Big Data Revealed

RADAR FOR YOUR DATA LAKE

The Intelligent Catalog for Your Data Lake

Congratulations, you have a data lake, and you even have some successes using it

But your have heard of failures and you are beginning to understand the frustrations of others and your own

Business moves at a fast pace, but big data is shackled by IT backlogs responsible to create access views

This blocks using big data for the truly important things, the surprises that shape markets

Exactly when is BigDataRevealed needed?

Data Scientist and Analyst and their Utilization and Shortage

Sources of the ***Zettabyte*** of Data collected in your Data Lakes

Cost of reporting on wrong / skewed data

Data Scientist and Analyst and their Utilization and Shortage

Data Scientists / Analysts

By 2018, U.S. companies will be short 1.5 million managers able to make data-based decisions. McKinsey put the national gap in data scientists and others with deep analytical expertise at 140,000 to 190,000 people by 2017. Glassdoor survey found the national median to be about $113,000, nearly double the median for a regular programmer.

CrowdFlower found that 83 percent of respondents said there weren't enough data scientists to go around, an increase from 79 percent reported a year earlier. However, if you're working in a metro area, such as New York City or Silicon Valley, the starting salary for a data scientist can exceed $200,000, a Stanford University director recently told Bloomberg.

Too much data, too many sources

Top use cases include big data projects, real-time analytics for operational insights, centralized data acquisition or staging from other systems. Massive quantities of data are being moved into these environments and often include highly sensitive payment card data (PCI), personally identifiable information (PII), and protected health information (PHI) . Fear is not unreasonable; the risk is high given what cyber attackers are after and the extreme damages that may result from a successful data breach.
http://c.ymcdn.com/sites/www.issa.org/resource/resmgr/JournalPDFs/feature0216.pdf

Cost of skewed and dirty data

It is estimated that poor data quality costs US companies $600 billion per year. It isn't just the potential for serious mistakes that poor data engenders, but also the painstaking amount of time and human effort that it takes to fix this data TechRepublic. Since the 2007 financial crisis, global systematically important banks have been under close regulatory scrutiny with the **regulatory fines amounting to more than $275bn as of March, 2016**. (BCG Report, 2016)

Data Scientists and Analysts Waste Time finding Metadata and Cataloguing the Eco-system

The New York Times **estimates that data scientists spend between 50 and 80 percent of their time working on mundane "janitorial" tasks.** While these tasks are necessary, they take valuable time away from the processes that provide tangible business value. IBMbigdatahub

Big Data Revealed
RADAR FOR YOUR DATA LAKE

Your Data Lake

Without BigDataRevealed

Once data arrives in the big data environment, it is stripped of the identity information from other environments.

Worse, it is optimized for storage, meaning that if you don't carefully manage it, you may never find the data again in your data lake.

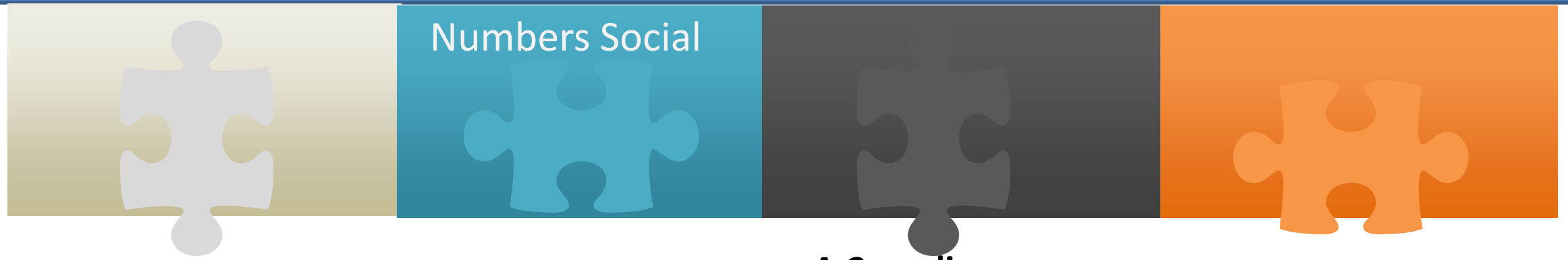
With BigDataRevealed

BigDataRevealed can accept, store and manage metadata created in other environments giving your Data Scientist an enormous head start. Non-technicians can add metadata to the environment.

For data without existing metadata, BigDataRevealed provides fully automated processes designed to identify critical data and assist in cataloging all other data.

Your have heard the horror stories of data scientists spending their time just searching for the information needed in their analysis. Let data scientists use their time productively so that managers can be armed with relevant, accurate information. We believe we have a better way, the BigDataRevealed intelligent catalog.

How your business benefits from the intelligent catalog



Those who need to eradicate PII

Legacy systems have stored information with a fair amount of creativity. Fields designed to hold comments may have been used for Credit Card or Social Security numbers, maybe medical diagnosis data or other personal identity information.

Once you replicate this data into Hadoop these fields become exposed / lost and potentially create huge liabilities for you and your company.

BigDataRevealed's extensive pattern matching capabilities scan all fields in all rows to identify these risks and provide a mechanism to locate the troublesome data..

A Data Scientist's Best Friend

Data Scientists require unambiguous clean data in their data lake if they are to create meaningful and reliable analytics.

BigDataRevealed processes every field in every row to insure data scientist know exactly what is contained in a file.

There is no possibility of confusion or of overlooking troublesome data that would result if only every 10th row were processed. Data Scientist's are freed to create their 'magic' by knowing exactly what is in their data lake.

A Compliance Officer's dream

Compliance is all about the identification of problematic patterns and repeat offenses of the same patterns. Quickly and completely findings these patterns is critical.

BigDataRevealed allows you to pre-schedule pattern detection processing so that no file will circumvent the process. The scheduled pattern matching capabilities of BigDataRevealed give the compliance officer the ability to track patterns and their results, Now and over time, delivering the Compliance Officer with a roadmap and complete view of the Compliance Officers potential nightmares.

A Product Manager's enabler

Product Managers are always looking for an edge in the market, whether it comes from social media, trade organizations or other sources as well of course from legacy data.

The ability to analyze patterns from incoming streams gives product managers the ability to make sense from the chaos that exists in the marketplace, and perhaps act on a pattern that much faster than their competitor thus decreasing the risk, harm and marketing nightmares of an unexpected, untimely hack.

Typical Big Data scenarios

Typically Late to Disruptions

Typical Intelligent Catalog Environments

Typical CAO Driven Quadrant

The Managed Swamp

For patterns that are known or relationships that are assured, the managed swamp can provide answers or identify cues that market sentiment is not supporting business as usual assumptions.

For PII and fraud analyses, known patterns can be researched to identify and remediate anomalies.

Typical Drivers

Data Scientist driven integration

Data Lineage

Information Usage Resistance

Market Sourced Data, usage organized

Using sentiment and proximity analysis of social media, blogs and other sources to determine whether the branding messages and promotional data is resonating in the marketplace.

Typical Drivers

Pattern based management

Centralization of metadata aligned to the data lake

Collaboration and Open Metadata Repository

Target Quadrant

Typical Mature Environment

The Data Swamp

Information, where ever it is sourced, is published to the big data environment for analysis. Thousands of tables make the data to big, each with a level of programming to make it appear as it were a traditional table with context provided by non-tabular information

Usage is sparse because no one can find anything in the data swamp due to poor folder naming standards and lack of File / Columnar Naming

Typical Drivers

Data Clutter

Cost

Complexity

IT

The opportunistic Swamp

High profile initiatives are destined to the big data environment, where IT gets involved to create views for long term opportunities.

The ability to use the big data environment for time sensitive uses is limited by the backlog and throughput of the team who creates views of the big data environment for consumption by non-technicians.

Typical Drivers

Internal/External Audits

Public Filings

Marketing Campaigns

Competitive Analysis

Opportunistic Quadrant

Typically Data Swamps

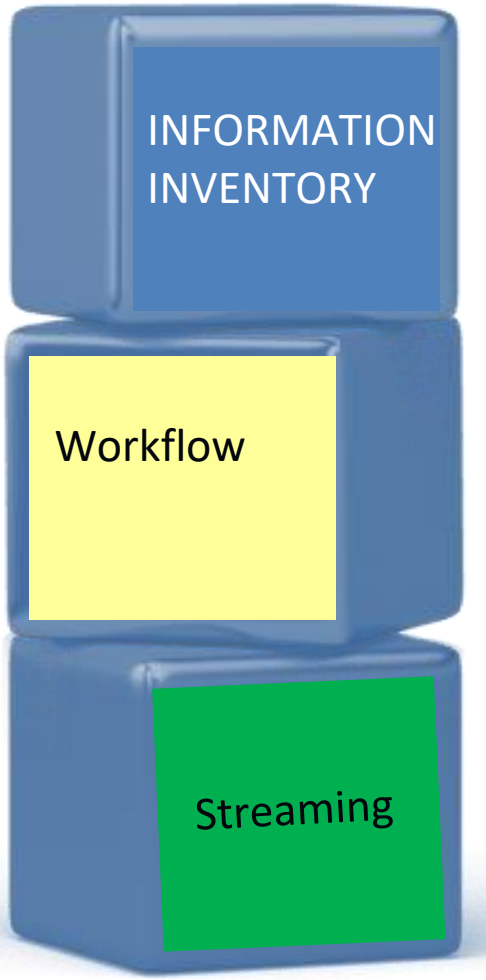
Thrives in disruption, but not stable markets

Which most typically represents your data lake?

Exactly when does BigDataRevealed help?

Your business climate changes regularly, so should your data lake

Information clutter without context is the biggest reason for the creation of data swamps



Information Inventory

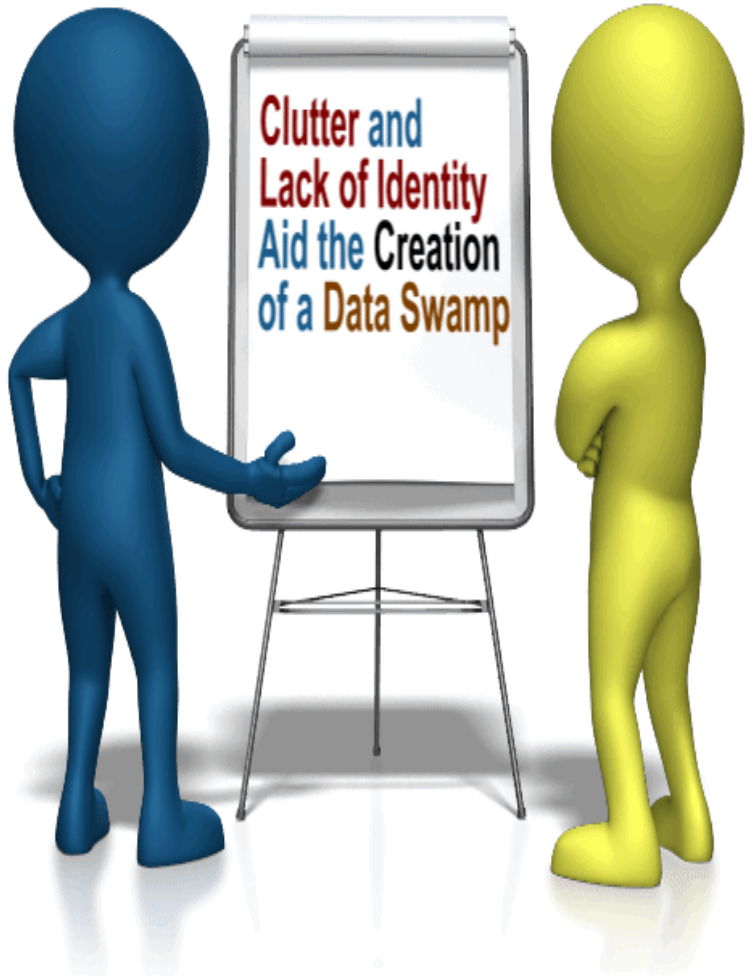
A dynamic information inventory, which constantly is in a state of flux to meet the current competitive landscape of the business community is required. BigDataRevealed, which provides an as exists catalog, with all the metadata and rules that apply to the information at hand. This prevents your dynamic data lake from becoming a swamp.

Workflow

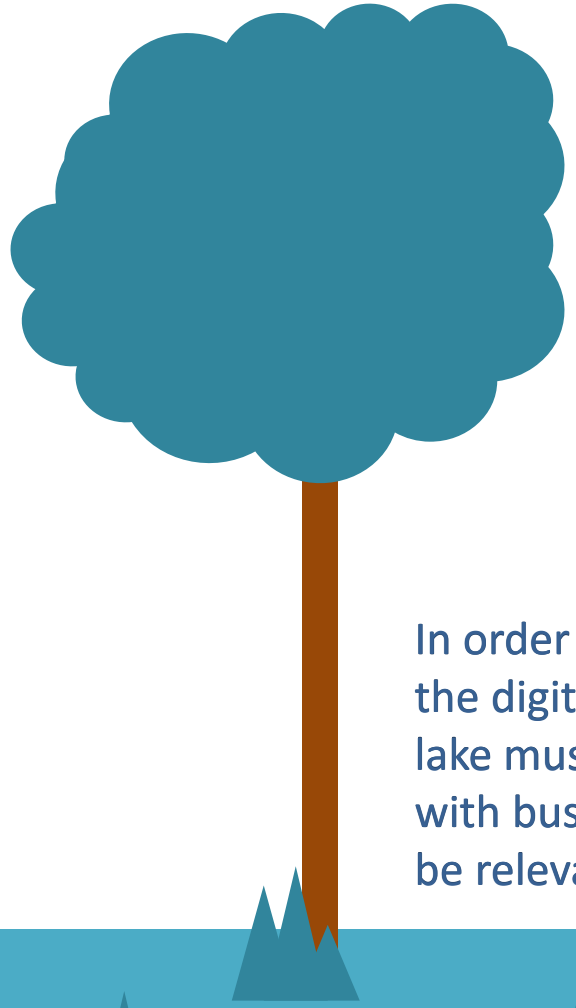
There will be items that require research. BigDataRevealed contains the workflow of items researched and an ability to build cases of repeat offenses so that questionable items are identified, eradicated, and do not become repeat offenses.

Streaming

You will have needs where the information that will make a difference is not housed in your environment and changes at a moments' notice. Inclusion of streams of data is mandatory, and the fact that BigDataRevealed is installed in the Hadoop environment means it will be able to keep up with the streams.



Getting your big data house in order



THE NEW ORDER

Because business process changes rapidly, the mapping of information will similarly change. This will require information models to be highly fluid & Metadata Driven.

CHANGE

Your data lake should be dynamic to support the current population of issues being tackled. BigDataRevealed's Intelligent catalog is the perfect solution for managing the dynamic properties of the data lake.

VISION

Algorithms serve as early warnings when processing information from outside the organization. They serve as triggers for action and patterns such as PII regulatory violations.

Information no longer mapped is clutter and will be either archived or de-emphasized or removed from the Data Lake.

CLUTTER

In order to participate in the digital age, the data lake must stay in lock step with business intents to be relevant and valuable.

MAPPING

The intelligent data catalog should be the foundation for your data lake as it ensures that your team can find data in a highly fluid data lake.

Big Data Revealed
RADAR FOR YOUR DATA LAKE

How to get started

Select and use a catalog that captures potential and actual use of information
(Note: The catalog must track, information, business models, processes, actors, resistance, etc.)

Share the BigDataRevealed Catalog with your third party Metadata / Catalog products for a more complete, current and accurate view of your data assets.

Information only has value if it is utilized to achieve a value proposition and valued end solution. This could be monetary, market share, loyalty, branding, auditable, predictable and easily accessible and understood.

Incentivize your CDO to relentlessly identify and preference uses of information that enhance organizational value
(Note: This is not a pitch to perform more analyses, produce more reports, nor store more information. If the information is not consumed in business processes, it generates cost but does not generate value.)

If the Integrity and accuracy and trustworthiness of your data is not gained, it is all for Naught.

The Data Warehousing Institute estimates that data quality problems cost U.S. businesses more than $600 billion a year

Assign a Data Analyst, Data Scientist along with a Privacy, Compliance and Risk Officer or Consultant and a Data Steward

What is the Intelligent Data Catalog?

Cataloguing and Metadata with History

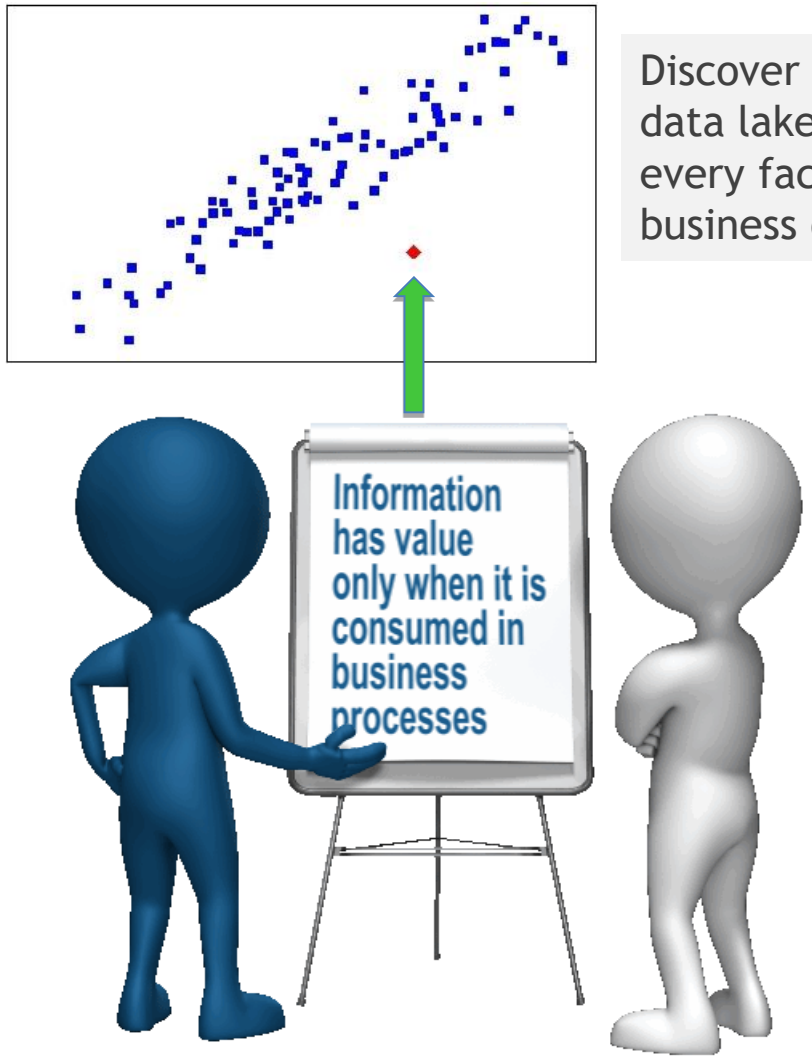
- Pattern Folder / File Column Locations
- Pattern(s) with % found
- Columnar Metadata naming with User Metadata
- Data Discovery of Data Completes and integrity
- Assist in originating Source Loading of Data
- Assist in determining masking and zone encryption

BigDataRevealed Hadoop Ecosystem Stored Catalog / Metadata

- Legacy
- IOT
- Third Party Data
- BI, Predictive Analytics, AI
- Lineage, Job Status by Users

Big Data Revealed
RADAR FOR YOUR DATA LAKE

Investigate and Discover Data



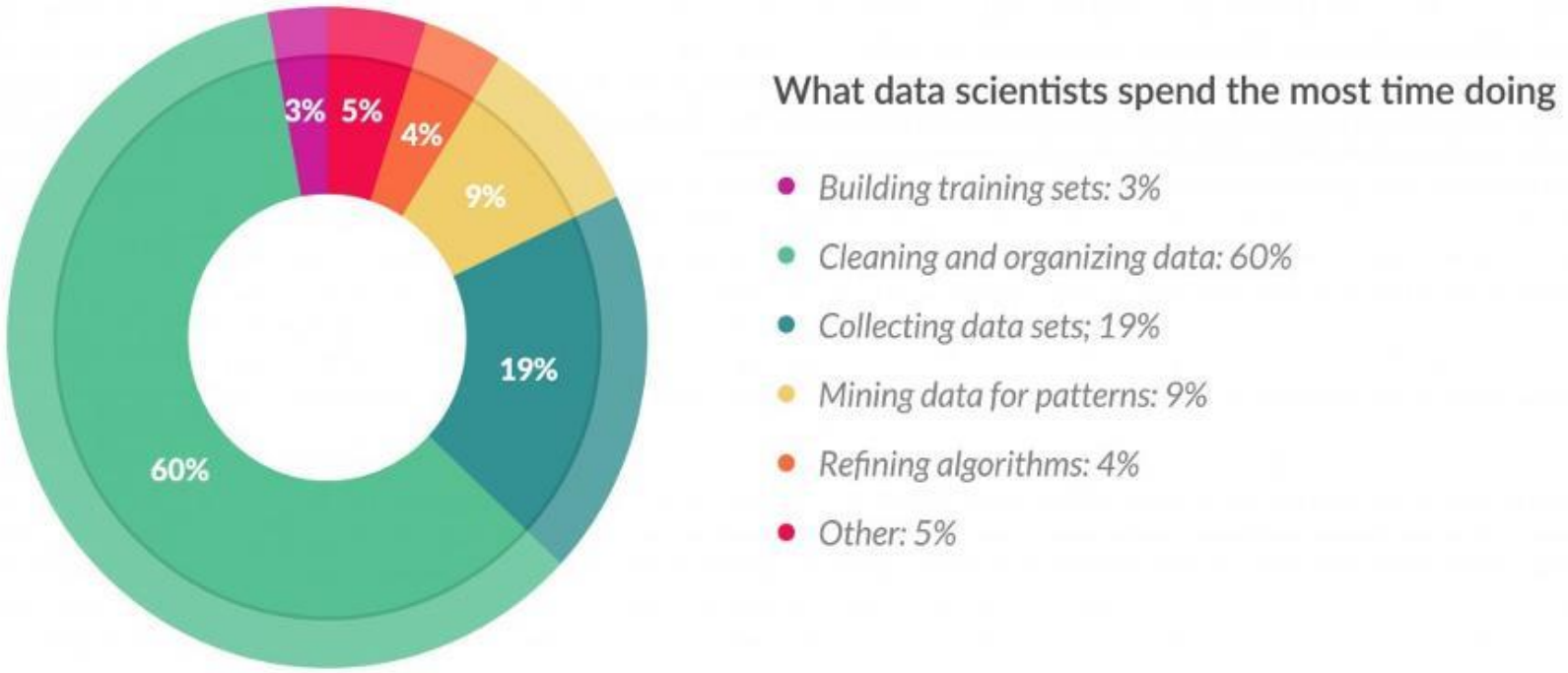
Discover patterns in the data lake that impact every facet of your business operations

There will be a shortage of talent necessary for organizations to take advantage of big data. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions. McKinsey & Company

Data preparation *accounts for about 80% of the work of data scientists*

The misalignment and clutter issues waste much of the precious time for critical decisions

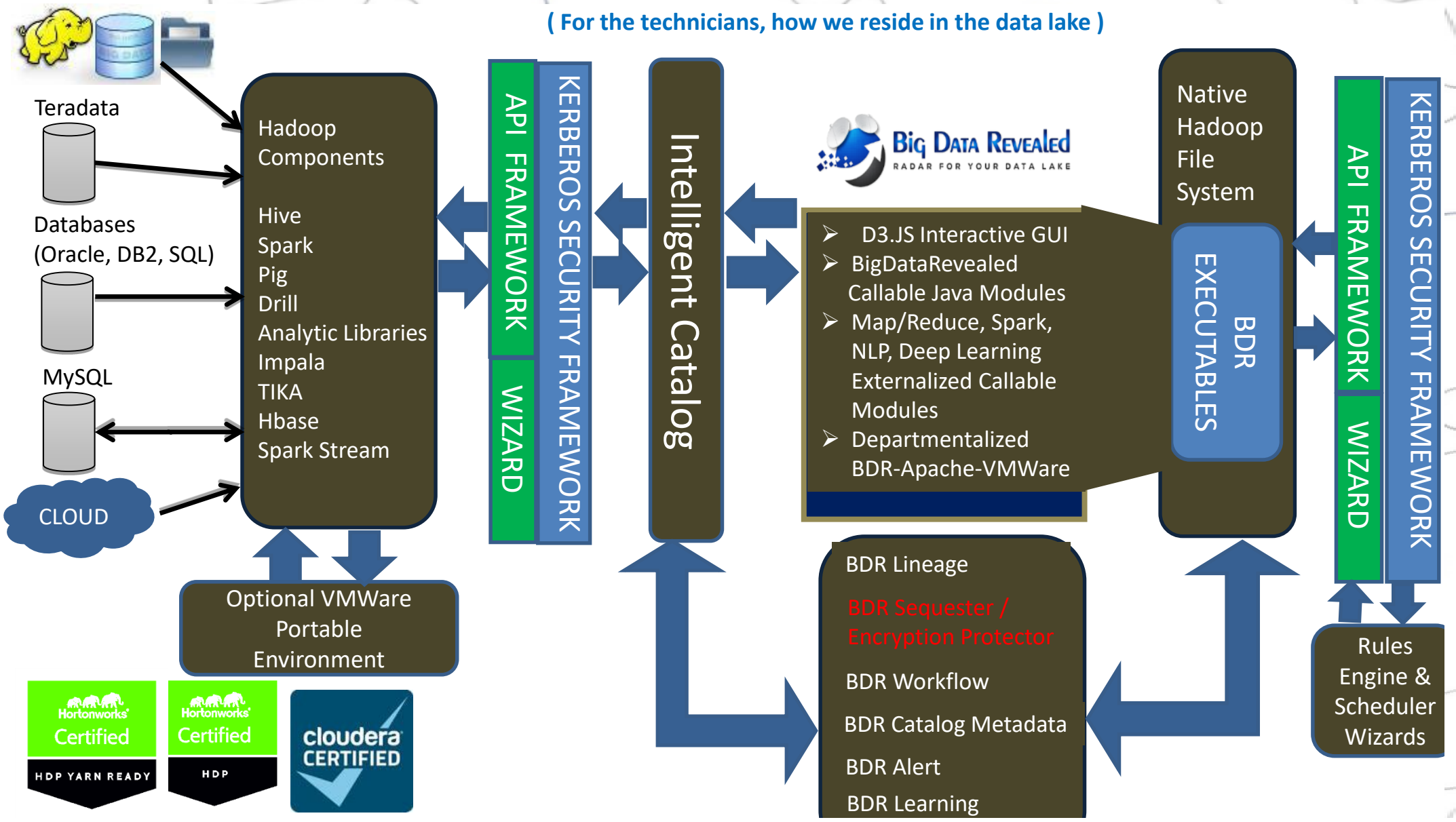
Need proof of the misalignment?



Data preparation accounts for about 80% of the work of data scientists

Data scientists spend 60% of their time on cleaning and organizing data. Collecting data sets comes second at 19% of their time, meaning data scientists spend around 80% of their time on preparing and managing data for analysis. [Forbes Article]

BDR Architecture - Powered with Apache™ Hadoop®



Big Data Revealed
RADAR FOR YOUR DATA LAKE

BigDataRevealed Data Discovery for Big Data Hadoop

BigDataRevealed Discovers and Isolates Personally Identifiable & Potentially Risky(Outlier/Anomaly) information/data in your Hadoop ecosystem!

Steven Meister
CTO & Founder, BigDataRevealed
steven.meister@bigdatarevealed.com
847-791-7838
www.bigdatarevealed.com