Projet de fin d'études – Résumé

L'objectif de ce document est de proposer un résumé/guide simple et concis des attentes du projet de fin d'étude du M2 Data Engineering au sujet de concevoir, développer et déployer une solution de traitement des données massives.

1) Résumé du projet : « Concevoir, développer et déployer une solution de traitement des données massives"

Ce projet consiste à concevoir, développer et déployer une solution de traitement des données massives (big data). Les étudiants devront mettre en place une architecture distribuée capable de traiter et analyser d'importants volumes de données en « temps réel » ou en batch. Ils devront également intégrer différentes sources de données, optimiser les pipelines, automatiser les tests et le déploiement.

Objectif finale :
Soutenance Orale et démonstration de la solution par binôme devant un jury et les autres apprenants.
Rapport écrit en binôme examiné par un jury.

Pour plus de détail, veuillez trouver en bas de ce document le descriptif complet exhaustif du projet.

2) Checklist technique pour valider les points du sujet :

- Conception de l'architecture distribuée :
  - Réaliser une veille technologique sur les frameworks big data
  - Choisir les environnements logiciels de traitement adaptés
  - Prévoir la résilience et la scalabilité de l'architecture

- ~~Streaming de données :~~
  - ~~Identifier les solutions de streaming temps réel~~
  - ~~Traiter les données en micro-batch sur une période donnée~~
  - ~~Traiter les données au fur et à mesure en temps réel~~

ps : la partie streaming étant complexe à mettre en place dans le scope du projet. Je vous recommande de ne pas la traiter (sujet déjà discuté avec l'équipe pédagogique)

- Transformation de données :
  - Utiliser des outils d'informatique décisionnelle (ETL...)
  - Intégrer des données de sources variées
  - Manipuler les données multidimensionnelles (OLAP Cubes)

- Pipelines de données :
  - Intégrer différentes sources de données
  - Optimiser les performances (monitoring, containerisation etc.)
  - Permettre l'ajout de nouvelles sources de données

- Automatisation :
  - Containeriser/pipelines
  - Mettre en place l'intégration et le déploiement continus (CI/CD)
  - Automatiser les différents tests (unitaires, d'intégration, etc.)

3) Plan proposé pour les étudiants :

Étape 1 : Étude des besoins et veille technologique

- Comprendre les besoins du client (fictif si hors cadre de l'alternance)
- Réaliser une veille sur les frameworks big data, streaming, ETL, etc.

Étape 2 : Conception de l'architecture distribuée
- Dimensionner les ressources nécessaires
- Choisir les technologies adaptées (Hadoop, Spark, Kafka, etc.)
- Définir l'architecture résiliente et scalable

Étape 3 : Développement des pipelines
- Implémenter l'ingestion des différentes sources
- Développer les transformations ETL
- Mettre en place le traitement batch
-

Étape 4 : Optimisation et monitoring
- Optimiser les performances des pipelines
- Mettre en place les indicateurs de monitoring

Étape 5 : Automatisation
- Containeriser
- Mettre en place l'intégration et déploiement continus
- Automatiser les différents tests

Étape 6 : Oral
- Préparation à la soutenance orale

4) Exemples d'idées de projet :

1. Système de recommandation pour une plateforme de streaming
- Analyser les données d'utilisation (vues, likes, recherches, etc.) en mode batch
- Entraîner périodiquement des modèles de machine learning pour faire des recommandations
- Mettre à jour régulièrement les recommandations selon l'activité récente

2. Détection de fraudes pour une plateforme e-commerce
- Ingérer des données de transactions, de livraisons, d'évaluations en mode batch
- Appliquer des règles métiers et de l'IA pour détecter les fraudes a posteriori
- Générer des rapports et des alertes pour les analystes

3. Plateforme d'analyse de données IoT industrielles
- Collecter périodiquement des données de capteurs de machines, équipements
- Analyser les données historiques pour détecter les anomalies, prédire les pannes
- Optimiser la maintenance prédictive planifiée des équipements

4. Observatoire de données environnementales
- Agréger des données de stations météo, satellites, capteurs de qualité de l'air
- Analyser en mode batch les tendances, événements extrêmes liés au climat
- Fournir des rapports et visualisations de données historiques

BC2 : Concevoir, développer et déployer une solution de traitement des données massives			
A1 : Conception de solution de traitement de données A2 : Développement de solution données A3 : Déploiement d'une solution de traitement des données massives A4 : Transformation de données issues de sources différentes	BC2.1. Concevoir en s'appuyant sur une veille technologique et mettre en œuvre une architecture distribuée répondant au besoin du client pour traiter les données massives en entreprise en utilisant les technologies de traitement	E2.1. Étude de cas réalisée en amont Conception d'une application de traitement distribué. E2.2. Évaluation : Présentation de l'architecture par binôme devant un jury et les autres apprenants.	*Une architecture de traitement distribué de données est proposée* *C1 :* Une veille technologique des Framework big data est réalisée *C2 :* L'architecture proposée répond au besoin exprimé par le client en termes de traitement *C3 :* Les ressources mobilisées sur le cluster de calcul en termes de puissance de calcul et de mémoire sont suffisantes *C4 :* Les environnements logiciels de traitement de données sont adaptés *C5 :* La solution proposée permet de traiter (en batch) et d'analyser l'ensemble de données disponibles *C6 :* L'architecture proposée est résiliente à la panne du système *C7 :* La solution assure une scalabilité horizontale
A5 : Optimisation de pipelines A6 : Application de systèmes appropriés en réponse à une demande A7 : Création et automatisation de tests	BC2.2. Implémenter un système distribué en utilisant des technologies de streaming identifiées à partir d'une veille *pour traiter des données* sur une période précise ou en temps quasi réel	E2.3. Étude de cas réalisée en amont Conception d'une application de traitement distribué. E2.4. Evaluation : Soutenance Orale et démonstration de la solution par binôme devant un jury et les autres apprenants. Rapport écrit en binôme examiné par un jury.	*Une solution de streaming distribué est mise place :* *C1 :* Une veille technologique permet d'identifier et mobiliser les solutions de streaming adaptées *C2*: La solution proposée permet de traiter (en micro batch) et d'analyser l'ensemble de données collectées sur une période données (en secondes ou millisecondes) *C3 :* La solution proposée permet de traiter (en temps réel) et d'analyser l'ensemble de données au fur et à mesure de leur disponibilité.

EXPERT EN INGENIERIE DE DONNEES - Efrei

	BC2.3. Transformer les données provenant de différentes sources en prenant en compte la variété de données pour faire de l'analytique à échelle (intégration, formatage, manipulation, stockage données multidimensionnelles)	E2.5. Étude de cas réalisée en amont Conception d'une solution d'intégration et de transformation de données. Évaluation : E2.6. Evaluation : Soutenance Orale et démonstration de la solution par binôme devant un jury et les autres apprenants. Rapport écrit en binôme examiné par un jury.	*Les données sont extraites, transformées et mise à disposition :* *C1* : L'utilisation des outils d'informatique décisionnelle permet de transformer et formater les données selon un format précis *C2* : Les outils utilisés permettent l'intégration de données *C3* : Les CUBES permettent la représentation et la manipulation des données multidimensionnelles
	BC2.4. Optimiser la performance des pipelines en utilisant les techniques d'intégration et de mise en scène adéquates pour le traitement des données massives	E2.7. Étude de cas réalisée en amont E2.8. Evaluation : Soutenance Orale et démonstration de la solution par binôme devant un jury et les autres apprenants. Rapport écrit en binôme examiné par un jury.	*Les pipelines sont développés* *C1* : Les pipelines développés permettent d'intégrer des sources de données provenant de différentes sources *C2* : - un accès rapide aux données - L'ajout de nouvelles sources de données *C3* : Les indicateurs mis en place permettent de mesurer la performance de transfert de données *C4* : L'ordonnanceur mobilisé est adapté aux technologies utilisées *C5* : La performance du pipeline de données est monitorée *C6* : Les pipelines permettent l'ajout de nouvelles sources de données

EXPERT EN INGENIERIE DE DONNEES - Efrei

	BC2.5. Automatiser la création, les tests, l'intégration et le déploiement des pipelines de données en s'appuyant sur une veille technologique qui permet d'identifier et de mobiliser les solutions pour maximiser l'efficience et réduire le 'time to market' tout en utilisant les technologies de containerisation et d'ordonnancement.	E2.9. Étude de cas réalisée en amont Automatiser la création, les tests, l'intégration et le déploiement des pipelines de données E2.10. Evaluation : Soutenance Orale et démonstration de la solution par binôme devant un jury et les autres apprenants. Rapport écrit en binôme examiné par un jury.	*La création, les tests, l'intégration et le déploiement des pipelines de données sont automatisées :* *C1* : La veille technologique permet d'identifier et mobiliser les technologies de contrôle de version, containerisation, testing, intégration et déploiement continue adaptées *C2*: Les pipelines développés sont containerisés *C3* : Les outils d'orchestration permettent l'intégration et le déploiement continu et rapide du produit *C4* : Les différents types de tests mobilisés permettent de vérifier et valider la solution déployée