Bonjour Héloïse et Osman,

Voici le document récapitulatif des attendus pour le projet MapReduce, avec une grille d'évaluation concise.

pouvez-vous transmettre ? merci !

---

1. Attendus du projet

Fonctionnalités obligatoires

- **Protocole TCP** : Le master synchronise les workers (envoie les adresses des nœuds et les splits à traiter), mais **ne transite jamais par les données**. Le shuffle est direct entre workers via une fonction de hachage (clé modulo nombre de machines).
- **Expérience Loi d'Amdahl** :
    - Fixer la taille totale des données.
    - Tester sur un nombre croissant de nœuds (minimum 3 machines).
    - **Chronométrer uniquement** les phases map, shuffle et reduce (exclure déploiement et récolte des résultats).
    - Tracer une **courbe** du speedup et en déduire le taux de parallélisation global.
    - **Extrapolation linéaire** : Pour le point (1,1), si un split fait 300 Mo, extrapoler pour X Mo (ex. 3000 Mo = 10× le temps).
    - **Round-robin** : À partir de 2 nœuds, distribuer les splits en alternance (ex. : split 1 → nœud 1, split 2 → nœud 2, etc.).
- **Deuxième MapReduce** : Tri réparti des mots par fréquence, puis par ordre alphabétique. Répartition des fréquences entre les nœuds (ex. : nœud 1 traite 1-10, nœud 2 traite 10-20, etc.), avec envoi des min/max de fréquences.
- **Déploiement** :
    - Un seul SCP sur une machine (NFS partagé = déploiement automatique).
    - Démarrage des serveurs via SSH (max 5 SSH/min depuis une machine source, pas de limite depuis une machine perso).
    - **Précision** : Si script PowerShell/Windows, le signaler (test sous Linux par défaut).

Rapport et rendu

- **Rapport** (5 pages max) :
    - Description du système et du protocole.
    - Courbe et analyse de la Loi d'Amdahl.
    - **Données brutes** : Fichier séparé, avec documentation claire pour les récupérer (ex. : chemin, format, commande pour les extraire).
    - **Pas de code** dans le rapport, mais indiquer où trouver les fonctions clés.
    - **Reproductibilité** : Je dois pouvoir relancer votre système et obtenir les mêmes résultats. Plus c'est facile à reproduire, plus c'est valorisé.
- **Code** : Archive ZIP avec tout le code et les scripts de déploiement.

2. Grille d'évaluation

Critère	**Points**	**Détails**
Fonctionnalité de base	0-10	Protocole, synchronisation, shuffle direct, déploiement fonctionnel.
Expérience Loi d'Amdahl	0-4	Courbe, analyse du plateau, données brutes exploitables et bien documentées.
Deuxième MapReduce (tri réparti)	0-3	Tri par fréquence/alphabétique, répartition correcte des fréquences.
Bonus (langdetect)	+2	Analyse de la répartition des langues dans CommonCrawl.
Qualité scientifique	0-3	Analyse critique, limites du système, comparaison pertinente avec Hadoop.

Deadline : 19 novembre 2025 (dépôt sur Moodle IP Paris).

Pour toute question, n'hésitez pas à me contacter.

Cordialement,