Managing data flows

Martyn Winn
Scientific Computing Dept.
STFC Daresbury Laboratory
Cheshire

8th May 2014

Overview

- Sensors $\Rightarrow$ continuous stream of data
  - Store / transmit / process *in situ*?
  - Do you need to keep all the data?
  - Data streams
- Mobile computing
  - Need for energy efficiency
  - Porting of apps to mobile devices
  - Efficient use of cores
- Telemetry / Networking
  - Drip feed data over 3G / 4G
- Ethics
  - IP
  - Privacy

To keep or not to keep

X-ray crystallography: reduce diffraction images to list of Bragg peak intensities

Next generation sequencing: reduce images to base calls

CERN: filter data for interesting signals

SKA: Never throw data away - cannot re-run universe

Data Streams

- A data stream is an ordered sequence of instances that can be read only a small number of times using limited computing and storage capabilities.
- In the data stream model, some or all of the input data that are to be operated on are not available for random access from disk or memory, but rather arrive as one or more continuous data streams.
- Examples of data streams include computer network traffic, phone conversations, ATM transactions, web searches, and sensor data.

Data are dynamic events rather than a static resource

Use Cases for Data Streams

- 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

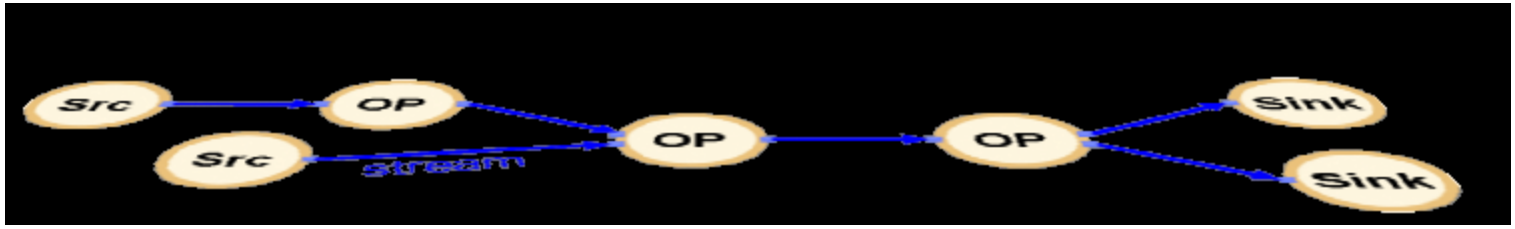
- Smart cities
  - Managing utilities

- Health monitoring
  - Real-time analytics and correlations on physiological data streams
  - Blood pressure, Temperature, EKG, Blood oxygen saturation etc.
- Environment
  - Pollution monitoring, pathogen alert

Data Stream Mining

- Process of extracting knowledge structures from continuous, rapid data records.
- Mining includes elements of machine learning / incremental learning.
- Anomaly detection, e.g. "Anomaly detection in streaming environmental sensor data: A data-driven modeling approach", Environmental Modelling & Software (2010), **25** (9) 1014 http://dx.doi.org/10.1016/j.envsoft.2009.08.010
  - to identify measurement errors in a windspeed data stream from Corpus Christi, Texas
- http://siam.omnibooksonline.com/2011datamining/data/papers/023.pdf#page=1
  - Example of single pass through genetic data. Efficient processing of static but large datasets

IBM InfoSphere Streams



- 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)

- Application written in Streams Processing Language (SPL)
- This can wrap C or Java code.

Practicalities

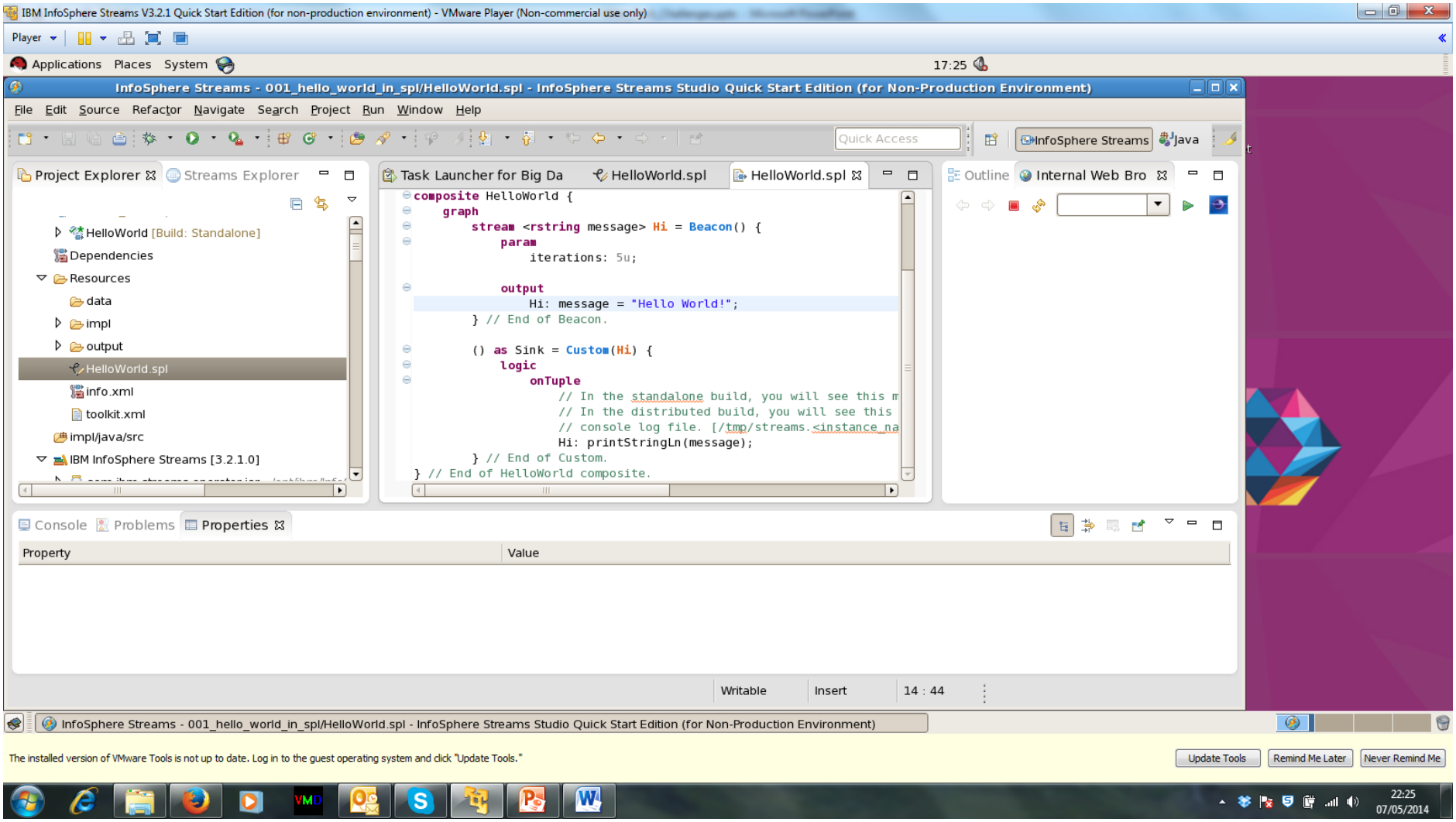
- Free download of InfoSphere Streams Quick Start Edition:

http://www-01.ibm.com/software/data/infosphere/streams/quick-start/downloads.html

- Stand-alone version or VMWare RedHat image (run in free VMWare player)
- Tried at RAL for monitoring ISIS (neutron source) data for error messages

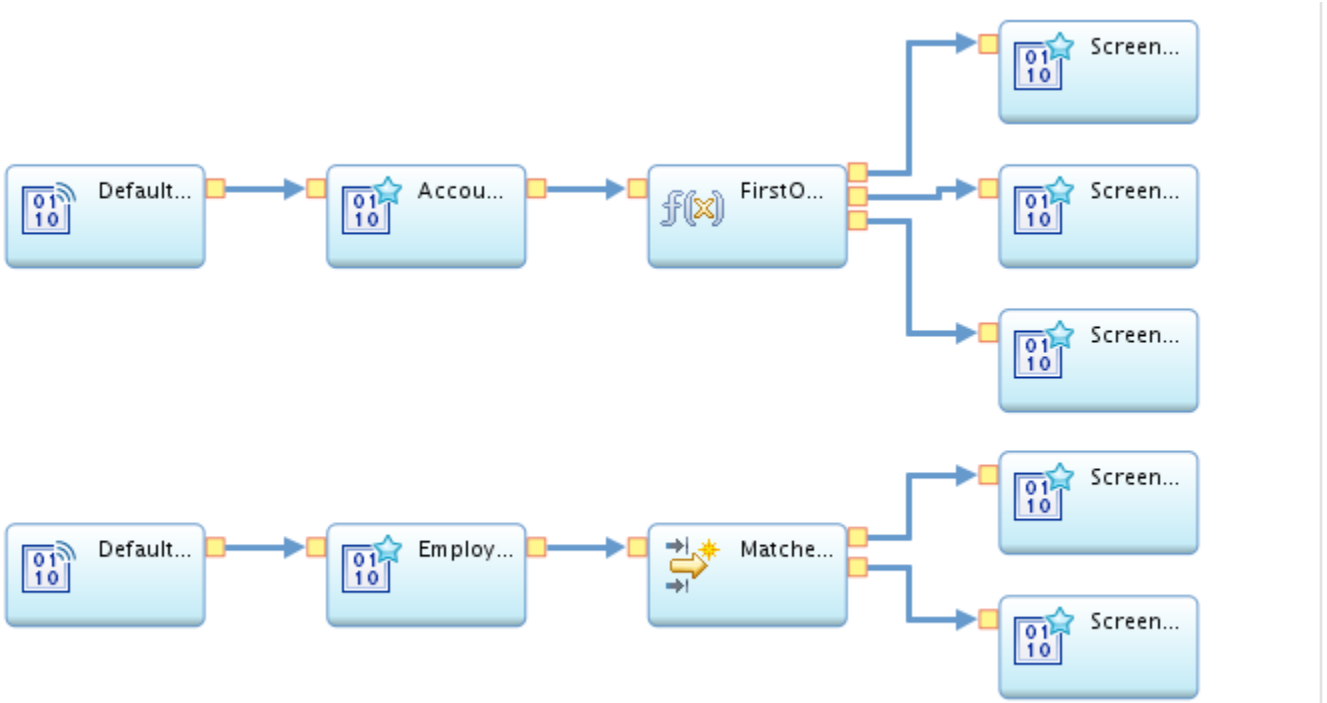
- Also available: OpenSourceStorm

http://storm.incubator.apache.org/



Eclipse environment for IBM InfoSphere Streams

Graphical editor for SPL workflows



Energy Efficient Computing

- For traditional computing, increasing emphasis on energy efficiency (electricity bills for large compute/data centres!)

- Current petaflop computers consume > 10 MW
- Desire to keep exaflop supercomputers < 20 MW
- Driving energy efficiency measures
- Supercomputing in mind, but may apply more widely?

Energy Efficient Computing

Green 500 (cf "Top500") provides a ranking of the most energy-efficient supercomputers in the world (performance-per-watt).

- Top ten of the Green500 use a similar architecture, i.e., Intel CPUs combined with NVIDIA GPUs

ARM processors

- Dominate mobile computing due to low power consumption
- Not previously in desktops / supercomputers due to low performance, but now being seriously looked at (e.g. plans at BSC)
- 64-bit ARMv8-A architecture, announced in October 2011. Released 2013 / 2014.
- AMD has taken out a license from ARM and will start designing both ARM chips and AMD64 chips, leaving only Intel designing solely its own proprietary designs.

Raspberry Pi etc

Raspberry Pi

Processor: ARMv6l (as in first iPhone)

Cotton Candy

Dual core ARM Cortex A9

Parrallella

Bioinformatics on ARM

4273π: Bioinformatics education on low cost ARM hardware - Daniel Barker (St Andrews),

http://www.biomedcentral.com/1471-2105/14/243/abstract

Commentary at:

http://www.qfab.org/2013/08/22/bionformatics-future-hardware-from-arm/

"much of the software used in bioinformatics is open source and can be easily compiled for the new ARM servers"

"if bioinformatics could standardise on a few good algorithms/applications for mainstream high-throughput analysis pipelines, then ARM chips could be extended with specialised hardware, in the same way that they are used for JPEG and movie compression on smartphones and cameras"

STFC Energy Efficient Computing centre

- £19m capital investment to finance research into energy efficient computing http://www.stfc.ac.uk/2557.aspx

- The STFC press release refers to:

  £19m “to firmly establish the UK as the world leader in energy efficient supercomputer software development to meet big data challenges.”

- Includes:
  - 1152 x 64-bit ARM cores
  - "deep fat fryer" (oil immersion)
  - Active Storage (reduce energy cost of moving data)

LSF - IBM Platform Computing

Work on improving usage of compute nodes.

Reduce power on inactive modes:

- Use of S-states (sleep) and C-states (cpu power)
- LSF monitors inactive nodes, and change state according to specified policy
- Yet knows they are available for job submission
- Implemented via IPMI (Intelligent Platform Management Interface)

CPU frequency management

- Energy consumption $\propto$ frequency^3
- Clock down, increase time, save energy (?)
- LSF can set CPU frequency "bsub -freq 1.8GHz ..."
- Also obtain energy consumption reports
- Relies on certain Linux kernel modules (e.g. ibmaem)
- Uses "on demand" CPU frequency governor
- Set energy policy and energy tags for applications

Data Processing Issues

- Compute on-site vs compute in cloud
- Data reduction or filtering
- Data streams vs static datasets
- Compute on-site needs energy efficiency
- Active control of compute
- Software, software, software