



T.P. n°2 : Projet "Isil-Data-Scientist"

Objectif :

- ◆ Architecture de type Big data avec intégration de méthodologies de projet et de diverses technologies
- ◆ Savoir mener une **étude de recherche** dans le domaine du Big data, en ce compris l'aspect simulation.
- ◆ Maîtriser les bases du développement avec une **carte à puces** de type Java Card ou carte d'identité électronique
- ◆ Intégrer les différentes **communications réseaux** entre les différents intervenants d'un modèle client-serveur.

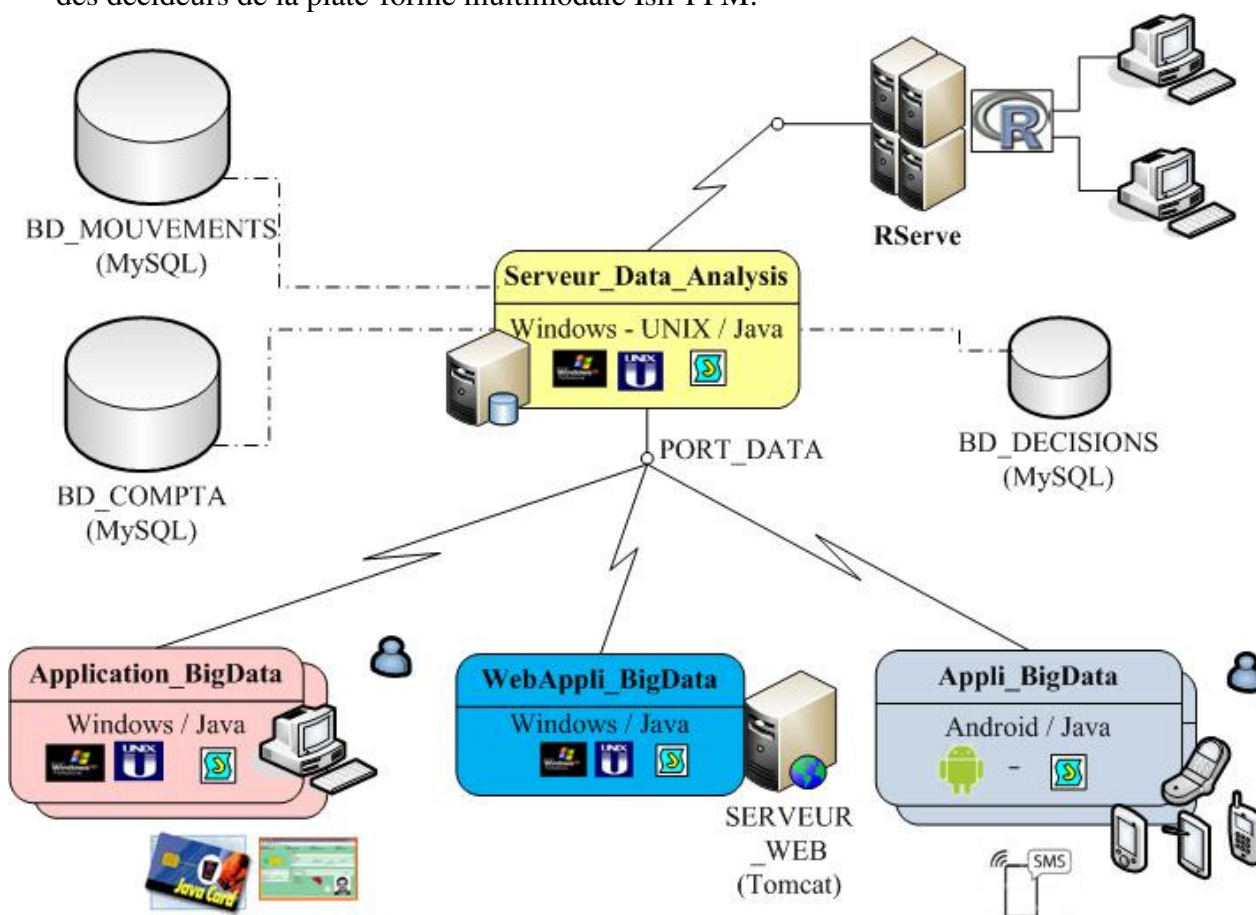
Travail à réaliser: **en équipe** et par étapes successives durant la période 11/2000—1/2021, avec produit final présenté conjointement lors de l'examen de janvier 2021

Composition de l'équipe

Nom	Prénom	Rôles principaux	Rôles secondaires

1. Une infrastructure de Data mining et Big data

Il s'agit de développer une petite infrastructure de traitement de Big data à destination des décideurs de la plate-forme multimodale Isil-PFM:



2. Data: Les données d'exploitation

Afin de situer la problématique, on imagine ici deux bases de données, l'une concernant la gestion de la plate-forme (BD_MOUVEMENTS) et l'autre répondant plutôt aux opérations comptables et financières (BD_COMPTA). Leur contenu possible n'est ici présenté qu'à titre exemplatif et peut être modifié.

2.1 La base BD MOUVEMENTS

Ses tables, si on admet un certain nombre d'approximations, de simplifications et d'omissions sans importance pour le projet tel que défini ici, seront en première analyse :

- ◆ **Parc** : l'occupation du parc
 - * coordonnées (x,y) de l'emplacement : 5,19
 - * identifiant du container: YABB-CHARL-A1B2C3
 - * flag l'état de l'emplacement : 2
 - * la date éventuelle de réservation (sous forme de chaîne de caractères); 15/9/2020
 - * la date d'arrivée (sous forme de chaîne de caractères) : 17/9/2020
 - * le poids (contenant + contenu) : 15.6
 - * la destination du container : Lourdes
 - * le fait qu'il atteindra cette destination par Bateau ou par Train : Bateau
- ◆ **Sociétés** : les différentes sociétés qui traitent avec la plate-forme :
 - * identifiant société;
 - * nom du contact;
 - * email du contact;
 - * numéro de téléphone de ce contact;
 - * adresse de la société (simple chaîne de caractères).
- ◆ **Containers**: les différents containers qui transitent ou ont transité par le parc :
 - * identifiant container;
 - * identifiant société propriétaire;
 - * nature du contenu;
 - * capacité
 - * dangers particuliers.
- ◆ **Transporteurs**: les différents transporteurs (routiers, fluviaux ou ferroviaires) qui ont traité avec la plate-forme :
 - * identifiant transporteur (immatriculation dans le cas d'un camion);
 - * identifiant société propriétaire;
 - * capacité
 - * caractéristiques techniques (sous forme d'une simple chaîne de caractères).
- ◆ **Mouvements** : tout mouvement d'un container en entrée ou en sortie :
 - * identifiant mouvement
 - * identifiant container
 - * identifiant transporteur entrant
 - * date arrivée
 - * identifiant transporteur sortant
 - * poids (contenant + contenu);
 - * date départ
 - * destination
- ◆ **Destinations** : les destinations desservies par la plate-forme :
 - * ville
 - * distance par bateau (depuis la plate-forme)
 - * distance par train
 - * distance par route

2.2 La base BD COMPTA

Les données comptables d'Isil-PFM se trouvent dans une base de données MySQL BD_COMPTA. Elle doit contenir les informations utiles concernant tout ce qui touche les mouvements financiers de la plate-forme (correspondant à l'utilisation des containers) et les salaires du personnel.

- ◆ **Personnel** : les membres du personnel de la plate-forme :
 - * identifiant matricule
 - * nom
 - * prénom
 - * login
 - * mot de passe
 - * adresse e-mail interne
 - * fonction (manutentionnaire, chef d'équipe, préposé(e), chef de poste, comptable, chef-comptable, directeur(rice), ...)
- ◆ **Items Facture** : une ligne de facture correspondant à un mouvement
 - * identifiant item (compteur)
 - * identifiant facture;
 - * identifiant mouvement;
 - * identifiant container;
 - * destination
 - * destination;
 - * prix (hors TVA);
- ◆ **Factures** : pour une société donnée, on facture tous les mouvements du mois écoulé; outre son numéro et ses Item_Factures, une facture comportera
 - * l'identifiant de la société;
 - * le mois et l'année considérés;
 - * le montant total hors TVA;
 - * le montant total à payer (donc TVA comprise);
 - * un flag "facture validée"
 - * l'identifiant du comptable valideur
 - * un flag "facture envoyée"
 - * le moyen d'envoi de la facture (mail ou papier)
 - * un flag "facture payée"
- ◆ **Tarifs** : prix à payer pour un emplacement et sa manipulation (le prix n'est pas le même selon qu'interviennent un camion et/ou un bateau et/ou un train - on peut se référer à <http://www.portdeliege.be/fr/tarification>) :
 - * identifiant prix
 - * type (occupation emplacement ou manipulation)
 - * matériel utilisé (pont roulant pour occupation, camion, bateau ou train pour manipulation)
 - * flag produit pétrolier
 - * date dernière mise à jour.

3. Analyse Big Data: les informations stratégiques de décision

3.1 Quelles données pour quelles variables ?

Il s'agit d'implémenter dans chacune des applications visibles sur le schéma et expliquées ci-dessous la possibilité de réaliser

- ◆ une exploration de données sur un ensemble de données d'exploitation - pour ne pas rendre ce TP démesuré, on demande d'implémenter une ACP et une ACM.
- ◆ un test d'hypothèse sur un ensemble de données se trouvant provenant des données d'exploitation - pour ne pas rendre ce TP démesuré, on demande d'implémenter une REG-CORR à 3 variables explicatives au plus et une ANOVA2.

Les questions que l'on peut se poser

Le tableau ci-dessous donne une idée de ce qui est attendu - la liste des questions n'est bien sûr pas exhaustive.

question posée	type de test pour analyser	fonction R ?

...		

3.2 Génération des données de simulation

Comme cela se pratique souvent dans le milieu professionnel, les tests préliminaires s'effectueront sur des données simulées en nombre important. La génération de ces données sera réalisée avec R et/ou Excel.

4. Développement: Les applications fournies aux décideurs

Un serveur **Serveur_Data_Analysis** (écrit en Java) est chargé de réaliser

- ♦ les accès simples aux données (avec JDBC - ce n'est pas vraiment ce qui nous intéresse le plus ici);
- ♦ les requêtes d'exploration des données de type **Analyses factorielles** (ACP, ACM, AFC, CAH, K-means) ou **Tests d'hypothèses classiques** (ANOVA 1 et 2, REG-CORR simple ou multiple, ANCOVA, CHI-CARRE); pour ce type de requête, le serveur utilise les services d'un serveur **RServe**; la synthèse du résultat de ces requêtes est écrite dans BD_DECISIONS.

Ces accès et requêtes sont formulés par trois types d'applications clientes:

- 1) une application installée **Application_BigData** qui
 - ♦ utilise des techniques de cartes à puces;
 - ♦ intègre la reconnaissance de client par carte d'identité électronique.
- 2) une application mobile Android **Appli_BigData** dont la reconnaissance de client s'effectue par un mécanisme OTP.
- 3) une application Web **WebAppli_BigData** dont la reconnaissance de client s'effectue par un digest simple sur mot de passe.

5. Technologie: Intégration des cartes à puces

L'accès à l'utilisation de l'application installée ne peut être obtenu qu'après **une identification lourde** : selon le type de l'utilisateur

- ♦ pour les employés de la société : une **carte à puce** donnée à chaque employé contient un code d'accès et un compteur du nombre d'accès déjà effectués; l'authentification consiste à envoyer au serveur un digest SHA-1 ou SHA-256 calculé par l'application sur le code, le login et le password, accompagné de la valeur du compteur; le serveur, au moyen des informations trouvées dans BD_COMPTA, calcule un deuxième digest pour comparaison, et compare également la valeur du compteur d'accès mémorisé dans cette base; l'accès est autorisé en cas de concordance des deux éléments.
- ♦ pour les personnes extérieures : l'authentification est basée sur une authentification lourde basée sur leur **carte d'identité électronique** : ils signent un challenge.

6. Recherche : les SMS sous Android

6.1 Les SMS

Comment envoyer/recevoir un **SMS** depuis une application Android. Application : **Appli_BigData** envoie un SMS pour permettre une authentification par code.

Pour démarrer, voir par exemple :

http://www.tutorialspoint.com/android/android_sending_sms.htm

6.2 Les mails sous Android

Un utilisateur reçoit par e-mail (si il le désire) envoyé par l'application Android les résultats d'analyse les plus importants.

Comment envoyer un mail dans une application Android ? Voir par exemple :

http://www.tutorialspoint.com/android/android_sending_email.htm

6.3 La sécurité logicielle sous Android

Comment envoyer des mails sécurisés ?

i) Spongy Castle : Bouncy Castle adapté à Android

Voir par exemple :

<https://rstyley.github.io/spongycastle/>

ii) Comment gérer les clés selon une architecture non hiérarchique PGP (c'est-à-dire basée sur la confiance entre utilisateurs, sans véritable CA) ? Une appli populaire : Android Privacy Guard, disponible sur Play Store (chercher "APG").

Voir par exemple :

<http://iceblog.free.fr/index.php?article95/tuto-crypter-signer-emails-android-windows>

https://info.securityinabox.org/fr/k9_apg_principal

<https://www.youtube.com/watch?v=b14h3aPpv3Y>