ソフトバンク接続率 No.1 達成の裏で
活躍したビッグデータ解析と
さらなる発展について

2015年6月3日

株式会社Agoop 代表取締役 柴山 和久、加藤 有祐

Redshift?

Why Redshift？

シンプルSQL

Hive?
Presto?

PostgreSQL

初期費用不要

容量単価が安い

導入が早い

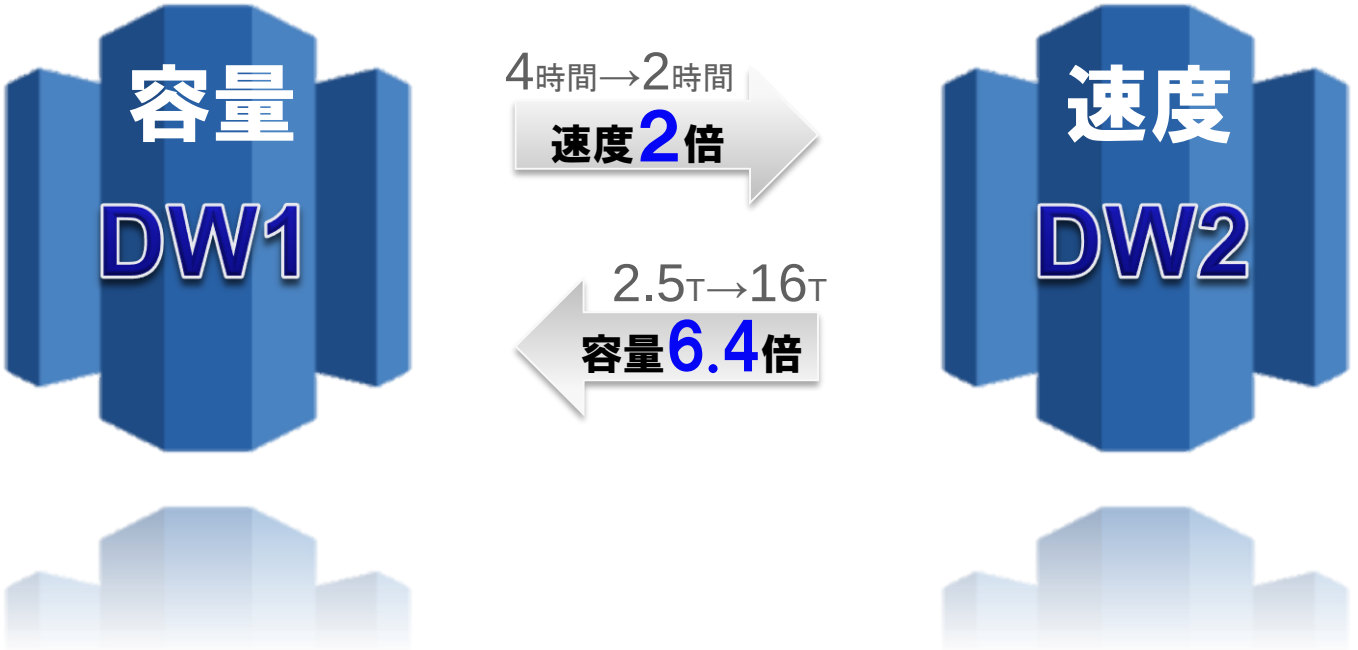
拡張性

操作性

日々110クエリ

解析30種類

DW1/DW2の使い分け

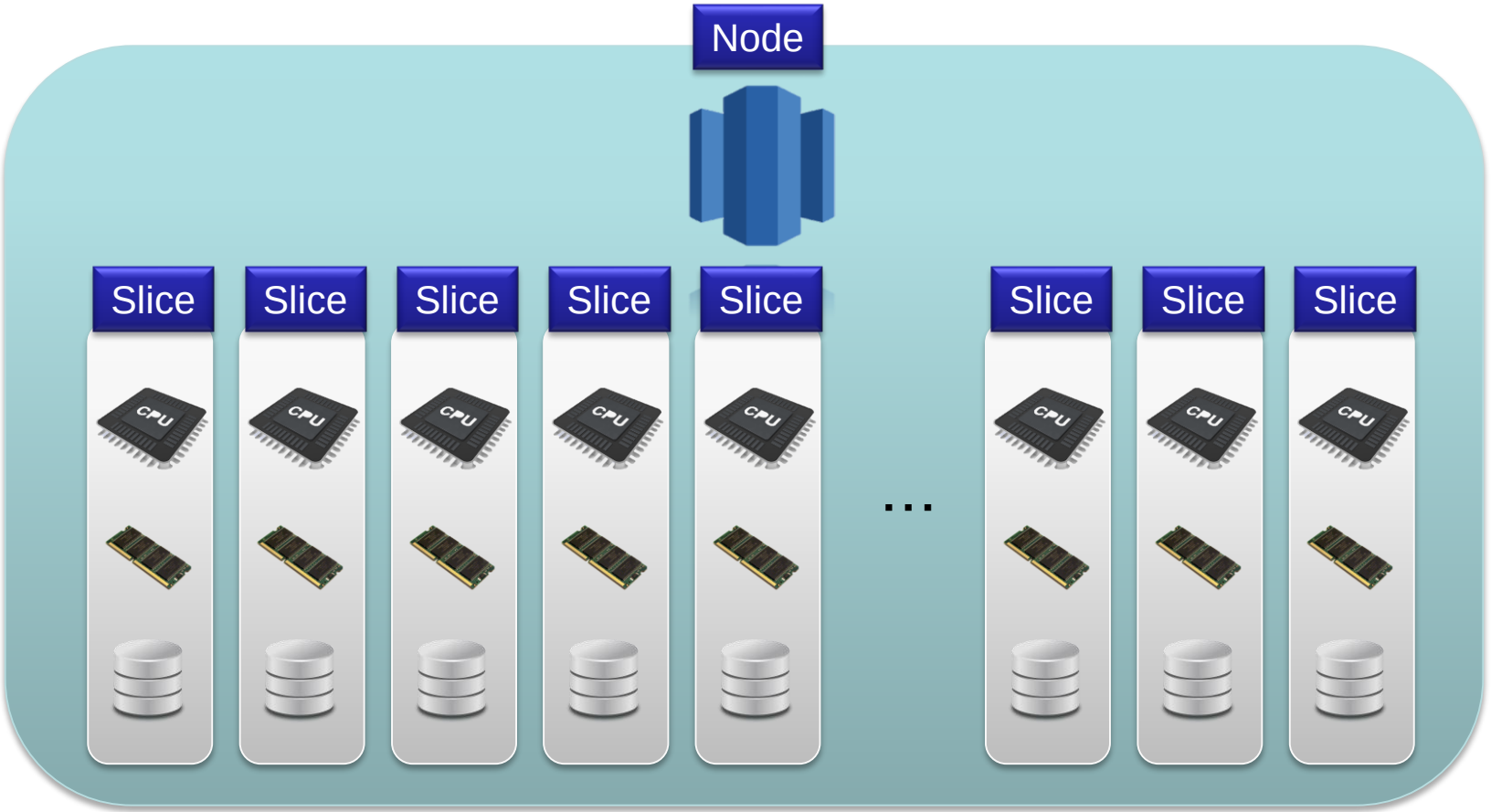


長期間のログ保存
アドホック解析

短期間のログ保存
日次解析


さらなる高速化へ

分散キーはパフォーマンスの最重要



分散キーはパフォーマンスの最重要

分散キーの見直しで処理速度2倍！

user_id	action_id
1	2
2	2
2	3
2	5
5	2

分散キーを正しくセットしないと

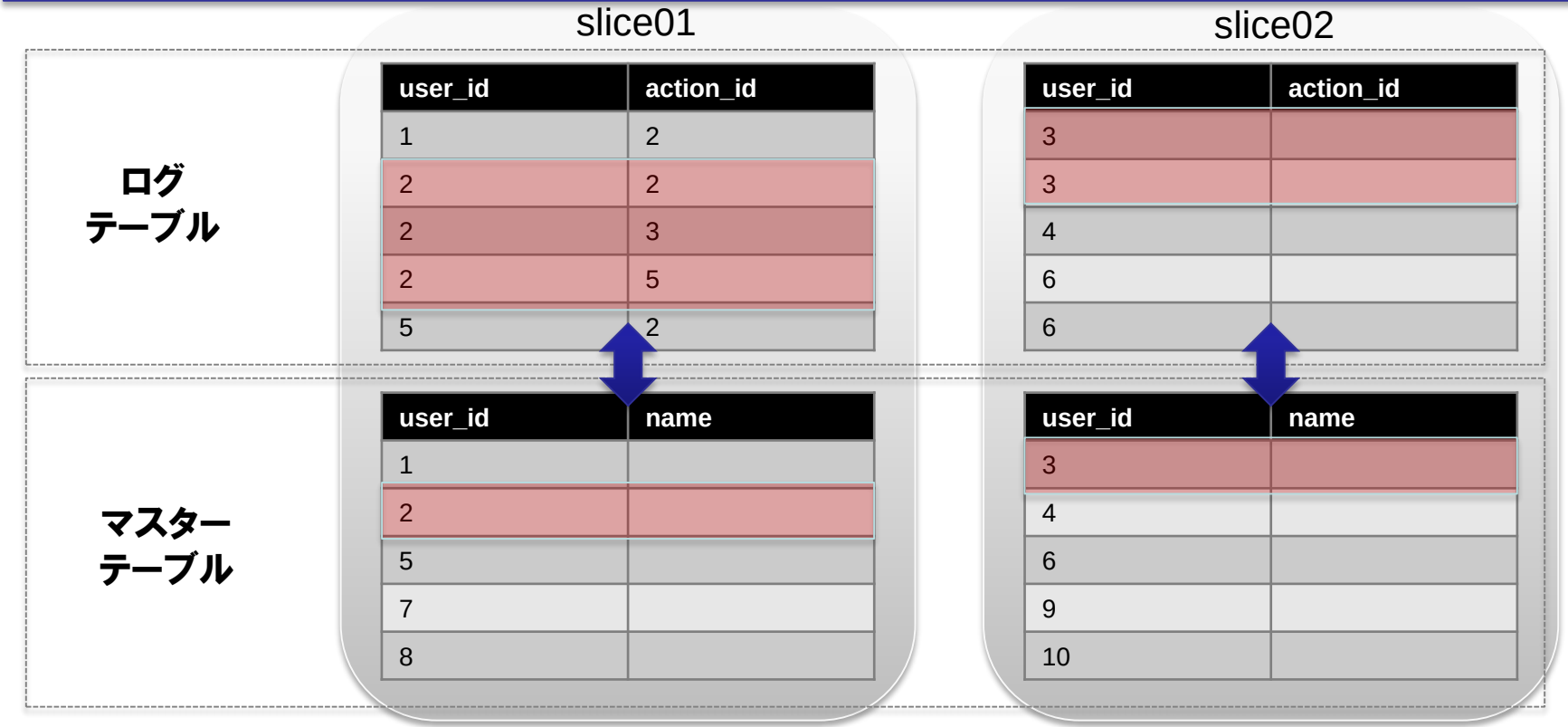
データの**再移動**が発生

分散とクエリがマッチすると**ノード追加**の**効果大**

- ID系をセット
- JOINで使う項目
- Group Byする項目

分散キー効果のイメージ

結合キー，GroupByにセットする ＝ 再分散しない



クエリチューニング（改善例）

DW1 x 2node クエリチューニング 2.5時間 → 40分

大量データテーブルのIDを振り直す

Before

Column1(uniqueid)
Column2
Column3
….
Column39
Column40

row_number()

ノード間の大量データ移動発生

INSERT INTO
SELECT

After

Column1(uniqueid)
row_number()

Tempテーブル

Column1(uniqueid)
Column2
Column3
….
Column39
Column40

JOINしてINSERT

Copyコマンドの高速化

Amazon S3

applicationlog01.gz
applicationlog02.gz
applicationlog03.gz
applicationlog04.gz
applicationlog05.gz
applicationlog06.gz
applicationlog07.gz
applicationlog08.gz

1ファイルずつCopy

applicationlog01.gz
applicationlog02.gz
applicationlog03.gz
applicationlog04.gz
applicationlog05.gz
applicationlog06.gz
applicationlog07.gz
applicationlog08.gz

一括指定でCopy

Redshift

2000万件（40GB）1時間 → 15分へ短縮