IBM InfoSphere DataStage 9.1

Presenter:  Tony Curcio .... IBM Product Manager and ‘Visionary’

November 2nd, 2012

DSXchange Webinar Series

Editor – Dennis James
IBM InfoSphere DataStage 9.1 Update

Questions and suggestions regarding presentation topics? - send to editor@dsxchange.com

Downloading the presentation

http://www.dsxchange.net/IIS91_whats_new.html

Replay will be available within one day with email with details

Pricing and configuration - send to editor@dsxchange.net Subject line: Pricing

For those that stay through the entire presentation, we have a extra give away!

Bonus Offer – Free premium membership for your DataStage Management! Submit your management’s email address and we will offer him access on your behalf.

Email Info@dsxchange.net subject line “Managers special”.

Join us all at Linkedin http://tinyurl.com/DSXmembers
A Detailed Look at IBM InfoSphere DataStage 9.1
Session Number INT-1124

Tony Curcio, Product Manager
Shyam Mudambi, Senior Development Manager
Please note

IBM’s statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM’s sole discretion.

Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user’s job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.
Agenda

• Review Recent Activity
• New Information Server 9.1 Data Integration Features
• Roadmap Peek
• Follow-ups
IBM InfoSphere Integration and Governance

**Business Driven**
- Industry Focused → Fastest ROI
- Total Governance → Trusted Data
- Agile Frameworks → Lowest TCO

**Smarter**
- Single Platform
- Unified Blueprints
- Cloud Ready

**Faster**
- Best Performance
- Fast Provisioning
- Rapid Discovery
The IBM Big Data Solution

New analytic applications drive the requirements for a big data platform

• Integrate and manage the full variety, velocity and volume of data
• Apply advanced analytics to information in its native form
• Visualize all available data for ad-hoc analysis
• Development environment for building new analytic applications
• Workload optimization and scheduling
• Security and Governance
Information Server Release Review
Data Integration Features

2011

8.5 FP1

- New Suite Installer
- Source Code Control Integration
- Looping Transformer
- Scalable XML
- Vertical Pivot
- Connector Multilink Support
- Distributed Transaction Stage
- zOS File Stage
- Connection Migration Tool
- Audit Logging
- Balanced Optimization
- Data Masking Stage
- BluePrint Director
- ISA Lite

2012

8.7 FP1

- Operations Console
- Big Data File Stage
- Parallel Job Debugger
- Job Log Designer Integration
- Netezza Connector & Balanced Optimization
- Real-time CDC Integration
- Data Rules Stage (with IA)
- Designer & Engine Performance
- Command Line Backup
- Security Features
- Metadata Asset Manager

2013

9.1 FP2

- New Suite Installer
- Source Code Control Integration
- Looping Transformer
- Scalable XML
- Vertical Pivot
- Connector Multilink Support
- Distributed Transaction Stage
- zOS File Stage
- Connection Migration Tool
- Audit Logging
- Balanced Optimization
- Data Masking Stage
- BluePrint Director
- ISA Lite
New and exciting partner apps
Several can be seen on the demo floor !!!

iKnow

Provides data integration automation, monitoring and analysis features to enable trusted information in complex environments.
IOD Booth #419
www.iknowsw.com

Intelcom

TCS Info ProcessWorks

A tool for ETL Process:
- Automation
- Standardization
- Control
- Monitoring
IOD Booth #G512
http://tinyurl.com/infoprocessworks
Email Teg.Ibmim@tcs.com

Presentation in the Expo area
Tues 6 to 6:30 pm

TWS for Apps 8.6

Automates and integrates DataStage jobs and sequencers within the larger TWS environment for maximum efficiency.


TestDrive

complete testing framework to design test cases, create test data, specify expected results, run test scripts, report test status and manage all of the testing artifacts over time
IOD Booth #321
http://www.compactbi.com/solutions/testdrive

Stream Quality Mgr

Enterprise DQM solution that provides a view into the quality of your data insuring the data is capable of meeting it’s intended purpose.
IOD Booth #G405
www.streamintegration.com

Tivoli

iDirector

Keep you informed of the status of your DataStage jobs and help you control the runs of jobs remotely
http://idirector.inwoo.co.kr/sites/index.html

iKnow

TCS Info ProcessWorks

TWS for Apps 8.6

TestDrive

Stream Quality Mgr

iDirector
Agenda

• Review Recent Activity
• New Information Server 9.1 Data Integration Features
  • Roadmap Peek
• Follow-ups
IBM InfoSphere Information Server 9.1

**Agile Integration**
Wherever your integration resides, integrate it quickly and flexibly

**Business Driven Governance**
Make decisions with confidence using trusted data at the point of impact

**Sustainable Quality**
Ensure information accuracy and quickly adapt to strategic business changes
## Six Dimensions of Capability Investment

<table>
<thead>
<tr>
<th>Productivity</th>
<th>Transformation</th>
<th>Connectivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rich user interface features that simplify the design process and metadata management requirements</td>
<td>Extensive set of pre-built objects that act on data to satisfy both simple &amp; complex data integration tasks</td>
<td>Native access to common industry databases and applications exploiting key features of each</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Performance</th>
<th>Operations</th>
<th>Administration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Runtime engine providing unlimited scalability for all integration patterns batch/real-time, ETL/ELT/DV/SOA</td>
<td>Simple management of the operational environment lending analytics for understanding and investigation.</td>
<td>Intuitive and robust features for installation, maintenance, configuration, security and resiliency</td>
</tr>
</tbody>
</table>
Information Server Home Page
Single starting place for access to Information Server Web UIs
Access new home page from your Start menu
Provides a central location for access to Information Server UIs
Productivity
Simplifying user experience to deliver business results faster

XML Schema Views
Focus on the business object
New XML “schema views” narrow the scope of the schema tree to enhance the design experience and improve the efficiency of the parsing and composing processes.

Rational Team Concert Integration
Enterprise control of assets
Extends the certification of our asset interchange integration with source code control systems to now include Rational Team Concert

InfoSphere Data Click
Load your data mart in 2 clicks
New web-based experience allows novice users to move data in batch or real-time from databases or InfoSphere MDM via a few simple clicks with full set of governance and run-time optimization.
Demand for Analytics

Thirst for Information

Speed of Trusted Information Delivery

- **Agility** users want to be empowered to access information on demand in minutes or hours (i.e. self-service data integration)

- **Governance** delivery cycles to satisfy requests for trusted information within a governed framework are typically days or week

- **Organizations can’t sacrifice data governance in the pursuit of agility**
InfoSphere Data Click
Agility & Governance for Self Service Analytics

1. **Built-in Governance:**
   Define and control scope & architecture

2. **Business Agility:**
   get your data in just 2 clicks

3. **Built-in Governance:**
   Review and manage environment & compliance
InfoSphere Data Click

Simple “two clicks” to data integration

- Preselected sources and targets minimize user interaction
- Checked tabs indicate that the required configuration is complete

1st Click

- Review configuration

2nd Click

Execution!!!
What did Data Click just do?

Running on the power of Information Server, Data Click Provides:

Automated creation of:
- target database schema
- bulk extract and load (with DataStage)
- real-time replication subscriptions (with CDC)

Managed workload for optimized performance of potentially 1000s of artifacts

Ensured governance around both data access as well as for metadata capture (to support data lineage and impact assessment)

Initial release support Oracle/DB2 sources and Netezza targets – others to follow
New Transformation Expressions
Enhanced String Functions
Includes the server functions for “ereplace” and “change” within the parallel transformer for better handling of string replacement.

IBM Operational Decision Mgmt Integration
Support Dynamic Business Changes
Direct integration with ODM (ILOG JRules) to provide unique rules management within the data integration framework. Where business users need to directly control the logic, this combines best of breed market capabilities to deliver the solution.

Dynamic File Creation (seq & bdfs)
Drive job logic through data values
New “Generate Multiple Files” property drives a dynamic nbr of output files based on data values, without needing multiple stages or recoding each time the output set number changes.
Dynamic File Generation

• **Overview**
  • Create multiple files formed from a base name
  • Drive file roll over from keys value or file size
  • Include the key value as part of the file name
  • Combine with a sequential source of file pattern to create an output file
  • Avoids job updates when a new output file is required

• **Patterns to Generate Multiple Files**
  1. w/ Root File String (default)
  2. w/ Root & Maximum File Size
  3. w/ Root & Key Column
  4. w/ Root & Key Column in File Name
  5. w/ Root, Key Column in File Name & Exclude Part String
Connectivity
Reaching new sources of information and exploiting new options

Java Integration Stage
Second generation extensibility
Offers simple user paradigm for integrating Java code and creates baseline for upcoming big data source support. Upcoming developerWorks articles/assets will accelerate how customers can reach new NoSQL and other big data sources.

InfoSphere Streams Integration
Open up Real-time Big Data
Direct data flow integration between InfoSphere DataStage and InfoSphere Streams to combine the power and reach of both platforms

Native Excel Access
Simple Worksheet Access
New stage provides direct Excel read capabilities on all platforms with rich features to support ranges, multiple worksheets and more
Native Excel Access

New Unstructured Stage provides:

- **3-Step Configuration** – enter file name, select data range and setup column name mapping
- **Native Excel File Support** – reads data from any supported OS (Windows, Linux, Unix)
- **XLS Version Variants** - includes both Excel 97-2003 OLE2 (.xls) and 2007 OOXML (.xlsx) file format
- **Simplified Access** - no need to define ODBC DSN for each excel file
- **Password Protected Files** – read data from password encrypted files
- **Multiple File Support** – ability to simultaneously process multiple files in different nodes
- **Multiple Sheet Support** - extract data from multiple sheets at a time
- **Runtime Column Propagation** – define column metadata
- **Excel Field Extraction** – ability to extract row numbers, comments, hyperlinks, formulas, document Properties, etc...
Excel Access

### Excel

- **Filename**: SampleData.xls

### DataStage

<table>
<thead>
<tr>
<th>RowNo</th>
<th>File</th>
<th>Sheet</th>
<th>OrderDate</th>
<th>Region</th>
<th>Rep</th>
<th>Item</th>
<th>Units</th>
<th>Unit Cost</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>SampleData.xls</td>
<td>SalesOrder</td>
<td>1/6/10</td>
<td>East</td>
<td>Jones</td>
<td>Pencil</td>
<td>95</td>
<td>1.99</td>
<td>189.05</td>
</tr>
<tr>
<td>5</td>
<td>SampleData.xls</td>
<td>SalesOrder</td>
<td>1/23/10</td>
<td>Central</td>
<td>Kivell</td>
<td>Binder</td>
<td>50</td>
<td>19.99</td>
<td>999.5</td>
</tr>
<tr>
<td>6</td>
<td>SampleData.xls</td>
<td>SalesOrder</td>
<td>2/9/10</td>
<td>Central</td>
<td>Jardine</td>
<td>Pencil</td>
<td>36</td>
<td>4.99</td>
<td>179.64</td>
</tr>
<tr>
<td>7</td>
<td>SampleData.xls</td>
<td>SalesOrder</td>
<td>2/26/10</td>
<td>Central</td>
<td>Gill</td>
<td>Pen</td>
<td>27</td>
<td>19.99</td>
<td>539.73</td>
</tr>
<tr>
<td>8</td>
<td>SampleData.xls</td>
<td>SalesOrder</td>
<td>3/15/10</td>
<td>West</td>
<td>Sorvino</td>
<td>Pencil</td>
<td>56</td>
<td>2.99</td>
<td>167.44</td>
</tr>
<tr>
<td>9</td>
<td>SampleData.xls</td>
<td>SalesOrder</td>
<td>4/1/10</td>
<td>East</td>
<td>Jones</td>
<td>Binder</td>
<td>60</td>
<td>4.99</td>
<td>299.4</td>
</tr>
<tr>
<td>10</td>
<td>SampleData.xls</td>
<td>SalesOrder</td>
<td>4/18/10</td>
<td>Central</td>
<td>Andrews</td>
<td>Pen</td>
<td>75</td>
<td>1.99</td>
<td>149.25</td>
</tr>
</tbody>
</table>

*Sheet name: SalesOrder*

*Hyperlink: mailto:andrewh@star.com*

*Comment: Check Price!*
Performance
Faster….. Oh yeah!!!

**DBMS Connector Boost**
Deliver job results faster
Performance optimizations including new “big buffer” optimizations which has increased bulk load performance in DB2 and Oracle Connector by more than 50% in many cases.

**Balanced Optimization for Hadoop**
Harness the power of Hadoop with DataStage
Allows a Balanced Optimizer customer to push levels of processing into Hadoop where locality of processing will boost performance.

**Balanced Optimization for DB2 z/OS**
Exploits the power of Z
Extends the data locality/ELT paradigm to facilitate z/OS based warehouses; aligns with and facilitates IBM IDAA strategy

**Streamlined Windows Processes**
Boosts performance and reduces overhead
Move from MKS emulation to native process forking api's to increase scalability on Windows and removing resource layers in that delivery.
New Hadoop-based Big Data Support

Big Data Integration Common Use Cases

Any to Big Data
“I need to mix in traditional sources into Hadoop so that I can run the analytical models I need.”

Big Data to Any
“Now that I know something new, how do I move this back into my applications and warehouses so that it is easily consumable.”

Big Data Hub
“I need to transform and cleanse information to make it (re)usable for analytics but can’t afford to move TBs across the network frequently.”
New Hadoop-based Big Data Support
Integrating Extreme Velocity, Variety and Volumes

- Continue Big Data investments within Information Server that began with Hadoop data file system (HDFS) integration
- Programmatic invocation of Big Insights and other MapReduce jobs through the Job Sequencer, combining DataStage and Hadoop for end-to-end workflow.
- Leverage the Designer User Interface and standard stage constructs to generate MapReduce jobs, giving users the capability to deal with BigData sources that would be more efficiently processed using Hadoop.
- Direct data flow integration between InfoSphere DataStage and InfoSphere Streams to combine the power and reach of both platforms
- Rich metadata support through all use cases.
Operations
Optimize and Manage the runtime environment

**Enhanced Operations Console Features**
Accelerate investigation & remediation
Full log display included in Job Run Detail dialogue so users can perform complete job analysis and adds run, stop, and reset actions for Operators to take action when required.

**Pre-built Operations Analytical Reports**
Gain New Insight
Pre-built reports in Cognos BI which showcase the rich operation metadata being gathered and whose schema was documented in release 8.7. Provides a great synergy story for Cognos customers & illustrates analytics to anyone using other BI tools.

**Workload Management**
Maximize performance in shared environment
Empower the data integration admin to set the policies in a shared services framework for how jobs should be run and resource thresholds should be honored to optimize system perf
Workload Management
Policy driven control of system resources in a shared service environment

- Allows the proactive mgmt of system resources where multiple teams share a common hardware infrastructure.
- Optimize hardware utilization and prioritize mission critical tasks.
- Throttle job activity where system resources exceed admin-specified thresholds.
- Assign the priority of any submitted job. Configurable across various Projects as well as at the job level.
- Manual overrides by privileged user to promote specific jobs to the top of the queue.
WLM Performance Benefits

• Greater % of CPU applied to data integration workload
  – Overloaded Systems consume disproportionate amount of “system cpu”
  – WLM minimizes “thrashing” and thereby maximizes “user cpu”

• Our performance tests brought a job stream that ran in 55 minutes down to 16 mins! (your mileage may vary)

![Bar chart showing runtime reduction with WLM]
Administration
Enterprise class support for the data integration and governance platform

PureSystems Private Cloud
Rapid deployment of optimized environments
New system pattern for Information Server Enterprise Edition provides unparalleled Private Cloud offering facilitating leaps in time to value

Migration User Interface
Reducing cost of upgrade
New Migration Assistant Export Wizard guides the user through an interview and collects information the environment to automatically create the response file for driving the migration.

New Support Features
Expediting Investigation
FFDC – generating stack traces for windows, linux and AIX platforms when a job fails; New Data Stage Runtime Trouble shooting guide for easy diagnosis and resolution of common issues
Expert Integrated Systems
Faster time to value on optimized systems

- Enables companies who are interested in cloud computing to achieve faster time to value for data intensive projects.

- Provides Data Integration and Data Quality capabilities to satisfy requirements for users developing solutions in dev/test as well as production environments.

- Simple provisioning capabilities to get started and scale the environment quickly.
Head to the Demo Room to learn more!

#20 - Big Data Integration with DataStage
- Balanced Optimization for Hadoop
- Seq File & BDFS Dynamic File Creation
- InfoSphere Streams Integration

#28 - Information Server Serviceability
- New Support Features

#29 - Getting up and running
- Migration User Interface

#32 - Information Server Data Click
- InfoSphere Data Click

#33 - Information Server Hypervisor Edition
- PureSystems Info Server Private Cloud

#44 – DataStage Managing and Monitoring
- Workload Management
- Enhanced Operations Console Features
- Operation Analytical Reports

#55 – Information Server Connectivity
- DBMS Connector Boost
- Operational Decision Mgmt / ILOG
- Native Excel Access
- Java Integration Stage
- XML Schema Views

#42 – What's New in Information Server 9.1
- For anything else, ask here for the right engineer who can help
Agenda

- Review Recent Activity
- New Information Server 9.1 Data Integration Features
- Roadmap Peek
- Follow-ups
Q&A on upgrade

When should I upgrade?
• Now! Information Server 9.1 offers tremendous benefits for customers on any prior release
• If you are already on 8.7, you will be able to move to 9.1 with an in-place upgrade package (*planned for release in Q1*)
• IBM offers a free Concierge Program to provide guidance as you move to the newest release
• To learn more, please send an email message to isconcierge@us.ibm.com

When will older versions be retired?
• InfoSphere DataStage 7.5.1 was at end-of-life in April 2012
• InfoSphere DataStage 7.5.2 and 7.5.3 will be at end-of-life in April 2013
• Information Server 8.0.1 and 8.1 are still supported and end-of-life not yet announced; however, no other maintenance fixpacks are scheduled
Communities

• On-line communities, User Groups, Technical Forums, Blogs, Social networks, and more
  o Find the community that interests you …
    o Information Management bit.ly/InfoMgmtCommunity
    o Business Analytics bit.ly/AnalyticsCommunity
    o Enterprise Content Management bit.ly/ECMCommunity

• IBM Champions
  o Recognizing individuals who have made the most outstanding contributions to Information Management, Business Analytics, and Enterprise Content Management communities
    o ibm.com/champion
Thank You!
Your Feedback is Important!

• Access SmartSite to complete your session surveys
  o Any web or mobile browser at iodsmartsite.com
  o Any SmartSite kiosk onsite
  o Each completed session survey increases your chance to win an Apple TV with daily drawing sponsored by Alliance Tech