DATAWATCH

The Self-Service Data Preparation eBook

Brought to you by Datawatch

DATAWATCH

Chapter 1:

The 5 Overlooked Requirements for Self-Service Data Preparation

1 Access to Multi-Structured and Streaming Data

Everyone can connect to relational data, CSV and other standard data you'd expect. But often the most important information you need is locked away in multi-structured documents and sources that seem impossible to use without rekeying the data. You may have production reports or third party documents that would add tremendous analytic value if only there were an easier way.

Datawatch has the unique capability to acquire and prepare data from multi-structured documents like PDF reports, log files, telemetry data, print spools, text, EDI, HTML, and more. In addition, we provide access to over 25 relational sources, Hadoop, NoSQL and streaming data sources.

2 Data Masking

The most common cause of data breaches comes from internal employees. Data discovery tools are a great way for users to build and share information, but many times the underlying data is not protected and includes personally identifiable data (like Social Security numbers), personal sensitive data (like medical procedures) or commercially sensitive data. There are a number of industry and government regulations where non-compliance has an estimated cost of $5.5 million per breach and can even result in legal action against you and your organization.

Datawatch provides data masking which hides, or obfuscates the original data with random characters or other data but allows it to still be usable for analytics. And the data can be un-masked by authorized users. Think of this as your "Get Out of Jail Free" card.

3 Automated Processes

Every time you prepare data, you should look to automate the data preparation routines whenever possible. And you should share your work with others in your organization so you can have more consistent data and be more productive.

Datawatch creates a reusable load plan every time you prepare data. You can then fully automate this process either based on a scheduled time, or when the source data changes or becomes available (for example, when you are using a report file as a data source, Datawatch "listens" to a folder and runs the process when the new report file is available). With an intuitive visual process flow interface, you can easily deliver the prepared data to specific users based on roles or move it into other systems including data warehouses or over 25+ relational sources. You can also share your preparation models with other users via the Datawatch Portal.

4 Integrated Data Prep and Data Discovery

Data preparation is usually a highly iterative process where you are constantly moving data into a visualization tool only to realize that you need to make some additional changes to the data. So you go back to your preparation tool, make the changes, and bring it back to visualize.

With Datawatch, you can simultaneously prepare and visualize data side-by-side to get the desired results quickly. Our visual discovery tool has built-in data preparation capabilities and you can open both the preparation canvas as well as the visualization canvas and see how your results look with every step of the process.

Reduced Risk with Governance

The move to self-service is all about speed and agility for the business user. But IT has given up some control and with that comes increased risk. You need to introduce some governance but do it in a frictionless way that also accounts for the fact that much of the data preparation is against non-managed data sources—like CSV extracts, PDF reports or third party data.

Datawatch provides an enterprise solution to securely store, manage and control access to:

- Source content used by analysts that does not reside in a database
- Prepared data sets
- Reusable models for data extraction and preparation routines
- Visualizations and dashboards

You have a complete audit log of document versions, changes made, and usage. You also have complete data lineage from within the visualization and can drill right down to the line in the source report if you are using something like a PDF report as a source. And all of this can be initiated in a self-service manner using a Web based interface.

©2015 Datawatch Corporation. All rights reserved. Datawatch, the Datawatch logo and other product names, logos, and tag lines are trademarks of Datawatch Corporation. Part ID: 1508MY_US

DATAWATCH

Chapter 2:

Bridge the Gap Between Business Agility and Governance

Self-service data preparation tools are rapidly being recognized as a necessary element to any data discovery or advanced analytics implementation. At Datawatch, we've been delivering this capability for twenty years and today it's in use at 93 of the Fortune 100. The Datawatch Managed Analytics Platform is an enterprise solution that bridges the gap between the ease-of-use and agility that business users demand together with the scalability, automation and governance needed by IT.

Bridge the Gap Between Self-Service Data Preparation and Enterprise Governance

Most organizations have well-formed strategies to govern data that reside in managed systems like enterprise applications or data warehouses.

One of the biggest benefits of self-service analytics lies in the ability to rapidly combine and analyze data from a variety of sources. However this approach also provides a serious governance challenge in that it is estimated that half of this data comes from sources not typically managed by IT. For example, analysts will pull from sources including CSV or text extracts from transactional systems; personal spreadsheets; third party reports; semi-structured content; and more. Issues then arise around version control, data breaches, reconciliation, auditing and more. What is required is an enterprise data preparation platform that can address these governance risks but do so in a frictionless manner so the business user has the speed and agility they require.

Governing "Non-Managed" Data

The Datawatch Managed Analytics Platform provides an enterprise solution for self-service data preparation. The Datawatch Server provides a content repository to securely store, manage and control access to:

- Source content that gets utilized by analysts that does not reside in a database like XLS/CSV extracts, PDF documents, log files
- Reusable models for data extraction and preparation routines
- Prepared data for portal based access
- Visualizations and dashboards created

This content repository provides the foundation for governing "non-managed" data. Layered on top of the repository are key governance capabilities including:

Data Retention – At a minimum you should have document version control for consistency, but you may also need to persist and archive the source data/documents to meet regulatory or business requirements.

Data Masking – The most common cause of data breaches comes from internal employees. Data discovery tools are a great way for users to build and share information, but many times the underlying data is not protected and includes personally identifiable data (like Social Security numbers), personal sensitive data (like medical procedures) or commercially sensitive data. There are a number of industry and government regulations where non-compliance has an estimated cost of $5.5 million per breach and can even result in legal action against you and your organization. Datawatch provides data masking which hides, or obfuscates the original data with random characters or other data but is still usable for analytics. And the data can be un-masked by authorized users.

Data Lineage – Because you have a repository for the underlying source content, you can then provide complete data lineage and drill right down to the cell level of the source document. This capability is critical when you need to audit or reconcile data.

Data Curation – Frequently used data sources or automated data preparation routines can be shared with others based on user roles. This way you instill confidence and consistency in the data used to make important business decisions.

Role-based Access – Prepared data sets can be segmented based on user roles to ensure that the right subset of data is delivered to authorized users.

Auditing – The Datawatch Server provides complete audit logging and reporting.

DATAWATCH

Chapter 3:

The Value of Extending Self-Service Data Preparation Through Automation

As an analyst, you know how difficult it can be to access, prepare and combine data from a variety of sources. But the real value from data discovery and advanced analytic tools comes only when you bring together the right information, in a timely and trusted manner.

You know you need a self-service tool that will empower you to do this quickly, so you can spend less time preparing data and more time analyzing it. The value you bring to your team is in the insights you provide, not the mundane time spent in preparation. And you're not the only person in your organization struggling with this task. Adding automation to your data preparation drives significant time and cost savings while still preserving the agility that you need.

Datawatch Monarch is the most widely used self-service data preparation solution in the market and has been for over the last 20 years. Today it is the trusted solution used by 93 of the Fortune 100 organizations. Building on our proven technology and decades of experience, the Datawatch Managed Analytics Platform can help you in the following ways:

Self-Service Automation

It starts with self-service data preparation being performed by an analyst without the need for IT intervention. Every step of the data preparation process is captured in reusable and human-readable workspaces that can be executed automatically going forward. Using a simple visual process designer, you decide when the process should start and where the prepared data should be delivered.

Don't Start From Scratch

Frequently used data sources or automated data preparation routines can easily be shared and discovered with other users. This way you won't be recreating the wheel – it already exists. Or you can take another trusted process or prepared data set and simply augment this data set for your own use.

Deliver the Right Data to Different Users

Prepared data sets can be segmented for individual users (like John or Mary) or for specific roles (like North America Sales or Europe Sales). With a single automated data preparation process you can deliver the right data to the right users every time. You can even mask or redact specific data fields for certain users. For example, if you are in a regulated industry like healthcare, you need to mask personally identifiable patient information but still need to roll up the data by individual patient using a pseudonym. And you may need to redact the social security number completely for anyone but a hospital administrator role. You now have the power to do this in a fully automated and highly governed way.

Update Your Warehouse Too

Prepared data sets can also be delivered to other systems like data warehouses or departmental data marts. The new data can be added to existing systems easily using industry standard database drivers to 25+ different systems.

Real-time Refresh

When new source data becomes available, your data preparation process can automatically be invoked. For example, if you use data from customer invoices, you can direct Datawatch to "listen" for any new invoices in a number of places, such as a content management system, or a shared directory or even your email inbox. This way you'll always have the most up to date information.

Instill Confidence and Trust

Business critical decisions are made every day based on the information you've prepared. Users need transparency on how the data was prepared and they can see by looking at the workspace you have shared. You can also create sanctioned data sets and processes that must be used by others if you need that level of control. And for those occasions where users are reviewing a dashboard and question "where did that outlier come from?" — you can provide complete data lineage and even allow them to drill right down to the underlying source of data such as a PDF of a customer invoice where the row of source data is highlighted.

About Datawatch

The Datawatch Managed Analytics Platform is an enterprise solution that bridges the gap between the ease-of-use and agility that business users demand, together with the scalability, automation and governance needed by IT.

Datawatch data preparation capabilities are available as a stand-alone offering for use with any third party analytic front-end, to load into a data warehouse/data mart, as well as tightly coupled with our visual authoring tool for quickly building visualization applications.

©2015 Datawatch Corporation. All rights reserved. Datawatch, the Datawatch logo and other product names, logos, and tag lines are trademarks of Datawatch Corporation. Part ID: 1508MY_US