


IBM Watson Data Platform

Jay Limburn

Distinguished Engineer & Director of Offering Management

October, 2017

 @jaylimburn

Why do companies struggle to deliver value?

Data

- Data resides in silos
- Detailed data was never stored
- Unstructured and external data wasn't considered
- Difficult to access

Skills

- Data Science skills are in low supply and high demand
- Nurturing new data professionals is challenging

Governance

- If the data isn't secure, self-service isn't a reality
- Understanding lineage and getting to a system of truth

Infrastructure

- Need an environment that enables a “fail fast” approach
- Discrete tools present barriers to progress



Knowledge workers spend

80%

of their time searching for the correct data

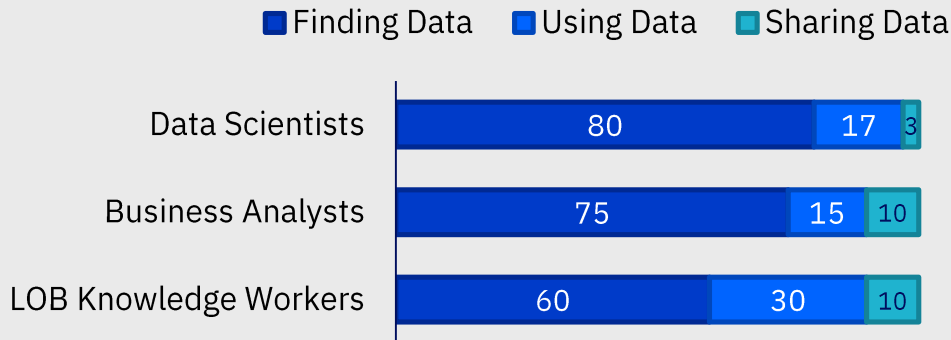
.....imagine if that time was spent doing Data Science and Analytics!

The Data Lake Fallacy

Enterprise Data Lakes are not delivering on their promise

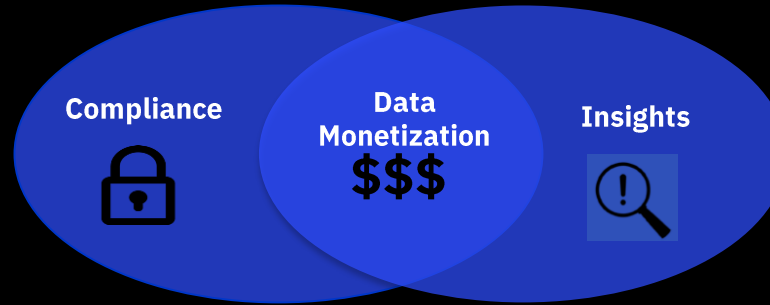
- Inability to easily find data
- Lack of trust in how the data will be used
- Obstacles to finding and sharing
- Too difficult to ingest sources

% Time spent working with data



Users spent significantly more time finding the correct data, rather than extracting value from it.

Chief Data Officers need answers!



WHERE is all my data?

Compliance

WHO is using it?

Compliance

HOW can I monetize it?

Insights

WHAT can I do to improve it?

Insights

Watson Data Platform

an integrated platform of tools, services and data that helps companies accelerate their shift to become data-driven organizations

Drive governance policy effectiveness while tracking how data is used and its value to the company

Access powerful tools to prepare data to tease out the insights they're looking for, without IT involvement

Data Steward

Data Scientist



Data Engineer

App Developer

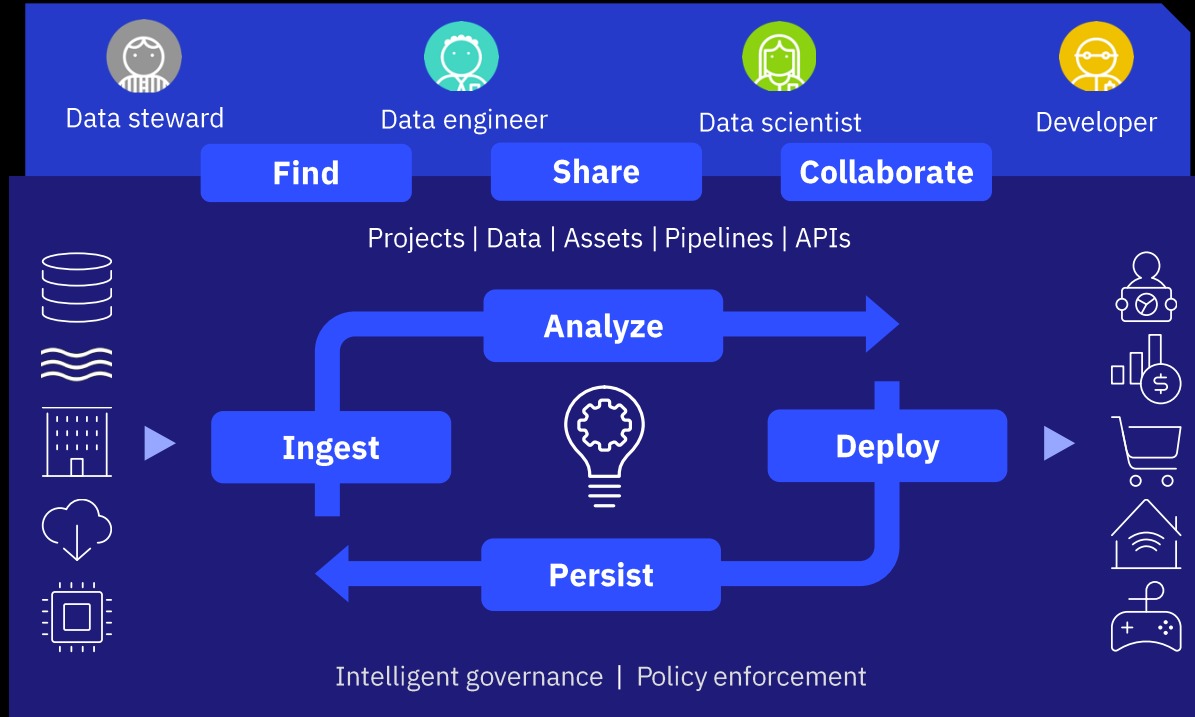
Easily build data pipelines that power dashboards and data platforms while ensuring high quality

Make the insights immediately actionable and add intelligence to apps in straightforward manner

The IBM foundation for data innovation



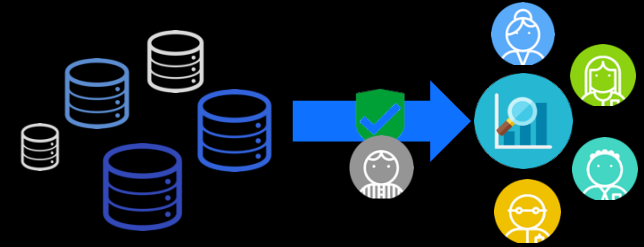
Making data into insights is a team sport



Our Core Tenets

1. Intelligent by Design
2. Collaborative for data Professionals
3. Self-service access to trusted data
4. Best in class streaming and real-time analytics
5. Open and Extensible

Democratized data and analytics



Projects and communities

enable team collaboration ensuring the enterprise builds upon the insights its members develop and springboards from those that peer organizations and experts share.

Common tools

including open source notebooks and tools, shapers and data visualizations that enable the team to quickly investigate any volume of data

Catalog

to organize, find and understand data that enables every project team to quickly discover and use data they can trust

Data access









to disparate data sources that makes it easy to connect to and gather data from anywhere without friction

Enterprise grade governance





























enabling regulatory compliance with automatic policy enforcement that enables data professionals to focus on delivering valuable insights rather than managing and organizing data

Data at your fingertips

In the IBM Cloud

 Cloudant NoSQL DB	 Compose for MongoDB
 Compose for PostgreSQL SQL	 Compose for Redis
 Db2 Warehouse on Cloud SQL	 Db2 on Cloud SQL
 Message Hub	 Streaming Analytics

Where your data is

IBM			
 IBM Informix	 PostgreSQL on Compose	 MySQL on Compose	 IBM Db2 for i
 IBM Cloudant	 IBM Cloud Object Storage	 Bluemix Object Storage	 IBM Db2
 IBM BigInsights HDFS	 IBM Db2 Hosted	 SoftLayer Object Storage	 IBM PureData for Analytics
 IBM Db2 for z/OS	 IBM Db2 Warehouse on Cloud		
Third Party			
 Cloudera Impala	 Salesforce.com	 Apache Hive	 Amazon Redshift
 Microsoft SQL Server	 Sybase IQ	 Sybase	 Oracle
 Amazon S3	 MySQL	 Hortonworks HDFS	 PostgreSQL
 Pivotal Greenplum	 Microsoft Azure SQL Database		

IBM Data Catalog Projects Catalog ^{IBM} Governance ^{IBM} Data Services US South Docs

Categories > Great Outdoors Catalog

Browse Usage Access control

Recently created assets Expand

Filter What are you looking for?

Showing 15 of 15 assets

NAME	OWNER	TAGS	TYPE	DATE ADDED
Retail Store PresenceData - Cloudant	Aksham Cloud Convergence	ret presence retail	Connection	Aug 6, 2017
Historical Weather Data - ObjectStore	Aksham Cloud Convergence	weather	Connection	Aug 6, 2017
AWS S3 Warehouse	Jay Limbun	S3 Sales Warehouse	Connection	Aug 30, 2017
Enterprise DataWarehouse - Clou	Aksham Cloud Convergence	Default... outbors	Connection	Aug 6, 2017
KOT Presence data from Retail stores	Aksham Cloud Convergence	KOT Presence retail	Data Asset	Aug 6, 2017
Product Line data	Aksham Cloud Convergence	product	Data Asset	Aug 6, 2017
Product Line lookup table	Aksham Cloud Convergence	outbors sales Datawar...	Data Asset	Aug 6, 2017
sales order dimension table	Aksham Cloud Convergence	outbors Datawa... sales	Data Asset	Aug 6, 2017
Product data	Aksham Cloud Convergence	product	Data Asset	Aug 6, 2017
Sales Fact table	Aksham Cloud Convergence	outbors sales	Data Asset	Aug 6, 2017
product survey	paal taylor		Data Asset	Aug 29, 2017
Q3 Sales Route Information	Jay Limbun	sales - q3	Data Asset	Aug 30, 2017
Employee Data	Jay Limbun	HR: Employee	Data Asset	Aug 16, 2017
forecast weather datafor-1	paal taylor	weather	Data Asset	Aug 18, 2017
statement fact	paal taylor		Data Asset	Aug 28, 2017

Unlock tribal knowledge to unleash your data professionals

Discover

Intelligent discovery of data, advanced classification and profiling to provide context

Catalog

A rich metadata index of all data, with social collaboration and enhanced findability

Govern

Powerful governance policy tools to control and protect access to data with visibility to data use

Active Draft Active

What policies, rules, and categories are you looking for?

Customer Information <p>Size: 20 Policies & 34 Rules</p> <p>Lucas ipsum dolor sit amet scyphus danti ubi- wen organa londo organa feli londo calissian barka, C-3p0 dooku ambla danti wedge mers. Calissian Moff ki londo yoda solo boba awen.</p>	Financial Statements and Records <p>Size: 20 Policies & 34 Rules</p> <p>Lucas ipsum dolor sit amet scyphus danti ubi- wen organa londo organa feli londo calissian barka, C-3p0 dooku ambla danti wedge mers.</p>
Information Disposal <p>Size: 20 Policies & 34 Rules</p> <p>Lucas ipsum dolor sit amet scyphus danti ubi- wen organa londo organa feli londo calissian barka, C-3p0 dooku ambla danti wedge mers. Calissian Moff ki londo yoda solo boba awen.</p>	Location Data <p>Size: 20 Policies & 34 Rules</p> <p>Lucas ipsum dolor sit amet scyphus danti ubi- wen organa londo organa feli londo calissian barka.</p>
Records & Retention - Internal Information	Security and Access Controls



Data Science Experience

```
# Load the historical Weather Data
weather = sqlContext.read.format('json').load(['swift://forecast.gowweather/'])
weather.registerTempTable('weatherdata')

# Create dataframe with data elements of interest for camping (TEMP, WIND, RAIN)
campweatherDF = sqlContext.sql('SELECT dailysummary.date, tenure as location, dailysummary.date, day,
date.day as day, dailysummary.meantemp as meantemp, dailysummary.meandepth as meandepth, daily
summary.rain, dailysummary.hail, dailysummary.snow from weatherdata')
campweatherDF.registerTempTable('campweatherdata')
campweatherDF.cache()

sqlContext.sql('SELECT * from campweatherdata').show(5)

+-----+
| location|year|month|day|meantemp|meandepth|precip|rain|hail|snow|
+-----+
| America/Los Angeles|2014| 04| 05| 91| 11| 1.86| 1| 0| 0|
| America/Washington|2014| 04| 05| 95| 18| 10.86| 1| 0| 0|
| America/New York|2014| 04| 05| 92| 11| 30.86| 0| 1| 1|
| America/Atlanta|2014| 04| 05| 42| 11| 6.86| 1| 0| 0|
| America/Bartford|2014| 04| 05| 44| 13| 9.86| 0| 0| 0|
+-----+
only showing top 5 rows

In [6]: # Categorize days by Temperature
daysByTempDF = sqlContext.sql('SELECT year, meantemp, CASE when meantemp <30 then 'COLD' \
when meantemp between 30 and 60 then 'COOL' when meantemp between 60 and 90 then 'WARM' \
when meantemp >90 then 'HOT' END as category, 1 as daycount from campweatherdata order by year, d
daysByTempDF.cache()

# Aggregate days categorized by Temperatures
aggDaysByTempCategory = daysByTempDF.groupBy(daysByTempDF.year, daysByTempDF.category).agg('count(*)'
as count), 'total_days')

# Visualize results
```

Ensemble Learning to Improve Machine...
AUTHOR: Stats and Bits | DATE: Aug 25, 2017
TOPIC: Machine Learning | FORWARD: Web page

Got zip code data? Prep it for analytics...
AUTHOR: Raj Bragh | DATE: Aug 21, 2017
TOPIC: Analytics 4 | FORWARD: Web page

How smart catalogs can turn the big data...
AUTHOR: Bluemix blog | DATE: Aug 23, 2017
TOPIC: Analytics v5 | FORWARD: Web page

Introduction to Neural Networks, Advantages...
AUTHOR: Jitendra Mahanta | DATE: Aug 02, 2017
TOPIC: Neural Networks | FORWARD: Web page

Select a technique
You cannot change label column, feature columns, model type, or validation split after adding an estimator.
You must first delete all estimators in order to make changes to these attributes.

Column value to predict (Label Col)
Item, Identifier

Feature columns
Item, Fat, Content, Item_Type, Item_Weight, Item_MRP

☒ Suggested technique.

- Binary Classification**
Classify new data into defined categories based on existing data. Choose if your label column contains two distinct categories.
- Multiclass Classification**
Classify new data into defined categories based on existing data. Choose if your label column contains a discrete number of categories.
- Regression**
Predict values from a continuous set of values. Choose if your label column contains a large number of values.

Validation Split
Train: 80 | Test: 20 | Holdout: 20

Making data science a team sport

Learn

Built-in learning to get started or go the distance with advanced tutorials

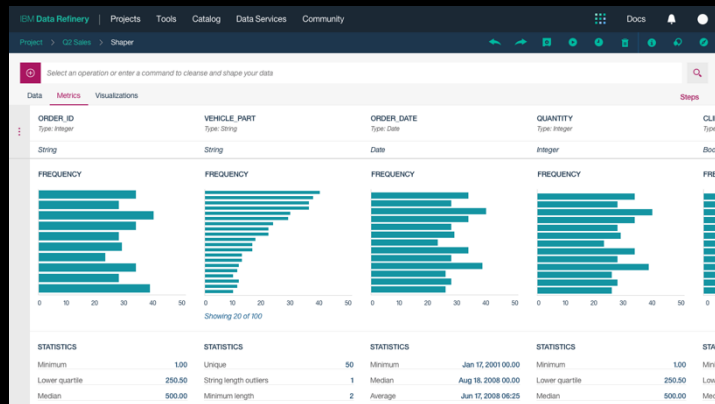
Create

The best of open source and IBM value-add to create state-of-the-art data products

Collaborate

Community and social features that provide meaningful collaboration

Data Refinery



A Breakthrough Approach to Explore and Prepare Data

Wrangle

Interactively explore, resolve quality issues, enrich, classify, standardize, summarize and join data

Flow

Create data flows visually, schedule for repeatability, monitor and notify

Adapt

Connect to 30+ cloud and on-premises stores and scale on demand with cataloging and governance

Project

Q1 Sales

Shaper

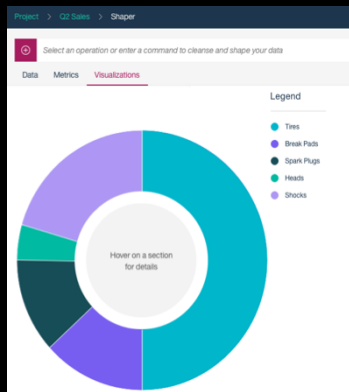
Select an operation or enter a command to cleanse and shape your data

Data

Metrics

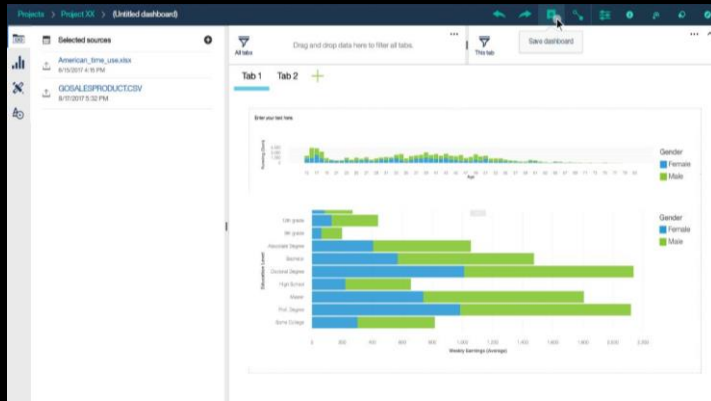
Visualizations

ORDER_ID	VEHICLE_PART	ORDER_DATE	QUANTITY	
Type: Integer	Type: String	Type: Date	Type: Integer	
String	String	Date	Integer	
1	77641	Break Pads	6/04/2017	10
2	77638	Break Pads	5/27/2017	-
3	77634	Break Pads	5/19/17	7
4	77632	Break Pads	5/04/2017	10
5	77631	Break Pads	5/1/2017	-
6	77628	Break Pads	4/29/2017	15
7	77625	Break Pads	4/27/2017	-
8	77622	Break Pads	4/24/17	14
9	77620	Break Pads	4/19/2017	11
10	77617	Break Pads	4/15/17	-
11	77615	Break Pads	4/11/2017	7
12	77611	Break Pads	3/29/2017	15
13	77632	Break Pads	4/28/17	27
14	77631	Break Pads	4/22/2017	9



Watson Data Platform

Dashboards



Making Insights Available to All

Visualize

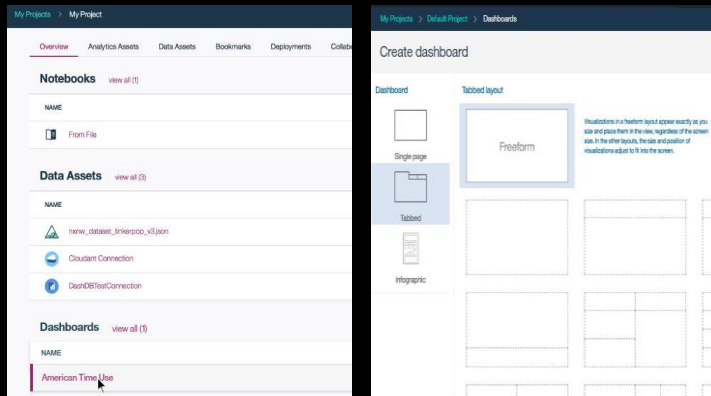
Represent analytic results as compelling interactive graphics.

Share

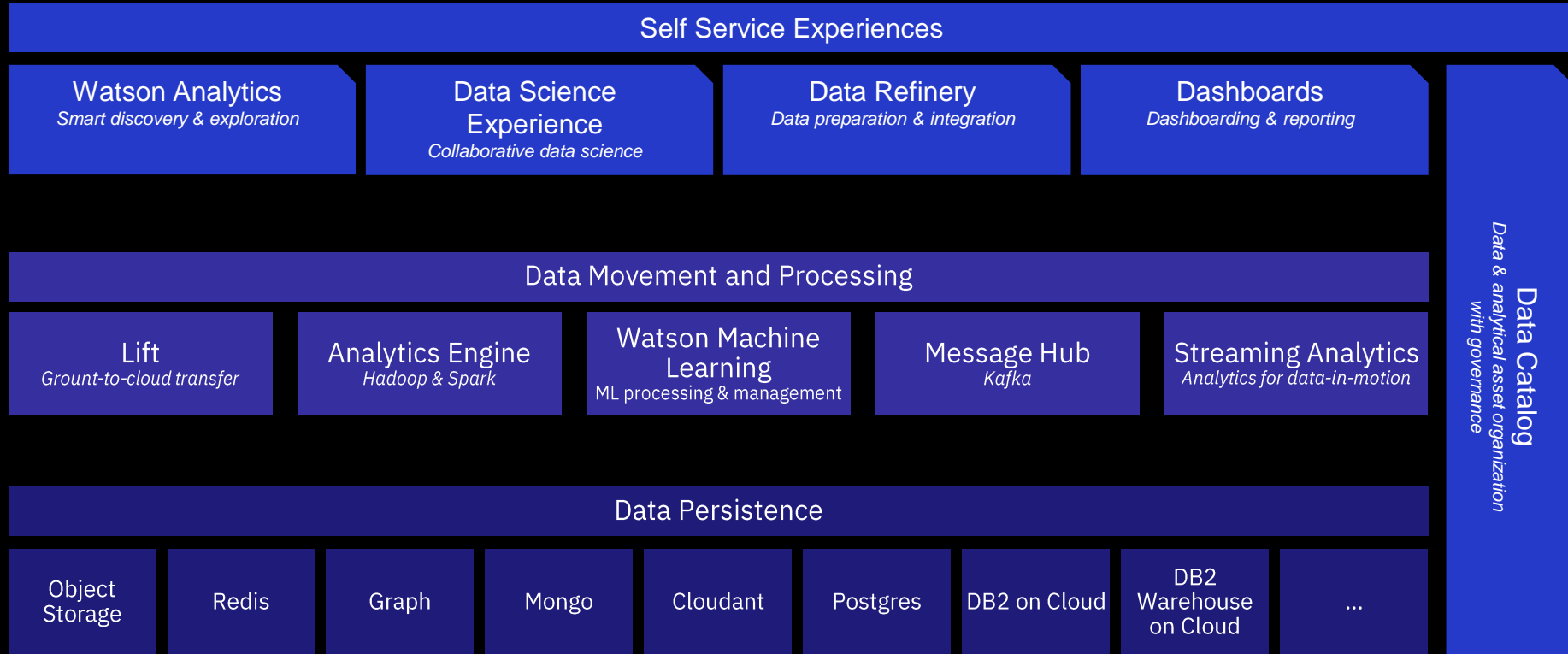
Deliver dashboards across the enterprise where analytical insights are required to make business decisions.

Discover

Utilize data assets from the entire organization with confidence by accessing them through the Data Catalog.



Services of the integrated platform



IBM Data Catalog
Powered by the Watson Data Platform

360 View of Data



Data Science

Allow Data Scientists to spend more time doing Data Science!

Business problem

Data Scientists can't easily locate the data they need. Projects typically start with an exhaustive effort to find locate and gain access to the data they need. What's worse is they don't even know what other data could help them find a different answer.

Solution

Watson Data Platform allows data to be easily indexed and classified. Intelligent search algorithms can quickly guide your Data Scientists to the best data for their purpose. Once found they can move the data into their sandboxes. Better still Data Scientists and share and collaborate on their models allowing for reuse and accelerating the initiation of new projects.



Compliance Office

Transform your business to become Data Driven

Business problem

Too much data, not enough use. How can the compliance office monetize their data, yet ensure it is protected and governed at the same time? The Compliance Team don't have the tools or understanding to drive the cultural change required to unlock their organizations value in their data.

Solution

Watson Data Platform provides an easy way to build and index of all the assets across your business. It builds an intelligent model for those assets and provides context around those asset to make them easy to find and understand how that data should be used, accessed and managed.



Business Analysis

Turn every worker into a knowledge worker that thinks about data first

Business problem

Tribal knowledge is locked away in department silos. Business analysts know this data is locked away but don't know how they can get hold of it to improve their analytics. They know their analytics are not as accurate as they need to be.

Solution

Watson Data Platform makes it easy for tribal knowledge to be shared and accessible across the enterprise. Every worker can easily collaborate and curate assets to aid the findability of data for all. It makes finding data a delightful online shopping experience for all and allows them to provide their own assets back into the marketplace.



Data Engineering

Optimize your Data Engineers by reducing duplication and providing self service access to data allowing them to focus on the most important jobs

Business problem

Your data engineers are working as hard as they can but it takes too long to provide data to the business when they request it. The complexities of moving, protecting and cleansing the data takes time and the business teams can't wait.

Solution

Watson Data platform provides self service access to data. The intelligent catalog allows all users to know what data exists and the tools to allow them to go and get it. The business users no longer need to wait on IT to get the data they need and the Data Engineers can focus on the business critical tasks essential for business growth.



Application Development

Supercharge your next generation application development

Business problem

Data is growing, changing, being duplicated, being moved. Application Developers add to this problem by creating new silos of data to power their applications. In addition they are unable to understand what data should be used to power their applications

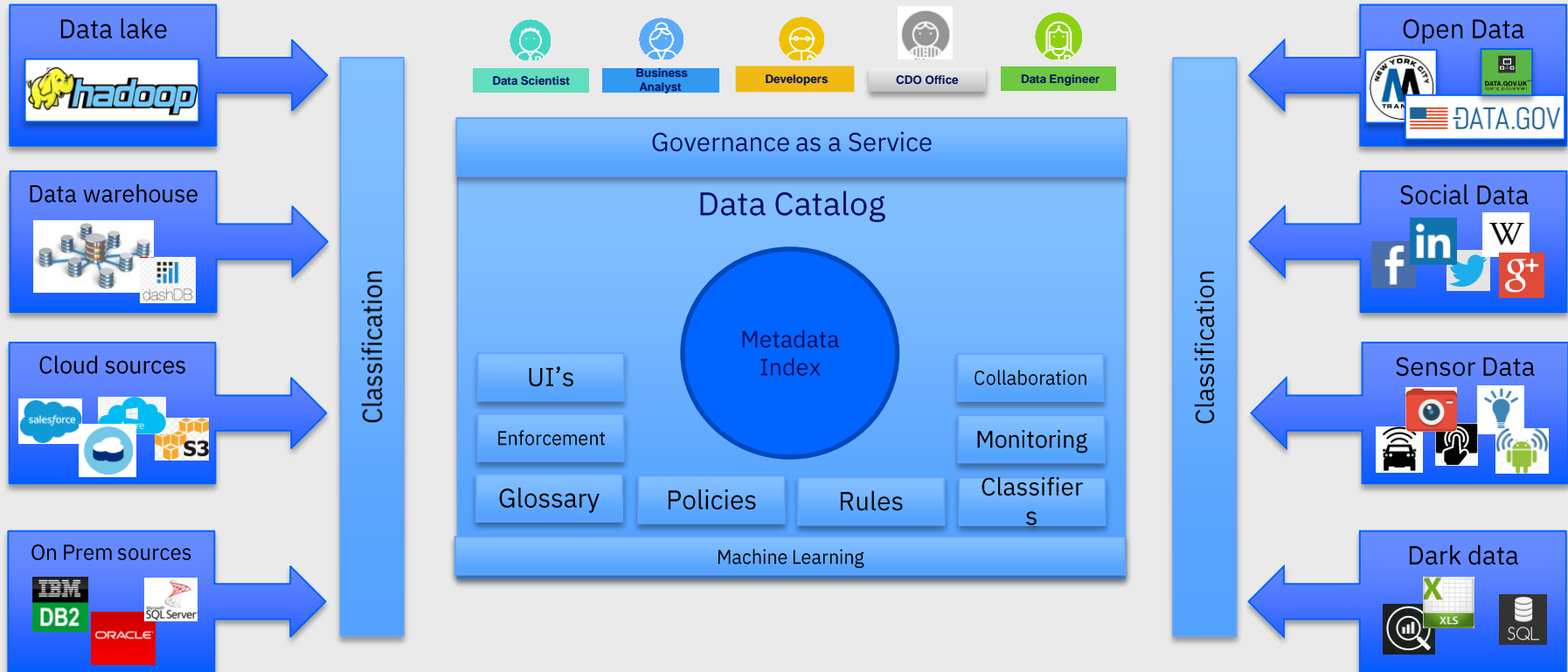
Solution

IBM Data Catalog provides an intelligent index of all data so that application developers can be guided towards the right assets for consumption within their next generation cognitive applications, but more importantly any new data is cataloged and made available to knowledge workers for further insights.



Intelligent Metadata Index

- Providing 360 view of **all** data for **all** users
- Data can reside in original systems, but users can discover it quicker join together for smarter analytics



Governance Framework

- Govern Data throughout its lifecycle
- Automatically discovery and provide context to data
- Ensure it is used in a compliant manner
- Aid findability of data
- Take action to improve the data and its use
- Empower your knowledge workers to be more effective with their tools

A fabric that supports the data governance objectives of our clients throughout the analytics lifecycle

Define

Policy Management

Rule Authoring

Term Definitions

Discover

Automated Metadata discovery

Manual ingest of data

Classify

Automatic profiling of data to relate to data policies

Assignment of Business Terms to Technical Assets

Share

Publish to catalog

Allow knowledge workers to curate and collaborate with data

Enforce

Automated policy enforcement throughout its lifecycle within the catalog

Monitor

Understand how data is used and found.

Take action to improve data reuse

IBM Data Catalog
Powered by the Watson Data Platform

ALL Users ALL data EVERYWHERE

.....Never before has so much information been
available to so many



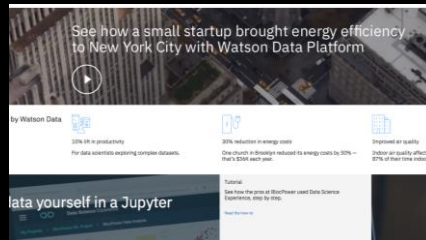
Continue your WDP Exploration

<https://www.ibm.com/analytics/us/en/watson-data-platform/>

Follow our thought leadership series:

<https://www.ibm.com/blogs/bluemix/tag/ibm-data-catalog/>

See how a small startup brought energy efficiency to NYC with WDP



<https://www.ibm.com/analytics/us/en/watson-data-platform/bloccpower/>

Browse & Share the WDP Tutorial

<https://ibm.biz/wdptutorial>

