Data Integration with IBM- An Indian Perspective

Aarti Borkar, Program Director, Information Server, IBM Inc;
Ankur K Das, Product Manager, IBM India Software Labs

InformationOnDemandIndia2011

The Premier Conference for Information Management
Manage. Analyze. Govern.

February 2, 2011
Hyatt Regency I Mumbai, India

Agenda

- Biggest Challenges for Data Integration
  - Indian Market : Challenges and Opportunities
- IBM Information Server
  - Understanding, Cleansing, Transforming and Delivering trusted data
- Key solutions to specific Data Integration Challenges
  - **Productivity** of and Flexibility for the developer
  - **Performance** and Speed of processing and delivery
  - **Accurate** information delivery

An effective technology strategy is mandatory to leverage the potential of this information

Customer Challenges

- Lack of understanding of source systems and relationships
- Lack of information governance
- Multiple versions of the truth
- Long project cycles/cost overruns
- Poor data quality undermines the business

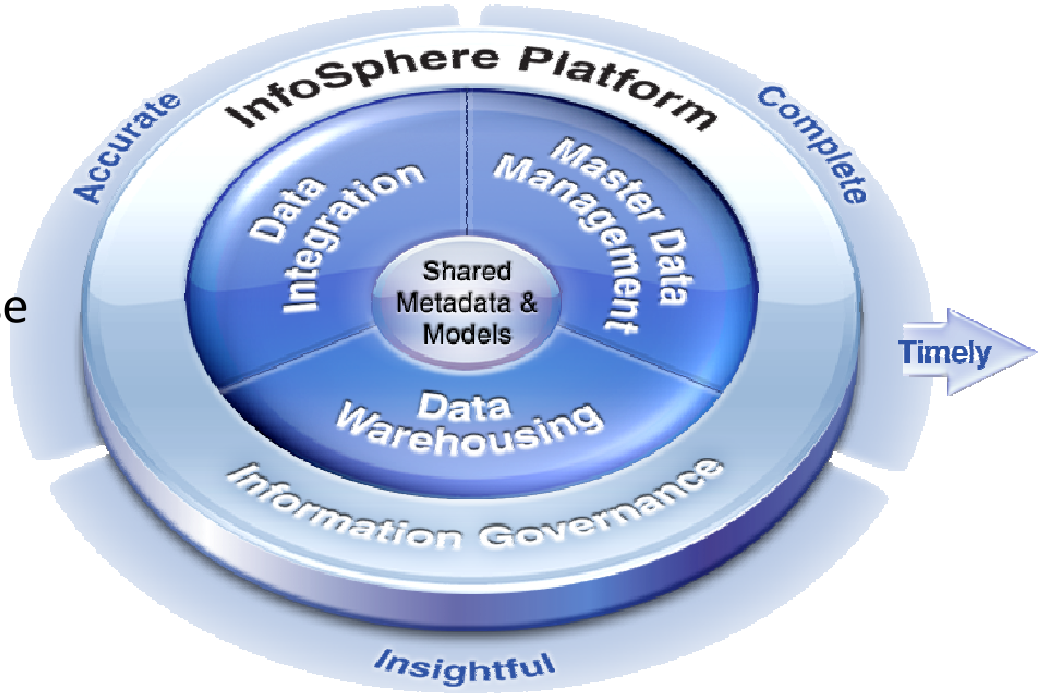
Strategic Initiatives	***IT Projects***
Profitability Analysis	Business Intelligence
Customer Cross-Sell	Master Data Management
Mergers & Acquisitions	Data Consolidation
Multi-Channel Service	Data Synchronization
Solvency II	Regulatory Compliance

Indian Market : Challenges and Opportunities

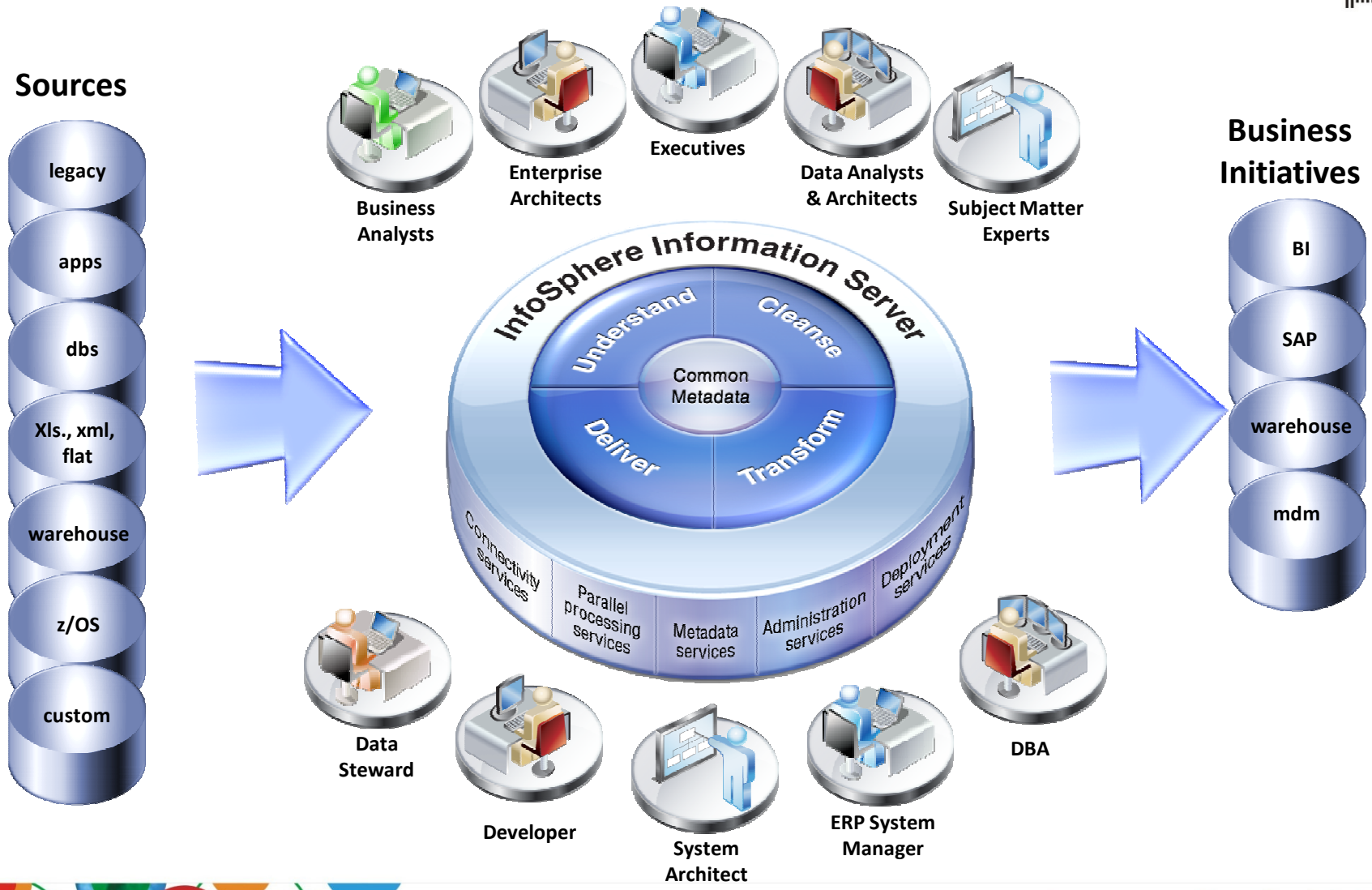
- Data volumes are massive
  - Need for scalable trusted platform
- Immediate reactions to market changes – global and local
  - Need for real time information, in the right form to the right person
- Harnessing the available skills in the local markets
  - Flexible platform to support developer skills
- Higher expectation of Time to Value for investment
  - Integrated collaborative platform with fast results
- World is flat – we all play in the same international market
  - Support for International standard and business standards

IBM InfoSphere is the cornerstone for delivering Trusted Information for Strategic Initiatives

- Manage and deliver trusted information
- Accelerate client value
- Promote collaboration
- Mitigate risk
- Reduce costs
- Scalable from project to enterprise



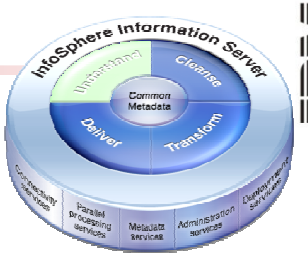
Align business and IT objectives using single platform that creates trusted information for use in key initiatives



Understand Your Information

InfoSphere Foundation Tools allow you to discover your data, design optimal data structures, and govern data over time

- Business challenges
  - Inconsistent data definitions between LOB and IT
  - No trust in the data
  - No insight into where the data lives, who uses it and meaning to the business
  - Missed revenue opportunities
- InfoSphere Foundation Tools facilitates collaboration between business and IT:
  - Discovers your data across systems
  - Designs your trusted information structure
  - Governs your information over time



"Based on information collected during the in-depth customer interviews, Forrester calculated a five-year risk adjusted ROI of 132% for the composite organization with a payback period of 1.23 years"

Forrester TEI Study – 2009

"We are leveraging the capabilities of Business Glossary to share information about the business and technical metadata stored in our research repositories so that users can have the relevant information at their fingertips."

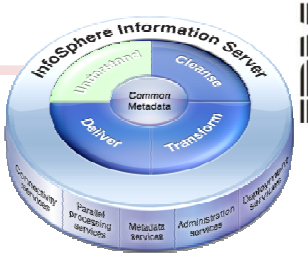
Rob Merriel, Director of Business Development, Melbourne Health

Cleanse Your Information

InfoSphere data quality capabilities ensure you have reliable, accurate information

- Business challenges
  - Unreliable data insight
  - Negative customer satisfaction
  - Inability to identify source of quality issues
  - No method to maintain high quality data
  - High costs due to poor data

- InfoSphere data quality:
  - Identifies source of data quality problems
  - Defines business rules to monitor and maintain quality
  - Removes duplicate data
  - Validates, standardizes and enriches data
  - Enables other key business initiatives, such as BI



IBM's data quality capabilities "continue to be positioned as enterprisewide data quality standards, and are being used in multiple projects in customer organizations."

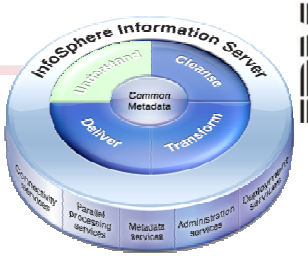
Gartner Data Quality Tools Magic Quadrant, June 2010

"We've standardized our customer name and address data using IBM InfoSphere QualityStage. This allows us to identify, match and merge duplicate records so we have a single view of our customers."

Noel Garry, Executive Manager Irish Life & Permanent

Transform Your Information

InfoSphere DataStage extracts, transforms and loads data between multiple sources and targets



- Business challenges
  - Multiple sources of heterogeneous data
  - Increasing volumes of data
  - Parallel processing/bulk data movements

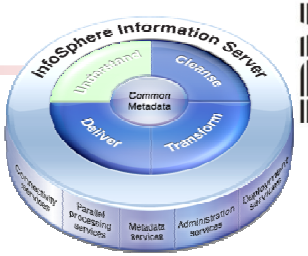
- InfoSphere DataStage:
  - Offers easy-to-use interface design tools
  - Supports massive scalability requirements
  - Transforms data from multiple, disparate information sources
  - Delivers data in batch or real time

IBM continues to demonstrate strong vision in the market for extensive data integration capabilities, while also executing well in increasing the adoption of the various components within existing IBM customers and beyond.

Gartner Data Quality Tools Magic Quadrant, June 2010

Deliver Your Information

InfoSphere data delivery provides timely, reliable movement of heterogeneous data



- Business challenges
  - High cost of capturing changes from data sources
  - Lengthy batch loading processes that hold up business workflows
  - Lack of timely data for trusted decision making
  - Inability to physically move data out of source systems due to security/privacy concerns

- InfoSphere data delivery capabilities:
  - Minimize processing costs on source systems
  - Shortens batch loading processes
  - Provide high performance data movement
  - Support a broad range of heterogeneous databases and data delivery styles

"The InfoSphere technology represented such clearly superior solutions that we elected to proceed with aggressive implementations and then move on to other projects"

Senior Staff Systems Architect

InfoSphere Blueprint Director

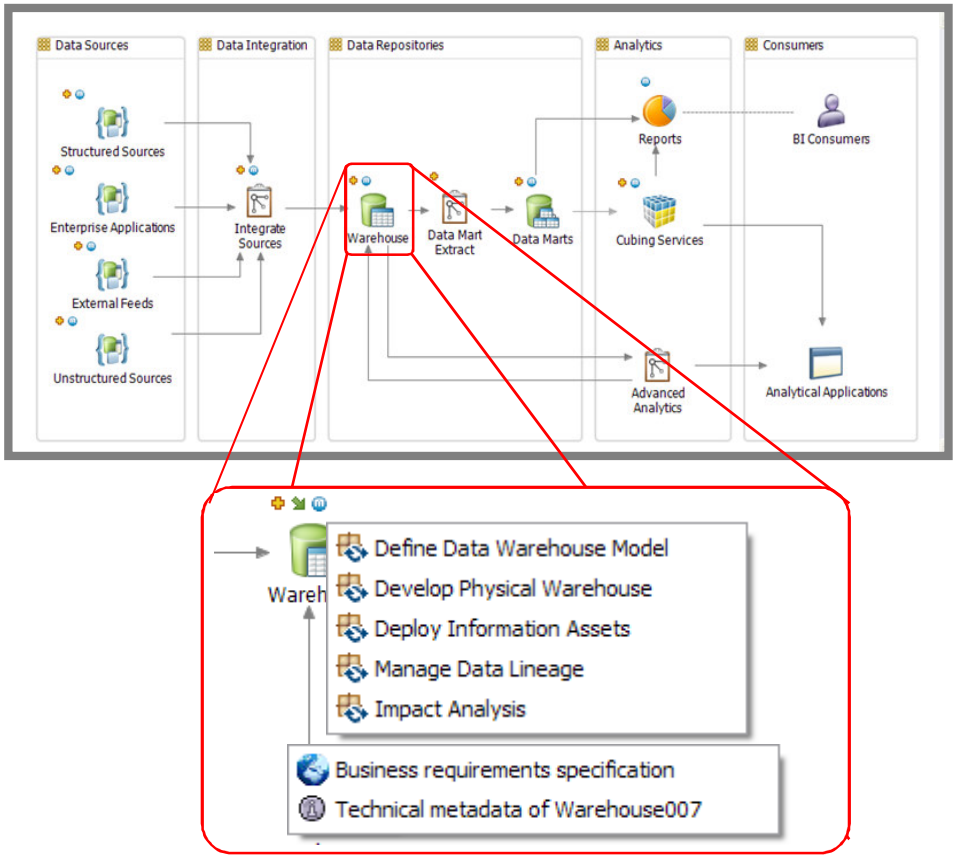
Vision • Execution • Completion

- A unique new paradigm for integration projects, allowing teams to define & manage the end-to-end information flow

- Improve predictability and ensure success of projects by linking blueprints to:
  - Reference architecture
  - Reusable best practices and methodology
  - Business and technical artifacts

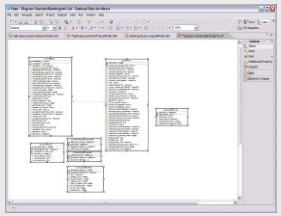
- Provides control and insight of the information roadmap and its evolution through a collaborative project lifecycle:
  - Vision, Execution, Completion

Before you start!



Optimized & Simplified Data Integration with Information Server

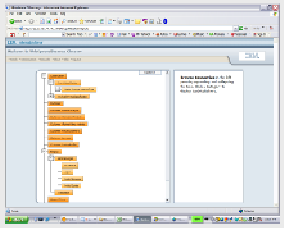
1 ***Establish Platform Import & Enhance Industry Model***



Data Architect

Populates

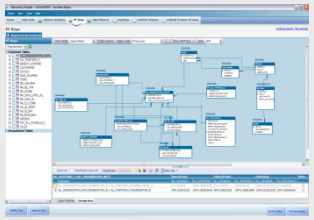
2 ***Define Business Requirement & Glossary***



Business Glossary

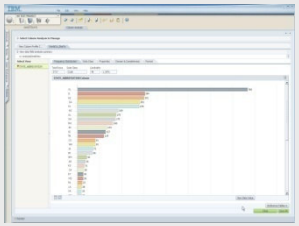
Links

3 ***Understand Data Relationships***



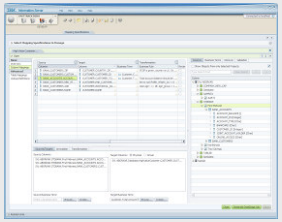
Discovery

4 ***Assess, Monitor, Manage Data Quality Rules***



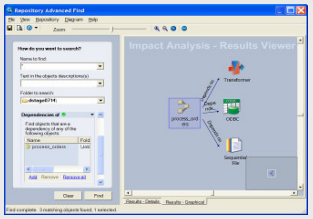
Information Analyzer

5 ***Map Sources to Target Model***



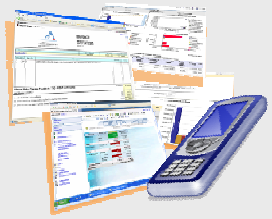
FastTrack

6 ***Generate Logic to Load Warehouse***



DataStage & QualityStage

7 ***Deliver Warehouse and Reports***



Warehouse and BI tools

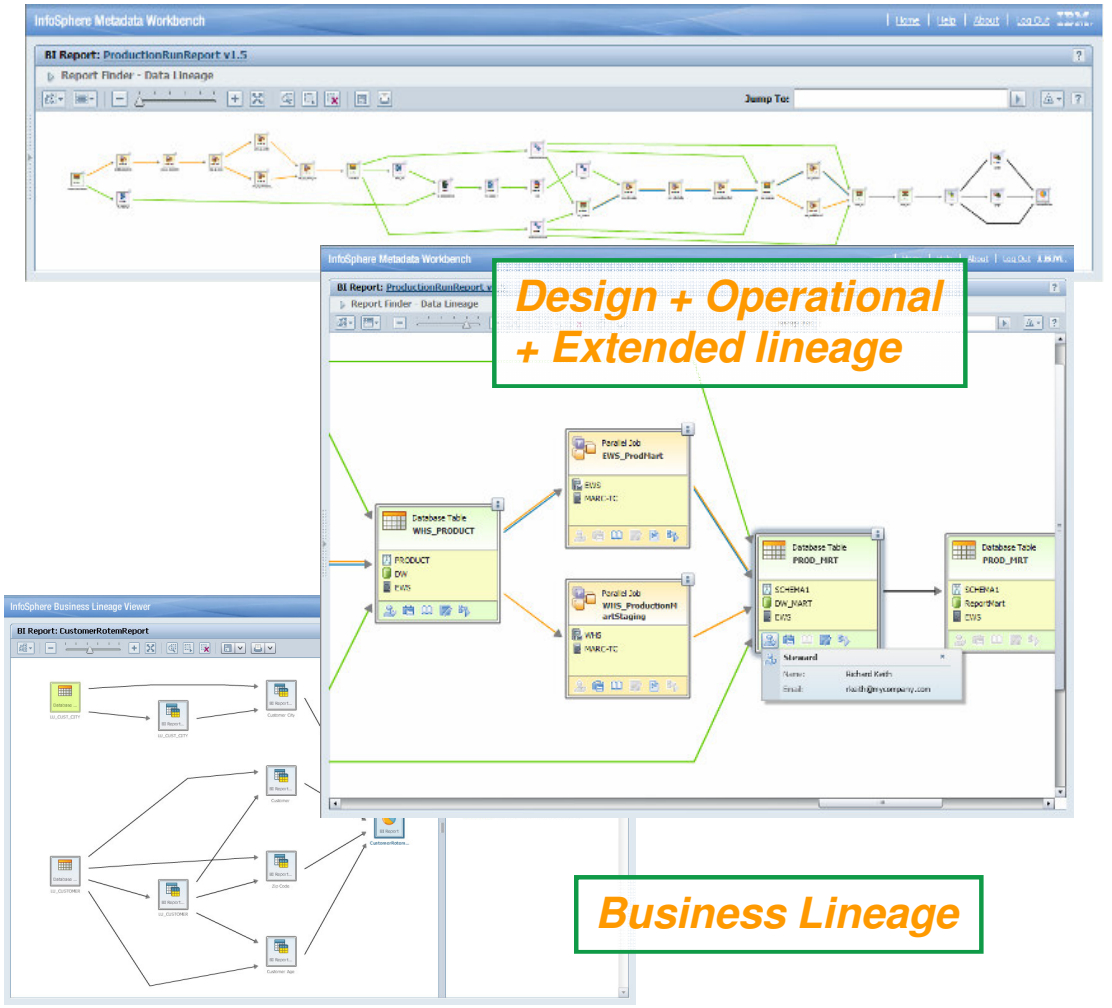
Metadata Server

Simplification & Content: reduces project time, risk and cost!

InfoSphere Metadata Workbench

- Understand the impact of making changes to information environment
- Visualize and trace information flows across enterprise landscape
- Access and report on operational metadata
- Provide audit information for data governance and build trust with LOB users

Trusted end to end view

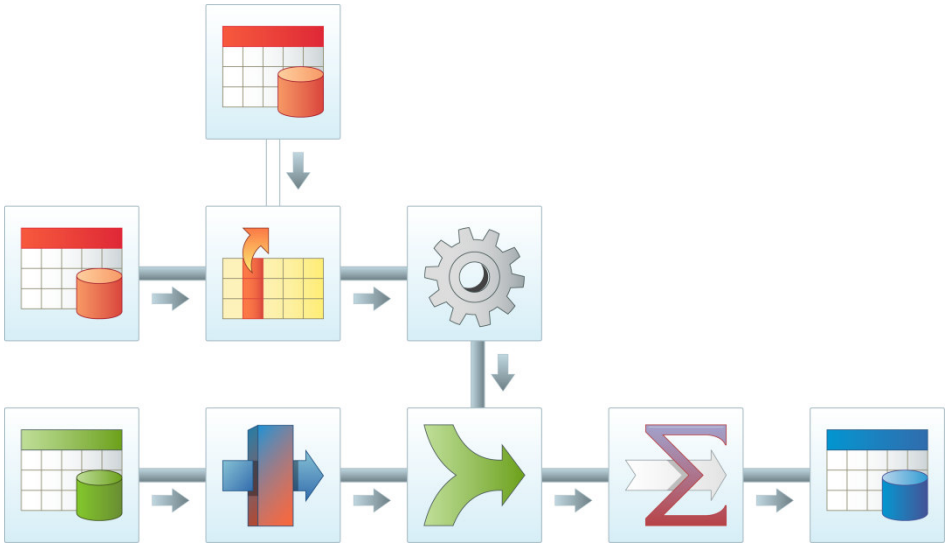


Agenda

- Biggest Challenges for Data Integration
  - Indian Market : Challenges and Opportunities
- IBM Information Server
  - Understanding, Cleansing, Transforming and Delivering trusted data
- Key solutions to specific Data Integration Challenges
  - **Productivity** of and Flexibility for the developer
  - **Performance** and Speed of processing and delivery
  - **Accurate** information delivery

Flexible & Powerful Data Integration with DataStage

- **Design your data integration process**
  - Flexible, graphic design tool generates process
  - Top down design metaphor
  - Transform, cleanse and integrate data,
  - De-normalize data
  - Aggregate data from multiple sources
- **Deploy integration processes**
  - Run and monitor jobs
  - Add and delete projects
  - Set job monitoring limits and user permissions
  - Enable server tracing
  - Test, debug, and deploy job designs
  - Produce job history reports



- **Manage all integration processes**
  - Browse and edit meta data
  - Import and export DataStage design components
  - Report audit trails
  - List contents of DataStage repository

Design Flexibility : Over 100 Components Available

Usual ETL Sources & Targets

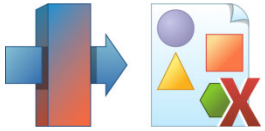
RDBMS, Sequential File, Data Set

Combining Data

Lookup, Joins, Merge, Aggregator

Transform Data

Transformer, Remove Duplicates



Ancillary

Row Generator, Peek, Sort

Extensible Framework

Wrapped

Specify an OS command or script existing Routines, Logic, Apps

Build Op

Wizard/Macro-driven development

Custom

API development

IBM Information Server Connectivity

Information Sources and Targets

Files

Web Content

Legacy Data

PeopleSoft

SAP

ORACLE

SIEBEL

DB2

Teradata

NETEZZA ....

Widest range of sources

- Enterprise Applications
- Mainframe, Mini-computer & Open Systems
- Flat Files, Hierarchical, Relational & Proprietary Databases
- Message Queues, EDI
- XML, and Programming Languages

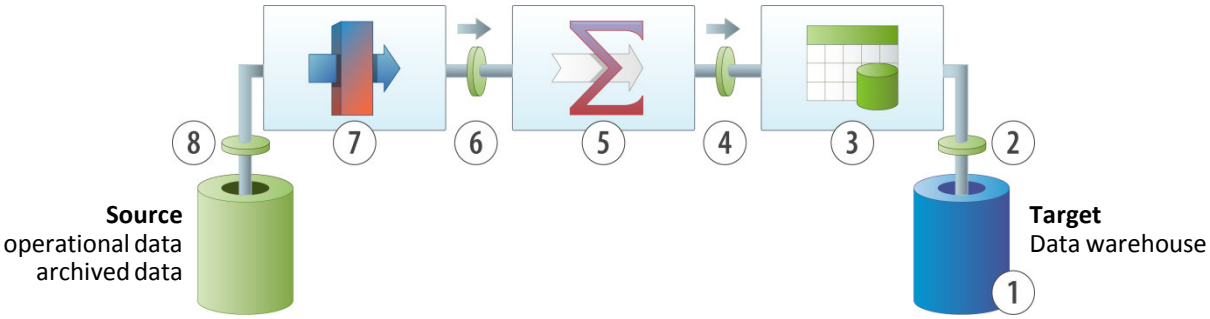
Broadest functionality

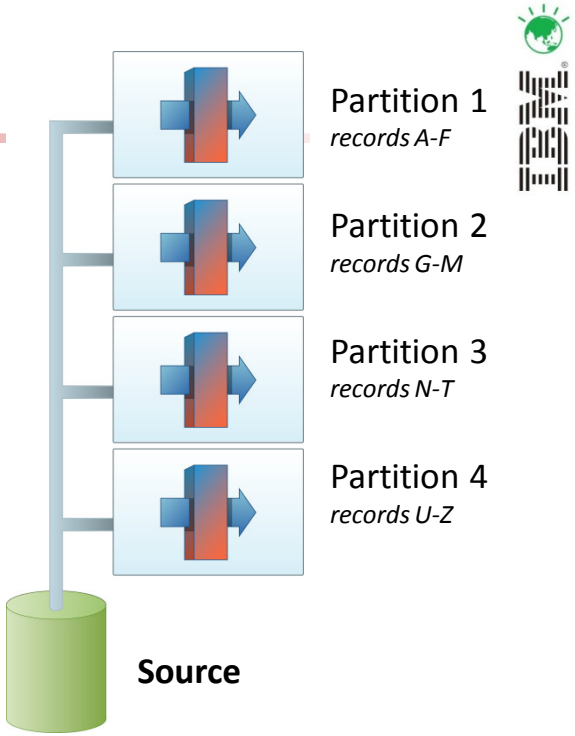
- Native Adapters, and Protocols
- Multi-byte Enabled
- Optimized parallel RDBMS interfaces
- Standards-based
- Batch, Business Objects, and Data Access
- Common query mechanisms
- Integrates source metadata

Extensive Changed Data Capture

- Real-time/push & batch/pull
- Active & Archive Log based
- Trigger & Time/Date Stamp based
- MQ, TCP/IP & FTP data delivery

Performance with the Information Server Engine
Data Pipelining and Data Partitioning



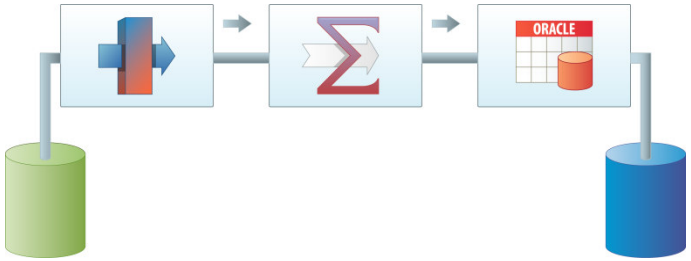


- Eliminate the write to disk and the read from disk between processes
- Start a downstream process while an upstream process is still running.
- This eliminates intermediate staging to disk, which is critical for big data.
- This also keeps the processors busy.
- Still have limits on scalability
- Think of a conveyor belt moving the records from process to process!

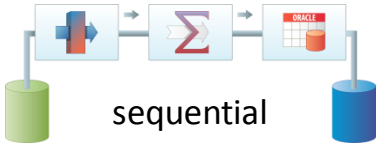
- Break up big data into partitions
- Run one partition on each processor
- 4X times faster on 4 processors; 100X faster on 100 processors
- Partitioning is specified per stage meaning lending to in-flight **dynamic repartitioning** between stages
- Types of partitioning
- DB2, Entire, Hash, Modulus, Random, Range, Round Robin, Same

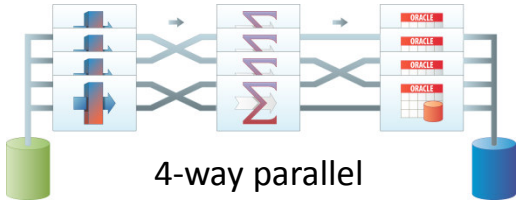
Parallel Runtime Execution

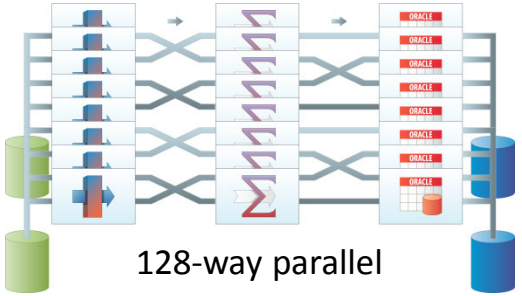
Application Assembly: One Dataflow Graph Created With the DataStage GUI



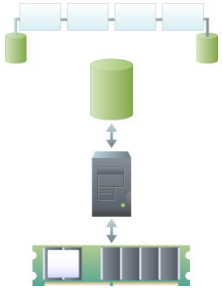
Application Execution: Sequential or Parallel

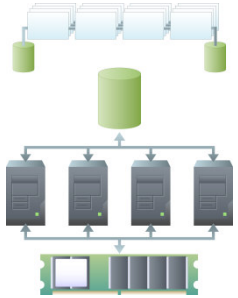


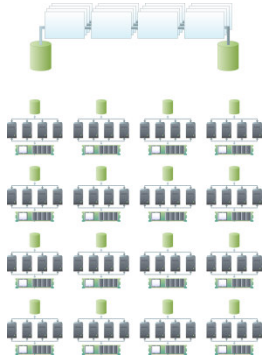




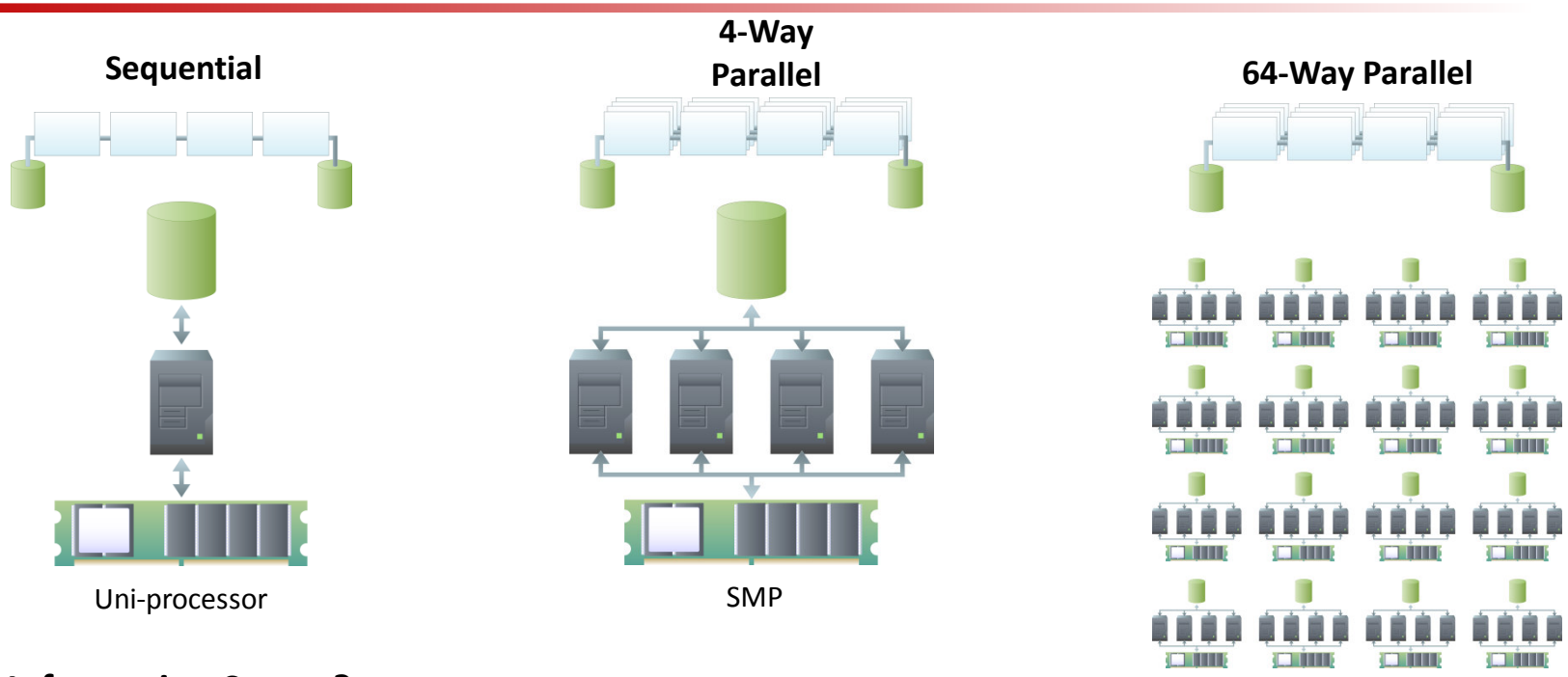
Hardware Platform







Unlimited Scalability



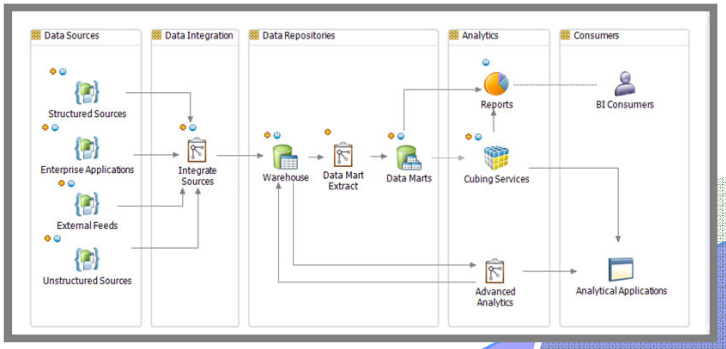
Why Information Server?

- Design sequentially, deploy in parallel
- Proven linear scalability
- Dynamic data partitioning and in-flight repartitioning of data
- Portable across SMP, Clustered, GRID, and MPP platforms
- Parallel RDBMS support, including IBM DB2 UDB, Oracle, & Teradata
- Codeless parallelization
- Incorporate and parallelize existing applications into data integration process

Business Benefits

- Meet business commitments through higher productivity
- Optimal hardware use
- Flexible execution options

Accurate and Trusted Information Delivered on Time



Heterogeneous Sources

CRM

Transform & Cleanse

Master Data

Data Warehouse

Reports & Dashboards

...

Trusted Information

- Blueprint
- Data Quality
- Lineage & impact analysis

QUESTIONS??

Thank You

InformationOnDemandIndia2011

The Premier Conference for Information Management

Manage. Analyze. Govern.

February 2, 2011

Hyatt Regency I Mumbai, India