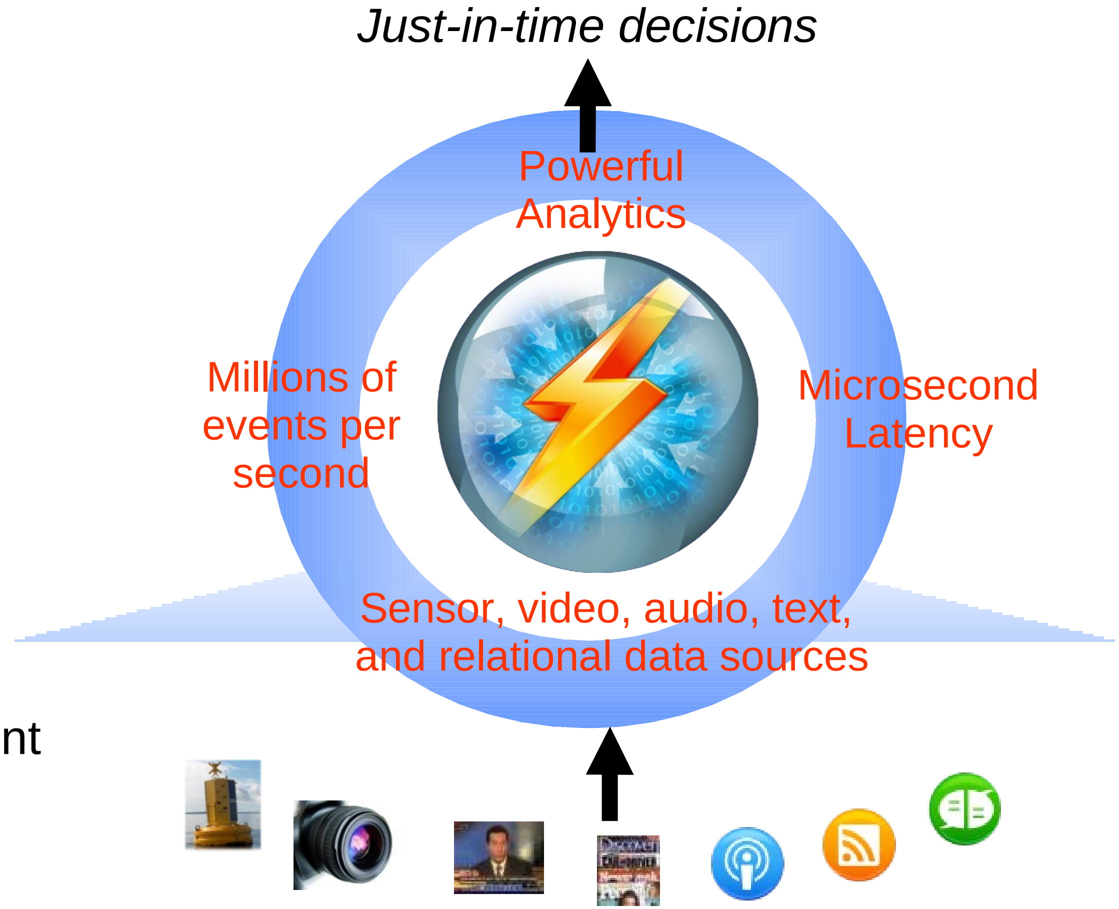
IBM InfoSphere Streams
Technical Overview

Jérôme Chailloux
Europe IOT - Sr. Technical Field Specialist - Big Data, Linux Advocate
jerome.chailloux@fr.ibm.com

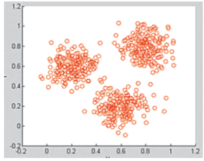
IBM InfoSphere Streams v3.0

A platform for real-time analytics on BIG data

- **Volume**
  - Terabytes per second
  - Petabytes per day
- **Variety**
  - All kinds of data
  - All kinds of analytics
- **Velocity**
  - Insights in microseconds
- **Agility**
  - Dynamically responsive
  - Rapid application development

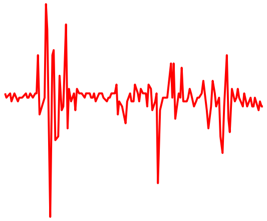

Just-in-time decisions
Powerful Analytics
Millions of events per second
Microsecond Latency
Sensor, video, audio, text, and relational data sources

Streams Analyzes All Kinds of Data



Mining in Microseconds

(included with Streams)



Acoustic

(IBM Research)
(Open Source)

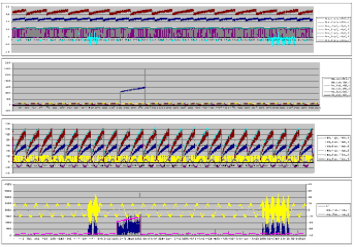
***New

Text
(listen, verb),

(radio, noun)

Simple & Advanced Text

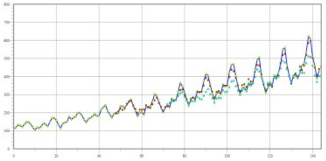
(included with Streams)



Advanced Mathematical Models

(IBM Research)

***New



Predictive

(included with Streams)

$$\sum_{population} R(s_t, a_t)$$

Statistics

(included with Streams)

***New

Geospatial

(included with Streams)

OpenCV

Image & Video

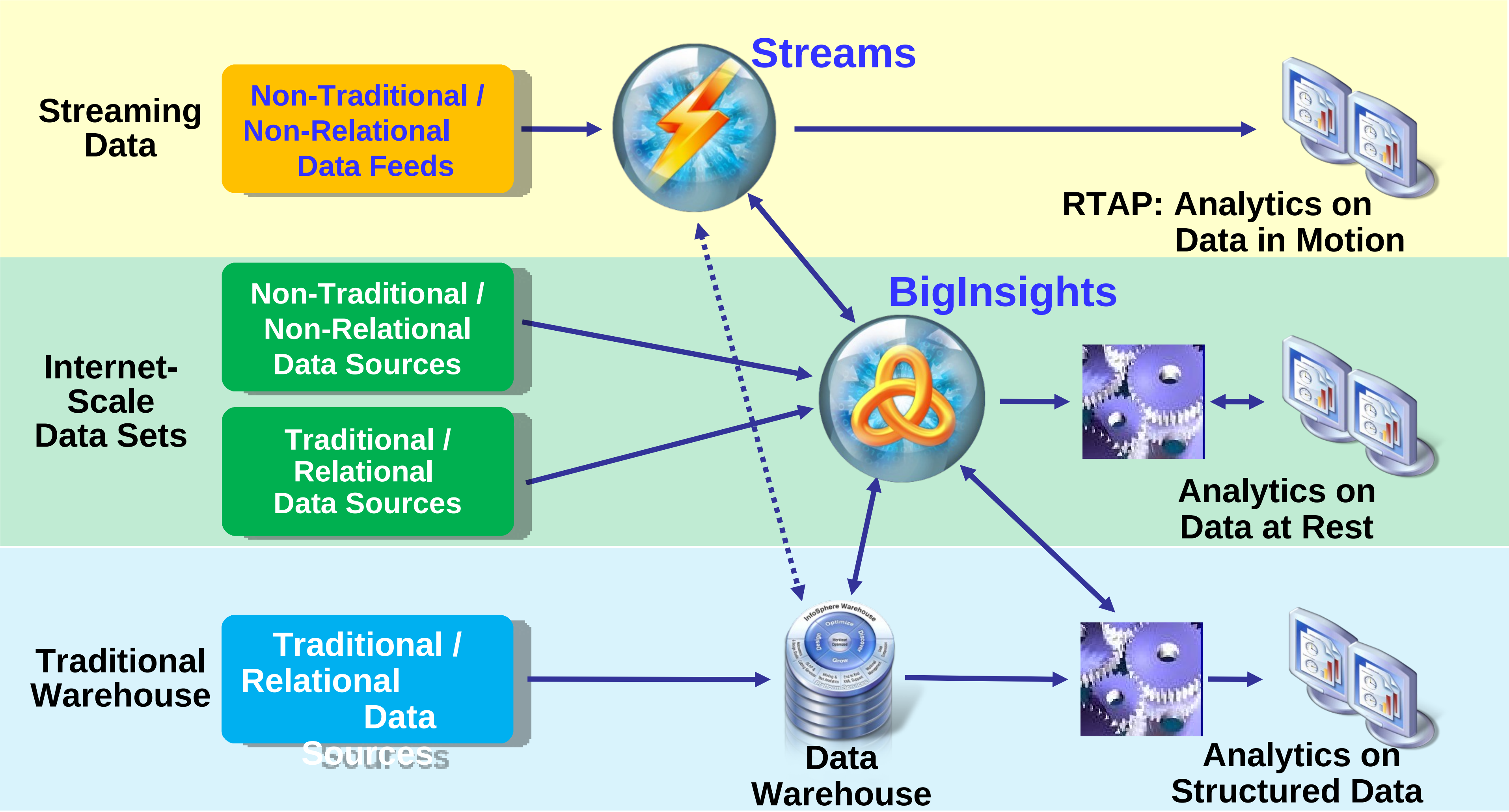
(
S

Categories of Problems Solved by Streams

- **Applications that require on-the-fly processing, filtering and analysis of streaming data**
  - Sensors: environmental, industrial, surveillance video, GPS, …
  - "Data exhaust": network/system/web server/app server log files
  - High-rate transaction data: financial transactions, call detail records

- **Criteria: two or more of the following**
  - Messages are processed in isolation or in limited data windows
  - Sources include non-traditional data (spatial, imagery, text, …)
  - Sources vary in connection methods, data rates, and processing requirements, presenting integration challenges
  - Data rates/volumes require the resources of multiple processing nodes
  - Analysis and response are needed with sub-millisecond latency
  - Data rates and volumes are too great for store-and-mine approaches

The Big Data Ecosystem: Interoperability is Key



Streaming Data

Non-Traditional / Non-Relational Data Feeds

Streams

RTAP: Analytics on Data in Motion

Internet-Scale Data Sets

Non-Traditional / Non-Relational Data Sources

Traditional / Relational Data Sources

BigInsights

Analytics on Data at Rest

Traditional Warehouse

Traditional / Relational Data Sources

Data Warehouse

Analytics on Structured Data

Streaming Analytics in Action

Natural Systems
- Wildfire management
- Water management

Transportation
- Intelligent traffic management

Manufacturing
- Process control for microchip fabrication

Health & Life Sciences
- Neonatal ICU monitoring
- Epidemic early warning system
- Remote healthcare monitoring

Telephony
- CDR processing
- Social analysis
- Churn prediction
- Geomapping

Stock Market
- Impact of weather on securities prices
- Analyze market data at ultra-low latencies

Law Enforcement, Defense & Cyber Security
- Real-time multimodal surveillance
- Situational awareness
- Cyber security detection

Fraud Prevention
- Detecting multi-party fraud
- Real time fraud prevention

e-Science
- Space weather prediction
- Detection of transient events
- Synchrotron atomic research

Other
- Smart Grid
- Text analysis
- Who’s talking to whom?
- ERP for commodities
- FPGA acceleration

Use Case: Law Enforcement and Security

- **Video surveillance, wire taps, communications, call records, etc.**
- **Millions of messages per second with low density of critical data**
- **Identify patterns and relationships among vast information sources**

"The **US Government** has been working with IBM Research since **2003** on a radical new approach to data analysis that enables **high speed**, **scalable** and **complex analytics** of **heterogeneous data streams** in motion. The project has been so successful that US Government will deploy additional installations to enable other agencies to achieve greater success in various future projects"
– US Government

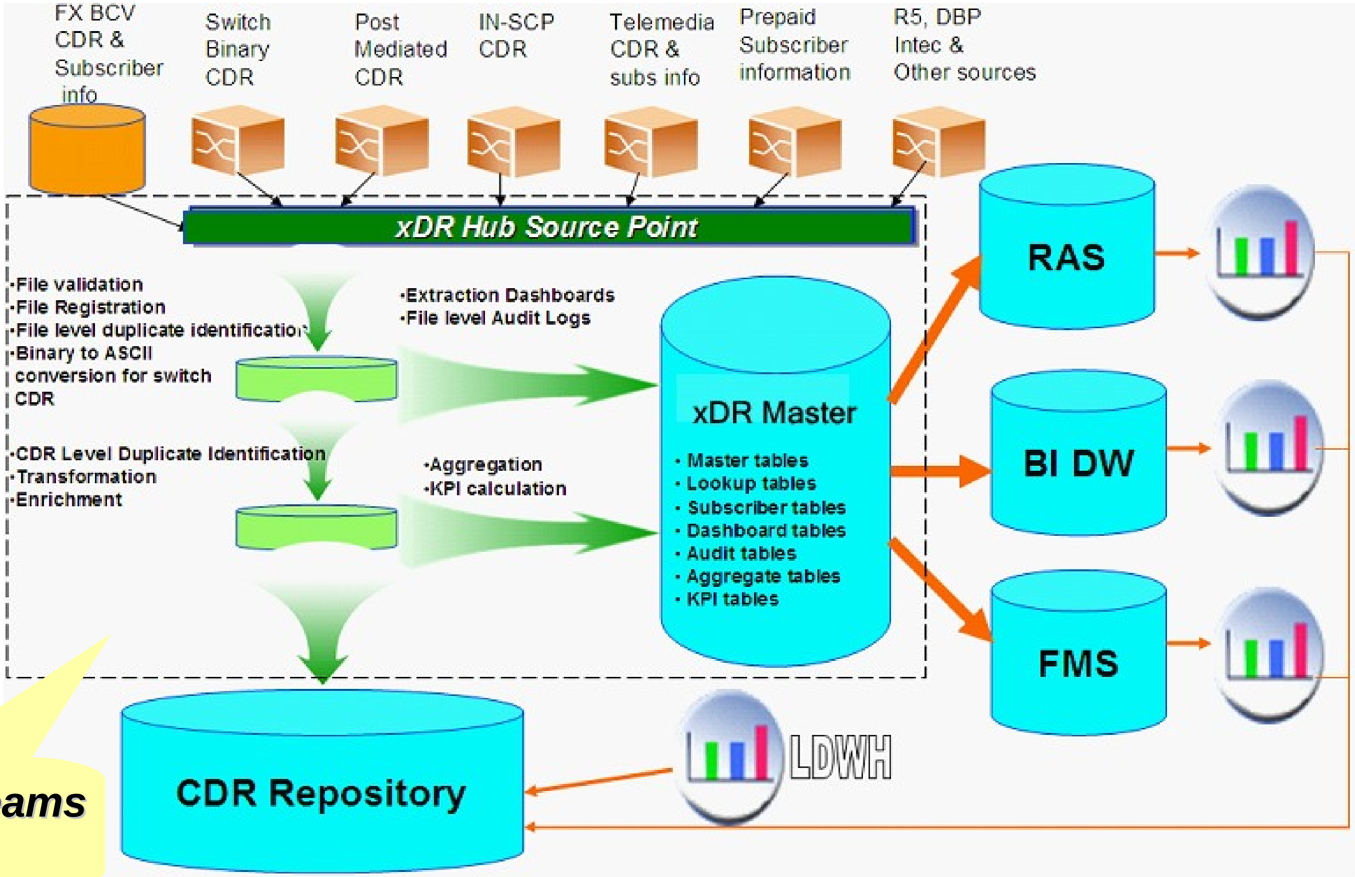
Predictive Analytics in a Neonatal ICU

- **Real-time analytics and correlations on physiological data streams**
  - Blood pressure, Temperature, EKG, Blood oxygen saturation etc.,
- **Early detection of the onset of potentially life-threatening conditions**
  - Up to 24 hours earlier than current medical practices
  - Early intervention leads to lower patient morbidity and better long term outcomes
- **Technology also enables physicians to verify new clinical hypotheses**

UOIT
CHALLENGE INNOVATE CONNECT

Smarter Faster Cheaper CDR Processing

6 Billion CDRs per day, dedups over 15 days, processing latency from 12 hours to a few seconds
6 machines (using ½ processor capacity)

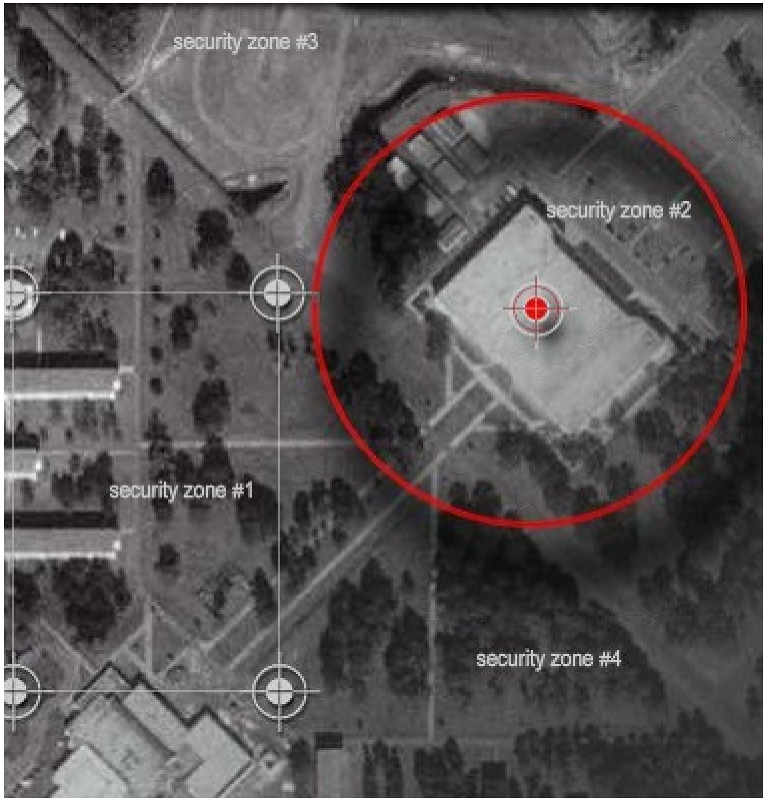


InfoSphere Streams xDR Hub

Key Requirements: Price/Performance and Scaling

Surveilance and Physical Security: TerraEchos (Business Partner)

- **Use scenario**
  - **State-of-the-art covert surveillance system based on Streams platform**
  - Acoustic signals from buried fiber optic cables are monitored, analyzed and reported in real time for necessary action
  - Currently designed to scale up to 1600 streams of raw binary data
- **Requirement**
  - Real-time processing of multi-modal signals (acoustics. video, etc)
  - Easy to expand, dynamic
  - 3.5M data elements per second
- **Winner 2010 IBM CTO Innovation Award**



TerraEchos

Streams for Real-Time Geomapping

Multiple GPS Data Sources

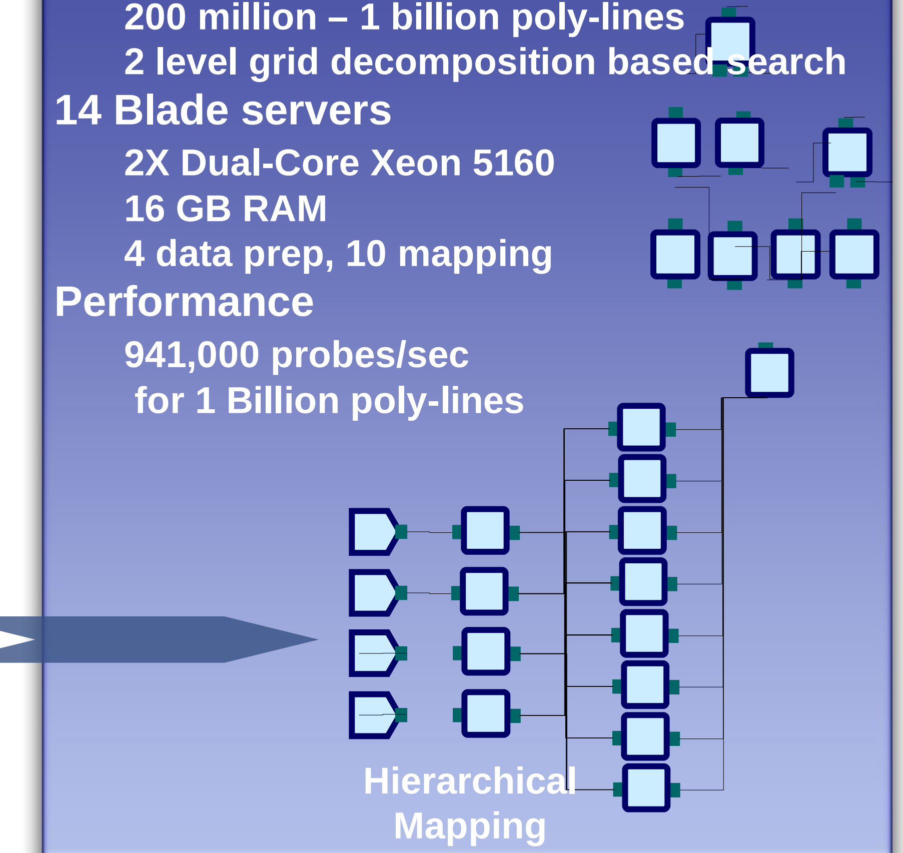
- 350-400K probe points / second per source
- Map probe point to nearest poly-line (Map)
- 200 million – 1 billion poly-lines
- 2 level grid decomposition based search

14 Blade servers

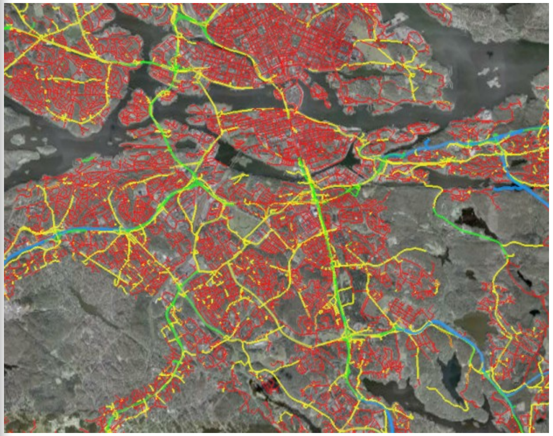
- 2X Dual-Core Xeon 5160
- 16 GB RAM
- 4 data prep, 10 mapping

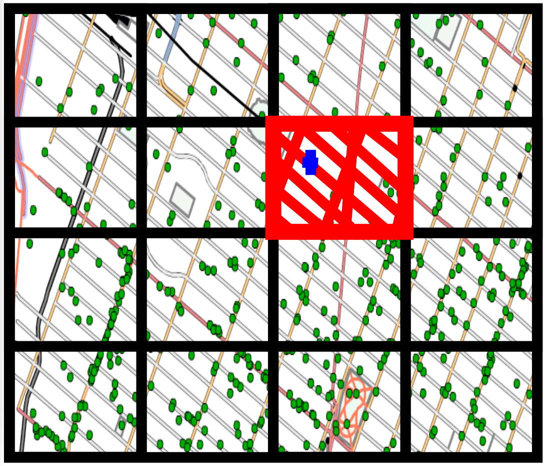
Performance

- 941,000 probes/sec
- for 1 Billion poly-lines

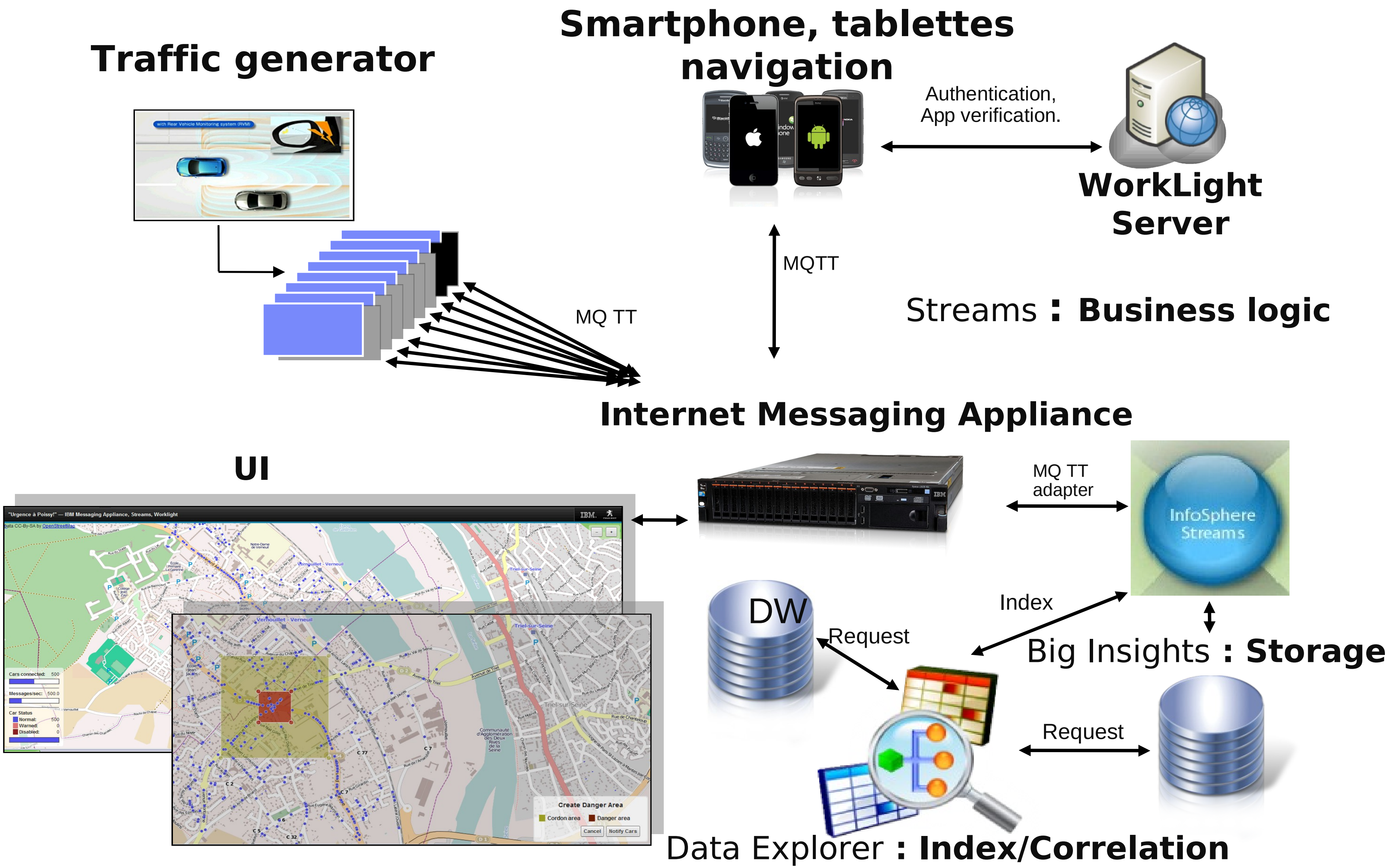


Real-time location profile





Connected Cars



Traffic generator

Smartphone, tablettes navigation

Authentication, App verification.

WorkLight Server

MQTT

MQ TT

Streams **: Business logic**

Internet Messaging Appliance

UI

MQ TT adapter

Index

Request

Big Insights **: Storage**

Request

Data Explorer **: Index/Correlation**

"Urgence à Poissy!" — IBM Messaging Appliance, Streams, Worklight

Data CC-By-SA by OpenStreetMap

Cars connected: 500

Messages/sec: 500.0

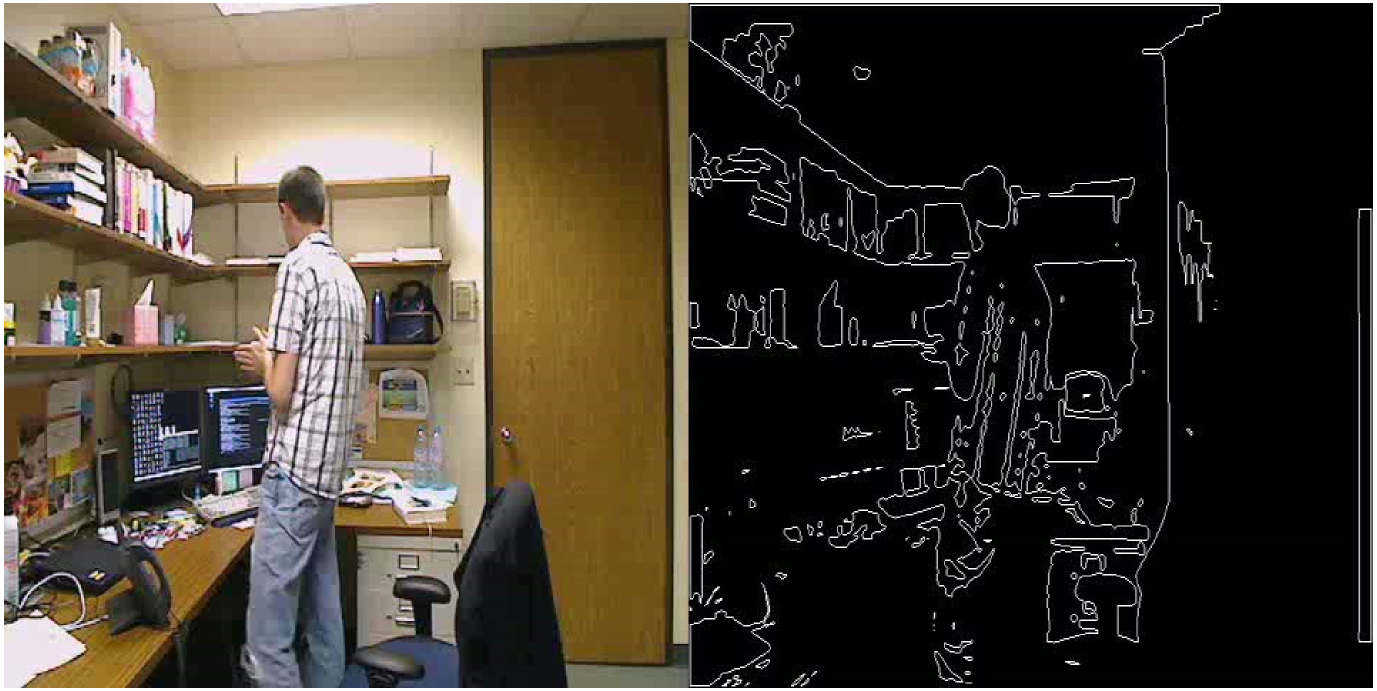
Car Status
- 🟦 Normal: 500
- 🟥 Warned: 0
- 🟥 Disabled: 0

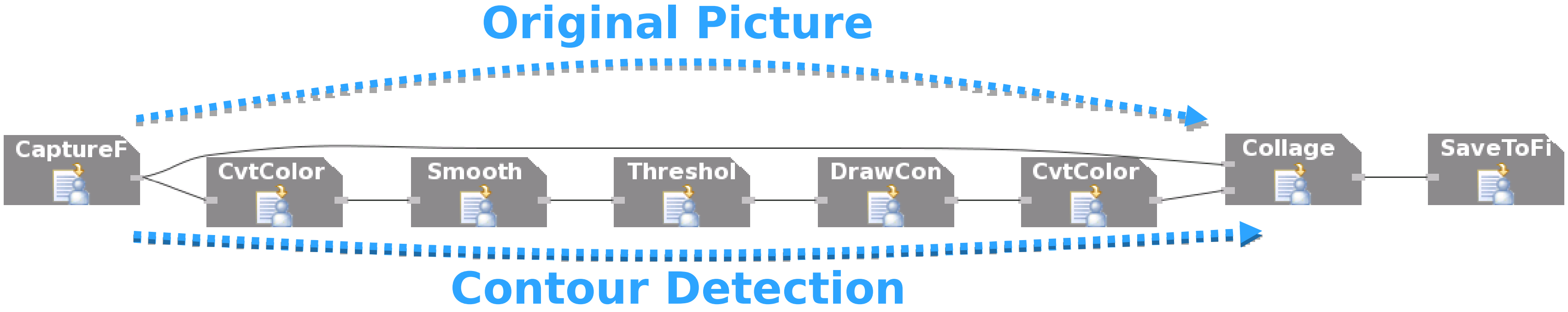
Create Danger Area

🟨 Cordon area 🟥 Danger area

Cancel | Notify Cars

Use Cases: Video Processing (Contour Detection)





Original Picture

CaptureF

CvtColor

Smooth

Threshol

DrawCon

CvtColor

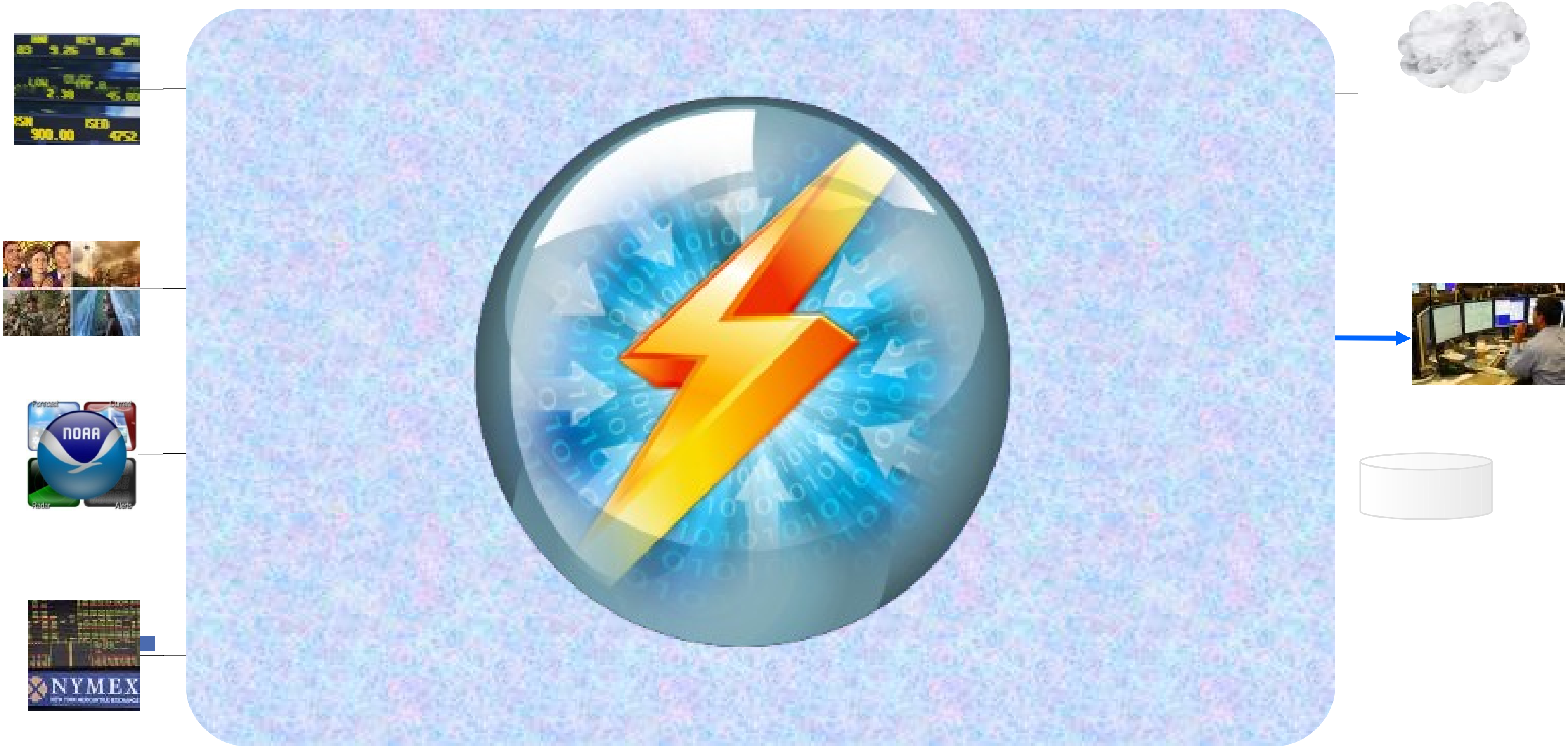
Collage

SaveToFi

Contour Detection

How Streams Works

→ Continuous ingestion
→ Continuous analysis



How Streams Works

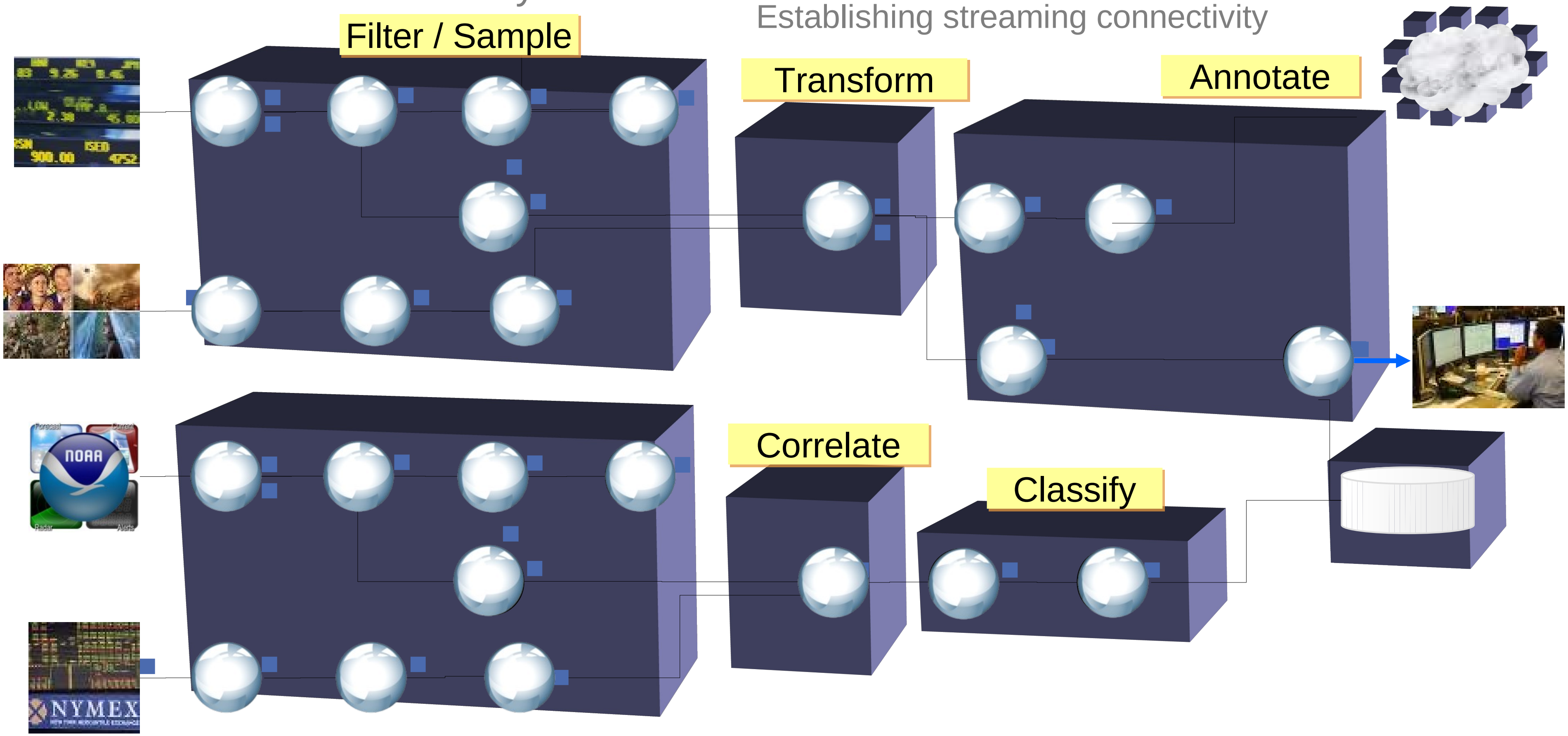
→ Continuous ingestion

→ Continuous analysis

Infrastructure provides services for

Scheduling analytics across hardware hosts,

Establishing streaming connectivity



Achieve scale:

By partitioning applications into software components

By distributing across stream-connected hardware hosts

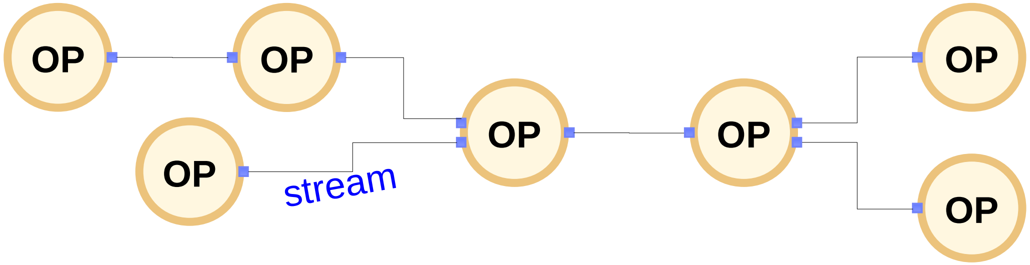
Where appropriate:

Elements can be *fused* together

for lower communication latency

Scalable Stream Processing

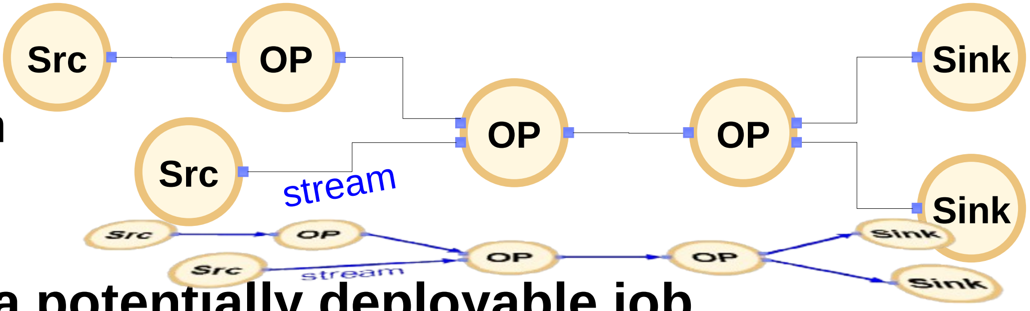
- **Streams programming model: construct a graph**



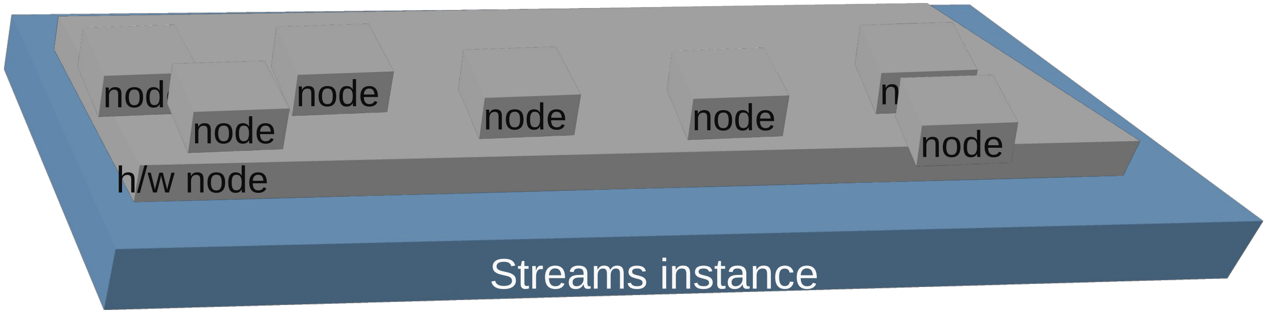
  - Mathematical concept
    - not a line -, bar -, or pie chart!
    - Also called a network
    - Familiar: for example, a tree structure is a graph
  - Consisting of **operators** and the **streams** that connect them
    - The vertices (or nodes) and edges of the mathematical graph
    - A directed graph: the edges have a direction (arrows)
- **Streams runtime model: distributed processes**
  - Single or multiple operators form a Processing Element (PE)
  - Compiler and runtime services make it easy to deploy PEs
    - On one machine
    - Across multiple hosts in a cluster when scaled-up processing is required
  - All links and data transport are handled by runtime services
    - Automatically
    - With manual placement directives where required

From Operators to Running Jobs

- **Streams application graph:**
  - A directed, possibly cyclic, graph
  - A collection of operators
  - Connected by streams

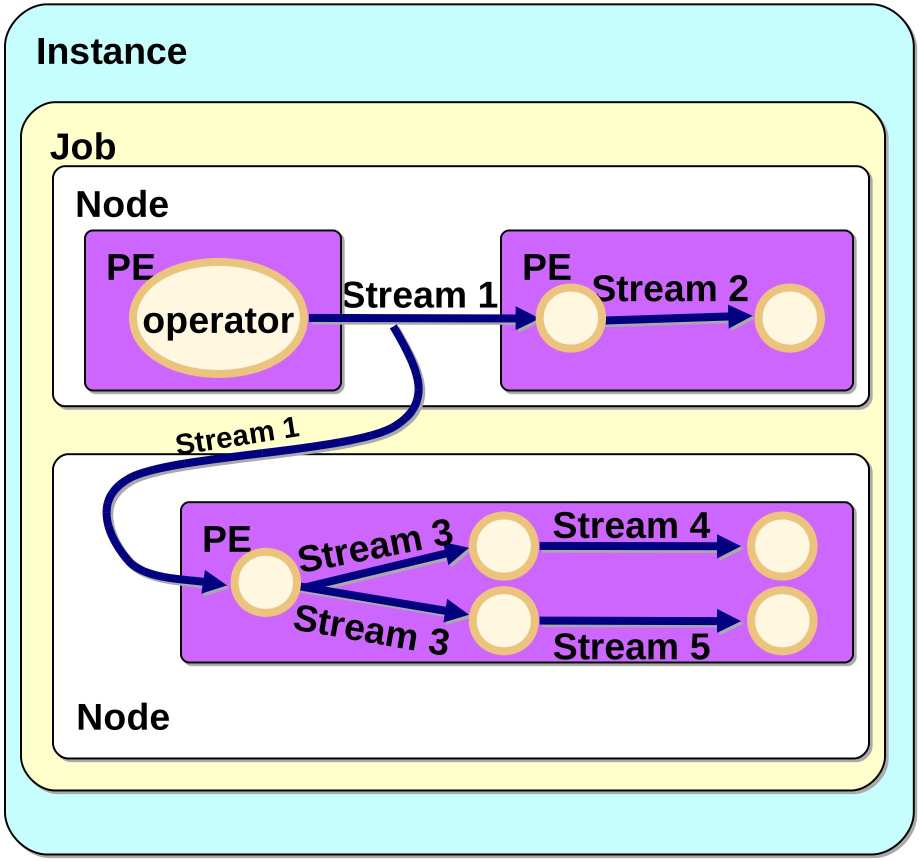


- **Each complete application is a potentially deployable job**
- **Jobs are deployed to a Streams runtime environment, known as a Streams Instance (or simply, an instance)**
- **An instance can include a single processing node (hardware)**
- **Or multiple processing nodes**

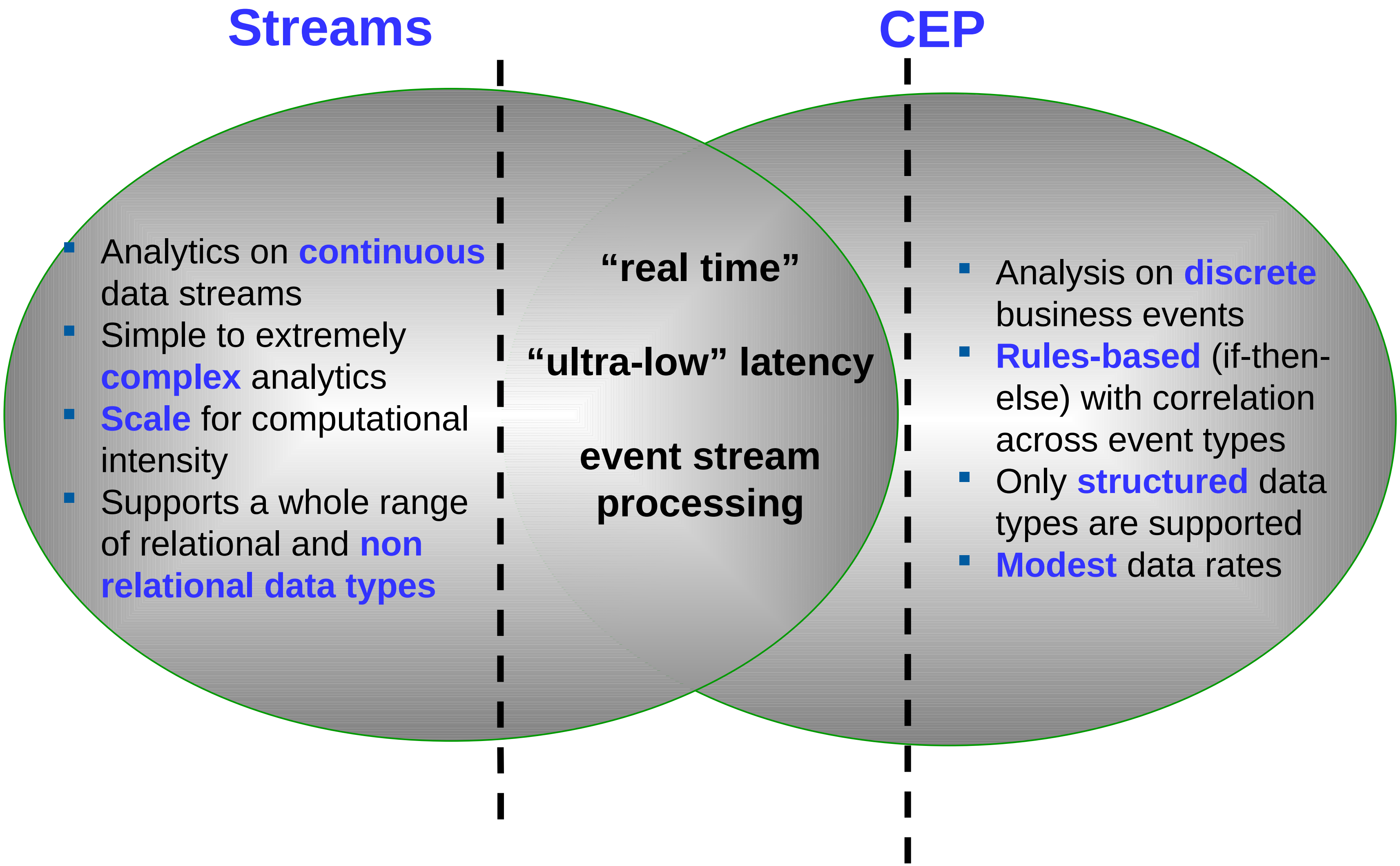


InfoSphere Streams Objects: Runtime View

- **Instance**
  - Runtime instantiation of InfoSphere Streams executing across one or more hosts
  - Collection of components and services
- **Processing Element (PE)**
  - Fundamental execution unit that is run by the Streams instance
  - Can encapsulate a single operator or many “fused” operators
- **Job**
  - A deployed Streams application executing in an instance
  - Consists of one or more PEs



Competition: Complex-Event Processing (CEP)



Streams

- Analytics on **continuous** data streams
- Simple to extremely **complex** analytics
- **Scale** for computational intensity
- Supports a whole range of relational and **non relational data types**

“real time”

“ultra-low” latency

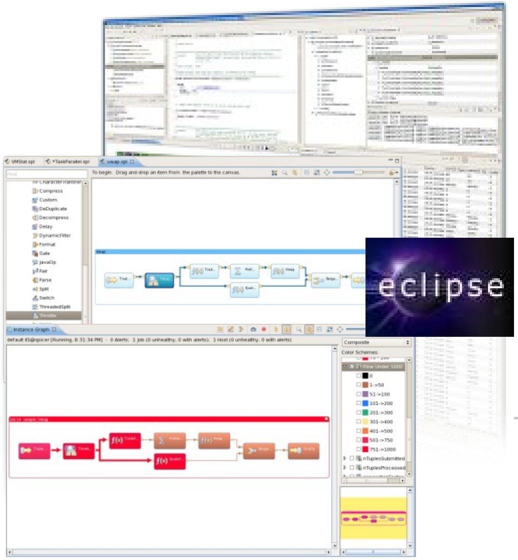
event stream processing

CEP

- Analysis on **discrete** business events
- **Rules-based** (if-then-else) with correlation across event types
- Only **structured** data types are supported
- **Modest** data rates

IBM InfoSphere Streams 3.0

Comprehensive tooling



- Eclipse IDE
- Web console
- **Drag & Drop editor**
- Instance graph
- **Streams visualization**
- Streams debugger

NEW

Scale-out architecture

- Clustered runtime for near-limitless capacity
- RHEL v5.3 and above
- CentOS v6.0 and above
- X86 & Power multicore hardware
- InfiniBand support
- Ethernet support

Sophisticated analytics with **toolkits & accelerators**

Front Office 3.0

- **Big Data, CEP**, Database, **Data Explorer (Big Data), DataStage**, Finance, **Geospatial**, Internet, **Messaging**, Mining, **SPSS**, Standard, Text, **TimeSeries** toolkits

NEW

- **Telco & Social Media accelerators**

What is Streams Processing Language?

- **Designed for stream computing**
  - Define a streaming-data flow graph
  - Rich set of data types to define tuple attributes
- **Declarative**
  - Operator invocations name the input and output streams
  - Referring to streams by name is enough to connect the graph
- **Procedural support**
  - Full-featured imperative language
  - Custom logic in operator invocations
  - Expressions in attribute assignments and parameter definitions
- **Extensible**
  - User-defined data types
  - Custom functions written in SPL or a native language (C++ or Java)
  - Custom operators written in SPL
  - User-defined operators written in a native language (C++ or Java)

Streams Standard Toolkit: Relational Operators

- **Relational operators**
  - Functor Perform tuple-level manipulations
  - Filter Remove some tuples from a stream
  - Aggregate Group and summarize incoming tuples
  - Sort Impose an order on incoming tuples in a stream
  - Join Correlate two streams
  - Punctor Insert window punctuation markers into a stream

Streams Standard Toolkit: Adapter Operators

- **Adapter operators**
  - FileSource Read data from files in formats such as csv, line, or binary
  - FileSink Write data to files in formats such as csv, line, or binary
  - DirectoryScan Detect files to be read as they appear in a given directory
  - TCPSource Read data from TCP sockets in various formats
  - TCPSink Write data to TCP sockets in various formats
  - UDPSource Read data from UDP sockets in various formats
  - UDPSink Write data to UDP sockets in various formats
  - Export Make a stream available to other jobs in the same instance
  - Import Connect to streams exported by other jobs
  - MetricsSink Create displayable metrics from numeric expressions

Streams Standard Toolkit: Utility Operators

- **Workload generation and custom logic**
  - Beacon Emit generated values; timing and number configurable
  - Custom Apply arbitrary SPL logic to produce tuples
- **Coordination and timing**
  - Throttle Make a stream flow at a specified rate
  - DelayTime-shift an entire stream relative to other streams
  - Gate Wait for acknowledgement from downstream operator
  - Switch Block or allow the flow of tuples based on control input

Streams Standard Toolkit: Utility Operators (cont’d)

- **Parallel flows**
  - Barrier Synchronize tuples from sequence-correlated streams
  - Pair Group tuples from multiple streams of same type
  - Split Forward tuples to output streams based on a predicate
  - ThreadedSplit Distribute tuples over output streams by availability
  - Union Construct an output tuple from each input tuple
  - DeDuplicate Suppress duplicate tuples seen within a given time period
- **Miscellaneous**
  - DynamicFilter Filter tuples based on criteria that can change while it runs
  - JavaOp General-purpose operator for encapsulating Java code
  - **Parse** Parse blob data for use with user-defined adapters
  - **Format** Format blob data for use with user-defined adapters
  - **Compress** Compress blob data
  - **Decompress** Decompress blob data
  - **CharacterTransform** Convert blob data from one encoding to another

XML Support: Built Into SPL

- **XML type**
  - Validated for syntax or schema
- **XMLParse operator**
  - With XPath expressions and functions to parse and manipulate XML data
  - Convert XML to tuples
- **XQuery functions**
  - Use XML data on the fly
- **Adapters support XML format**
  - Standard Toolkit
  - Database Toolkit
    - Supports DB2 pureXML

```
<catalog>
  <book price="30.99">
    <title>This is a boring title</title>
    <author>John Smith</author>
    <author>Howard Hughes</author>
    <reference quality="-1">
      <book>The first reference</book>
    </reference>
    <reference quality="100">
      <book>Another Book</book>
    </reference>
  </book>
</catalog>
```

Example: Extract information about books from an XML catalog

```
stream<BookInfo> X = XMLParse(XML) {
  param trigger : "/catalog/book" ;
        parsing : permissive;     // log and ignore
errors
  output X : title      = XPath("title/text()"),
             authors    = XPathList("author/text()"),
             price      = (decimal32) XPath ("@price"),
             references = XPathList("reference",
                   {quality = (int32) XPath("@quality"),
                    book    = XPath("book/text()") });
}
```

Streams Extensibility: Toolkits

- **Like packages, plugins, addons, extenders, etc.**
  - Reusable sets of **operators**, **types**, and **functions**
  - What can be included in a toolkit?
    - Primitive and composite operators
    - User-defined types
    - Native and SPL functions
    - Sample applications, utilities
    - Tools, documentation, data, etc.
  - Versioning is supported
  - Define dependencies on other versioned assets (toolkits, Streams itself)
- **Base for creating cross-domain and domain-specific applications**

- **Developed by IBM, partners, end-customers**
  - Complete APIs and tools available
  - Same power as the “built-in” Standard Toolkit

Integration: the Internet and Database Toolkits

- **Integration with traditional sources and consumers**

- **Internet Toolkit**
  - InetSource periodically retrieve data from HTTP, FTP, RSS, and files
- **Database Toolkit**
  - ODBC databases: DB2, Informix, Oracle, solidDb, MySQL, SQLServer, Netezza
  - ODBCAppend Insert rows into an SQL database table
  - ODBCEnrich Extend streaming data based on lookups in database tables
  - ODBCRun Perform SQL queries with parameters from input tuples
  - ODBCSource Read data from a SQL database
  - SolidDBEnrich Perform table lookups in an in-memory database
  - DB2SplitDB Split a stream by DB2 partition key
  - DB2PartitionedAppend Write data to table in specified DB2 partition
  - NetezzaLoad Perform high-speed loads into a Netezza database
  - NetezzaPrepareLoad Convert tuple to delimited string for Netezza loads

Integration: the Big Data Toolkit

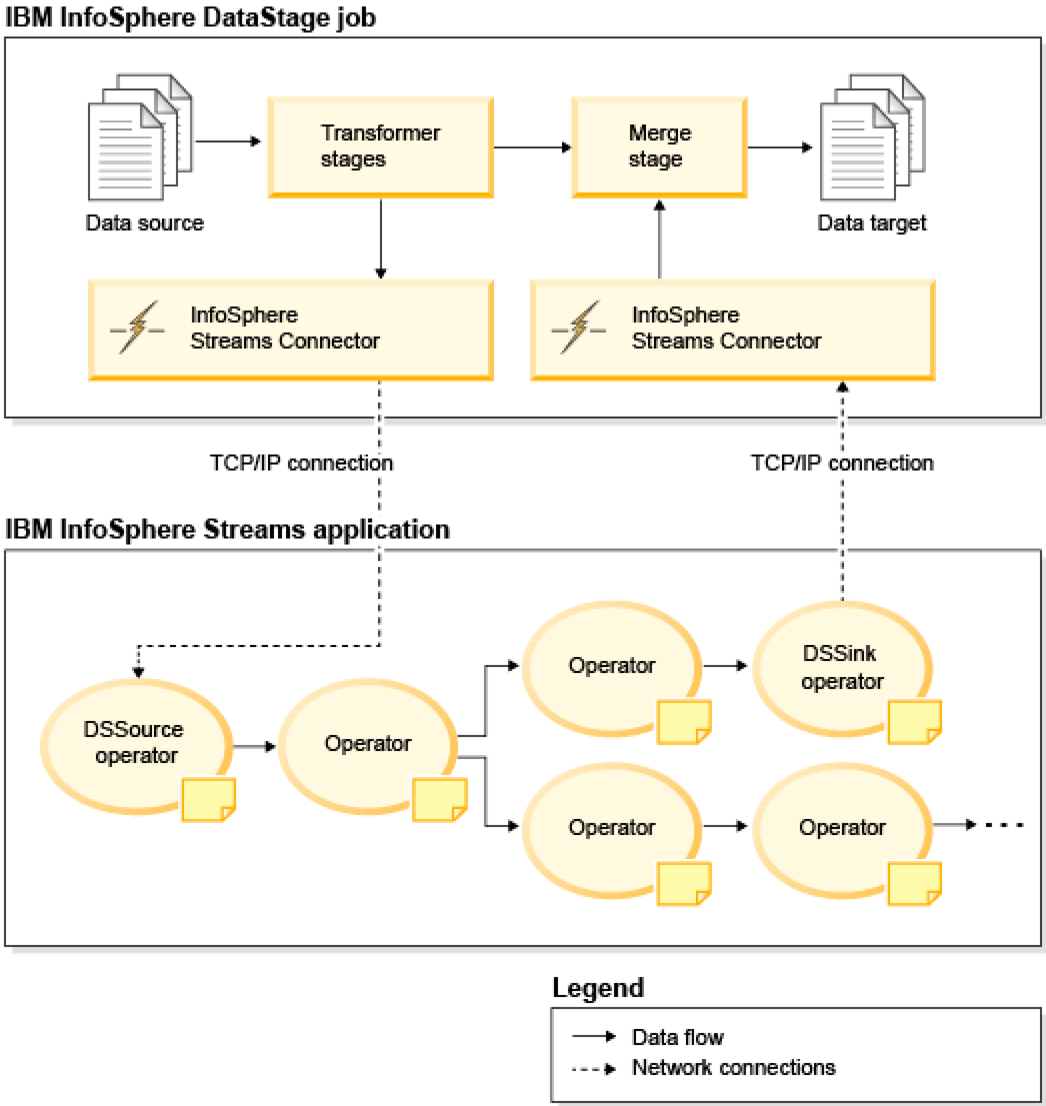
- **Integration with IBM's Big Data Platform**

- **Data Explorer**
  - DataExplorerPush Insert records into a Data Explorer index

- **BigInsights: Hadoop Distributed File System**
  - HDFSDirectoryScan Like DirectoryScan, only for HDFS
  - HDFSFileSource Like FileSource, only for HDFS
  - HDFSFileSink Like FileSink, only for HDFS
  - HDFSSplit Write batches of data in parallel to HDFS

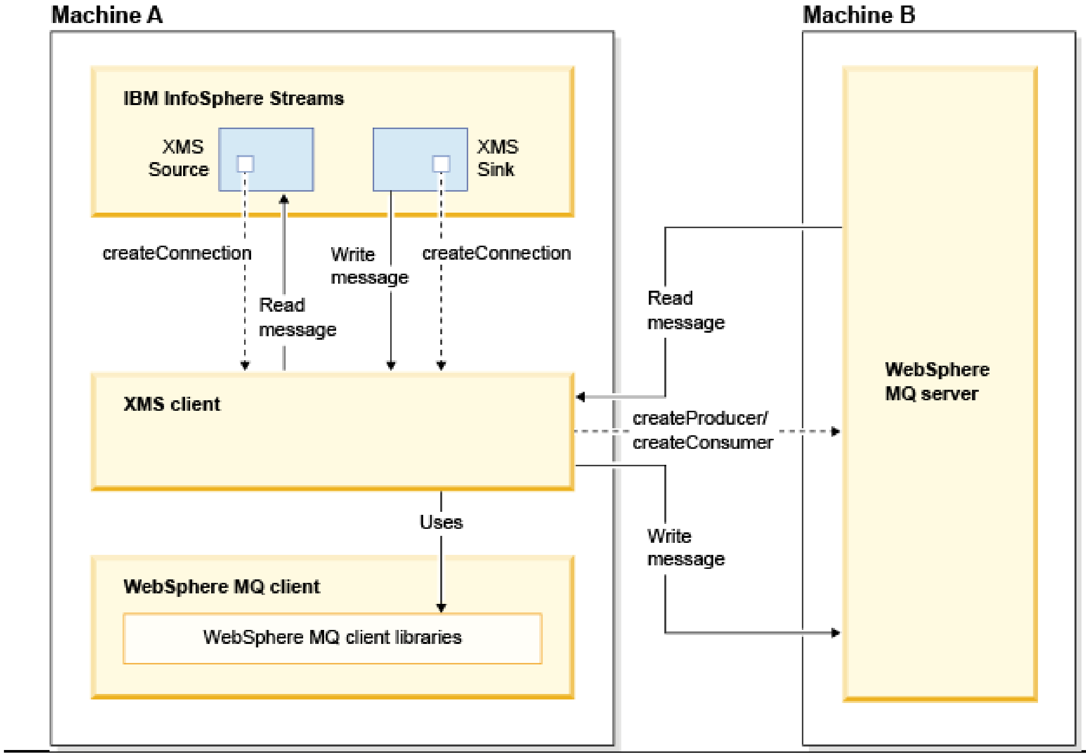
Integration: the DataStage Integration Toolkit

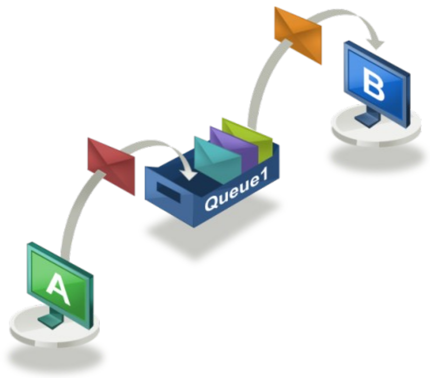
- **Streams real-time analytics and DataStage information integration**
  - Perform deeper analysis on the data as part of the information integration flow
  - Get more timely results and offload some analytics load from the warehouse
- **Operators and tooling**
  - Adapters to exchange data between Streams and DataStage
    - DSSource
    - DSSink
  - Tooling to generate integration code
  - DataStage provides Streams connectors



Integration: the Messaging Toolkit

- **Integrate with IBM WebSphere MQ**
  - – Create a stream from a subscription to an MQ topic or queue
  - – Publish a stream to an MQ series topic or queue
- **Operators**
  - – XMSSource Read data from an MQ queue or topic
  - – XMSSink Send messages to applications that use WebSphere MQ





Integration and Analytics: the Text Toolkit

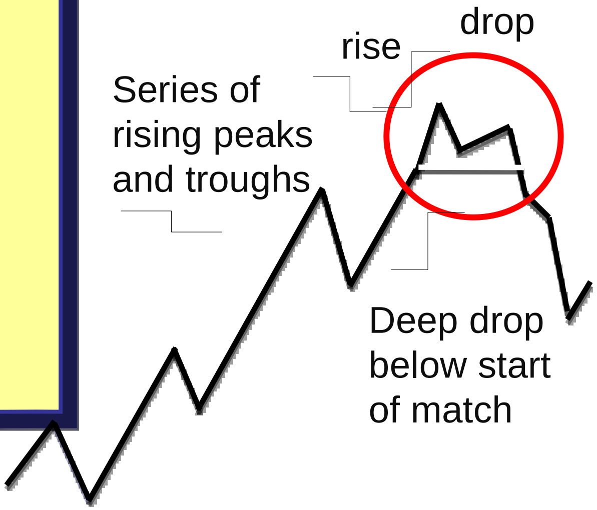
- **Derive structured information from unstructured text**
  - Apply extractors
    - Programs that encode rules for extracting information from text
    - Written in AQL (Annotation Query Language)
    - Developed against a static repository of text data
    - AQL files can be combined into modules (directories)
    - Modules can be compiled into TAM files
    - **NOTE: Old-style AOG files not supported by BigInsights 2.0 and this toolkit**
      - **Can be used in Streams 3.0 with the Deprecated Text Toolkit**
- **Streams Studio plugins for developing AQL queries**
  - Same tooling as in BigInsights
- **Operator and utility**
  - TextExtract   Run AQL queries (module or TAM) over text documents
  - createtypes script
    - Create stream type definitions that match the output views of the extractors.
- **Plays a major role in accelerators**
  - SDA: Social Data Analytics
  - MDA: Machine Data Analytics

Analytics: The Complex Event Processing Toolkit

- **MatchRegex** **Use patterns to detect composite events**
  - In streams of simple events (tuples)
  - Easy-to-use regex-style pattern match of user-defined predicates
- **Integration in Streams allows CEP-style processing with high performance and rich analytics**

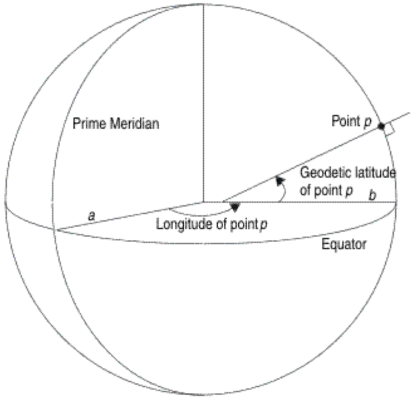
Example: detect “M-shape” patterns in stock prices: double-top formations

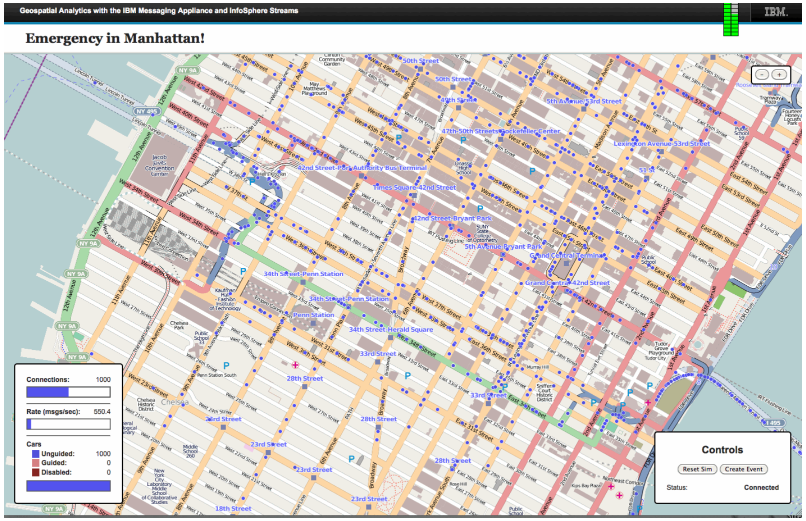
```
stream<MatchT> Matches = MatchRegex(Quotes) {
  param
    pattern : ". rise+ drop+ rise+ drop* deep" ;
    partitionBy : symbol ;
    predicates : {
      rise = price >  First(price) && price >= Last(price),
      drop = price >= First(price) && price <  Last(price),
      deep = price <  First(price) && price <
Last(price) };
  output
    Matches : symbol = symbol, seqNum = First(seqNum),
              count = Count(), maxPrice = Max(price);
}
```


drop
rise
Series of rising peaks and troughs
Deep drop below start of match

Analytics: The Geospatial Toolkit

- **High-performance analysis and processing of geospatial data**
- **Enables Location Based Services (LBS)**
  - Smarter Transportation: monitor traffic speed and density
  - Geofencing: detect when objects enter or leave a specified area
- **Geospatial data types**
  - e.g. Point, LineString, Polygon
- **Geospatial functions**
  - e.g. Distance, Map point to LineString, IsContained, etc.




Geospatial Analytics with the IBM Messaging Appliance and InfoSphere Streams

Emergency in Manhattan!

Connections: 1000

Rate (msgs/sec): 550.4

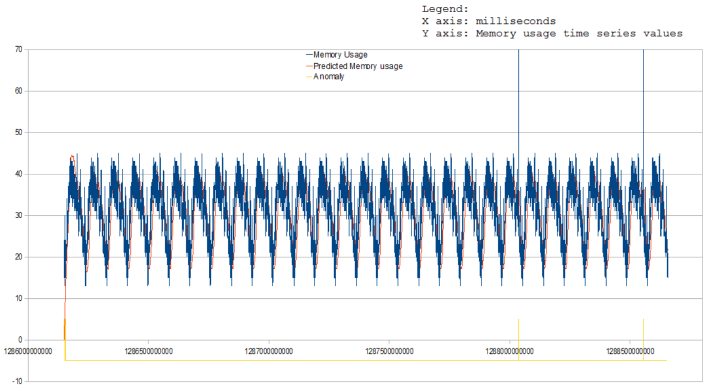
Cars
- 🟦 Unguided: 1000
- 🟥 Guided: 0
- 🟥 Disabled: 0

Controls

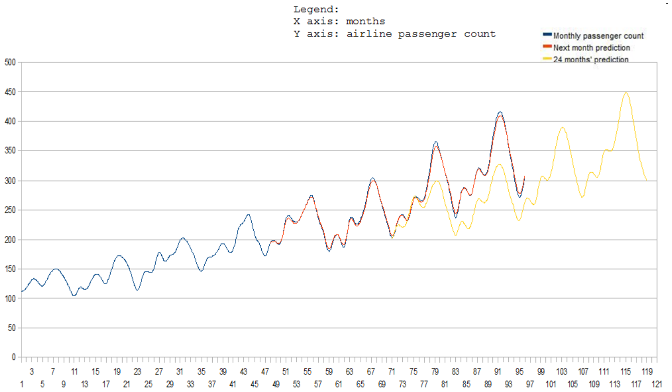
Status: Connected

Analytics: The TimeSeries Toolkit

- **Apply Digital Signal Processing techniques**
  - Find patterns and anomalies in real time
  - Predict future values in real time
- **A rich set of functionality for working with time series data**
  - Generation
    - Synthesize specific wave forms (e.g., sawtooth, sine)
  - Preprocessing
    - Preparation and conditioning (ReSample, TSWindowing)
  - Analysis
    - Statistics, correlations, decomposition and transformation
  - Modeling
    - Prediction, regression and tracking (e.g. Holt-Winters, GAMLearner)



The time series simulates memory consumption from a computer . FMP is used for prediction and anomaly detection



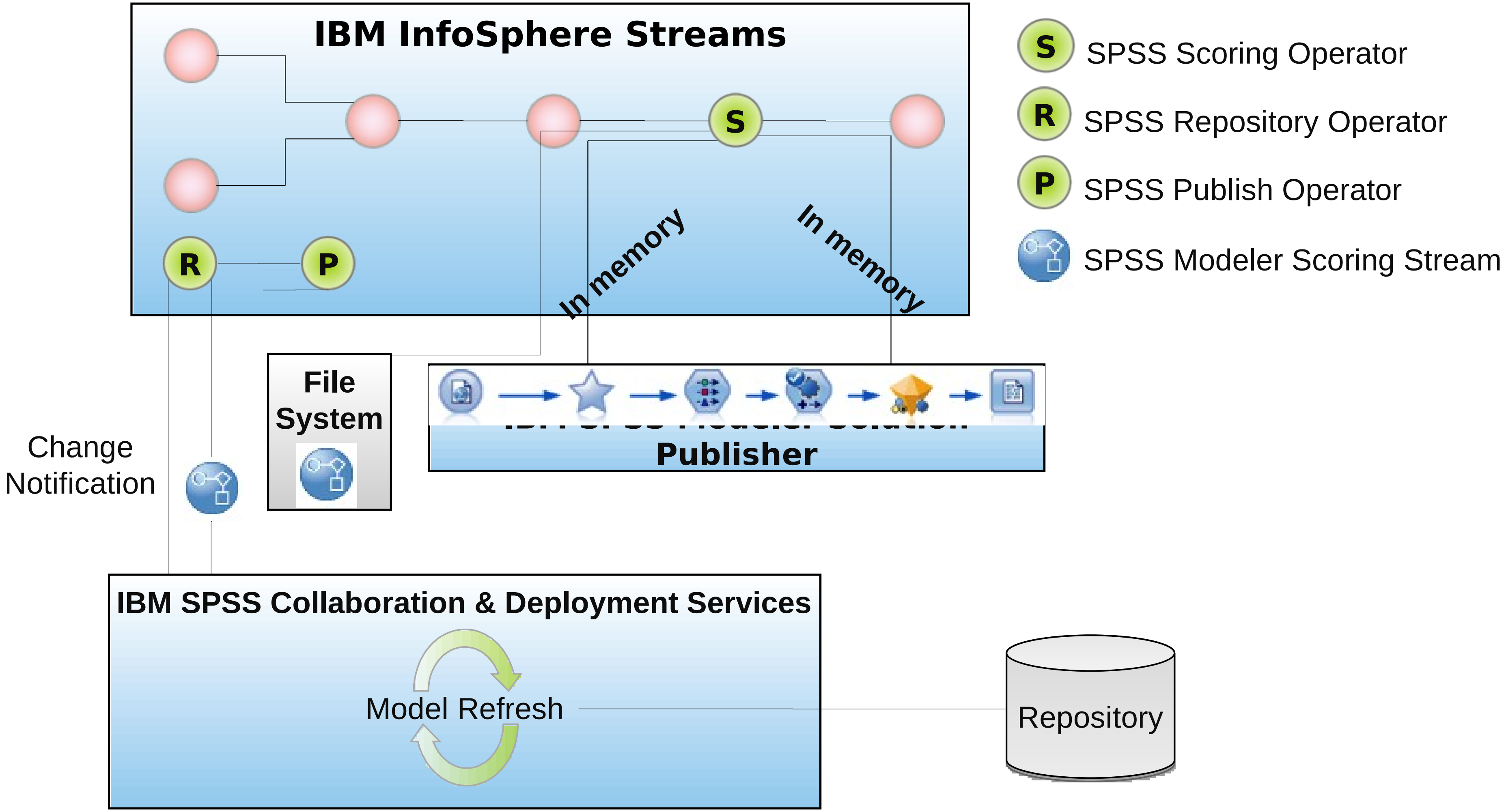
Holt Winters algorithm used for predicting next month and next 24 months ahead airline passengers count

Analytics: the Mining Toolkit

- **Enables scoring of real-time data in a Streams application**
  - Scoring is performed against a predefined model (in PMML file)
  - Supports a variety of model types and scoring algorithms
- **Predictive Model Markup Language (PMML)**
  - Standard for statistical and data mining models
  - Produced by Analytics tools: SPSS, ISAS, etc.
  - XML representation defined by http://www.dmg.org/
- **Scoring operators**
  - Classification Assign tuple to a class and report confidence
  - Clustering Assign tuple to a cluster and compute score
  - Regression Calculate predicted value and standard deviation
  - Associations Identify the applicable rule and report consequent (rule head), support, and confidence
- **Supports dynamic replacement of the PMML model used by an operator**

Analytics: Streams + SPSS Real Time Scoring Service

- **Included with SPSS, not with Streams**



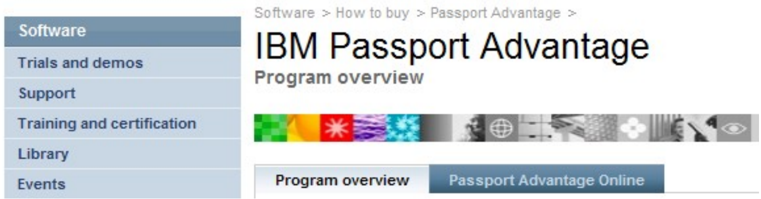
Analytics: the Financial Services Toolkit

- **Adapters**
  - Financial Information Exchange (FIX)
  - WebSphere Front Office for Financial Markets (WFO)
  - WebSphere MQ Low-Latency Messaging (LLM) Adapters
- **Types and Functions**
  - OptionType (put, call), TxType (buy, sell), Trade, Quote, OptionQuote, Order
  - Coefficient of Correlation
  - “The Greeks” (put/call values, delta, theta, etc.)
- **Operators**
  - Based on QuantLib financial analytics open source package.
  - Compute theoretical value of an option (European, American)
- **Example applications**
  - Equities Trading
  - Options Trading

RESOURCE LINKS

Where to get Streams

- **Software available via**
  - Passport Advantage site

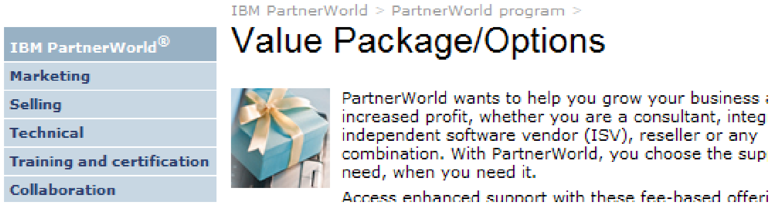


Software > How to buy > Passport Advantage >

IBM Passport Advantage

Program overview

  - PartnerWorld



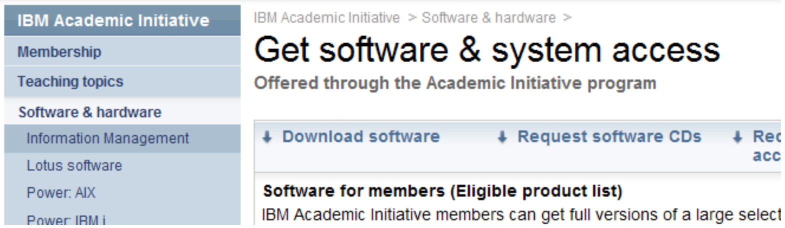
IBM PartnerWorld > PartnerWorld program >

Value Package/Options

PartnerWorld wants to help you grow your business a increased profit, whether you are a consultant, integ independent software vendor (ISV), reseller or any combination. With PartnerWorld, you choose the sup need, when you need it.

Access enhanced support with these fee-based offeri

  - Academic Initiative



IBM Academic Initiative > Software & hardware >

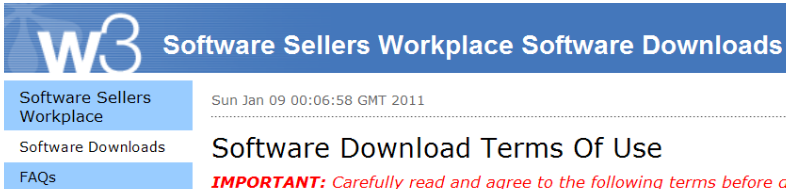
Get software & system access

Offered through the Academic Initiative program

Software for members (Eligible product list)

IBM Academic Initiative members can get full versions of a large select

  - IBM Internal



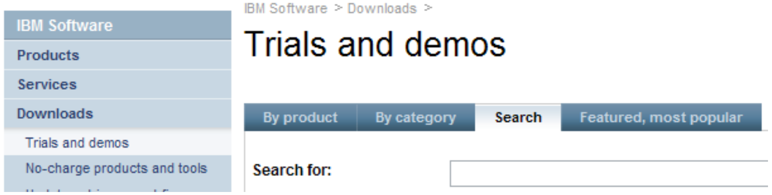
Software Sellers Workplace Software Downloads

Sun Jan 09 00:06:58 GMT 2011

Software Download Terms Of Use

IMPORTANT: *Carefully read and agree to the following terms before c*

  - 90-day trial code
    - IBM Trials and Demos portal
    - ibm.com -> Support & downloads ->Download -> Trials and demos



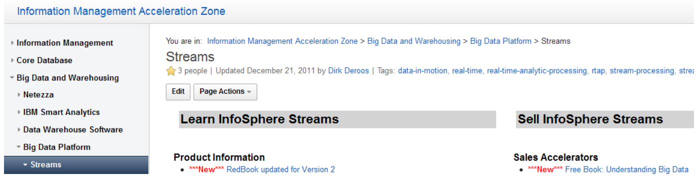
IBM Software > Downloads >

Trials and demos

Search for:

More Streams links

- **Streams in the Information Management Acceleration Zone**



Information Management Acceleration Zone

You are in: Information Management Acceleration Zone > Big Data and Warehousing > Big Data Platform > Streams

Streams

3 people | Updated December 21, 2011 by Dirk Deroos | Tags: data-in-motion, real-time, real-time-analytic-processing, rtap, stream-processing, strea

Learn InfoSphere Streams

Product Information

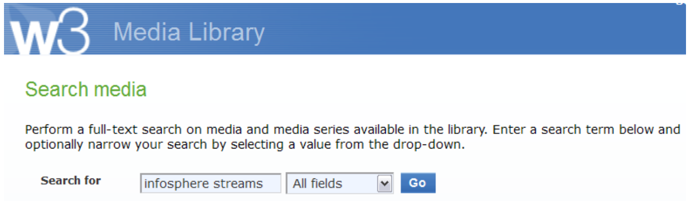
- ***New*** RedBook updated for Version 2

Sell InfoSphere Streams

Sales Accelerators

- ***New*** Free Book: Understanding Big Data

- **Streams in the Media Library (recorded sessions)**
  - Search on “streams”



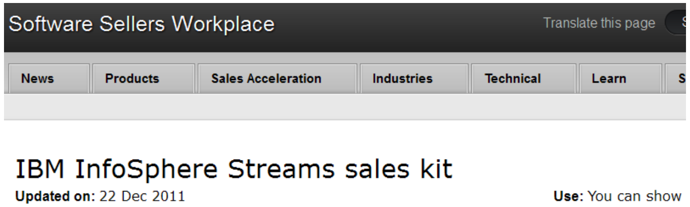
W3 Media Library

Search media

Perform a full-text search on media and media series available in the library. Enter a search term below and optionally narrow your search by selecting a value from the drop-down.

Search for: infosphere streams | All fields | Go

- **The Streams Sales Kit**
  - Software Sellers Workplace (the old eXtreme Leverage)



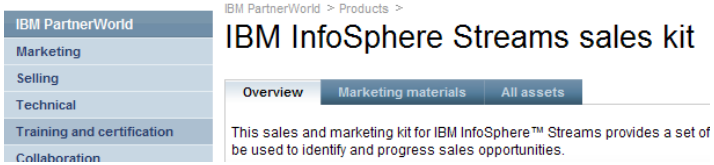
Software Sellers Workplace

IBM InfoSphere Streams sales kit

Updated on: 22 Dec 2011

Use: You can show

  - PartnerWorld

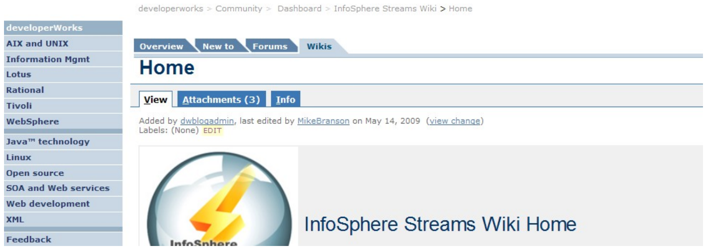


IBM PartnerWorld > Products >

IBM InfoSphere Streams sales kit

This sales and marketing kit for IBM InfoSphere™ Streams provides a set of be used to identify and progress sales opportunities.

- **InfoSphere Streams Wiki on developerWorks**
  - Discussion forum, Streams Exchange



developerworks > Community > Dashboard > InfoSphere Streams Wiki > Home

Home

Added by dwblogadmin, last edited by MikeBranson on May 14, 2009 (view change)
Labels: (None) EDIT

InfoSphere Streams Wiki Home

Thank You