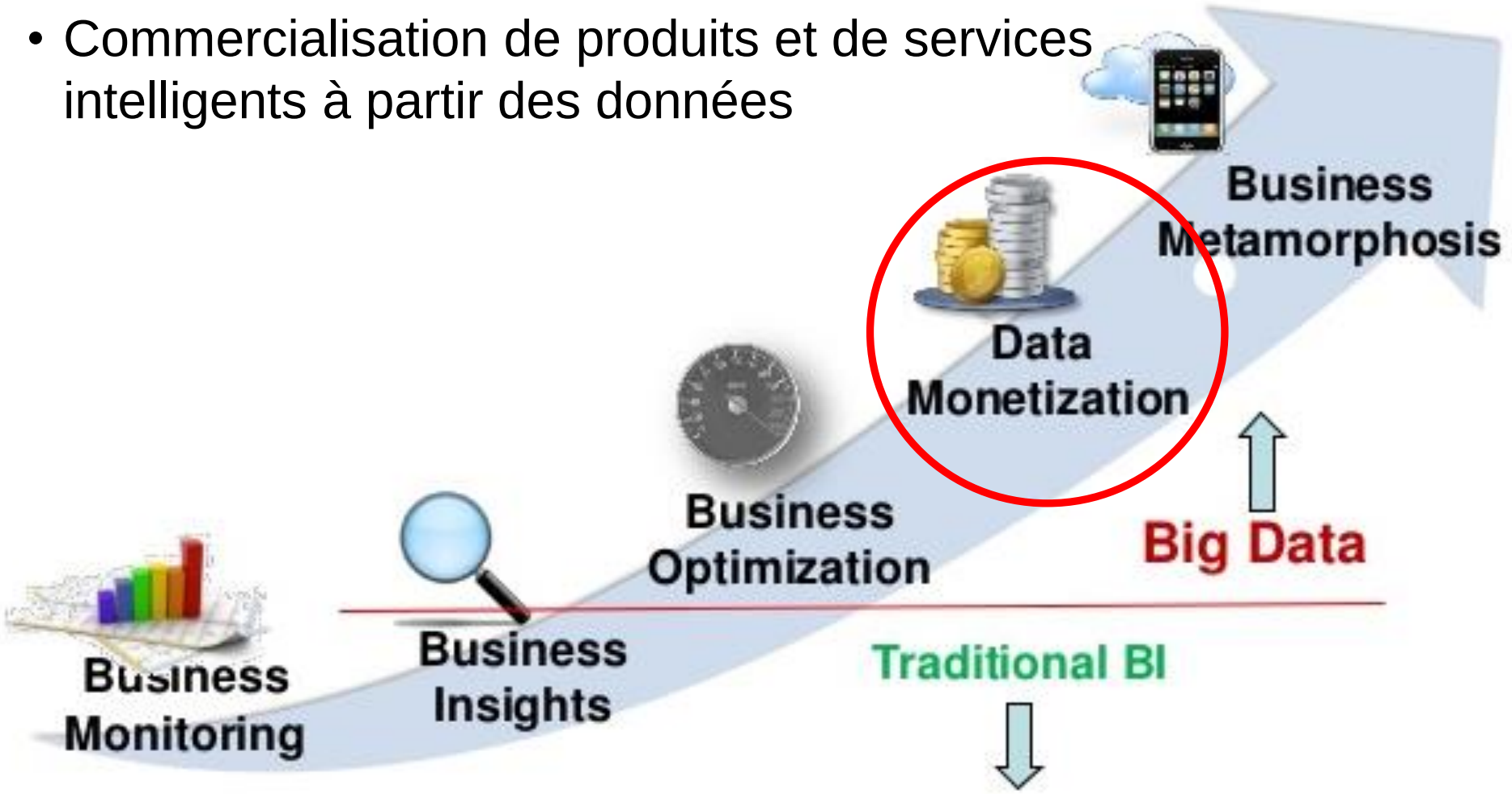
Voiture autonome

Démarche:
Collecter des images des caméras embarquées et les données GPS en temps réel

Objectif:
La Toyota Prius peut circuler en toute sécurité sur la route sans l'intervention d'êtres humains

Modèle de maturité en Big Data

- Commercialisation des données
- Commercialisation de produits et de services intelligents à partir des données



Monétisation des données

- Amélioration des processus internes
  - Performance marketing et commerciale
  - Performance opérationnelle et financière

- Générer un revenu
  - Commercialisation des données brutes, enrichies ou indicateurs
  - Commercialisation de produits et de services intelligents

Type de données et utilisation

Brute	Enrichie	Indicateur	Prédiction	Service
Données brutes de logs, capteurs, formulaires	Données validées, nettoyées, complétées	Données agrégées, calculées	Données prédites, probabilités	Données « intelligentes »
Audit, détection d'incidents	Analyse détaillée, gestion, reporting	KPI, score, alerte, dashboard, infographie	Anticipation, recommandation, alerte préventive	Processus adaptatif, règles de gestion adaptative

Médiamétrie

Démarche:
Connaître à chaque instant ce qui est visionné sur le téléviseur en rapprochant les données "log" que nous communique Canal Satellite en temps réel avec nos données panel TV

Objectif:
Mieux cerner l'usage des chaînes dont l'audience est de plus en plus fragmentée.

Météo France

Démarche:
Les données météo représentent tout simplement la plus grande plate-forme Big Data existante.
Elle deviendra encore plus grande lorsqu'elle sera enrichie de toutes les données hyper locales issues des mobiles des gens ou des thermostats connectés

Objectif:
Améliorer la qualité des services payants (Bulletins départementaux à sept jours , Certificats en cas d'intempéries Viginet, information par mobile)

Bloomberg

Démarche:
Introduction du web sémantique pour proposer à ses clients un service EDM (Enterprise Data Management)

Objectif:
Améliorer la qualité de l'information économique et financière

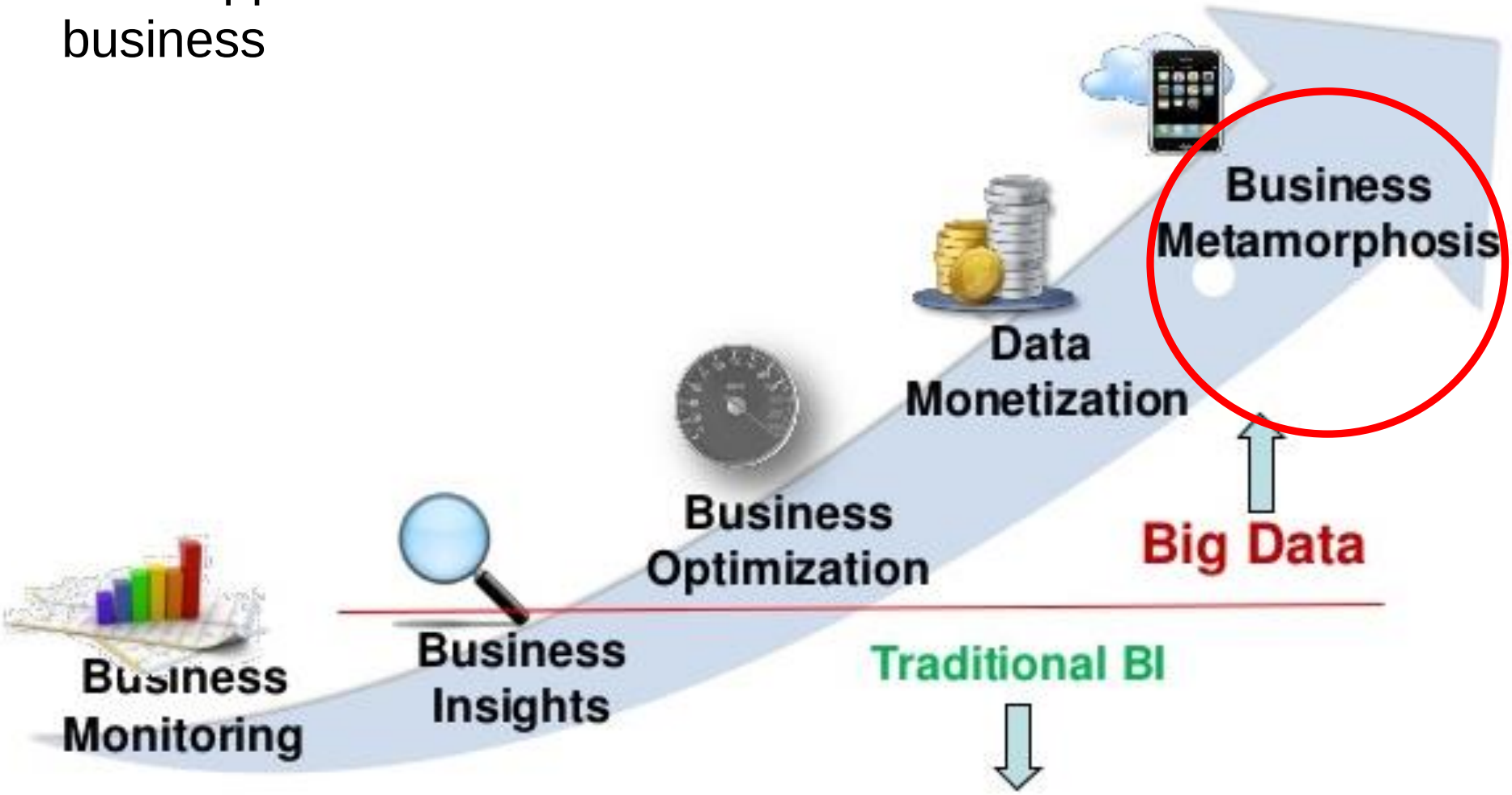
1000mercis

Démarche:
Collecte de centaines de millions d'impression et de cookies en temps réel puis utilisation d'un algorithme de prédictions des conversions

Objectif:
Abaisser de 30% le cout d'acquisition des campagnes Sofinco

Modèle de maturité en Big Data

- Développement de nouveaux business



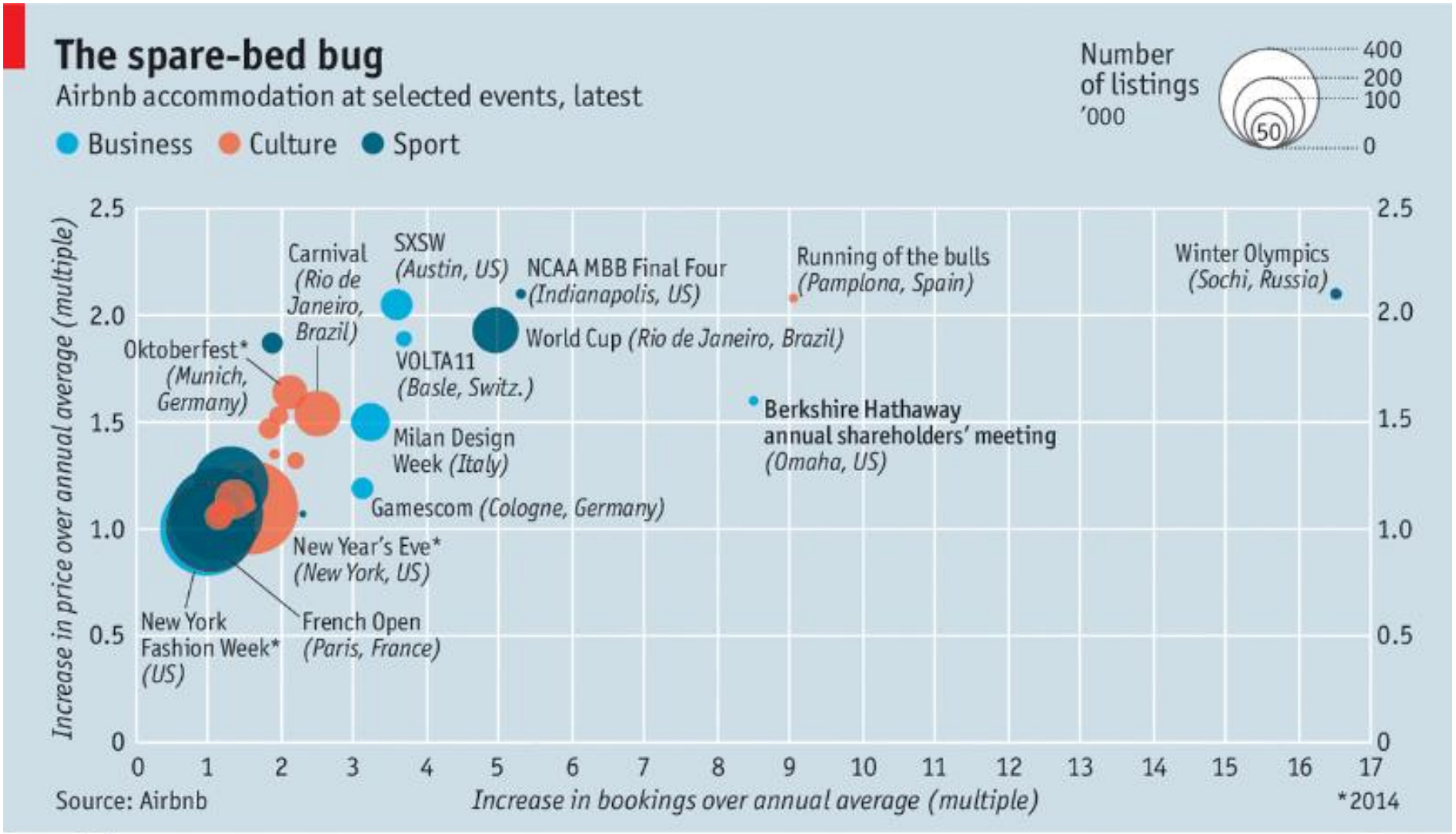
Business Metamorphosis (Energie)

- Les fournisseurs d'énergie peuvent proposer des services d'optimisation de l'énergie pour le particulier tels que:
  - les marques d'appareils à acheter,
  - La météo
  - Les conditions de l'eau locale
  - Les couts d'énergie

Business Metamorphosis (Distributeurs)

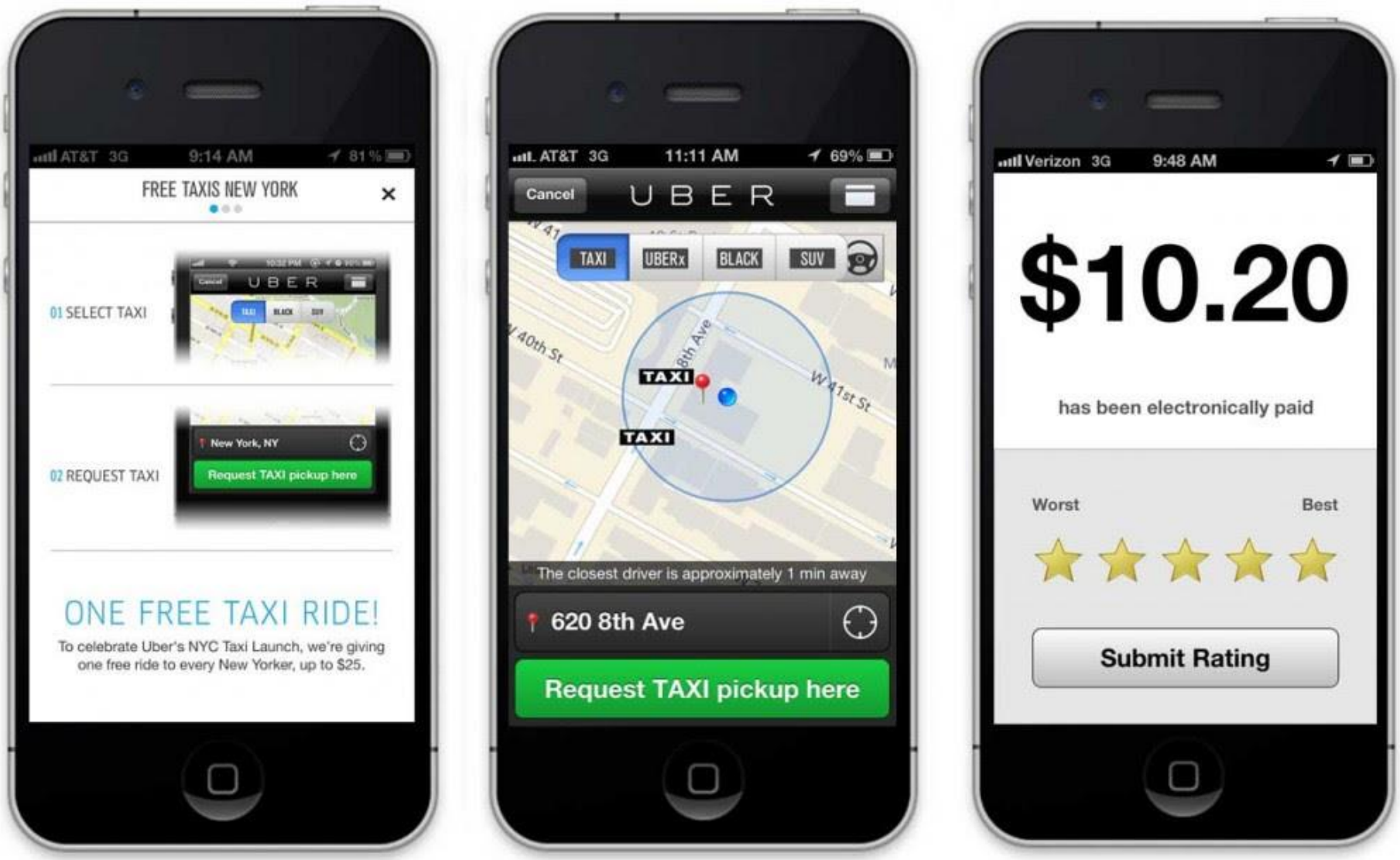
- Les distributeurs peuvent proposer des services d'optimisation des achats tels que:
  - les produits convenant à vos attentes
  - Les produits des concurrents

Airbnb – louer au moment d'évènements



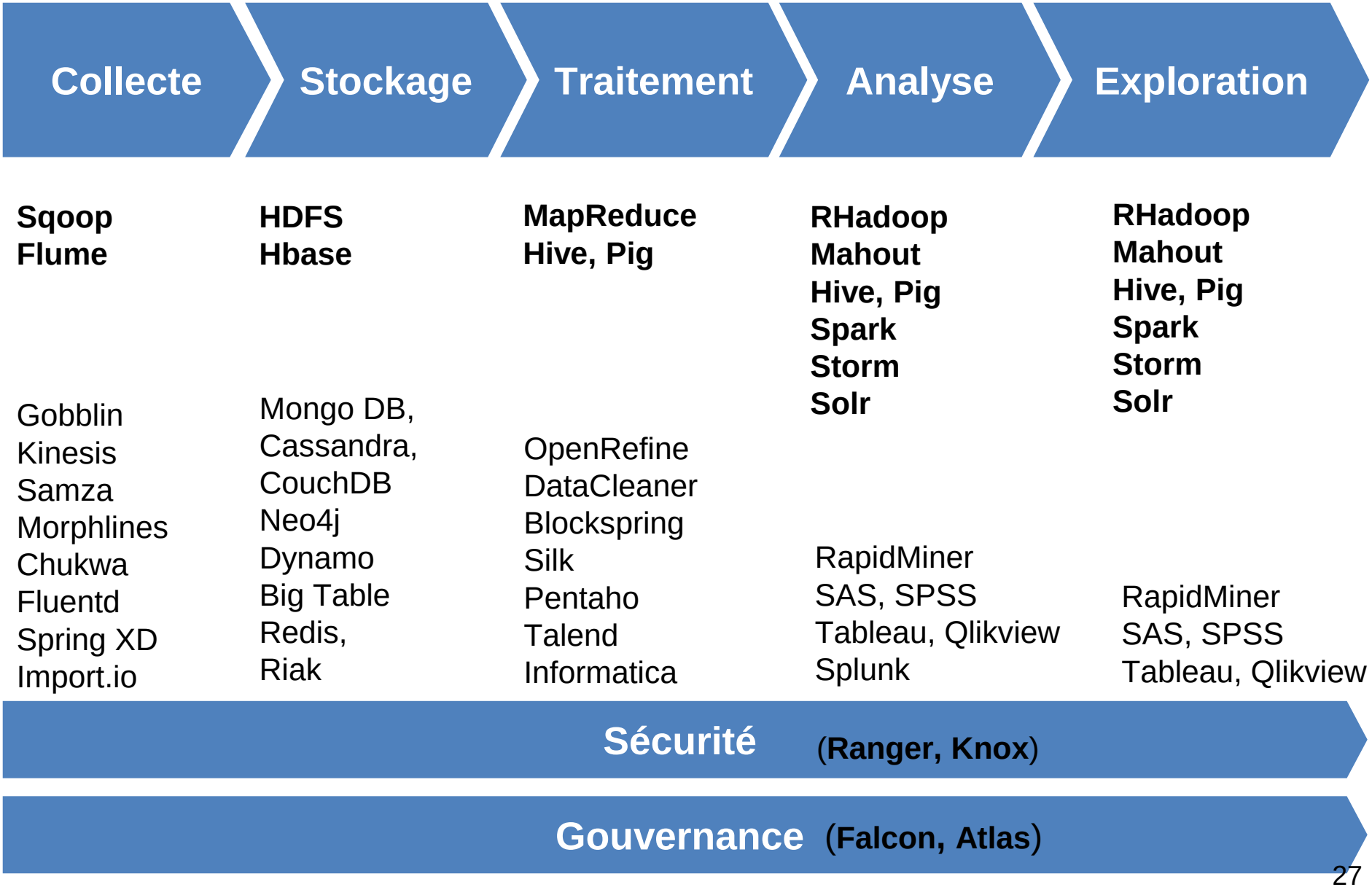
Economist.com

Uber – Partage d'informations de transport



Evaluation des outils dédiés au Big Data

Cycle de vie des Big Data



Cycle de vie des Big Data avec Hadoop

Collecte

- Sqoop (transfer des données de bases relationnelles vers HDFS)
- Flume (transfert des données de logs vers HDFS)
- Gobblin (outil universel d'extraction, de transformation et de chargement)
- Amazon Kinesis (outil cloud pour collecter des clics de sites web, des transactions financières, messages, etc…)
- Samza (real time)

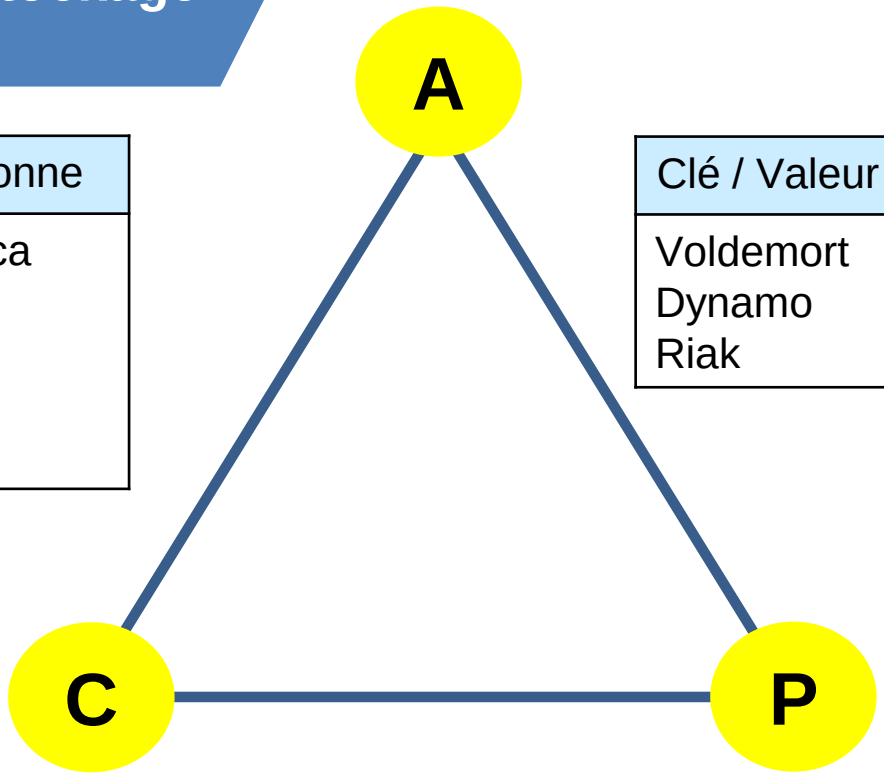
Cycle de vie des Big Data avec Hadoop

Collecte

- Cloudera Morphlines (extrait, transforme, charge les données vers Solr, HDFS, Hbase)
- Chukwa (collecte et analyse des données de logs)
- Fluentd (outil universel d'extraction , de transformation et de chargement)
- Spring XD (collecte et transfert en mode batch ou temps réel des données vers Hadoop)
- Import.io (numéro 1 extraction des données du web)

Cycle de vie des Big Data avec Hadoop

Stockage



Relationnel	Colonne
MySQL Postgres Aster Data Greenplum Neo4j	Vertica

Clé / Valeur	Colonne	Document
Voldemort Dynamo Riak	Cassandra	CouchDB SimpleDB

Clé / Valeur	Colonne	Document
Redis MemcacheDB	Hbase Hypertable BigTable	MongoDB Terrastore

Cycle de vie des Big Data avec Hadoop

Data Cleansing

- Open Refine
- DataCleaner
- Hive, Pig

Data Integration

- Blockspring (connect to AWS, Import.io, Tableau,)
- Silk (intégration et transformation de données)
- Pentaho (big data integration & business analytics)
- Talend (big data integration, data management)
- Informatica (big data integration)

Cycle de vie des Big Data avec Hadoop

- RHadoop
- Mahout
- Hive
- Pig
- Spark
- Storm
- Solr

Cycle de vie des Big Data avec Hadoop

- RapidMiner (predictive analysis)
- SAS
- IBM SPSS Modeler
- Tableau
- Qlikview
- Splunk

Cycle de vie des Big Data avec Hadoop

- RHadoop
- Mahout
- Hive
- Pig
- Spark
- Storm
- Solr

Cycle de vie des Big Data avec Hadoop

- RapidMiner (predictive analysis)
- SAS
- IBM SPSS Modeler
- Tableau
- Qlikview
- Splunk

Data languages

- R
- Python
- Scala
- RegEx (Regular Expressions peuvent extraire, manipuler et changer les données)
- XPath (query language pour sélectionner certains nœuds d'un document XML).

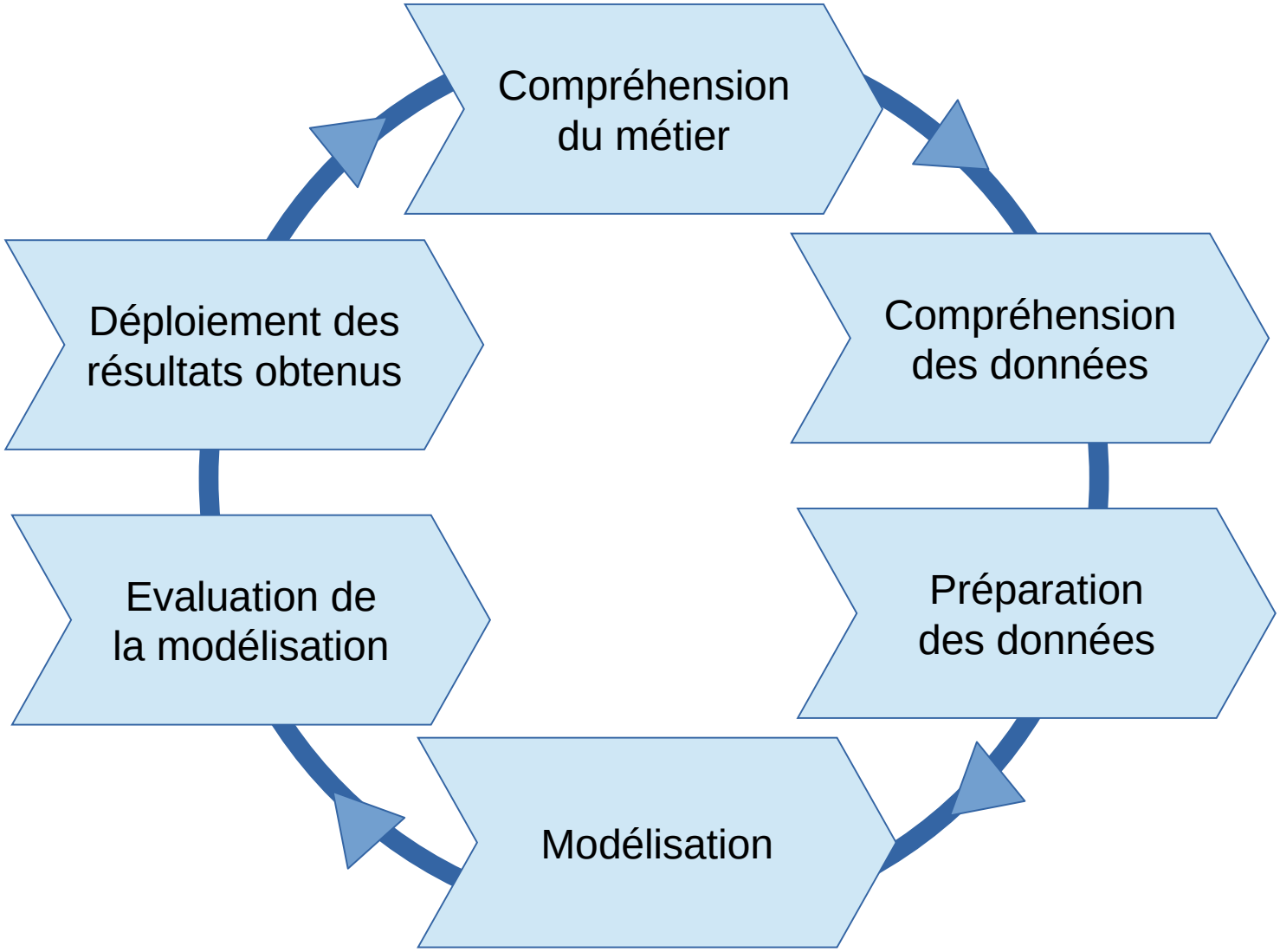
Une methode analytique innovante

Utilisation du Data Mining

Ensemble des techniques et méthodes :

- destinées à l'exploration et l'analyse
- traitant de grandes bases de données
- en vue de détecter dans ces données des règles

Les étapes de Data Mining



Compréhension du métier

- Définir les objectifs du projet et les contraintes
- Traduire en un problème de data mining
- Préparer une stratégie initiale

Compréhension des données

- Recueillir les données
- Utiliser l'analyse exploratoire
- Évaluer la qualité des données
- Sélectionner des sous-ensembles intéressants

Préparation des données

- Préparer l'ensemble final des données qui va être utilisé
- Sélectionner les cas et les variables à analyser
- Transformer si nécessaire certaines données
- Supprimer si nécessaire certaines données

Modélisation

- Sélectionner les techniques de modélisation
- Calibrer les paramètres des techniques de modélisation
- Revoir si nécessaire la préparation des données

Evaluation de la modélisation

- Evaluer la qualité des résultats obtenus
- Déterminer si les résultats atteignent les objectifs fixés
- Décider si on passe au déploiement

Déploiement des résultats obtenus

- Prendre les décisions en conséquence des résultats
- Préparer la collecte des informations futures

Les 2 types de technique de Data Mining

- Analyse descriptive ou Apprentissage non supervisé
- Analyse prédictive ou Apprentissage supervisé

Analyse descriptive (1/2)

- Vise à mettre en évidence des informations présentes mais cachées par le volume des données
  - Segmentation de clientèle,
  - Recherche d'associations de produits sur les tickets de caisse.

- But: Réduire, résumer et synthétiser les données,

- Pas de variable «cible» à prédire.

Analyse descriptive (2/2)

- Visualisation

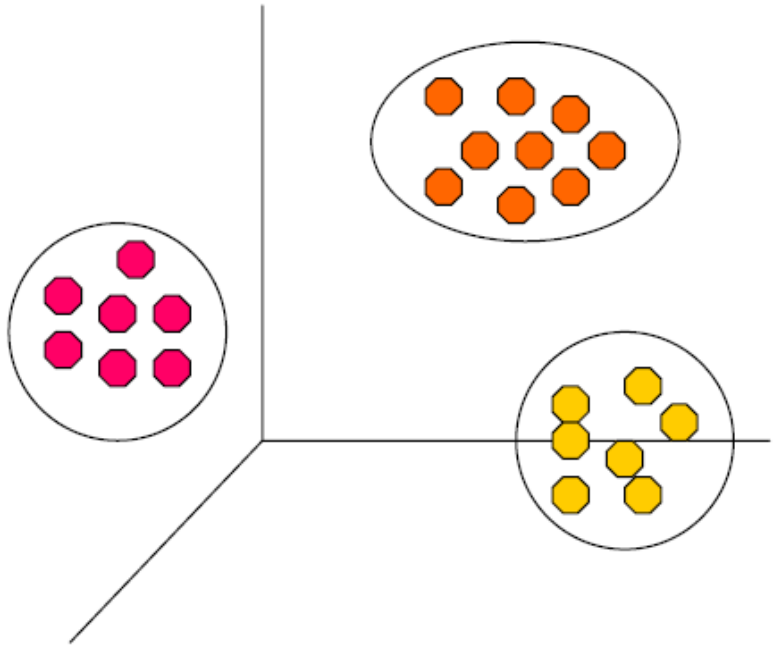
- Analyse en Composantes Principales (ACP), analyse factorielle et des correspondances

- Classification non supervisée (clustering).

Objectif du clustering

Minimiser les distances intra-cluster

Maximiser les distances inter-clusters



Cas d'utilisation du Clustering

Marketing
segmentation du marché en découvrant des groupes de clients distincts à partir de bases de données d'achats

Assurance
identification de groupes d'assurés distincts associés à un nombre important de déclarations

Analyse prédictive (1/2)

- Explique les données par rapport à une connaissance
  - âge des individus,
  - catégorie socio-professionnelle,
  - niveau de formation.

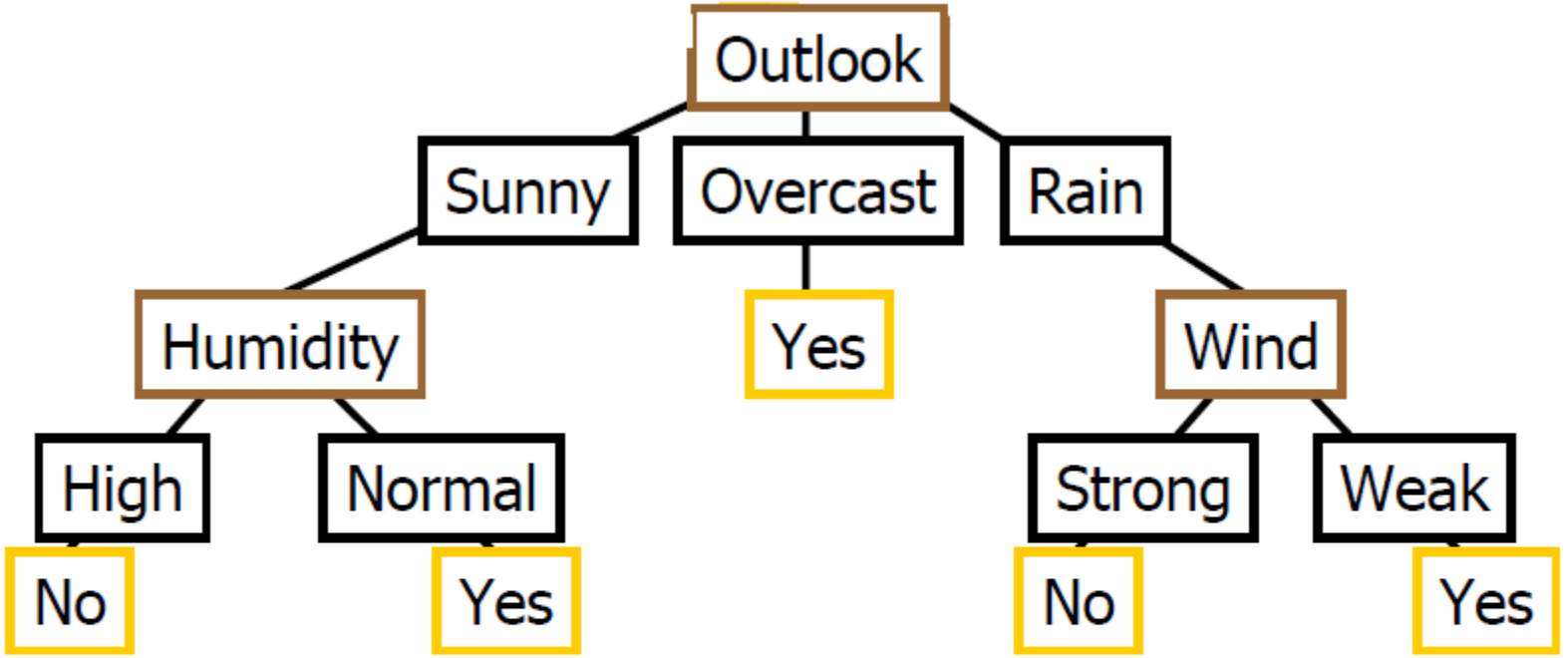
- Permettre de prendre des décisions lors de l'arrivée de nouvelles données,

- Une variable «cible» à prédire.

Analyse prédictive (2/2)

- Arbre de décision (Decision trees)
- Foret aléatoire (Random forests)
- Régression logistique
- boosting,
- support vector machines,
- linear

Des arbres de décision aux règles



R_1: If (Outlook=Sunny) $\wedge$ (Humidity=High) Then PlayTennis=No
R_2: If (Outlook=Sunny) $\wedge$ (Humidity=Normal) Then PlayTennis=Yes
R_3: If (Outlook=Overcast) Then PlayTennis=Yes
R_4: If (Outlook=Rain) $\wedge$ (Wind=Strong) Then PlayTennis=No
R_5: If (Outlook=Rain) $\wedge$ (Wind=Weak) Then PlayTennis=Yes

Cas d’utilisation de l’analyse prédictive

- Détection de fraude
- Marketing teléphonique

Analyse statistique du Big Data avec RHadoop

Sommaire pour RHadoop

- Introduction de RHadoop
- Package rhdfs
- Package rhbase
- Package plyrmr
- Package rmr2
- Package ravro

Cycle de vie des Big Data avec Hadoop

Collecte	Stockage	Traitement	Analyse	Exploration
Sqoop Flume	HDFS Hbase	MapReduce Hive, Pig	**RHadoop** Mahout Hive, Pig Spark Storm Solr	RHadoop Mahout Hive, Pig Spark Storm Solr

Analyse statistique du Big Data

+

Une solution possible

Définition de R

- Logiciel permettant de faire des analyses statistiques et de produire des graphiques
- Langage de programmation complet
- Pour des ensembles réduits de données

R n’est par fait pour

- Programmer Mapper
- Programmer Reducer
- Manipuler de gros volumes de données

Pourquoi Hadoop avec R ?

- Développement plus facile et plus productif de Map Reduce
- Réduction de 50 % de lignes de code par rapport à Java
- Transition facilitée au Big Data pour les analystes R

Introduction à RHadoop

- Intègre le langage R et Hadoop
- Déplace l'algorithme d'exécution vers les données
- Permet d'accéder à des librairies statistiques
- Accélère le traitement en l'exécutant en parallèle

Composition de Rhadoop

Une série de 5 packages R principaux permettant de gérer et d'analyser des données avec Hadoop.

Package	Description
rhdfs	Interaction avec HDFS de Hadoop
rhbase	Accès à NoSQL Hbase
plyrmr	Manipuler de gros volumes de données
rmr2	Fonctions liées à MapReduce
ravro	Lire et écrire des fichiers avro

rhdfs

- Connexion avec HFDS

- Développeurs en R peuvent afficher, lire, écrire et modifier des fichiers stockés dans HDFS à partir de R

- Installer ce package que sur le nœud où s'exécute le client R

Package Rhdfs: Initialization, Utility & Directory

- Initialization
  - hdfs.init(),
  - hdfs.defaults()
- Utility
  - hdfs.ls(),
  - hdfs.list.files(),
  - hdfs.file.info(),
  - hdfs.exists()
- Directory
  - hdfs.dircreate(),
  - hdfs.mkdir()

Package Rhdfs: File manipulation

- hdfs.copy(),
- hdfs.move(),
- hdfs.rename(),
- hdfs.delete(),
- hdfs.rm(),
- hdfs.del(),
- hdfs.chown(),
- hdfs.put(),
- hdfs.get()

Package Rhdfs: File Read / Write

- hdfs.file(),
- hdfs.write(),
- hdfs.close(),
- hdfs.flush(),
- hdfs.read(),
- hdfs.seek(),
- hdfs.tell(),
- hdfs.line.reader(),
- hdfs.read.text.file()

rhbase

- Connexion avec Hbase en utilisant le serveur Thrift

- Développeurs en R peuvent afficher, lire, écrire et modifier des tables stockés dans Hbase à partir de R

- Installer ce package que sur le nœud où s'exécute le client R

Package Rhbase: Initialization & Utility

- Initialization
  - hb.defaults()
  - hb.init()
- Utility
  - hb.list.tables()

Package Rhbase: Table Manipulation

- hb.new.table()
- hb.delete.table()
- hb.describe.table()
- hb.set.table.mode()
- hb.regions.table()

Package Rhbase: Read / Write

- hb.insert()
- hb.get()
- hb.delete()
- hb.insert.data.frame()
- hb.get.data.frame()
- hb.scan()
- hb.scan.ex()

plyrmr

- Manipuler de gros volumes de données stockés dans HDFS (comme avec plyr, reshape2)

- Installer ce package sur chaque nœud du cluster

Package plyrmr

- Data manipulation
  - bind.cols « add new columns »
  - select « select columns »
  - where « select rows »
  - transmute « all of the above plus summaries »
- From reshape2
  - melt and dcast « convert between long and wide data frame »
- Summary
  - count
  - quantile
  - sample
- Extract
  - top.k
  - bottom.k

rmr2

- Permet de lancer des fonctions MapReduce à partir de R

- Installer ce package sur chaque nœud du cluster

ravro

- Permet de lire et écrire des fichiers avro à partir de R

- Installer ce package que sur le nœud où s'exécute le client R

Conclusion avec Rhadoop

Utilisation de Rhadoop

- Exploration des données
- Exécution de taches en parallèle
- Trier des données
- Echantillonner des données
- Joindre des données

RHadoop

Ce qu'il faut retenir…

- 4 stratégies Big Data (Business Insight, Business Optimization, Data Monetization, Business Metamorphosis)
- Large choix d'outils Big Data pour chaque étape de la gestion du Big Data
- Rhadoop une bonne alternative d'analyse statistique du Big Data si vous possédez une large bibliothèque de scripts R

Merci