Atelier 2 : Analyse Big data avec Hive

Etape 1 : Afficher les 10 premiers visiteurs de la maison blanche

1.1 Créer un nouveau fichier **whitehouse.hive**

1.2 Nous allons nous intéresser aux visiteurs de la maison blanche en fonction de leurs dates d'arrivées.

Pour commencer, créer une requête qui récupère toutes les lignes de la table **wh_visits** en éliminant les lignes ayant **time_of_arrival** vide.

```
select * from wh_visits where time_of_arrival != "";
```

1.3 Pour trouver la première visite, il faut trier le résultat. Cela nécessite de convertir le time_of_arrival dans un timestamp en utilisant la fonction **unix_timestamp** comme suit :

```
unix_timestamp(time_of_arrival,'MM/dd/yyyy hh:mm')
```

Ajouter à la requête ci-dessus un order by sur unix_timestamp :

```
order by unix_timestamp(time_of_arrival,'MM/dd/yyyy hh:mm')
```

1.4 Limiter le résultat à 10 lignes.

```
limit 10;
```

Requête finale :

```
select * from wh_visits
where time_of_arrival != ""
order by unix_timestamp(time_of_arrival,  'MM/dd/yyyy hh:mm')
limit 10;
```

1.5 Sauvegarder puis relancer le script **whitehouse.hive :**

```
hive -f whitehouse.hive
```

1.6 Les résultats devraient être de 10 visiteurs, et la première visite devrait être de Charles Kahn le 3/5/2009.

Résultat :

KAHN	CHARLES	3/5/2009 12:30	3/5/2009 12:30	WH	
KEEHAN	CAROL	3/5/2009 12:30	3/5/2009 12:30	WH	
DALEY	WILLIAM	3/17/2009 18:30	3/17/2009 18:30	WH	
KEEHAN	CAROL	3/17/2009 18:30	3/17/2009 18:30	WH	ST. PATRICK'S RECEPTION GUESTS
DAVIS	RICHARD	3/27/2009 11:16	3/27/2009 12:00	WH	
PAESE	MICHAEL	3/27/2009 11:24	3/27/2009 12:00	WH	
DAVIS	RICHARD	5/1/2009 15:45	5/1/2009 15:45	WH	
EDWARDS	ERIC	5/22/2009 14:15	5/22/2009 14:15	WH	
JOHNSON	JAMES	5/22/2009 14:15	5/22/2009 14:15	WH	
YZAGGUIRRE	RAUL	8/12/2009 9:30	8/12/2009 9:30	WH	

Etape 2 : Trouver la dernière visite

2.1 Adapter la requête **whitehouse.hive** pour trouver la dernière visite de la maison blanche

```
select * from wh_visits
where time_of_arrival != ""
order by unix_timestamp(time_of_arrival,  'MM/dd/yyyy hh:mm') desc
limit 10;
```

2.2 Exécuter la requête à nouveau, et vous devriez voir que la visite la plus récente a été Jackie Walker le 18/03/2011.

```
hive -f whitehouse.hive
```

Résultat :

```
Total MapReduce CPU Time Spent: 13 seconds 10 msec
OK
WALKER      JACKIE    3/18/2011 13:27      3/18/2011 13:00      WH
BATES       AARON     3/17/2011 18:56      3/17/2011 18:00      WH
PENA        LISA      3/17/2011 17:35      3/17/2011 17:40      WH
DALEY       NORA      3/17/2011 17:02      3/17/2011 18:00      WH
KRIS        JODY      3/17/2011 16:25      3/17/2011 16:30      WH
KRIS        DAVID     3/17/2011 16:25      3/17/2011 16:30      WH
KRIS        AUDREY    3/17/2011 16:24      3/17/2011 16:30      WH
KRIS        HANNAH    3/17/2011 16:24      3/17/2011 16:30      WH
BLANTON     THOMAS    3/16/2011 14:09      3/16/2011 14:15      WH
MCDERMOT PATRICE      3/16/2011 14:09      3/16/2011 14:15      WH
Time taken: 110.566 seconds, Fetched: 10 row(s)
```

Étape 3:Trouver les commentaires les plus communs

Dans cette étape, il faut explorer le champ **info_comment** qui indique le motif de la visite de la maison blanche.

Déterminer donc les 10 commentaires les plus communs (motifs de visites les plus récurrents). Il faut utiliser certaines des fonctions d'agrégats de Hive pour réaliser la requête.

3.1 Créer un nouveau fichier **comments.hive**

3.2 Créer la requête permettant d'afficher les 10 de commentaires les plus communs et le nombre d'occurrences pour chaque commentaire en éliminant les **info_comment** vide.

3.3 Sauvegarder et exécuter le fichier **comments.hive**.

```
hive -f comments.hive
```

```
select count(*) as comment_count, info_comment
from wh_visits
group by info_comment
order by comment_count DESC
limit 10;
```

Résultat :

```
OK
9036
1253    HOLIDAY BALL ATTENDEES/
894     WHO EOP RECEP 2
700     WHO EOP 1 RECEPTION/
601     RESIDENCE STAFF HOLIDAY RECEPTION/
586     PRESS RECEPTION ONE (1)/
580     GENERAL RECEPTION 1
540     HANUKKAH RECEPTION./
540     GEN RECEP 5/
516     GENERAL RECEPTION 3
```

3.4 Vous constatez que le blanc est le commentaire le plus en commun, puil vient en 2eme position **le HOLIDAY BALL ATTENDEES**.

Modifiez la requête afin qu'elle ignore les commentaires vides. Vous devriez voir le commentaire "GEN RECEP 6 /" qui s'affiche.

```
hive -f comments.hive
```

```
select count(*) as comment_count, info_comment
from wh_visits
where info_comment != ""
group by info_comment
order by comment_count DESC
limit 10;
```

Résultat :

```
OK
1253    HOLIDAY BALL ATTENDEES/
894     WHO EOP RECEP 2
700     WHO EOP 1 RECEPTION/
601     RESIDENCE STAFF HOLIDAY RECEPTION/
586     PRESS RECEPTION ONE (1)/
580     GENERAL RECEPTION 1
540     GEN RECEP 5/
540     HANUKKAH RECEPTION./
516     GENERAL RECEPTION 3
498     GEN RECEP 6/
```

Etape 4 : Trouver les commentaires les moins fréquents

Adapter la requête précédente pour identifier les 10 commentaires les moins fréquents.

Le résultat de la requête devrait ressembler à:

```
1 merged to u59031
1 WHO EOP/
1 WHO EOP RECLEAR
1 WAITING FOR SUPERMAN VISIT
1 ST. PATRICK'S RECEPTION GUESTS
1 SCIENCE FAIR
1 RES PARTY/
1 PRIVATE MEETING
```

1 PRIVATE LUNCH
1 POTUS PHOTO W/ US ATTORNEYS/

Note : Cela semble exact puisque 1 est le moins de fois qu'un commentaire peut apparaître.

De plus cette requête relève que Superman a visité le président au moins une fois.

```
select count(*) as comment_count, info_comment
from wh_visits
where info_comment != ""
group by info_comment
order by comment_count ASC
limit 10;
```

Étape 5: Analyser les incohérences des données

3.1 En analysant les résultats des commentaires les plus fréquents et les moins fréquents, il apparaît que Plusieurs variations de RECEPTION GÉNÉRALE se produisent dans le top 10.

A cette étape, vous allez essayer de déterminer le nombre de visites réelles impliquant une réception générale en essayant de nettoyer certaines de ces incohérences dans les données.

Des incohérences comme celles-ci sont très courantes dans les grosses masses de données, nous allons donc contourner cela en utilisant leurs propres abréviations.

3.2 Modifier le script **comments.hive** en cherchant les commentaires les plus fréquents qui contiennent la chaîne "RECEP". Changer aussi, la limit de 10 à 30 lignes ;

```
select count(*) as comment_count, info_comment
from wh_visits
where info_comment like "%RECEP%"
group by info_comment
order by comment_count DESC
limit 30;
```

Résultat :

```
OK
894     WHO EOP RECEP 2
700     WHO EOP 1 RECEPTION/
601     RESIDENCE STAFF HOLIDAY RECEPTION/
586     PRESS RECEPTION ONE (1)/
580     GENERAL RECEPTION 1
540     HANUKKAH RECEPTION./
540     GEN RECEP 5/
516     GENERAL RECEPTION 3
498     GEN RECEP 6/
438     GEN RECEP 4
421     KENNEDY CENTER HONORS RECEPTION
404     DIPLOMATIC CORPS RECEPTION
403     HOLIDAY RECEPTION
173     NATIONAL MEDALS OF SCIENCE CEREMONY AND RECEPTION/
100     MARINE BAND MUSICIANS FOR A HOLIDAY RECEPTION
```

```
        43      WHO EOP RECEP/
        40      MARINE BAND MEMBERS FOR A HOLIDAY RECEPTION.
        34      MARINE BAND FOR CONGRESSIONAL RECEPTION EVENT
        34      MARINE BAND FOR A HOLIDAY RECEPTION
        33      MARINE BAND FOR THE HANUKKAH RECEPTION.
        31      HOLIDAY RECEPTION/
        31      GENERAL RECEPTION 2
        23      GENERAL RECEPTION 3
        20      WHO EOP HOLIDAY RECEP/
        20      GENERAL RECEPTION 6
        20      GENERAL RECEPTION 5
        13      GENERAL RECEPTION 1
        12      MARINE BAND FOR A HOLIDAY RECEPTION.  THEY WILL NOT BE WITHIN ARMS REACH OF THE
PRESIDENT.
        10      STAFF RECEP
        9       MARINE BAND FOR THE HOLIDAY RECEPTION/
```

3.3 Pour avoir le résultat qui correspond à la notification RECEPTION GÉNÉRALE, il faut modifier la clause Where dans **comments.hive** pour inclure **"%GEN%**": C'est-à-dire info_comment contient à la fois **RECEP** et **GEN**

```
select count(*) as comment_count, info_comment
from wh_visits
where info_comment like "%GEN%RECEP%"
group by info_comment
order by comment_count DESC
limit 30;
```

Résultat:

```
OK
580     GENERAL RECEPTION 1
540     GEN RECEP 5/
516     GENERAL RECEPTION 3
498     GEN RECEP 6/
438     GEN RECEP 4
31      GENERAL RECEPTION 2
23      GENERAL RECEPTION 3
20      GENERAL RECEPTION 6
20      GENERAL RECEPTION 5
13      GENERAL RECEPTION 1
8       GENERAL RECEP 5
3       GENERAL RECEPTION 6/
3       GENERAL RECEPTION 6
2       GENERAL RECEPTION 2/
1       GENERAL RECEPTION 1 /
1       GENERAL RECEPTION 1/
```

3.4 Compter le nombre total de visite incluant **GEN** et **RECEP** dans **info_comment**

```
select count(*) as comment_count
from wh_visits
where info_comment like "%GEN%RECEP%";
```

```
Total MapReduce CPU Time Spent: 12 seconds 560 msec
```

```
OK
2697
Time taken: 111.434 seconds, Fetched: 1 row(s)
```

3.5 Notez qu'il ya **2 697 visites** à la maison blanche avec **GEN RECEP** dans le commentaire, ce qui représente environ **12%** des **21 819** visites totales du POTUS dans notre ensemble de données.

Plus important encore, ces résultats montrent que notre première requête de 1 253 participants à HOLIDAY BALL ATTENDEES ne signifie pas que HOLIDAY BALL ATTENDEES est la raison la plus probable pour visiter Le président.

Plus de deux fois plus de visiteurs sont là pour une réception générale. Ce Type d'analyse est commun dans les grandes données, et il montre comment les analystes de données doivent être créatifs lors de leurs analyses.

Etape 6 : Vérification de résultat :

6.1 Nous avons 12% des visiteurs du POTUS pour une réception générale, mais il y a Beaucoup de déclarations dans les commentaires qui ont contenu l'WHO et l'EOP. Modifier la requête à partir de dernière étape et afficher les 30 commentaires qui contiennent«WHO» et «EOP».

Le résultat devrait ressembler à:

```
OK
894         WHO EOP RECEP 2
700         WHO EOP 1 RECEPTION/
43  WHO EOP RECEP/
20  WHO EOP HOLIDAY RECEP/
13  WHO/EOP #2/
8   WHO EOP RECEPTION
7   WHO EOP RECEP
1   WHO EOP/
1   WHO EOP RECLEAR
Time taken: 193.54 seconds, Fetched: 9 row(s)
```

```
select count(*) as comment_count , info_comment
from wh_visits
where info_comment like "%WHO%EOP%"
group by info_comment
order by comment_count DESC
limit 30;
```

6.2 Exécuter une requête qui compte le nombre d'enregistrements avec WHO et l'EOP dans les commentaires:

```
select count(*) as comment_count
from wh_visits
where info_comment like "%WHO%EOP%"
;
```

Résultat :
Total MapReduce CPU Time Spent: 12 seconds 510 msec
OK
1687
Time taken: 106.352 seconds, Fetched: 1 row(s)

Vous devriez obtenir 1,687 visites, soit 7,7% des visiteurs du POTUS. Donc, les réceptions générales semblent encore être le motif le plus fréquent.

Étape 7:Trouver les personnes qui ont le plus visité la maison blanche

7.1 Ecriver une requête hive qui permet d'afficher les 20 premières personnes qui ont visité le plus lePOTUS.

```
select fname, lname, count(*) as visit_count from wh_visits
group by fname, lname
order by visit_count desc
limit 20;
```

7.2 Pour vérifier votre script, voici les 20 premières personnes qui ont visité la maison blanche, avec le nombre de visites:

```
OK
ALAN    PRATHER         16
CHRISTOPHER     FRANKE          15
ANNAMARIA       MOTTOLA         15
ROBERT          BOGUSLAW        14
CHARLES         POWERS          14
SARAH   HART    12
JACKIE  WALKER          12
JASON   FETTIG  12
SHENGTSUNG      WANG    12
FERN    SATO    12
DIANA   FISH    12
JANET   BAILEY  11
PETER   WILSON          11
GLENN   DEWEY   11
MARCIO          BOTELHO         11
DONNA   WILLINGHAM      11
DAVID   AXELROD         10
CLAUDIA         CHUDACOFF       10
VALERIE         JARRETT         10
MICHAEL         COLBURN         10
Time taken: 195.863 seconds, Fetched: 20 row(s)
```

Résultat :

Vous avez écrit plusieurs requêtes Hive pour analyser les données des visiteurs de la Maison Blanche. Le but est que vous soyez à l'aise avec les requêtes et manipulation des données dans Hive.

Vous serez en mesures d'interroger et traiter des grandes volumétries de données stockées dans Hive