

# OCljp

HEATWAVEを利用した重複チェック – BIGQUERYも

**Oracle Cloud Infrastructure**

**2021/1/19 OCljp#14**

**Fullenergy**

Akishi Matsumoto

# 自己紹介（松本 昭史） ocijp発起人の一人です

## <職歴>

- 1997.04 サポートエンジニア（シリコングラフィックス・クレイ）
- **2001.05 Oracle DBA**（フリーランス）
- 2007.09 ミドルウェアエンジニア（フルエナジー）
- 2014.04 **AWSクラウドアーキテクト**（フルエナジー）
- **2018.11 OCIクラウドアーキテクト**（フルエナジー）

## <OCI関連の登壇>

- 2018.02 OPNパートナー向けセミナー
- 2018.07 NTTオンラインセミナー
- 2018.07 Fujitsu Forum 2018
- 2019.02 Oracle Cloud 九州
- **2019.07～ ocijpユーザグループ**
- 2019.08 Shing Star
- 2020.10 しゃちほこオラクル倶楽部



ā'pěks  
(#orclapex)



# 株式会社フルエナジー

- クラウドインテグレーター
- 2007年9月14日設立(14年目)
- 東京都港区南青山2-22-17
- **エンジニア率 93%**

- **AWS契約 7社 22アカウント**
- **OCI契約 30社 38テナンシー**

- 事業内容: **クラウド設計/構築/移行・リモート運用支援 (24x365監視)**  
DB移行/IDCS2要素認証導入/WAF導入/CDN導入/クラウドDR導入  
システム負荷テスト(クラウドサイジング支援)/Apexアプリ開発支援  
閉域網接続支援(Tokaiコミュニケーションズ・NTT東日本・Coltテクノロジーと協業)  
Fastly再販・Forti-VM販売・クラウド持込用MS RDS-SAL/CALの提供  
ISVとして自社サービスをOCIで複数稼働中(Fastconnect 2本も利用)



# 本日のネタ

## ■ 弊社で請けたタスク:



## シリアルナンバーの重複チェック

※2つの方法を用いてダブルチェックすること

※データは、テキストファイルで提供 (17GB)

シリアルナンバー: **14桁の整数** (よく缶コーヒーとかに貼ってあるシールの中身)

1回目のキャンペーン: 6.4億

2回目のキャンペーン: 4.8億 **※好評の為、新規発行。但し、1回目との重複があるかもしれない。**

14桁 ⇒ 1000兆通りを作れるから、重複する確率は低い。。。

もっと良い方法をご存じの方は是非ご連絡ください。

# 1つ目の方法

## ■ Linuxコマンドを使った重複チェック

<前提条件>

- 1.物理メモリが多いインスタンスの /dev/shm 領域で作業をする。
- 2.寝てる間に動いてくれることを期待して、crondから実行するようにする。
- 3.プロセスの優先順位を上げるため、rootユーザで作業をする。

## ■ 並び替え

```
# cat テキストファイル |sort > /dev/shm/210119/all_serial.txt
```

## ■ 重複のカウント

```
# uniq -c /dev/shm/210119/all_serial.txt > /dev/shm/210119/Duplicate
```

## ■ 重複したシリアルを抽出

```
# fgrep -v "      1 " /dev/shm/210119/Duplicate > /dev/shm/210119/Result
```

```
# cat /dev/shm/210119/Result
```

もっと良い方法をご存じの方は是非ご連絡ください。

# 1つ目の方法の結果

- 重複レコード数: 1878件
- 所要時間: 24時間
- リソース: OCI Compute vm.standard.E3flex (4core 64GBmemory)
- 費用: 約573円(24H)

# 2つ目の方法

## ■ MySQLを使った重複チェック

<前提条件>

- 1.DBAじゃない人でもできる簡単な手段を使う。
- 2.サイズの大きなMySQL Database Serviceを使う。（安価に済ませるため）

## ■ データのロード

```
# mysql --defaults-extra-file=/home/opc/mds.cnf -e ¥  
"LOAD DATA local INFILE 'serial.txt' ignore INTO TABLE testdb.table1 (serial)"
```

## ■ 重複したシリアルを抽出&カウント

```
mysql> SELECT serial, COUNT(serial) FROM table1 GROUP BY SERIAL HAVING COUNT(serial) > 1;
```

もっと良い方法をご存じの方は是非ご連絡ください。



# Oracleが提供するMySQLのクラウド・データベース

## MySQL Database Service (MDS)

- MySQLの開発元であるOracleが提供するDBaaS
- 業界唯一の有償版MySQLベースのサービス  
※他社MySQL系サービスは無償版MySQLがベース
- Amazon RDS (MySQL)の1/3以下のコスト
- 東京を含む8リージョンで展開中、年内に11リージョン
- オンプレ向けの無償版、有償版と同一バージョン
- E3 VM (fixed shapes)とE2 BMで提供中
- 高可用性構成は近日提供開始予定

|                                   |          |                   |
|-----------------------------------|----------|-------------------|
| MySQL Database—Standard—E3        | \$0.038  | OCPU per hour     |
| MySQL Database—Standard—E3—Memory | \$0.0022 | Gigabyte per hour |

## HeatWave

- OCI専用のMDSのクエリ・アクセラレーター
- Oracle LabsのProject RAPIDの成果を活用  
[https://labs.oracle.com/pls/apex/f?p=labs:49:::P49\\_PROJECT\\_ID:14](https://labs.oracle.com/pls/apex/f?p=labs:49:::P49_PROJECT_ID:14)
- アプリケーションからは単に分析処理の応答性能が高いMySQLを利用しているように見える
- TPC-Hのクエリの処理性能は
  - Amazon Auroraの1,000倍以上高速、コスト1/3以下
  - Amazon Redshiftの2.7倍高速、コスト1/3以下
- メルカリ、DMM.com, SCSK, スクウェア・エニックスがベータプログラムに参加し性能や運用性を検証

|                                          |          |               |
|------------------------------------------|----------|---------------|
| MySQL Database for Analytics—Standard—E3 | \$0.3536 | Node Per Hour |
| *minimum 3 nodes (\$789 per months)      |          |               |

出典: <https://www.mysql.com/jp/news-and-events/seminar/downloads.html>



# 2つ目の方法（補足：MDS構築）



DBシステムの基本情報の指定

名前  
mysql1

説明  
MySQL DBシステムの説明の記述

配置とハードウェアの構成

コンパートメントの選択  
fullenergy240 (root) (ルート)/perpetual/akishi

可用性ドメインの選択  
RQKF-AP-TOKYO-1-AD-1

☐ このDBシステムのフォルト・ドメインの選択  
フォルト・ドメインを選択しない場合、最適な配置がOracle

シェイプの選択  
VM.Standard.E2.8  
CPUコア数: 8  
メモリー・サイズ: 64 GB

[構成オプションの表示](#)

データ・ストレージ・サイズ(GB)  
1000

メンテナンス・ウィンドウの開始時間 オプション  
なし

データベースの基本情報の指定

管理者資格証明の作成

ユーザー名 ⓘ  
mysqadm

パスワード  
.....

パスワードの確認  
.....

ネットワーク情報の指定

[VCNとサブネットを作成します。](#)

の仮想クラウド・ネットワーク (コンパートメント)  
.....

のサブネット (コンパートメントの変更)  
..... (リージョナル)

ホスト名 ⓘ  
MySQL DBシステムのDNSホスト名の定義

MySQLポート オプション  
3306

MySQL Xプロトコル・ポート オプション  
33060

MySQL DBシステムの作成

DBシステム情報  
データベース情報  
3 バックアップ情報

バックアップの構成

☐ 自動バックアップの有効化  
自動バックアップを有効化します。保存

前 作成 取消

[使用条件およびプライバシー](#) [Cookieの設定](#)

## 2つ目の方法の結果

- 重複レコード数: 1878件 (1878 rows in set (2 days 18 hours 11 min 59.18 sec))
- 所要時間: 66時間11分59秒 ※準備に+2時間程度
- リソース: vm.standard.E2.8 (8core 64GBmemory)
- 費用: 約3,487円(68H)

# 3つ目の方法（2021/01/15）

## ■ Heatwaveを使った重複チェック

<前提条件>

- 1.DBAじゃない人でもできる簡単な手段を使う。
- 2.Heatwaveを使う。（短時間で済ませるため）

## ■ データのロード

```
# mysql --defaults-extra-file=/home/opc/mds.cnf -e ¥  
"LOAD DATA local INFILE 'serial.txt' ignore INTO TABLE testdb.table1 (serial)"
```

## ■ Heatwaveの起動（GUIから）とHeatwave設定

```
mysql> ALTER TABLE table1 SECONDARY_ENGINE = RAPID;  
mysql> ALTER TABLE table1 SECONDARY_LOAD;
```

## ■ 重複したシリアルを抽出&カウント

```
mysql> SELECT serial, COUNT(serial) FROM table1 GROUP BY SERIAL HAVING COUNT(serial) > 1;
```

もっと良い方法をご存じの方は是非ご連絡ください。

# 3つ目の方法（補足：Heatwave構築①）



### Configure placement and hardware

Select a Compartment  
diecoci (root)/central/d2cpn

Select an Availability Domain  
pzcw:AP-TOKYO-1-AD-1

☐ Choose a Fault Domain for this DB System  
If you do not select a Fault Domain, Oracle will choose the best p

Select a Shape

MySQL.Analytics.VM.Standard.E3

CPU Core Count: 16

Memory Size: 512 GB

[Show Configuration Options](#)

Data Storage Size (GB)  
1000

Maintenance Window Start Time *Optional*  
None

### データベースの基本情報の指定

#### 管理者資格証明の作成

ユーザー名 ⓘ  
mysqladm

パスワード  
\*\*\*\*\*

パスワードの確認  
\*\*\*\*\*

#### ネットワーク情報の指定

VCNとサブネットを作成します。

の仮想クラウド・ネットワーク (コンパートメント)

のサブネット (コンパートメントの変更)

(リージョナル)

ホスト名 ⓘ  
MySQL DBシステムのDNSホスト名の定義

MySQLポート オプション  
3306

MySQL Xプロトコル・ポート オプション  
33060

### MySQL DBシステムの作成

DBシステム情報  
データベース情報  
3 バックアップ情報

バックアップの構成

☐ 自動バックアップの有効化  
自動バックアップを有効化します。保存

前 作成 取消

使用条件およびプライバシー Cookieの設定

Qiitaがお勧め : <https://qiita.com/sugimount/items/9ecf8800cd59ba2a635c>

# 3つ目の方法（補足：Heatwave構築②）

Edit Start Stop Restart More Actions

MySQL DB System Information

OCID: ...n2vokxi2yq [Show](#) [Copy](#)

Description: [REDACTED]

MySQL Version: 8.0.22-u4-cloud

Shape: MySQL.Analytics.VM.Standard.E3

Configuration: [MySQL.Analytics.VM.Standard.E3.Built-in](#)

Storage Size: 1,000 GB

State: Active

HeatWave Cluster Attached: No

Automatic Backups: Enabled

Maintenance Window Start: Wednesday 16:38

More Actions

- Edit Backup Configuration
- Add HeatWave Cluster
- Add Tags
- Delete

## Add HeatWave Cluster

### Configure HeatWave Cluster

Select a Shape

MySQL.Analytics.VM.Standard.E3

CPU Core Count: 16

Memory Size: 512 GB

Node Count

2

Number between 2 and 24

Total Memory Size: 1 TB

Estimate Node Count

This operation can take several minutes to complete.

Add Heatwave Cluster Cancel

## HeatWave Cluster Information

Edit Start Stop Restart Delete

Cluster Size: 2 Nodes

Shape: MySQL.Analytics.VM.Standard.E3

State: ● Creating

## HeatWave Nodes

| Node ID        | State      |
|----------------|------------|
| cluster-node-1 | ● Creating |
| cluster-node-2 | ● Creating |

Qiitaがお勧め : <https://qiita.com/sugimount/items/9ecf8800cd59ba2a635c>

# 3つ目の方法の結果

- 重複レコード数: 1878件 (1878 rows in set (32.71 sec))
- 所要時間: **32.71秒** ※準備に+2時間程度
- リソース: MySQL.Analytics.VM.Standard.E3(16core 512GBmemory) x1  
MySQL.Analytics.VM.Standard.E3(16core 512GBmemory) x2
- 費用: 約293円(2H)

(2つ目の方法と比較して)

**Heatwave**を利用したほうが、**7286倍**のスピードで処理が完了した!!

238319.18秒 ⇒ 32.71秒

しかも費用も、**12分の1**でした!!

3,487円 ⇒ 293円



# 4つ目の方法（2021/01/15）

## ■ 更にサイズの大きなMDSを使った重複チェック

<前提条件>

- 1.DBAじゃない人でもできる簡単な手段を使う。
- 2.HeatwaveをDisableにして1台で実施する。（単に比較するため）

## ■ データのロード

```
# mysql --defaults-extra-file=/home/opc/mds.cnf -e ¥  
"LOAD DATA local INFILE 'serial.txt' ignore INTO TABLE testdb.table1 (serial)"
```

## ■ Heatwaveの停止（GUIから）

## ■ 重複したシリアルを抽出&カウント

```
mysql> SELECT /*+ SET_VAR(use_secondary_engine = OFF) */ serial, COUNT(serial) FROM table1  
GROUP BY SERIAL HAVING COUNT(serial) > 1;
```

もっと良い方法をご存じの方は是非ご連絡ください。

# 4つ目の方法の結果

- 重複レコード数: 1878件 (1878 rows in set (2 days 10 hours 5 min 30.46 sec))
- 所要時間: **58時間5分30秒** ※準備に2時間程度
- リソース: MySQL.Analytics.VM.Standard.E3(16core 512GBmemory)
- 費用: 約2,933円(60H)

Heatwaveを利用したほうが、**6393倍**のスピードで処理が完了した!!

209130.46秒 ⇒ 32.71秒

しかも費用も、**10分の1**でした!!

2,933円 ⇒ 293円

# Appendix.

# 5つ目の方法（2021/01/17）

## ■ Google Bigqueryを使った重複チェック

<前提条件>

1.DBAじゃない人でもできる簡単な手段を使う。

## ■ データのロード

```
# bq load --autodetect --source_format=CSV SERIAL2020.TABLE1 ./serial.txt
```

## ■ 重複したシリアルを抽出&カウント

```
> select int64_field_0, count(1) from xxxxxxxx.SERIAL2020.TABLE1 GROUP BY int64_field_0 HAVING  
COUNT(int64_field_0) > 1;
```

もっと良い方法をご存じの方は是非ご連絡ください。

# 5つ目の方法の結果

- 重複レコード数: 1878件      クエリ完了 (経過時間: 13.6 秒、処理されたバイト数: 8.3 GB)
- 所要時間: **13.6秒** ※準備に1時間程度
- リソース: 非公開
- 費用: Always free内利用(1H)

The screenshot displays the Google Cloud Platform BigQuery interface. The left sidebar shows the navigation menu with options like 'BigQuery', 'SQL ワークスペース', 'データ転送', 'スケジュールされたクエリ', '予約', and 'BI Engine'. The main area is titled 'エクスプローラ' (Explorer) and shows a query editor with the following SQL query:

```
1 select int64_field_0, count(1) from [REDACTED].SERIAL2020.TABLE1 GROUP BY int64_field_0 HAVING COUNT(int64_field_0) > 1;
```

Below the query editor, the 'クエリ結果' (Query Results) section is visible. It shows the query execution status: 'クエリ完了 (経過時間: 13.6 秒、処理されたバイト数: 8.3 GB)'. The results are displayed in a table with two columns: 'int64\_field\_0' and 'f0\_'. The first row shows a value of '1' in the 'int64\_field\_0' column and '2' in the 'f0\_' column.

| 行 | int64_field_0 | f0_ |
|---|---------------|-----|
| 1 | 1             | 2   |

# 総括

- すでにCSVやJSONとしてファイルがある場合は、Google Bigquery優位。
- データがMySQLに入っている状態であれば、Heatwave優位。  
⇒最小構成なので、オラクル社がLimit解除してくれたら、再トライします！
- もうAurOraを使うモチベーションもコストも無し…  
⇒挑戦者求む！

**ご清聴ありがとうございました。**



# We're Hiring!

**プロフェッショナル複業でも挑戦可！**



FULLENERGY