ATTUNITY

Efficient and Real Time Data Integration
With Change Data Capture

Efficiency . Low Latency . No Batch Windows . Lower Costs.

an Attunity White Paper

Change Data Capture (CDC) is a strategic component in the data integration infrastructure that dramatically improves efficiencies, eliminates the needs for batch windows, deals with the growth of data volumes, delivers information in real-time, and reduces costs. Attunity Stream provides an enterprise-class CDC solution that complements and works seamlessly with existing ETL, data integration and event processing technologies.

Change Data Capture – Efficient and Real-time Data Integration
Attunity Stream Product White Paper – February 2009

Table of Contents

The Need for Change Data Capture (CDC)

Information in general and Business Intelligence (BI) in particular, is at the center of the IT infrastructure of the largest global organizations, enabling them to understand business trends, improve decision making and support day-to-day operations. Traditionally, Data Integration, Data Propagation, and ETL (extract, transform and load) processes run on a periodic basis (weekly, daily) and use a bulk data movement approach that moves and integrates the entire source data to a target data source. This approach is limited in its ability to deal with new trends and business requirements, some of which include continuous refreshing of a DW, real time business intelligence and operational data stores. Change-Data-Capture (CDC) technology provides a solution to IT operational challenges and the new business models and indeed has become a strategic requirement for many Data Integration (DI) initiatives.

Trends and Changing Business Requirements

The following table lists key trends and emerging requirements that affect the data integration market and demonstrates the limitations of traditional ETL and Data Integration:

	Trends and new Requirements	Traditional ETL/DI
1.	***Data Warehouses required to support tactical and operational BI applications***. Required for timeliness of the information and integration with operational data stores.	Data Warehouses support reporting and analytic BI applications for strategic decision making.
2.	***Business users need up-to-date information***, with low latency ranging from hours to minutes or even seconds. This requirement is fueled by competitive pressures, new business models, and customer satisfaction requirements.	The data in the Data Warehouse is a few days old, typically updated daily or weekly. Information in a DW is normally in a summary format and does not effectively address operational requirements.
3.	***Data volumes are doubling every 1-2 years***, making the option of moving the entire source data impractical, or even unfeasible.	Move all the source data. Traditionally, the entire source database is extracted, transformed and then loaded to the target Data Warehouse or Data Mart.
4.	***Batch Windows are shrinking***. Continuous uptime and the need to use every bit of resources for transaction processing impacts the available time for batch windows.	Use 'Batch Windows' for ETL. These windows represents periods of time when the operational system is not working but rather processing data for ETL purposes.

What is CDC?

Change Data Capture (CDC) is an innovative approach to data integration, based on the identification, capture, and delivery of changes made to enterprise data sources. By processing only the changes, CDC makes the data integration process more efficient and reduces its cost by reducing the latency between the time a change in the data occurs and the time the same change is made available to the business user (see above figure.)

Many companies have home-grown CDC solutions that are typically limited in scope and are costly and difficult to maintain. Today, a new breed of pure-play CDC software is available that supports many different business intelligence initiatives and data integration processes and is a strategic component of the data integration infrastructure and compliments other technologies such as ETL and EAI.

The value of CDC

CDC allows IT managers address many of the information management challenges, some of which are listed above, as we as new trends and requirements:

1. ***Delivers data on-demand and in near real-time***, providing business users with the most current information supporting tactical and operational BI applications.

2. ***Dramatically increasing efficiencies*** by moving only the data that has changed and addressing the growth in data volumes head-on with continuous data feeds that can be filtered and scheduled based on needs and requirements

3. ***Eliminates the need for Batch Windows***, Data is captured and processed while the systems keep working. Again, continuous data feeds can virtually eliminate the “batch window” bottleneck.

4. ***Strategic solution***, Meaning a comprehensive platform that works with other integration products and can be used for many initiatives, including BI/DW, CPM, BAM, Migrations, Consolidations, and others.

5. ***Cost savings, CDC can substantially reduce IT operational costs in terms of human resources required for a given integration project, as well as on-going costs related to system and storage requirements***

Attunity Stream – providing Enterprise Class CDC

Attunity Stream is a CDC software solution that supports many enterprise data sources including Mainframe, AS/400 and Oracle databases (with its Universal Data Access). It works seamlessly with all the leading ETL and EAI tools and has been proven to dramatically improve the efficiencies of existing ETL by 500% and more.

To learn more about the Attunity products, visit us at www.attunity.com or email info@attunity.com.

The rest of this document provides in-depth information about the required functions of a CDC solution and how Attunity enables enterprises to leverage CDC today for more efficient ETL and more effective BI.

Use Cases for CDC

Attunity's Strategic CDC technology is a comprehensive solution that addresses many business requirements and can be applied to solve a variety of real time and on-demand initiative that can substantially improve return on investment. Following are examples of how CDC can be implemented to solve real business problems:

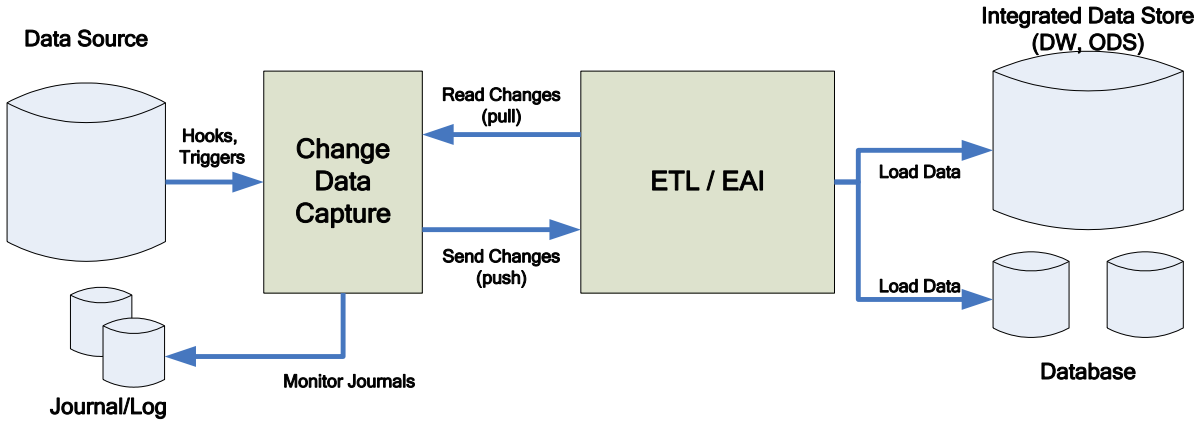
1. **ETL and Data Warehousing**
2. **Building Operational Data Stores (ODS)**
3. **Real-time Dashboards**
4. **Data Propagation**
5. **Data Synchronization**
6. **Application Integration**
7. **BAM – Business Activity Monitoring**
8. **Data Quality**

ETL and Data Warehousing (DW)

A common case for using CDC involves the process of moving information into a DW. Traditionally, DW updates are processed with an ETL (Extract, Transform, and Load) tool. ETL is a software program that extracts the data from the source system, transforms and cleanses the data, and then loads it into the DW. These processes require that the operational system(s) be put off line for a given period of time. This period of time is referred to as a "Batch Window", typically measured in hours and sometimes days, during which the system is busy with moving the data and cannot perform operational and other mission critical functions. Given the limitation of this 'bulk' approach, most IT shops update their DW only daily, and often on a weekly basis.

Given the need of many companies for up-to-the-minute information, they have started looking for ways to update their DW in real-time, considerably reducing latency. EAI tools are sometimes considered for accomplishing the same goal. CDC, however, provides a new approach to moving information into a DW, and can work seamlessly with ETL or EAI tools.

The following diagram illustrates what a CDC solution looks like:



CDC delivers changes to an ETL or EAI tool in batch or real-time, allowing to dramatically improve the efficiency of the entire process, reduce or totally eliminate batch windows, delivery information in low latency, and reduce the associated costs including CPU cycles, storage, network bandwidth and human resources.

Building Operational Data Stores (ODS)

An ODS is an integrated data repository addressing a specific business area (marketing, finance, support and maintenance, etc.) and providing complete and current information that can be used by business users and BI applications. The ODS is stored in a relational database and receives updates from a dedicated set of programs, ETL or EAI tools. Given the business need fulfilled by an operational data store, it requires timely updates.

CDC provides an efficient mechanism to keep an ODS up-to-date, by identifying and delivering the changes on a continuous basis, rather then periodically querying the entire data base for changes. In addition, CDC can push changes in near real-time to support ODS applications that have very low latency requirements.

Real-time Dashboards

Dashboards provide managers with selected metrics, known as Key Performance Indicators (KPI), that measure the performance of various business operations. These metrics can represent sales trends, margin tracking, financial triggers and others that alert the business user to a business condition requiring attention. Dashboards however, are only as good as the metrics they provide and the process of updating these metrics is typically done at the data integration level. Metrics that are updated infrequently mean that the user might become aware of a business problem too late. After all, how far will you get with a fuel gauge that is updates every 2 hours?

Implementing a CDC will provides a method of identifying changes to different data sources that are required in order to measure a certain KPI. Attunity's CDC solution comes with data filters that can process data changes in a given order and to re-calculate new values of the KPI close to the time of the change. This in turn can provide greater and timelier visibility to the business user resulting in much faster and effective response to changing business conditions.

Data Propagation

Data propagation address to the need to have one or more copies of the data from a given data source. Common examples include, making production data available for reporting purposes and accessible by various departments; distributing data from a source system to multiple data centers, and often to multiple ODS to improve response time.

With CDC, the process of propagating data can be made much more efficient and reduce the latency in making new data available.

Data Synchronization

Data synchronization is typically required in order to keep two or more systems in sync and up to date. A common reason for data synchronization is system migration (as a result of business consolidation; down sizing, etc…) where running two systems in parallel for a period of time, sometimes for several years, is required. During that period, transactions that are captured by one system need to be updated in the other as well. Another common reason for data synchronization is a merger or acquisition (M&A) where systems used by one company overlap with others, and both needs to be kept up to date.

CDC allows capturing changes to as they occur and process them using such tools as EAI, ETL, or homegrown programs. By incorporating CDC, data synchronization can be made efficient and real-time.

Application Integration

Application integration is a process that allows the integration of a company's data across its various application systems and thereby enabling these systems to "talk" to one another through a messaging layer. Application integration provides a "unified view" of an enterprise's business and its applications, seeing how existing applications, legacy and otherwise, fit into the new view and devising ways to efficiently reuse what already exists while adding new applications and data.

CDC can take change events and generate messages that can be processed by Application Integration products (also known as EAI tools) as part of orchestrated processes. Such events can trigger an automated process for updating one or more applications, or for processing a set of rules and execute an automated business process.

BAM – Business Activity Monitoring

BAM monitors the activity of business processes for various purposes including process optimization, performance management, and exception identification. BAM relies on receiving data about business events from many source systems for the purpose of monitoring and analyzing business activities. Given the dynamic nature of business activities, BAM products, in order to provide quality business insight, require real-time data feeds from multiple production systems.

CDC provides a mechanism to identify activities at the data level. These activities can be filtered based on a certain selection criteria and then delivered to a BAM application for processing. For example, it can provide records that maintain processing states or changed inventory levels in a manufacturing environment.

Data Quality

IT/Business Problem Solved: **Continuous monitoring of data quality.**

Many BI initiatives fail because of poor data quality. As a result, organizations are looking for ways to improve the quality of data available to business users. In some cases, data is cleansed as part of the process of creating a new data store (e.g. DW). In others, processes are put in place to clean up the source data.

With CDC, companies can identify and capture changes to source systems as they happen and immediately feed them into data quality processes. Such processes can then apply quality rules and recommend whether clean up activities are required. This shortens the time for ‘bad’ data to reside in the system.

What to look for in a CDC Solution?

As the previous section demonstrates, CDC can be used for many types of initiatives, and enterprises should take a strategic view of CDC to make sure that it can address current and future requirements. This section provides an overview of key capabilities and functions required in a strategic CDC solution, and ones that you should look for when evaluating any CDC technology. These include:

1. Capture Changes from Many Data Source Types, on Many Platforms
2. Non-Intrusive Change Capture
3. Change Filtering
4. Batch and Near Real-Time Delivery
5. Handle Non-Relational Data (e.g. Mainframe)
6. Guaranteed Delivery
7. Interoperability with Leading ETL/EAI Tools
8. Recoverability
9. Performance and Throughput
10. Ease of Use

Capture Changes to Many Data Source Types, on Many Platforms

The IT landscape in medium and large organizations includes many platforms and data sources that run the various operational systems. There are many benefits in choosing a CDC solution that supports many data sources on many platforms, from Windows and UNIX, to Mainframe. By adopting and standardizing on a strategic platform rather than point products, organizations can minimize the number of integration products they use, have a standard and common architecture, shorten learning curves, reduce required skill sets, reuse trained personnel for many projects and reduce total cost of ownership.

Non-Intrusive Change Capture

A critical factor in CDC is the impact on the source system. There are various degrees of impact and customers should be careful when evaluating CDC solutions and look for those that minimize the impact on the source system, including its maintenance, operations, and cost.

The most invasive approach requires changes to the applications that actually make the changes. There may be hundreds or more such applications and the cost of making and maintaining such a solution is prohibitive. Less invasive approaches make changes to the schemas of the source tables, typically by

adding timestamp fields. This approach impacts the source system in terms of its storage requirements and the processing required in order to update the timestamps.

Other approaches use triggers within the source system to capture and process changes. While these appear to have no impact on the applications that make the changes, they do have a significant impact on the system, which includes additional processing time that takes away from the resources that are available for the operational system and it impact capacity planning. In addition, the use of triggers can introduce error into a system that is already in operations adding the need to manage the triggers.

The most non-invasive approach uses system or database logs or journals. In this approach, the CDC mechanism is not accessing the source system, but rather its log. The solution can be managed separately and it does not share resources with the operational system.

Change Filtering

A key goal of CDC is to improve efficiency by reducing the amount of data that needs to be processed to a minimum. CDC solutions can provide filters that allow to reduce the amount of information and deliver only the relevant records. Such filters enable the delivery of records based on the type of change (i.e. inserts, updates), deliver records only when a change happened to specific fields, or specify a subset of fields from the original record that are required for processing.

These filters can achieve maximum benefit when they run close to the source, by reducing the amount of records that traverse the network.

Batch and Near Real-Time Delivery

Different applications that use CDC may have different latency requirements. A robust CDC solution provides the ability to work efficiently with different data delivery models supporting the different latency requirements of such applications.

For example, a DW can be updated once a day or every four hours. The update is typically performed by an ETL tools and there are many records to process. It makes sense in this case to read many changes at once and process them in batch.

In another case, two systems need to be synchronized in near real-time. The coordination of the process may be done with a message-driven EAI tool, where messages are handled individually, as soon as possible. This case calls for a real-time type of delivery.

Handle Non-Relational Data (e.g. Mainframe)

Information residing in legacy systems such as VSAM, IMS or Adabas, is usually mission critical. This data is more difficult to manage because of its non-relational structure. Companies that have these type data sources should look for CDC solutions that are able to deal with Non Relational sources effectively.

A CDC solution that can capture non-relational data should be able to deliver the information in a way that can be easily processed by other tools such as ETL or EAI. If changes are processed by an ETL tool that uses SQL, look for a solution that can normalize the non-relational data and provide a relational metadata model. If changes are processed by an EAI tool, typically in XML, look for a solution that can map the legacy data source into an XML document with a corresponding XML schema that represents the original record hierarchy. A robust CDC solution should support both.

Guaranteed Delivery

All integration processes require a certain level of guaranteed delivery. While some can use ‘at least once’ guarantee other may require ‘once and only once’. For example, ‘at least once’ guarantee may be acceptable for loading an ODS since processing a record twice does not cause a problem. However, a process that synchronizes applications and updates sales orders must guarantee the processing of a new order ‘once and only once’. There is usually a tradeoff, since a stricter guarantee requires more resources.

A CDC solution needs to provide a robust architecture where the desired level of guarantee can be achieved.

Interoperability with ETL/EAI Tools

Different tools are often used for different initiatives and a big organization will typically have at least one EAI tool (e.g. BizTalk Server), at least one ETL tools (e.g. IBM/Ascential DataStage), a BAM tool, etc. When choosing a strategic CDC solution, companies should make sure that the CDC solution the select is interoperable with the existing products and has an open architecture that will allow it to integrate with future integration technologies.

Some ETL products include CDC capabilities, but these capabilities are limited to the ETL being used and do not interoperate with other tools. Always look for CDC solutions that use open interfaces like SQL and XML. Preference should be given to CDC vendors that have certified their solution with the ETL or EAI tool that you are using or planning to use.

Recoverability

Data integration takes place over a network and over a certain period of time. As outage can happen, it is important that the CDC solution will be fault-resilient and able to recover. CDC solutions should provide support for recovery procedures, including automated restart as well as manual reset to a specific point in time where the system stopped processing.

Performance and Throughput

Always an important factor, a CDC solution needs to support the performance and throughput capabilities that will allow processing the accumulated changes in the desired timeframe. There are many factors that impact performance and throughout, including network traffic, storage for staging the changes, communication protocols, etc. When considering CDC solution, evaluate the technologies and architectures that the vendor put in place to improve performance and increase potential throughput.

Ease of Use

Integration in general and data integration in particular is not an easy task. It requires domain expertise, technology know-how and the ability to map between systems that were not designed to work together. By making a solution easy to use, a CDC solution can reduce required skill set and accelerate implementation timeframes. These in turn impact the total cost of ownership (TCO).

When evaluating ease of use, look for intuitive solutions that assist the user with task-oriented guides, and simplify configuration with wizards and utilities.

Attunity Stream – a Strategic Platform for Enterprise CDC

Attunity Stream is a change data capture (CDC) solution and a part of the Attunity Integration Suite data integration platform. Attunity Stream was designed as a strategic CDC solution and is unique in its scope of functionality, breadth of support, and proven interoperability with leading ETL and EAI tools.

Features and Capabilities

The following list provides an overview of the features and capabilities that Attunity Steam provides, mapped to the selection criteria mentioned earlier:

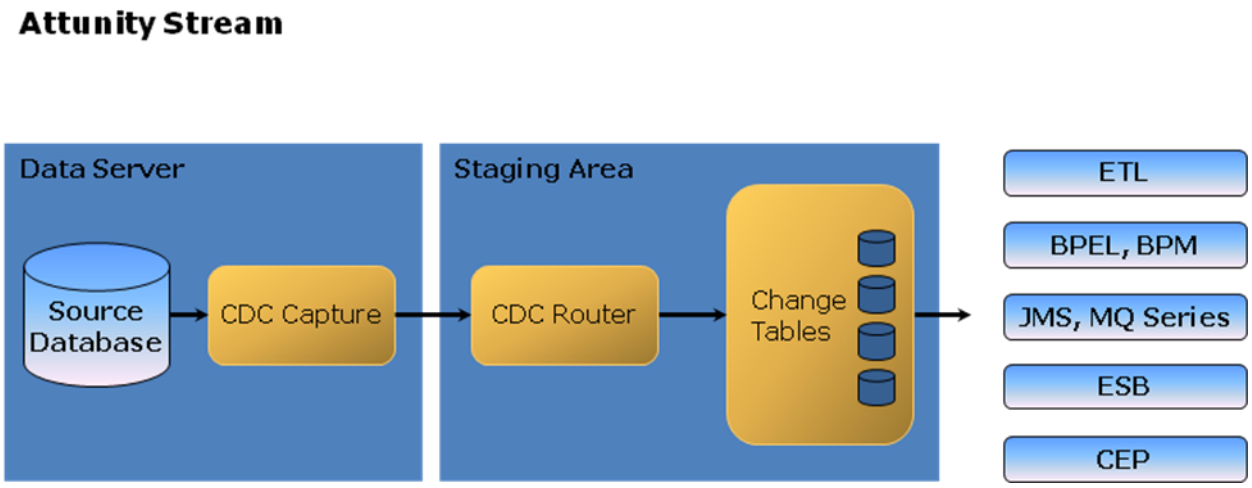
	Selection Criteria	Attunity Stream Features and Capabilities
1.	**CDC to Many Data Sources, Cross Platform**	Support over 10 data sources, relational and non-relational, on many enterprise platforms including: • Adabas (Mainframe, Open Systems) • VSAM (CICS and batch) • DB2 Mainframe • DB2/400 on IBM iSeries (AS/400) • Enscribe on HP NonStop (tandem) • IMS (online and batch) • Oracle 9i, 10g, 11g • SQL/MP on HP NonStop (tandem) • SQL Server 2000 and 2005
2.	**Non-Intrusive CDC**	Minimal intrusion is a key design guideline in Attunity Stream. Logs and journals as the primary source for change capture.
3.	**Change Filtering**	Attunity Stream provides a robust filtering service, allowing filtering changes on the data server for optimal efficiency. The solution supports: • Filter tables • Filter fields • Filter operations (insert, update, delete) • Filter on specific field changes • Filter on specific field content
4.	**Batch and Near Real-Time Delivery**	Attunity Stream provides flexible change delivery models that support both batch and real-time approaches. In batch, Attunity Stream enables an application to get all the accumulated changes. In real-time, Attunity Stream can push one or more change messages as soon as they occur.

5.	**Handle Non-Relational Data**	The Attunity Integration Suite is unique in its ability to handle non-relational data, providing a relational representation with support to various SQL interfaces. Attunity Stream automatically generates a relational schema for non-relational records (e.g. VSAM) such that SQL consumers such as ETL tools can process the information 'table by table'. The mapping maintains the hierarchical relationships by automatically adding fields that maintain the parent-child relationships. Attunity Stream can automatically generate an XML schema and provide the legacy data as an XML document.
6.	**Guaranteed Delivery**	Attunity Stream provides intelligent mechanisms for guaranteed delivery. It moves changes while persisting the 'position' of the last successful delivery.
7.	**Interoperability with Leading ETL/EAI Tools**	Attunity Stream has an open architecture and it supports many standard interfaces that provides seamless interoperability with ETL and EAI tools, as well as with many other enterprise applications. For SQL – Attunity Stream provides ODBC, OLEDB, ADO, ADO.NET and JDBC clients. For XML – Attunity Stream supports .NET, JCA, and 3GL clients. In addition, Attunity Stream has been tested and proven with leading ETL tools, including: • Microsoft SQL Server Integration Services (SSIS) • BusinessObjects Data Integrator • IBM WebSphere DataStage • Oracle Data Integrator • Talend • Cognos DecisionStream • Informatica PowerCenter • SAS Data Integration Server • OpenText/Hummingbird Genio • Embarcadero DT/Studio Attunity Stream also provides dedicated interfaces to SOA and BPM tools including: • Microsoft BizTalk Server • Oracle SOA Suite (ESB, BPEL)

8.	**Recoverability**	Attunity Stream is fault resilient and provides robust functions for restart and recovery. At all times, it maintains the last 'position' of changes that are retrieved from the source, such that it can always restart smoothly. In addition, it enables change consumers to 'reposition' the stream of changes as part of a system recovery process.
9.	**Performance and Throughput**	Attunity Stream provides two architectures for CDC, for optimal performance in a SQL environment (ETL) and an EAI environment (XML events). In addition, it provides binary-xml technology that reduces network traffic for increased throughput. At customer sites, Attunity Stream was proven to finish data movement processes at 500% faster then alternative products.
10.	**Ease of Use**	Attunity Stream solutions are configured using the Attunity Studio, a wizard-driven and task-oriented graphical user interface. With an integrated environment, the user can quickly set up solutions that span a distributed environment, including data source metadata, data access, change capture.

Solution Components

The following diagram provides a high level architecture of the Attunity Stream solution and its components:



The key components in the solution include:

Change Capture Agents

Attunity Change Capture Agents are 'live' software components that are responsible for the identification and capture of changes to operational data stores. Attunity offers change capture agents for many data sources on many enterprise platforms.

Staging Area

The CDC staging area improves performance, flexibility and recoverability by providing a place to store changed data and to apply intelligent filters and services. The staging area makes it easy to support multiple change consumers and to control the lifecycle of changed data.

CDC Consumer Interfaces

The CDC consumer interfaces allow ETL, EAI, and homegrown applications to easily consume changed records. Using standards-based APIs, applications can poll changes using SQL, or listen and wait for changes using XML-based messaging.

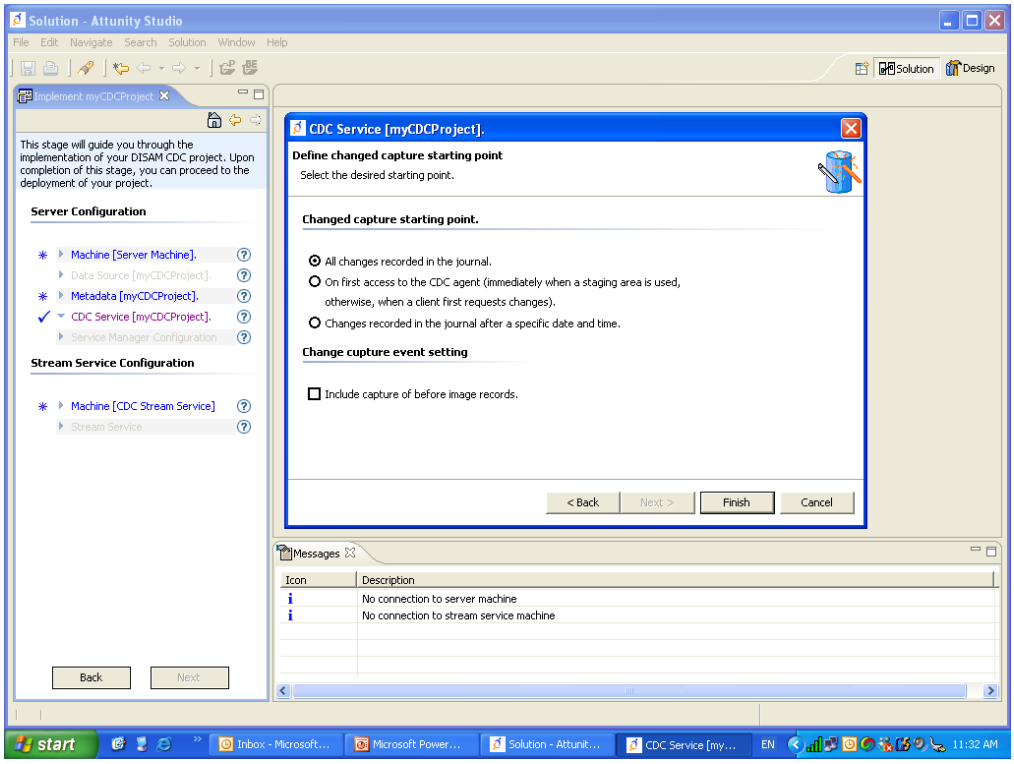
Metadata Management

Attunity provides metadata management that allows you to import and control the data models of the data being captured. For non-relational data sources, Attunity further enables you to define mappings to an enhanced relational data model.

Attunity Studio

The Attunity Studio is a GUI for configuring and managing the entire Attunity product line. Graphical and wizard-based, the Attunity Studio provides a productive environment for defining metadata, configuring adapters, deploying the solution, and managing it in production. The Attunity Studio provides a single console to any Attunity Server installed throughout the enterprise.

The following snapshot illustrates the Attunity Stream solution guide and configuration wizards:



About Attunity

Attunity is a leading provider of real-time event capture and data integration software. Using our software solutions, Attunity's customers enjoy dramatic business benefits by driving down the cost of managing their operational systems, creating flexible, service-based architectures for increased business agility, and by detecting critical actionable business events, as they happen, for faster business execution.

Attunity has supplied innovative software solutions to its enterprise-class customers for nearly 20 years and has successful deployments at thousands of organizations worldwide. Attunity provides software directly and indirectly through a number of strategic and OEM agreements with partners such as Microsoft, Oracle, IBM, HP and SAP/Business Objects. Headquartered in Boston, Attunity serves its customers via offices in North America, Europe, and Asia Pacific and through a network of local partners. For more information, please visit www.attunity.com or email info@attunity.com.