III-les KPIs:

Performance Indicators qui pourraient être pertinents pour évaluer les performances et l'efficacité de système :

1. Temps de traitement moyen (MapReduce/Spark) : Mesure le temps moyen nécessaire pour traiter un ensemble de données à travers MapReduce ou Spark. Cela peut aider à évaluer la performance générale de votre système de traitement distribué.
2. Utilisation des ressources (CPU, mémoire) : Surveiller l'utilisation des ressources matérielles telles que le CPU et la mémoire pendant l'exécution des tâches MapReduce ou Spark. Cela peut aider à optimiser l'allocation des ressources et à identifier les goulets d'étranglement.
3. Taux d'échec des tâches : Mesure le pourcentage de tâches qui échouent lors de leur exécution. Cela peut indiquer des problèmes de code, de configuration ou de ressources qui doivent être résolus.
4. Débit de traitement des données : Mesure la quantité de données traitées par unité de temps (par exemple, octets par seconde). Cela peut être utilisé pour évaluer la capacité de traitement de votre système dans des conditions de charge différentes.
5. Temps de latence : Mesure le temps écoulé entre la soumission d'une tâche et la réception des résultats. Cela peut être critique pour les applications nécessitant des réponses en temps réel.
6. Efficacité de la compression des données : Mesure le taux de compression des données avant et après le traitement, ce qui peut avoir un impact significatif sur les performances et l'utilisation des ressources.
7. Équilibrage de charge : Surveiller la distribution des tâches sur les nœuds du cluster pour s'assurer qu'elles sont réparties de manière équitable et que les ressources sont utilisées efficacement.
8. Scalabilité : Mesure la capacité du système à maintenir ses performances à mesure que la taille des données ou la charge de travail augmentent.
9. Salaire moyen par entreprise : Calculé en groupant les données par le nom de l'entreprise et en calculant la moyenne du salaire pour chaque entreprise.
10. Affichage du schéma des données : Permet de comprendre la structure et les types de données présents dans le dataset.
11. Affichage des premières lignes du DataFrame : Donne un aperçu des données en affichant les premières lignes du DataFrame.
12. Comptage total des lignes dans le DataFrame : Donne le nombre total de lignes présentes dans le DataFrame.
13. Statistiques descriptives pour les colonnes "Salary" et "Salaries Reported" : Fournit des statistiques telles que la moyenne, l'écart type, le minimum, le maximum, etc., pour ces colonnes spécifiques.

14. Filtrage des lignes avec un salaire supérieur à 100 000 : Permet de sélectionner les lignes où la valeur de la colonne "Salary" est supérieure à 100 000.
15. Tri du DataFrame par salaire décroissant : Trie le DataFrame en fonction de la colonne "Salary" de manière décroissante.
16. Comptage du nombre d'emplois par emplacement : Calculé en groupant les données par emplacement et en comptant le nombre d'emplois pour chaque emplacement.
17. Comptage du nombre d'emplois par titre de poste : Calculé en groupant les données par titre de poste et en comptant le nombre d'emplois pour chaque titre de poste.
18. Filtrage des emplois à temps plein : Sélectionne les lignes où la valeur de la colonne "Employment Status" est "Full-time", afin de ne montrer que les emplois à temps plein.