Génie Logiciel et Projet
Cahier des charges

Projet n°33 : BIG DATA
Simulation pédagogique des principes fondamentaux du Big Data

Université : CY Cergy Paris Université
Formation : Licence Informatique – L2
Auteurs : TUNC Murat – TURKASLAN Emre – BOUMAIZA Sami
Date : 24/01/2026

1. Introduction au projet

Les systèmes informatiques modernes produisent en continu des quantités importantes de données. Dans le cas des réseaux sociaux, chaque action (visionnage, like, commentaire, partage) génère des informations exploitables. Les plateformes de type TikTok s'appuient sur ces données pour analyser les tendances et proposer des recommandations personnalisées.

Le Big Data regroupe des méthodes permettant de stocker, traiter et analyser des volumes massifs de données, générées rapidement et sous des formats variés. Ces données dépassent les capacités classiques de traitement sur une seule machine et nécessitent des approches distribuées.

Ce projet consiste à réaliser une application pédagogique de simulation Big Data. L'objectif n'est pas de reproduire une infrastructure industrielle réelle, mais de permettre à l'utilisateur de comprendre les principes fondamentaux de ce domaine à travers des scénarios simples, progressifs et interactifs.

Le fil conducteur retenu est un réseau social (exemple : TikTok). Les données simulées représentent des vidéos publiées et des interactions associées (vues, likes, commentaires, hashtags). Ce choix rend les notions plus concrètes et facilite la compréhension des traitements.

Le projet s'organise autour de trois simulations complémentaires : (1) visualisation du flux de données et des étapes Big Data, (2) simulation de la répartition des données et des requêtes sur plusieurs machines, (3) simulation d'un mécanisme d'intelligence simplifié (type recommandation) inspiré des systèmes de réseaux sociaux.

1.1 Motivation du choix du projet

Le choix du projet Big Data s'explique par l'importance de ce domaine dans de nombreuses applications actuelles (réseaux sociaux, commerce en ligne, services numériques). Il représente un sujet concret et utile pour comprendre les limites des systèmes classiques et l'intérêt du traitement distribué.

Le scénario « réseau social » permet d'illustrer facilement le volume de données, leur arrivée continue et l'exploitation des résultats (statistiques, tendances, recommandations). Le projet permet également de mettre en pratique des notions de génie logiciel, notamment la séparation entre le moteur de simulation (traitement) et l'IHM (affichage et interactions).2

2. Spécifications du projet

2.1 Notions de base (définitions)

Big Data : Ensemble de techniques permettant de gérer des volumes massifs de données, générées rapidement et sous des formes variées, afin d'en extraire des informations utiles.

Les 3V : Volume (quantité importante), Vélocité (arrivée continue/rapide), Variété (différents formats : nombres, textes, catégories).

Donnée simulée : Élément généré par l'application représentant une interaction sur un réseau social (ex. une vidéo, une vue, un like, un commentaire, un hashtag).

Requête : Question posée au système afin d'obtenir un résultat à partir des données (ex. « Top 10 vidéos les plus vues » ou « hashtags les plus utilisés »).

Traitement distribué : Organisation consistant à répartir le stockage et/ou le calcul sur plusieurs machines (nœuds) au lieu d'une seule.

2.2 Périmètre du projet

Le projet simule un système Big Data appliqué à un réseau social. La simulation se concentre sur la compréhension des étapes de traitement et sur la visualisation pédagogique.

Le projet ne comprend pas : le stockage réel en base de données distribuée, la gestion de comptes utilisateurs, la lecture de vraies vidéos, ni la collecte de données réelles.

2.3 Contraintes générales

- Contraintes pédagogiques : mécanismes simples, résultats compréhensibles, explications via l'IHM, possibilité de relancer la simulation.
- Contraintes techniques : application exécutable sur un ordinateur standard, temps de réponse raisonnable, architecture modulaire (moteur séparé de l'IHM).
- Contraintes de qualité : affichages clairs, interactions stables, messages d'erreur explicites, cohérence du scénario TikTok sur l'ensemble des simulations.

2.4 Données manipulées (réseau social type TikTok)

Une vidéo est représentée par un identifiant, une date/temps de publication simulé, une catégorie ou thème, et une liste de hashtags.

Les interactions associées à une vidéo sont représentées par des compteurs : nombre de vues, nombre de likes, nombre de commentaires, nombre de partages.

Un flux de données correspond à l'arrivée progressive de nouvelles interactions. Par exemple, les vues augmentent dans le temps et les likes/commentaires apparaissent avec un certain rythme.

2.5 Simulation 1 – Fonctionnement visuel des flux Big Data

- Objectif : montrer la chaîne de traitement Big Data de manière visuelle, depuis la génération des données jusqu'à l'affichage de résultats simples.
- Entrée : paramètres de simulation fournis par l'utilisateur (nombre initial de vidéos, durée de simulation, vitesse d'arrivée des interactions).
- Traitement : le moteur génère des vidéos et crée des interactions (vues/likes/commentaires) au fil du temps. Les données sont ensuite regroupées et préparées pour l'analyse.
- Sorties : affichage du volume de données générées, évolution temporelle des interactions, et premiers indicateurs (moyenne de vues, top vidéos).
- Résultat attendu : l'utilisateur comprend qu'un flux de données arrive en continu et qu'un traitement est nécessaire avant l'analyse.

2.6 Simulation 2 – Répartition des machines et exécution des requêtes

- Objectif : simuler le principe de répartition des données et des requêtes sur plusieurs machines (cluster) afin d'expliquer le traitement distribué.
- Le système simule N nœuds (machines). Chaque nœud reçoit une partie des données (par exemple un ensemble de vidéos ou un intervalle temporel).
- Une requête utilisateur est découpée en sous-tâches exécutées sur les nœuds, puis les résultats partiels sont regroupés pour produire la réponse finale.
- Exemples de requêtes supportées : Top 10 vidéos (vues), Top hashtags, total de likes sur une période simulée, moyenne de vues par catégorie.
- Affichage attendu : visualisation de la répartition (données par nœud), état des nœuds pendant l'exécution, et résultat final après agrégation.

2.7 Simulation 3 – Intelligence simplifiée (recommandation)

- Objectif : montrer de manière pédagogique comment une plateforme peut exploiter les données pour recommander du contenu.
- Le moteur calcule un score de popularité pour chaque vidéo à partir de ses interactions (exemple de score : vues + 2×likes + 3×commentaires).
- Le système applique une règle de mise en avant : les vidéos avec un score élevé apparaissent en priorité dans la liste « recommandée ».
- Le système simule un mécanisme de récompense : lorsqu'une vidéo est recommandée, elle reçoit davantage de vues simulées.
- Résultat attendu : l'utilisateur observe la création de tendances et comprend que la recommandation influence les données futures.

2.8 Simulation 4 – Apprentissage automatique (Machine Learning)

- Objectif : Illustrer le principe de l'apprentissage supervisé en montrant comment le Big Data permet d'entraîner un modèle prédictif pour anticiper la viralité d'un contenu.

- Principe : Le système utilise l'historique des données générées lors des simulations précédentes (dataset d'entraînement) pour identifier des corrélations entre les caractéristiques d'une vidéo (catégorie, heure, hashtags) et son succès final (nombre de vues).
- Fonctionnement : L'utilisateur lance une phase d'entraînement sur les données passées. Ensuite, lors de la génération de nouvelles vidéos, l'IA attribue un « score de viralité prédictif » avant même que les interactions ne commencent.
- Sorties : Affichage de la précision du modèle (pourcentage de réussite), mise en évidence des critères déterminants (ex : « le hashtag #Danse augmente les vues de 20% ») et comparaison graphique entre la courbe de vues prédite et la courbe réelle.
- Résultat attendu : L'utilisateur comprend que l'accumulation de données (Big Data) est le carburant nécessaire pour entraîner une IA capable de prédire des tendances futures.

2.9 Fonctionnalités attendues (point de vue utilisateur)

- Lancer une simulation : Choisir des paramètres (nombre de vidéos, durée, vitesse d'arrivée). Démarrer la génération des données et afficher le démarrage du flux.
- Mettre en pause / reprendre : Interrompre temporairement la génération et reprendre sans perdre l'état courant.
- Réinitialiser la simulation : Réinitialiser toutes les données à un état initial (zéro interaction) et revenir à l'écran de paramètres.
- Visualiser les flux : Afficher l'évolution des compteurs (vues/likes/commentaires) et le volume total de données générées.
- Exécuter une requête : Sélectionner une requête prédéfinie (Top vidéos, Top hashtags, etc.), lancer le traitement et obtenir le résultat.
- Simuler la répartition des machines : Choisir le nombre de nœuds, visualiser la distribution des données et suivre l'exécution distribuée d'une requête.
- Observer l'intelligence (recommandation) : Activer/désactiver la recommandation et comparer la liste des vidéos populaires avec la liste recommandée.
- Consulter un résumé : Afficher un récapitulatif en fin de simulation (statistiques globales, tendances détectées, vidéos dominantes).
- Entraîner et tester l'IA : Lancer l'apprentissage sur les données existantes et visualiser les prédictions sur les nouvelles vidéos générées.

2.10 Affichages attendus (IHM graphique)

- Zone de contrôle : boutons Démarrer, Pause, Reprendre, Réinitialiser ; sélection de la simulation 1/2/3.
- Panneau de paramètres : nombre de vidéos, nombre de nœuds (simulation 2), vitesse du flux, durée simulée.
- Tableau des vidéos : liste avec identifiant, vues, likes, commentaires, hashtags principaux, score (simulation 3).
- Panneau de résultats : top vidéos, top hashtags, statistiques globales (total, moyenne, maximum).
- Panneau de suivi distribué : répartition des données par nœud et état de calcul (en cours/terminé).

- Panneau Machine Learning : Graphique comparatif « Prédiction vs Réalité », barre de progression de l'entraînement, et affichage des facteurs d'influence (poids des variables).

2.11 Messages et comportements attendus

- Lors du démarrage, l'application affiche un message indiquant que la génération des données est en cours.
- Lors d'une pause, l'application conserve les données et bloque uniquement la progression du temps simulé.
- Lors d'une requête, l'application affiche l'état de traitement puis affiche le résultat final.
- En cas de paramètres incohérents (ex. valeur négative), l'application affiche un message d'erreur explicite et refuse le lancement.

3. Conclusion

Ce document de spécification décrit une application pédagogique simulant les principes fondamentaux du Big Data à partir d'un scénario de réseau social type TikTok. Les trois simulations proposées permettent de comprendre successivement les flux de données, la répartition des traitements et l'exploitation des données via une intelligence simplifiée.

Ce cahier des charges servira de base pour la conception détaillée (données et IHM), l'implémentation du moteur, la réalisation des traitements et la validation par des tests.