

Accélération du traitement de données avioniques : Hadoop@PireCloud

François Thiebolt(1), Aurélien Ortiz(2), Congduc Pham(3), Jean-Marc Pierson(4)

(1) thiebolt@irit.fr, IRIT – UMR 5505

(2) aortiz@serviware.com, ServiWare Toulouse

(3) congduc.pham@univ-pau.fr, Université de Pau et des pays de l'Adour

(4) pierson@irit.fr, IRIT – UMR 5505

Overview :

During a flight, every aircraft monitors and collects numerous data like speed, altitude, oil pressure, temperature, acceleration and so on. A small aircraft (up to 10 seats) generates something like 2GB of data every year which is very small comparing to a company managing more than 200 Airbus A320.

Every time a plane has landed, the in-flight data are collected for an off-board analysis to determine whether the aircraft requires any maintenance session. Efficiency and accuracy of this analysis is very important to keep the aircrafts flying at their maximum.

2Moro is a french company specialized in the analysis of such data. To face the ever increasing amount of in-flight data and to improve the responsiveness of their tools, we proposed them to switch from a monolithic approach to a cloud solution featuring the Hadoop framework [HAD].

This work was part of the Aragon-Aquitaine project named « OMNI-DATA » [CPH]

Les besoins :

Produire de la valeur ajoutée à partir du traitement de données de vol passe par un cycle d'analyse et de comparaison avec des états antérieurs. Le tableau de bord associé à chaque appareil donne ainsi à l'utilisateur une vision claire de l'état de sa flotte et lui permet d'optimiser la planification des sessions de maintenance.

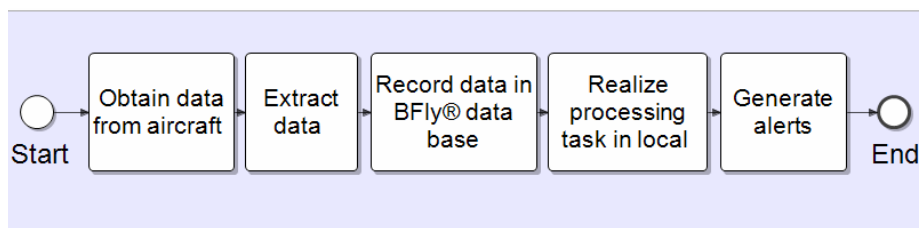


Fig. I : Approche monolithique du cycle de traitement des données de vol.

Dans ce workflow (fig. I), les données de vol collectées sont converties sous une forme exploitable pour être ensuite ajoutées à la base contenant l'ensemble des paramètres de vols. Le cycle d'analyse qui s'exécute sur un seul nœud de calcul extrait les informations et enregistre les résultats dans cette même base.

Avec un nombre croissant d'appareils et des volumes de données générées par chacun d'eux en constante augmentation, la nécessité d'une approche parallèle du traitement des informations de vol s'impose.

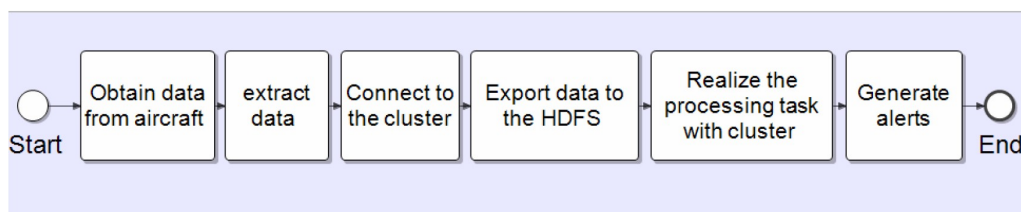


Fig. II : Nouvelle approche MapReduce du traitement des données de vol.

La répartition des données dans le système de fichiers HDFS et l'exploitation parallèle de plusieurs nœuds de calcul (fig. II) va permettre d'augmenter considérablement la réactivité du cycle d'analyse et de ce fait de produire plus rapidement de la valeur ajoutée pour l'utilisateur final.

Maintenant, l'exploitation du parallélisme dans le traitement des données de vol passe par la mise en œuvre du paradigme MapReduce ce qui signifie la ré-écriture du cœur de l'outil d'analyse.

Hadoop @ PireCloud :

Sur la période allant de 2009 à 2012, le programme européen Interreg IV A du POCTEFA a cofinancé la mise en œuvre d'une grille de calcul entre l'Espagne et la France. L'objectif était d'aider et d'accompagner les entreprises à s'approprier des technologies de traitement de données déjà couramment utilisées dans le monde de la recherche. PireGrid[Pir02] a été cette plateforme de calcul opérationnelle sur 4 sites dans les régions d'Aragon, de Navarre, d'Aquitaine et de Midi-Pyrénées.

Par la suite, pour compléter l'offre aux entreprises, PireGrid est devenu PireCloud avec la mise à disposition d'une plateforme de Cloud notamment sur le site de Pau. Pour ce faire, nous avons mis en place l'outil de management de machines virtuelles « Open Nebula » [OpN] ainsi qu'un accès par VPN. Pour gagner en accessibilité, nous avons rendu visible depuis Internet les machines virtuelles déployées au sein du cluster. Enfin, pour conserver les fonctionnalités de la grille, les services Glite [GLI] (v3.2) ont été virtualisés ce qui permet à la plateforme de Pau de fonctionner simultanément dans les deux modes.

C'est donc dans ce contexte PireCloud qu'ont été menées les expérimentations avec la société 2Moro (fig. III).

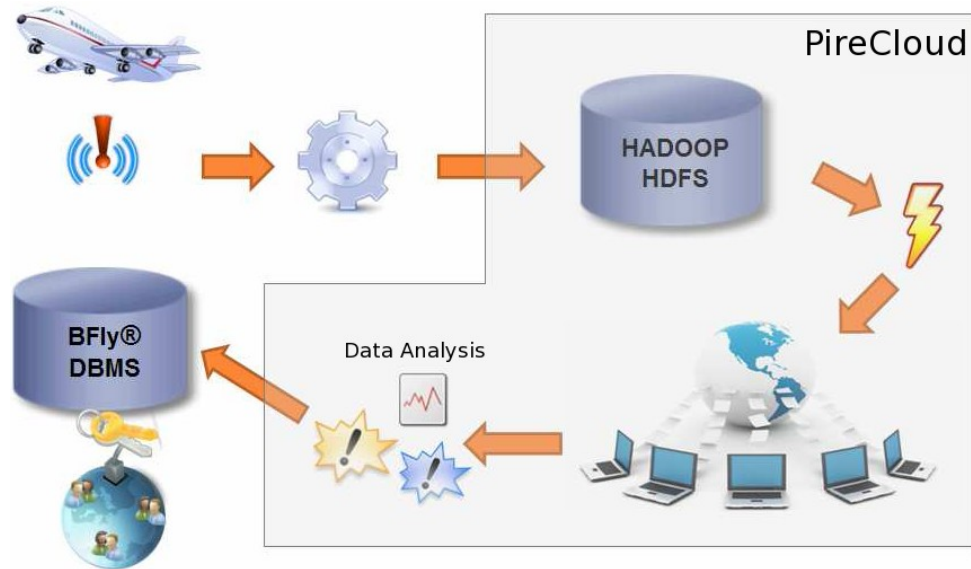


Fig III : Déploiement des outils 2Moro sur PireCloud avec Hadoop

Pour valider la pertinence de l'analyse de données de vol sur une plateforme de type Cloud, la société 2Moro a fait appel à l'outil de modélisation de workflow Intalio [INT].

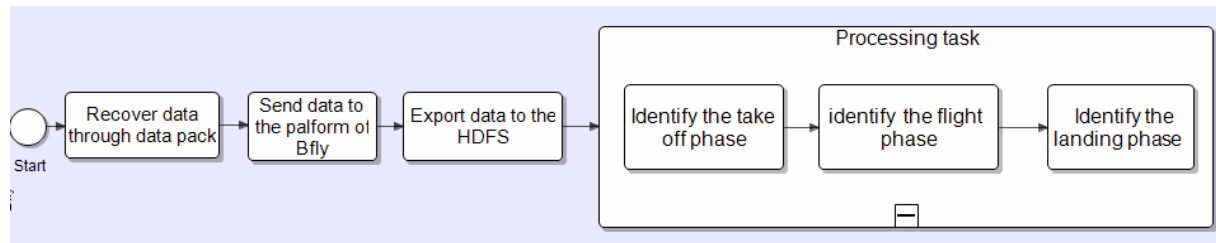


Fig IV : Vue d'ensemble du workflow d'analyse de données de vol

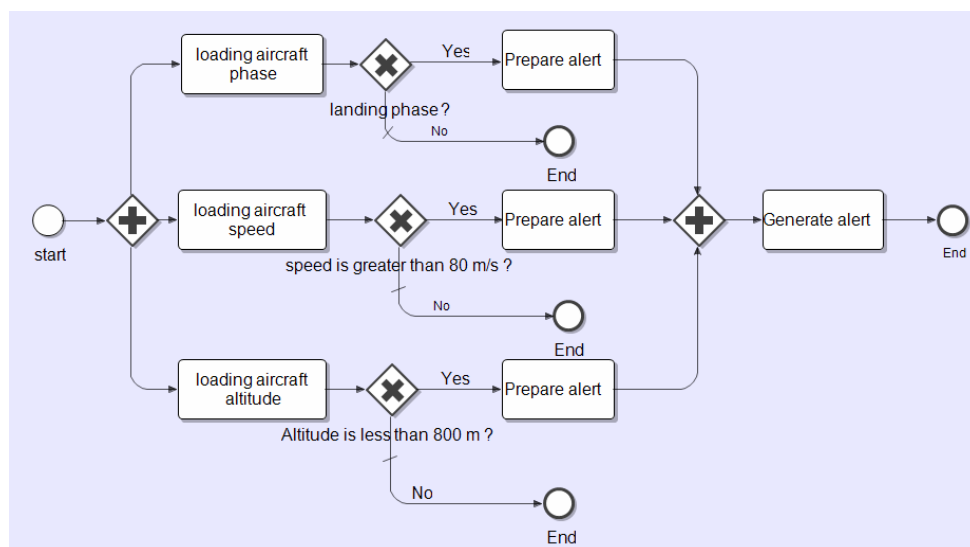
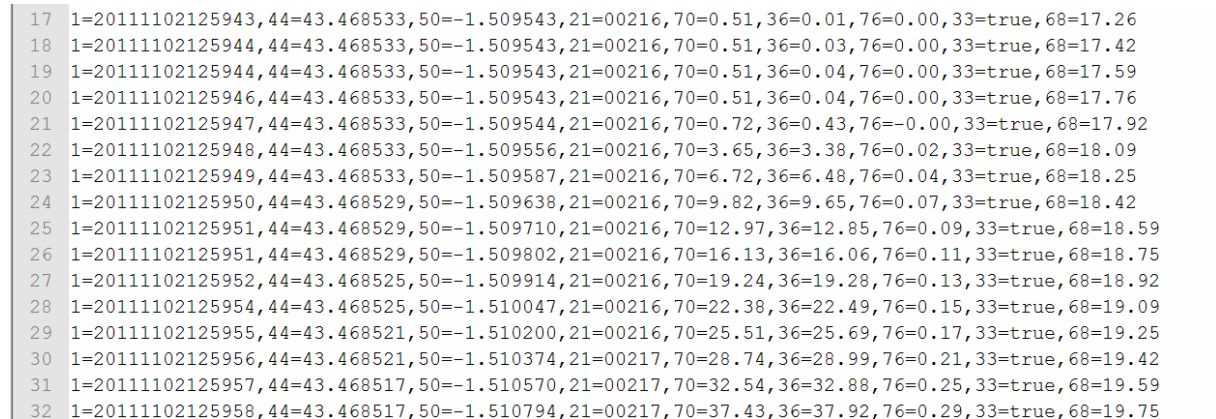


Fig V : Génération d'alertes relatives à la phase d'atterrissage.

Le workflow global (fig. IV) fait apparaître des traitements spécifiques liés aux différentes phases de vol. L'analyse simplifiée de la phase d'atterrissage (fig. V) va ainsi générer des alertes en fonction de la vitesse et de l'altitude.

Résultats :

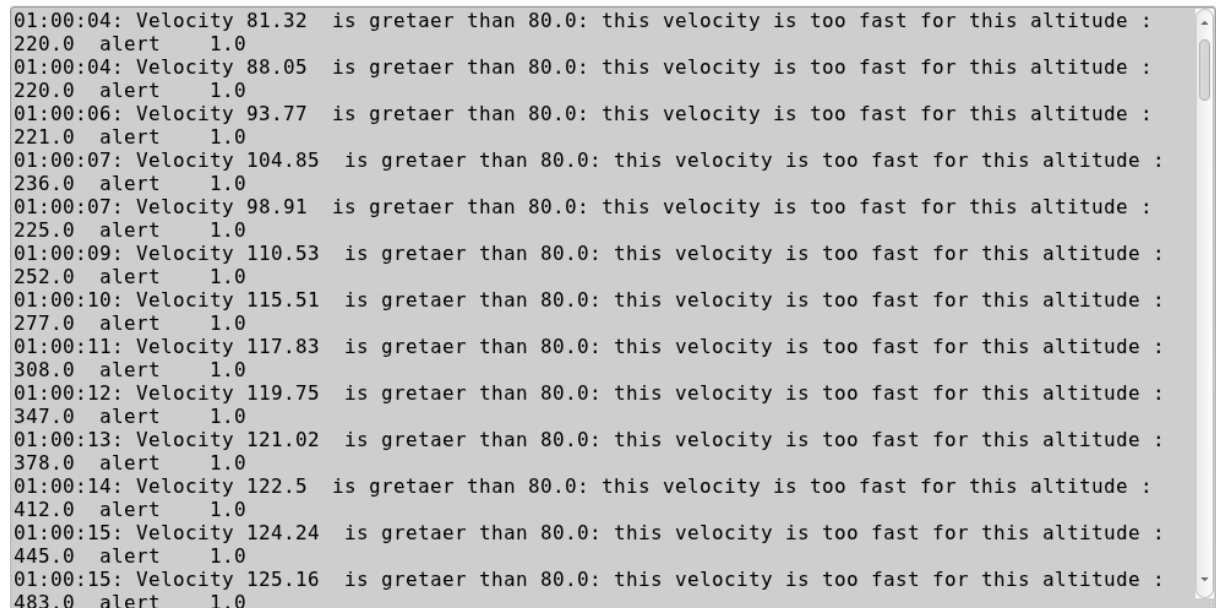
Le site PireCloud de Pau dispose de 16 lames ce qui représente un total de 128 processeurs Xeon 5520 2.26Ghz et 768Go de RAM. L'environnement de test fait appel à 4 machines virtuelles Ubuntu (2 cpu, 4Go ram par VM) déployées via OpenNebula v3.0. La configuration Hadoop sur ces VMs fait état d'un nœud maître et de 3 nœuds esclaves.



17	1=20111102125943,44=43.468533,50=-1.509543,21=00216,70=0.51,36=0.01,76=0.00,33=true,68=17.26
18	1=20111102125944,44=43.468533,50=-1.509543,21=00216,70=0.51,36=0.03,76=0.00,33=true,68=17.42
19	1=20111102125944,44=43.468533,50=-1.509543,21=00216,70=0.51,36=0.04,76=0.00,33=true,68=17.59
20	1=20111102125946,44=43.468533,50=-1.509543,21=00216,70=0.51,36=0.04,76=0.00,33=true,68=17.76
21	1=20111102125947,44=43.468533,50=-1.509544,21=00216,70=0.72,36=0.43,76=-0.00,33=true,68=17.92
22	1=20111102125948,44=43.468533,50=-1.509556,21=00216,70=3.65,36=3.38,76=0.02,33=true,68=18.09
23	1=20111102125949,44=43.468533,50=-1.509587,21=00216,70=6.72,36=6.48,76=0.04,33=true,68=18.25
24	1=20111102125950,44=43.468529,50=-1.509638,21=00216,70=9.82,36=9.65,76=0.07,33=true,68=18.42
25	1=20111102125951,44=43.468529,50=-1.509710,21=00216,70=12.97,36=12.85,76=0.09,33=true,68=18.59
26	1=20111102125951,44=43.468529,50=-1.509802,21=00216,70=16.13,36=16.06,76=0.11,33=true,68=18.75
27	1=20111102125952,44=43.468525,50=-1.509914,21=00216,70=19.24,36=19.28,76=0.13,33=true,68=18.92
28	1=20111102125954,44=43.468525,50=-1.510047,21=00216,70=22.38,36=22.49,76=0.15,33=true,68=19.09
29	1=20111102125955,44=43.468521,50=-1.510200,21=00216,70=25.51,36=25.69,76=0.17,33=true,68=19.25
30	1=20111102125956,44=43.468521,50=-1.510374,21=00217,70=28.74,36=28.99,76=0.21,33=true,68=19.42
31	1=20111102125957,44=43.468517,50=-1.510570,21=00217,70=32.54,36=32.88,76=0.25,33=true,68=19.59
32	1=20111102125958,44=43.468517,50=-1.510794,21=00217,70=37.43,36=37.92,76=0.29,33=true,68=19.75

Fig VI : Jeu de données de vol.

Les données d'entrée (fig. VI) fournies par 2Moro ont été réparties sous la forme de 4 fichiers de 64Mo dans le système de fichiers HDFS. Le code 36 représente la vitesse.



```
01:00:04: Velocity 81.32 is gretaer than 80.0: this velocity is too fast for this altitude :  
220.0 alert 1.0  
01:00:04: Velocity 88.05 is gretaer than 80.0: this velocity is too fast for this altitude :  
220.0 alert 1.0  
01:00:06: Velocity 93.77 is gretaer than 80.0: this velocity is too fast for this altitude :  
221.0 alert 1.0  
01:00:07: Velocity 104.85 is gretaer than 80.0: this velocity is too fast for this altitude :  
236.0 alert 1.0  
01:00:07: Velocity 98.91 is gretaer than 80.0: this velocity is too fast for this altitude :  
225.0 alert 1.0  
01:00:09: Velocity 110.53 is gretaer than 80.0: this velocity is too fast for this altitude :  
252.0 alert 1.0  
01:00:10: Velocity 115.51 is gretaer than 80.0: this velocity is too fast for this altitude :  
277.0 alert 1.0  
01:00:11: Velocity 117.83 is gretaer than 80.0: this velocity is too fast for this altitude :  
308.0 alert 1.0  
01:00:12: Velocity 119.75 is gretaer than 80.0: this velocity is too fast for this altitude :  
347.0 alert 1.0  
01:00:13: Velocity 121.02 is gretaer than 80.0: this velocity is too fast for this altitude :  
378.0 alert 1.0  
01:00:14: Velocity 122.5 is gretaer than 80.0: this velocity is too fast for this altitude :  
412.0 alert 1.0  
01:00:15: Velocity 124.24 is gretaer than 80.0: this velocity is too fast for this altitude :  
445.0 alert 1.0  
01:00:15: Velocity 125.16 is gretaer than 80.0: this velocity is too fast for this altitude :  
483.0 alert 1.0
```

Fig. VII : Extrait des résultats relatifs à l'analyse de la phase d'atterrissage.

La figure VII présente un extrait des résultats d'analyses relatifs à la phase d'atterrissage, les autres résultats n'ont pas été présentés pour la clarté du document.

Dans cette configuration de test, le traitement des données de vol s'est ainsi déroulé globalement trois fois plus vite qu'avec l'ancienne approche basée sur une solution faisant appel à un seul serveur accédant à une base de données centralisée. L'objectif de la démonstration sur une plateforme de type Cloud a ainsi été atteint.

Perspectives :

Ces résultats préliminaires ont pour l'essentiel permis à la société 2Moro de mettre en évidence la pertinence d'une approche basée sur le traitement de données de vol via le paradigme MapReduce dans une infrastructure de type Cloud, ce

qui correspond en tous points à l'objectif global du projet PireGrid/PireCloud.

Par ailleurs, cette plateforme dispose également d'un système performant de monitoring de l'énergie. Il permet de suivre l'évolution de la consommation de chaque nœud avec une précision de l'ordre de la seconde. Cela permet d'évaluer finement les économies d'énergies réellement apportées par les différentes politiques de placement/migration de VMs.

Prochainement, le GIS France-Grilles va disposer d'une nouvelle plateforme de Cloud sur le site de Toulouse nommée CloudMIP. Cette plateforme de 32 lames compte 256 processeurs XEON 5620 @ 2.4Ghz pour 1.5To de ram. Elle dispose d'un serveur NFS de 5To et plus de 9To d'espace disque hautes performances répartis dans les lames. Ces disques seront virtualisés et mis à disposition des utilisateurs sous la forme d'un espace de nom global (glusterFS). Cette plateforme, à peu de choses près identique à celle de PireCloud/Pau, dispose du même système de monitoring de l'énergie, de la dernière version de l'outil IaaS OpenNebula (v.3.6) et proposera prochainement l'interface graphique de management de VMs SunStone (OpenNebula GUI).

Références :

[CPH] : Projet « OMNI DATA » <http://web.univ-pau.fr/~cpham/>

[GLI] : Middleware de grille GLite <http://glite.cern.ch/>

[HAD] : Hadoop framework <http://hadoop.apache.org/>

[INT] : Outil de type BPMS (Business Process Management Software) <http://bpms.intalio.com/>

[OpN] : Outil de management de machines virtuelles <http://www.opennebula.org/>

[Pir02] : Projet PireGrid <http://www.piregrid.eu/>