時系列ストリームデータ統合利用システム

System for integration of time series streaming data

筑波大学システム情報工学研究科コンピュータサイエンス専攻　研究代表者：北川博之
研究協力者：渡辺陽介，山田真一

研究背景

- 多種多様な大量の**データストリーム**が利用可能
  - データ放送，センサーネットワーク，ログ情報，Web
  - 時々刻々と変化する情報を逐次送信
- データストリームに対する高度利用要求の増大
  - 大勢の利用者から大量の問合せ処理要求が発生

⇒ **ストリームに対する各種処理要求を実現するシステムが求められている**

研究目的

ストリームに対する問合せ処理を実現するエンジンの開発

1. 問合せ処理を実現するシステムの開発
2. ストリームデータをアーカイブし，新規到着データと過去データを統合利用できる仕組みの開発
3. ストリーム問合せ処理の最適化

こんな問題を解決！

ストリームをモニタリングするアプリケーションを効率よく開発したい！

⇒ 本システムはストリームデータに対する問合せ言語を提供します
さらに，アプリケーション開発をサポートするためのAPIを提供します

履歴データは大量にあるが，活用する術がない！

⇒ 本システムは既存のデータベースとストリームデータを統合することができます
履歴データから，ストリームのパターンとマッチするパターンを抽出できます

ストリームに対する処理要求が多すぎて応答が遅い！

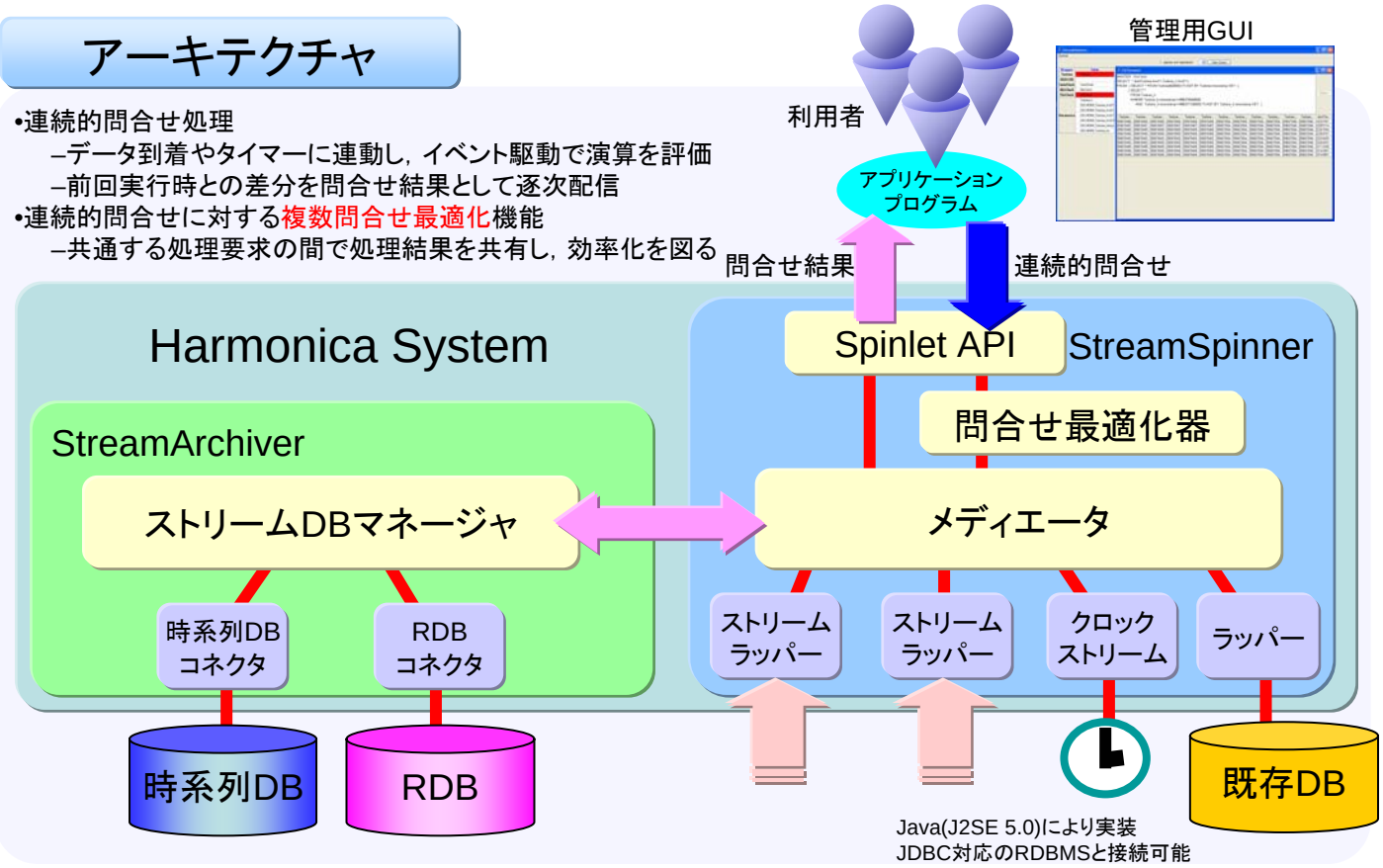
⇒ 本システムの問合せ最適化機構により，大量の処理要求が与えられても大丈夫！

想定される応用例

- システム機器や環境情報のモニタリング及び予測
- パーソナライズされた情報配信
- ユビキタスデータ管理等

アーキテクチャ

- 連続的問合せ処理
  - データ到着やタイマーに連動し，イベント駆動で演算を評価
  - 前回実行時との差分を問合せ結果として逐次配信
- 連続的問合せに対する**複数問合せ最適化**機能
  - 共通する処理要求の間で処理結果を共有し，効率化を図る



データモデル

ストリームのモデル化

ストリームSのヒストリ部
StreamArchiver が保持する

timestamp	・・・	value
9754434…	・・・	87.2
9754434…	・・・	87.2

古くなったデータはヒストリ部へ送られる

ストリームSのカレント部
StreamSpinner が保持する

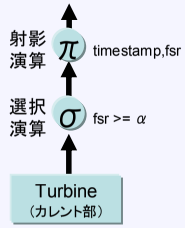
timestamp	・・・	value
9867664…	・・・	87.2
9867664…	・・・	87.3

新規到着データが順次追加される

ストリームに対する問合せ

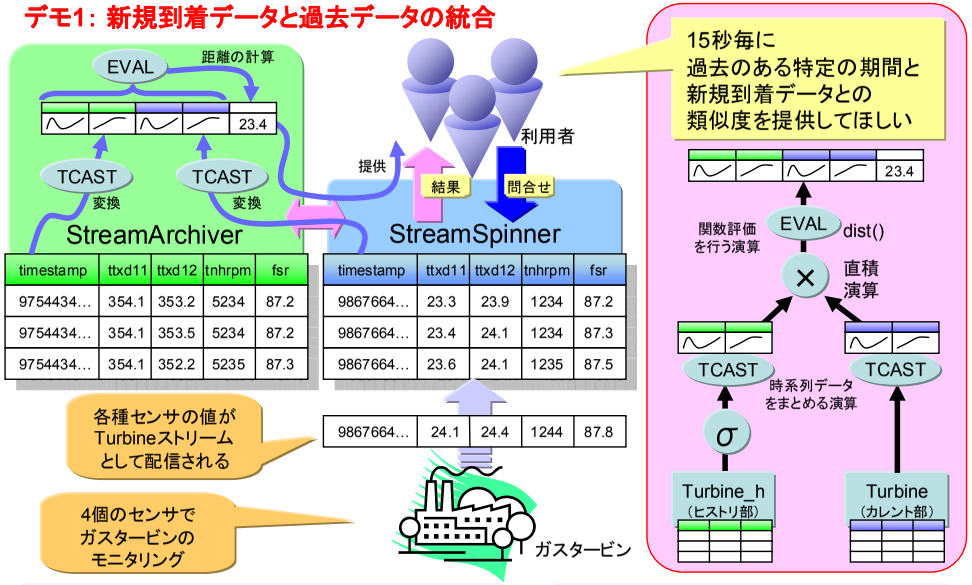
SQLライクな問合せ言語SpinQLを提供

```
MASTER Turbine
SELECT timestamp, fsr
FROM Turbine[1]
WHERE fsr >= α
```



MASTER節は
問合せ実行のきっかけを
与えるストリームを指定する

デモ1：新規到着データと過去データの統合



timestamp	ttxd11	ttxd12	tnhrpm	fsr
9754434…	354.1	353.2	5234	87.2
9754434…	354.1	353.5	5234	87.2
9754434…	354.1	352.2	5235	87.3

timestamp	ttxd11	ttxd12	tnhrpm	fsr
9867664…	23.3	23.9	1234	87.2
9867664…	23.4	24.1	1234	87.3
9867664…	23.6	24.1	1235	87.5
9867664…	24.1	24.4	1244	87.8

TimeSeries型

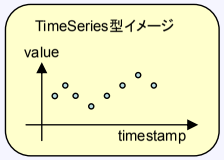
時系列データに対する解析機能をリレーションの枠組み内で扱うための抽象データ型

Timestamp型の値と数値型の値の対の配列で構成

ttxd11	ttxd12
{(98…,34.3),…}	{(98…,33.2),…}

TCAST　RCAST

timestamp	ttxd11	ttxd12
9867664..	34.3	33.2
9867664..	34.4	33.8
9867664..	34.7	34.2



TimeSeries型とリレーションの相互変換するための演算として
TCAST・RCAST演算を導入

EVAL演算

与えられたリレーションから，属性を1つ追加した
リレーションを作り出す演算

追加された属性には，各タプルを関数で評価した
結果が入る

ttxd11	ttxd12	dist(ttxd11,ttxd12)
{(98…,34.3),…}	{(98…,33.2),…}	1.349

EVAL dist(ttxd11,ttxd12)

ttxd11	ttxd12
{(98…,34.3),…}	{(98…,33.2),…}

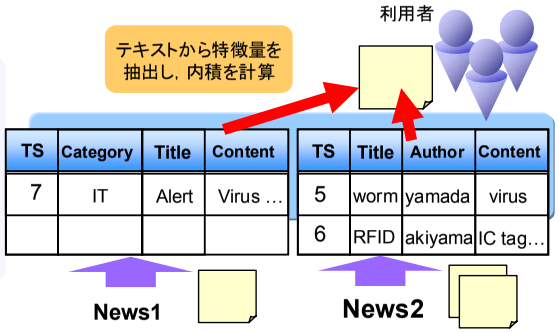
関数の例
dist・・・2つのTimeSeries型の距離を計算
sim・・・2つのテキストストリームの類似度を計算

デモ2：テキストデータの統合

要求例

ニュースサイトから記事が届いたときに，同じ日に届いた別のニュースサイトの記事の中から，類似度0.4以上のものをつけて提供して欲しい

```
MASTER News1
SELECT *
FROM News1[now], News2[24hour]
WHERE sim(News1.Content, News2.Content) >= 0.2
```



TS	Category	Title	Content
7	IT	Alert	Virus …

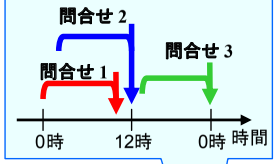
TS	Title	Author	Content
5	worm	yamada	virus
6	RFID	akiyama	IC tag…

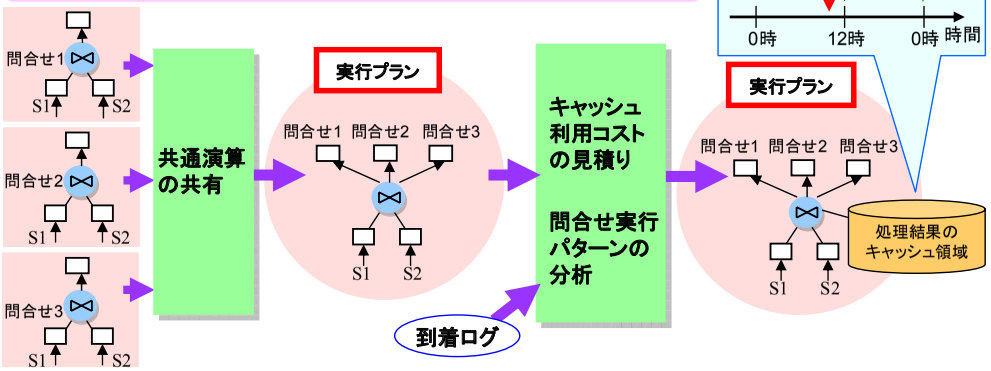
複数問合せ最適化

基本的アイデア

- 共通演算を持つ問合せのうち，近いタイミングで起動する問合せは多くの処理結果が共有可能
  - 処理結果をキャッシュ領域に保存し，別の問合せのために再利用する
- 最適化手順
  1. 問合せから共通演算を抽出し，実行プランを生成
  2. ログを元に各問合せのキャッシュ利用のコストを見積る

- 問合せ1の処理結果は問合せ2のためにキャッシュに保存
- 問合せ3は他の問合せと処理結果を共有しない





評価実験

実験環境

CPU	Pentium 4 3GHz
メモリ	1.5GB
OS	Windows XP
言語	Java(J2SE5.0)
RDBMS	HSQLDB 1.8.0

実験データ

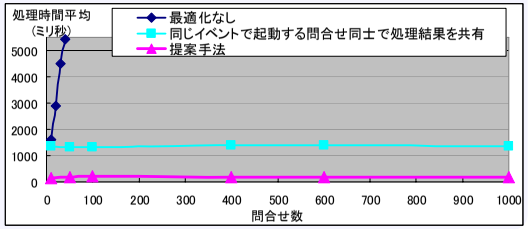
人工的に生成した擬似データ
- ストリームS：毎秒100タプル配信
- RDB：100個のタプル

評価問合せ
- ストリームとRDBをIDで結合する
- 10種類のクロックからひとつを指定
- 各問合せは1分周期で起動

```
MASTER Clock1min
SELECT *
FROM S[1min], RDB
WHERE S.ID = RDB.ID
```

実験結果

複数個(10～1000)の問合せを10分間実行した場合の処理時間を計測し，平均値を計算



関係する助成制度

科学技術研究費補助金　特定領域研究(2) (#16016205)
基盤研究(B) (#15300027)
特別研究員奨励費 (15・330)

科学技術信仰機構戦略的創造研究推進事業CREST
「自律連合型基盤システムの構築」

主な公表論文

"A Multiple Continuous Query Optimization Method Based on Query Execution Pattern Analysis", Proc. DASFAA2004, pp.443-456.

「連続的問合せに対する複数問合せ最適化手法」
電子情報通信学会論文誌, Vol. J-87-D-I, No. 10, pp. 873-886, 2004年.

「時系列データベースとストリーム処理の統合のためのデータモデルの提案」
情報処理学会研究報告 Vol.2005, No.68(2005-DBS-137(II)), pp.299-306.
電子情報通信学会技術研究報告 Vol.105, No.172, pp.25-30.