RAPPORT DE CONCEPTUALISATION
Système intégré d’analyse automatique de l’agenda médiatique français

Ce document présente une architecture complète, modulaire et scalable permettant la collecte, l’analyse et l’agrégation quotidienne de l’actualité provenant de trois sources :
- chaînes TV françaises d’information (via RSS),
- presse écrite (via RSS),
- réseaux sociaux.

Le système est conçu pour fonctionner **une fois par jour**, après la **dernière ingestion**, afin d’optimiser les coûts de calcul.

1. OBJECTIF GÉNÉRAL

L’objectif du système est de produire quotidiennement :

- les **top 10 sujets** les plus traités,
- les **top 10 mots clés** les plus évoqués,
- la comparaison des lignes éditoriales,
- les narratifs dominants,
- la détection des biais médiatiques,
- l’identification des sujets peu ou non couverts,
- la mesure de la durée de vie médiatique d’un thème,
- la détection des emballements médiatiques ("spikes").

L’ensemble est historisé dans une base structurée permettant des analyses temporelles.

2. SOURCES MÉDIATIQUES COUVERTES

2.1. Télévision (RSS)
- BFMTV
- CNEWS
- Franceinfo
- TF1 Info / LCI

2.2. Presse écrite (RSS)
15+ journaux nationaux, dont : Le Monde, Le Figaro, Libération, Le Parisien, 20 Minutes…

2.3. Réseaux sociaux
- Reddit (subreddits FR)
- Mastodon (instances FR publiques)
- YouTube News (titres vidéos “actualité France”)

Ces trois sources garantissent une couverture complète du paysage informationnel français.

3. ARCHITECTURE GLOBALE

Le système repose sur trois couches :

3.1 DATA LAKE
- **Raw** : données brutes (RSS XML, JSON réseaux sociaux).
- **Cleaned** : normalisation + nettoyage NLP.
- **Curated** : données enrichies (mots clés, sujets, embeddings…).

3.2 DATA WAREHOUSE
Tables analytiques organisées par date, source et thème :
- articles
- articles_clean
- keywords_daily
- topics_daily
- narratives_comparison
- media_bias
- spikes
- topic_lifetime

3.3 SERVICES
- API FastAPI
- Dashboard (Metabase, Superset, PowerBI)

4. PIPELINES D’INGESTION (OPTIMISÉES)

Le seul élément exécuté plusieurs fois par jour est **l’ingestion**, car les flux RSS sont dynamiques.

4.1 Fréquence d’ingestion optimisée
Pour réduire les coûts computationnels :

Télévision (RSS TV)
- **2 fois par jour** : 13h et 23h

Presse écrite
- **1 fois par jour** : 23h

Réseaux sociaux
- **2 fois par jour** : 12h et 22h

4.2 Logique d’ingestion
Chaque ingestion :
- lit les flux
- identifie les nouveaux articles
- déduplique
- stocke dans `raw`

Aucune re-ingestion inutile.

5. PIPELINE UNIQUE DE TRAITEMENT (POST-INGESTION)

Une fois la **dernière ingestion terminée**, une **chaîne unique** de traitement démarre :

1. Nettoyage NLP
2. Extraction entités nommées
3. Calcul top 10 mots par média et global
4. Calcul top 10 sujets via TF-IDF ou clustering
5. Analyse des biais médiatiques
6. Comparaison des narratifs entre chaînes
7. Mesure de la durée de vie des sujets
8. Détection des emballements médiatiques
9. Génération des agrégats journaliers
10. Mise à jour API + Dashboard

Toutes ces tâches ne tournent **qu'une seule fois par jour**, automatiquement.

6. MODULES ANALYTIQUES INTÉGRÉS

6.1. Analyse des biais médiatiques
Comparaison de l’angle, du vocabulaire, des acteurs mentionnés.

6.2. Carte des invisibles
Sujets présents dans réseaux sociaux mais absents des médias TV/presse.

6.3. Narratifs
Comparaison vectorielle des descriptions d’un même sujet entre chaînes (CNEWS vs BFMTV vs LCI vs Franceinfo).

6.4. Analyse thématique
Répartition des thèmes majeurs : sécurité, immigration, économie, climat…

6.5. Durée de vie médiatique
Nombre de jours consécutifs où un thème reste d’actualité.

6.6. Emballements médiatiques
Détection de pics soudains (spikes) dans la fréquence d’un sujet.

6.7. Distinction factualité / opinion
Analyse du niveau d’éditorialisation des contenus.

7. IMPACT ET UTILITÉ DU SYSTÈME

Ce projet permet :

- de cartographier l’agenda médiatique français,
- de révéler les biais éditoriaux,
- de comparer les lignes des chaînes TV,
- d’exposer les omissions médiatiques,
- d’offrir un outil transparent aux citoyens et chercheurs,
- de produire un historique consultable à vie.

C’est un outil puissant de veille médiatique et d’analyse sociopolitique.

8. AMÉLIORATIONS FUTURES

Non incluses dans cette version mais prévues :

- Analyse des **conflits d’intérêts** entre médias / entreprises / politiques.
- Comparaison **France vs Étranger** (BBC, ARD, CNN…)

Ces modules demandent financement, car ils nécessitent des bases supplémentaires.

9. CONCLUSION

Ce projet constitue l’un des systèmes Data Engineering les plus complets pour l'analyse de l’actualité :

- ingestion automatisée multi-sources,
- traitement NLP avancé,
- analyses thématiques, narratives et temporelles,
- dashboard et API centralisés,
- architecture scalable et modulaire.

Il offre une vision unique et objective du fonctionnement médiatique français.