Cloud Data Management

Teodor Macicas

L'Evolution du Web

- Le Web 1.0:
  - Création de l'Internet.
  - 1989 (CERN) Tim Berners-Lee.
  - 1991 Diffusion dans le monde entier.
  - Navigation entre des pages statiques.
  - Pas d'interactions entre les utilisateurs.

Die Evolution des Webs

- Web 2.0
  - Pages dynamiques
    - ASP (Active Server Page)
    - PHP
      - Personal Home Page / Form Interpreter
      - PHP : Hypertext Preprocessor
  - Intéraction entre les utilisateurs
    - Chat
    - Forum
    - Messagerie

Die Evolution des Webs

- Web 2.0 (suite)
  - Web des personnes
  - Réseaux sociaux
    - Facebook
    - Twitter
    - LinkedIn
    - etc.

Die Evolution des Webs

- Web 3.0
  - Web des objets
    - RFID
    - Code barre
    - Réalité augmentée
  - Web Sémantique
    - Les programmes inter-agissent entre eux
    - RDF (Ressource Description Framework)
    - OWL (Ontologie)
    - etc.

Die Evolution des Webs

- Quelles sont les conséquences de ces évolutions?
  - Plus de facilités à créer du contenu Web:
    - "60 heures de vidéo sont mises en ligne chaque minute, soit une heure de vidéo mise en ligne sur YouTube chaque seconde." –Youtube Press
  - Changements technologiques rapides.
    - Une poussé vers la réutilisation de code et l'adoption de plateformes existantes.

… Pourquoi réinventer la roue?

Chaque Minute sur Internet

Technologie?
Le Cloud!

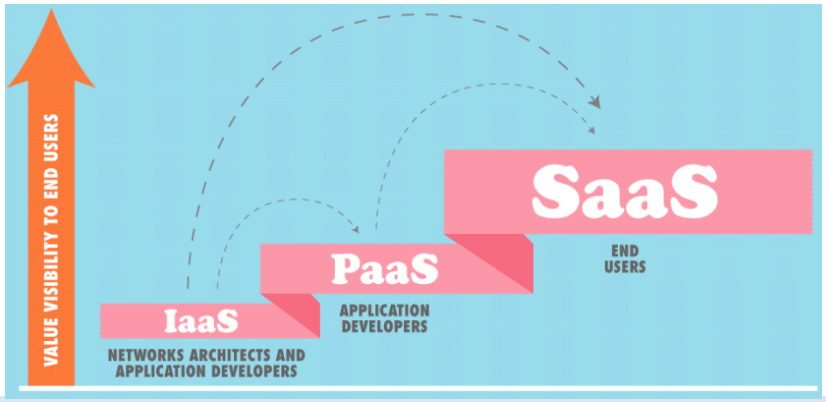
CLOUD DATA MANAGEMENT

Le Cloud, c'est quoi?

- Un concept qui pousse vers la dématérialisation des machines et des logiciels.
- Le calcul informatique se transforme en un service:
  - Payable par unité de temps, puissance et espace.
  - Gratuit, mais en contre partie les fournisseurs vous connaissent un peu plus.

Was ist die Cloud

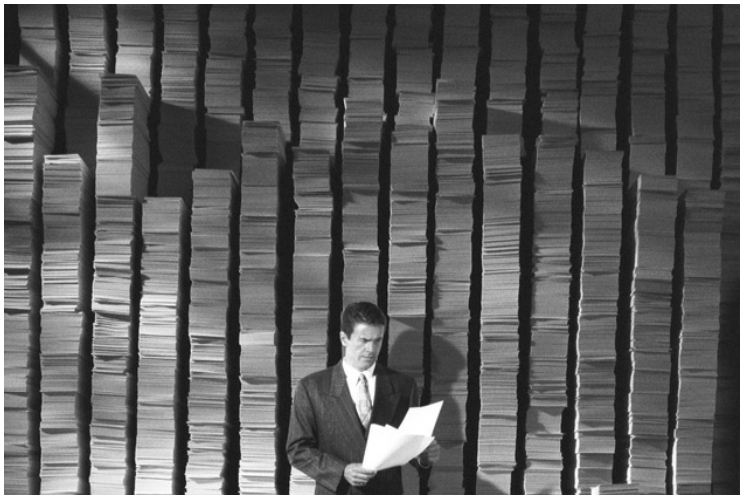
- Plusieurs niveaux de Cloud
  - **Infrastructure as a Service** – Mise à disposition de l'infrastructure (e.g., machines)
  - **Platform as a Service** – Amazon EC2, déploiement de puissance de calcul
  - **Software as a Service** – Mise à disposition de services



Was ist die Cloud

- Exemples:
  - Vos emails (gmail, ymail).
  - Vos documents. (Google docs, dropbox)
  - Votre code source (github.com, Google code).
  - Votre PC (Amazon Cloud, Chrome PC)
  - ... Même vous et vos amis êtes sur le Cloud (Devinez).

Comment s'y retrouver?



- Des moteurs de recherche pour indexer les milliards de pages web.
- Des systèmes de recommandation.
- L'agrégation de contenu, exemple: portails d'information et flux RSS.

Programme pour aujourd'hui

- Introduction au Software-as-a-Service
  - Yahoo Pipes
- Introduction au Platform-as-a-Service
  - Map-Reduce dans le navigateur

Introduction aux flux RSS

- Toujours plus d'information (Blogs, journaux, forums, etc.)
- Garder un œil sur toutes les sources d'information devient impossible
- Pourquoi l'information ne viendrait-elle pas directement vers vous plutôt que vous deviez aller vers l'information?

Introduction aux flux RSS

- Certains sites Web proposent de délivrer l'information sous forme de streaming (flux RSS).
- Il suffit de souscrire à vos flux préférés et spécifier des filtres, la mise à jour se fait à chaque nouvel évènement.
- Il est surtout possible de mélanger les flux de différentes sources.
- Un flux est consulté à l'aide d'un lecteur RSS.

Introduction aux flux RSS

flux sources site ou service

S'abonner à un fil d'info existant
Exploiter le flux sortant d'un service que j'utilise
Créer un flux à partir d'une page statique
Créer et alimenter un flux à partir d'un courriel
Créer et alimenter un flux en publiant des billets
Créer et alimenter un flux par IM

agrégation - lecture mixage - filtrage

Lire mes flux à partir d'un agrégateur web
Combiner plusieurs flux pour n'en faire qu'un
Mixer, filtrer, et publier mes flux
Créer une page de démarrage personnelle
Créer un portfolio personnalisé
Optimiser l'utilisation de mes flux

conversion réutilisation

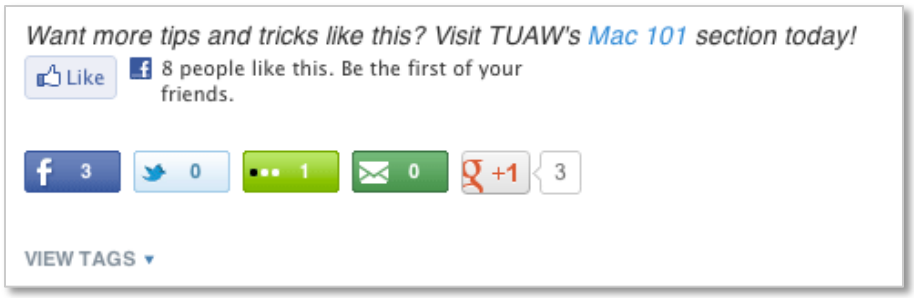
Recevoir le contenu d'un flux par courriel ou IM
Transformer le contenu d'un flux en PDF
Vocaliser le contenu d'un fil d'info
Republier le contenu d'un flux sur mon site
Créer un moteur de recherche personnalisé
Créer un nuage de mots clés à partir d'un flux

L'information à travers les réseaux sociaux

- L'effet médiatique des réseaux sociaux tel que Facebook permet une plus grande propagation de l'information.
- Un simple bouton "J'aime","Je partage" sont les principaux moteurs de cet effet viral.

L'information à travers les réseaux sociaux

- Un plugin social est une portion de code a intégrer sur une page HTML, et qui permet de partager cet même page avec ses amis.
- De plus en plus de sites web intègrent les plugins des réseaux sociaux sur leurs interfaces.
  - Pourquoi ?

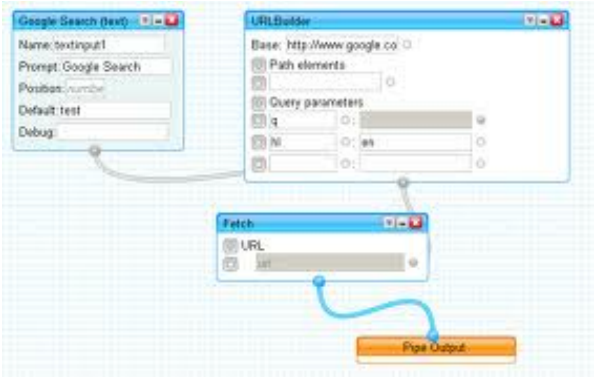


Yahoo! Pipes

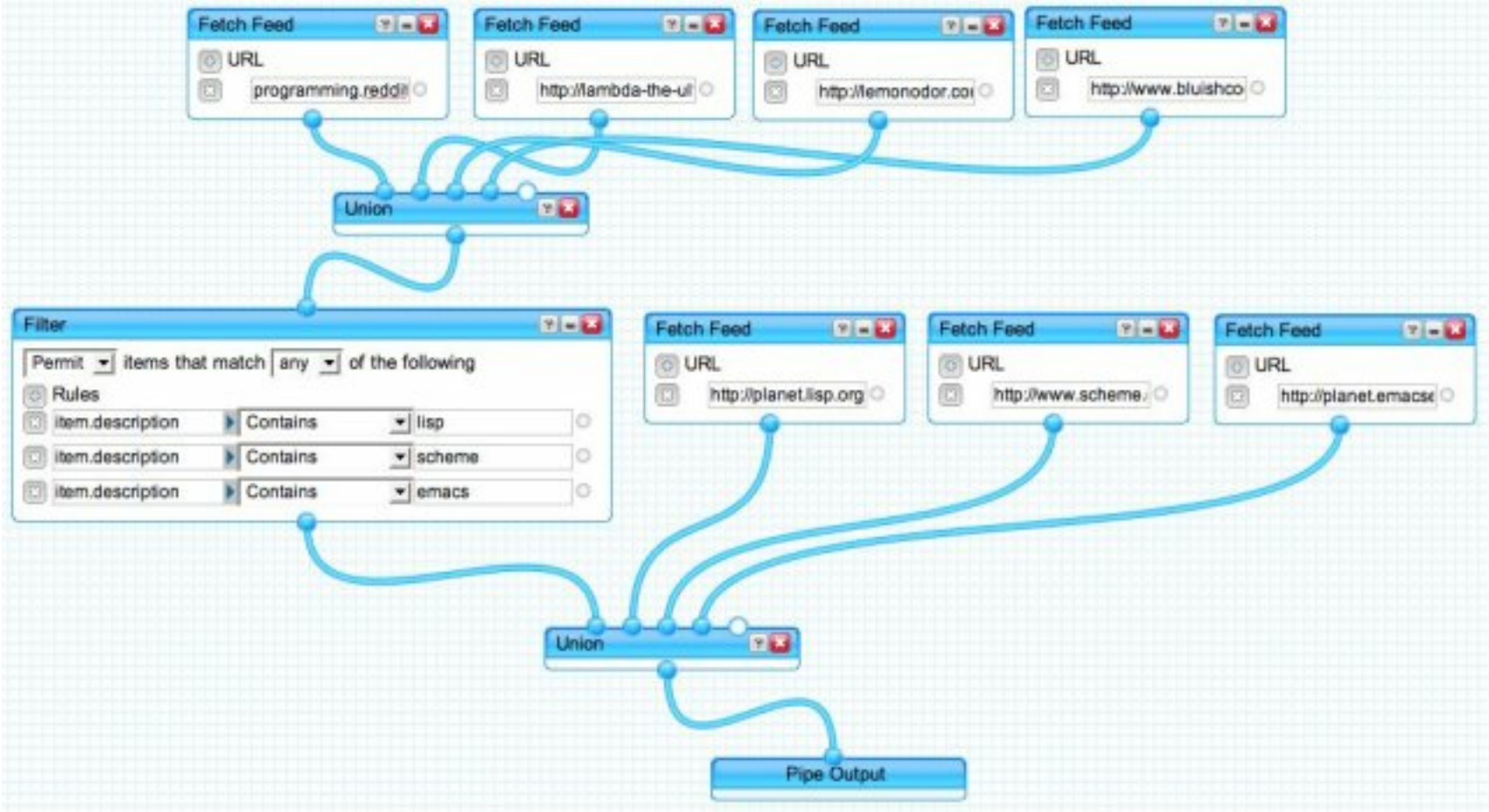
- Yahoo! Pipes est un service Cloud gratuit qui permet de mixer du contenu Web (RSS, XML, etc.).
- L'avantage est de créer une source de données personnalisée qui pourrait être partagée.

Yahoo! Pipes

- «Pipes Editor» est l'outil visuel mis à disposition par Yahoo! qui permet de:
  - Ajouter une ou plusieurs sources d'information.
  - Appliquer des filtres, ordonner et transformer l'information.
  - Rediriger le tout vers une sortie commune «Output».

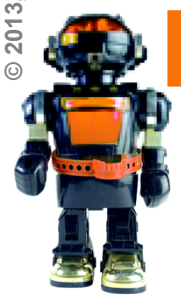


Yahoo! Pipes



Traitement parallèle et massif de données dans le cloud

MAP REDUCE

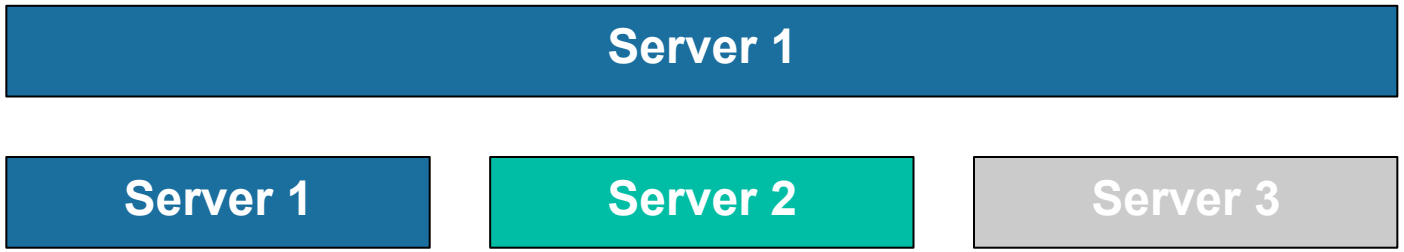


Map Reduce

- Chaque jour, des quantités inimaginables de données sont produites dans les centres de données de Google, Facebook & Co
  - Facebook compte environ 180.000 serveurs dans son infrastructure
  - Le clusert HDF de Facebook avait en Juillet 2012 100 PB de mémoire ~ 200.000 disques durs de 500 GB

Map Reduce

- Comment peut-on travailler avec ces quantités astronomiques de données?
  - Un disque dur lit en séquentiel 100MB/s
    - 100PB/100MB/s =>**32 ans de lecture**!
- But: traitement massivement parallèle
  - Parallélisation des données



Map Reduce – Pour quoi faire?

- Analyser de sentiments sur Twitter...
- Proposer de nouveaux amis sur Facebook...
- Analyser des données scientifiques (e.g., au CERN)...
- Analyser des logs pour prévenir des fraudes

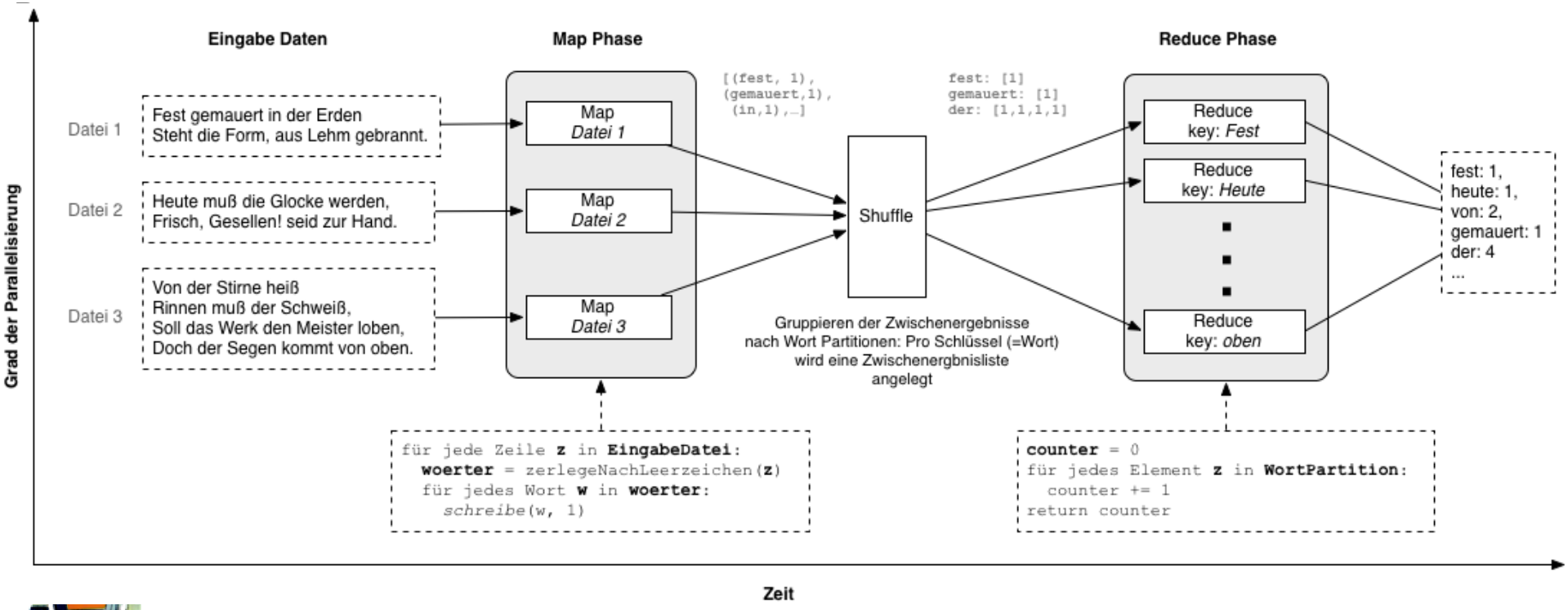
Map Reduce

MapReduce est un modèle de programmation qui permet de gérer ***d'énormes quantités de données*** *de manière* ***parallèle****. Les processus MapReduce sont divisés en deux étapes. La phase* ***MAP*** *filtre et trie les données, tandis que la phase* ***REDUCE*** *agrège les données résultant du MAP.*

Map Reduce

- map() – Applique une fonction à une liste de données en entrée afin de créer une nouvelle liste de données en sortie
  - Mais: taille(Entrée) n'est pas forcément égale à taile(Sorte)
  - Ex.: map(["abc", "cdef", "a"){|e| e.size } = [3,4,1]
- reduce() – Applique une fonction à une liste de données en entrée de manière à les réduire à un seul élément en sortie
  - Ex.: reduce([3,4,1]) {|m,v| m + v} = 8

Map Reduce: Exemple



Exercices

- Yahoo Pipes
  - Créer un tuyau (i.e., une *Pipe*) qui recherche quelque chose sur le site de lematin.ch et renvoie les résultats sous forme de flux RSS!
- Map Reduce
  - Analyse des données Twitter!