Master "Sciences des Données et Systèmes. Intelligents (SDSI)"

Analyse de logs avec Pig

Réalisé par :
Meryem Ouafi
Wiam CHOUKOUD
Youssra Andoulla
Encadré par :
Pr.Mourad fariss

Année universitaire 2023 - 2024

Table des matières

1 Introduction **2**

2 Objectif du projet **2**

3 Les technologies utilisees **3**
- 3.1 Hadoop 3
- 3.2 HDFS (Hadoop Distributed File System) 3
- 3.3 Apache Pig 3

4 Les donnees qui l'on utilisent : **4**

5 Partie pratique : **4**
- 5.1 Importation des donnes dans hdfs : 4
- 5.2 Lancer pig 5
- 5.3 Faire des analyses : 5
  - 5.3.1 analyse 1 :Compter le nombre de requêtes pour chaque adresse IP 5
  - 5.3.2 Analyse 2 : de la répartition des codes de statut HTTP 6
  - 5.3.3 Analyse3 :Analyse des navigateurs utilisés : 6
  - 5.3.4 Analyse4 : Trouver le code d'état le plus fréquente : 7
  - 5.3.5 Analyse 5 :Trouver le code d'état le moins fréquente : 7
  - 5.3.6 Analyse 6 : Calculer le nombre total de requêtes par code d'état : 7
  - 5.3.7 Analyse des requêtes par pays 7

6 Conclusion **8**

1 Introduction

L'analyse des logs de serveur web est une pratique cruciale dans le monde de l'informatique et du web. Ces fichiers journaux, générés automatiquement par les serveurs web, enregistrent chaque interaction entre le serveur et les utilisateurs. Ils contiennent une mine d'informations sur les requêtes HTTP, les adresses IP des visiteurs, les navigateurs utilisés, les erreurs rencontrées et bien plus encore. L'objectif de cette analyse est de comprendre le comportement des utilisateurs, de diagnostiquer les problèmes de performance, d'identifier les menaces potentielles à la sécurité et d'optimiser l'expérience utilisateur. En examinant les logs de serveur web, les administrateurs système et les spécialistes du marketing peuvent obtenir des informations précieuses sur la popularité du site, les pages les plus visitées, les sources de trafic, les erreurs fréquentes, et bien d'autres aspects essentiels à la gestion efficace d'un site web.

figure 1 :L'analyse des logs

2 Objectif du projet

L'objectif de ce projet est d'utiliser Apache Pig pour analyser des logs de serveur web simples. Nous chercherons à extraire des informations significatives à partir de ces données brutes afin de comprendre le comportement des utilisateurs, diagnostiquer les problèmes potentiels de performance du serveur, et optimiser l'expérience utilisateur. En explorant les différentes analyses que nous pouvons effectuer, nous visons à démontrer comment Apache Pig peut être un outil puissant pour traiter et analyser les logs de serveur web, offrant ainsi des insights précieux pour la gestion et l'optimisation des sites web.

figure 2 :Les fichiers de logs

3 Les technologies utilisees

3.1 Hadoop

Nous utiliserons la plateforme Hadoop pour le stockage et le traitement distribué des données. Hadoop permet de gérer de grands volumes de données de manière évolutive et offre un environnement robuste pour exécuter des analyses sur des clusters de serveurs.

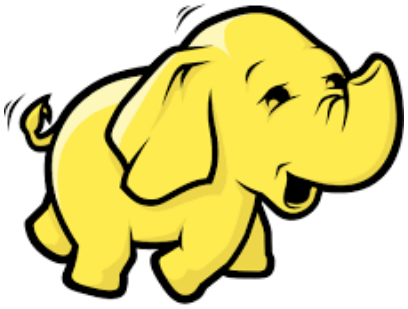


figure 3 : Icone de Hadoop

3.2 HDFS (Hadoop Distributed File System)

Nous stockerons les logs de serveur web dans le système de fichiers distribué Hadoop (HDFS). HDFS permet de stocker les données de manière distribuée sur plusieurs nœuds, offrant ainsi une haute disponibilité et une tolérance aux pannes.

figure 4 :Icone de HDFS

3.3 Apache Pig

Nous utiliserons Apache Pig, un langage de traitement de données parallèle pour Hadoop, pour analyser les logs de serveur web. Pig offre une syntaxe simple et expressive pour effectuer des transformations de données complexes sur de grands ensembles de données de manière efficace.

figure 5 :Icone de Apache Pig

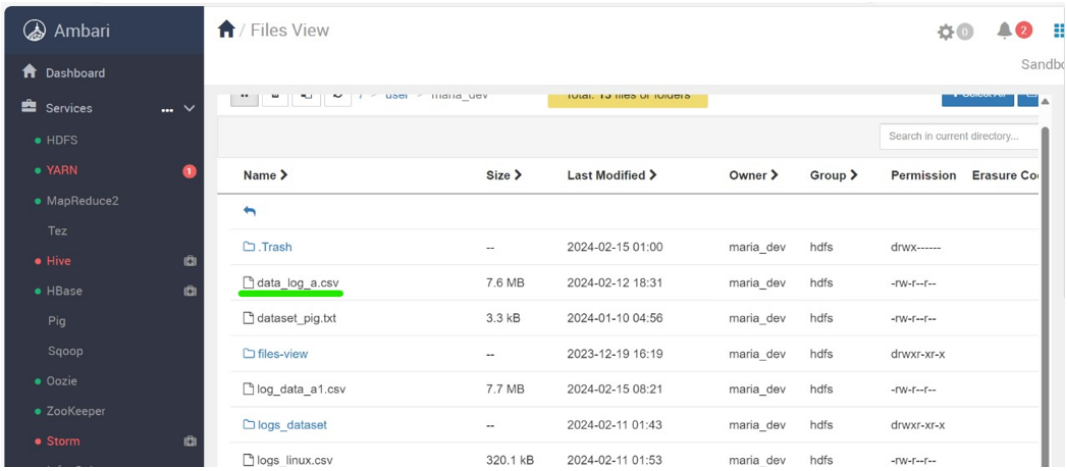
4 Les donnees qui l'on utilisent :

Les données que nous utilisons dans ce projet proviennent des logs d'un serveur web. Chaque ligne de données représente une interaction entre le serveur web et un utilisateur. Voici les différentes colonnes présentes dans nos données :
• IP : L'adresse IP de l'utilisateur qui a effectué la requête.
• Timestamp : L'horodatage de la requête, indiquant quand elle a été effectuée.
• Request : Le type de requête HTTP effectuée, par exemple GET, POST, DELETE, etc.
• Navigateur : Le navigateur web utilisé par l'utilisateur pour accéder au site.
• Plateforme : Le système d'exploitation ou la plateforme utilisée par l'utilisateur, par exemple Android, Windows, iOS, etc.
• Country : Le pays à partir duquel la requête a été effectuée.
• Statucode : Le code de statut HTTP retourné par le serveur en réponse à la requête, indiquant si la requête a réussi, a été redirigée ou a rencontré une erreur.
Ces données fournissent des informations essentielles sur les interactions des utilisateurs avec le site, telles que les cara ctéristiques de leurs appareils, les pays d'où proviennent les visiteurs, et les éventuelles erreurs rencontrées lors de l'accès au site.

5 Partie pratique :

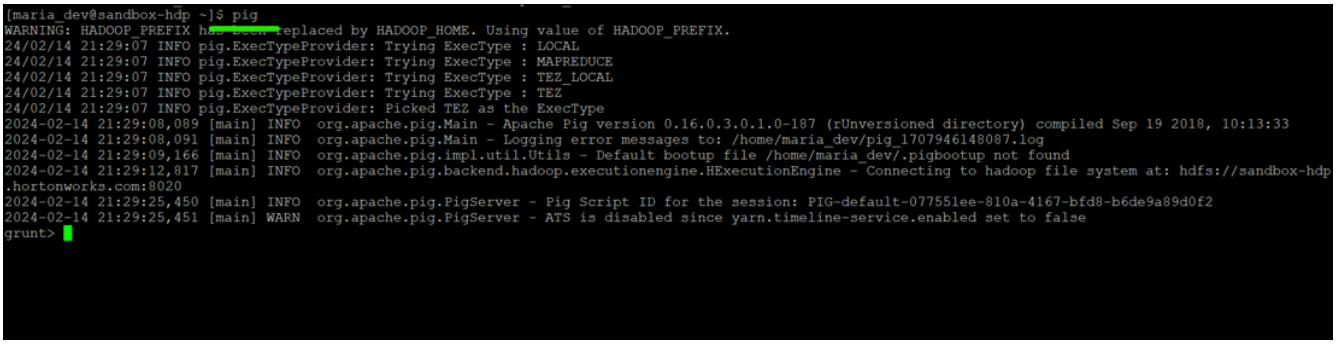
5.1 Importation des donnes dans hdfs :

Importation de dataset dans hdfs :

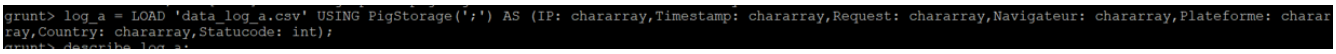


5.2 Lancer pig

Lancer pig :



Charger les donnees de hdfs a pig :

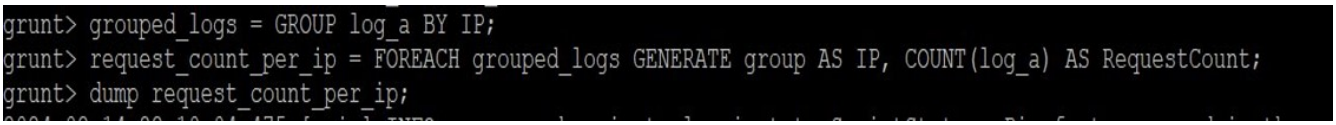


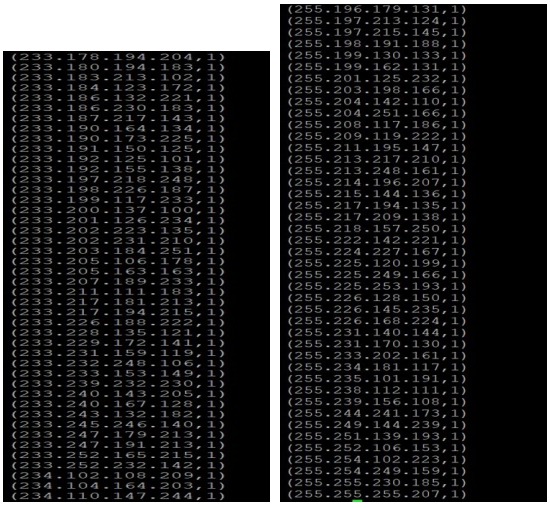
Afficher la description :

5.3 Faire des analyses :

5.3.1 analyse 1 :Compter le nombre de requêtes pour chaque adresse IP

Regrouper les logs par adresse IP pour compter le nombre de requêtes pour chaque adresse IP





5.3.2 Analyse 2 : de la répartition des codes de statut HTTP

Examiner la répartition des codes de statut HTTP dans les logs du serveur web. En regroupant les logs par code de statut HTTP

```
grunt> grouped_status_codes = GROUP log_a BY Statucode;
grunt>
```

```
grunt>  status_code_counts = FOREACH grouped_status_codes GENERATE group AS StatusCode, COUNT(log_a) AS Re
grunt>
```

les utilisateurs accèdent principalement au site web à partir d'appareils Android et Windows, avec respectivement 40 162 et 39 810 requêtes. Cependant, on observe également une présence significative d'utilisateurs de Mac OS X, bien que moins importante en termes de volume, avec 9 938 requêtes.

```
(404,42867)
(500,6370)
(501,1)
(502,26007)
(503,24754)
(504,1)
```

5.3.3 Analyse3 :Analyse des navigateurs utilisés :

la répartition des navigateurs utilisés par les utilisateurs pour accéder au site web.

```
grunt> grouped_platforms = GROUP log_a BY Plateforme;
grunt> platform_request_counts = FOREACH grouped_platforms GENERATE group AS Platform, COUNT(log_a) AS RequestCount;
grunt>
```

```
grunt> DUMP platform_request_counts;
```

```
grunt> DUMP status_code_counts;
```

le code de statut HTTP le plus fréquent est 404, avec un total de 42 867 occurrences. Les codes 502 et 503 sont également significativement présents, suggérant potentiellement des erreurs ou des problèmes de connectivité sur le serveur. les utilisateurs accèdent principalement au site web à partir d'appareils Android et Windows, avec respectivement 40 162 et 39 810 requêtes. Cependant, on observe également une présence significative d'utilisateurs de Mac OS X, bien que moins importante en termes de volume, avec 9 938 requêtes. Cette répartition des plateformes utilisées souligne l'importance de l'optimisation de l'expérience utilisateur pour une variété d'appareils et

de systèmes d'exploitation.

```
(Android,40162)
(Windows,39810)
(Mac OS X,9938)
```

5.3.4 Analyse4 : Trouver le code d'état le plus fréquente :

```
grunt>  most_frequent_status = FOREACH (GROUP log_a ALL) GENERATE MAX(log_a.Statucode) AS most_frequent_status;
grunt>
grunt>
grunt>
grunt> dump most_frequent_status;
```

```
(504)
grunt>
```

5.3.5 Analyse 5 :Trouver le code d'état le moins fréquente :

```
grunt> least_frequent_status = FOREACH (GROUP log_a ALL) GENERATE MIN(log_a.Statucode) AS least_frequent_status;
grunt>
grunt>
grunt> dump least_frequent_status;
```

```
(404)
grunt>
```

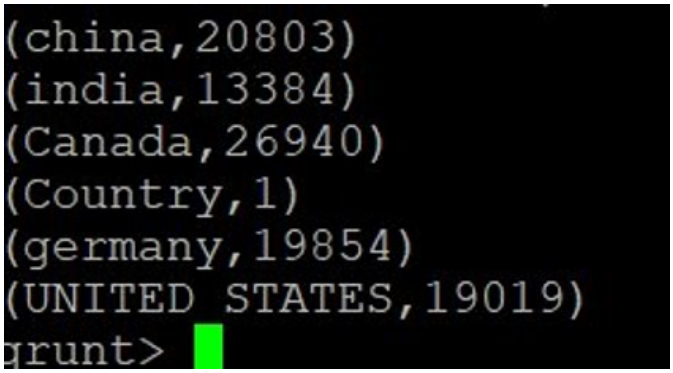
5.3.6 Analyse 6 : Calculer le nombre total de requêtes par code d'état :

```
grunt> total_requests_by_status = FOREACH (GROUP log_a BY Statucode) GENERATE group AS status_code, SUM(log_a.Statucode) AS total_requests;
grunt>
grunt>
grunt> dump total_requests_by_status;
```

```
(404,17318268)
(500,3185000)
(501,501)
(502,13055514)
(503,12451262)
(504,504)
```

5.3.7 Analyse des requêtes par pays

```
grunt> grouped_countries = GROUP log_a BY Country;
grunt> country_request_counts = FOREACH grouped_countries GENERATE group AS Country, COUNT(log_a) AS RequestCount;
grunt>
grunt>
grunt>
grunt>
```



6 Conclusion

L'analyse des logs d'un serveur web à l'aide d'Apache Pig offre une vision détaillée et précieuse du comportement des utilisateurs ainsi que des performances du site. En examinant les données telles que les adresses IP, les codes de statut HTTP, les navigateurs utilisés et les pays d'origine des visiteurs, nous pouvons tirer des enseignements précieux pour améliorer l'expérience utilisateur, optimiser les performances du site et prendre des décisions stratégiques.

À travers ce projet, nous avons pu mettre en œuvre différentes analyses telles que le comptage des requêtes par adresse IP, l'analyse des codes de statut HTTP, la répartition des navigateurs utilisés et l'analyse géographique des requêtes par pays. Ces analyses nous ont permis de mieux comprendre le comportement des utilisateurs, d'identifier les zones d'amélioration potentielles et de prendre des décisions éclairées pour optimiser le fonctionnement du site web.

En conclusion, l'utilisation d'Apache Pig pour l'analyse des logs de serveur web s'avère être un outil puissant pour extraire des informations pertinentes à partir de grandes quantités de données de manière efficace et éviter ainsi les erreurs et les incohérences. En continuant à explorer et à analyser ces données, les entreprises peuvent améliorer leur compréhension de leur audience en ligne et prendre des mesures pour optimiser leur présence sur le web