

Welcome to BLM's Data Tool Quality Presentation. I am Teresa Alred, the Data Quality Tool Manager. Today I'm going to give you a little bit of history and overview of BLM's Data Quality Tool.

BLM's Data Management Group in Denver began using software to analyze data seven years ago. The tool used at that time was good, but we soon realized we needed something better.

We built our requirements and researched products. The tool is now used not only by Data Management Group but throughout the Bureau of Land Management. We have trained users for over a year now with great success.

In improving your data quality, BLM's Data Quality Tool is a software package that enables you to discover what kind of data is actually in your database. It helps you reduce redundancies and incomplete data. It validates name and addresses, builds business rules, and develops standardized data.

When you leave this training, these are the objectives you should take away with you:

- Understand how to use the Data Quality Tool;

- Access your data;

- Know how to profile and look at measurements;

- Know how to validate addresses;

- Improve your overall data quality;

And then it will prepare you for the hands-on Data Quality Tool training in the future.

What is the Data Quality Tool? The tool assists in doing data analysis and discovery. The tool will help you find problems with your data. It goes beyond SQL searches. For example, the tool allows you to see if there's spaces in front of your data, or many spaces in front of your data, finds quotes, nulls and zeros where there shouldn't be. It connects directly to Oracle, Informix, and Access or flat files. It supports the Information Data Quality Act Section 515.

Now let's talk about data profiling and how the Data Quality Tool works to support that. What is data profiling? Data profiling assists in finding essential features of your data; assists in breaking down components and seeing what your data really looks like; displays measurements and patterns and how to discover those patterns and frequencies; looks at table comparisons.

Now we're looking at tool measurements. Data profiling measures your data for minimum values that start with zero or special characters and maximum values

that start with the last letter in the alphabet. At the top here you're going to see what some of those measurements are. They're null counts, blank counts, minimum value, maximum value, unique count and uniqueness. Those are just some of the measurements within the tool. Down below you're going to see column profiling, frequency distribution, pattern frequency distribution, percentiles, outliers and notes. Here you're going to see the values that are in the address and the counts. These are addresses within the database and it would be very hard to find these with an SQL query. Here you're going to see quotes and double quotes. These would be very hard to find with a query, but you could do this very quickly with this tool. You would not be able to know if there are one or ten spaces, or if there are question marks or quotes or asterisks or pound signs very quickly. This tool immediately finds them and displays them for you.

Now we're going to look at patterns. Patterns are used in different ways that the data can display like phone numbers. Here you're going to look at the phone field, and there are 9's that represent number values and A's that represent the actual character values. And here are all the different ways that patterns are displayed down below.

Now we're going to talk about tool drill down. The drill down tool allows you to drill down into values and find the whole records and counts with that value. Here we're looking at the address field. The tool assists you to find names with the same addresses and different phone numbers, as you can see above here. This will assist you to find any missing or incomplete data as you will see in the city column or state column. It also allows you to export into an Excel or text file spreadsheet.

Table comparison allows you to look at two tables with like values and compare them. Here we're looking at contacts table and sales table, the like values address. You can also drill down into that information once looking at the common information below. This allows you to look at values and display the overlapping values in both. So here you're looking at contacts and sales and both are common. There are 6,626 that are common in both. Here you're going to see that you can also export or see these in Excel spreadsheet or flat file.

What are we doing currently with the Data Quality Tool? Well, CPS has cleansed 1.5 million name and address records down to 85,000 records, preparing for the FBMS conversion. Currently has implemented this conversion and are monitoring and maintaining new customer addresses daily with the Data Flux web service. Wild Horse and Burros is cleansing their data and have developed a web service for name and address and is currently sharing this web

service with all of BLM. Property and Fleet has prepared for FBMS conversion and has standardized 54,000 records for manufacture names where no standards had previously existed. LR-2000 Mining Materials and [?] have created business rules to implement into their applications.

In conclusion, improving your data quality includes: Reduced redundancy and improving your data quality and you learned how to validate name and address. This demo showed you how to unearth savings and benefits to BLM, and now you can see how improving your data quality will help BLM make better business decisions with your data.

Now let's walk through a live application. This will show you how easy it is to use the Data Quality Tool.

To find out more information, please contact our project management office.

Hello. This is Teresa Alred, and I'm giving you a live demonstration of the Data Flux Data Quality Tool, the DF Power Studio. This does not take the place of the DOI hands-on training but only to show you how easy it is to use this tool.

This is the front screen of the Data Flux Data Quality Tool. We're going to go to the Tools, and here you're going to see different tools used here, but we're going to go to Profile Configurator.

First, we will be demonstrating how connections are brought in with the tool using ODBC. This can be set up by you or a system administrator. Those are easily brought in by the tools ODBC data source administrator. You can set up your drivers. You can bring in Informix, Oracle, Access databases. There are many databases that you can use to be brought in and you can ask your system administrator by looking at the drivers.

We're going to cancel out of this and we're going to look at the list of databases that were already established. Here you're going to see database with tables listed here below. When we click on one of these table and highlight them, then you go into the information, the fields that are on the right side. We will begin the process to do data profiling next.

Now we've begun profiling by looking at your database tables. We picked the sales table and we're looking at address in sales table. Down below we're looking at frequency distribution. Here, as you can see, we have a list of values and the counts. So if we were to sort on this information, you're going to see here that there are double quotes. We found some addresses with double quotes in them. And there are 12 of those. There are null values. So there are some addresses that are absent or incomplete and they're missing information. We also have a lot of different addresses here. They could be the same. Here

you're looking at two addresses that one has a period after the "n" and it's spelled out "avenue" where the other one is not. So this can quickly find duplicate addresses.

We're going to go down and look at another field, the Contact field, and here you're going to see where there's null values again down below. In the frequency distribution, also "John Doe" was spelled two different ways. So this is going to quickly find this information for you. This is just a quick demonstration of how you can find redundancies in your data.

Now we're going to create an architect job. This is a job flow, and I will demonstrate how quick it is to validate names and addresses with the USPS postal codes. You go to Task, Base, Architect. Here you can see that you have different choices on your menu. We're going to go to Data Inputs and look at Data Source. We're going to pull this over here to the right. It is a job flow. We're going to look at that data source, which is actually your database. We're going to right click on Properties. We're going to find that database, and next the table that you want to look at, and say, OK. We're going to bring out over all the fields in that table and say OK. Now you've established a connection to your database and data sources.

Next we're going to look at the menu type here Enrichment. Go into Address Validation, pull over the job flow, connect the two together, right click on Properties again, make a Suggest. This will fill in the field types from the database and then also you want to come down to the bottom of the screen, pick everything with U.S. These are the U.S. Postal Service information that you're going to pull over here, and we're going to go all the way to U.S. Result Codes.

We're going to bring those over to the right. We're going to say OK. And we're going to preview this information first, and I got an error message. So we have to right click, go in here, into your Properties, go into Options. Let's take off Cast Compliance. That requires a special license that we don't have right now. We're going to preview that information again.

This will take a few minutes. This concludes the Data Quality Tool live demo. I would like to see you next at the DOI Data Quality Tool training.