

あなたの知らないS3 Selectの世界



山田 雄

自己紹介

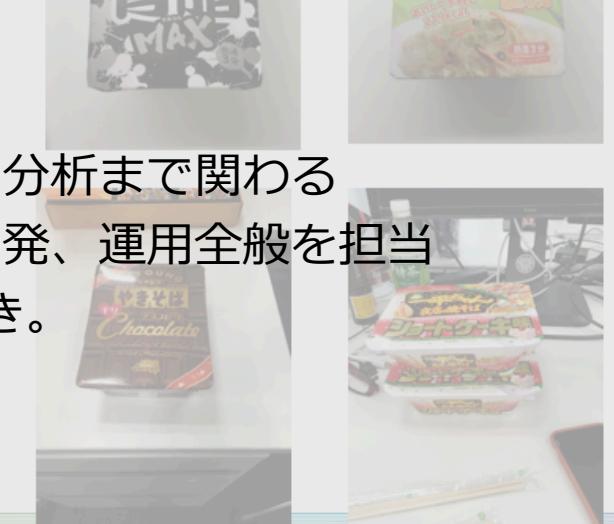
■山田 雄 (ヤマダ ユウ)



株式会社 リクルートライフスタイル
ネットビジネス本部
データ基盤T

Twitter:@nii_yan
GitHub:<https://github.com/yu-yamada>

- ・以前はメールマーケティング用基盤の作成からデータ分析まで関わる
- 現在はリクルートライフスタイルの共通分析基盤の開発、運用全般を担当
- ビックデータ、Ruby、ビール、カップ焼きそばが好き。



最初に

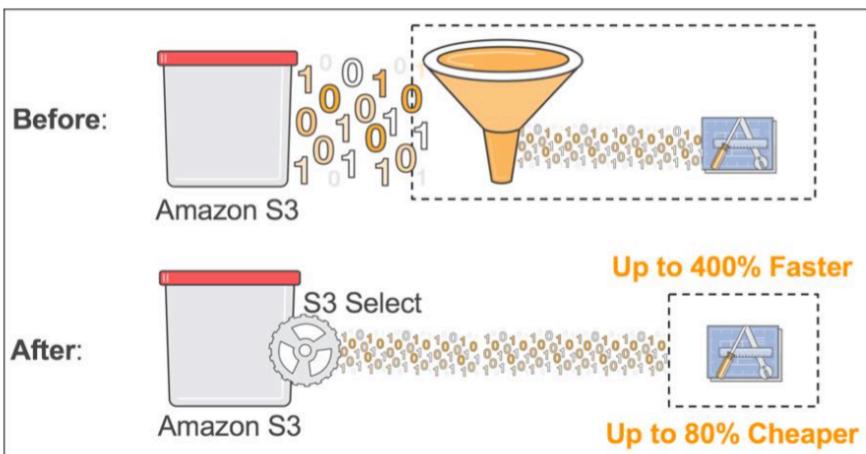
プレビューを触ったのを前提に書こうとしたのですが、4/5になんとGAされたので書いてる内容が色々ぶれています





<Amazon S3 機能アップデート>

オブジェクト全体ではなく、特定のコンテンツをSelect可能なS3 Select
を発表（プレビュー）



<S3 Selectの概要>

• 特徴

- ファイル全体をダウンロードせず必要なデータのみをシンプルなSQLでクエリ可能に
- 標準SQLを使って、オブジェクトからデータをフィルタ可能
- アプリケーションが必要とするデータのみを取得することで、大幅なパフォーマンス向上が期待できる
- AthenaやRedshift/EMRがS3 Selectをサポート予定
- プレビュー期間は無料で利用できCSV/JSONをサポート。圧縮はGZIPのみ対応。暗号化ファイルは現時点では非対応

• 注意点

- 米国東部（バージニア北部）、米国東部（オハイオ）、米国西部（オレゴン）、欧州（アイルランド）、アジアパシフィック（シンガポール）リージョンで利用可能

FileDelimiterを指定できる
ので、CSVだけではなく、
TSV,PSVなどもいける

特徴

- SDK(~~preview~~では追加lib必要) or CLI(~~preview~~では使えないまだ見当たらない) or APIから使用可能
- ~~GUI無い~~
~~AWSコンソールで探しても何も出ません…~~
128MBまでのファイルならコンソールから使用可能
- 今の所S3Select用のロールもない
S3のロールで動く
- IPアドレスはS3selectを叩いているIPアドレスでS3にアクセスしてるぽい
- Presto connectorもあるが、EMR用
- 1ファイルに対してのみ実行可能

コンソール

hoge.csv 最新バージョン ▾

「select from」が日本語だと残念な感じに

概要

プロパティ

アクセス権限

次から選択

S3 Select では、SQL 式を使用して 1 つの CSV または JSON ファイルからレコードを抽出できます。S3 Select 一側の暗号化ファイルがサポートされます。コンソールを使用して、最大 128 MB のソースファイルから最大 40 大きなファイルまたはより多くのレコードを操作するには、API を使用します。 [詳細](#)

より複雑な SQL 式を必要とするデータを S3 で分析する場合は、以下を参照してください。 [Amazon Athena](#)

ファイル形式 [i](#)

CSV

JSON

区切り記号

Comma

Tab

Custom

ヘッダー行 [i](#)

ファイルにはヘッダー行があります

サンプルクエリ

```
select * from S3Object limit 100
```

```
select _1, _3 from S3Object where _3 > '100'
```

```
select count(*) from S3Object
```

```
select min(cast(_1 as int)) from s3object
```

アクセスログ例

```
REST.POST.SELECT test/hoge.tsv "POST  
/XXXXbucket/test/hoge.tsv?select  
HTTP/1.1" 200 - 70 - 69 - "-" "Boto3/1.5.25  
Python/3.6.1 Darwin/15.3.0  
Botocore/1.8.39" -
```

サンプルコード

```
response = s3.select_object_content(←
    Bucket='buket_name',←
    Key='test/s3select_test.csv',←
    SelectRequest={←
        'ExpressionType': 'SQL',←
        'Expression': 'Select * from S3Object s',←
        'InputSerialization': {←
            'CompressionType': 'NONE',←
            'CSV': {←
                'FileHeaderInfo': 'Use',←
                'RecordDelimiter': '\n',←
                'FieldDelimiter': ',',←
            }←
        },←
        'OutputSerialization': {←
            'CSV': {←
                'RecordDelimiter': '\n',←
                'FieldDelimiter': ',',←
            }←
        }←
    }←
)}
```

サポートされている型一覧

Type	Example
null	NULL
missing	MISSING
bool	FALSE
int	10000
string	'xyz'
float	CAST(0.456 AS FLOAT)
decimal	123.456
timestamp	CAST('2007-04- 05T14:30Z' AS TIMESTAMP)
tuple(struct)	{'a':5}
list(array)	[1, 'a', FALSE]

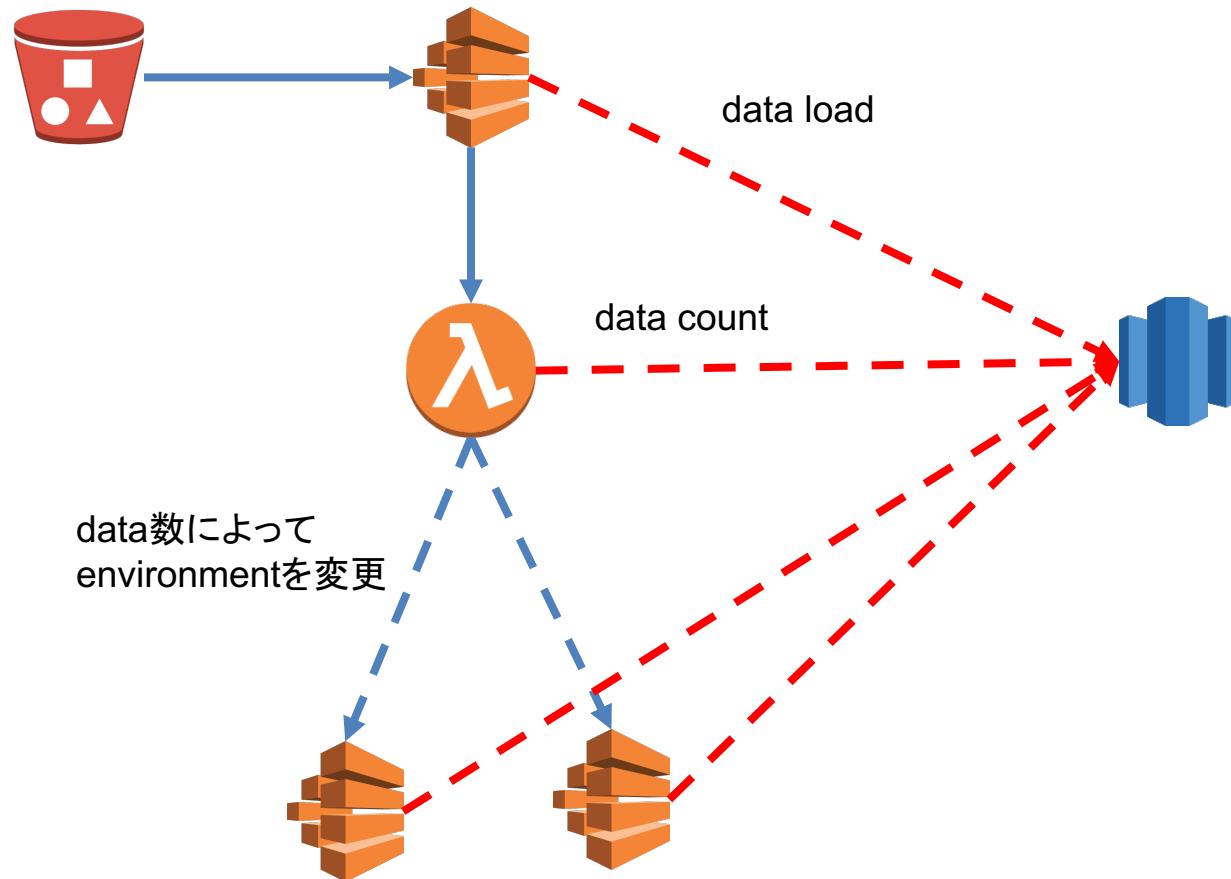
読み込みファイル設定可能オプション

Name	Default	Valid values	note
CompressionType	NONE	NONE or GZIP	圧縮方式
CSV or JSON		CSV or JSON	ファイルフォーマット
RecordDelimiter	¥n		改行文字
FieldDelimiter	,		区切り文字
QuoteCharacter	"		エスケープ文字
quoteEscapeCharacter	"		quoteのエスケープ文字
FileHeaderInfo	None(?)	None Use Ignore	ヘッダー情報
Comments	#		コメント文字

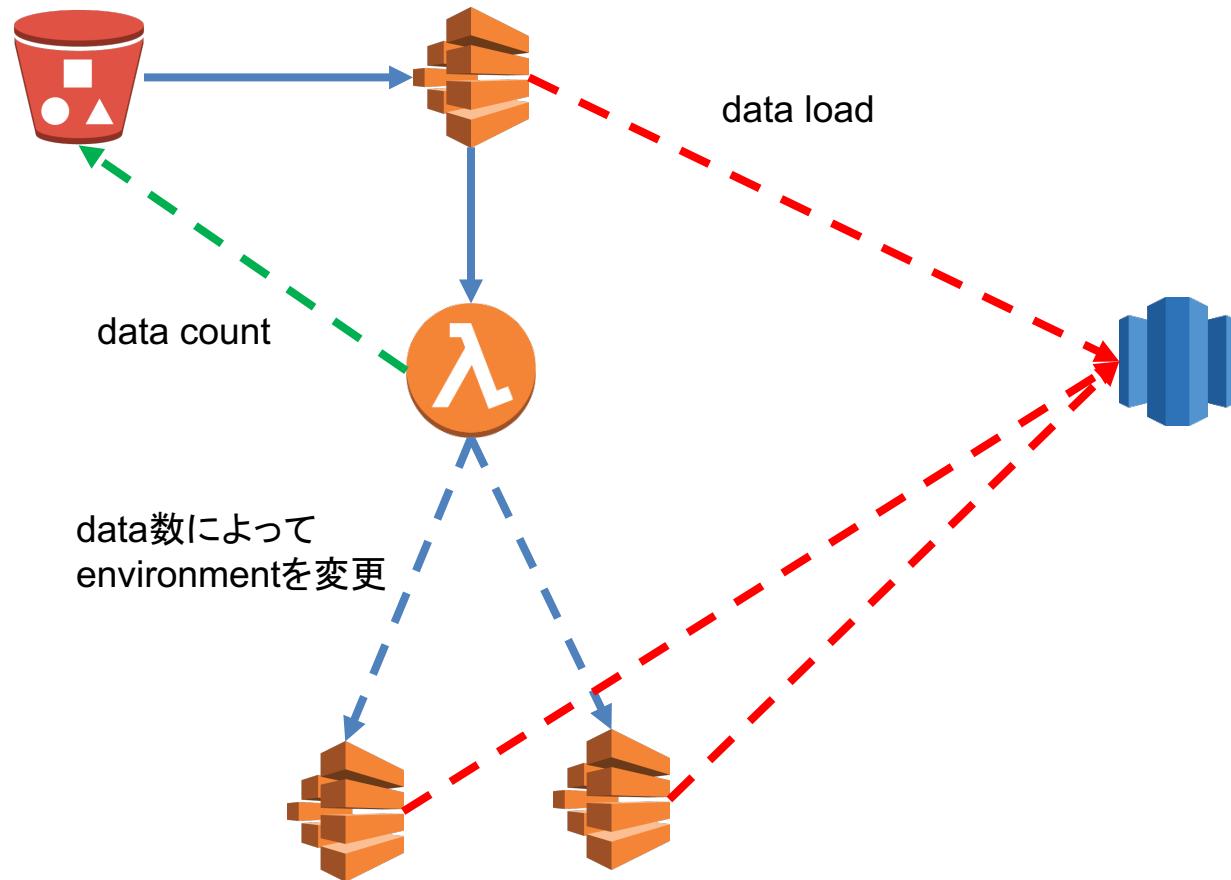
料金体系

- スキャンした合計データ量
 - GBあたり0.002USD
- S3 Selectによって戻されたデータ量
 - GBあたり0.0007USD

こんな感じに利用出来そう



こんな感じに利用出来そう



こんな感じに利用出来そう

s3://examplebucket/20180222/スキーマ名/テーブル名/…

- as is

ディレクトリ名やファイル名に意味を持たせる

- to be

ファイルの中身にも意味を持たせる

Demo

宣伝

4月
18

【RLSMeetup#8】 パブリッククラウドがドライブさせるリクルートライフスタイルのプロダクト開発 ★

クラウドアーキテクト

主催：リクルートライフスタイル



RECRUIT
リクルートライフスタイル
MeetUp

ハッシュタグ : #RLSMeetup