

ビッグデータ分析における 集計処理速度とコストの最適化手法

氏名：湯浅 雅人 e-mail：m.yuasa0617@gmail.com

開発における問題点

近年、ビッグデータ分析のためのデータ集計基盤は、集計速度向上や安定性向上を目的に、オンプレミス環境からクラウド環境への移行が盛んに行われている。しかし、クラウド化により集計処理の速度向上（スケールUP）が容易になる反面、再現なくコストが増加し得る。

手法・ツールの適用による解決

1. 集計システム（集計処理）に対する要求を整理
2. 集計システム（集計処理）の数理モデル化
3. モデルに対して最適化 & 求根手法の適用
上記手法によりビッグデータ分析のための集計システムに対する課題を解決する。

集計システムに対する要求

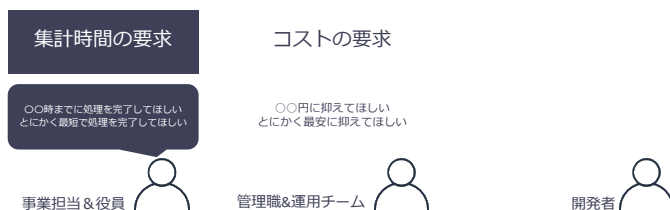
対象システム（AWS Glue）

本修了制作ではAWSが提供するデータ分析用データ集計（ETL）サービスGlueが対象



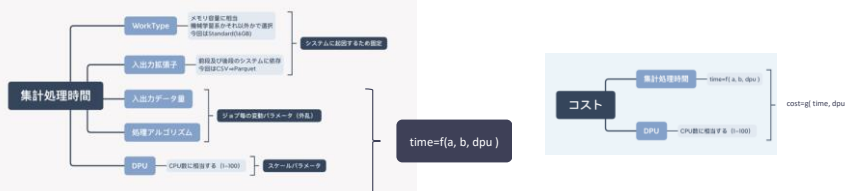
システムへの要求

集計データ提供に対する時間及びコストの要求を整理



集計システムのモデル化

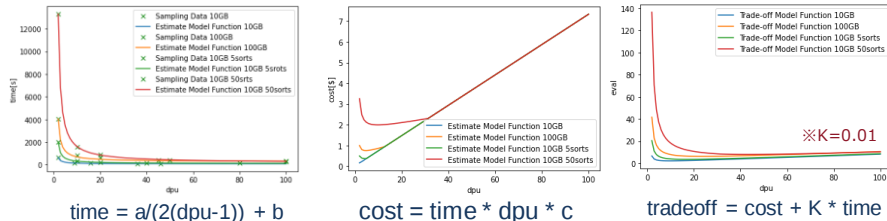
集計時間 & コストの影響因子



集計時間 & コスト & トレードオフのモデル化

上記影響因子（パラメータ）を変動させて、集計時間 & コストを計測

⇒ 実測結果をもとに集計時間およびコストをモデル化 & トレードオフ評価関数を定義



集計システムの最適化

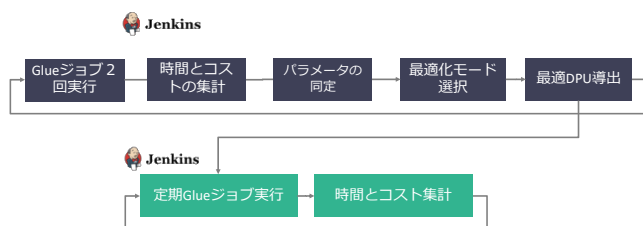
最適化及び求根手法の適用

本事例では探索すべきパラメータ空間小
⇒ 全探索で解導出が可能



最適化を実現するためのプロトタイプシステム

最適化検証のために以下のシステムを構築し、最適化パラメータ導出と検証実験を行った。



最適化検証

検証パターン

以下の処理パターンに対して最適化実施

NO.	入出力データ量	アルゴリズム負荷
1	データ量小 10GB	負荷小 拡張子変更
2	データ量大 100GB	負荷小 拡張子変更
3	データ量小 10GB	負荷大 拡張子変更 + ソート50回

検証結果

コスト最適化、集計速度最適化、トレードオフ最適化の結果

データ量	ソート回数	最適化	コスト	集計時間
10GB	0	コスト最適化	\$0.16	641.0秒
小	小	トレードオフ..	\$0.59	160.6秒
10GB	50	時間最適化	\$7.33	86.2秒
小	大	コスト最適化	\$2.00	2,044.0秒
10GB	0	トレードオフ..	\$3.08	487.8秒
100GB	0	コスト最適化	\$0.80	2,169.4秒
大	小	トレードオフ..	\$1.61	460.7秒
100GB	50	時間最適化	\$7.33	819.0秒

まとめ & 今後の取り組み

まとめ：

各要求に対して最適化の実現

今後の取り組み：

本番環境を想定したシステム構成検討
複数集計ジョブの組み合わせ最適化

組み合わせ最適の例

