ia·rbre

Mika **INISAN**
LIRIS

23 octobre 2025

Synthèse :
Data pipelines
Chaînes de traitement de données

MÉTROPOLE GRAND LYON

TelesCoop

Université Lumière Lyon 2

FRANCE 2030

BANQUE des TERRITOIRES | Caisse des Dépôts GROUPE

Mika **INISAN**
LIRIS

Sommaire

Mika **INISAN**
LIRIS

01

Objectifs de la présentation

Mika **INISAN**
LIRIS

Objectifs de la présentation

Restituer les principaux enseignements de la synthèse :

approches, défis, tendances, meilleures pratiques...

Partager une base commune de connaissance :

enjeux, problématiques techniques et stratégiques...

Être un point de départ favorisant les échanges pour ajuster et affiner le projet :

questions, idées, compléments, approfondissements, confrontation avec les attentes...

Mika **INISAN**
LIRIS

02

Objectifs de la synthèse

Mika **INISAN**

Objectifs de la synthèse

- Positionner la solution développée par rapport à l'état de l'art (convergence, divergence, opportunités d'ajustement, axes différenciants)
- Tirer parti des travaux existants pour accélérer les développements et éviter de repartir de zéro
- Identifier des leviers d'innovation et fonctionnalités, susciter des idées nouvelles
- Réduire l'incertitude, éclairer les décisions, valider et justifier les choix a priori et a posteriori
- Anticiper les futurs besoins, démontrer l'impact

03

Méthode de recherche employée

Mika **INISAN**

Méthode de recherche employée

01
Identifier le problème

02
Formaliser, contextualiser, problématiser

03
Faire l'état de l'art

04
Émettre un positionnement pour IA.rbre

05
Développement d'un POC (si nécessaire)

06
Produire les livrables de recherche

04

Rappels sur IA.rbre et ses objectifs

Mika **INISAN**
LIRIS

Rappels sur IA.rbre et ses objectifs

Rappels sur IA.rbre

- **Finalités** :
  - Aider à la décision
  - Servir les services métiers et le grand public grâce aux besoins généralisables pris en compte
- **Propriété intellectuelle** : commun numérique
- **Formes** :
  - Outils de visualisation et de scénarisation
  - Produit de données
- **Adjectifs** : frugalité, innovation, basé sur la science, open data / source / innovation / science, données FAIR

Mika **INISAN**
LIRIS

Rappels sur IA.rbre et ses objectifs

Objectifs fonctionnels

- Intégrer les données territoriales disponibles pour agrandir la connaissance et la pertinence de l'analyse :

  Sources : acteurs multiples du territoire, publics et privés.
  DataGrandLyon, IGN, exploitant de réseaux (TCL, ENEDIS, SFR...)...

  Contenu : données de réseaux, de chantiers, d'occupation de sols, inventaire du végétal urbain existant...

  Moyen d'acquisition : en ligne, issues d'une démarche DICT, locales à la métropole...

  Types : modèles, données territoriales, vérité terrain, données issues de modèles...

  Formats : raster, vectoriel, GeoTiff, Shapefile, GeoPackage, CityGML, CSV, HDF5, MLmodel...

- Transférer la connaissance d'une métropole ayant beaucoup de données, sur une métropole en ayant moins

Mika **INISAN**
LIRIS

Rappels sur IA.rbre et ses objectifs

Objectifs fonctionnels

- Faire des analyses complexes dont ML, AI :

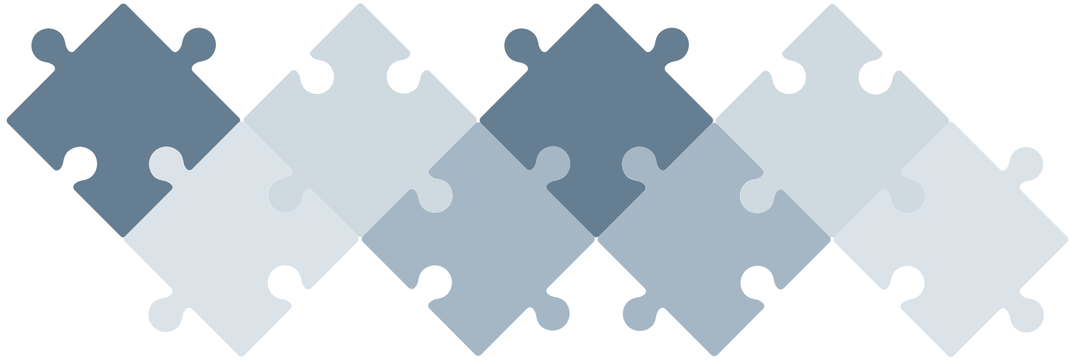
  - Amélioration des données par croisement
  - Génération de nouvelles données à partir de celles existantes
  - Réunification, réconciliation des données
  - Croisement des données
  - Substituer les données pour le calcul des facteurs
  - Calcul des facteurs de faisabilité
  - Transformer les données ingérées en indices intermédiaires
  - Pondération possiblement automatique des facteurs
  - Génération des calques basiques (plantabilité) et thématiques
  - Exploration de scénario, données dans le temps, évolution temporelle des villes
  - A/B testing de modèles

Mika **INISAN**
LIRIS

Rappels sur IA.rbre et ses objectifs

Objectifs non fonctionnels

Disponibilité
Accessibilité
Qualité
Fiabilité
Réplicabilité
Réutilisabilité
Reproductibilité
Explicabilité
Transparence



Adaptabilité
Flexibilité
Comparabilité
Conformité
Traçabilité
Scalabilité
Tolérance à l'erreur
Facilité de gestion
Performance

Mika **INISAN**
LIRIS

05

Notions sur les data pipelines

Mika **INISAN**
LIRIS

Notions sur les data pipelines

Introduction

Structure logicielle qui permet le déplacement et la manipulation systématiques de données provenant de sources potentiellement multiples et hétérogènes, vers des destinations variées.

Se compose d’un ensemble d'étapes, possiblement automatisées, interconnectées à travers lesquelles les données passent de manière séquentielle, la sortie d'une étape servant d'entrée pour la suivante.

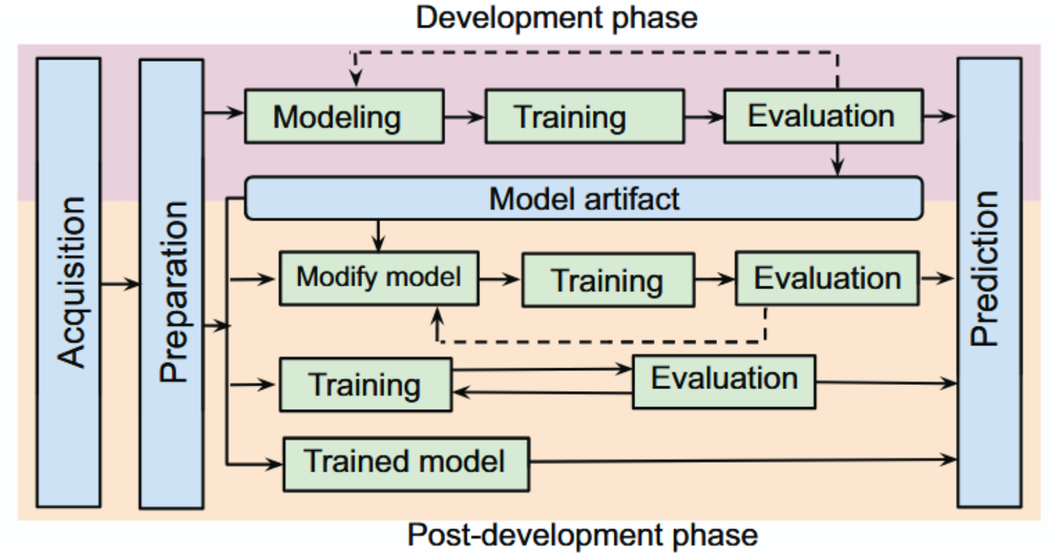
Pas de standardisation et de terminologie largement acceptée dans la littérature scentifique [1]

[1] S. Biswas, M. Wardat, and H. Rajan, “The art and practice of data science pipelines,” *Proceedings of the 44th International Conference on Software Engineering*, pp. 2091–2103, May 2022

Mika **INISAN**
LIRIS

Notions sur les data pipelines

Exemple et représentation



Exemple que nous détaillerons par la suite

Représentation : graphe orienté acyclique
(Directed acyclic graph, DAG)

Mika **INISAN**
LIRIS

Notions sur les data pipelines

Caractéristiques

Formats des données	Non structurés, semi-structurés, structurés
Sources des données	Locales, distantes ; centralisées, distribuées ; API, fichiers, base de données, origines d'un crawling ou scraping...
Destinations	Stockage, applications (outils de visualisation...), autre data pipelines, modèles d'apprentissage automatique...
Modes de déclenchement	Manuel, programmé ; ponctuel, récurrent, en réponse à des stimuli basés sur des événements
Modes d'ingestion de données	Par lots (« batch »), continu (« streaming »), hybride
Modes d'exécution	Séquentiel, parallèle ; centralisé, distribué

Mika **INISAN**
LIRIS

Notions sur les data pipelines

Étapes

Extraction	Transformation	Chargement	Filtrage	Fusion
Sélection	Ingestion	Acquisition	Exploration	Agrégation
Validation	Enrichissement	Stockage	Visualisation	...

Étapes spécifiques à l'analyse (ML, AI...) :

Feature engineering	Modélisation	Entraînement	Évaluation	...

Mika **INISAN**

Notions sur les data pipelines

Rôles et objectifs

Les chaînes de traitement de données, ou data pipelines, occupent une place de plus en plus centrale au sein des systèmes de gestion de données contemporains grâce à leur **rôle structurant**. Elles constituent la fondation des activités de traitement, d'analyse et de prise de décision.

Elles permettent de **contrôler toutes les opérations** liées aux données et d'**orchestrer l'ensemble du flux** (traitement, transfert, stockage) de manière rationalisée de la source à la destination.

En **remplaçant des opérations manuelles** susceptibles d'introduire des erreurs **par des processus systématiques** et reproductibles.

Elles **décomposent les analyses complexes** de grands ensembles de données **en une série de tâches plus simples**.

Les propriétés des implémentations des différentes étapes visent à encourager la réutilisation, la composition flexible et la configurabilité pour des usages spécifiques.

Mika **INISAN**
LIRIS

Notions sur les data pipelines

Rôles et objectifs

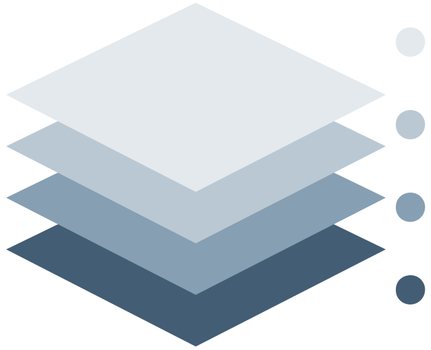
- Atténuer les goulots d'étranglement et les délais
- Renforcer la scalabilité (notamment face au volume de données)
- Simplifier la conception et le déploiement des services de traitement des données
- Reproductibilité, traçabilité, fiabilité, tolérance aux pannes
- Améliorer la cohérence des systèmes d'information
- Augmenter la vitesse de bout en bout, l'efficacité
- Réduire la latence dans le développement des produits de données
- Accès temps réel à l'information actualisée
- Une meilleure prise de décision

06

Meilleures pratiques

Mika **INISAN**
LIRIS

Meilleures pratiques



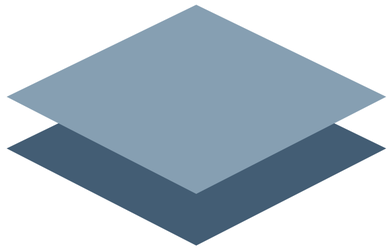
- Contrôle de la dette technique
- Évolutivité, flexibilité, robustesse
- Faciliter maintenance, entretien, dépannage, documentation...
- Meilleure testabilité, sécurisabilité, reproductibilité, réutilisabilité

Concevoir de manière modulaire (séparer, isoler clairement les responsabilités au sein des étapes, de minimiser les redondances)

- *Une attention particulière doit être portée sur l'interface entre les étapes*
- *L'utilisation d'un métamodèle peut simplifier la phase de conception*

Mika **INISAN**
LIRIS

Meilleures pratiques



- Recentrer l'effort sur des missions à plus forte valeur ajoutée que celles de gestion de données
- Minimiser l'erreur humaine, la charge de travail et les coûts liés au volume, la vélocité et la variété des données

Maximiser l'automatisation des étapes et tâches

Mika **INISAN**
LIRIS

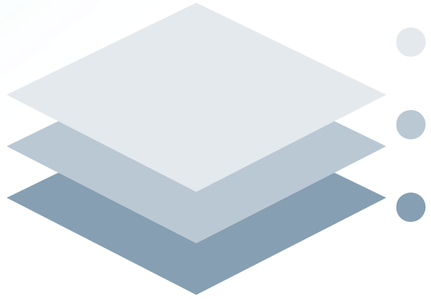
Meilleures pratiques

Maximiser la tolérance aux défaillances et le recouvrement (matérielles, algorithmiques, erreurs métier)

Mettre en place :
- Un monitoring (latence, vitesse de transfert, taux d'erreur ; logs d'erreurs, rapports, tableaux de bord…)
- Une détection de défaillance
- Des règles de validation et de réjection entière ou partielle (exemple : espace de staging)
- Des stratégies d'atténuation (exemple : versionnement)
- La levée d'alerte et d'événements automatiques

Mika **INISAN**

Meilleures pratiques



- Reproductibilité, traçabilité
- Pouvoir d'appliquer des traitements différents à l'avenir (e.g., comparaisons)
- Éviter de refaire l'intégralité des traitements en cas d'erreur ou de problème de qualité intermédiaire

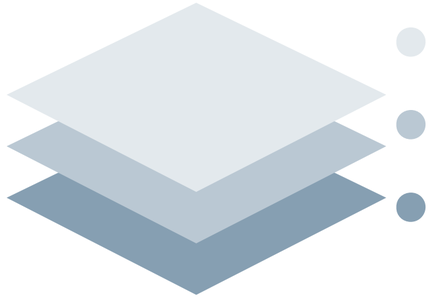
Stocker les données initiales et les données intermédiaires issues de transformation en plus des données finales

Séparer les stockages initiaux, intermédiaires et finaux

Minimiser la duplication qui amène à des complexités de gestion et une incertitude sur l'état des données comme l'actualisation

Mika **INISAN**
LIRIS

Meilleures pratiques



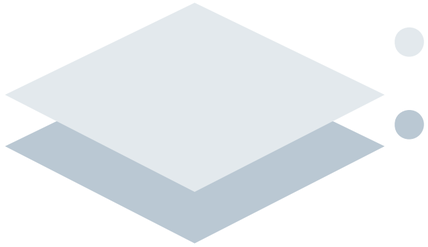
- Minimiser les coûts d'ingestion et maximiser la frugalité
- Traçabilité
- Éviter l'explosion des temps d'exécution, de l'espace de stockage pour des données qui sont potentiellement inutilisables

Faire un suivi rigoureux de l'origine des données (e.g, via des métadonnées)

Faire du data profiling pour choisir précisément les données à intégrer, par exemple, les données les plus récentes par rapport à celles déjà intégrées (intégration incrémentale)

Mika **INISAN**
LIRIS

Meilleures pratiques



- Être flexible à la variété des données (formats, structures...)
- Supporter l'évolution

Choisir une architecture qui privilégie une compatibilité et une évolution automatique du schéma

Choisir des outils pérennes dans le temps, faciles d'utilisation, faciles à l'intégration, offrant un bon monitoring, sécurisés et scalables

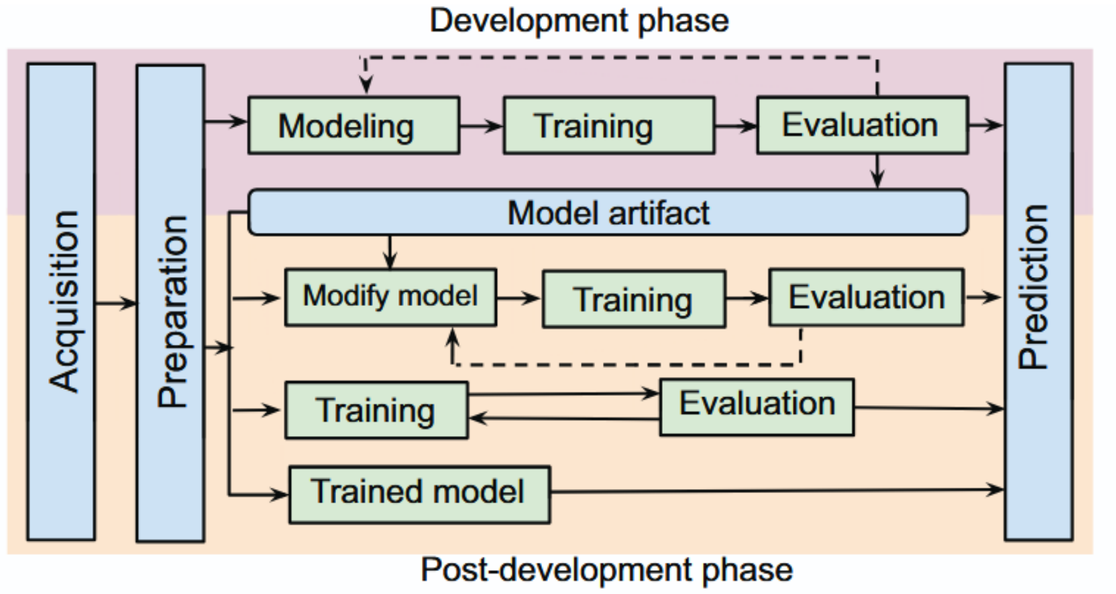
07

Data pipeline pour IA.rbre

Mika **INISAN**
LIRIS

Data pipeline pour IA.rbre

Architecture pour la réplicabilité



Architecture représentative des projets d'envergure [1]

- **Cas de donneés abondantes** : entraîner les modèles sur des données prétraitées, stocker les artifacts afin de prédire les facteurs à partir de ceux-ci
- **Cas de données de faible qualité, insuffisantes** : réutiliser directement des modèles déjà entraînés sur une autre ville pour effectuer des prédictions
- **Cas intermédiaire** : surentraîner des modèles existants sur les données locales, puis effectuer la prédiction

[1] S. Biswas, M. Wardat, and H. Rajan, "The art and practice of data science pipelines," *Proceedings of the 44th International Conference on Software Engineering*, pp. 2091–2103, May 2022

Mika **INISAN**
LIRIS

Data pipeline pour IA.rbre

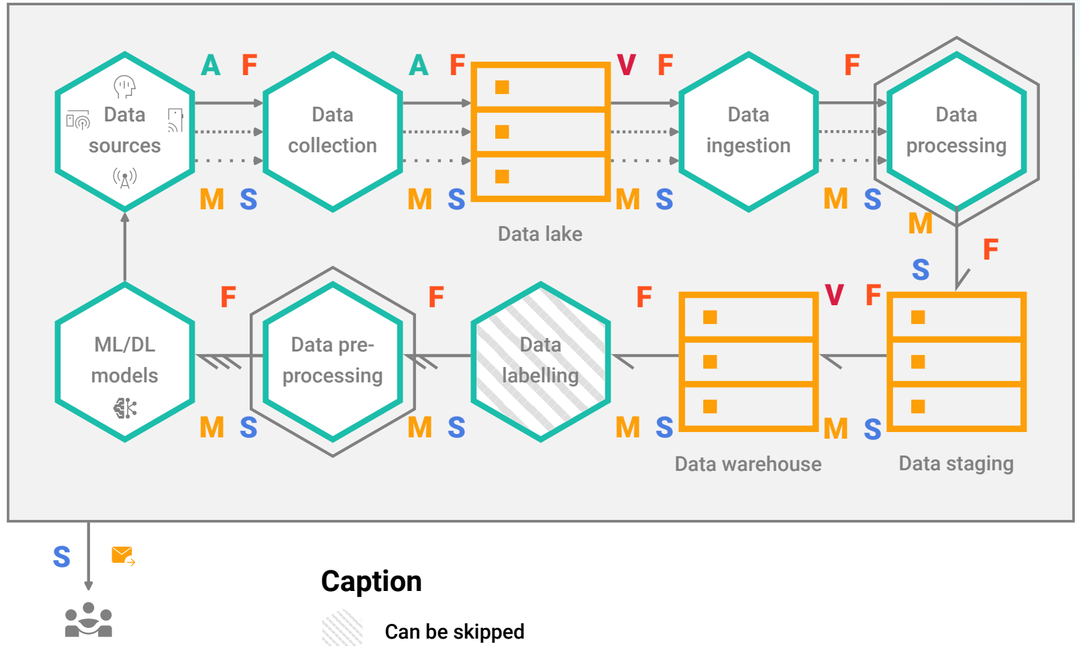
Architecture pour la tolérance aux failles

Nodes

Capabilities	Types
Data generation	Human generated, Machine generated
Data collection	
Data ingestion	
Data processing	
Data storage	
Data labelling	
Data preprocessing	
Data reception	ML/DL models, Dashboard

Connectors

Capabilities		Types
Data transportation		Continuous, Intermittent, Batches, Structured data, Labelled data, Preprocessed data
Mitigation	M	
Monitoring	F	
Authentication	A	
Sending alarm	S	
Validation	V	



Modèle conceptuel de la data pipeline et son métamodèle selon [2]

[2] A. Raj, J. Bosch, H. H. Olsson, and T. J. Wang, "Modelling Data Pipelines," *46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pp. 13–20, Aug. 2020

Mika **INISAN**

LIRIS

Des questions ? Des idées ?

- Synthèse complète sur le *GitHub*

  VCityTeam/UD-IArbre-Research/data-pipelines

- Partagez vos retours directement via les Issues *GitHub*

Merci !