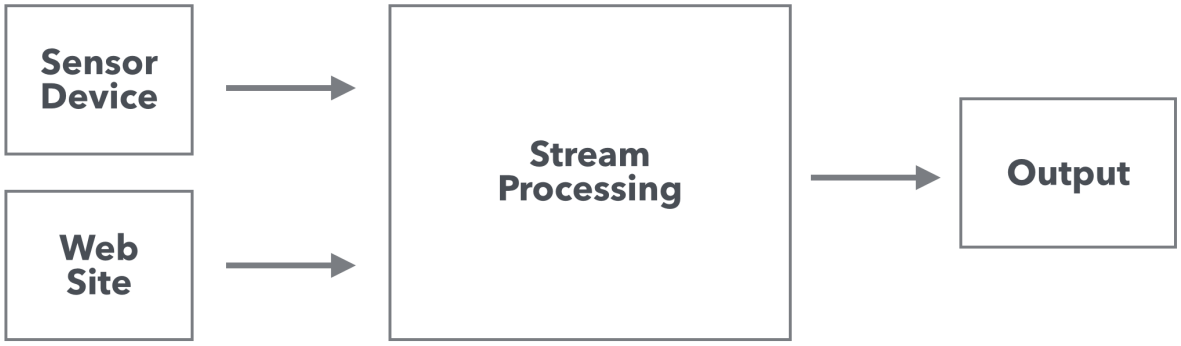
ストリーム処理ってなんだろう

ストリーム処理ってなに？

まずはこれから扱うストリーム処理がどういったものなのかを整理してみましょう。
本誌では以下の定義で話を進めていこうかと思います。

> 刻々と生成されるデータを短い時間幅で取り込んで処理する

例えば、短い間隔でデータを生成するものといえばセンサーなどがイメージしやすいのではないでしょうか。
はたまた、ECサイトを例にとると、そのサイトに訪れた人のログなども短い間隔でその時その時データを生成するものとも考えることができます。
センサーデータの例でいうと、センサーデータが送信させるたびに閾値をチェックしたり、直近の平均値をその都度計算したりといったように、
あとでまとめて処理するのではなく、短い時間幅で処理することを本誌でのストリーム処理とします。



ストリーム処理とよく比較されるのがバッチ処理です。
バッチ処理は発生データをある程度ためたのち、まとめて処理するといった特徴があります。
例えばECサイト上での1日のログを夜間にまとめて処理して、ユーザの傾向を把握するといった具体です。

ストリーム処理もバッチ処理もデータを処理してある結果を生成するという点では同じですが、
ストリーム処理が優れている点はなんでしょうか。

それは、即時に結果をえることができる点です。(なんだかそのままになってしまいました笑
例えば、クレジッドカードの利用履歴を分析する例を考えてみましょう。
クレジッドカード会社ではカードの利用履歴から突然これまでより大金を利用したなどを検知し、
なりすましなどによるクレジッドカード詐欺の可能性を検出しています。

[alt] | *target*

こういった処理はこれまでバッチ処理で行われていたのですが、ストリーム処理でこの処理を実現すること
より早期にクレジットカード詐欺の可能性を検出することが可能となります。

どういうところがポイントになるか

ストリーム処理を実現するOSSはたくさんあるのですが、できることや処理の方式が違うものなどそれぞれに特徴があります。　それぞれのプロダクトにはざくっと以下の2つのポイントで特徴があらわれてきます。

- 少しずつまとめて処理するのか/一つずつ処理するのか
- 発生データをどこまで厳密に処理できるのか
- どこまで厳密にデータを届けられるのか

ひとつずつ見ていきましょう。

少しずつまとめて処理するのか/ひとつずつ処理するのか

基本的なストリーム処理の動作は、処理対象のデータをよりリアルタイムに処理していくことになるのですが、 その処理するデータの扱い方については大きく 2 つのやり方があります。

処理のやり方	説明	主要プロダクト
micro-batch方式 (少しずつまとめて処理する)	hoge	Spark Streaming,
tuple-at-a-time方式 (ひとつずつ処理する)	fuga	Storm,

micro-
batch方式では、きたデータをその都度処理するのでなく、短いインターバルでバッチ処理を実施します。
<図を入れる>

Spark　Streamingではこのmicro-batchが利用されており、RDDと呼ばれるほげほげに対して短いインターバルで高速にバッチ処理を実施することでストリーム処理を実現しています。
<図を入れる>

かたやtuple-at-atimeはきたデータをひとつずつ処理していきます。
そのため、よりストリーム処理のイメージに近いのはこちらの方式かもしれません。
<図を入れる>

Stormはtuple-at-timeでリアルタイムにきたデータを処理していきます。
<図を入れる>

発生データをどこまで厳密に処理できるのか

さきほどはどのようにきたデータを処理するのかといった観点で説明しました。
次は、きたデータをどのように云々

種類	説明	主要プロダクト
イベント発生時刻	hoghoge	hogehoge
データ到着時刻	hoghoge	hoghoge

micro-

batch方式では、きたデータをその都度処理するのでなく、短いインターバルでバッチ処理を実施します
。
<図を入れる>

Spark Streamingではこのmicro-batchが利用されており、RDDと呼ばれるほげほげに対して短いインターバルで高速にバッチ処理を実施することでストリーム処理を実現しています。
<図を入れる>

かたやtuple-at-atimeはきたデータをひとつずつ処理していきます。
そのため、よりストリーム処理のイメージに近いのはこちらの方式かもしれません。
<図を入れる>

Stormはtuple-at-timeでリアルタイムにきたデータを処理していきます。
<図を入れる>

データの偏りがあってSKEWの考え方があって、ほげほげほげほげほgへおgへおhごえhごえhげおhごえhごえhごえhごえ
ほげほげほげほげほgへおげほげほげほげほげほげほげgほえほえhごえhごえhごえhごhごえほgへおgへおgへおhごhげ
ほげほげほげほげほgへおgへおgへおhげおhごえhげおhごえhごえhごえhごえhごえhごえhごえほげほgへおgへおh/ごえh

メッセージ信頼性

メッセージの信頼性とは云々

種類	説明	主要プロダクト
at most once	hoghoge	hogehoge
at least once	hoghoge	hoghoge
exactly once	hogehoge	hogehgoe

イベント発生時刻はウンウン
ほげ
ほげ
ほげ

<図をいれる>

データ到着時刻は云々
ほげ
ほげ
ほげ

<図をいれる>

一般的にskew(データ偏り)が云々
ほげ

ほげ

ほげ

<図を入れる>