



RetEx sur le déploiement d'une Infrastructure ETL

Talend Enterprise Data Infrastructure

Julien Bouillot - Avril 2019

INSTITUT NATIONAL SUPÉRIEUR DES SCIENCES AGRONOMIQUES, DE L'ALIMENTATION ET DE L'ENVIRONNEMENT

MEMBRE DE



- Contexte
- Besoin métier
- Architecture Technique
 - Déploiement Pré-Production
 - Déploiement Production
- Coût du projet
- Problématiques rencontrées
- Avenir de la plateforme

Contexte



- DSI AgroSup Dijon
- 4 services
 - 3 services de développement (~30 personnes)
 - SI Interne
 - Web
 - Enseignement Agricole
 - Service MARSS (13 personnes)
 - Maintenance Interne
 - Infrastructure / Réseaux
 - Systèmes

- Expérience avec Talend Open Studio depuis ~10 ans
 - Turnover des agents
- Hétérogénéité des traitements
 - Code
 - Exécution
 - Suivi
- Besoins récurrents de transformation de données entre dispositifs

Besoin Métier



Choix de la solution Talend EDI :

- Fait (uniquement) par le service métier ayant le plus d'expérience avec l'outil
- Mise en place d'une infrastructure de Pré-Production pour évaluer l'adéquation au besoin
 - 2 licences utilisateur

Avantages attendus :

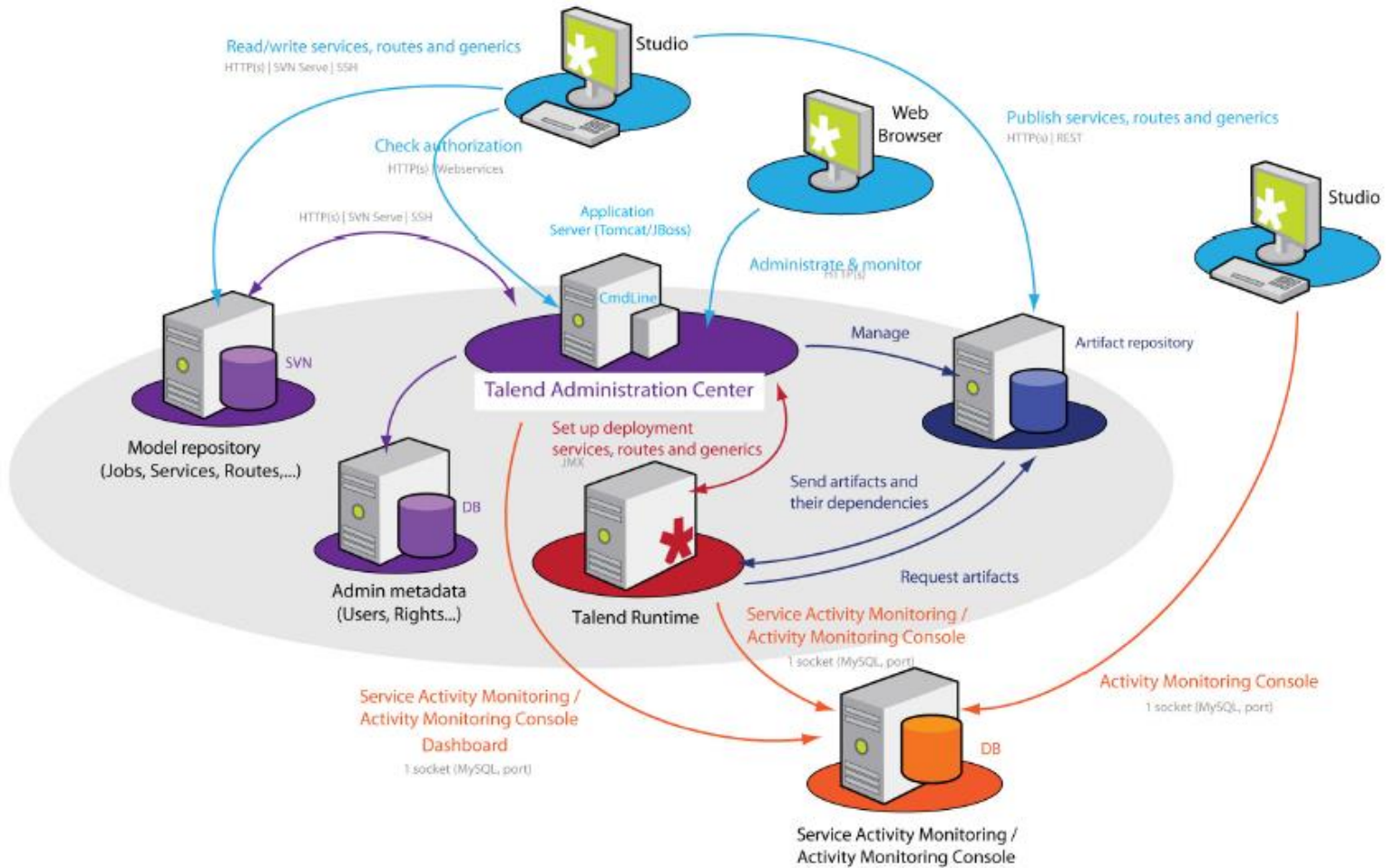
- Mise en commun de composants récurrents
- Homogénéisation des pratiques de développement
- Suivi des exécutions : orchestration, supervision
- Support éditeur

Architecture Technique



- Intégration de l'équipe MARSS
 - Réunion de lancement du projet (avril 2015)
- Fourniture du guide d'installation de la plateforme
- Pas d'éléments d'architecture précis

Figure 1.3. Architecture de base + Serveurs d'exécution + Monitoring

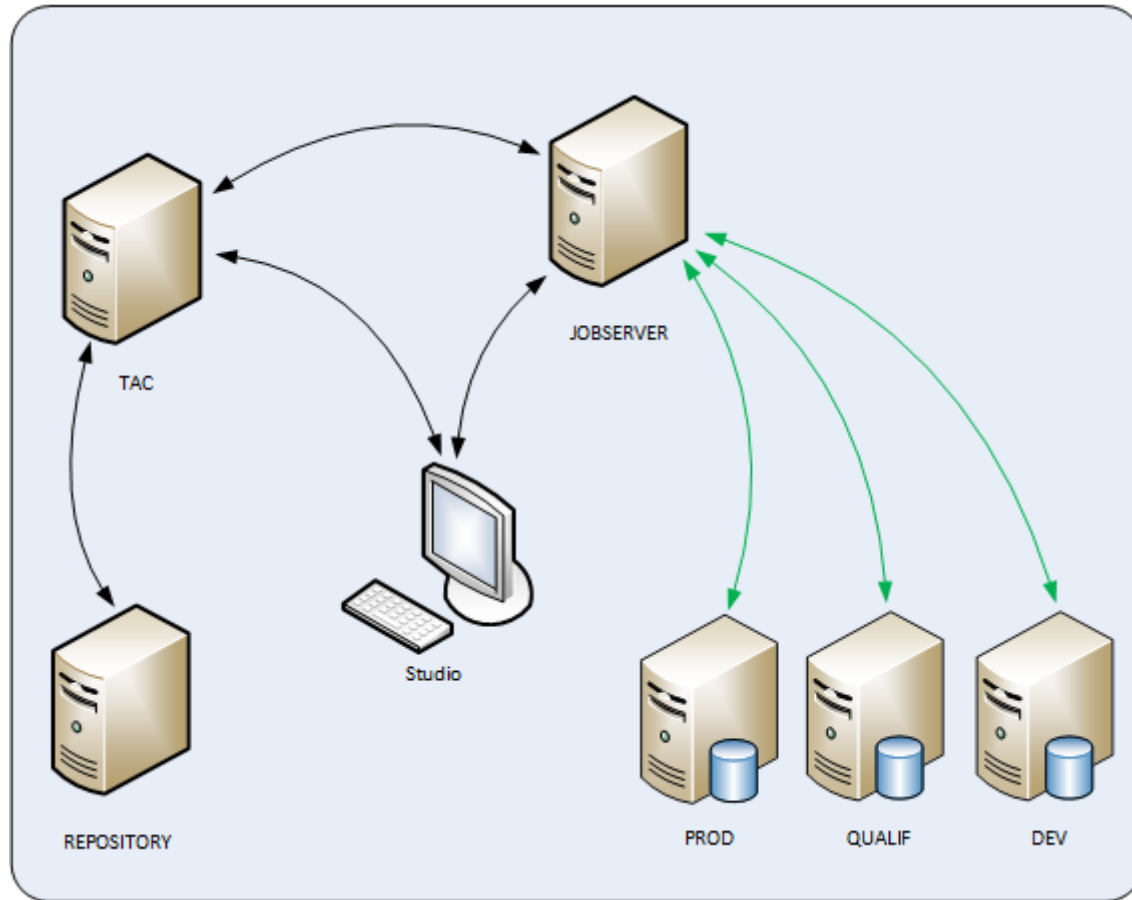


Déploiement Pré-Production



Architecture basée sur la documentation fournie :

- 3 VMs
 - Applicatifs Talend + Base SQL
 - Dépôt de code SVN
 - Serveur d'exécution de Jobs
- Système Debian
- Java 1.7 (2015)



- Installation chaotique
 - Documentation inégale
 - Incomplète pour partie
 - Inadaptée au déploiement recommandé
- Livraison Septembre 2015

Installation Infra Talend PreProd

Sommaire [masquer]

- 1 Caractéristiques Machines
 - 1.1 Serveur Talend Administration Center
 - 1.2 Serveur Repository
 - 1.3 Job Server
- 2 Récupération Sources Licence & Documentation
 - 2.1 Licence
 - 2.2 Sources d'installation
 - 2.3 Documentation
- 3 Installation
 - 3.1 Prérequis
 - 3.1.1 unzip
 - 3.1.2 Java
 - 3.1.3 Fichier */etc/hosts*
 - 3.2 Talend Repository
 - 3.2.1 SVN
 - 3.2.2 Archiva
 - 3.2.3 Nexus
 - 3.3 Talend Administration Center
 - 3.3.1 Prérequis
 - 3.3.1.1 postgresSQL
 - 3.3.2 TAC
 - 3.3.2.1 Setup
 - 3.3.2.2 Démarrage automatique de TAC
 - 3.3.3 Première connexion / Configuration accès BDD
 - 3.3.4 Configuration
 - 3.3.4.1 Audit
 - 3.3.4.2 CommandLine/primary
 - 3.3.4.2.1 Démarrage automatique de CommandLine
 - 3.3.4.3 Artifact Repository
 - 3.3.4.4 Job Conductor
 - 3.3.4.5 Logging
 - 3.3.4.5.1 Démarrage automatique du service Logging
 - 3.3.4.6 Software Update
 - 3.3.4.7 Svn
 - 3.3.4.8 Talend Suite
 - 3.3.4.9 Backup
 - 3.4 Talend Job Server
 - 3.4.1 Installation
 - 3.4.2 Configuration
 - 3.4.3 Configuration dans Talend Administration Center
 - 3.5 Talend Activity Monitoring Console
 - 3.5.1 **APARTE** : *Surveillance de la taille de la base amc*
 - 3.5.2 Création de la base de données
 - 3.5.3 Autorisation de connexion à la base *amc*
 - 3.5.4 Ajout du driver de connexion à la base de données
 - 3.5.5 Configuration dans Talend Administration Center

Bilan avant passage production (mi-2017) :

- Adaptations techniques à prévoir
 - Sauvegarde/Restauration
 - Sécurisation des flux
- Evaluation métier incomplète
 - Performances
 - Processus

Déploiement Production



- Passage d'une version majeure (5.6 -> 6.1)
 - Procédure non fournie
 - Conservation des données
- Modification de l'architecture
 - Découpage composants
 - Dimensionnement adapté

- Recours à une prestation de l'éditeur
 - Validation du déploiement (architecture, procédure)
 - Formation utilisateurs
- 5 jours (août 2017)
 - 2 jours dédiés à l'infrastructure
 - 3 jours formation exploitation métier

Résultat de la prestation

- Validation des éléments techniques
- Connaissance relative du déploiement
- Limites de la prestation

Sommaire [masquer]

1 Prérequis communs

1.1 Configuration proxy

1.1.1 Solution approuvée

1.1.2 Mise en place testée

1.2 unzip

1.3 Java

1.3.1 NOTE : a considérer

1.4 Fichier */etc/hosts*

2 Talend Repository

2.1 SVN

2.2 Nexus

3 BDD

4 JobServer

4.1 Installation

4.2 Configuration

4.3 Authentification Utilisateurs

4.4 Appender LogServer

5 Talend Administration Center (TAC)

5.1 Modification variable système

5.2 Setup

5.3 Ajout du driver de connexion à la base de données AMC

5.4 Finalisation installation

5.4.1 Configuration

5.4.1.1 Audit

5.4.1.2 CommandLine/primary

5.4.1.3 Monitoring

5.4.1.4 Artifact Repository

5.4.1.5 Job Conductor

5.4.1.6 Logging

5.4.1.7 Software Update

5.4.1.8 Svn

5.4.1.9 User Libraries

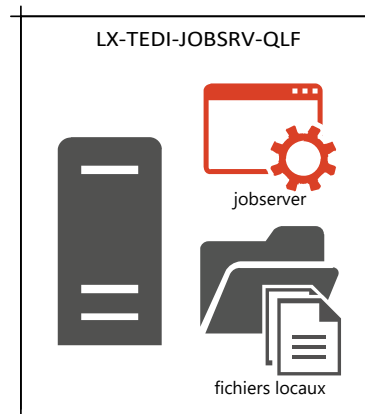
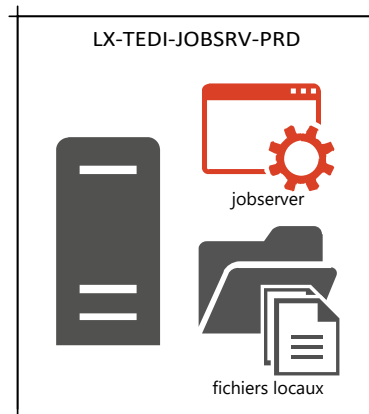
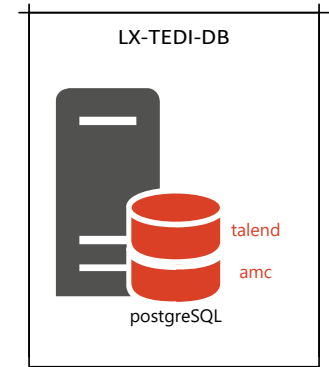
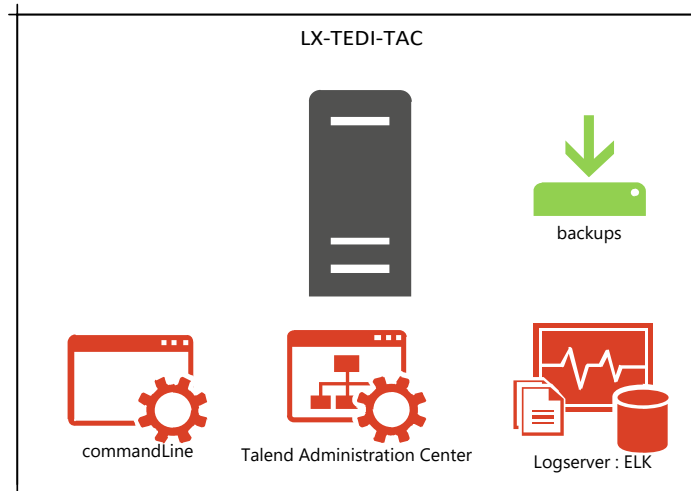
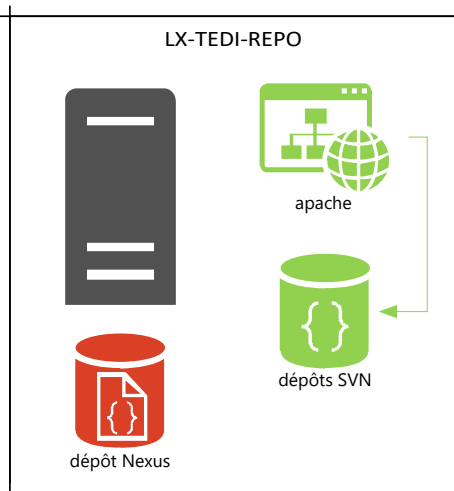
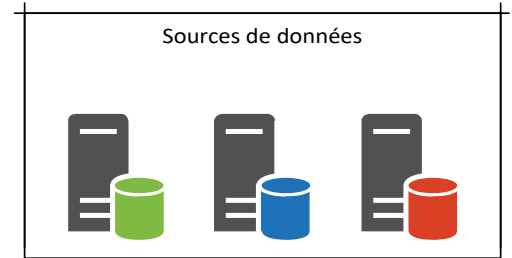
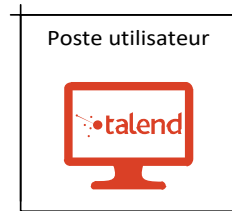
5.4.1.10 Backup

5.4.1.10.1 NOTE : Fonction dépréciée

5.5 Configuration JobServer

5.6 Appender LogServer

6 SVG Liens HTTPs Kibana



Processus d'utilisation de la plateforme

- Développement en local depuis le client TOS
 - Debug, Tests
- Intégration à la plateforme
 - Portage dans le client TIS
 - Création du projet
 - Import du job existant
 - Exécution à distance sur un jobserver
 - Mise en production
 - Exécution planifiée
 - Suivi

Coût du projet



- Licence / Maintenance
 - Environ 20 k€ / an pour 2 licences
- Temps homme
 - 35 j en 2015
 - 11 j en 2016
 - 56 j en 2017
 - 16 j en 2019
- 5 serveurs

Problématiques rencontrées



- Compréhension de l'outil
 - Fonctions des composants
 - Echanges réseaux
 - Ex : job avec connexion reposant sur certificat
- Documentation très inégale
- Support éditeur
 - Ex : Procédure mise à jour

- Talend
 - Installation Infra Talend PreProd
 - Infra Prod
 - Migration 5.6 - 6.4
 - Installation
 - Exploitation
 - Gestion des comptes
 - Exploitation de la plateforme
 - CommandLine
 - Expiration licence
 - Indisponibilité du service
 - Echanges nécessitant certificats

Sommaire [masquer]

1 Infrastructure de la plateforme

1.1 Machines & fonctions

1.1.1 lx-tedi-tac

1.1.1.1 Talend Administration Centrer (TAC)

1.1.1.2 LogServ

1.1.1.3 CMDline

1.1.1.4 Backups

1.1.2 lx-tedi-db

1.1.3 lx-tedi-repo

1.1.3.1 Repo SVN

1.1.3.2 Nexus : Dépôt d'artefacts

1.1.4 JobServers

1.2 Services

1.2.1 lx-tedi-tac

1.2.1.1 Talend Administration Centrer (TAC)

1.2.1.2 CMDline

1.2.1.3 LogServ

1.2.2 lx-tedi-db

1.2.3 lx-tedi-repo

1.2.3.1 Repo SVN

1.2.3.2 Nexus

2 Vérifications d'usage

2.1 Système

2.1.1 Etat des services

2.1.2 Espace disque

3 Cas d'erreurs

3.1 CommandLine indisponible

3.2 Recherche problèmes droits projets

3.3 Disque plein

- Intégration des composants de la plateforme
 - Création des comptes, Authentification
 - Cache du serveur d'exécution
 - Gestion des licences
- Déploiement en production
 - Suivi de version Java
 - Sécurisation des flux
 - Sauvegarde/Restauration des données
 - Journalisation des composants

- Fonctionnement interne
 - Administration fonctionnelle de la plateforme
 - Exemples : création de comptes utilisateurs
- Debug
- Limites métier
 - licences
 - Composants logiciels

Avenir de la plateforme



Migration future :

- Versions
 - Application (majeure)
 - OS (majeure)
 - Java
- Sera impulsée par la demande des services métier
- Nouvelle révision de l'architecture de l'infrastructure

Conclusion



- De l'importance du PoC
- Implication des équipes techniques dès les premières phases projet

Merci de votre attention

Avez-vous des questions ?