IBM Blog Article

I wrote the following article for IBM's Big Data Hub blog.

_____

Avoiding the "Data Everglades" with IBM BigIntegrate and BigQuality

Remember the story of Juan Ponce de Leon, the 16th century Spanish conquistador who discovered Florida? According to legend, he was searching for the Fountain of Youth, a mythical spring with waters that could turn an old man young again.

In one version of the story, after landing on the Florida coast, Ponce de Leon and his crew moved inland and discovered Lake Okeechobee, the freshwater lake in the center of Florida's panhandle. They traveled across the lake on canoes, thinking it would be smooth sailing from there on. But they didn't know that Lake Okeechobee drains into the Everglades. When they reached the south end of the lake, Ponce de Leon and his crew suddenly found themselves bogged down in the cypress swamps, and had to turn back. They never found the Fountain of Youth.

Like Ponce De Leon, many data managers and data analysts start off with "clear data lakes" from various sources. But when they dump their data into a Hadoop or traditional database, those "data lakes" suddenly turn into a dense "swamp" of unstructured data. Data analysts often find themselves wading through the "Data Everglades," in search of the fabled "Fountain of Business Intelligence" that big data is supposed to provide.

Into the "Soup" of Hadoop

When they do a data dump into Hadoop, data managers typically do one of two things:

1. They take the "do nothing" approach and dump the unstructured data into the Hadoop database, figuring they will deal with data integration problems as they come up later.

2. After dumping the data, they use Master Data Management (MDM) to extract the data, run it through a cleanse/governance program, and then load it back into the database.

Data managers often get bogged down trying to clean up this "swamp" of data. It's a complex and time-consuming process that involves hand-coding applications to extract and transform certain kinds of data, cleansing "dirty data" from public sources such as social media, and constantly checking data quality.

BigIntegrate (BI) Ensures a Clean Data Lake

IBM's BigIntegrate/BigQuality (BI/BQ) data solutions package helps you avoid the "Data Everglades," by keeping your "data lakes" clear and transparent, and making them easier to navigate. You can design your transformation or cleansing once and run it on your Hadoop Cluster, or your traditional engine or database, instead of wasting valuable time with hand-coded integration and quality checking.

BigIntegrate includes high-performance native connectors that connect to a wide range of Hadoop and traditional enterprise data sources. Built on IBM's InfoSphere Information Server, BI offers a scalable platform to integrate, transform, govern, and deliver big data.

BI features hundreds of built-in transformation functions, allowing you to aggregate and deliver your data sources to any Hadoop or traditional database. It also provides tools for managing and monitoring runtime environments, to help keep your "data lakes" free of useless digital garbage. A simple, web-based dashboard and an information governance catalog help you understand and manage metadata that can choke your critical information like swamp algae.

BigQuality (BQ) serves up crystal-clear data

IBM BigQuality acts as an advanced filtration and delivery system for big data. It automates data analysis, and lets you analyze, cleanse, and monitor data in any place, and across systems. Pre-configured rules let you maintain data quality, and data monitoring tools help you to spot inconsistencies across records and tables. Also, BQ's automated discovery tools help you to quickly discover data insights, identify patterns and trends, and align quality indicators according to your business policy tastes.

Unlike Ponce De Leon, you don't have to venture into the murky "Data Everglades" to find that "Fountain of Business Intelligence" that will help your business to stay young and prosperous. IBM's BigIntegrate and BigQuality will help you to maintain "clear data lakes" – and it will be smoother sailing from there. *Click here to find out more.*