October 22-23, 2007

High Volume Data Flows:

Introduction to Pervasive DataRush™

Jim Falgout

Solutions Architect, Integration Products

AGILE INTEGRATION – FOR EVERYONE!

Agenda

- Market Forces
- DataRush Overview
- DataRush Architecture
- Lighthouse customer program
- Demo

Tectonic Shift – Multi-core

"…the world is on the leading edge of probably the most important architectural transition in the history of the microprocessor"

– Justin Rattner, Intel CTO

Quad-cores are a Reality

*”With four execution cores, the Intel® Core™ 2 Quad processor …makes the most of **highly threaded applications**. Whether you're creating multimedia, annihilating your gaming enemies, or running compute-intensive applications at one time, new quad-core processing will **change the way you do everything**. Pioneer the new world of quad-core and unleash the power of **multithreading.**”*

– Intel ad copy

Mix In a Few Disks …

“A single hard drive with four terabytes of storage (4TB) could be a reality by 2011, thanks to a nanotechnology breakthrough by Japanese firm Hitachi.”

Quote from BBC article
“Drive advance fuels terabyte era”
October 15, 2007

A Desktop Supercomputer

- Not too distant future:
  - Dual processor (4 or more cores each)
  - 4+ Terabytes of disk space
  - Tens of Gigabytes of memory
  - All for under $5k (?)
  - Order on dell.com or pick one up at Fry's on your way home from work …

Bigger than a Desktop

- Network attached computing
  - Hundreds of cores
  - Hundreds of Gigabytes of memory
  - Hardware transactional memory (HTM)
  - Compact solution (few slots in a rack)
  - Running 64-bit Java!
- Years in the future?
  - Visit www.azulsystems.com

The Result

- Increased compute power
  - Not by traditional clock speed increase
  - Increased processor cores
  - Large memory systems
- Increased disk storage
  - Ability to store more data than ever before
- Where there is ability …demand will grow!
  - Maintain larger data stores
  - Mine data for more intelligence
  - Process more data in less time

The Software Paradox

- Most batch software is not ready for multi-core
  - Scripts
  - Single-threaded Java/C++
  - Will not scale
    - Will not run faster as more cores are added
  - May run slower on multi-core processor
    - Single core speeds have slowed

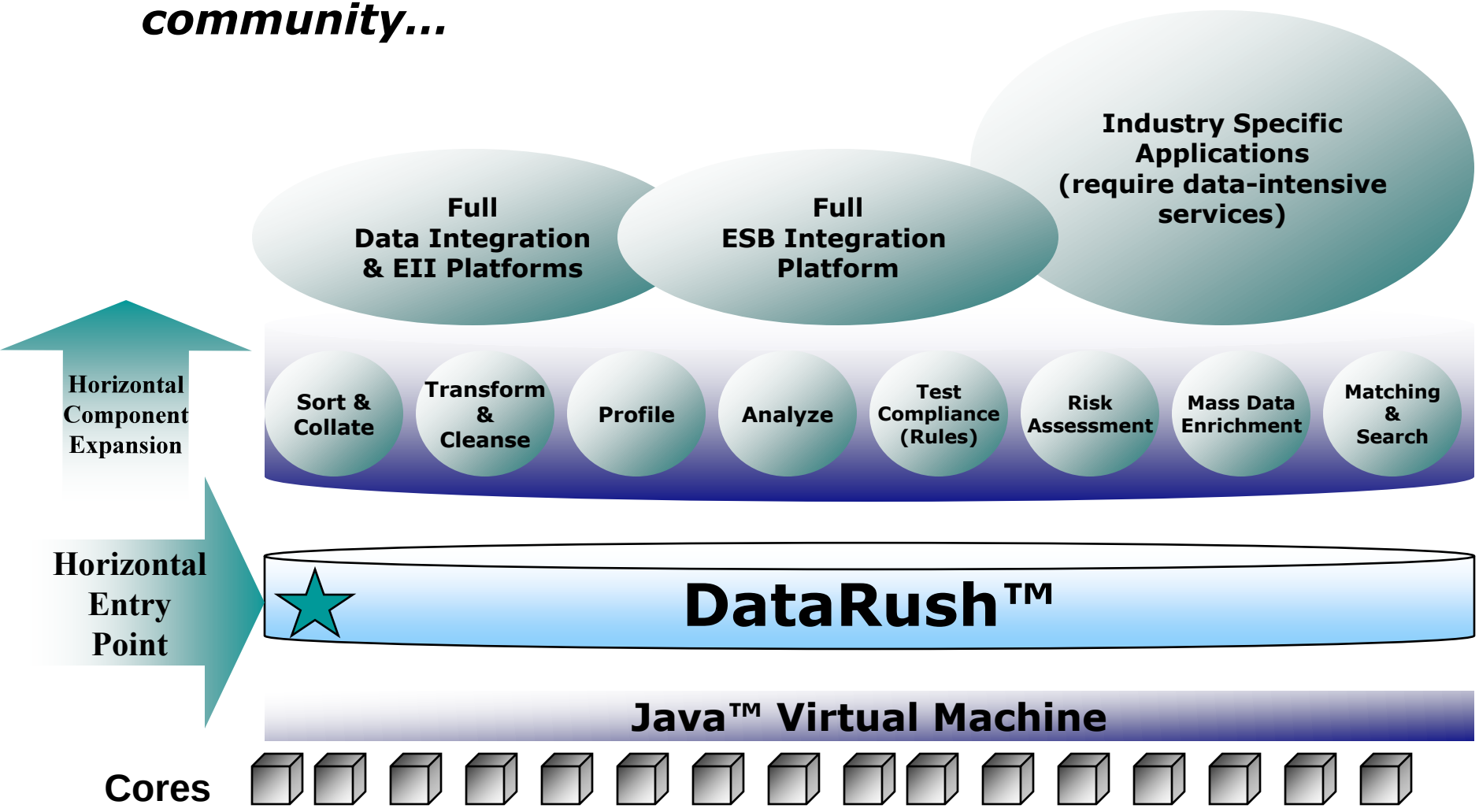
DataRush™ Answers the Challenge

- Scalable Java based engine and SDK for data-intensive multi-core SMP systems
- Build components that are “parallelized” by a powerful framework at runtime
- Integrates fully with Business Integrator 9.X to handle data profiling and certain types of transforms

Placing high-performance computing in the hands of all developers

DataRush™ Direction

For the enterprise architect and Java developer community…



DataRush Overview

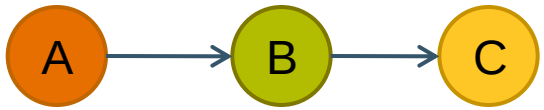
- Extensible, component-based architecture for developing highly scalable Java applications
- Hide complexity of parallel programming
  - Framework handles locking, threading and memory
  - Java components don't require locking/threading code
- Provide high level programming constructs
  - XML based dataflow language
  - Productivity via use of composition and "customizers"
  - Eclipse-based studio
- Dynamic scalability
  - Take advantage of more CPU resources as available
  - Scaling up AND down (small dept level systems)

DataRush Makes It Easier, but It Is Not a Black Box or Auto-Transmogrifier

- Requires some rewrite, not just a port or recompile
- Requires a new way of describing the process – drxml
- Uses a new graphical design approach in Eclipse IDE
- Requires a fundamentally new way of thinking about the application – a pipelined architecture

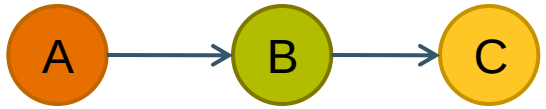
DataRush implements dataflow

- Based on Kahn networks; Parks scheduling
- Provides functional parallelism
- Program represented as a directed graph
  - Nodes represent a computational unit
  - Edge represents a one-way FIFO data queue
  - Nodes have 0..* input and output edges
  - Nodes communicate only over these edges

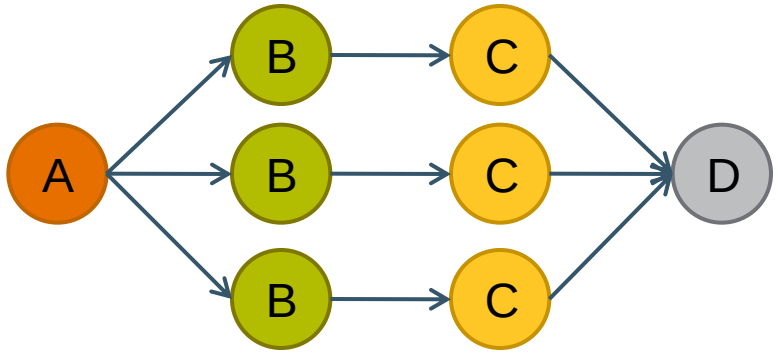


Parallelization Techniques in DataRush

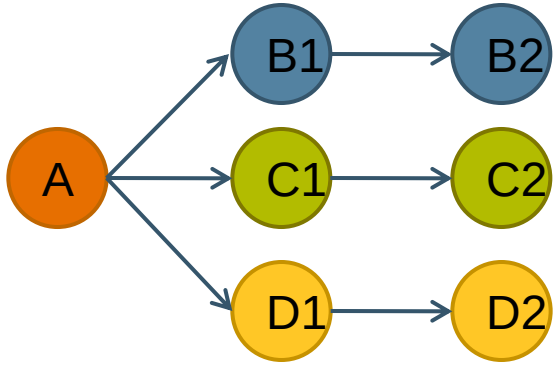
- Pipeline



- Horizontal Partitioning



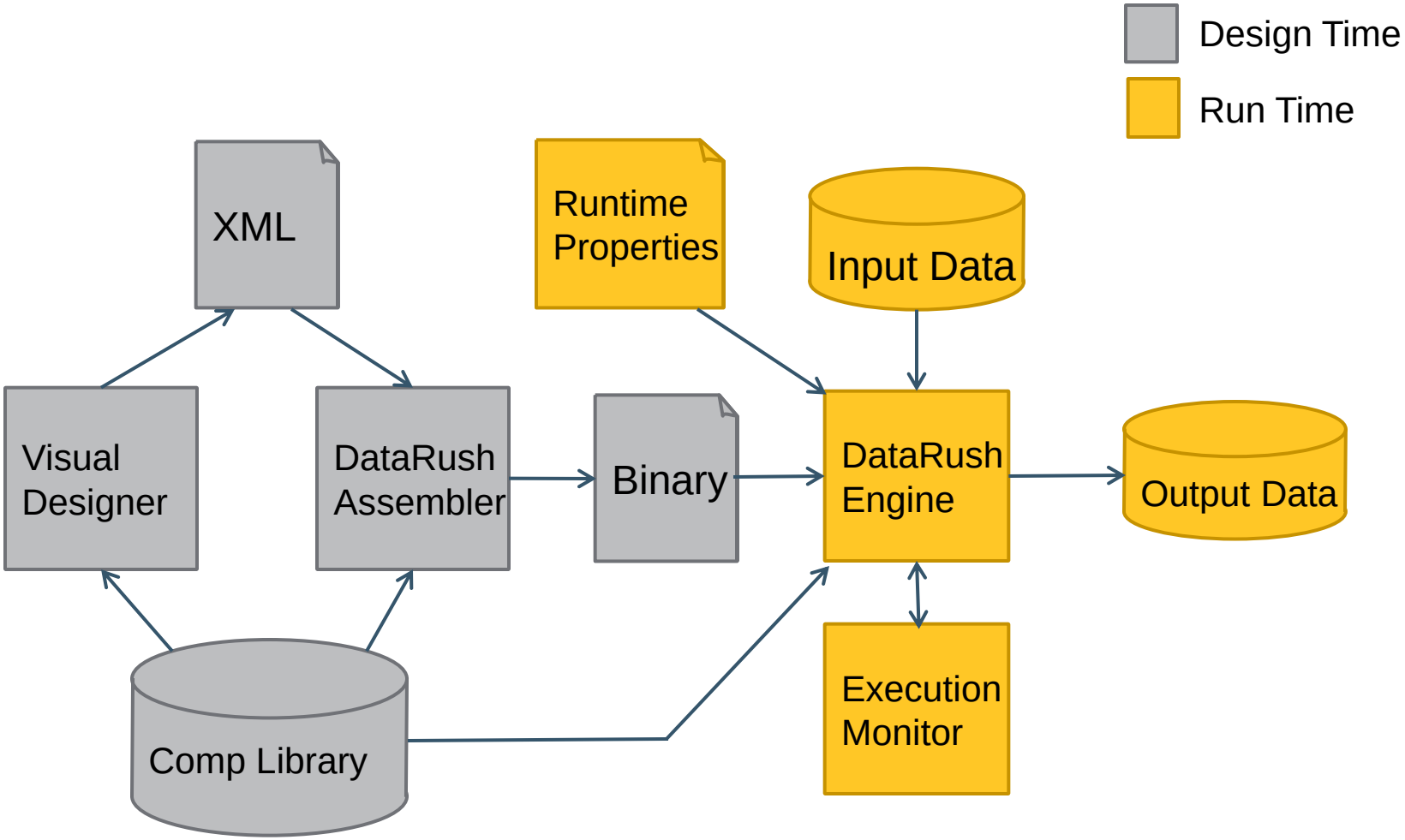
- Vertical Partitioning



DataRush Architectural Overview

- Design Time
  - Eclipse IDE
  - Large and growing component library
    - DataRush components (customer extensible)
    - Foundational library including: I/O, sort, merge, join, …
  - DataRush assembler
    - Validates structure and dependencies
    - Generates binary form
- Run Time
  - DataRush compiler
    - Builds dataflow plan
    - Propagates port types and schemas (outputs to inputs)
    - Runs customizers
  - DataRush executor
    - Executes dataflow plan
    - Multi-threaded dataflow engine
    - Initiates/monitors dataflow graph execution
      - JMX console

DataRush Architecture Block Diagram



Terminology

- Assembly
  - Composite dataflow operator; DataRush XML
  - Defined contract (ports, properties)
- Process
  - Basic unit; written in Java
- Customizer
  - Assembly contained code generator (helper)
  - Invoked at compile time
  - Extends and/or refines assembly

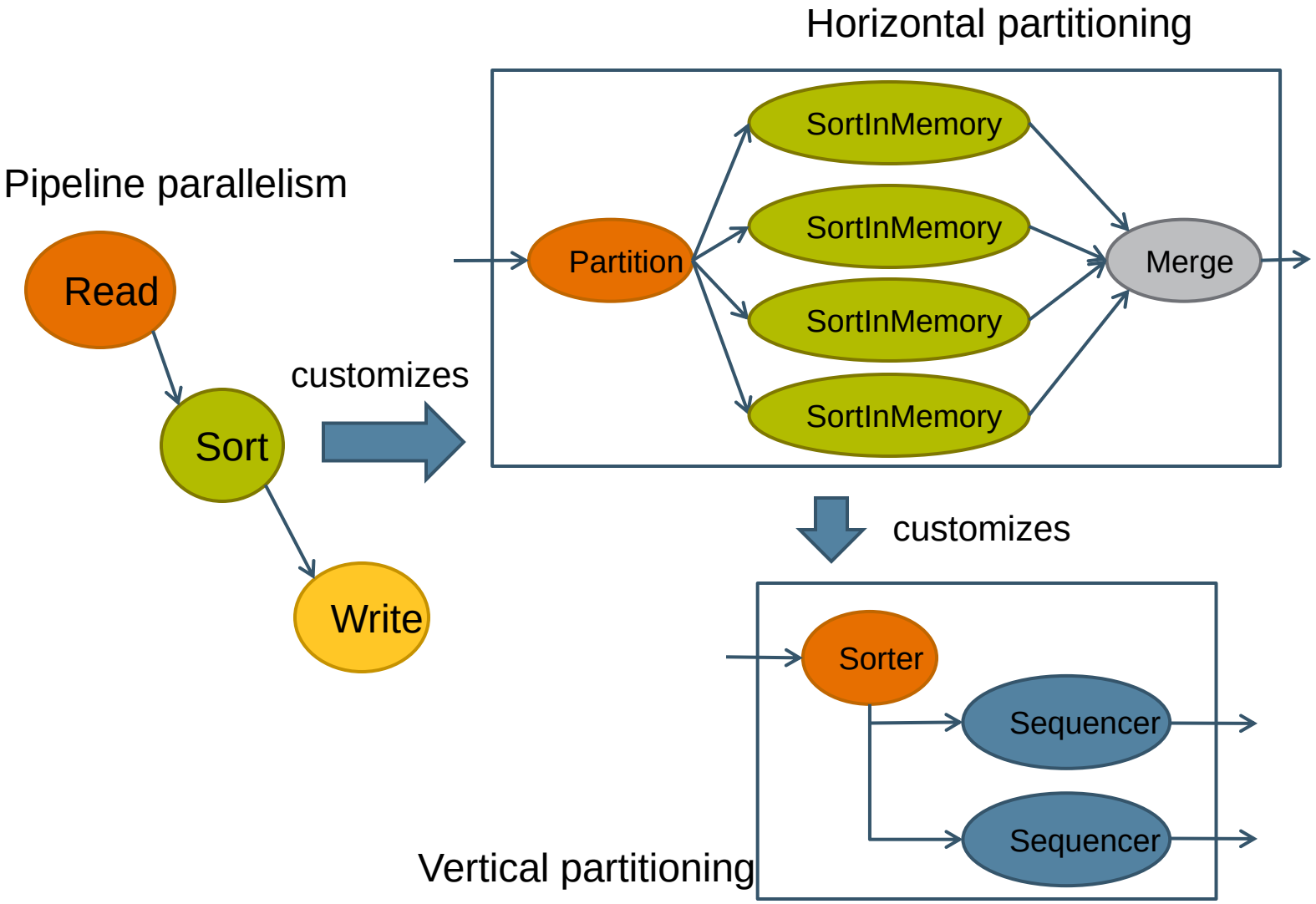
Developing with DataRush

- User builds Assemblies
  - Utilizes component library
  - Extends with own components (Java)
  - Parallelizes with a customizer (Java)
- Assembles with DataRush assembler
- Runs with DataRush engine
  - Compilation step: customizers invoked
  - Executor executes generated dataflow graph

Building Parallelized Components

- Graph generation with Customizers
  - Customizers can be used to:
    - Create horizontal partitioning
    - Create vertical partitioning
    - Support dynamic port type mappings
  - Written in Java as compiler helpers
  - Run at compile time
  - Can create additions to the dataflow graph at compile time

Customizers – code generation



Benchmark of Simple Pipeline

Using K-means clustering algorithm

	Non-threaded	Threaded
1 core	79.9 secs	132.0 secs
2 cores		48.4
4 cores		24.9
8 cores		13.6

Source: Benchmark on HP 585 with 4 dual core AMD Opteron processors

DataRush Progress

- Embedded in Pervasive DataProfiler
  - Successful commercial product since 2005
  - In use by DoD, state government offices and several large BPO organizations
- Proof of concept work
  - Integrated into BI Process Designer as an MCF component
  - Used in conjunction with DataIntegrator
  - DR for data validation; DI for transformation
- Beta 2 release now available
  - Download at http://www.pervasivedatarush.com

New in Beta 2

- New User Interface (Eclipse 3.3 based)
- New JMX performance monitoring
- Additions to operator library including support for multiple scripting languages
- Using Java 6 for enhanced parallel performance
- Supported Platforms
  - Windows XP, Server 2003, Vista
  - Linux: RedHat, Suse, Azul
  - Solaris

Light House Customers Wanted

Ideal situation:

- Java
- 4 cores or more
- Gigabytes of data
- Long running batch jobs

Innovators and Early Adopters seeking a business advantage from new approaches to difficult performance challenges…

Contacts

- We welcome your feedback and encourage you to post questions and comments at www.pervasivedatarush.com
- Or contact:

Steve Hochschild
shochschild@pervasive.com
Business Development

Jim Falgout
jfalgout@pervasive.com
Software Architect

DataRush Demo

FUTURE-PROOFING YOUR APPLICATIONS

Q&A

FUTURE-PROOFING YOUR APPLICATIONS