解説 2

ストリーム・コンピューティング時代を開く基盤ソフトウェアIBM InfoSphere™ Streams

ストリーム・コンピューティングとは、複数の情報ソースより時々刻々到達するデータを複合的に分析・判断し、迅速な意思決定を可能にする計算パラダイムのことであり、データ爆発の時代といわれる今日、業界の注目を集めているものです。IBM 基礎研究部門では、2003年という早い時期から、ストリーム・コンピューティングの可能性に着目し、プロジェクトを立ち上げ、研究開発と実証実験を重ねてきました。その成果は、昨年 IBM InfoSphere Streams という製品に結実しました。本稿では、まず、InfoSphere Streams の特長を、開発環境、実行環境、およびツールキットという観点から解説します。次いで、これまでの実証実験の中から、金融、通信、学術におけるものを紹介します。最後に、InfoSphere Streams の今後の展開についてご紹介します。

Article 2

IBM InfoSphere Streams: Software Platform for the Stream Computing Era

Stream Computing continuously analyzes massive amounts of incoming data streams, enabling faster decision-making. Facing an explosive growth of data, the industry is paying increasingly more attention to this technology. With forethought on its significant potential, IBM Research launched a project on Stream Computing as early as 2003, building infrastructure and conducting experiments with select clients. These efforts eventually lead to the announcement of an IBM product, “IBM InfoSphere Streams,” on May 13, 2009. In this article, we first describe IBM InfoSphere Streams from its development environment, runtime environment and toolkits. We then present three experiments, one each from financial markets, telecommunications, and science. Finally, we futurize Stream Computing and IBM InfoSphere Streams.

❶ はじめに

ストリーム・コンピューティングとは、複数の情報ソースより時々刻々到達するデータを、複合的に分析・判断し、迅速な意思決定を可能にする計算パラダイムのことです。代表的な応用例としては、証券市場からの取引データを分析・判断し、自動売買を行う「アルゴリズム・トレード・システム」や、監視カメラからの画像データを分析・判断し、異常を検知する「ビデオ監視システム」などがあります。

情報ソースより時々刻々到達するデータのことをストリーム・データと呼びますが、今日の情報ソースの多様化により、ストリーム・データの量は増加の一途をたどっています。例えば、GPS 機能搭載の携帯電話からの位置情報、自動改札機からの入出場情報、情報配信会社からのニュース、SNS（Social Networking Service）におけるユーザーの行動履歴、Twitter でのつぶやきなど、枚挙にいとまがありません。データ量の増加はまさに爆発的という形容が相応しく、ストリーム・コンピューティングの潜在的応用範囲もまた爆発的に拡大しているといってよいでしょう。

IBM 基礎研究部門では、2003 年という早い時期から、ストリーム・コンピューティングの可能性に着目し、研究プロジェクトを立ち上げ、ストリーム・コンピューティングのためのプラットフォームを構築し、実証実験による検証を重ねてきました。その成果は昨年、InfoSphere Streams という製品に結実し、5 月 13 日に当該製品に関する新聞発表が行われ、国内外で大きな関心を呼びました［1］。

現在 IBM では、地球規模の課題を IT の活用により解決するという Smarter Planet というコーポレート・ビジョンを提唱していますが、それを裏付けるものとして、3 つの技術的進歩を掲げています。すなわち、あらゆるものにデジタル機器が搭載され、リアルタイムで状態を把握すること（Instrumented）、それらが互いに接続されて連携可能な状態になること（Interconnected）、

高度な予測能力を持つこと（Intelligent）です。これらはそのままストリーム・コンピューティングの世界をほうふつさせるものであり、「InfoSphere Streams は Smarter Planet 実現の基盤である」と多くの人が考えているゆえんでもあります。

本稿では、まず InfoSphere Streams の特長を紹介し、次いで、これまでの実証実験の中から、金融、通信、学術におけるものを紹介します。最後に、InfoSphere Streams によって可能となるストリーム・コンピューティングの今後について論考します。

❷ InfoSphere Streams の特長

InfoSphere Streams は大別して、開発環境、実行環境、およびツールキットから構成されます。

InfoSphere Streams における開発環境では、独自に開発したストリーム・コンピューティング用のプログラミング言語を用いてプログラムを開発します。この言語は SPL（Streams Processing Language）と呼ばれ、オペレーターを組み合わせてプログラムを構成します。1 つのオペレーターは 1 つの処理単位に相当し、プログラムは処理の流れに即して構成されます。そのため直観的で分かりやすく、簡潔なものとなっています。

SPL が提供する基本オペレーターの幾つかを紹介しますと、フィルタリング処理（データの間引き）を行う Functor オペレーター、複数の入力ストリームを 1 つの出力ストリームに結合する Join オペレーター、1 つの入力ストリームを複数の出力ストリームに分散させる Split オペレーターなどがあります。またスライディング・ウィンドウをサポートする Aggregate オペレーターは、ストリーム・コンピューティングらしいもので、例えば指定の数のデータが到着するたびに、合計を取ったり、最大値を取ったり、平均値を取ったりするなどの処理を記述することができます。株価の移動平均の計算などは、このオペレーターを使うことで一行で記述することができます。情報ソースからのデータ取得を担当する Source オペレーターは、ファイル、データベース、ネットワーク・ポートなどさまざまな情報ソースをサポートしており、業界に特化した情報ソースとして、IBM WebSphere® Front Office（市場データの配信プラットフォーム）と接続することも可能です。

さらに Streams Studio と呼ばれる統合開発環境も提供されており、SPL によるプログラムの作成、編集、テスト、デバッグなどを強力に支援します。例えば、Application Graph View では、プログラムをグラフ表示する機能があり（**図 1**）、開発者のプログラム理解を支援します。また、Streams Live Graph では、走行中のプログラムをモニタリングすることが可能で（**図 2**）、性能上のボトルネック発見を支援します。

実行環境について述べますと、InfoSphere Streams は、1 台から最大 125 台までのクラスター環境をサポートしています。SPL で書かれたプログラムは、クラスター環境に適切に分配されて実行され、複数のプログラムを同時に実行することも可能です。データ量が少ないうちは小さなクラスター環境で実行し、データ量の増加とともにクラスター環境を増強していくというのが典型的なシナリオであり、その際 SPL のプログラムは再コンパイルす

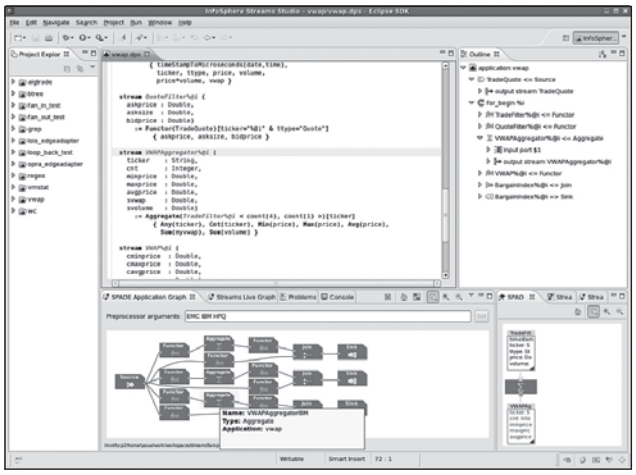

図1. Streams Studioによるプログラム開発

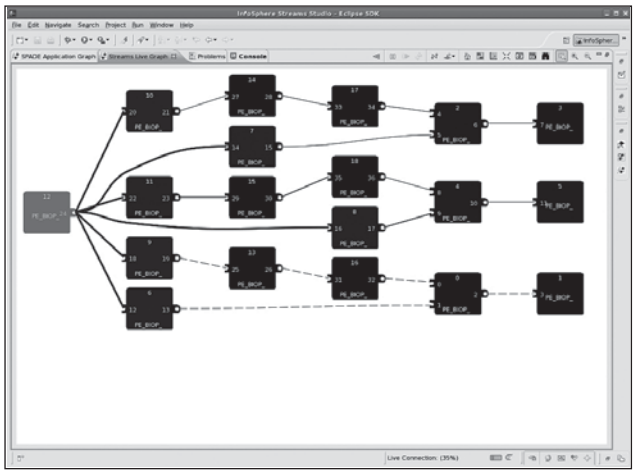

図2. Streams Live Graphでのモニタリング

るだけでよく、書き換える必要はありません。

最後にツールキットについて紹介します。ストリーム・コンピューティングの応用を考えるとき、不正検知、顧客選別、（携帯電話の）解約防止、侵入検知、といったものが考えられますが、いずれもストリーム・データに対する高度な分析が必要であり、上述の基本オペレーターから組み上げていくのは得策ではありません。そこで、InfoSphere Streams Data Mining Toolkitというものが提供されており、決定木、単純ベイズ、ロジスティック回帰、人口分析クラスタリング、線型回帰、多項式回帰といったデータ・マイニング・アルゴリズムがサポートされています。さらに、業界に特化したツールキットとして、証券業界向けのInfoSphere Streams Financial Services Toolkitというツールキットも提供されています。

❸ 実証実験

Streamsを用いた実証実験は世界各国で行われており、半導体製造工場における実時間異常検知、医療分野における患者モニタリング、河川の水質管理モニタリング、ボットネットなどサイバー・セキュリティー対策など多岐にわたっています。本章では通信業界、証券業界、および学術分野における事例を紹介します。

3.1 通信業界の事例

通信業界において国外の企業と実施した実証実験について紹介します。

この実証実験における課題は、通話の際に発生するCDR（Call Detail Records：通話明細レコード）データ量が加入者数の増加に伴い肥大化し、1日数十億データの処理に膨大な時間がかかっていることです。CDRデータは、月次の請求処理や使用パターンの分析による料金設定、マーケティング分析などに使用されますが、この前処理として（1）バイナリー・データからASCII（ASN.1）データへの変換、（2）数百の変換ルールに及ぶCDRの加工処理、（3）交換機やネットワーク障害におけるCDRデータの重複検出といった処理が必要とされます。この実証実験では（1）、（2）、（3）の処理をInfoSphere Streamsを用いて実装し、高速に処理できることを実証しました。

高速化以外にも幾つかの効果がありました。従来のバッチ処理ではデータを一旦ストレージに蓄えてから分析処理を施しますが、InfoSphere Streamsではデータが到着し次第、順次処理していくことになります。それによりバッチ処理と比較して、処理が平準化され、また、必要なデータのみを蓄えることとなるため、ストレージ量を削減することができました。さらに、SPL言語による開発生産性の向上も確認されました。

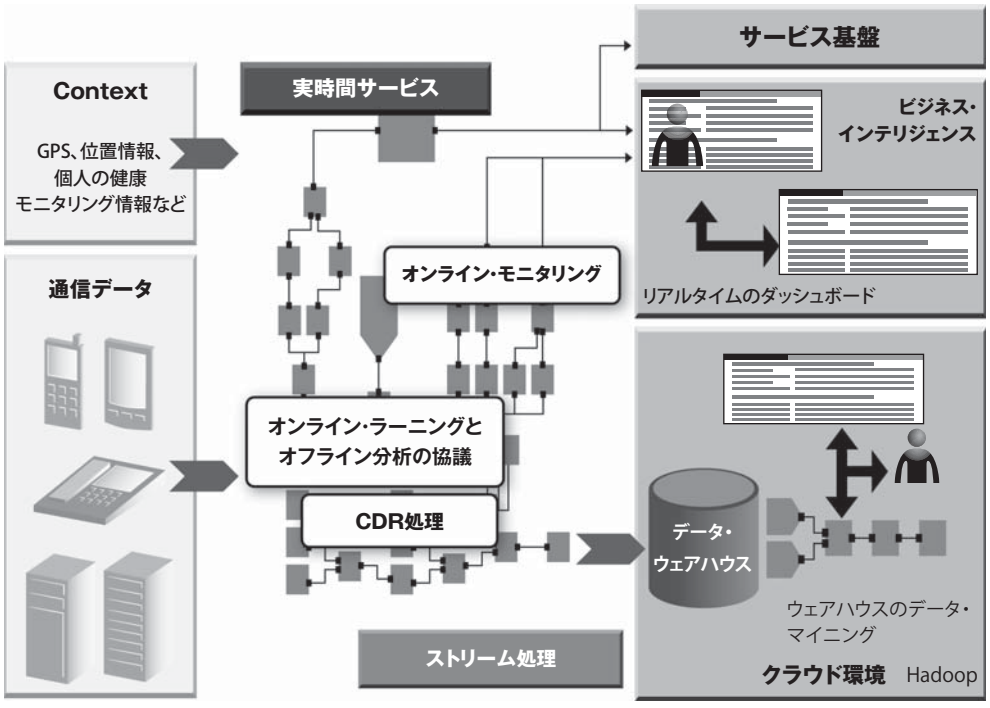

図3. InfoSphere Streamsの通信業界への適用

通信業界においてはほかにも多くの応用例が考えられます。今日、携帯電話には、GPS、電子決済、Web ブラウザーなどの機能が搭載され、国内だけでも 1 億台以上の端末から発信される総データ量は膨大なものとなっています。これらのデータを複合的に用いて、契約者の行動分析や嗜好分析を行い、マーケティング・データとして活用するような、いわゆる「ライフログ」のエリアが盛んになりつつありますが、これも InfoSphere Streams を有効活用できる領域といえるでしょう。また、**図 3** に示すように、携帯電話などの GPS 位置情報を用いた実時間サービスや、オフラインで行ったソーシャル・ネットワーク分析の結果から、よりタイムリーに契約者へキャンペーン案内を送付するなどの応用例も考えられます。

3.2 証券業界の事例

証券業界での実証実験の例として、ここではトロントドミニオン銀行様と行った事例［3］について紹介します。

この実証実験における課題としては、アルゴリズム・トレード・システムにおいて、増え続けるデータ量に対する既存システムの処理能力の限界が挙げられていました。特に既存システムでは、性能向上のための複雑なコンフィグレーションをすべて人手で行わなければならない点や、入力データに対するフィルタリング処理など特定の処理に負荷が偏ってしまい、コンピューティング・リソースを有効に活用することが困難な状況となっていました。

このため、本実証実験では、InfoSphere Streams を用いたストリーム・コンピューティングの実装によって、以下の可能性を明らかにすることが目的となりました。

- 指数的に増加するデータ・レートに対して、低レイテンシーでのリアルタイム処理の実現
- 競争力を維持するために、新しい取引アルゴリズムの開発から展開までを短期間で実現
- 取引システムの高い可用性と耐故障性の実現

実装では、まずシステムの中で真に低レイテンシーでの処理が必要な処理を明確化し（**図 4**）、その部分を SPL で実装しました。さらに、このプログラムを IBM のスーパーコンピューターである Blue Gene® 上で動作させ、強力な計算能力と高速ネットワークを併せて活用することで、低レイテンシー処理を実現することを目指しました。

その結果、最終的には、実際のデータ転送速度の 15 倍もの速度で市況データを受信しながら、同時に 151ms という低レイテンシーで処理を行うことができることを実証しました。

現在では、InfoSphere Streams の高速化に加え InfiniBand や Low Latency Messaging（LLM）などの高速ネットワーク技術の活用により、IBM

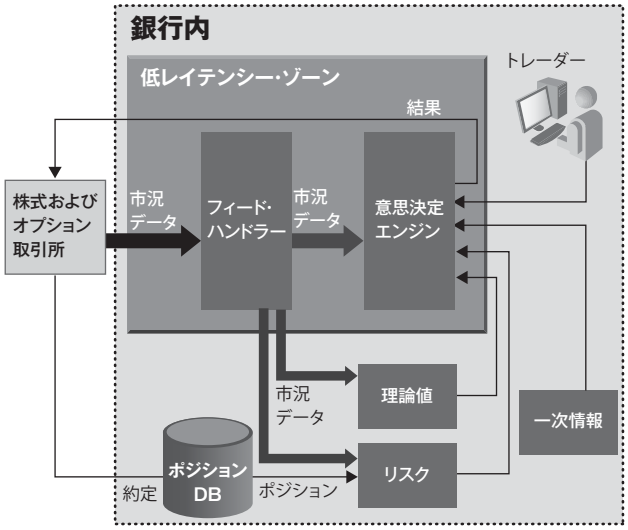

図4. 典型的なアルゴリズム・トレード・システム

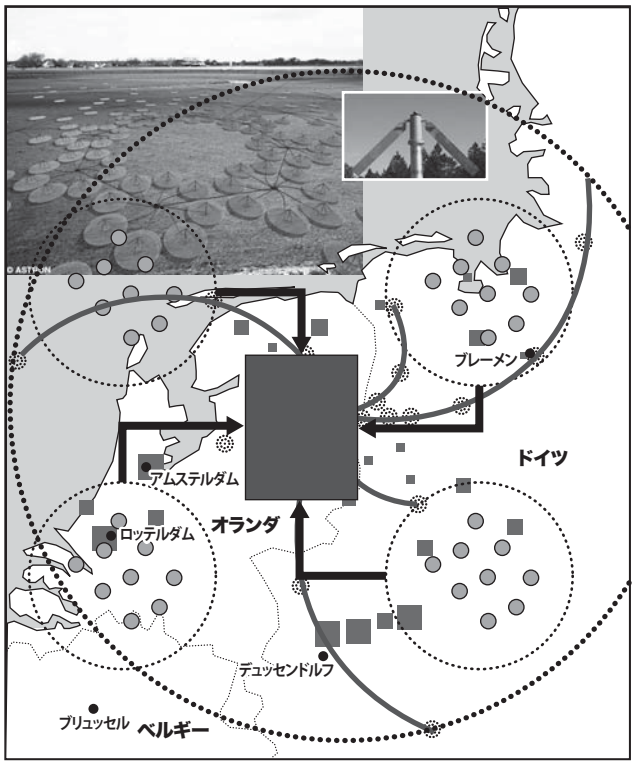

図5. LOFARシステム

BladeCenter® HS22といったインテル®・ブレードのクラスターでも同様な性能が達成されています。

3.3 学術分野の事例

学術分野においては、電波天文学での活用を通じて科学技術計算での有効性を評価する実証実験を行っています。

近年の電波天文学では、従来の単一で巨大なアンテナによる観測から、広域に分散した多数の小さなアンテナを用いて宇宙から飛来する電波を受信し、それらの信号を合成することで、より解像度の高いイメージを生成する方法へと観測手法が進化しつつあります。例えば、LOISプロジェクト［4］は、スカンジナビア半島において、LOFAR（Low Frequency Array）システム（**図5**）という2万5千個のアンテナを直径350kmの領域に分散配置するシステムの実現を目指しています。

このような多数のアンテナを用いたシステムでは、すべてのアンテナから送られてくるデータは非常に膨大で、本実証実験では、毎時22.7 Tバイトものデータを分析することが見込まれています。従来のように一度ストレージに蓄積してから解析をするアプローチでは、それを保存するための膨大な記憶装置が必要となり、加えてデータの保存自体にも時間がかかり、実際に観測してから天文学者がその結果を手に入れるまで長い時間を要してしまいます。

そこでLOISプロジェクトでは、InfoSphere Streamsを用いて、アンテナから送られてくるノイズの多い天体物理データを、適時に解析しマイニングするシステムの実現に向けた検証を行っています。検証の結果、例えば観測データがわずか10分程度の時間で利用可能となれば、研究者はよりインタラクティブな観測が可能となります。また、アンテナから送られてくるストリーム・データに対して、大規模クラスターを活用した解析処理を容易に作成できることが実証できれば、さまざまなデータ解析を用いた実験を効率的に行うことが可能となります。

❹ 今後の展開

ストリーム・コンピューティングとは、ストリーム・データを複合的に分析・判断し、迅速な意思決定を可能にする計算パラダイムです。その適用分野は爆発的に拡大していますが、ストリーム・データが大量であればあるほど、分析・判断が高度であればあるほど、そして意思決定までの時間が短ければ短いほど、チャレンジングな適用分野であるといえます。InfoSphere Streamsは、こうした最もチャレンジングな分野への適用を念頭において研究開発と実証実験を重ねてきており、それらに十分耐え得る開発環境と実行環境とツールキットを装備しています。

プログラミング言語であるSPLは、ストリーム・コンピューティングのために開発された言語であるだけに、ストリーム・コンピューティングを行うプログラムの高効率な開発が可能です。SPLは決して汎用言語ではありませんが、1970年代に登場したSQLがデータベースにおいて果たしたような役割を、ストリーム・コンピューティングにおいて果たすかもしれません。

さらに、既存のファイルやデータベースに格納されているデータをストリーム・データのように読むという視点で考えれば、いわゆるバッチ処理などもInfoSphere Streamsの適用範囲になる可能性すらあります。アプリケーションによっては、SPLによって簡潔に記述できる可能性もあり、SPLで簡潔に記述できれば、InfoSphere Streamsのスケーラブルな実行環境で非常に効率よく走らせることができます。東京基礎研究所では、現在、このような方向で研究活動を行っており、研究段階ではありますが、幾つかのアプリケーションで良好な結果を得ています。伝統的なストリーム・コンピューティングの枠を超えた領域への適用も期待されるゆえんです。

今日のストリーム・データの爆発は、まさに未曾有のことであり、新しい時代の到来を意味しているといっても過言ではありません。まったく新しい発想から、まったく新しい利活用が次々と実践されていくことでしょう。Twitterのつぶやきから景況感をつかもうとする人が現れるかもしれませんし、SNSでの行動履歴とGPS位置情報履歴を組み合わせて顧客選別を試みる人が出てくるかもしれません。InfoSphere Streamsは、こうした新しい時代の基盤プラットフォームとしての役割を果たしていくものと期待しています。

[参考文献]

[1] IBM Press Release, IBM Ushers In Era Of Stream Computing, http://www.ibm.com/press/us/en/pressrelease/27508.wss

[2] IBM Announcement Letter, IBM InfoSphere Streams V1.2.0 supports highly complex heterogeneous data analysis, http://www.ibm.com/common/ssi/cgi-bin/ssialias?subtype=ca&infotype=an&appname=iSource&supplier=897&letternum=ENUS210-037

[3] IBM Press Release, TD Bank Financial Group Partners With IBM on Breakthrough Supercomputing System, http://www.ibm.com/press/us/en/pressrelease/23793.wss

[4] LOFAR Outrigger In Scandinavia (LOIS) project: http://www.lois-space.net/index.html

[5] Square Kilometre Array (SKA) project: http://www.skatelescope.org/

日本アイ・ビー・エム株式会社
東京基礎研究所
インフラストラクチャー・ソフトウェア担当
シニア・テクニカル・スタッフ・メンバー

小野寺 民也 **Tamiya Onodera**

[プロフィール]
1988 年、日本 IBM 入社。以来、同社東京基礎研究所にて、オブジェクト指向言語の設計および実装の研究に従事。理学博士。ACM Senior Member。情報処理学会、日本ソフトウェア科学会、各会員。
http://www.trl.ibm.com/people/onodera/

日本アイ・ビー・エム株式会社
東京基礎研究所
インフラストラクチャー・ソフトウェア
アドバイザリー・リサーチャー

安江 俊明 **Toshiaki Yasue**

[プロフィール]
1995 年、日本 IBM 入社。以来同社東京基礎研究所にて、Java™ Just-in-Time コンパイラーのための最適化技術やミドルウェアの高速化手法の研究に従事。工学博士。ACM。情報処理学会、電子情報通信学会、各会員。
http://www.trl.ibm.com/people/yasue/

日本アイ・ビー・エム株式会社
東京基礎研究所
インフラストラクチャー・ソフトウェア
スタッフ・リサーチャー

鈴村 豊太郎 **Toyotaro Suzumura**

[プロフィール]
2004 年、日本 IBM 東京基礎研究所入社。専門研究分野は、グリッドなどの広域分散並列処理、XML/Web サービスや PHP 言語処理系の高速化、ストリーム・コンピューティング、理学博士。2009 年 4 月より、東京工業大学情報理工学研究科客員准教授と兼務。
http://www.trl.ibm.com/people/suzumura/