IMS Data Integration with Hadoop

Karen Durward
InfoSphere Product Manager
17/03/2015

z/OS Structured Data Integration for Big Data

- The Big Data Landscape
  - **Introduction to Hadoop**
    - What, Why, How
  - **The IMS community cares about Hadoop because …**
    - Exponential value at the intersection of Hadoop and Structured Data

- z/OS Data Integration with Hadoop
  - **The Requirements and The Journey**
    - Hadoop on z?
    - z/OS data integration
  - **InfoSphere System z Connector for Hadoop**
    - Fast, Easy, Low Investment z/OS data delivery to HDFS and Hive
  - **InfoSphere Data Replication**
    - Keeping your Big Data Up-to-Date

Hadoop … Created to Go Beyond Traditional Database Technologies

Foundation for Exploratory Analytics

- Hadoop was pioneered by Google and Yahoo! to address issues they were having with then-available database technology:
  - Data volumes could not be cost effectively managed using database technologies
  - Analyzing larger volumes of data can provide better results than sampling smaller amounts
  - Insights needed to be mined from unstructured data types
  - Data was being explored to understand its potential value to the business

Typical Use Cases

Analyze a Variety of Information
Novel analytics on a broad set of mixed information that could not be analyzed before

Analyze Extreme Volumes of Information
Cost-efficiently process and analyze petabytes of information

Discovery and Experimentation
Quick and easy sandbox to explore data and determine its value

3

What is Hadoop?

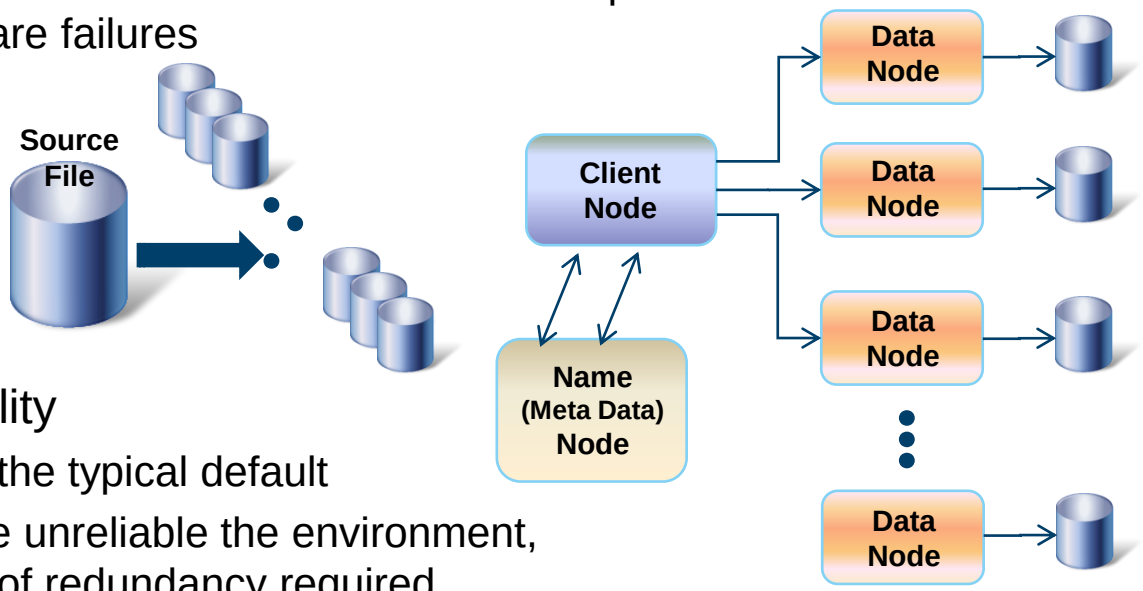
Divide and Conquer!

- Rapidly evolving open source software framework for creating and using hardware clusters to process vast amounts of data!
- 2.# version framework consists of:
  - **Common Core**: the basic modules (libraries and utilities) on which all components are built
  - **Hadoop Distributed File System (HDFS)**: manage data stored on multiple machines for very high aggregate bandwidth across the "cluster" of machines
  - **MapReduce**: programming model to support the high data volume processing of data in the cluster
  - **"Yet Another Resource Negotiator" (YARN)**: platform to manage the cluster's compute resources, de-coupling Hadoop workload and resource management

Typical Hadoop Data Flow

1. Load Data into an HDFS cluster

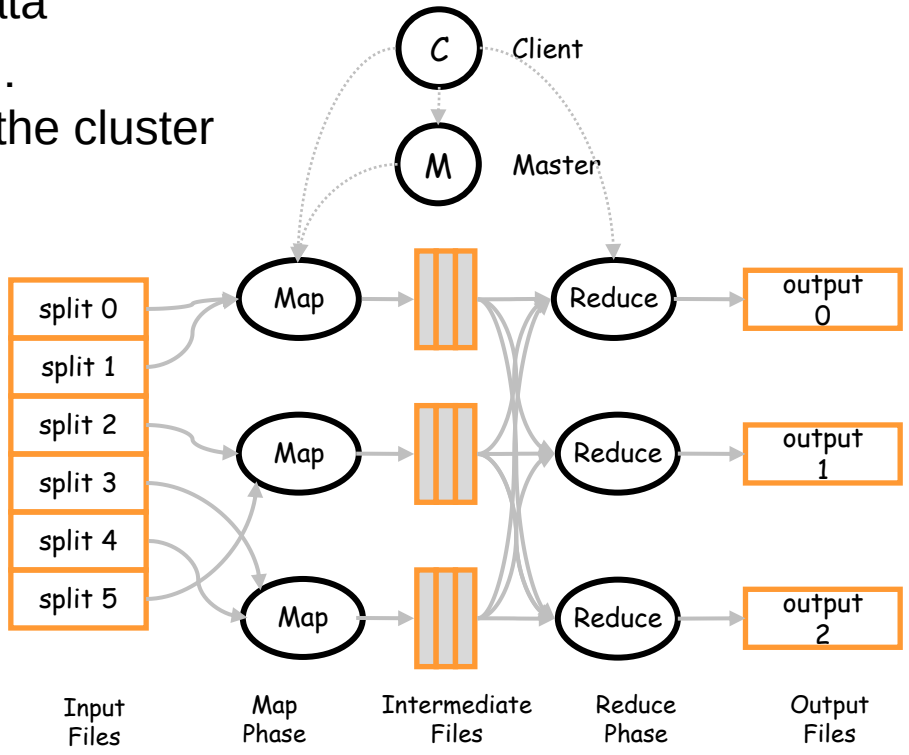
- Optimize for parallelism and reliability
  - Break the source into large blocks (typically 64MB) so that each block can be written (and read) independently
  - Each block is written to a node … the more nodes the more dispersion of the data for future parallel processing
  - Redundant copies of each block are maintained on separate nodes to protect against hardware failures



- Redundancy for reliability
  - Configurable, three is the typical default
  - Simplistically, the more unreliable the environment, the higher the degree of redundancy required

Typical Hadoop Data Flow

1. Load Data into an HDFS cluster
2. **Analyze the data in the cluster using a MapReduce "program"**
   - Map: Drive analysis of the data
   - Reduce: Construct a result …
     Writing it back into the cluster
3. Use the result!



Hadoop Cluster Configuration Implications

- Redundancy drives Network Traffic
  - With three-way redundancy, each terabyte of data results in three terabytes of network traffic
- Parallelism drives Performance
  - **Scale OUT** (more nodes) and/or files OUT (more blocks)
    - Spread blocks of data across more nodes so more blocks can be read in parallel
    - Can spread a file to more nodes if you have enough nodes
    - Network activity spread across many nodes/files
  - **Scale UP** (more CPUs and/or memory per node rather than more nodes)
    - Increases the density of each node
    - More network activity concentrated on each node

The Landscape is Rapidly Evolving

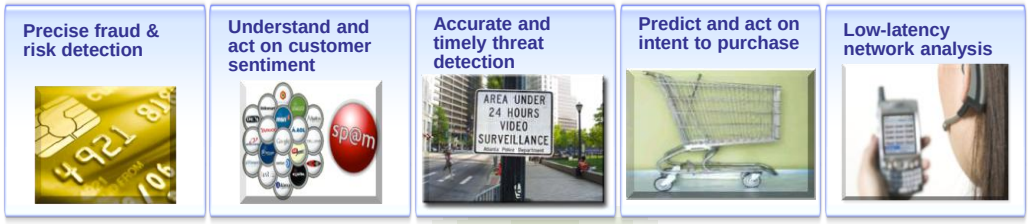
More Apache Frameworks and Products You'll Hear About

- **HiveTM** **Apache data warehouse framework accessible using HiveQL**
- **Spark** **In-memory framework providing an alternative to MapReduce**
- HBaseTM Apache Hadoop database
- Pig High level platform for creating long and deep Hadoop source programs
- Zookeeper Infrastructure & services enabling synchronization across large clusters
- Flume Collect and integrate data into Hadoop coordinating Web/app services, …
- Oozie Workflow processing connecting multiple types of Hadoop source jobs

z/OS Structured Data Integration for Big Data

- The Big Data Landscape
  - **Introduction to Hadoop**
    - What, Why, How
  - **The IMS community cares about Hadoop because …**
    - Exponential value at the intersection of Hadoop and Structured Data
- Starting on the z/OS data integration with Hadoop journey
  - **InfoSphere System z Connector for Hadoop**
    - Fast, Easy, Low Investment z/OS data delivery to HDFS and Hive
- Keeping Big Data Current
  - **InfoSphere Data Replication**
    - Continuous incremental updates

Imagine the possibility of leveraging all of your information assets

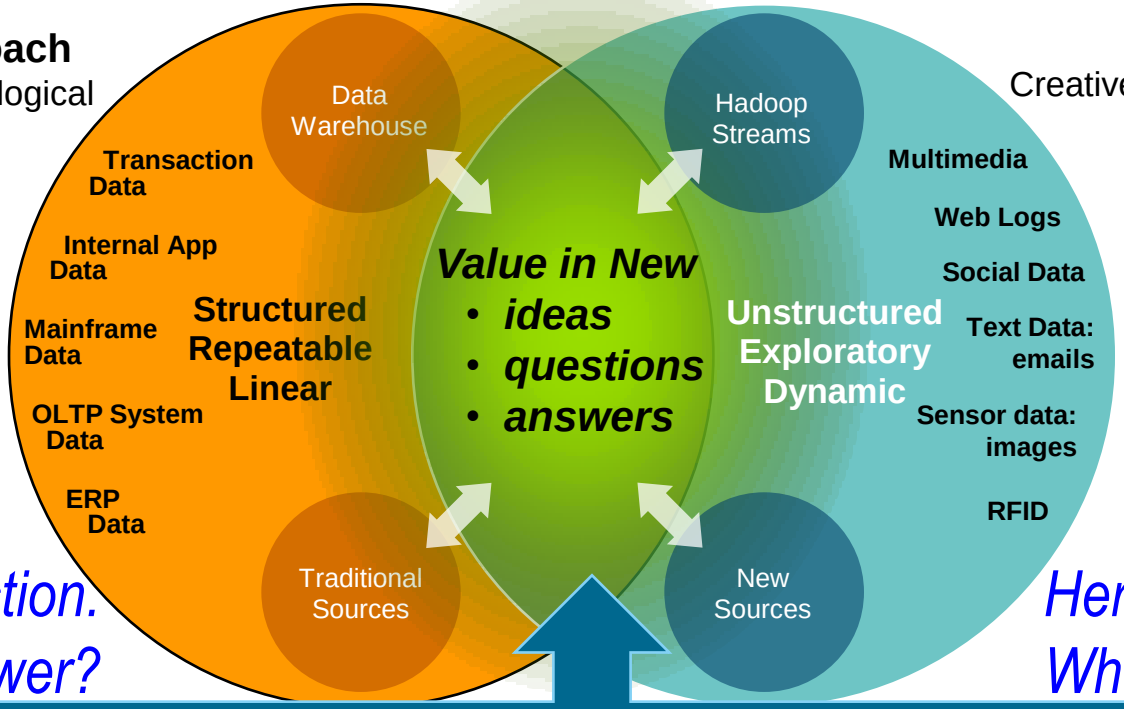


Traditional Approach
Structured, analytical, logical

Data: Rich, historical, private, structured.
Customers, history, transactions

New Approach
Creative, holistic thought, intuition

Data: Intimate, unstructured.
Social, mobile, GPS, web, photos, video, email, logs



Here's a question.
What's the answer?

Here's some data.
What does it tell me?

Transformational **benefit comes from integrating "new" data and methods with traditional ones!**

This is Driving the Shifting Sands of Enterprise IT

New ways of thinking for transformative economics

Traditional Approach

- Vertical Infrastructure
- Design schemas in advance
- What data should I keep?
- What reports do I need?
- ETL, down-sample, aggregate
- On-premise

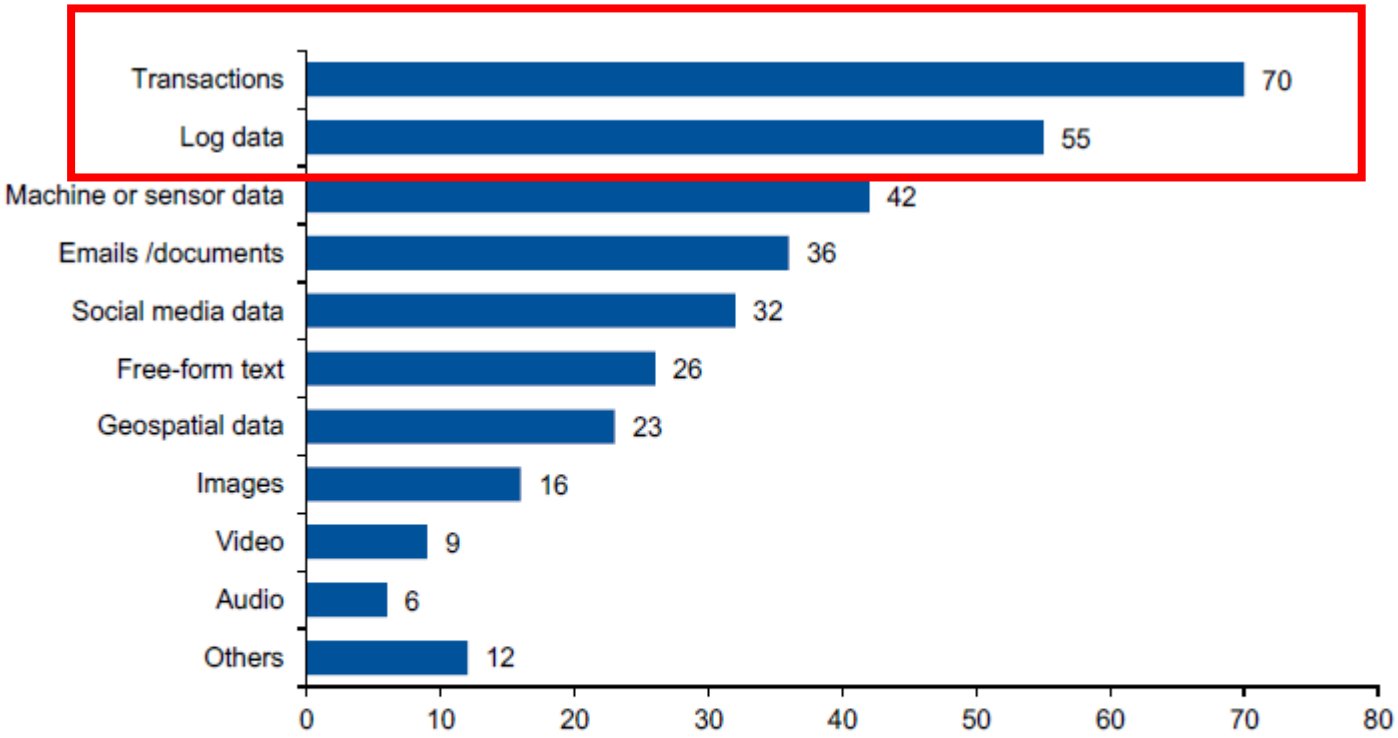
New Approach

- Distributed data grids
- Evolve schemas on-the-fly
- Keep everything, just in case
- Test theories, model on-the-fly
- Knowledge from raw-data
- On-premise, cloud, hybrid

11

Transaction & Log Data Dominate Big Data Deployments

Very High Proportion of this Data Resides on System z



N=465 (multiple responses allowed)

Source: Gartner (September, 2013)

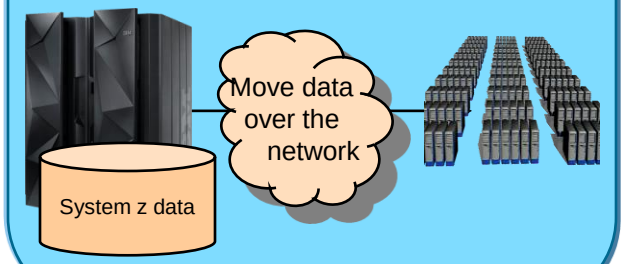
Gartner research note "Survey Analysis - Big Data Adoption in 2013 Shows Substance Behind the Hype" Sept 12 2013
Analyst(s): Lisa Kart, Nick Heudecker, Frank Buytendijk

Hadoop and System z

The Requirements and The Journey

Hadoop Topology Choices for System z Data

Processing done outside z (Extract and move data)



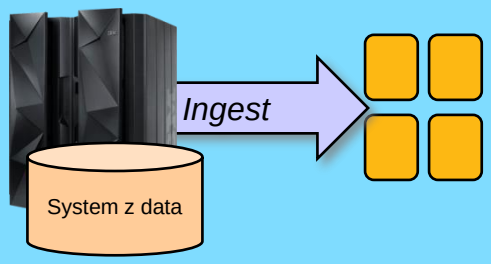
Petabytes possible

Additional infrastructure.

Challenges with:
- scale
- governance
- ingestion.

Data is outside System z control.

Processing done on z (Hadoop cluster on zLinux)



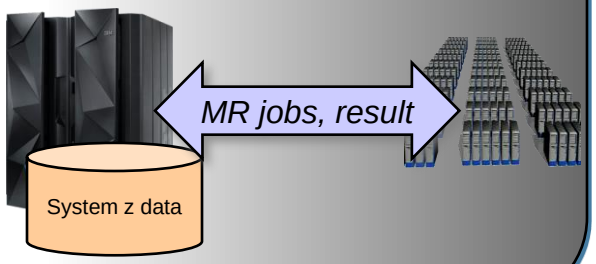
Gigabytes to Terabytes reasonable

Rapidly provision new node(s)

Near linear scale.

System z is the control point.

Extract and consume (z controls the process)



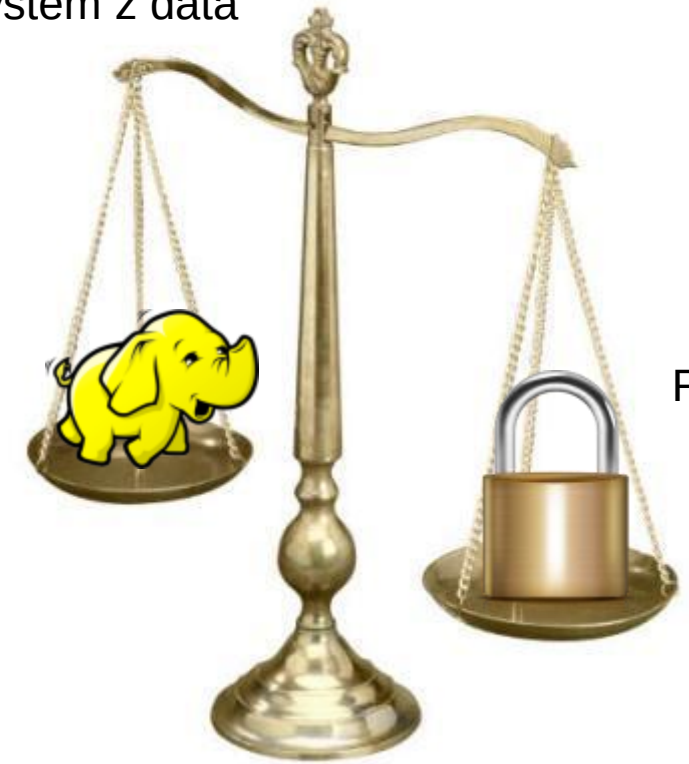
External data NOT routed through System z.

System z governance for the result set.

System z is the control point.

The most likely driver for Hadoop on System z

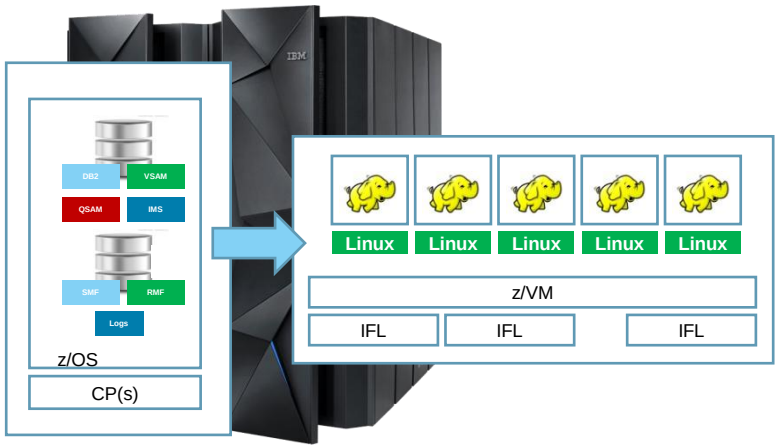
Value of exploratory analytic models applied to System z data



Risk associated with numerous copies of sensitive data dispersed across commodity servers with broad access

Hadoop on System z … What makes sense when?

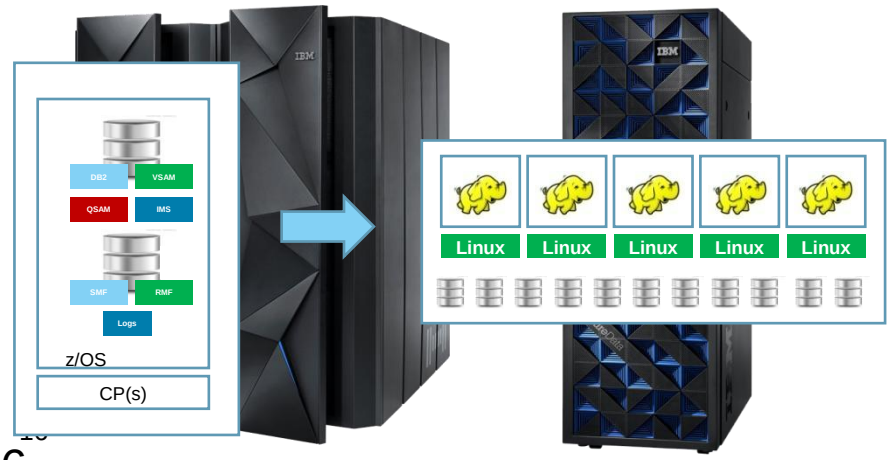
Case 1: Hadoop on the Mainframe



- *Data originates mostly on the mainframe (Log files, database extracts) and data security is important*
- *Z governance & security models needed*
- *Network volume or security concerns*
- *Moderate volumes – 100 GB to 10s of TBs*
- *Hadoop value from rich exploratory analytics*

  (Hybrid Transaction-Analytic appliances for traditional analytics)

Case 2: Hadoop off the Mainframe



- *Most data originates off of the mainframe*
- *Security less of a concern since data is not "trusted" anyway*
- *Very large data sets – 100s of TB to PBs*
- *Hadoop is valued for ability to economically manage large datasets*
- *Desire to leverage lowest cost processing and potentially cloud elasticity*

IBM InfoSphere BigInsights Uniquely Offers …

Multiple technology options & deployment models

Intel Servers

IBM Power

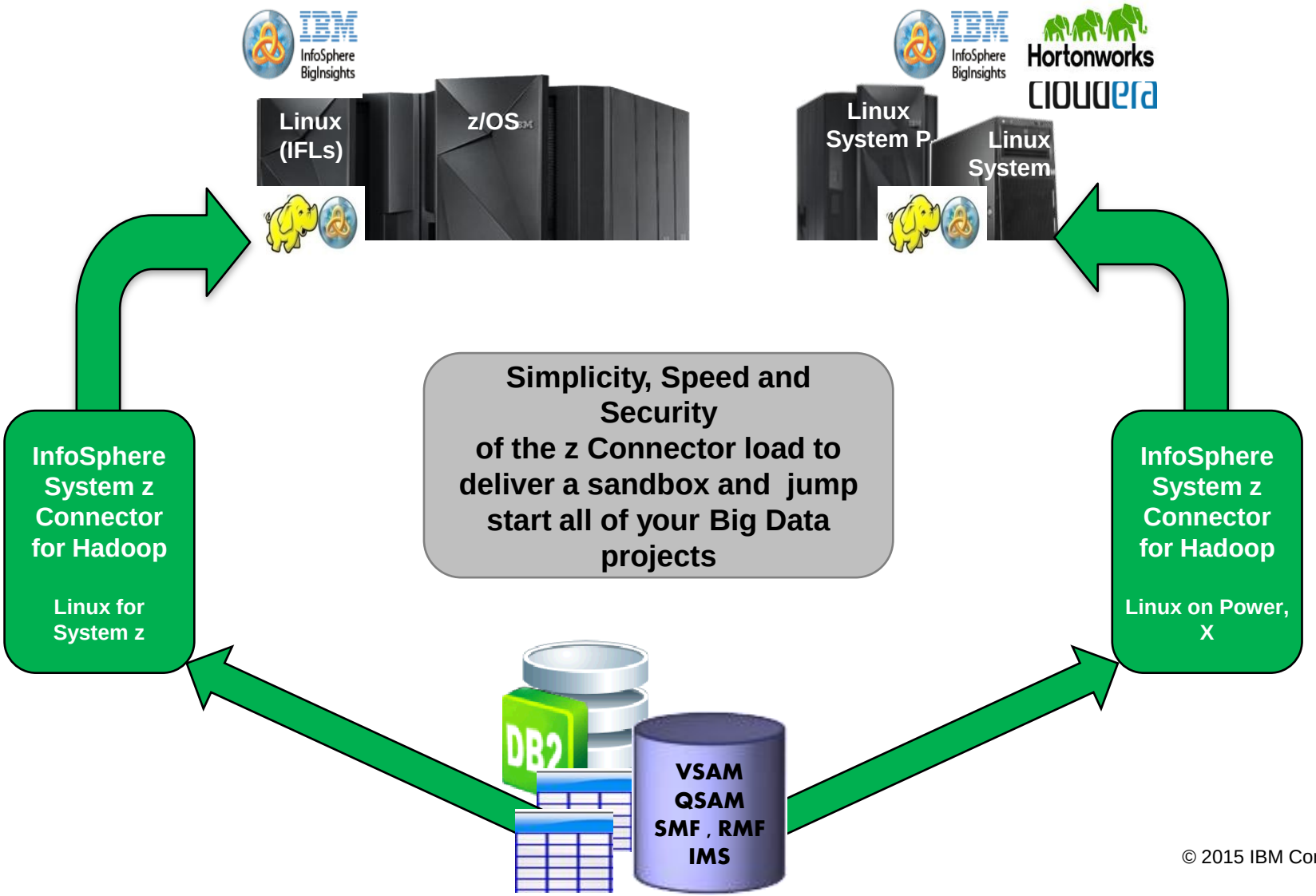
IBM System z

On Cloud

IBM SOFTLAYER an IBM Company

IBM's InfoSphere z/OS Data Integration with Hadoop

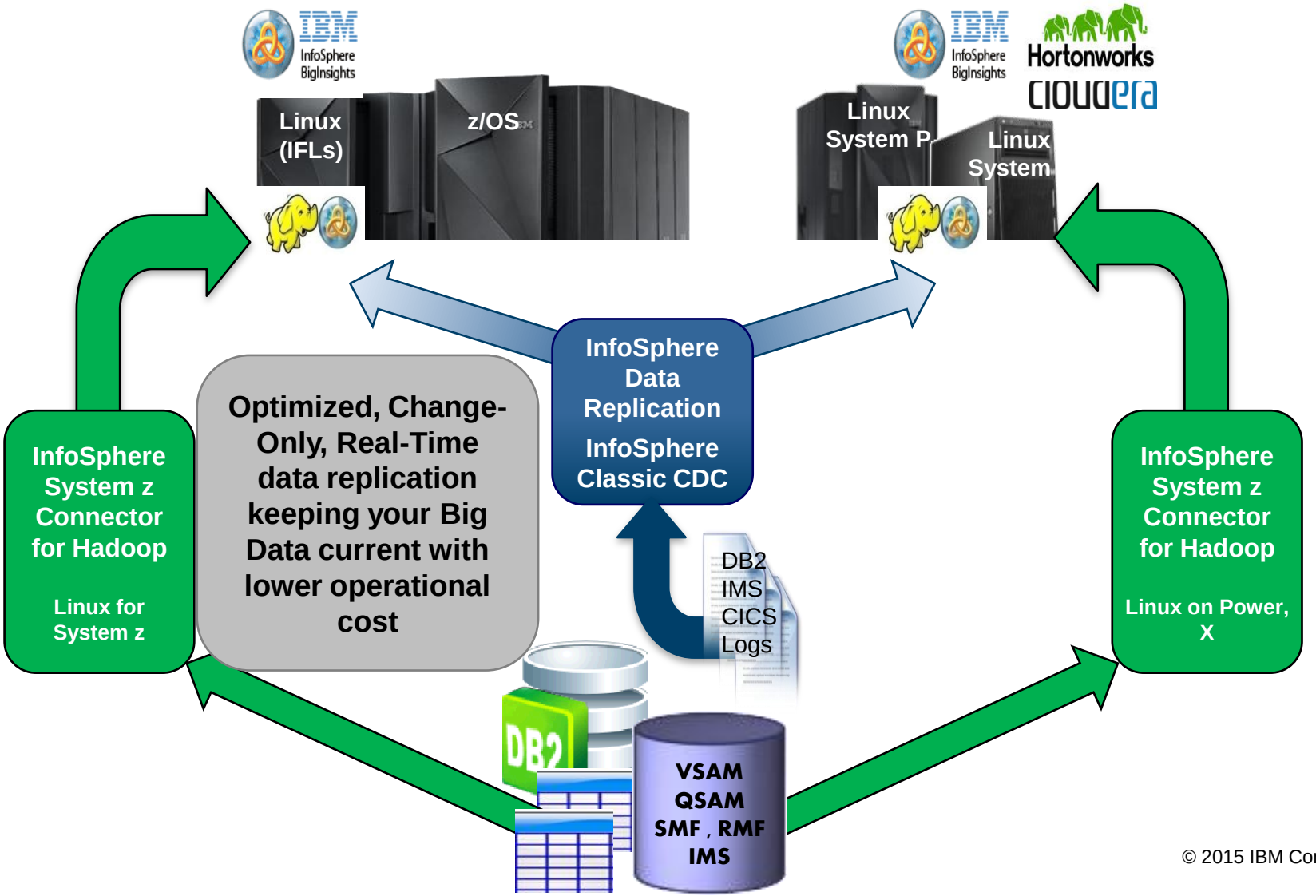
From the Sandbox to the Enterprise



18

IBM's InfoSphere z/OS Data Integration with Hadoop

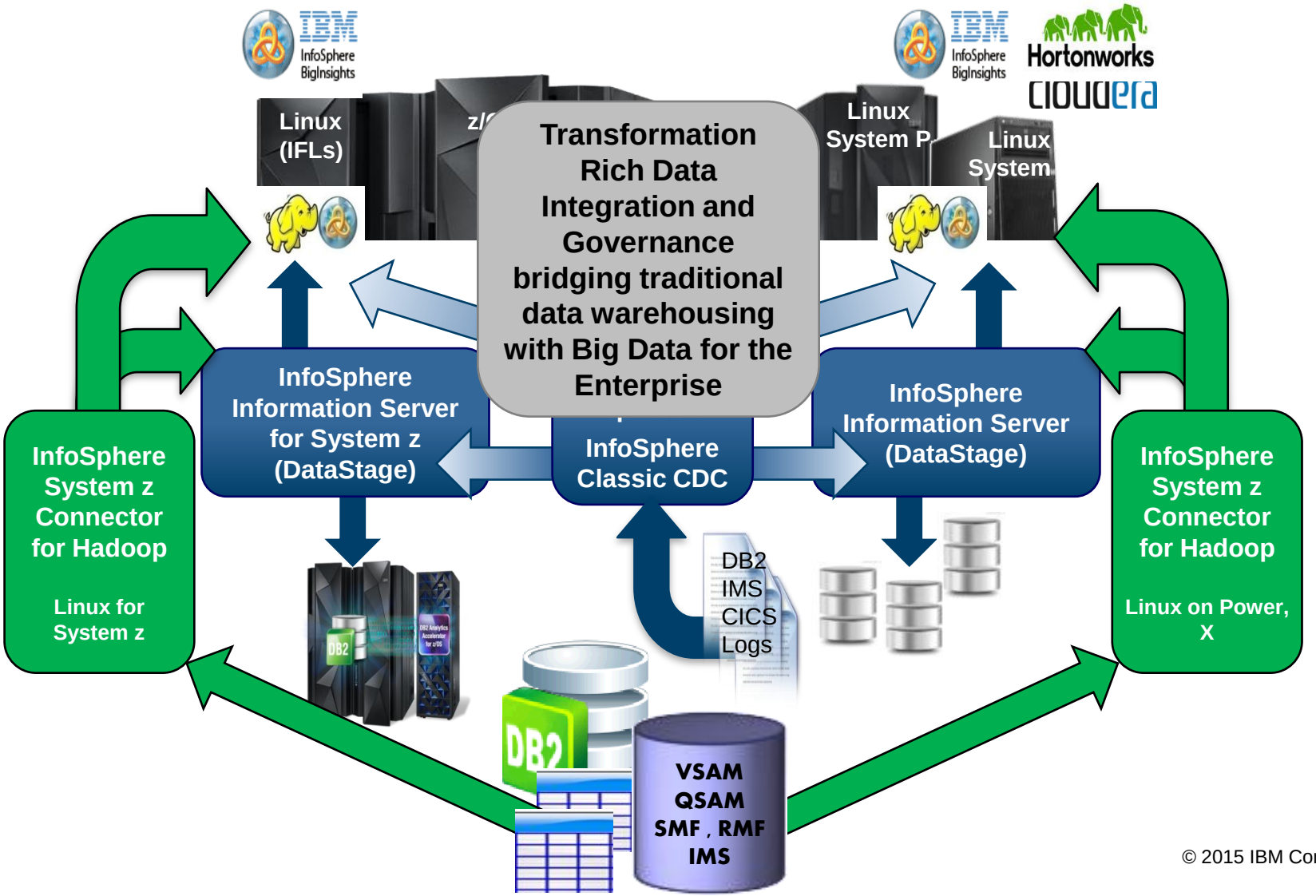
From the Sandbox to the Enterprise



19

IBM's InfoSphere z/OS Data Integration with Hadoop

From the Sandbox to the Enterprise



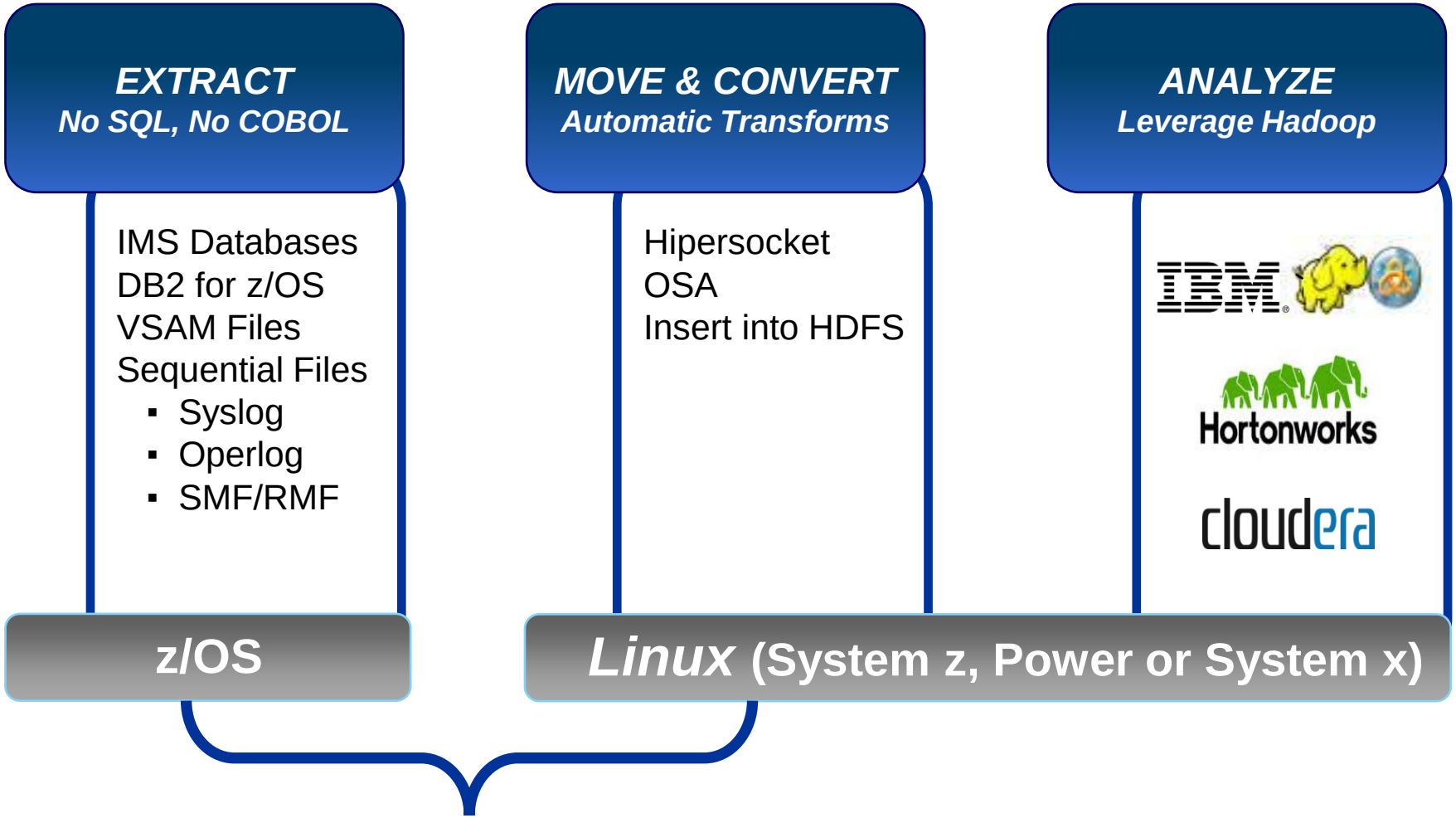
20

z/OS Data Integration with Hadoop

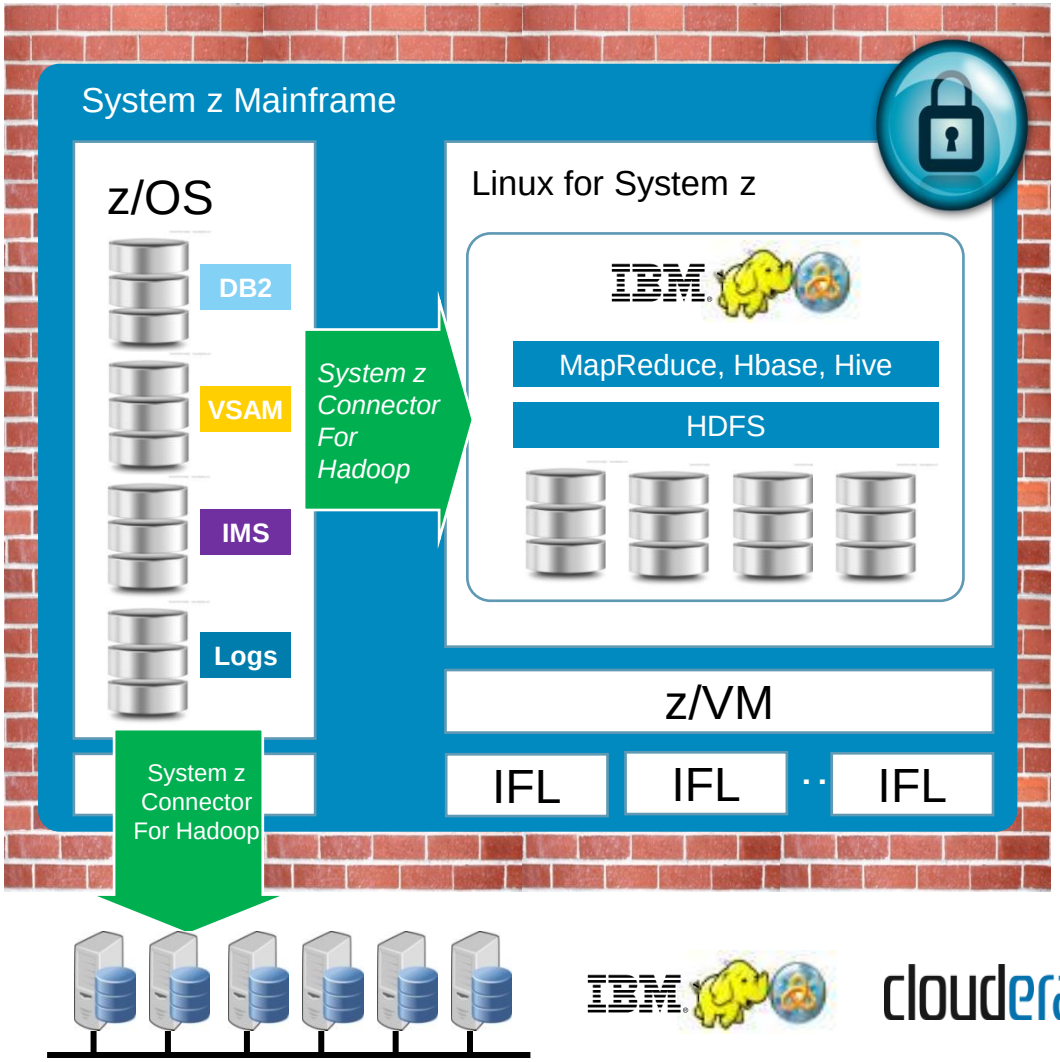
The InfoSphere System z Connector for Hadoop

IBM InfoSphere System z Connector for Hadoop

Setup in Hours, Generated Basic Transforms, Interactive or Scheduled



IBM InfoSphere System z Connector for Hadoop



- Leverage z data with Hadoop on your platform of choice
  - ✓IBM System z for security
  - ✓Power Systems
  - ✓Intel Servers
- Point and click or batch self-service data access
- Lower cost processing & storage

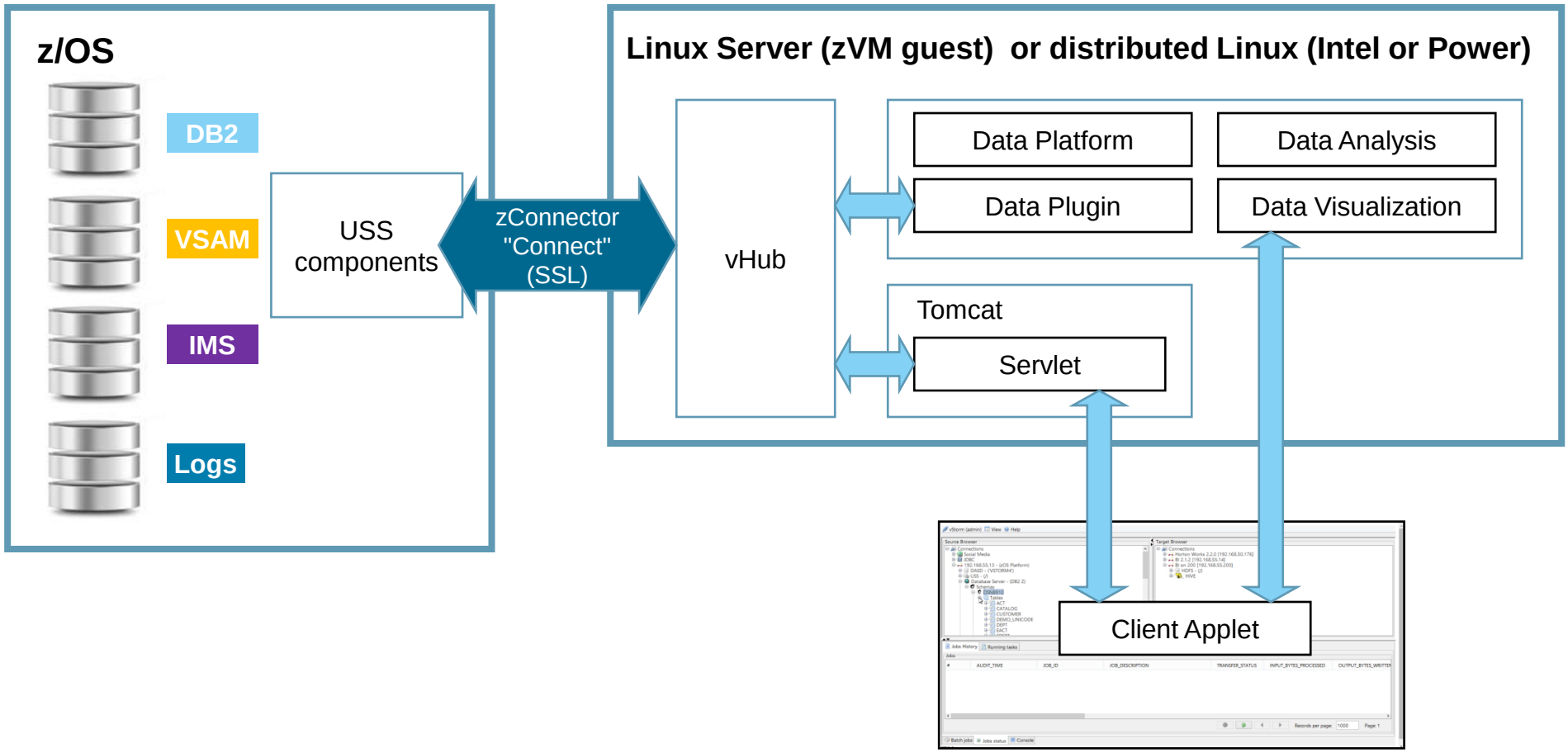
IBM InfoSphere System z Connector for Hadoop

Key Product Features

- Supports multiple Hadoop distributions – on and off the mainframe
- Multiple source formats: IMS, DB2, VSAM/QSAM, Log files
- HiperSockets and 10 GbE Data Transfer
- Drag-and-drop interface – no programming required
- Multiple destinations: Hadoop, Linux File Systems, Streaming endpoints
- Define multiple data transfer profiles
- Streaming interface – filter and translate data & columns on the fly, on the target
- Secure pipe – RACF integration
- Preserves metadata, landing mainframe data in Hive tables

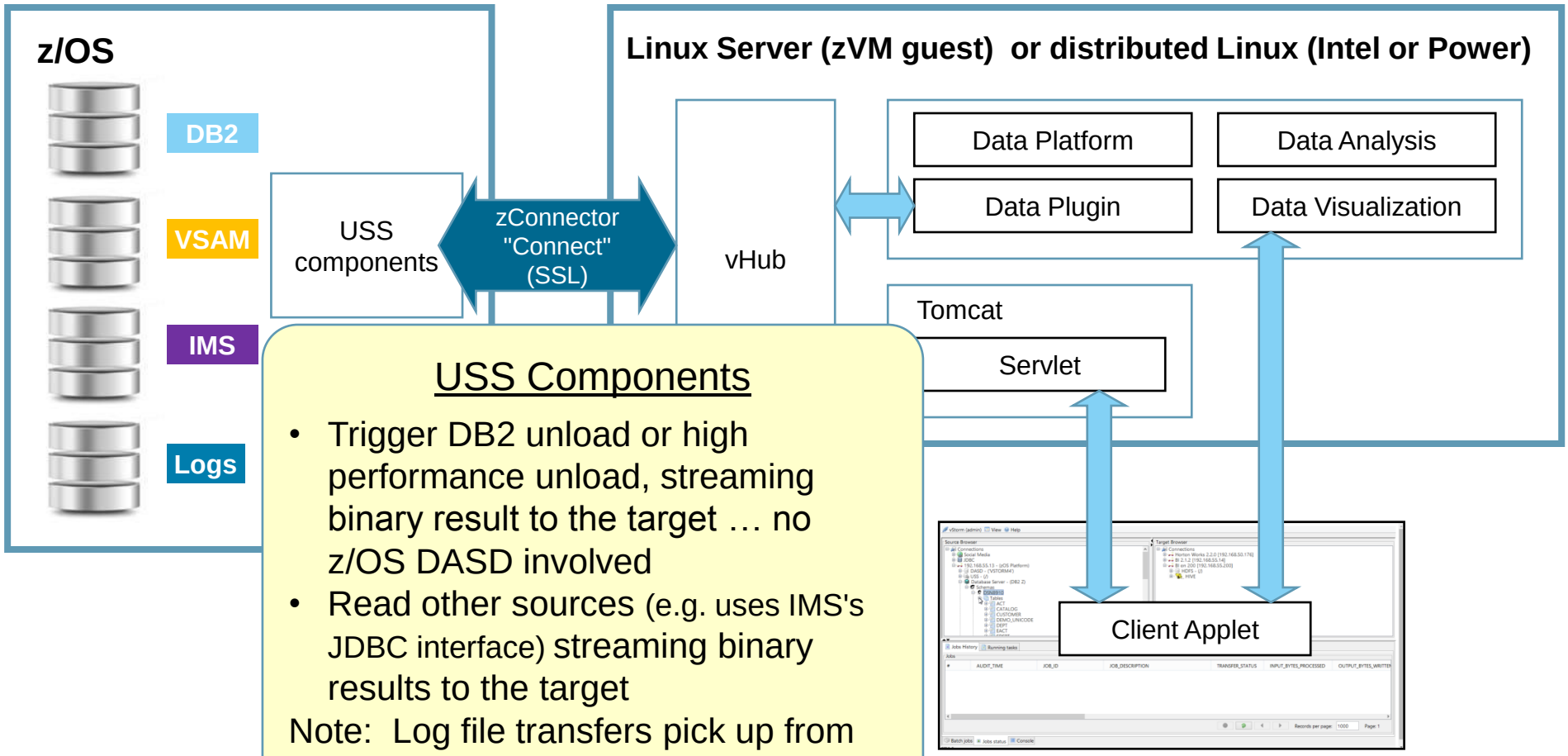
IBM InfoSphere System z Connector for Hadoop

Architectural view



IBM InfoSphere System z Connector for Hadoop

Architectural view



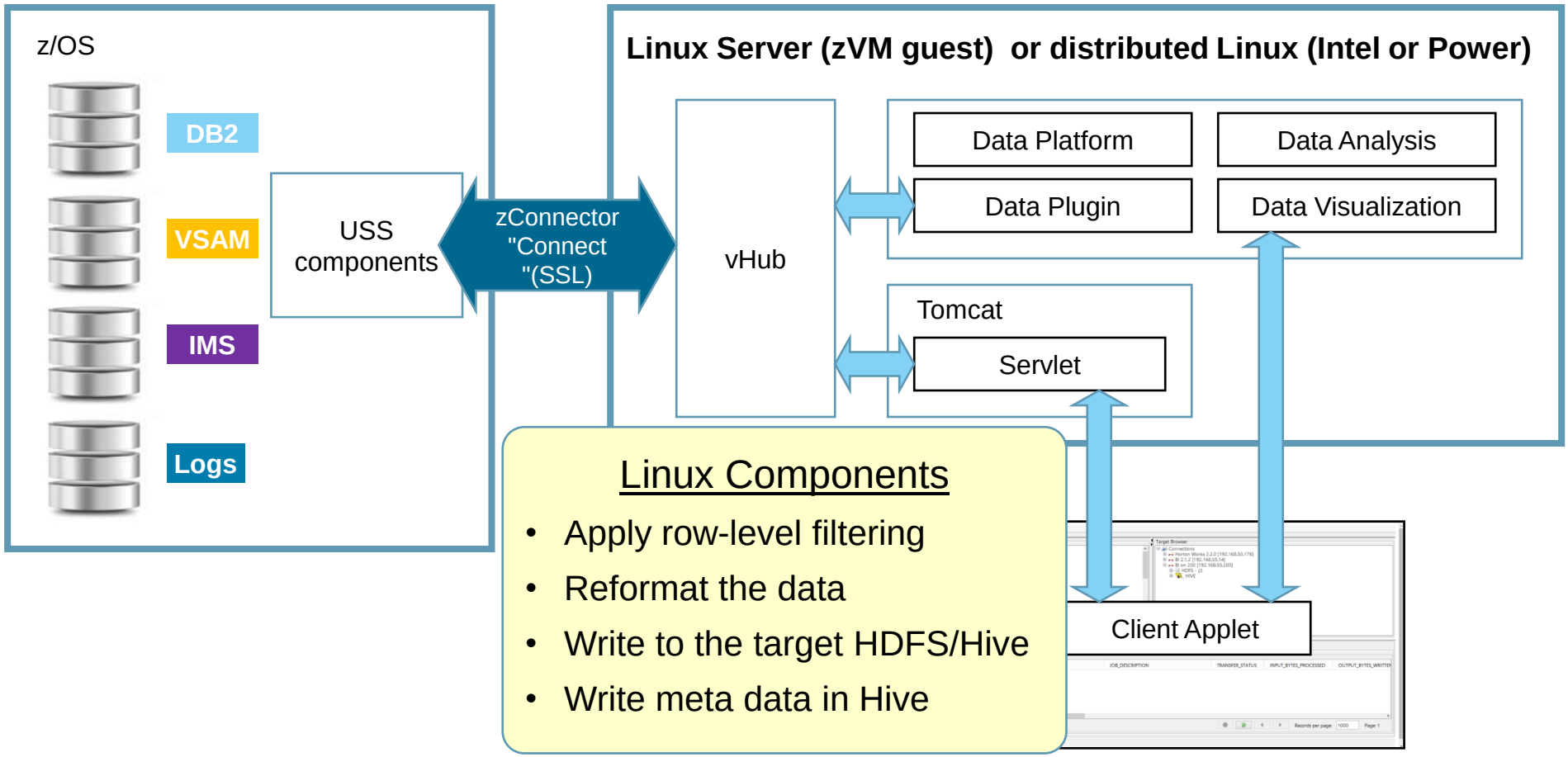
USS Components

- Trigger DB2 unload or high performance unload, streaming binary result to the target … no z/OS DASD involved
- Read other sources (e.g. uses IMS's JDBC interface) streaming binary results to the target

Note: Log file transfers pick up from the date/time stamp of the previous transfer automating target updates

IBM InfoSphere System z Connector for Hadoop

Architectural view



InfoSphere System z Connector for Hadoop Summary

- **A secure pipe for data**
  - RACF integration – standard credentials
  - Data streamed over secure channel using hardware crypto
- **A Rapid Deployment**
  - Integrating z/OS data in a few hours

- **Easy to use ingestion engine**
  - Light-weight; no programming required
  - Native data collectors accessed via a graphical user interface
  - Wide variety of data sources supported
  - Conversions handled automatically
  - Streaming technology does not load z/OS engines nor require DASD for staging

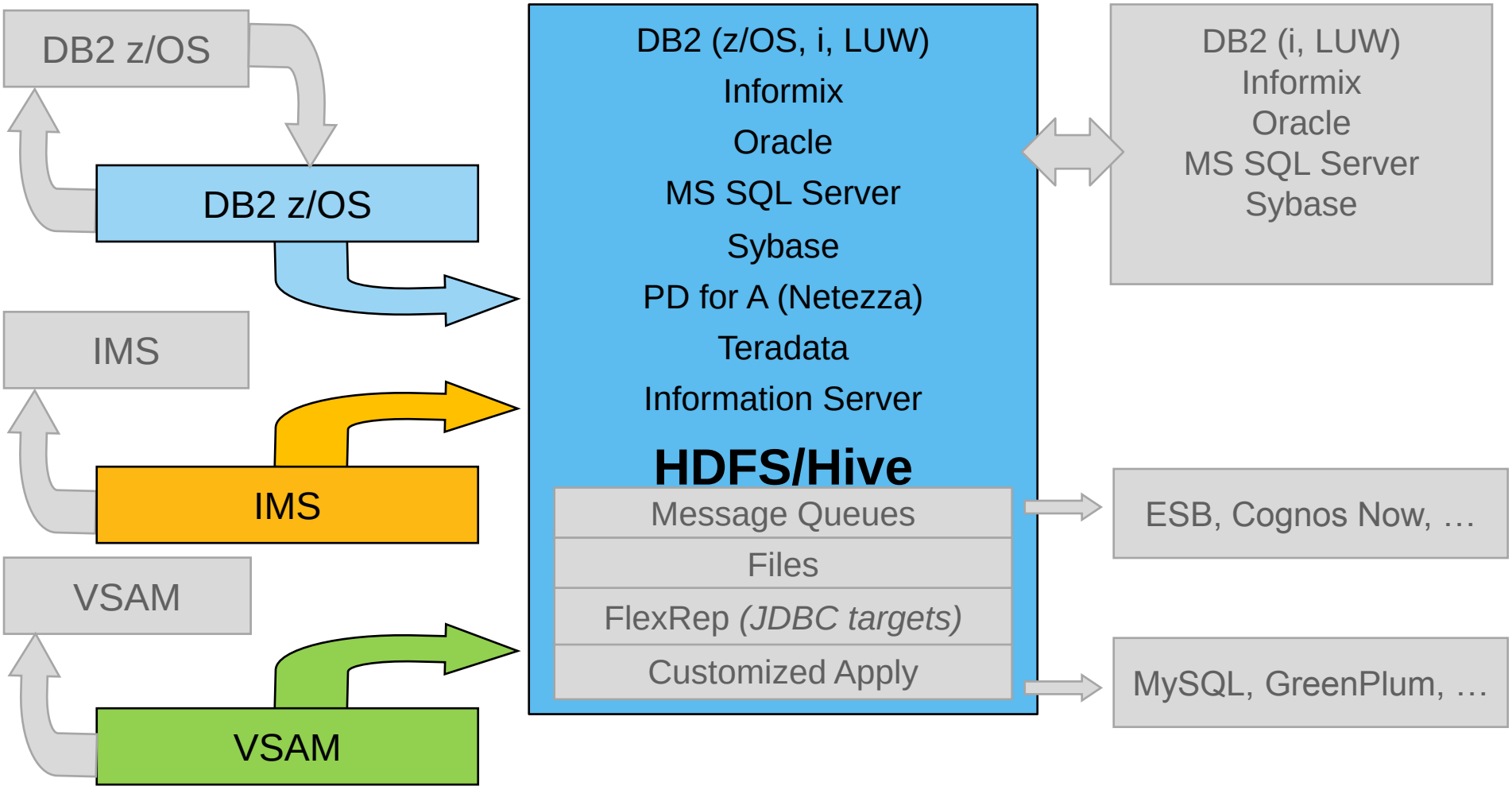
Best Use Cases

- HDFS/Hive sandbox for initial deployments … explore your data
  - ✓ ***Easy to setup … Hours, not Days!***
- Operational Analytics using z/OS log data (SMF, RMF, …)
  - ✓ ***Exploring operational data using Hadoop on day one!***
- Moderate volumes (100s of GBs to 10s of TBs) of transactional data
  - ✓ ***Source of the data is z/OS***
  - ✓ ***Security may be a primary concern***

z/OS Data Integration with Hadoop

Keeping your Hadoop Data Current

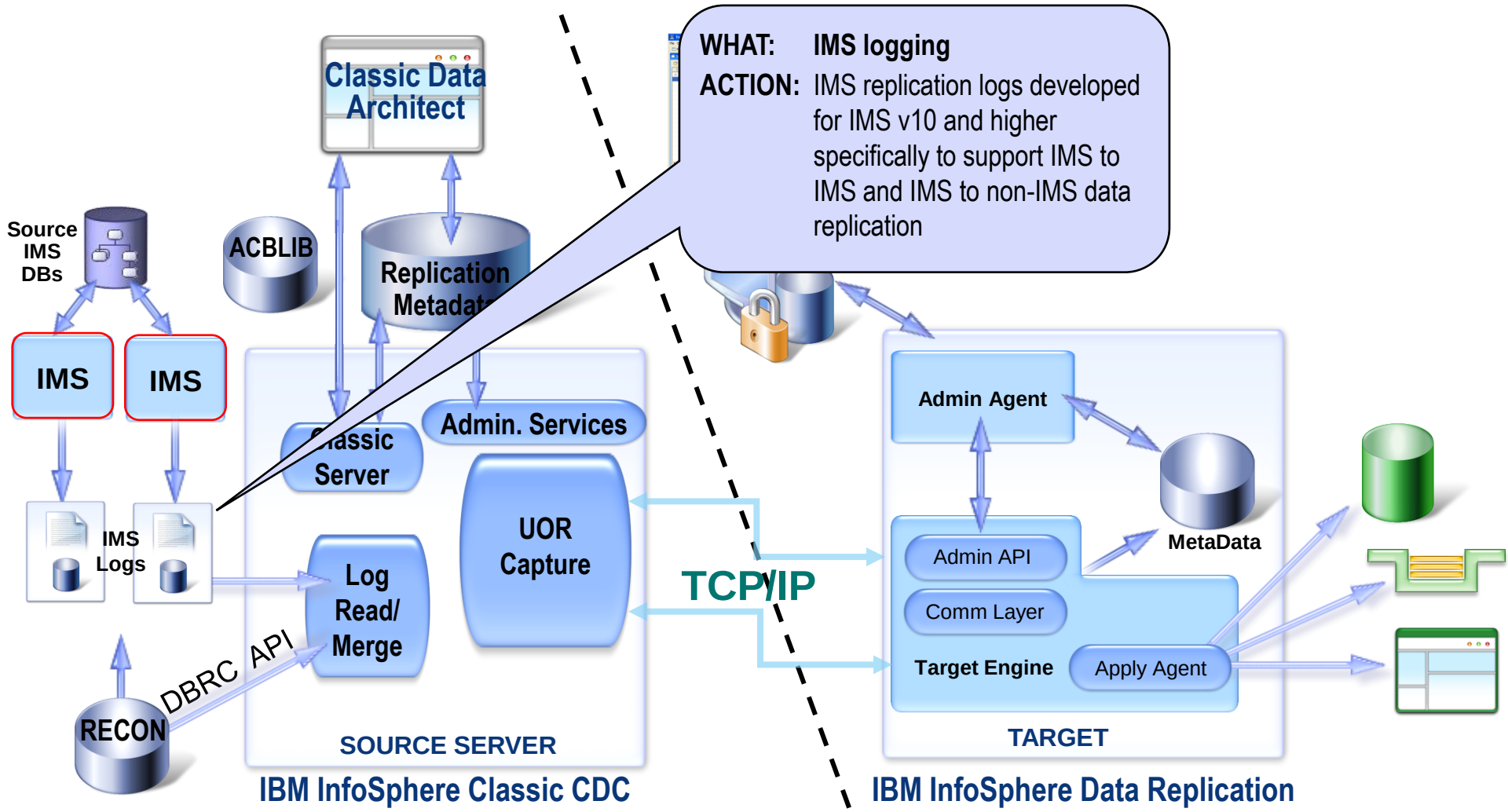
IBM's InfoSphere Data Replication (IIDR) Coverage



30

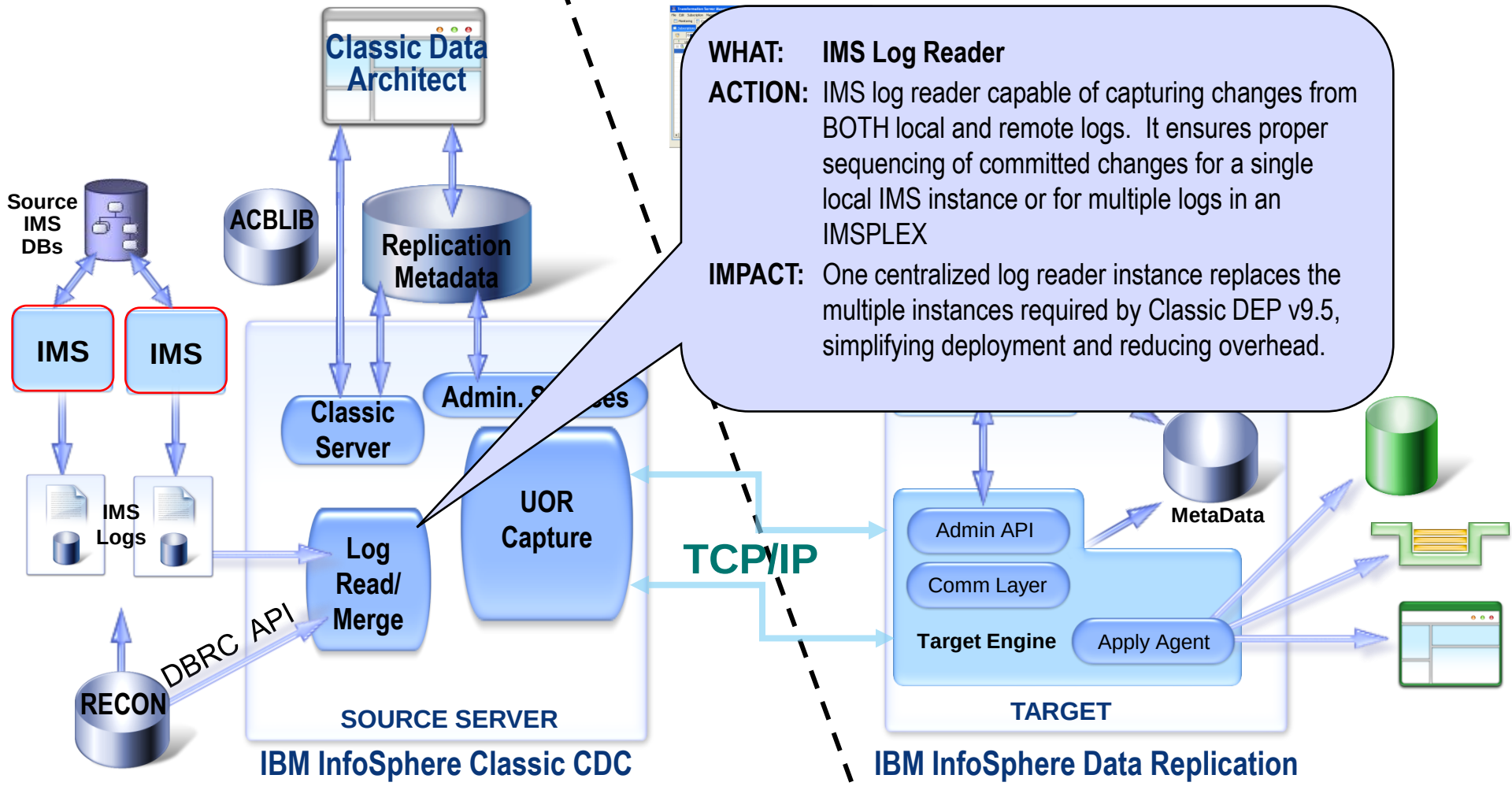
IMS to Hadoop

Read logs - Send committed changes - Apply changes



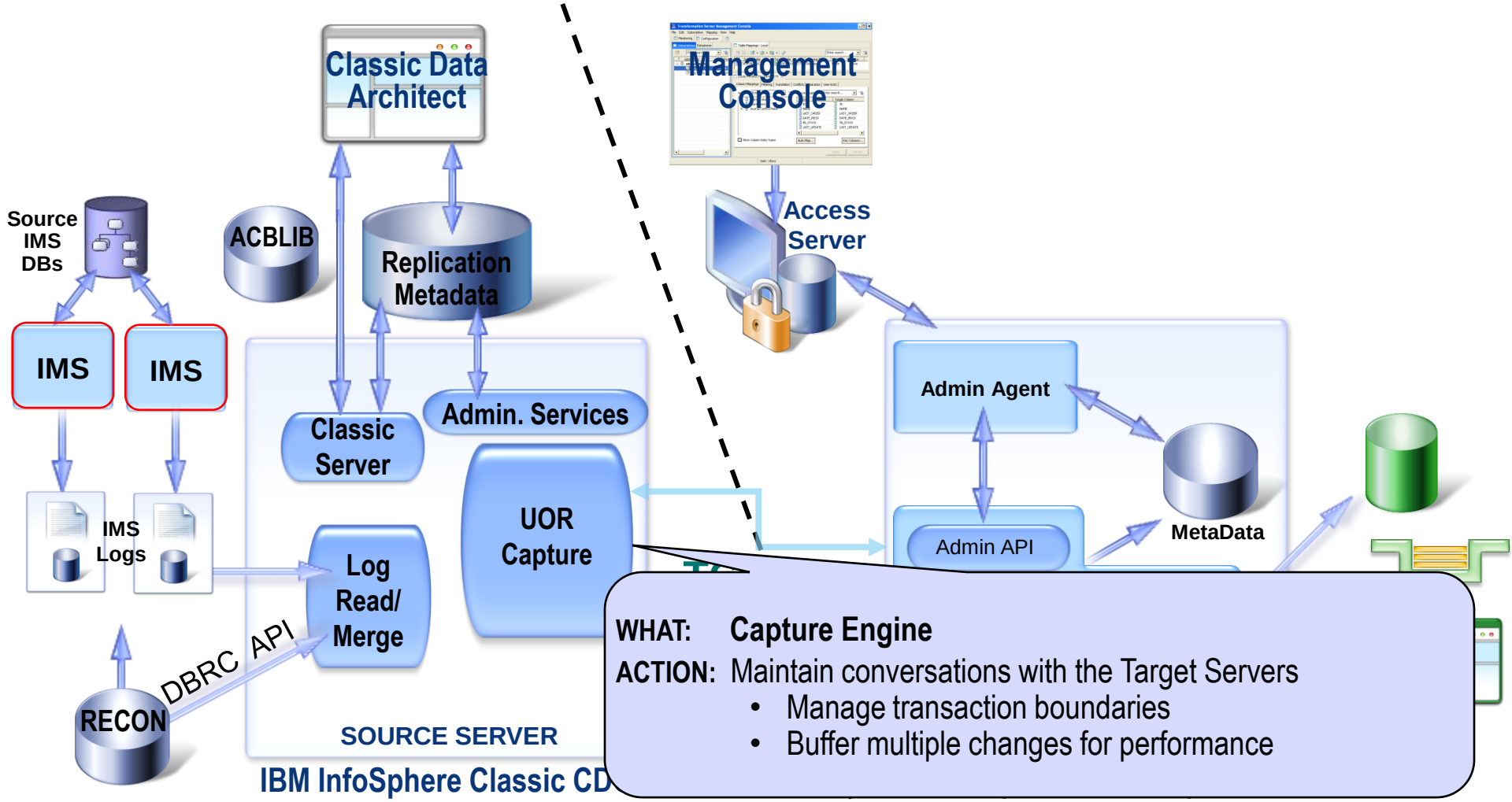
IMS to Non-IMS Data Replication

Read logs - Send committed changes - Apply changes



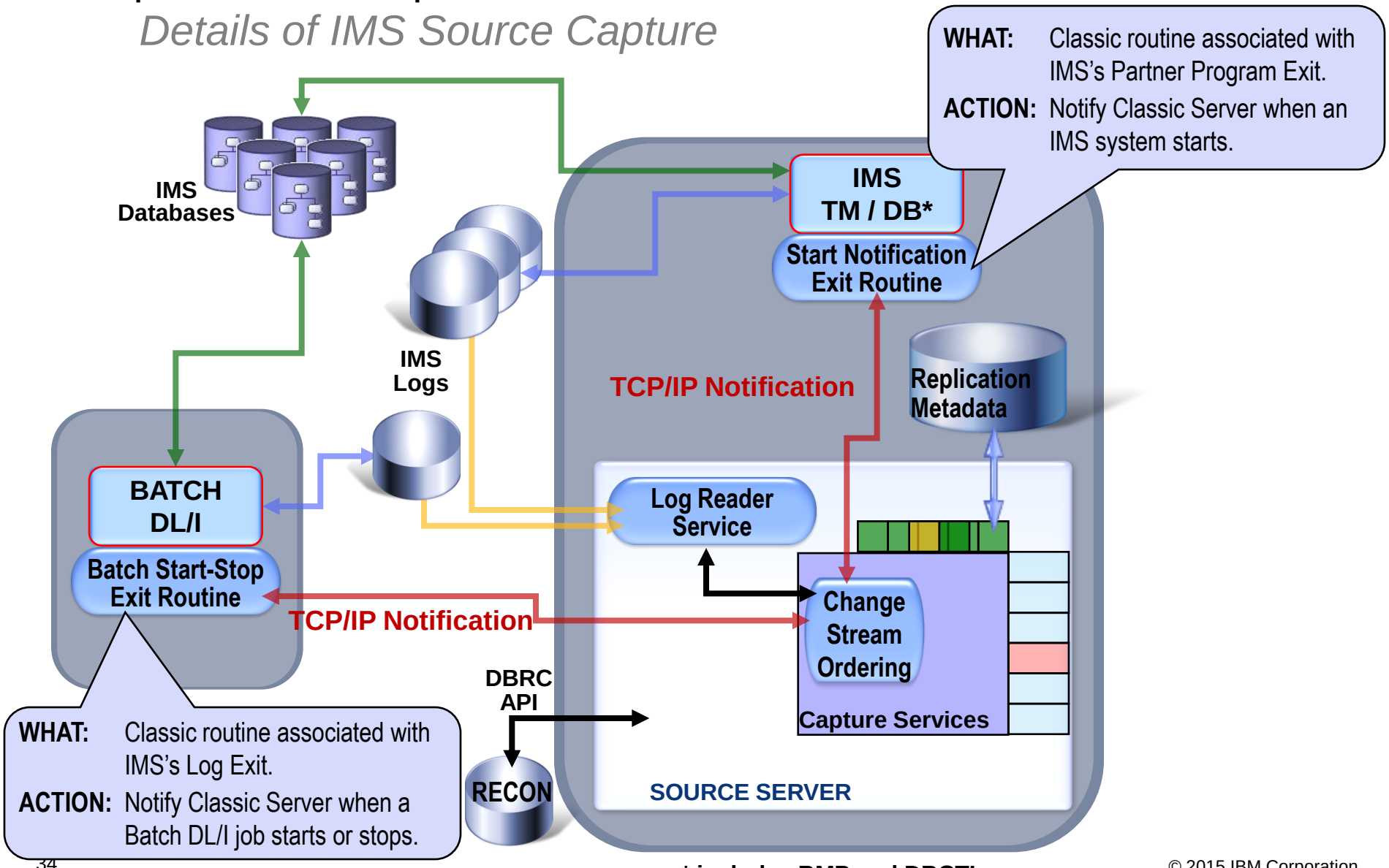
IMS to Non-IMS Data Replication

Read logs - Send committed changes - Apply changes



InfoSphere Data Replication for IMS for z/OS

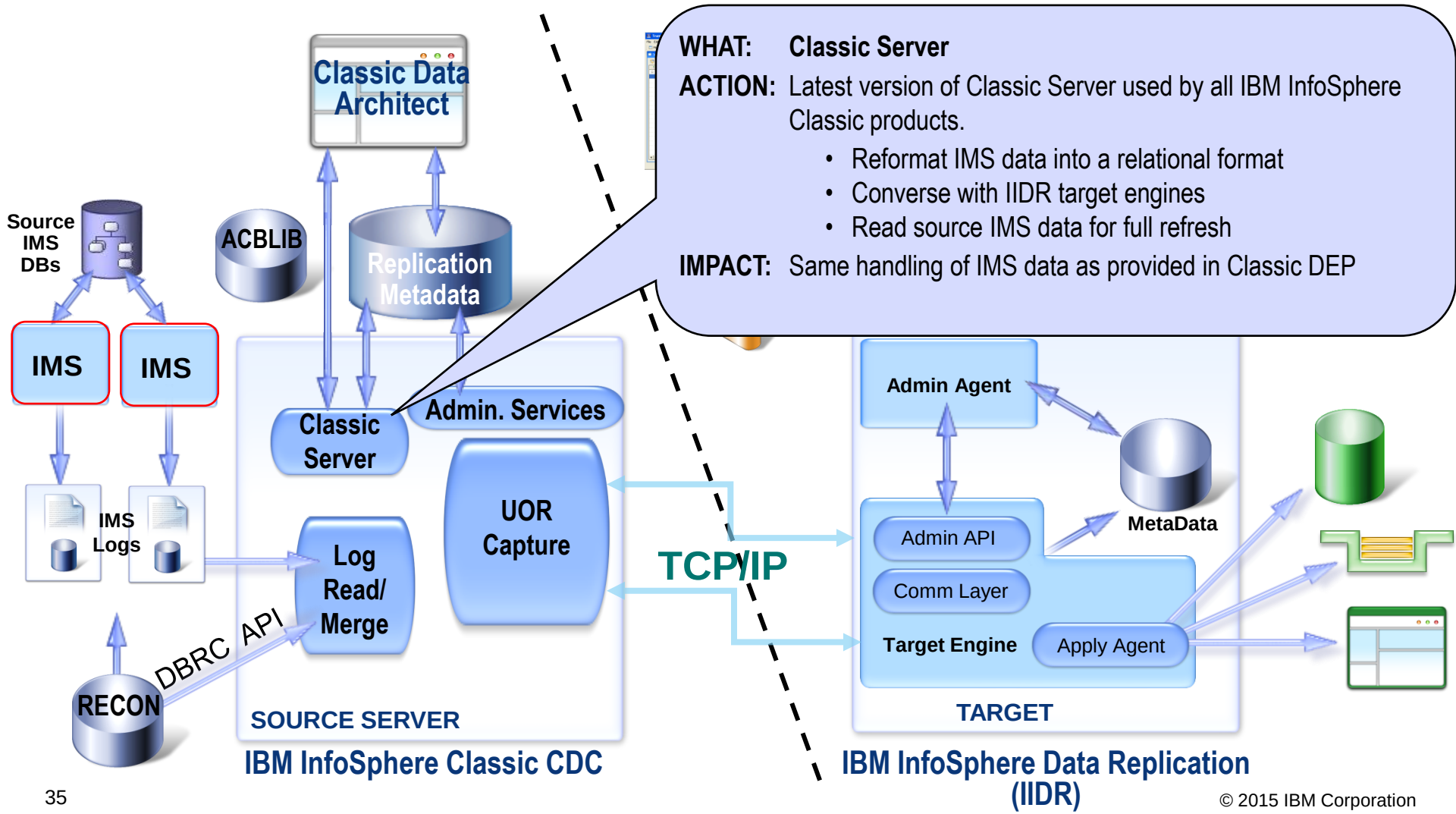
Details of IMS Source Capture



* includes BMP and DBCTL

IMS to Non-IMS Data Replication

Read logs - Send committed changes - Apply changes



IMS to Non-IMS Data Replication

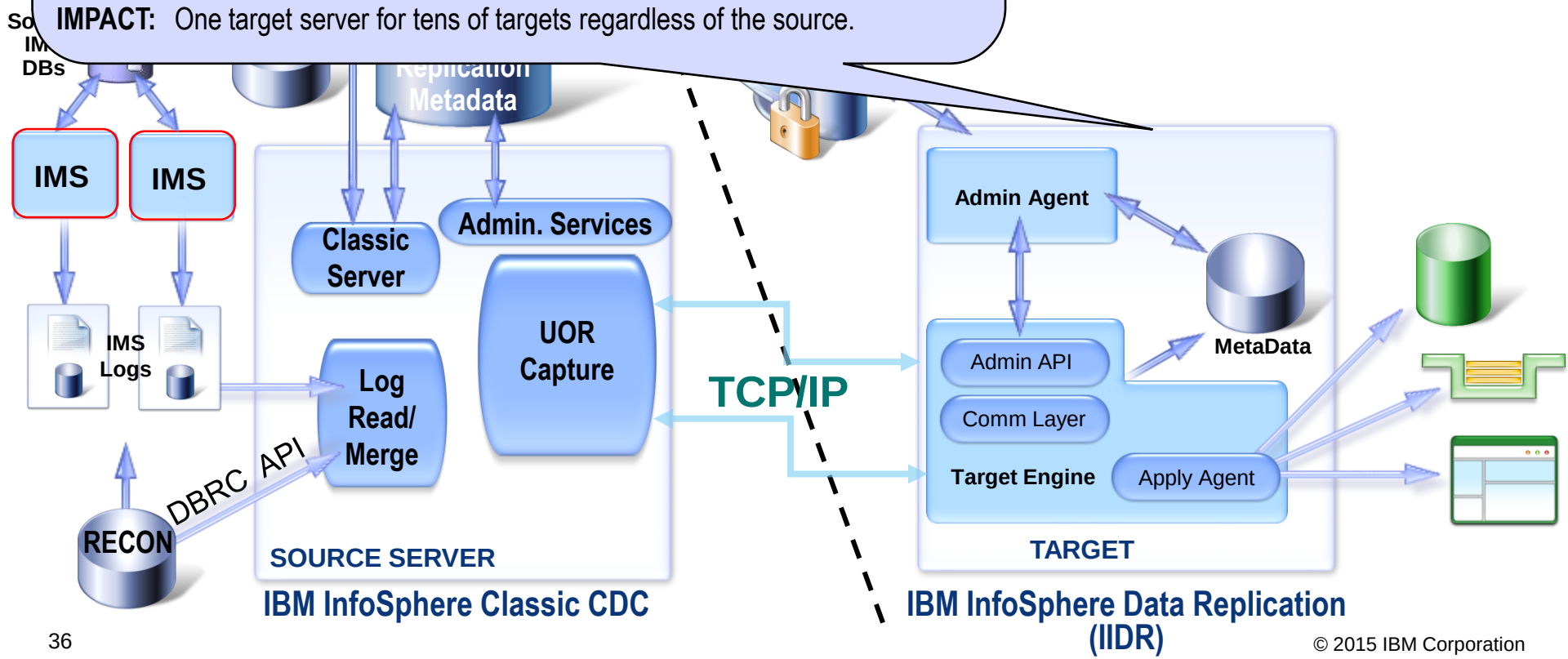
- Apply changes

WHAT: IIDR Target Server

ACTION: Apply changes to the target(s) while maintaining restart information in a local bookmark for recovery purposes

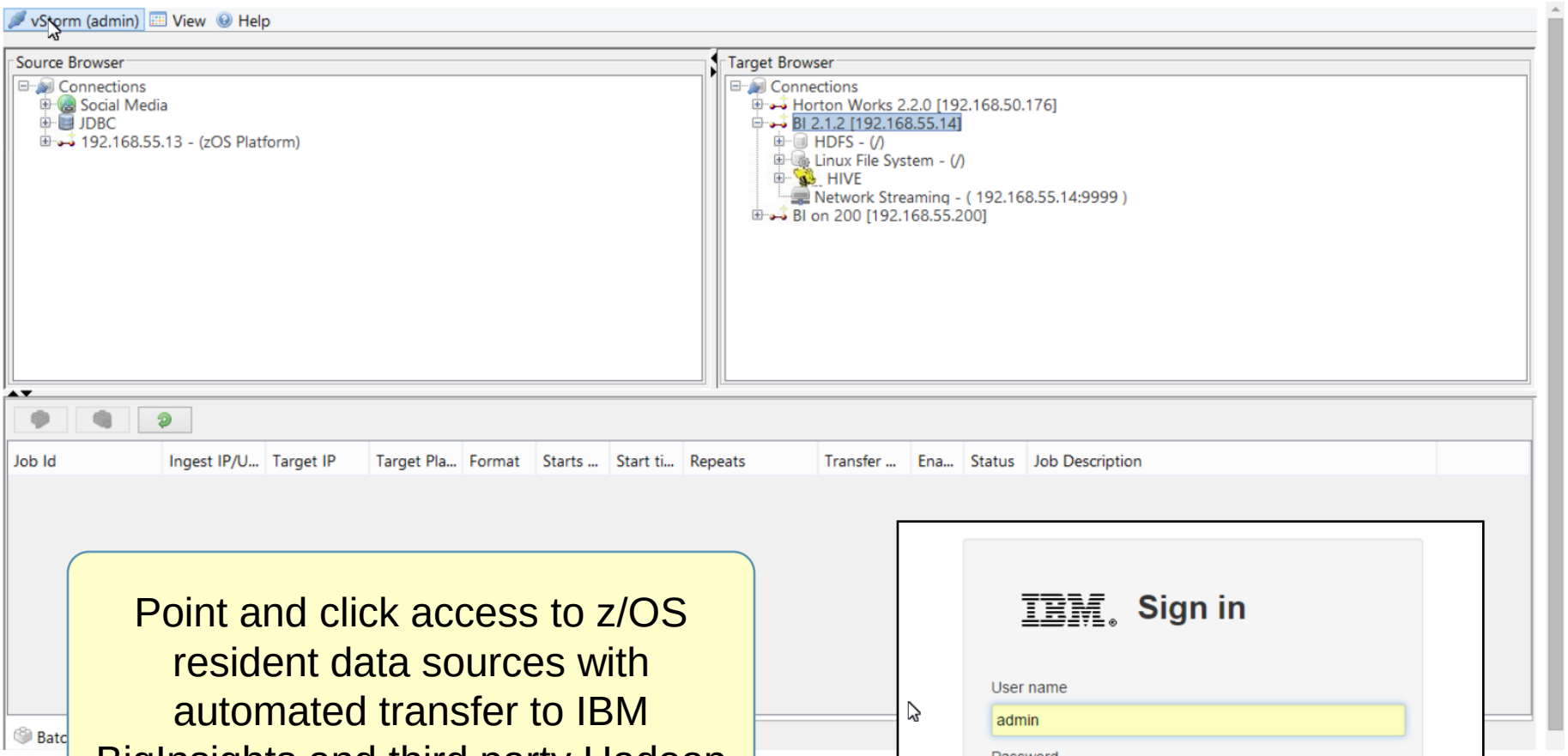
- Additional target-based transformations can be applied
- Integration with many other InfoSphere solutions
- Available for z/OS, Linux on System z, Linux, Unix, and Windows

IMPACT: One target server for tens of targets regardless of the source.



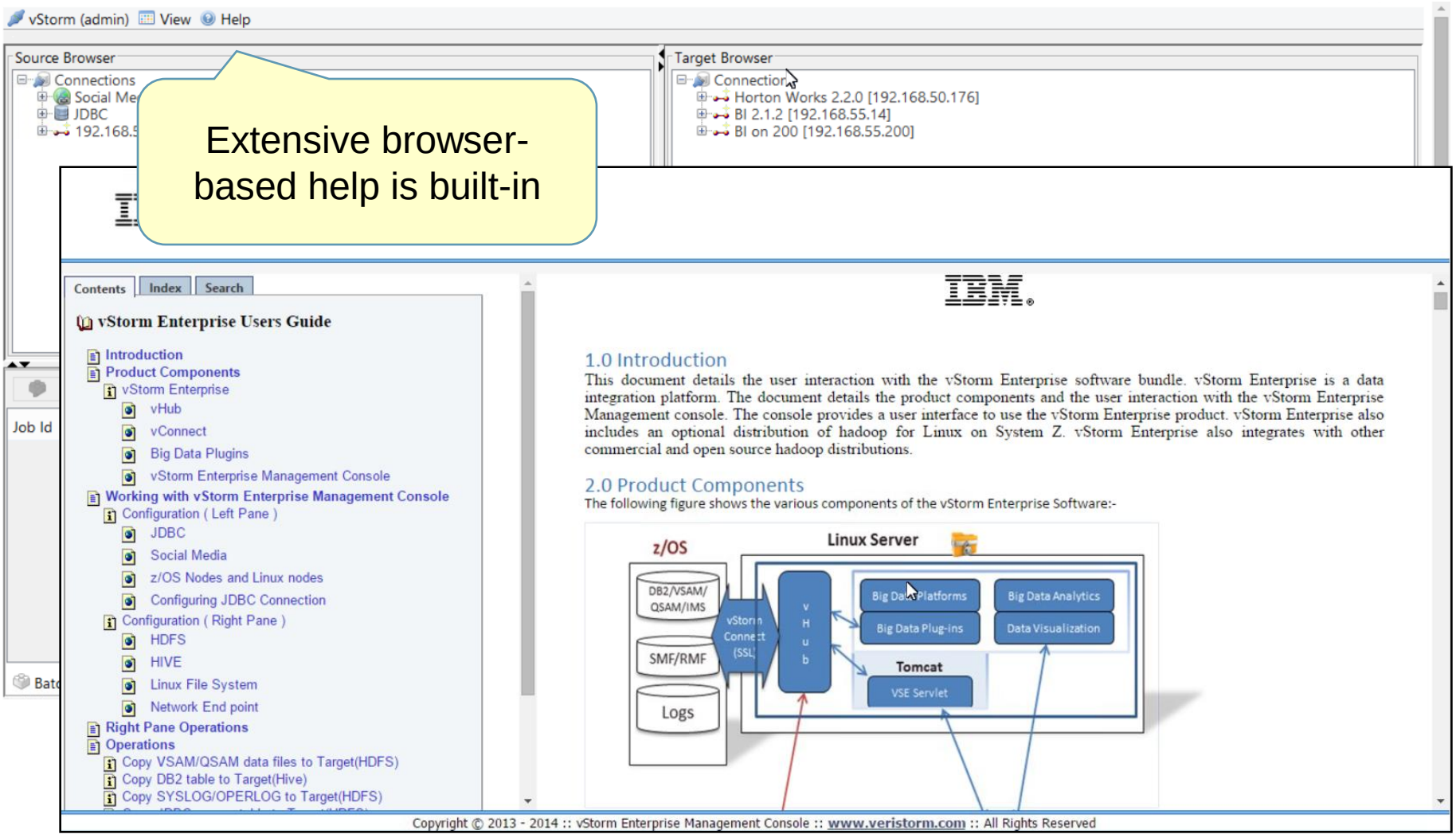
Thank you!

IBM InfoSphere System z Connector for Hadoop

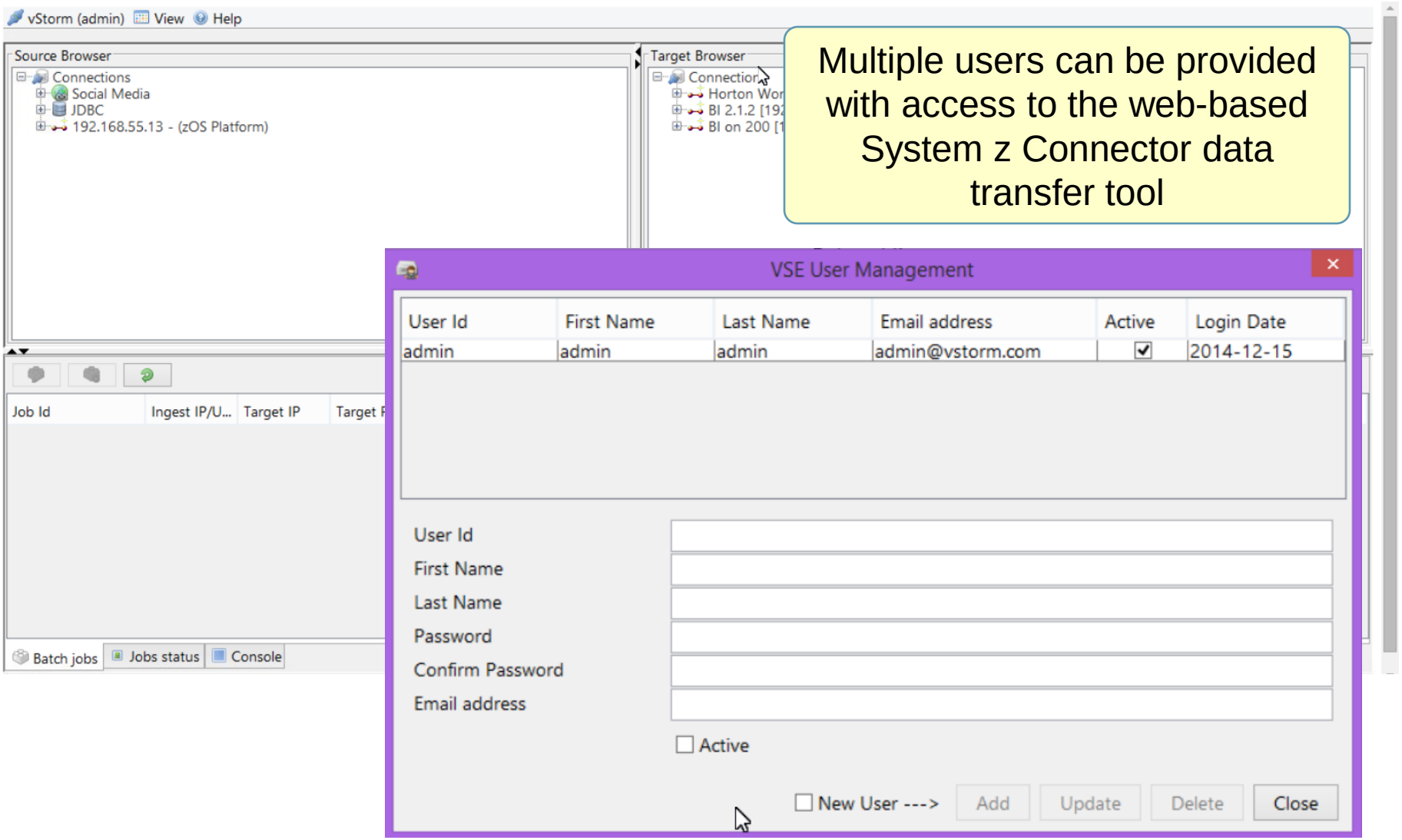


Point and click access to z/OS resident data sources with automated transfer to IBM BigInsights and third party Hadoop clusters.

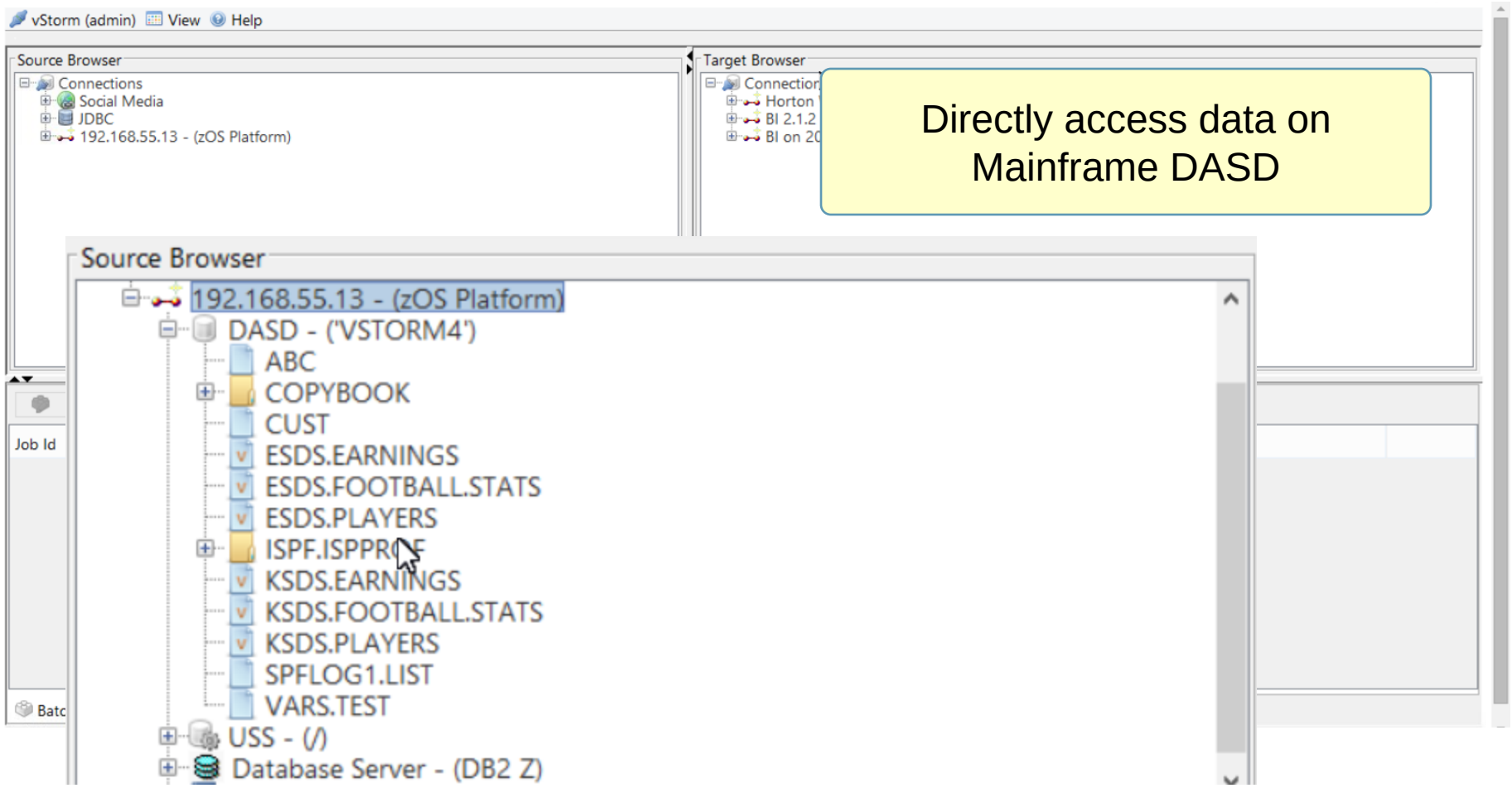
IBM InfoSphere System z Connector for Hadoop



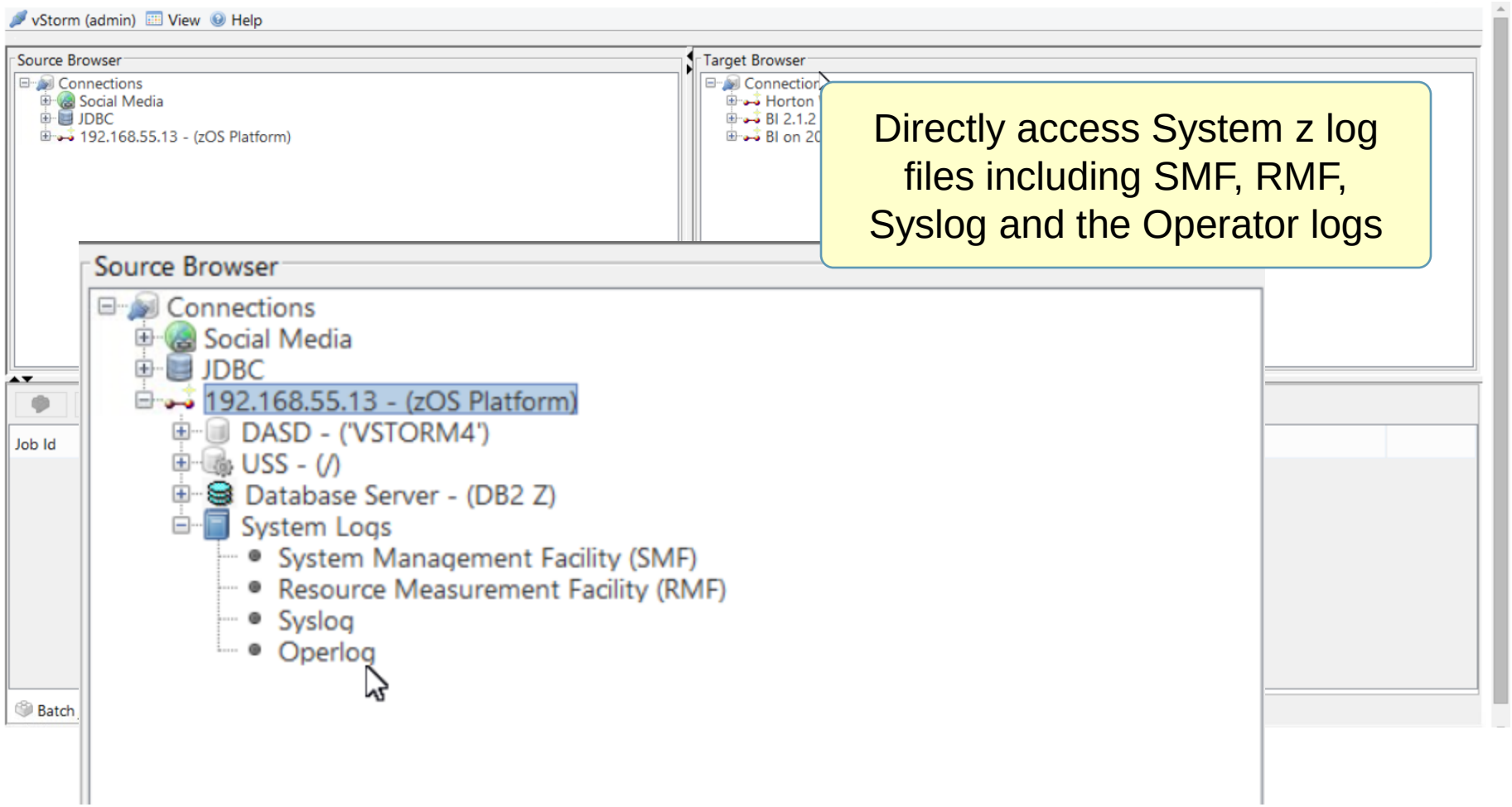
IBM InfoSphere System z Connector for Hadoop



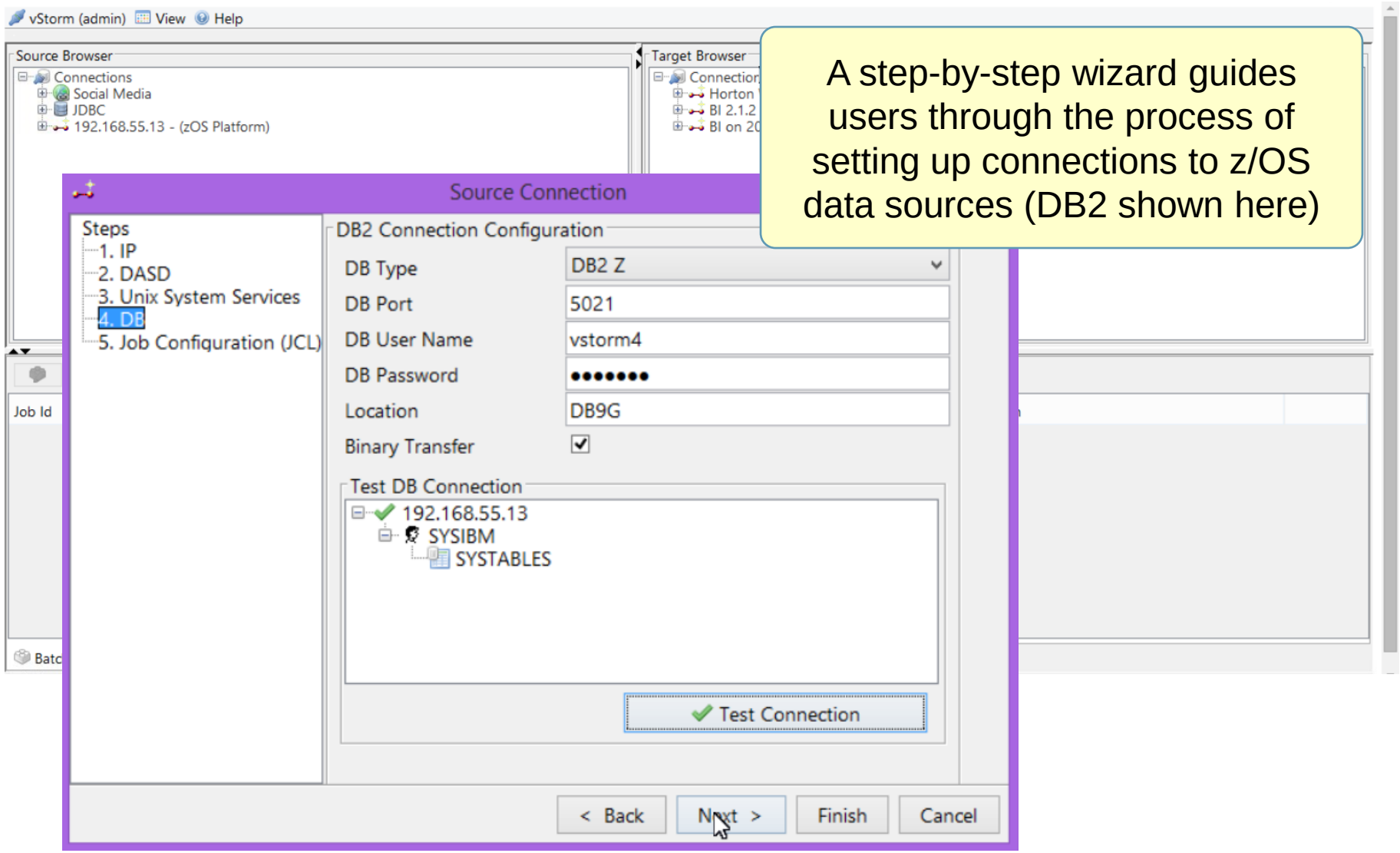
IBM InfoSphere System z Connector for Hadoop



IBM InfoSphere System z Connector for Hadoop

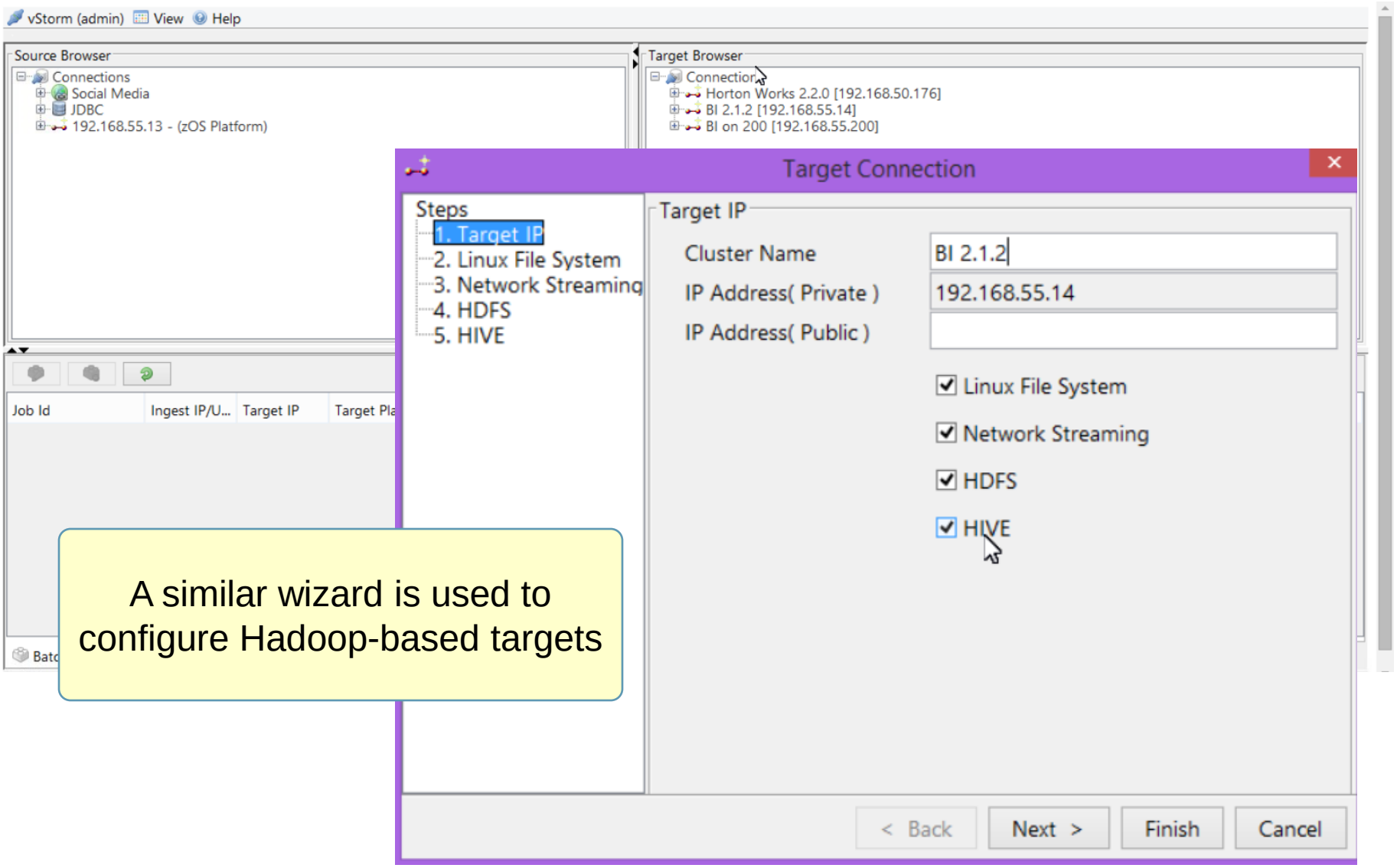


IBM InfoSphere System z Connector for Hadoop



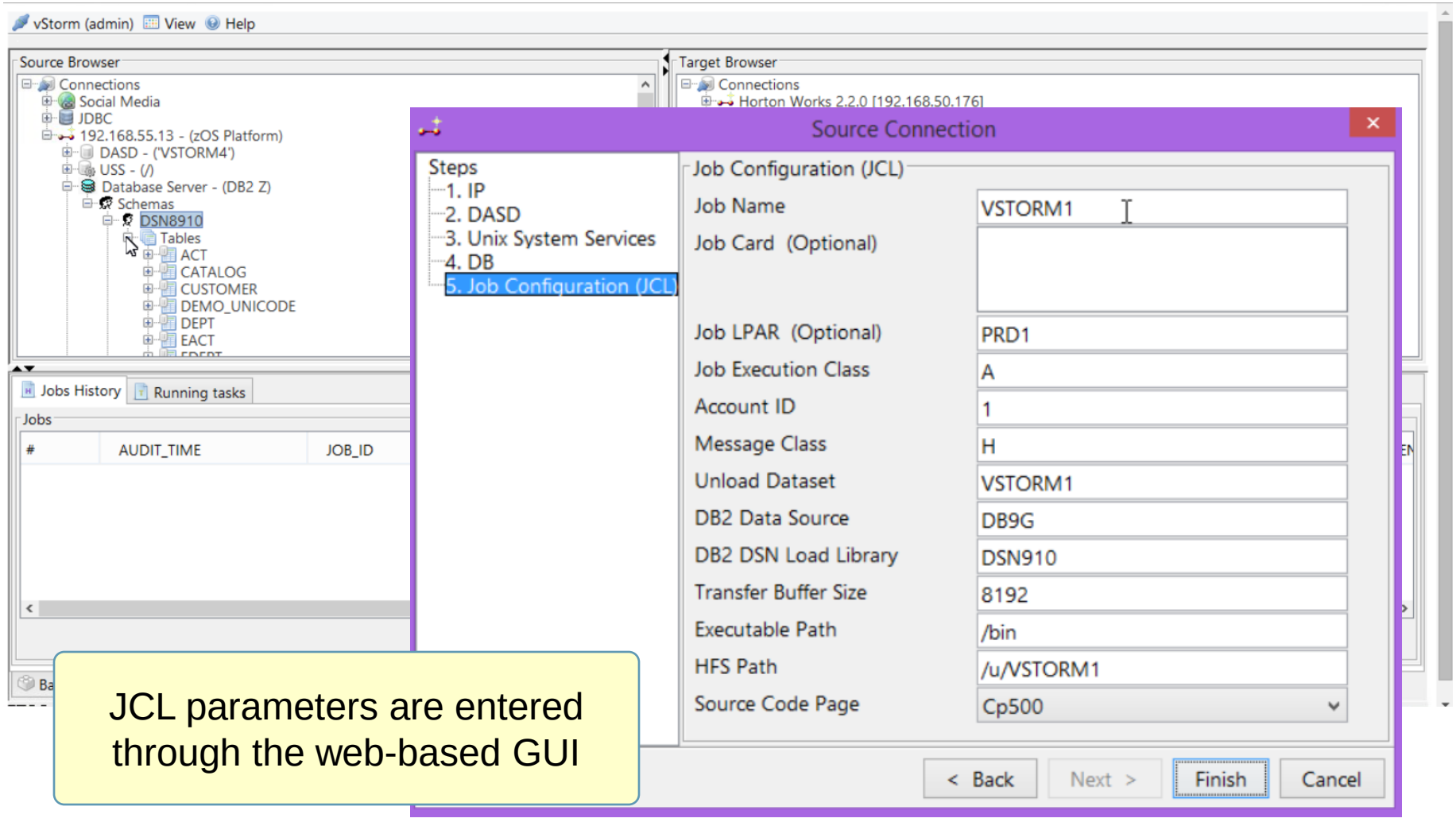
A step-by-step wizard guides users through the process of setting up connections to z/OS data sources (DB2 shown here)

IBM InfoSphere System z Connector for Hadoop



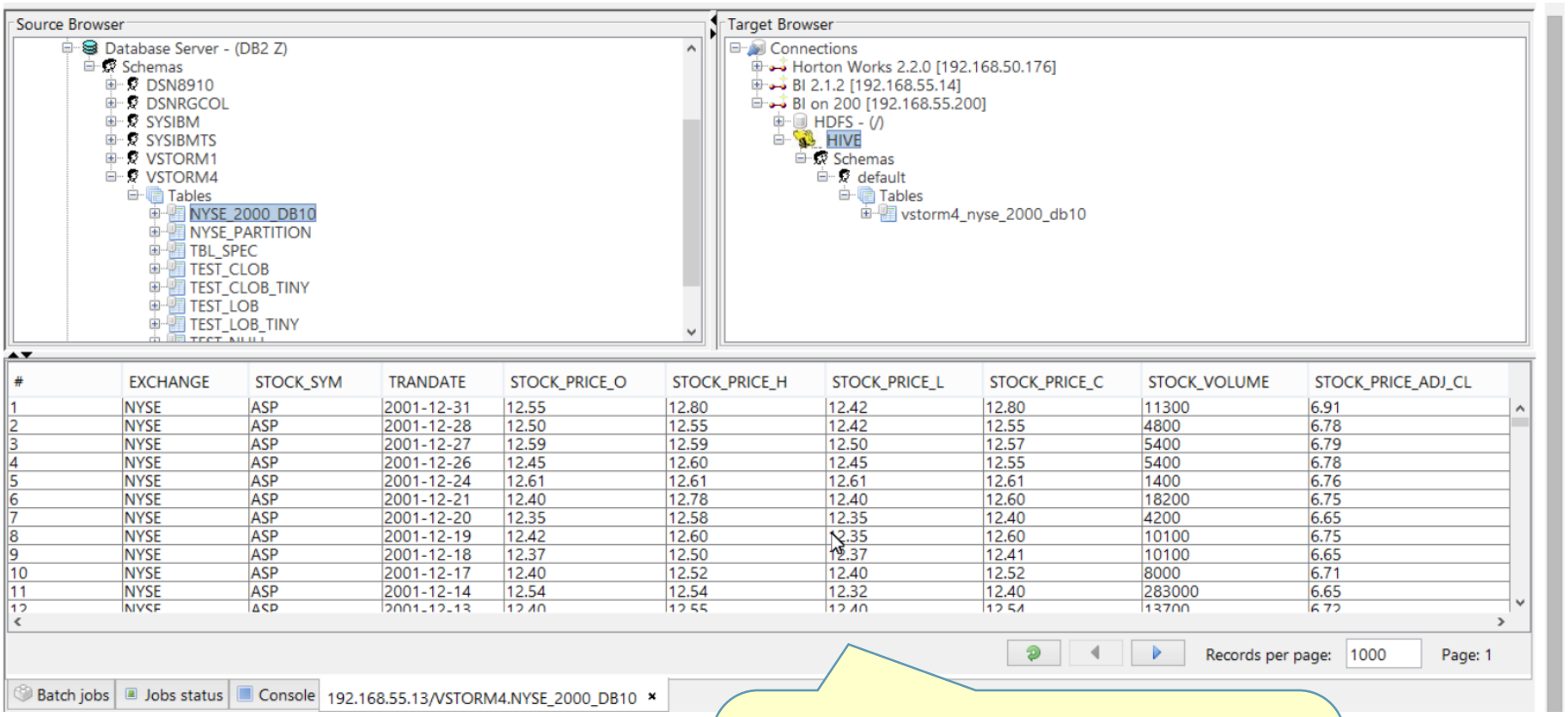
A similar wizard is used to configure Hadoop-based targets

IBM InfoSphere System z Connector for Hadoop



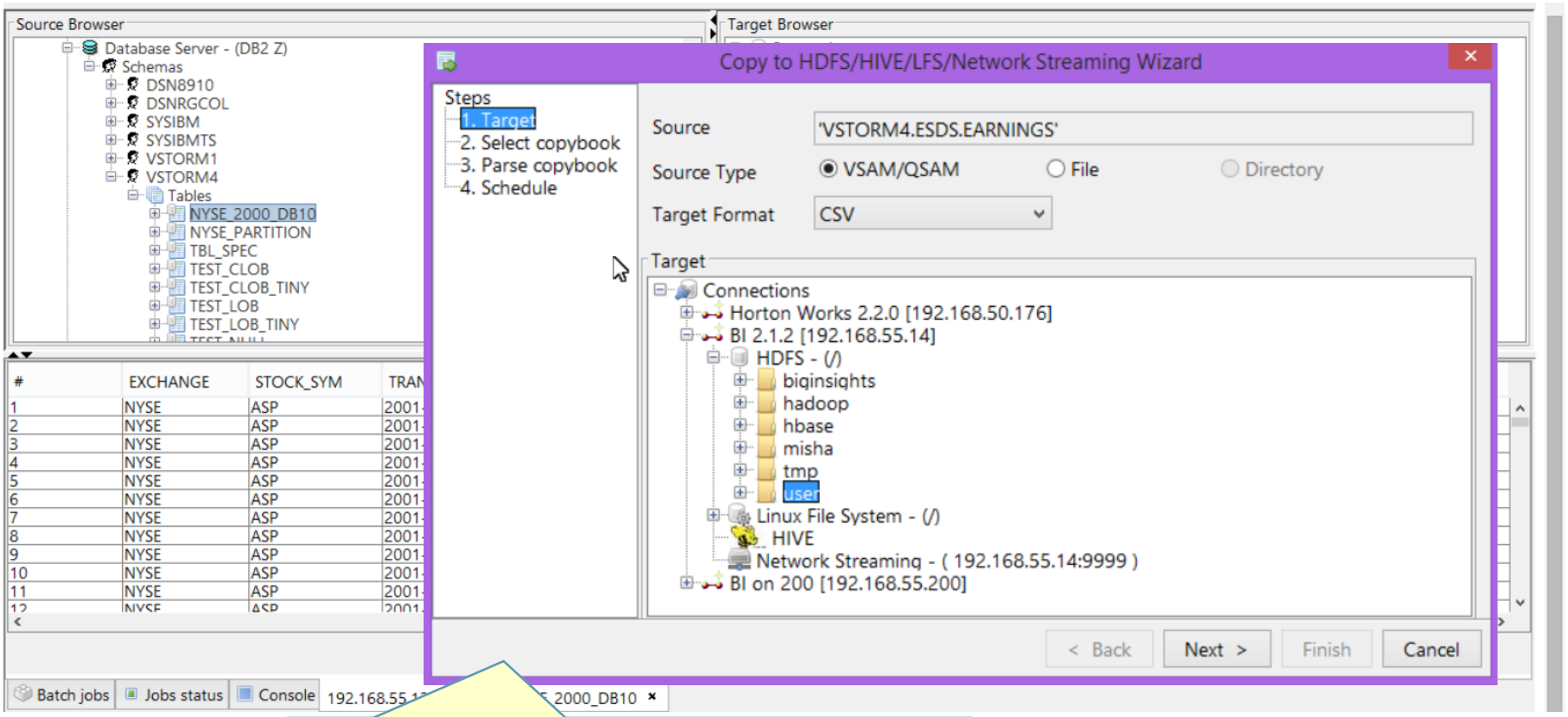
JCL parameters are entered through the web-based GUI

IBM InfoSphere System z Connector for Hadoop



Browse the contents of mainframe data sources from within the System z Connector for Hadoop interface

IBM InfoSphere System z Connector for Hadoop



Copy source data to HDFS, the Linux File System, HIVE or stream data directly to a receiving port on a Linux system or virtual machine in your chosen format

When it comes to realizing time to value:

Commercial Framework Differences Matter

Capabilities

Software tooling to build higher quality, more maintainable applications quickly and cost efficiently

Infrastructure

Deploy on reliable, cost-efficient infrastructure that matches quality-of-service requirements