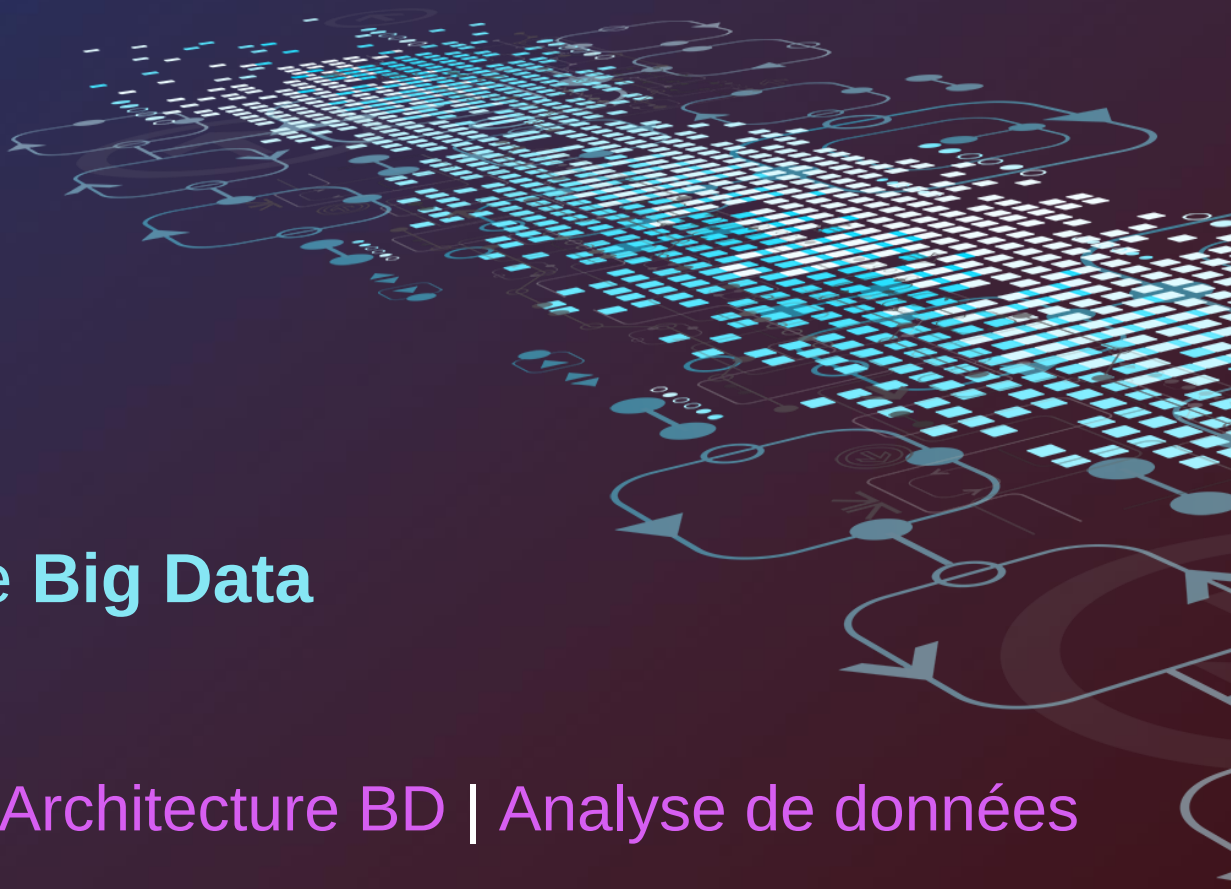
GRANDE ÉCOLE D'INGÉNIEURS

TRIMANE
THE DATA INTELLIGENCE COMPANY



Ingénierie **Big Data**

TP/TD **3**
Modules : Architecture BD | Analyse de données

Proposé et encadré par : *Bilel SDIRI, PhD*

Le but de ce TP et de vous familiariser avec les différentes fonctionnalités et opérateurs du pipeline d'agrégation. Les questions de ce TP sont basées sur la collection « *movies* » qui contient des informations sur un ensemble varié de films. Commencez le TP par la création d'une base de données intitulée « tp3 » contenant la collection « *movies* » importée depuis le fichier *collectionTP3_films.json*, fourni par l'enseignant.

Framework d'agrégation

Le but de ce TP et de vous familiariser avec les différentes fonctionnalités et opérateurs du pipeline d'agrégation de MongoDB. Les questions de ce TP sont basées sur la collection « *movies* » qui contient des informations sur un ensemble varié de films. Commencez le TP par la création d'une base de données intitulée « tp3 » contenant la collection « *movies* » importée depuis le fichier *collectionTP3_films.json*, fourni par l'enseignant.

Requêtes à faire

1. Affichez la note maximale, minimale, moyenne et l'écart-type des évaluations *imdb* (`imdb.rating`) des films ayant au moins un prix.

2. Affichez le nombre de documents ayant la même note d'évaluation *imdb* (`imdb.rating`). L'affichage se fait par ordre décroissants du nombre de films par évaluation.

3. Affichez le nombre de films disponibles en anglais et en japonais, ayant une évaluation *imdb* (`imdb.rating`) égale au moins à 7, le genre ne contient pas « Crime » ou « Horror », et le champ `rated` est soit « PG » ou « G ». Affectez le pipeline à une variable comme suit :

   ```
   var pipeline = [{ $etape1 : <traitement1> }, …]
   ```

   Afin de compter le nombre de documents du pipeline, utilisez la commande suivante :

   ```
   db.<nomCollection>.aggregate(pipeline).itcount()
   ```

4. Affichez le titre et l'année de sortie des films dont le titre est composé d'un seul mot. On considère que «ABC-CD » est un seul mot, et «ABC CD » contient deux mots.

5. Affichez le nombre de films et la moyenne des `metacritic` en fonction du nombre de directeurs ayant effectué la réalisation. L'affichage se fait par ordre décroissant du nombre de directeurs par film.

6. Les utilisateurs ont identifié les acteurs suivants comme étant leurs favoris : *"Sandra Bullock", "Tom Hanks", "Julia Roberts", "Kevin Spacey",* et *"George Clooney"*. Pour les films sortis aux États-Unis (USA) avec un `tomatoes.viewer.rating` supérieur strictement à 2, calculez un nouveau champ intitulé `nbr_favs` représentant le nombre d'acteurs favoris qui apparaissent dans le *casting* du film. Triez les résultats en fonction de `nbr_favs`, `tomatoes.viewer.rating` et du titre dans l'ordre décroissant. Quel est le titre du 39éme film dans le résultat de l'agrégation ?

7. Calculez l'évaluation moyenne (i.e. rating) pour chaque film sorti à partir de 1990, disponible en anglais, ayant un `imdb.rating` minimum égale à 1, et un `imdb.votes` minimal égale à 1. Vous serez emmenés à effectuer une remise à l'échelle des votes *imdb* et de normaliser les notes d'évaluation *imdb (*`rating`*)*. Quel est le film ayant la note d'évaluation normalisée minimale ? [demandez la formule de normalisation à l'enseignant]