トップエスイー修了制作

データサンプリングと回帰分析による Hiveクエリ実行時間の予測

株式会社東芝　　石川和典　　kazunori3.ishikawa@toshiba.co.jp

開発における問題点

ビッグデータ分析において広く用いられる Hiveクエリ (**HQL**)

- 集計対象のデータ量が大きい場合，実行には長時間を要する
- HQLのパフォーマンスは分散実行環境に依存

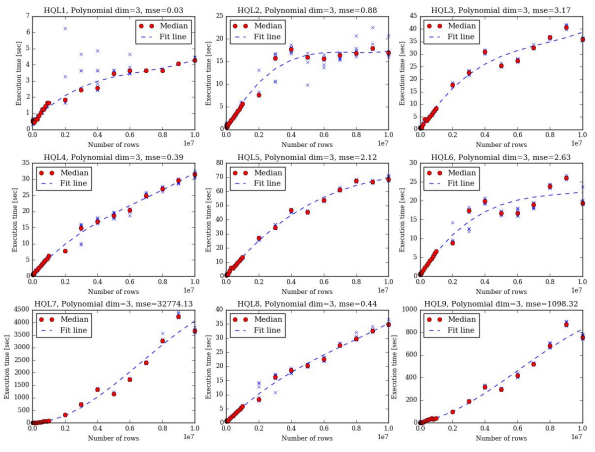
手法・ツールの適用による解決

HQLの実行時間を動的解析により予測

- テーブルをランダムにサンプリング → サンプルテーブルを作成
- 複数のサンプルテーブルの実行時間から，目的のテーブルでの実行時間を回帰分析により予測

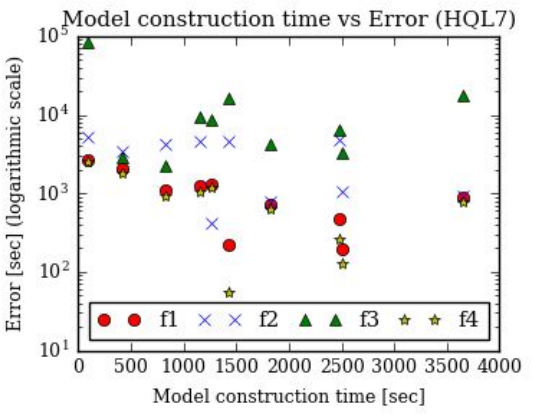
予測を行うための調査

RQ1. レコード数とHQL実行時間に規則性はあるか？→ ある



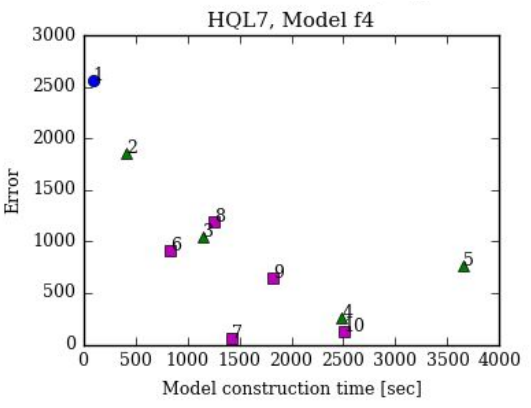
様々なHQLのテーブル行数に対する実行時間と，3次関数へのフィッティング結果

RQ2. 予測にはどの回帰モデルが適切か？→ $f(x) = ax\ log(x) + b$



$f_1(x) = ax + b$,　$f_2(x) = ax^2 + bx + c$,
$f_3(x) = ax^3 + bx^2 + cx + d$,　$f_4(x) = ax\ log(x) + b$

RQ3. モデルの学習データとして，どのようなサンプルセットが有利か？→ レコード数の多いサンプルを含むもの



●: 基準，▲サンプルの種類を増加，
■: サンプルのレコード数を増加

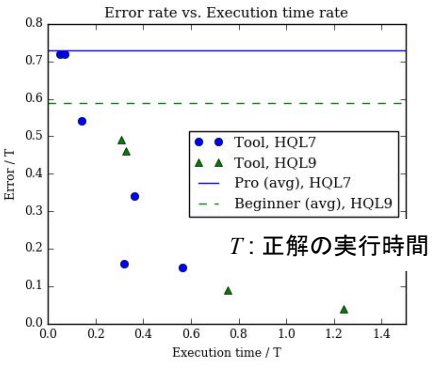
自動予測ツール

入力: HQLファイル，対象テーブル，制限時間

```
$ python pphq.py ./test.hql test_table 600
```

回帰モデル: $f(x) = ax\ log(x) + b$

1. 初期サンプルセット $\{r_1=10^5, r_2=10^6\}$ → 実行時間 t_1, t_2 を計測
2. $t_1, t_2, f(x)$ と残りの制限時間から r_3 を決定 → 実行時間 t_3 を計測
3. $r_1, r_2, r_3, t_1, t_2, t_3$ から $f(x)$ をフィッティング → $f(\mathrm{R})$ を予測値とする



T: 正解の実行時間

HQL	ツール		専門家	初学者
	Time [sec]	Error [%]	Error [%]	Error [%]
7	186	72	64	2.7兆
7	1176	16	84	84
Avg(7)	921	49	73	1.3兆
9	232	49	104	100
9	570	9.3	17	17
Avg(9)	497	27	316	59

- 実行時間に応じて予測値の誤差が減少
- 実行時間によっては専門家を超える予測値
  - 平均で専門家を超える予測値