Base de données du SIRTA
Flux en temps réel

SIRTA
Site Instrumental de Recherche par Télédétection Atmosphérique

M.A. Drouin
marc-antoine.drouin@lmd.polytechnique.fr

Rappels sur la base SIRTA

Base de données SIRTA : les chiffres

- 17 ans
- 14 To
- 15 millions de fichiers
- ~600 fichiers produits par jour
- 80 flux de données

Base de données SIRTA : l'organisation (1/2)

- Partie **privée**
  - Données brutes (sortie d'instruments)
  - Données de campagne, algorithmes en test ...
  - Accès restreint (demande d'autorisation nécesaire)
- Partie **publique**
  - Données formatées (netCDF ou CSV)
  - Niveau 1 et plus
  - Grandeurs géophysiques et produits dérivés
  - Accès libre aux données
- chemins d'accès
  - niveau_données/instrument/année/mois/jour/fichiers_données.nc

Base de données SIRTA : l'organisation (2/2)

Nommage des fichiers

basta_1a_cldradLz1R012m_v03_20181127_000000_1440.nc

- basta : nom de l'instrument
- 1a : niveau de données
- cldradLz1R012m : champs libre de description
- v03 : version des données
- 20181127 : date de début des données
- 000000 : premier pas de temps des données
- 1440 : durée du fichier en minutes

Base de données SIRTA : l'accès (1/2)

- Depuis le mésocentre IPSL
  - Ouverture de compte https://mesocentre.ipsl.fr
  - Base publique
    - Dans **/bdd/SIRTA/pub**
    - Accessible à tous les utilisateurs avec un compte
  - Base privée
    - Dans **/bdd/SIRTA/priv**
    - Nécessite une autorisation
- Sur le site web du SIRTA http://sirta.ipsl.fr
  - Outils de recherche et de visualisation
  - Formulaire de demande de données

Base de données SIRTA : l'accès (2/2)

- Par FTP
  - Base publique
    - **ftp://ftp.sirta.ipsl.polytechnique.fr**
    - login : **sirta_access**
    - pas de mot de passe
  - Base privée
    - Sur demande (contacter J.C. Dupont)
- Par HTTP
  - Base publique uniquement
  - http://sirta.ipsl.polytechnique.fr/bdd/pub/basesirta
  - Seuls un lien “complet” vers un fichier fonctionne

Flux en temps réel

Temps réel : pourquoi ?

- Mise en valeur des données de l’observatoire
  - QLs visibles sur le site web le jour des mesures
- Amélioration du suivi de l’instrumentation
  - Détection des problèmes plus tôt
  - Visualisation des paramètres instrumentaux
  - Mise en place d’un système d’alertes automatiques
- Envoi des données aux réseaux (E-profile)

Flux de données "classique"

- Collecte des données 1 fois/jour
- Traitement déclenchés au cours de la nuit
- Permet de limiter le nombre de fichiers
  - 1 fichier par jour par instrument/mesure
- Système fonctionne et est stable
  - Pas de volonté de le changer
  - Mais forte demande des PIs pour accès temps réel

Tests temps réel avec le système classique

- On augmente la fréquence de récupération des fichiers
  - Toutes les heures pour la plupart des données
  - Plus fréquemment en cas de besoin (5 minutes pour E-profile)
- Problèmes
  - jusqu’à 288 fichiers/jours au lieu de 1
  - Difficile à gérer par les systèmes de fichiers
  - Mal géré par le système de sauvegarde
  - Permet seulement de voir les données sur le site
  - Pas de solution apportées pour la surveillance
  - Arrivée à 23 millions de fichiers en 2017

TEMPS RÉEL : CAHIER DES CHARGES FONCTIONNEL

- 2 solutions dépendant des besoins
- Affichage sur le site web
  - Création d’un flux parallèle
  - Collecte des données au mieux toutes les 10 minutes
  - Données en cache sur le serveur de traitement pendant 48h
  - Après 24h, la collecte classique prend le relai
  - On garde le même système de traitement
- Suivi instrumental
  - Création d’un autre flux parallèle
  - Collecte au mieux toutes les minutes
  - Utilisation d’une base de données dédiée
  - Conservation des données pendant 30j

Temps réel : collecte

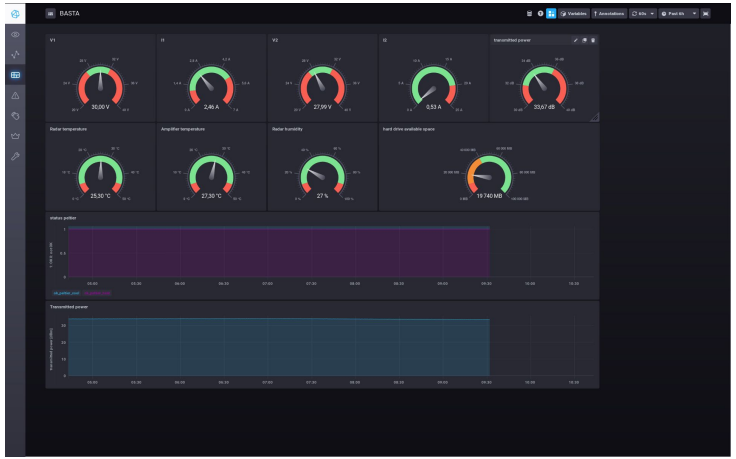
- Conditions requises
  - Que les instruments permettent la collecte en temps réel
  - Pas de perturbations de la collecte “classique”
- Centrales Campbell
  - collecte par HTTP
  - Toutes les 10 minutes pour visualisation
  - Toutes les minutes pour suivi instrumental
  - Fonctionne avec tous les modèles récents
- Autres instruments
  - Fichiers qui grossissent (CHM15k, hatpro)
  - Double collecte (ex: CL31)
    - 1 fichier créé toutes les 5 minutes
    - 1 fichier créé toutes les 24 heures

Suivi instrumental : quels outils (1/2)?

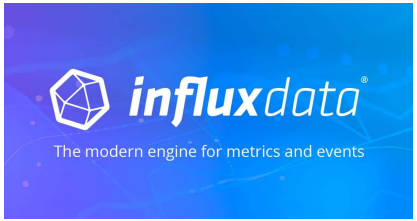
- Installation/mise à jour facile
- Utilisation facile
- Communauté active autour de l'outil
- Documentations et exemples disponibles
- Mutualisation avec le LMD
  - surveillance infrastructure

Suivi instrumental : quels outils (2/2)?

- Nombreux outils testés
- Bases de données:
  - prometheus (compliqué, documentation peu claire)
  - nagios (trop vieux)
- visualisation
  - chronograf
    - Manque de fonctionnalités
    - Compatible uniquement avec influxdb
- alertes
  - kapacitor
    - Compatible uniquement avec influxdb
    - Manque de fonctionnalités



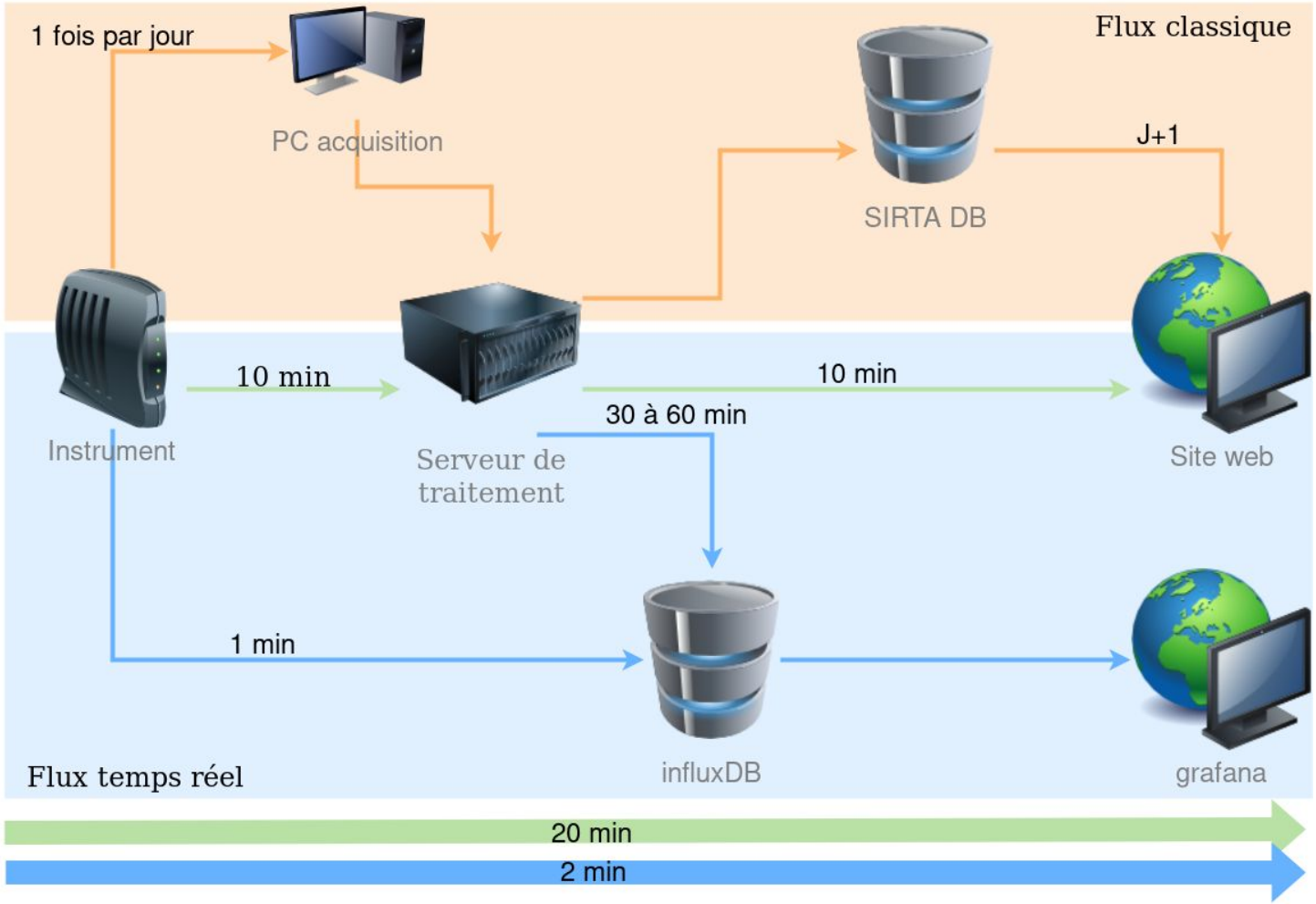
InfluxDB



- https://influxdata.com
- Base de données de série temporelle (TSDB)
- Opensource et gratuit mais fonctions limitées
- Bonne documentation
- Installation et mise à jour simple
- Actuellement en 1.7.x version 2.x.x cette année
- Clients disponibles dans (presque) tous les langages
  - client python compatible avec pandas
- Facile à alimenter en données
- Temps de rétention des données géré automatiquement
- Quelques fonctionnalités encore manquantes

GRAFANA

- https://grafana.com/
- Outils dédié à la visualisation de Séries temporelles
- Facile à installer et configurer
- Compatible avec nombreuses TSDB
- Gestion des utilisateurs/groupes
- Nombreux types de visualisations
- Facile à prendre en main
- Produit mature
- Fonction d'alerte automatique inclus
  - email, slack …
- Beaucoup d'exemples, beaucoup de plugins


1 fois par jour
PC acquisition
Flux classique
SIRTA DB
J+1
Instrument
10 min
Serveur de traitement
10 min
30 à 60 min
Site web
1 min
influxDB
grafana
Flux temps réel
20 min
2 min

Démo

Ce qu'il reste à faire

- Changer la manière dont sont stockées les données dans influxdb
  - Meilleur accès aux données dans grafana
  - Intercomparaison entre les instruments
  - Dashboard plus généralisable
  - Simplification en cas de nouvel instrument du même type
- Améliorer les scripts de collecte
  - Essayer d'utiliser les outils de collecte d'influxdata (telegraf)
- Suivre d'autres instruments
- Continuer à définir les règles d'alertes
- Mettre en place des écrans de visualisation

Merci
Des questions ?