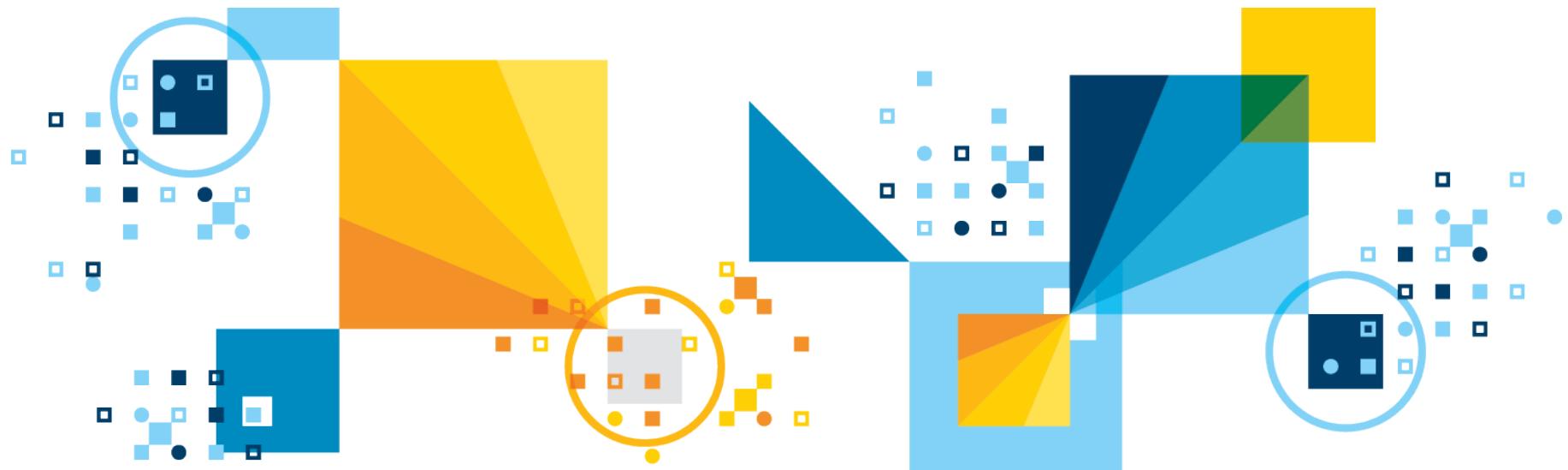
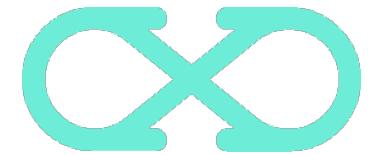


IBM Data Science Experience Overview

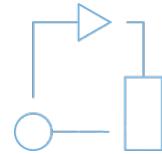


Contents

- **IBM Watson Data Platform**
- **Data Science Experience**
- **Projects and Community**
 - Projects
 - GitHub Integration
 - Community Cards
- **IBM Value-Add**
 - IBM Analytics for Apache Spark
 - Object Storage
 - Watson Data Platform Connectors
- **Open Source**
 - RStudio
 - Shiny Web Application Framework
- **IBM Machine Learning**
- **Appendix**

Introducing IBM Watson Data Platform

The first data and analytics platform for the Cognitive Business



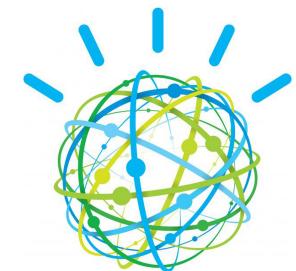
Platform.



Method.



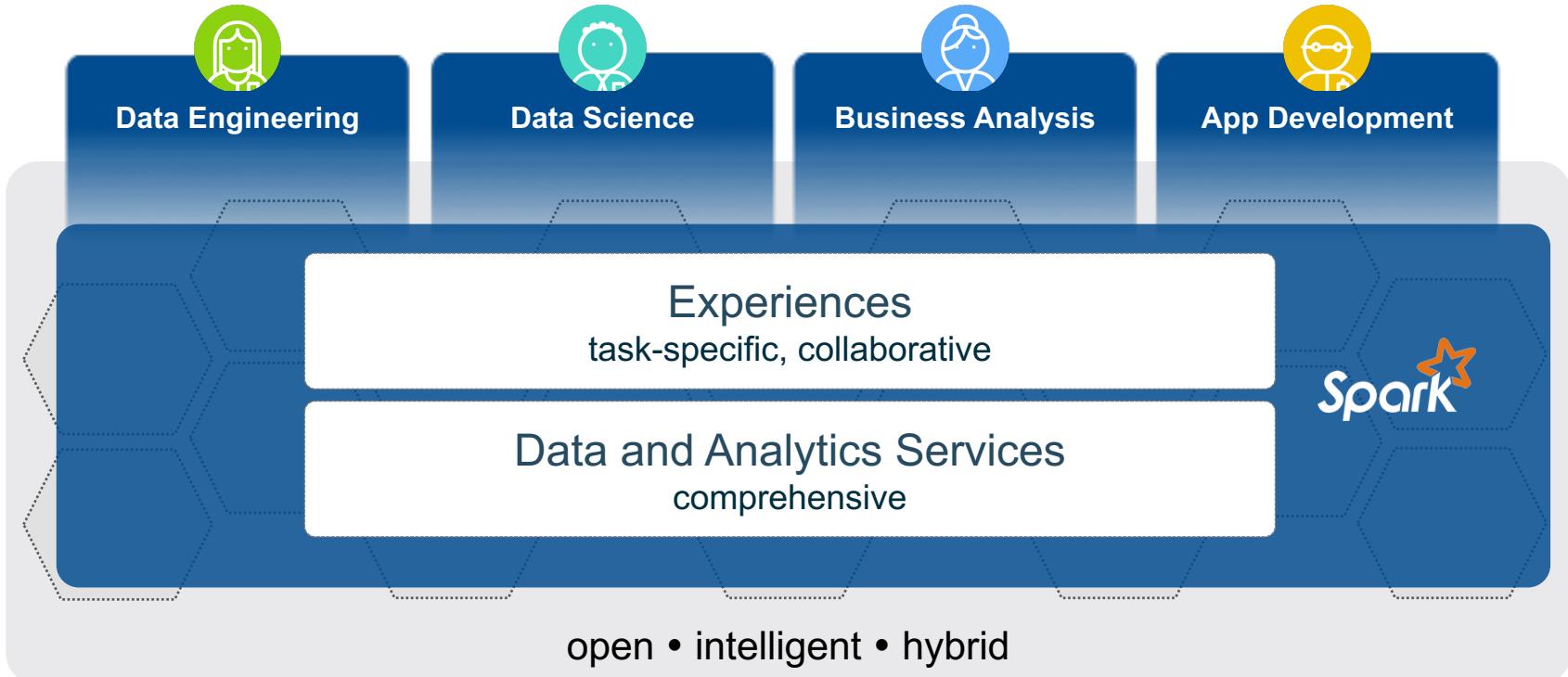
Ecosystem.



<http://ibm.co/makedatasimple>

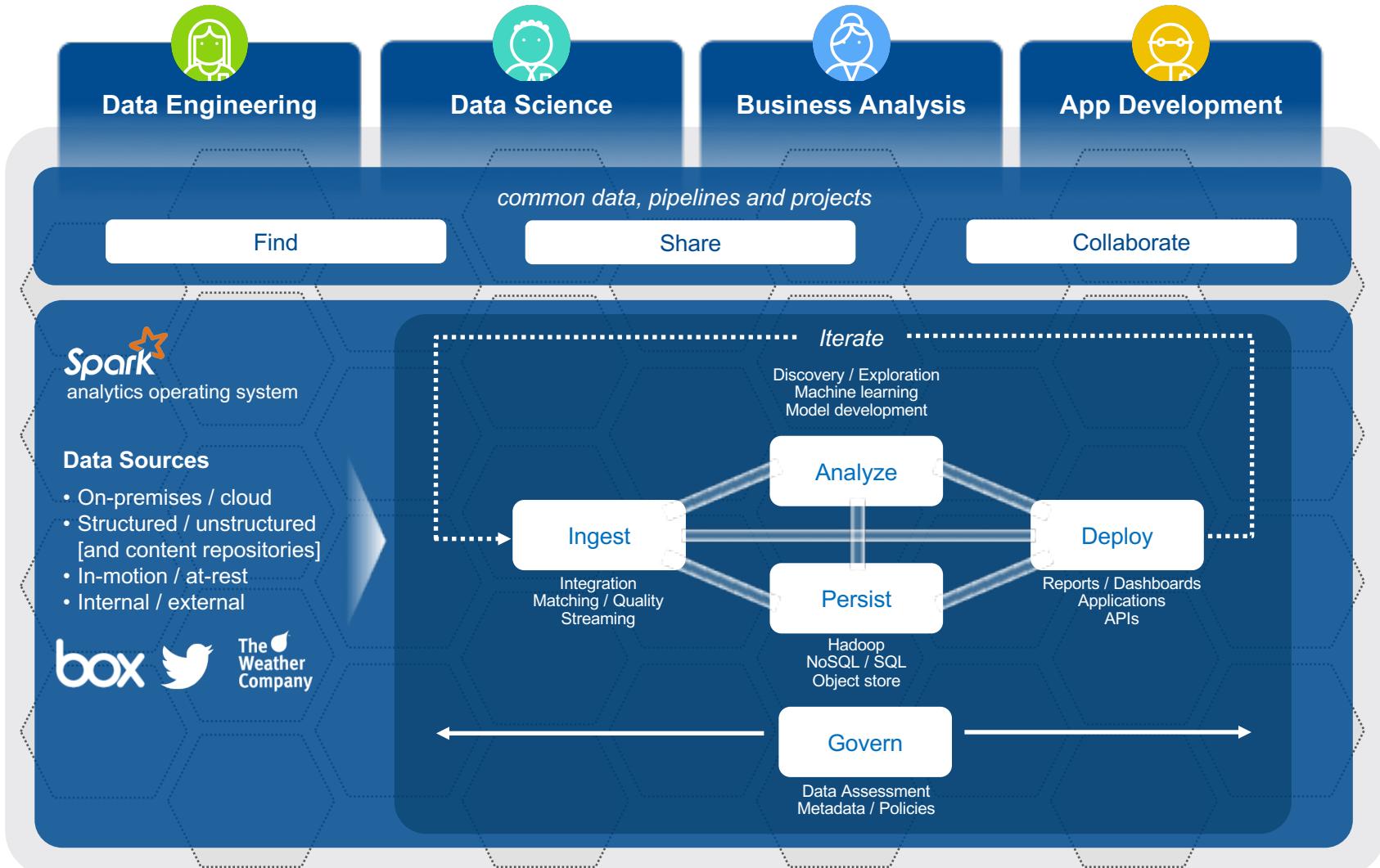
IBM Watson Data Platform

Experience New Ways To Put Data To Work



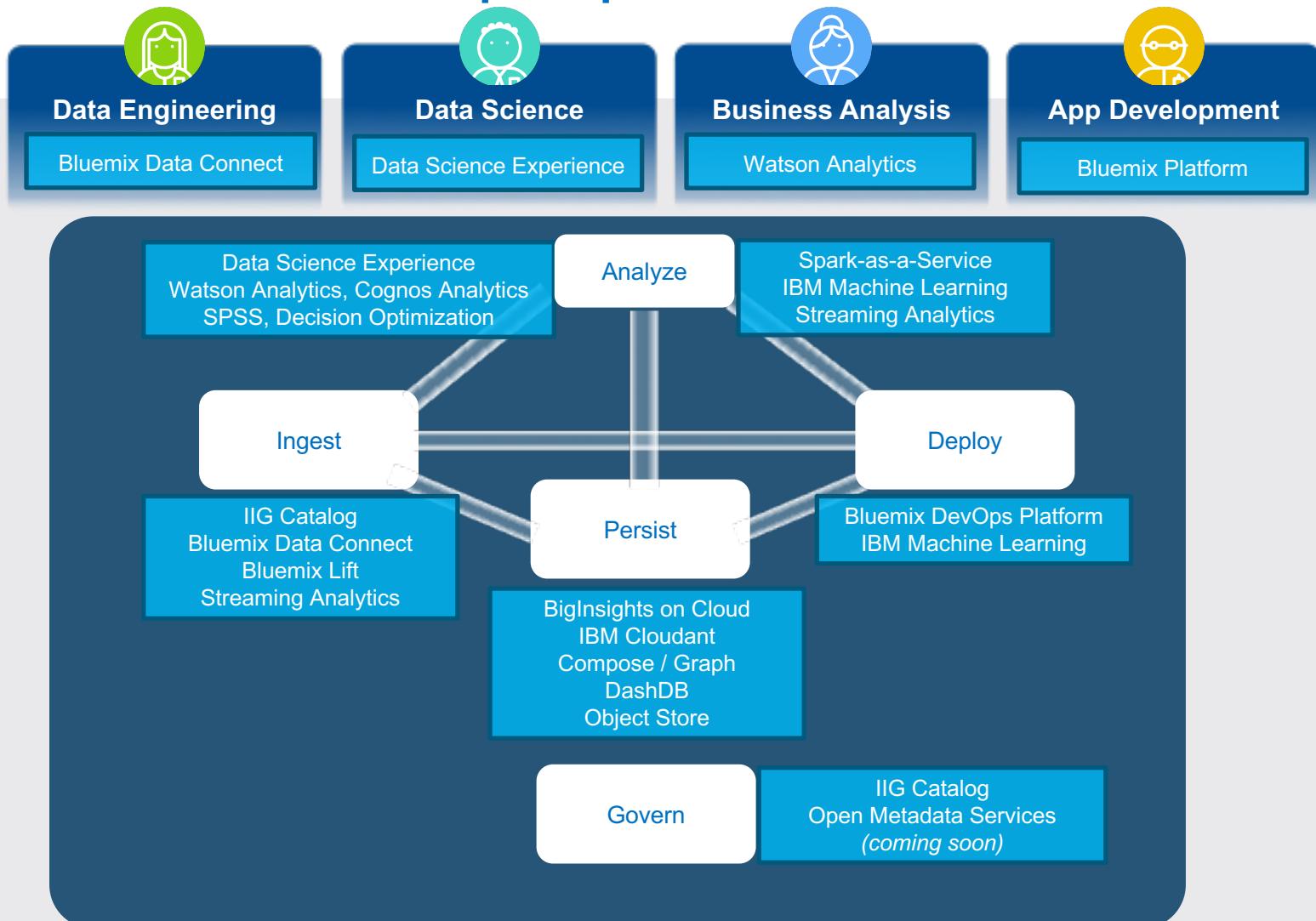
IBM Watson Data Platform

Connects Users to Data and Analytics



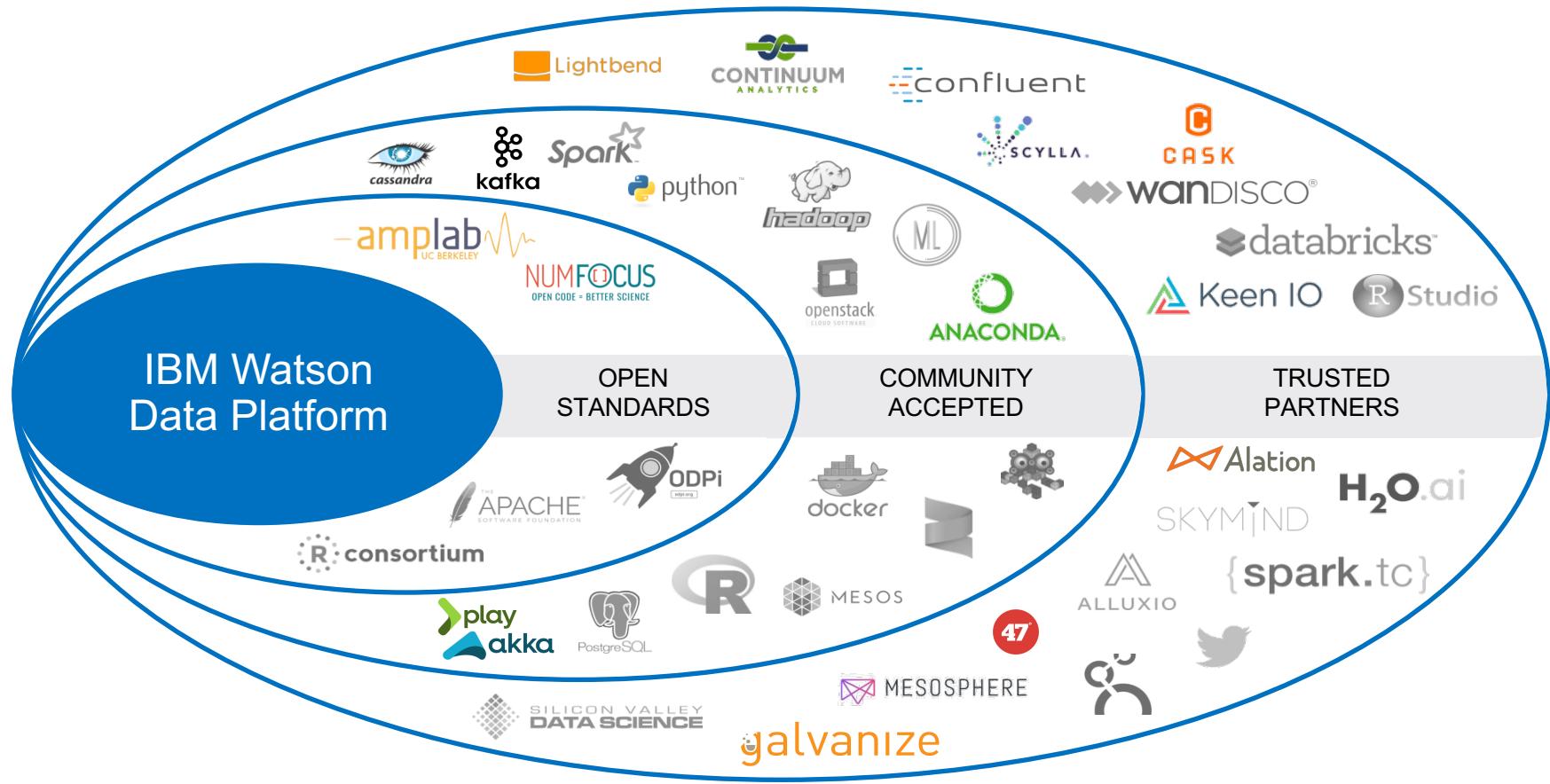
IBM Watson Data Platform

A closer look at what makes up the platform

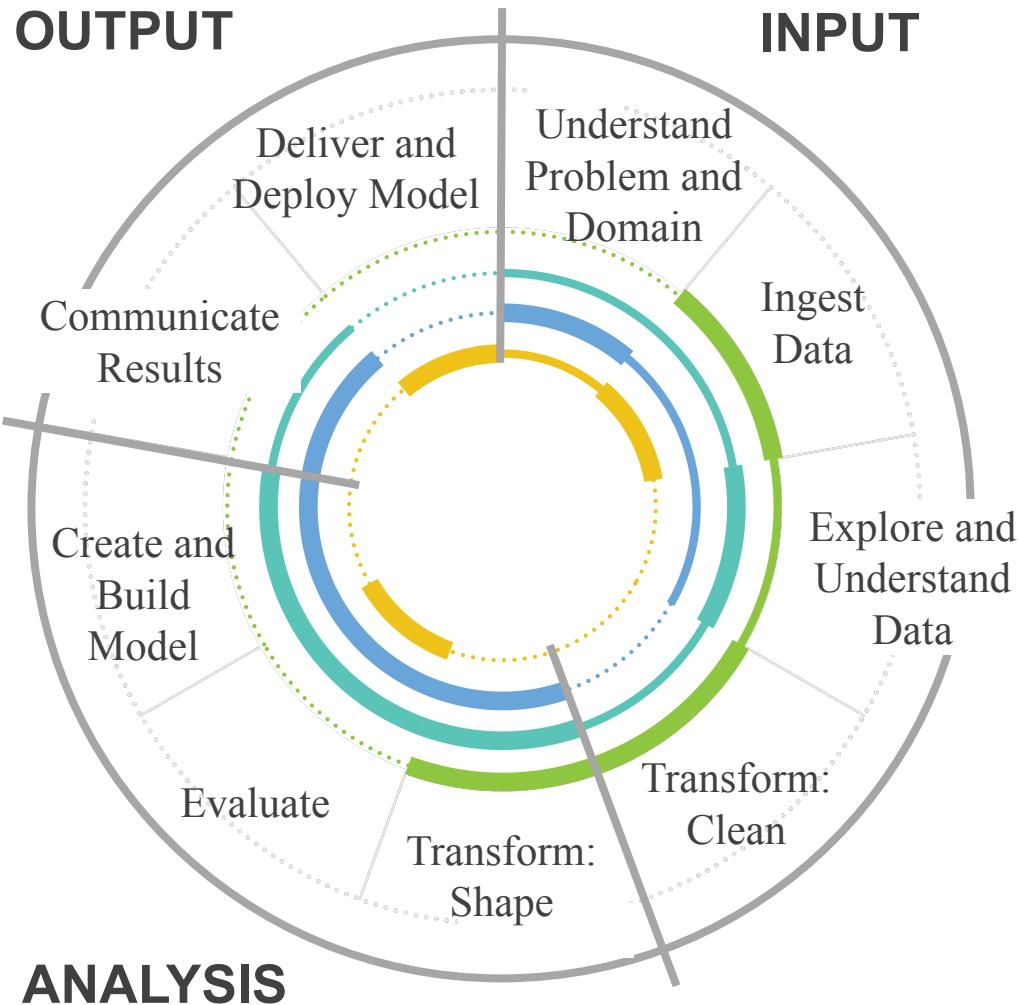


IBM Watson Data Platform Partner Ecosystem

The Open Community To Innovate Faster With Data



Tailored Experiences and User Collaboration



Data Engineer

Architects how data is organized & ensures operability
[Bluemix Data Connect](#)



Data Scientist

Gets deep into the data to draw hidden insights for the business
[Data Science Experience](#)



Business Analyst

Works with data to apply insights to the business strategy
[Watson Analytics](#)



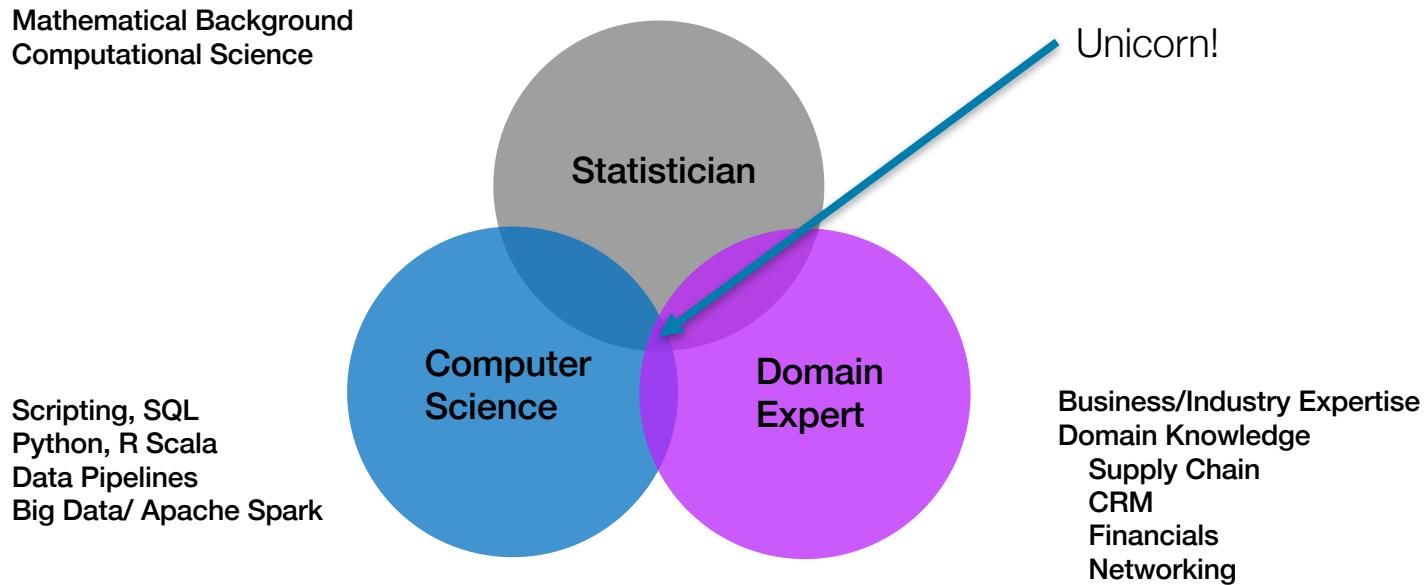
App Developer

Plugs into data and models & writes code to build apps
[Bluemix](#)



A Deeper look at a Data Scientist

Data Scientists Combine Skills across areas of Expertise



A Data Science Professional vary in a combinations of these skills



Data Scientist, Sexiest job in the 21st Century, but one of toughest...

▪ Rigid toolset

- Have to choose one and only one approach
- Cannot easily connect all of the capabilities required
- Difficult to navigate between the various tools used

▪ Fragmented and time consuming

- Using multiple disjoint environments
- Separate on-ramp/community for each tool/environment
- Does not have meta data or data lineage

▪ Analytical Silo

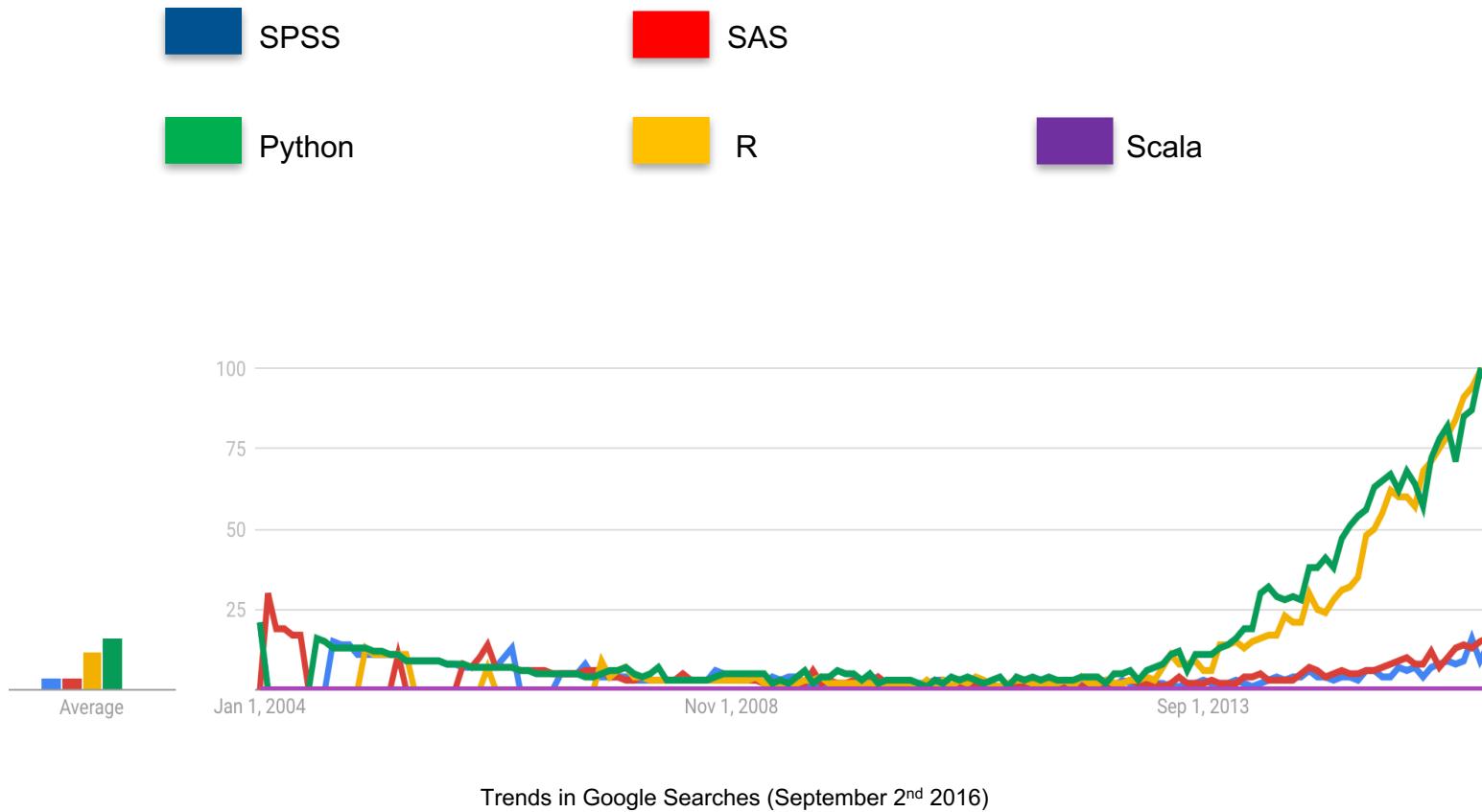
- Difficult to maintain and version control project assets
- Limited means of collaborating with teams
- Results are difficult to share

Data Science Platforms

Gartner February 2017 Magic Quadrant



Google Trends, *Open Source Data Science Technologies*



Open Source - Data Science Ecosystem



{ R }

{



} Big Data

{



} Python

Introducing the Data Science Experience



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



External URL: <http://datascience.ibm.com>
Internal ZACS Page: <https://ibm.biz/BdrDv5>

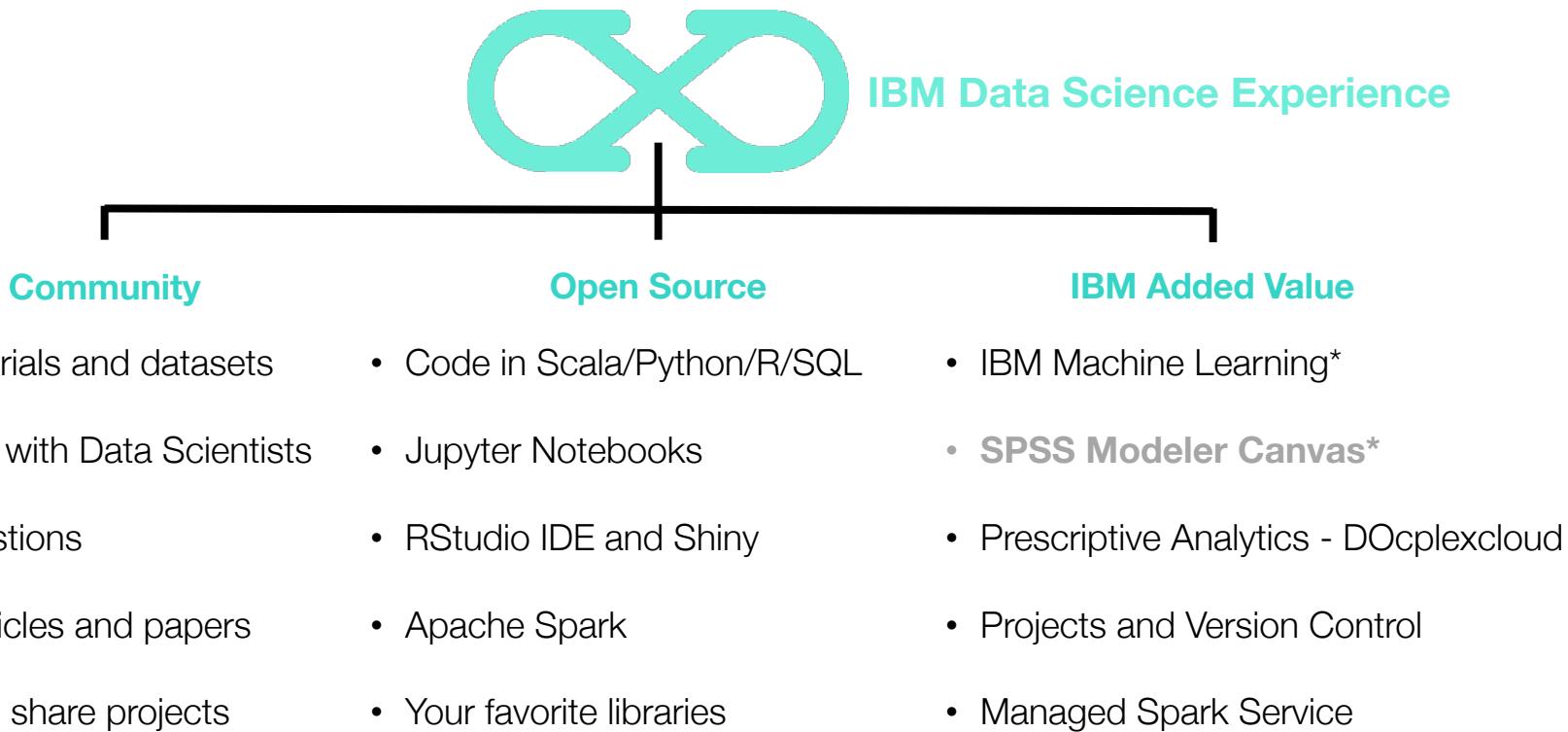
IBM Data Science Experience

ALL YOUR TOOLS IN ONE PLACE

IBM Data Science Experience is an environment that brings together everything that a Data Scientist needs. It includes the most popular Open Source tools and IBM unique value-add functionalities with community and social features, integrated as a first class citizen to make Data Scientists more successful.



Core Attributes of the Data Science Experience



Powered by IBM **Watson Data Platform**

* Closed beta



IBM Data Science Experience
<https://youtu.be/1HjzkLRdP5k>



Collaborate Using Projects

☰  Data Science Experience ▾

My Projects > New Sales campaign

Overview Analytics Assets Data Assets Bookmarks Collaborators Settings

Notebooks [view all \(2\)](#)

NAME	SHARED	STATUS	LANGUAGE
 Retail Sales Analysis v2			Python 2.7
 Machine Learning using R			R 3.3.0

Data Assets [view all \(23\)](#)

NAME	TYPE
Great Outdoors Orders for BBBT Ritika	Catalog File
Great Outdoors Orders for BBBT Ritika	Catalog File
 gchn-daily-by_year-format.rtf	RTF
 Presence Data (Cloudant NoSQL)	Connection
 Sales Data (dashDB)	Connection

Bookmarks [view all \(3\)](#)

ARTICLE
From Machine Learning to Learning M...
Nov 10, 2016 

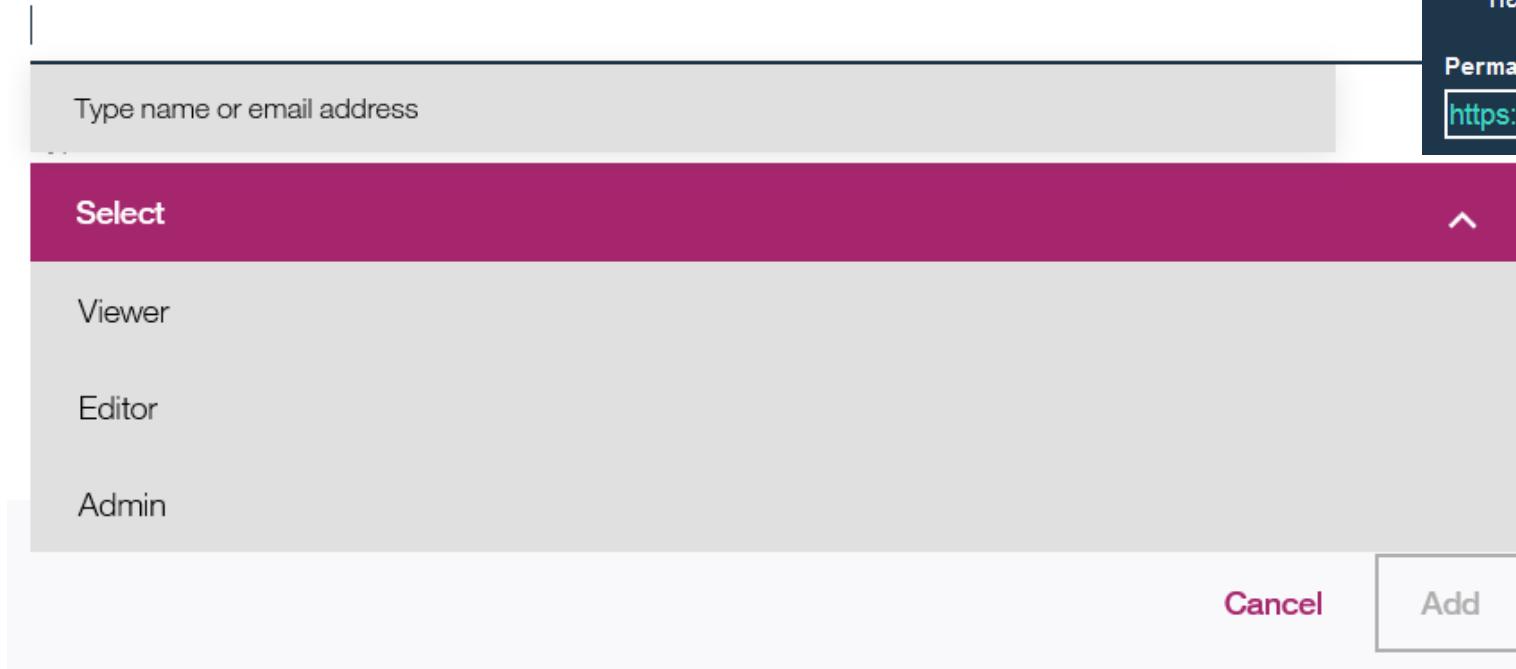
NOTEBOOK
Use deep learning for image classifica...
Oct 4, 2016 

TUTORIAL
Analyze open data sets using pandas ...
Oct 19, 2016 

Features for sharing, forking, and reusing Project assets to increase your data science team's productivity

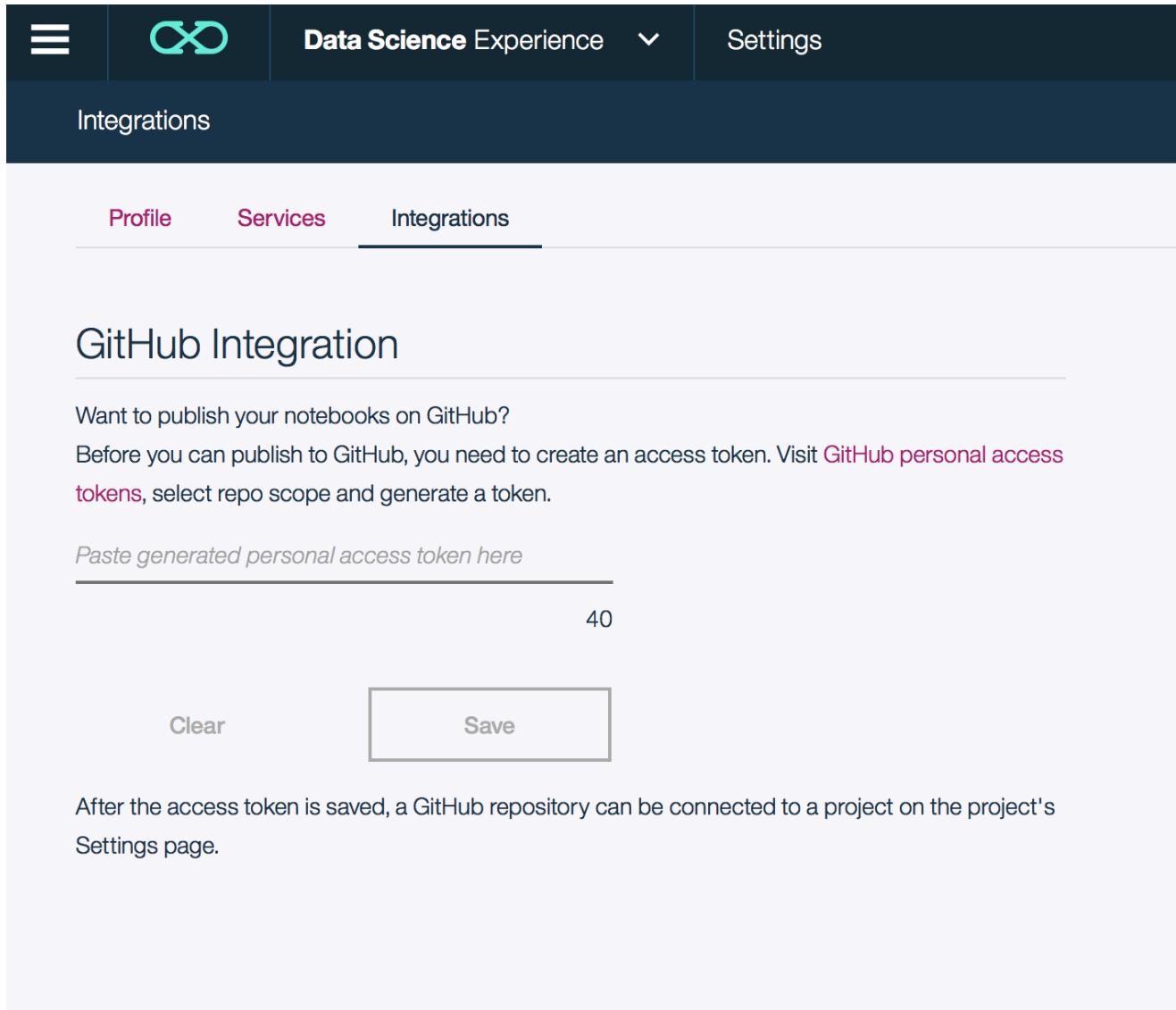
Add New Collaborator

Add users to your project for collaboration. Users with write access can add services to your project...



The screenshot shows a user interface for adding new collaborators to a project. At the top, there is a search bar with the placeholder "Type name or email address". Below the search bar, a red header bar contains the word "Select". Underneath this, there are three options: "Viewer", "Editor", and "Admin". In the bottom right corner of the main area, there are two buttons: "Cancel" and "Add". To the right of the main interface, a sidebar titled "Sharing" provides information about sharing notebooks and includes a checked checkbox for "Share with anyone who has the link". It also displays a "Permalink to view notebook" link: <https://apsportal.ibm.com/an>.

GitHub Integration



The screenshot shows the IBM Data Science Experience interface with the "Integrations" tab selected. The "GitHub Integration" section is displayed, prompting the user to paste a generated personal access token. A text input field contains the placeholder text "Paste generated personal access token here". Below the input field is a character count indicator showing "40". At the bottom are "Clear" and "Save" buttons.

Want to publish your notebooks on GitHub?

Before you can publish to GitHub, you need to create an access token. Visit [GitHub personal access tokens](#), select repo scope and generate a token.

Paste generated personal access token here

40

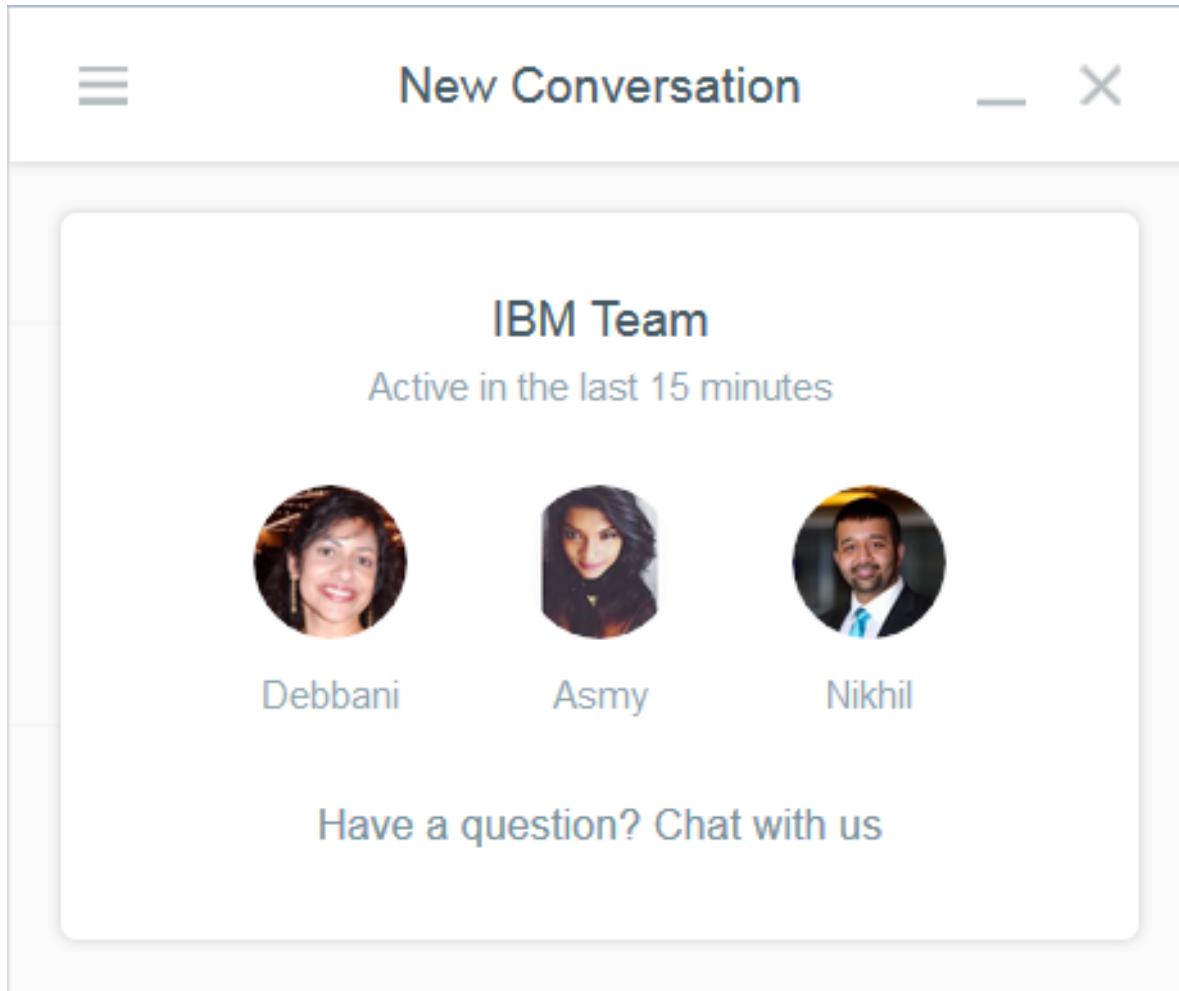
Clear Save

After the access token is saved, a GitHub repository can be connected to a project on the project's Settings page.

Community Cards provide in-context learning for users

<p>ARTICLE How can data scientists collaborate to build...</p> <p>SOURCE IBM DATE Jun 24, 2016</p>	<p>ARTICLE What is machine learning?</p> <p>SOURCE IBM DATE Jun 24, 2016</p>	<p>NOTEBOOK Insights from Twitter data about car makers</p> <p>SOURCE IBM DATE Jun 22, 2016</p>
<p>NOTEBOOK Insights from New York car accident reports</p> <p>SOURCE IBM DATE Jun 16, 2016</p>	<p>DATA SET Country Surface Area (sq. km)</p> <p>SOURCE IBM DATE Jun 16, 2016</p>	<p>NOTEBOOK Improved Flight delay prediction</p> <p>SOURCE IBM DATE Jun 06, 2016</p>
<p>NOTEBOOK Load data from different sources</p> <p>SOURCE IBM DATE Jun 02, 2016</p>	<p>NOTEBOOK Learn basics about notebooks and Apache Spark</p> <p>SOURCE IBM DATE Jun 02, 2016</p>	<p>NOTEBOOK Analyze precipitation data</p> <p>SOURCE IBM DATE Jun 02, 2016</p>

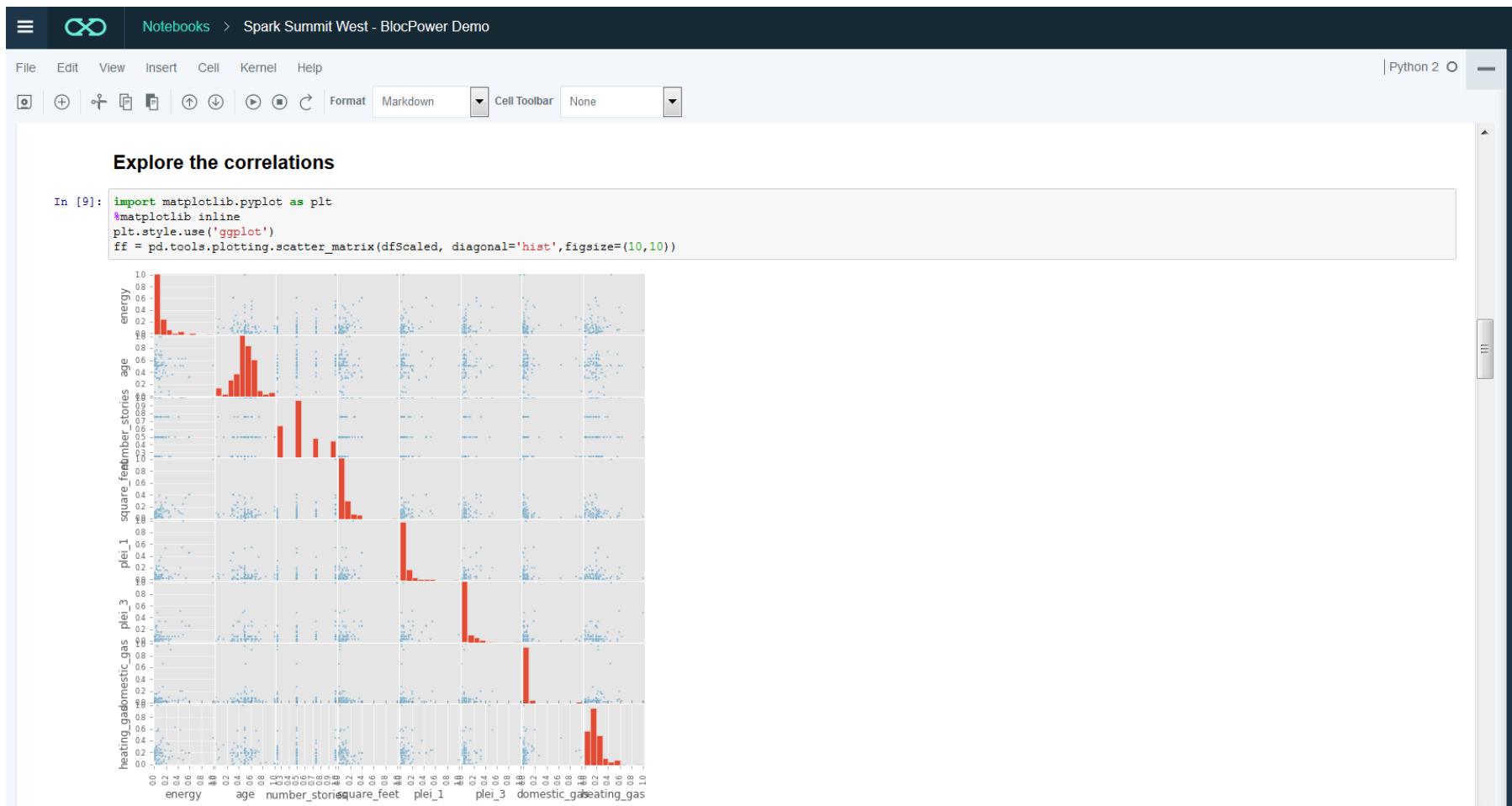
Live chat on Intercom for support from the IBM team and to provide your feedback on how we can improve DSX



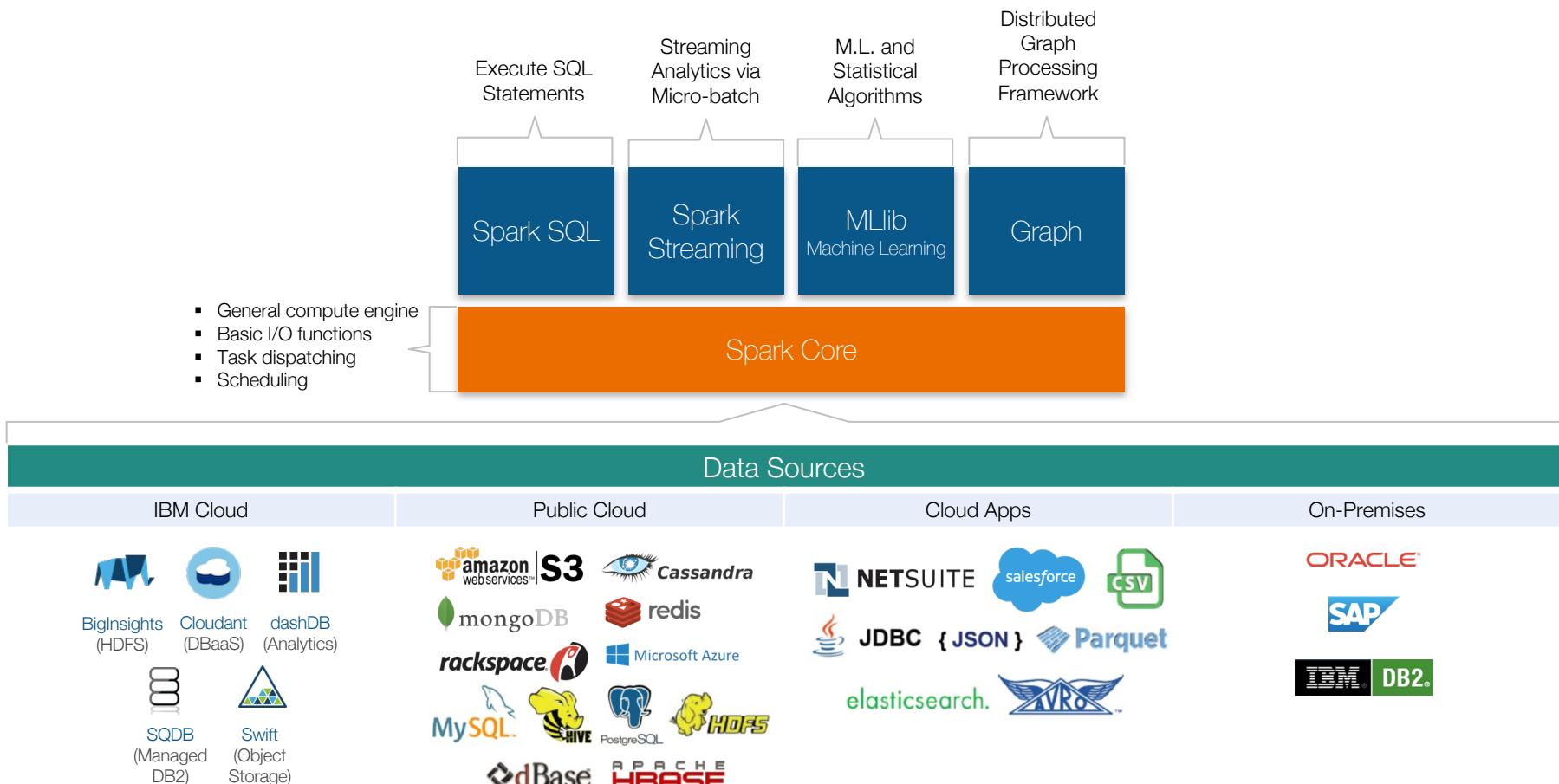
The screenshot shows the 'New Conversation' screen in Intercom. At the top, it says 'New Conversation' with a minimize and close button. Below that, it displays 'IBM Team' which was 'Active in the last 15 minutes'. Three team members are listed with their names and profile pictures: Debbani, Asmy, and Nikhil. At the bottom, there is a call-to-action button labeled 'Have a question? Chat with us'.



Integrated Jupyter Notebooks for interactive and collaborative development - seamless execution on Spark



From a Notebook you can use IBM's managed Spark Service to blend multiple data types, sources, and workloads



The Spark Service uses Bluemix Object Storage as its preferred data store for building performant applications

- Object storage provides **inexpensive, scalable and self-healing** retention of massive amounts of unstructured data
- Every object exists at the same level in a **flat address space**
- Bluemix Object Storage has a **drag-and-drop** upload and **Swift API** for programmatic access



Object Storage
IBM

Supported Data Sources for DSX via on-premises and cloud Connections



Cloud Sources	On-Premises Sources	Cloud Targets	On-Premises Targets
Amazon Redshift	Apache Hive	Amazon S3	IBM DB2® LUW
Amazon S3	Cloudera Impala	Bluemix Object Storage	IBM Pure Data for Analytics®
Apache Hive	IBM DB2® LUW	IBM Cloudant™	Teradata
Bluemix Object Storage	IBM Informix®	IBM dashDB	
IBM BigInsights™ on Cloud	IBM Pure Data for Analytics®	IBM BigInsights™ on Cloud	
IBM Cloudant™	Microsoft SQL Server	IBM DB2® on Cloud	
IBM dashDB	MySQL Enterprise Edition	IBM SQL Database	
IBM DB2® on Cloud	Oracle	IBM Watson™ Analytics	
IBM SQL Database	Pivotal Greenplum	PostgreSQL on Compose	
Microsoft Azure	PostgreSQL	SoftLayer Object Storage	
PostgreSQL on Compose	Sybase		
Salesforce	Sybase IQ		
SoftLayer Object Storage	Teradata		

IBM Decision Optimization for DSX via API calls to DOcplexcloud

Bank Marketing Campaign. - x https://apsportal.ibm.com/analytics/notebooks/1ef3b1b9-5d97-4c6e-89ee-3efb53210d07 Vincent

Notebooks > Bank Marketing Campaign. **Marketing Campaign Planning demo**

File Edit View Insert Cell Kernel Help Python 2

Let's create the optimization model to select the best ways to contact customers and stay within the limited budget.

Step 1: Set up the prescriptive engine

- Subscribe to the Decision Optimization on Cloud solve service [here](#).
- Get the service URL and your personal API key and enter your credentials here:

First import docplex and set the credentials to solve the model using IBM ILOG CPLEX Optimizer on Cloud. docplex is already imported in the previous cell.

```
In [12]: import sys
import docplex.mp
```

```
In [13]: url = "https://api-ooas.docloud.ibmcloud.com/job_manager/rest/v1/"
key = "api_f550300e-8e52"
```

Step 2: Set up the prescriptive model

Create the model

```
In [14]: from docplex.mp.model import Model
mdl = Model(name="marketing_campaign")
```

Warning: CPLEX DLL not found and model has no DOcloud credentials. Provide credentials at solve time

Define the decision variables

- The integer decision variables channelVars, represent whether or not a customer will be made an offer for a particular product via a particular channel.
- The integer decision variable totalOffers represents the total number of offers made.
- The continuous variable budgetSpent represents the total cost of the offers made.

```
In [15]: offersR = xrange(0, len(offers))
productsR = xrange(0, len(products))
channelsR = xrange(0, len(channels))

channelVars = mdl.binary_var_cube(offersR, productsR, channelsR)
totalOffers = mdl.integer_var()
budgetSpent = mdl.continuous_var()
budgetMax = mdl.integer_var(lb=availableBudget, ub=availableBudget, name="budgetMax")
```

```
In [16]: print("we created %d decision variables for this problem" %(len(offersR)*len(productsR)*len(channelsR)+1+1))
```

Decision Optimization on Cloud (DOcplexcloud) credentials used inside DSX

- (1) Purchase DOcplexcloud on IBM Cloud Marketplace
- (2) Receive credentials
- (3) Enter credentials into DSX

Plenty of samples and tutorials available within DSX

DSX has RStudio built into the experience thanks to our strategic partnership

The screenshot shows the RStudio interface with the following components:

- Code Editor:** The top-left pane displays the R script "RStudioTest.R*". A tooltip for the "annotate" argument in the "calendarPlot" function is shown, explaining its purpose: "This option controls what appears on each day of the calendar. Can be: "date" - shows day of the month; "wd" - shows vector-averaged wind direction, or "ws" - shows vector-averaged wind direction scaled by wind speed." The tooltip also includes a note: "Press F1 for additional help".
- Console:** The bottom-left pane shows the R console output for the "RStudioTest.R" script.
- Workspace:** The top-right pane shows the workspace, indicating "bloomsbury" with 50804 obs. of 18 variables.
- Plots:** The main right-hand area displays a "calendar plot" for O₃ in 2006. The plot uses a color scale from light yellow (low values) to dark red (high values). A vertical color bar on the right indicates the scale from 20 to 100. The plot shows monthly cycles with higher concentrations occurring in the summer months (July/August).

```

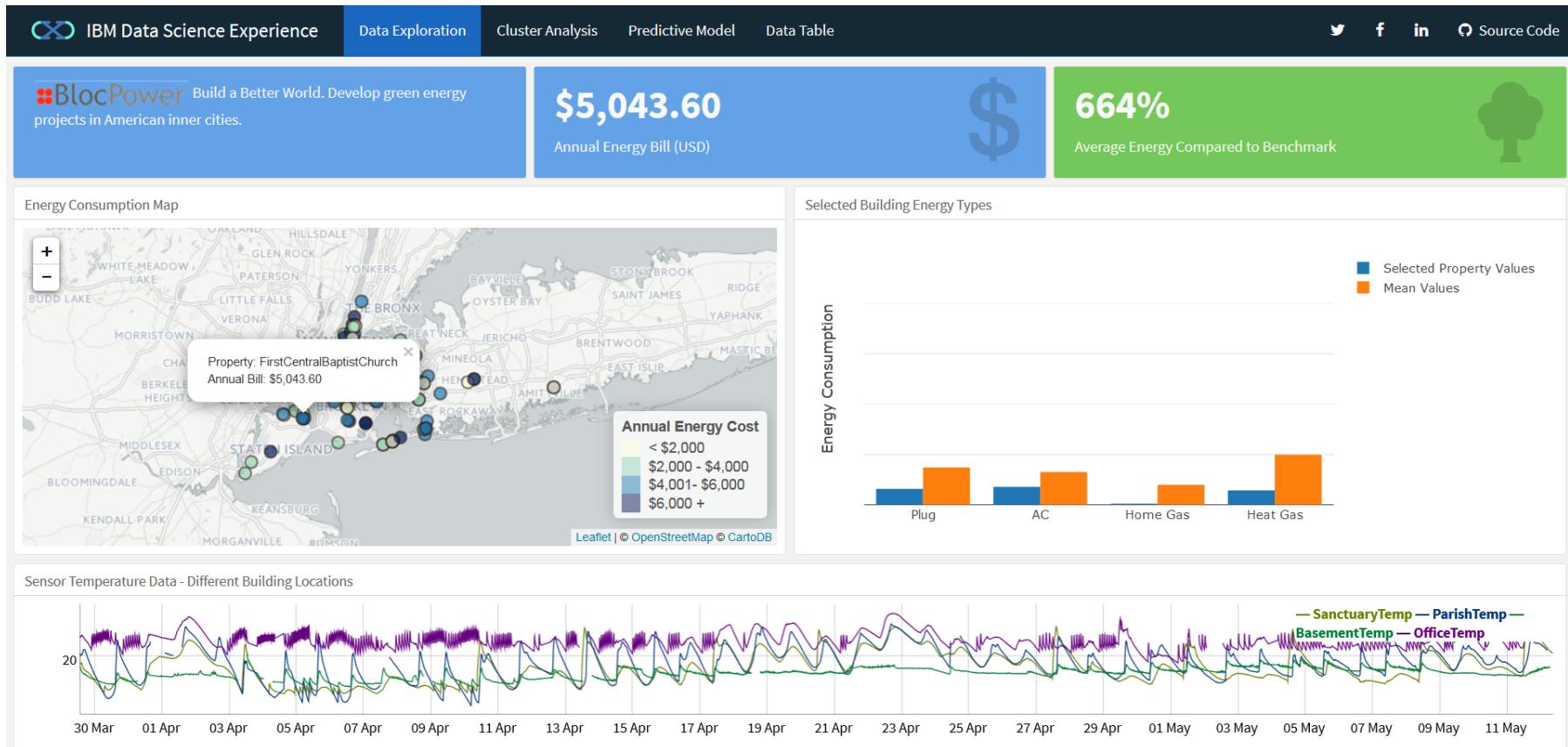
File Edit View Workspace Plots Help
RStudioTest.R*
Source on Save Run Line(s) Run All
library(openair)
## import some example data
bloomsbury <- importKCL(site = "b10", year = 2005:2010, met = TRUE)
## have a look at the data
summary(plot(bloomsbury))
## trend in o3 by wd
smoothTrend(bloomsbury, pollutant = "o3", deseason = TRUE, type = "wd")
## polarPlot of nox
polarPlot(bloomsbury, pollutant = "o3", type = "daylight")
## calendar plot
calendarPlot(bloomsbury, pollutant = "o3",)

mydata<-
pollutant<
years<
type<
annotate<
statistic<
cols<
limits<
Press F1 for additional help

Console ~/RStudioTest.R
>
>
>
>
> library(openair)
## import some example data
bloomsbury <- importKCL(site = "b10", year = 2005:2010, met = TRUE)
## have a look at the data
summary(plot(bloomsbury))
## trend in o3 by wd
smoothTrend(bloomsbury, pollutant = "o3", deseason = TRUE, type = "wd")
## polarPlot of nox
polarPlot(bloomsbury, pollutant = "o3", type = "daylight")

NOTE - mass units are used
ug/m3 for NOx, NO2, SO2, O3; mg/m3 for CO
PM10_raw is raw data multiplied by 1.3
Warning message:
In importKCL(site = "b10", year = 2005:2010, met = TRUE) :
  Some of the more recent data may not be ratified.
  date1   date2    nox     no2      o3      so2      co    pm10_raw    pm10    pm25    site
  "POSIXct" "POSIXct" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "factor"
  code      ws      wd      solar    rain     temp     bp     rhum
  "character" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
> calendarPlot(bloomsbury, pollutant = "o3", year = 2006)
  
```

With RStudio you can create Shiny web applications to make your analysis accessible to the business



Modelling Energy Usage in NYC – BlocPower



— Tooraji Arvajeh,
Chief Engineering Officer,
BlocPower

"BlocPower operation is diverse from outreach and targeting, origination of investment-grade clean energy projects to financing projects through our crowdfunding marketplace. Data is the underlying tool of our operation and IBM's Data Science Experience will facilitate a closer integration across it and help our business scale up faster."



Blog Link: <http://ibm.co/29KLbvU>

Use Shiny apps to share your analysis with business users

BlocPower Demo v3 - Blue

<https://apsx-dev.stage1.ng.bluemix.net/analytics/>

In [8]:

```
credentials_1['filename'] = 'CDD-MDU-Features.py'
dfCH = pd.read_csv(get_file_content(credentials_1['filename']))
```

1. Energy Usage (kWh) Metrics

Clean and Prepare Data

In [9]:

```
# energy usage in (kWh)
energy, age, num_stories, sq_feet, plenty_of_sunlight, domestic_gas, heating_gas
```

Build Feature Matrix, Fill Missing Values

In [10]:

```
matrix = np.transpose(np.matrix([energy, domestic_gas, heating_gas]))
cols = ['energy_usage', 'age', 'number_stories', 'sq_feet', 'plenty_of_sunlight', 'domestic_gas', 'heating_gas']
feat = pd.DataFrame(data=matrix)
# fill missing values with average
feat = feat.fillna(feat.mean())
# scale data
scaler = preprocessing.MaxAbsScaler()
feat = scaler.fit_transform(feat)
dfScaled = pd.DataFrame(feat, columns=cols)
```

Explore Correlations

In [11]:

```
plt.style.use('ggplot')
ff = pd.tools.plotting.scatter_matrix(dfScaled)
```

Data & Analytics Portal

<https://apsnginxrstage1.mybluemix.net/node1/rstudio40008/p/4456/shinyDemo.Rmd#data-exploration>

IBM Data Science Experience

BlocPower Build a Better World. Develop green energy projects in American inner cities.

23643 Annual Enery Use (kwh)

\$4,255.74 Annual Energy Bill (USD)

Energy Consumption Map

Location Values

Energy Consumption Type	Selected Property Values	Mean Values
Plug Load	~0.60	~0.18
AC Costs	~0.15	~0.15
Domestic ~	~0.05	~0.10
Heating G...	~0.20	~0.20

Sensor Temperature Data - Different Building Locations

Interactively explore the analysis of your data science team

BlocPower Demo v3 - Blue

The clustering model help us identify buildings.

- Note in the above figure that most buildings are green (purple cluster)
- Buildings that are part of the brown, yellow and light blue ones are part of the purple cluster
- Labels are re-coded into a binary variable where 1 means building is part of the purple cluster (green cluster) and 0 otherwise (purple cluster)

```
In [19]: # binary variable to identify inefficient buildings
label_binary = []
for v in labels:
    label_binary.append(0 if (v == 0) else 1)
label_binary = np.asarray(label_binary)
```

Classification Model Identify Inefficient Buildings

```
In [20]: # train classifier
log = linear_model.LogisticRegression(tol = 0.001)
log.fit(featReduced, label_binary)
accuracy = log.score(featReduced, label_binary)
y_pred = log.predict(featReduced)

In [21]: print "Model Accuracy: ", accuracy
Model Accuracy:  0.893203883495

In [22]: def plot_confusion_matrix(cm, title='Confusion matrix', cmap=plt.cm.Greens):
    plt.imshow(cm, interpolation='nearest', cmap=cmap)
    plt.title(title)
    plt.colorbar()
    tick_marks = np.arange(2)
    plt.xticks(tick_marks, ['Efficient', 'Inefficient'])
    plt.yticks(tick_marks, ['Efficient', 'Inefficient'])
    plt.tight_layout()
    plt.ylabel('True label')
    plt.xlabel('Predicted label')

In [23]: from sklearn.metrics import confusion_matrix
confusion_matrix(y_true, y_pred)
```

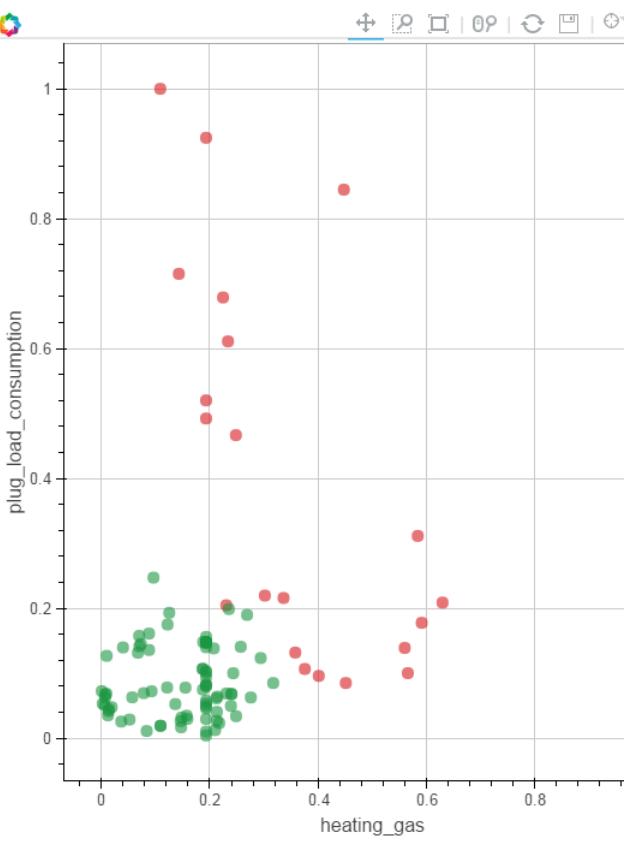
Data & Analytics Portal

IBM Data Science Experience

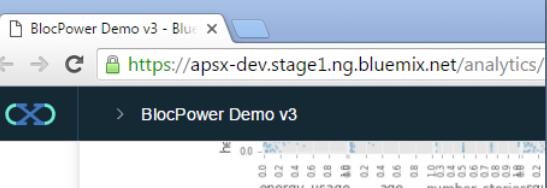
Energy Consumption Map



Clusters by Heating and Plug Consumption



Adjust parameters on-the-fly and visualize model predictions



Linear Regression Model.

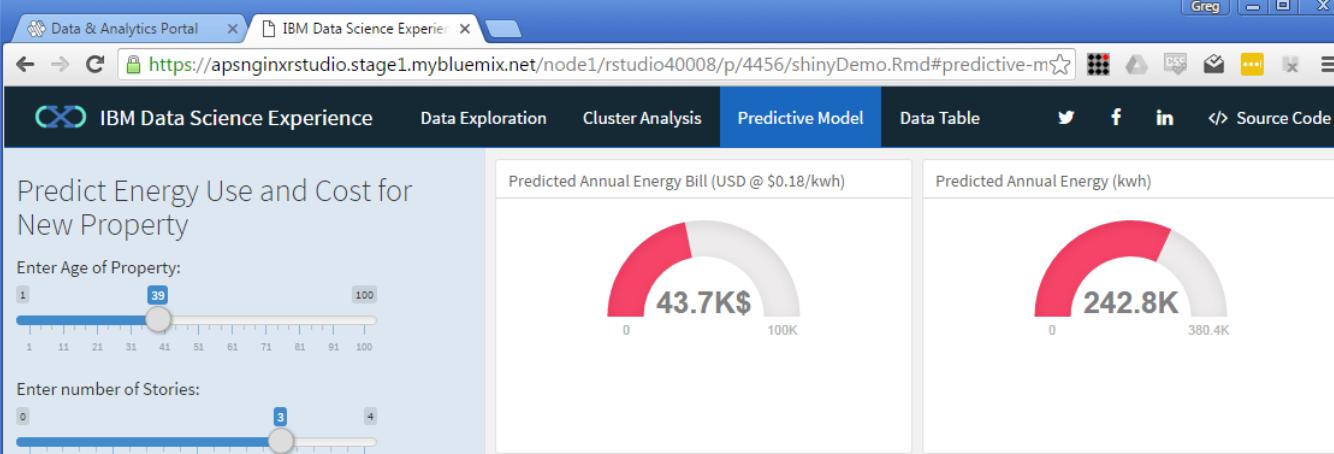
Hypothesis: energy usage (kWh) can be predicted by characteristics:

- age of the building
- square feet
- number of stories
- number of plugged equipment, ...

```
In [12]: features = dfScaled.columns.tolist()
response = ['energy_usage']
features.remove(response[0])
# prepare data for regression
lr = linear_model.LinearRegression(fit_intercept=True)
y = np.asarray(dfScaled[response])
X = dfScaled[features]
# run regression
regr = lr.fit(X,y)
coefs = regr.coef_[0]
# collect regression results
dataRegQ = []
dataRegQ.append(['Intercept', regr.intercept_])
for i in range(len(features)):
    dataRegQ.append(([features[i]], coefs[i]))
yh = regr.predict(X)
print 'R-Squared: ', r2_score(y,yh)
pd.DataFrame(dataRegQ,columns=['feature_name', 'coefficient'])
```

R-Squared: 0.725632741591

```
Out[12]: feature_name coefficient
0 Intercept -0.092789
1 age 0.139789
2 number_stories 0.059749
3 square_feet 0.734468
4 plug_equipment 0.330050
5 domestic_gas 0.208283
6 0.105810
```



Predict Energy Use and Cost for New Property

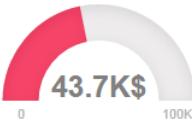
Enter Age of Property: 39

Enter number of Stories: 3

Enter Property Square Footage: 38,000

Enter Age of Property: 29

Predicted Annual Energy Bill (USD @ \$0.18/kwh)



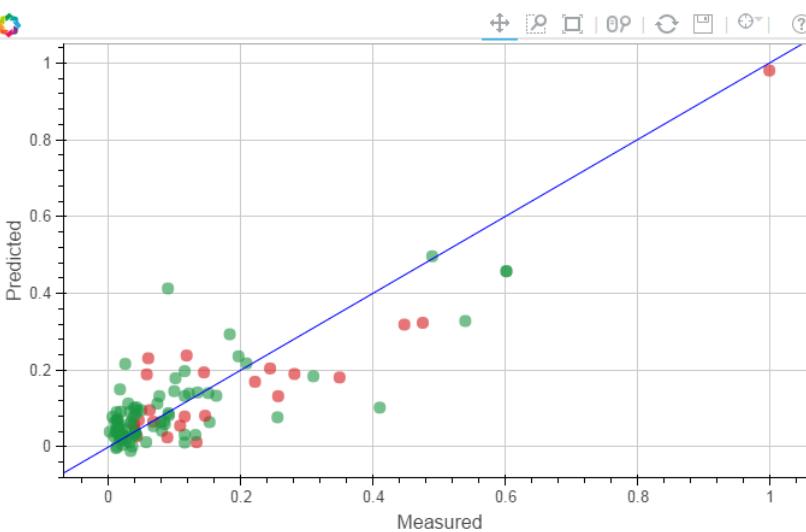
43.7K\$

Predicted Annual Energy (kwh)



242.8K

Regression Results: Actual Vs. Predicted Values

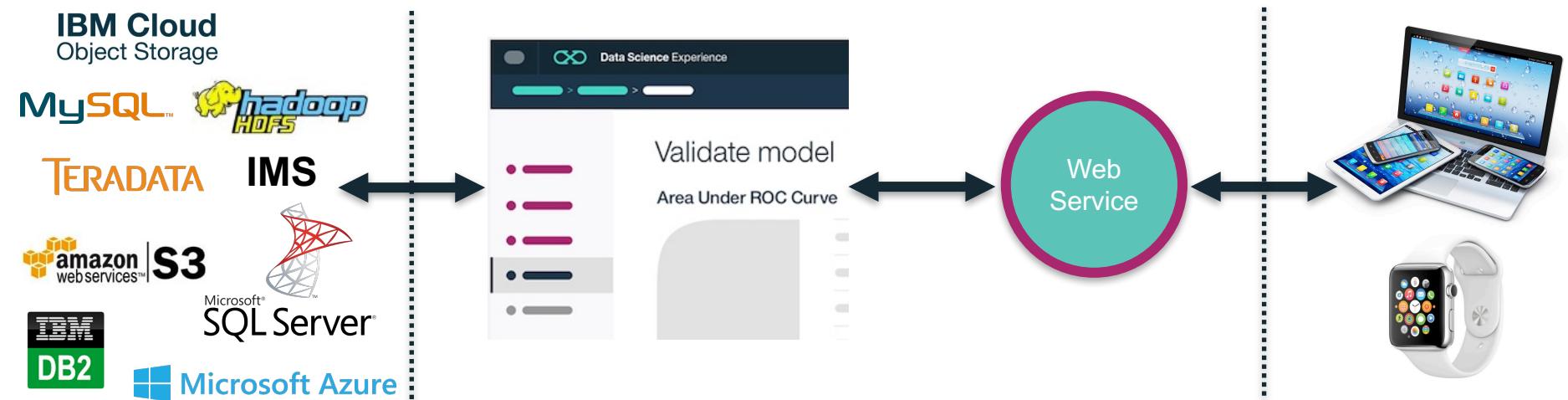


The scatter plot shows a strong positive linear relationship between Predicted and Measured values. A blue diagonal line represents the identity line (y=x). Data points are colored green and red, showing a clear separation between the two classes.

IBM MACHINE LEARNING

Operationalize insights with IBM Machine Learning

IBM Machine Learning



Data Access:

- Easily connect to Behind-the-Firewall and Public Cloud Data
- Catalogued and Governed Controls through Watson Data Platform

Creating Models:

- Single UI and API for creating ML Models on various Runtimes
- Auto-Modelling and Hyperparameter Optimization

Web Service:

- Real-time, Streaming, and Batch Deployment
- Continuous Monitoring and Feedback Loop

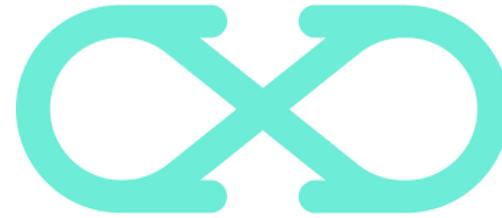
Intelligent Apps:

- Integrate ML models with apps, websites, etc.
- Continuously Improve and Adapt with Self-Learning

IBM Machine Learning – Functionality for All!



**IBM Machine Learning
(on Bluemix)**



**Data Science Experience
with IBM ML**



**IBM Machine Learning for
z/OS (with DSX)**



App Developer



Data Scientist



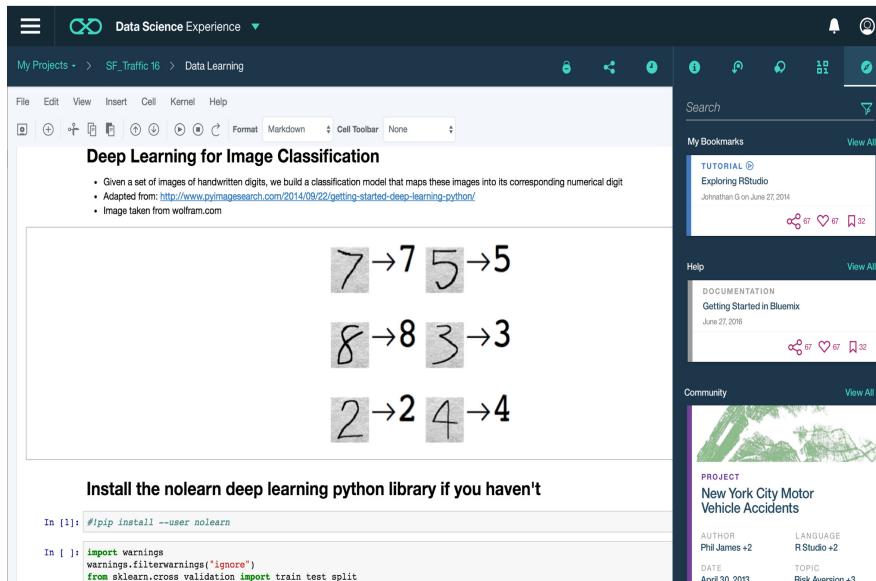
Data Scientist

IBM Machine Learning in Data Science Experience

IBM Machine Learning is provisioned by default in Data Science Experience

- Enables Data Scientists to deploy machine learning models as web services
- Single UI for creating, collaborating, deploying, monitoring, and feedback
- Accessible via API, Wizard GUI, and Canvas

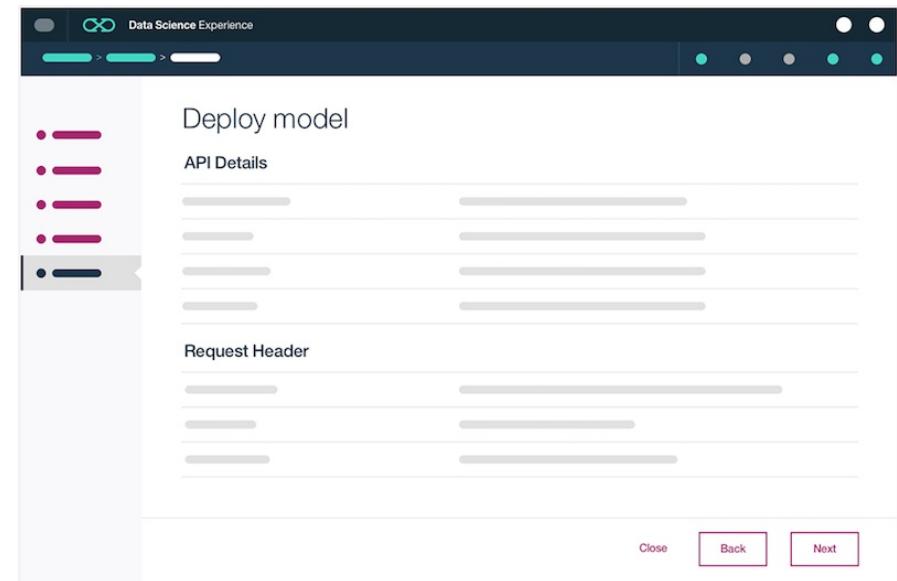
API for Jupyter Notebooks



The screenshot shows a Jupyter Notebook interface within the Data Science Experience. The notebook displays handwritten digits and their corresponding numerical labels, such as 7 → 7 and 5 → 5. Below the notebook, there is a snippet of Python code for installing the nolearn deep learning python library.

```
In [1]: #!pip install --user nolearn
In [2]: import warnings
         warnings.filterwarnings("ignore")
         from sklearn.cross_validation import train_test_split
```

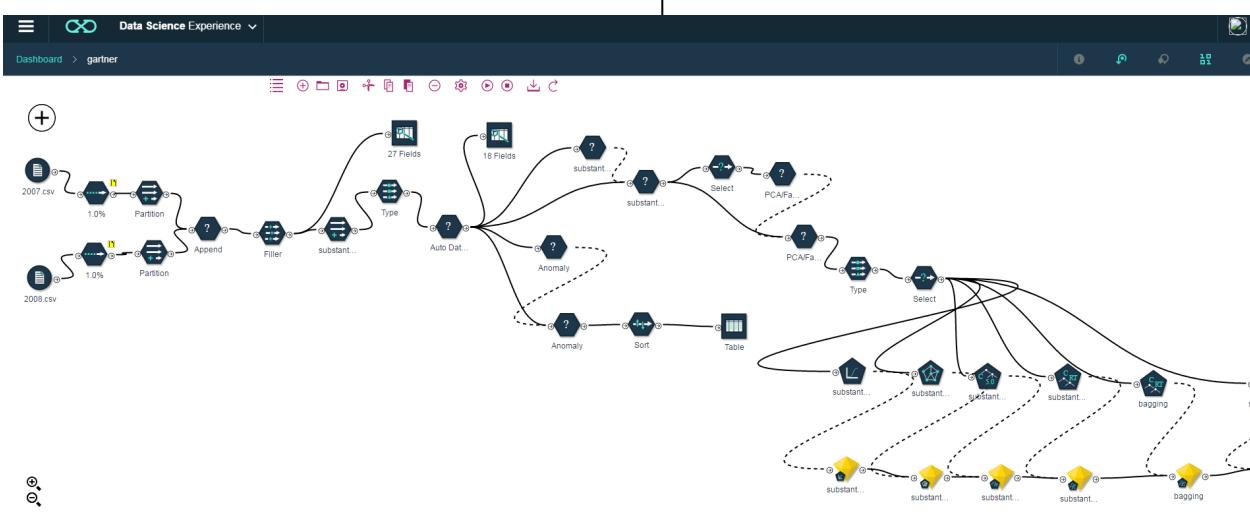
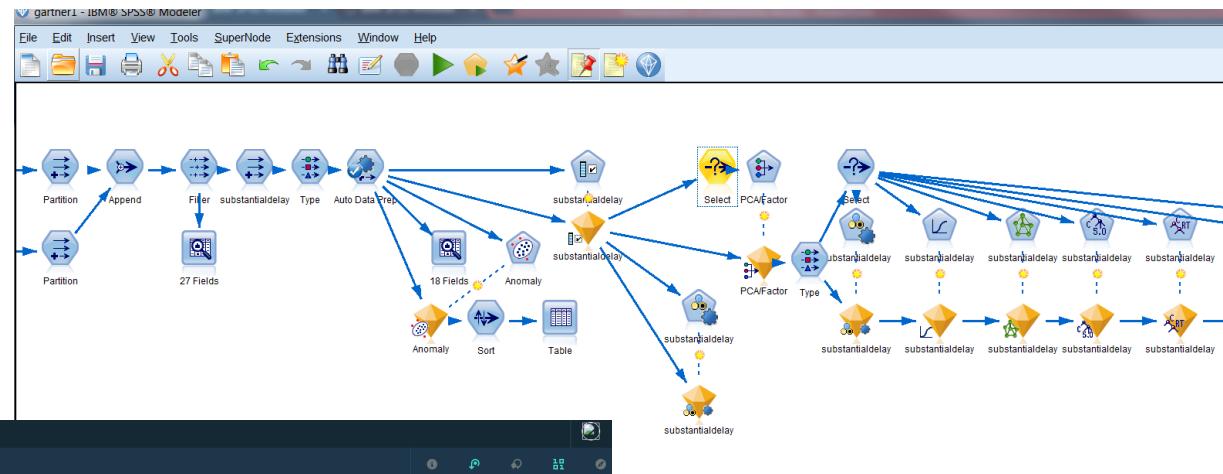
Wizard GUI



The screenshot shows the 'Deploy model' screen of the Wizard GUI. It includes sections for 'API Details' and 'Request Header', each containing several input fields represented by horizontal bars. At the bottom right, there are buttons for 'Close', 'Back', and 'Next'.

The DSX Canvas is compatible with existing SPSS Modeler Streams and can be deployed via IBM Machine Learning

- DSX Canvas has compatibility with legacy SPSS Modeler streams
- Multiple execution runtimes: SPSS Modeler, SparkML (coming soon)
- Support for R/Python/SQL code in the pipeline



DSX Canvas and Canvas One-Click Deployment via IBM ML are Roadmap Items

- One-click pipeline deployment from DSX Canvas (left) via IBM Machine Learning

Finding Data Science Experience Opportunities

1 Target the Right Buyer

- Data Scientists or Analytics Leaders who are tasked with building data-driven teams or intelligent applications leveraging data science or machine learning
- Decision makers such as Director of Analytics, Big Data Center of Excellence Leaders, VP of Data Science, LOB, CTO, Chief Data Officer and so on
- Existing analytics tools customers looking to grow their teams and expand their portfolio

2 Ask Qualifying Questions

- Do you have Data Science/Analytics teams in place?
- How do the various analytics team members collaborate?
- How do you manage open source tools for analytics?
- Do you now or in the future plan to leverage Apache Spark?
- How do you deal with data scientists needs for more tools in their portfolio?
- How do you create and deploy machine learning models today?



3 Listen for Key Words and Phrases

- Open Source Analytics, Collaborative Data Science
- RStudio, Jupyter Notebooks, R, Python, Scala, GitHub Integration, Microservices
- Single interface for predictive and prescriptive models
- Analyze large volumes of data, Apache Spark
- Scalable Pipelines and Machine Learning
- SPSS on Cloud, One-click Machine Learning
- Common project framework to manage the entire analytics lifecycle
- Shiny Applications, Decision Optimization

4 Align with other Analytics offerings

- IBM Machine Learning is available within DSX
- We will very soon have a Canvas that will support SPSS Modeler Streams and Machine Learning Pipelines
- We support R, Python and Scala so that data scientists can use their language of choice to leverage Apache Spark
- To address prescriptive analytics use cases we can leverage the DOcplexcloud API through notebooks in DSX
- As you can see DSX is not only very tightly integrated with our entire portfolio of analytics tools and services but also fits in very well into the broader Watson Data Platform

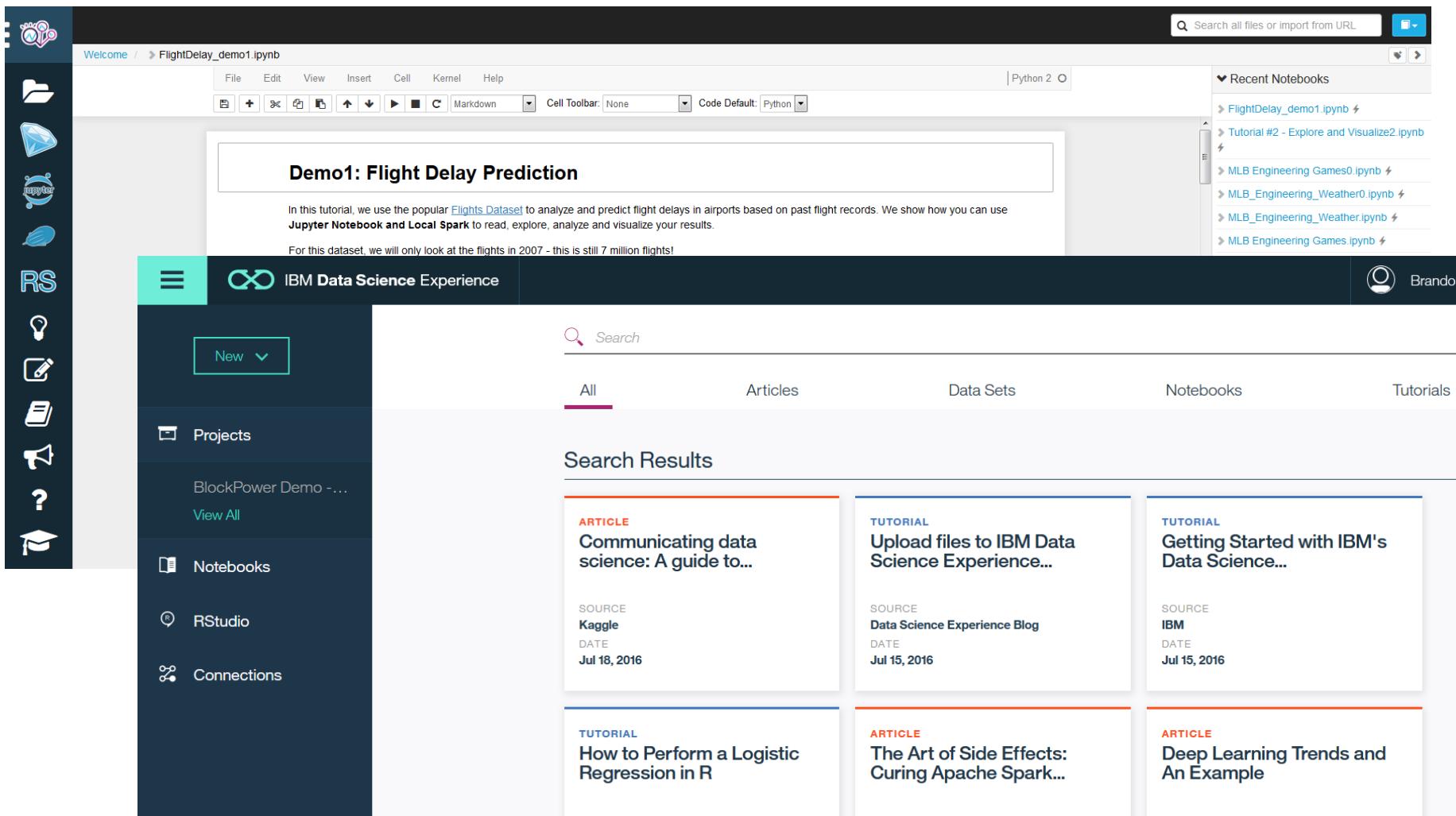
Get Started with Data Science Experience Today!

Calling all Data Science Professionals!

- Data Science Experience seamlessly integrates with the broader Watson Data Platform and is our primary experience for Data Science Professionals
- Our mission is to win the **hearts and minds** of Data Scientists
- DSX provides an integrated and collaborative environment that brings together, in a single environment, the tools and project framework needed to successfully make data science a team sport
- IBM Data Science Experience is a **freemium model** with Enterprise Plans available - **sign up** to try it out at datascience.ibm.com

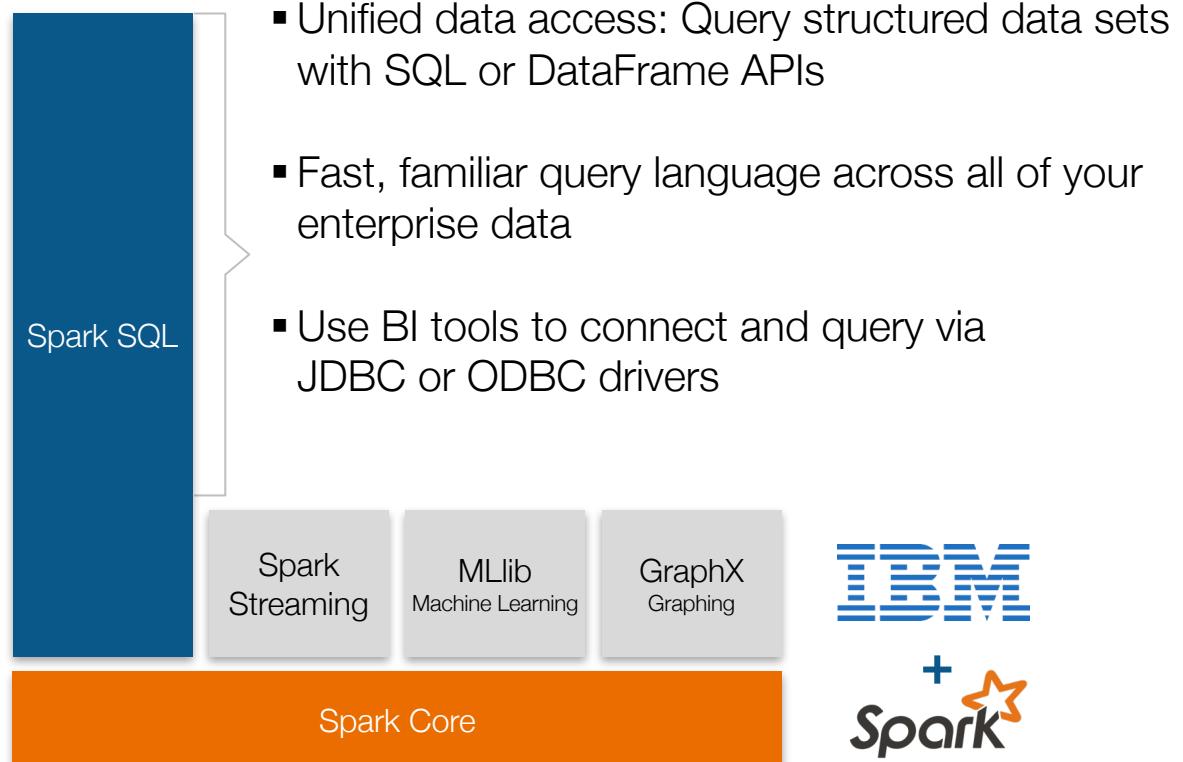
APPENDIX

Data Scientist Workbench vs Data Science Experience

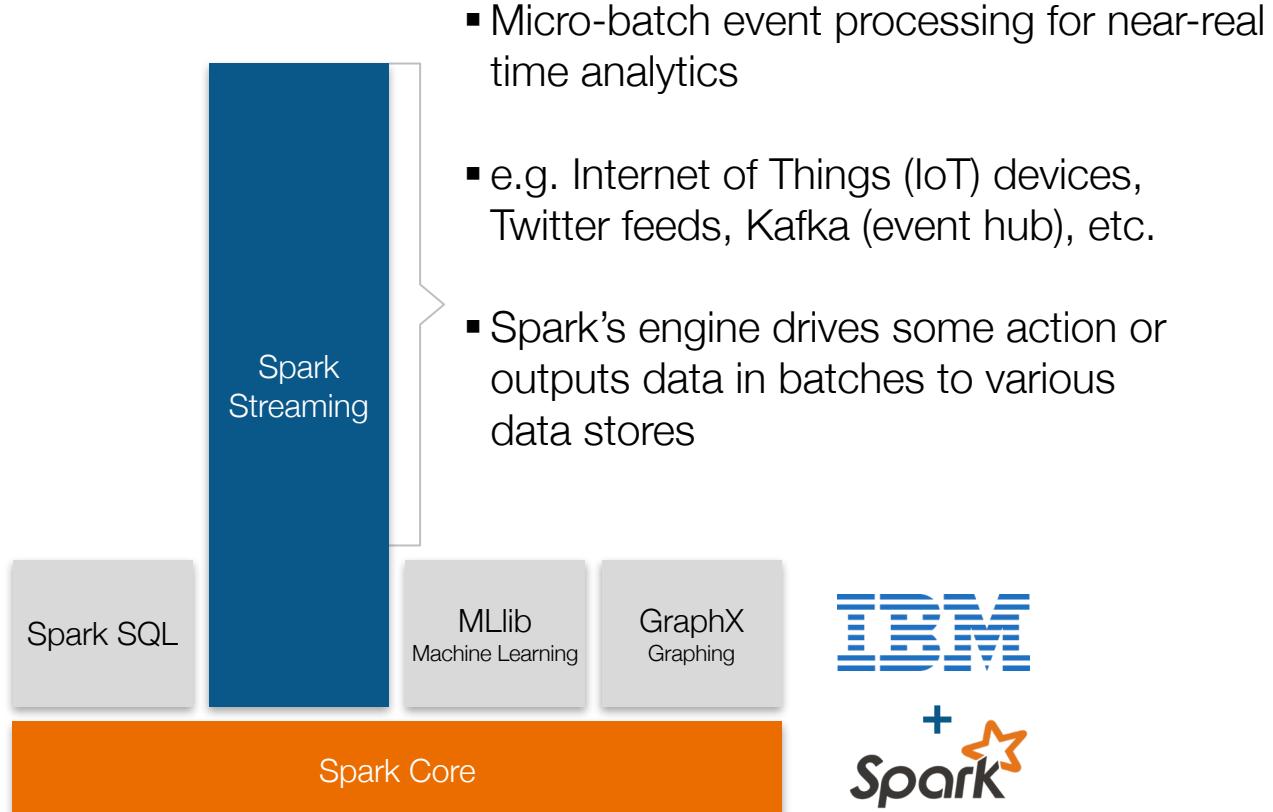


The screenshot shows the IBM Data Science Experience interface. On the left is a sidebar with various icons and links: Projects (BlockPower Demo - ..., View All), Notebooks, RStudio, and Connections. The main area has a header bar with tabs for Python 2, Cell Toolbar: None, and Code Default: Python. A search bar at the top right says "Search all files or import from URL". To the right of the search bar is a "Recent Notebooks" list containing "FlightDelay_demo1.ipynb", "Tutorial #2 - Explore and Visualize2.ipynb", "MLB_Engineering_Games0.ipynb", "MLB_Engineering_Weather.ipynb", "MLB_Engineering_Weather.ipynb", and "MLB_Engineering_Games.ipynb". The central content area displays a notebook titled "Demo1: Flight Delay Prediction" which uses the "Flights Dataset" to predict flight delays. Below the notebook is a search results page with a "Search" bar and tabs for All, Articles, Data Sets, Notebooks, and Tutorials. The "All" tab is selected. The search results show several articles and tutorials, such as "Communicating data science: A guide to...", "Upload files to IBM Data Science Experience...", "Getting Started with IBM's Data Science...", "How to Perform a Logistic Regression in R", "The Art of Side Effects: Curing Apache Spark...", and "Deep Learning Trends and An Example". Each result includes a "SOURCE" link (e.g., Kaggle, Data Science Experience Blog, IBM) and a "DATE" (e.g., Jul 18, 2016, Jul 15, 2016).

Select Libraries to Meet Use-Case Challenges

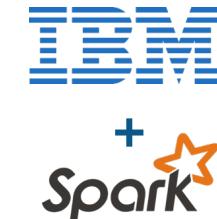
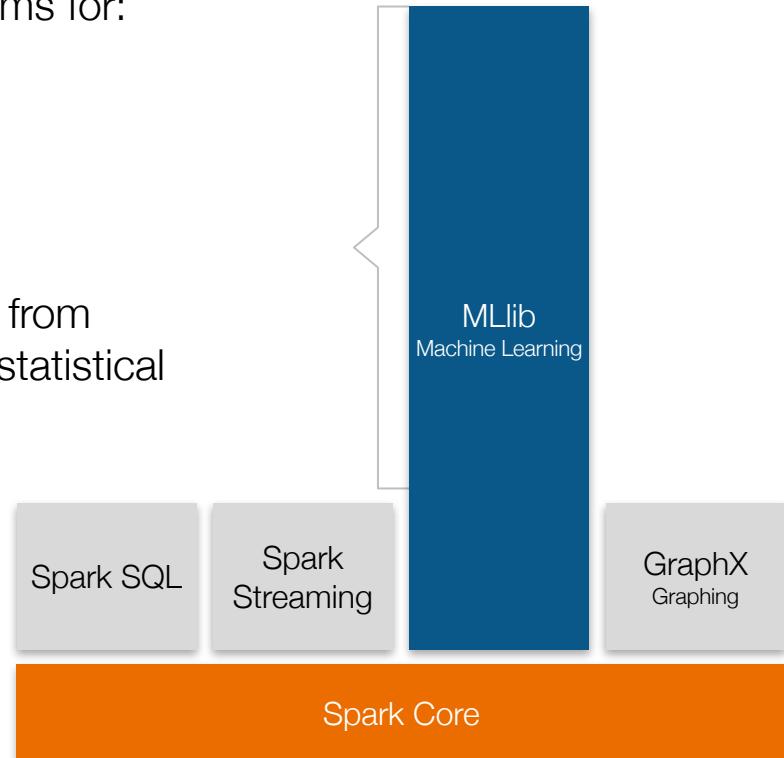


Select Libraries to Meet Use-Case Challenges



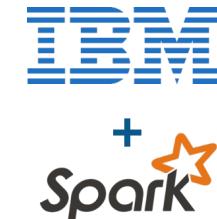
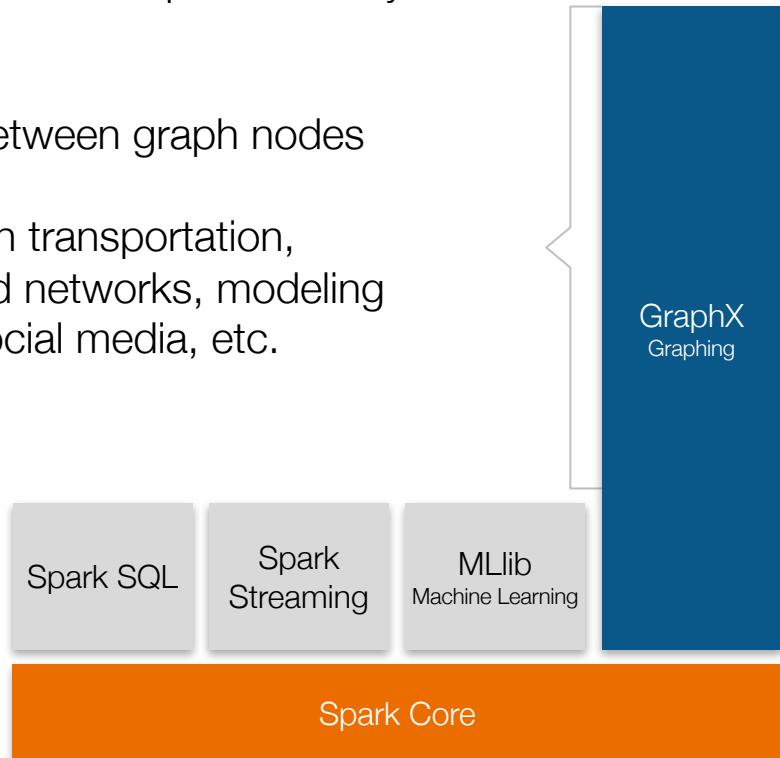
Select Libraries to Meet Use-Case Challenges

- Predictive and prescriptive analytics
- Machine learning algorithms for:
 - Clustering
 - Classification
 - Regression
 - etc.
- Smart application design from pre-built, out-of-the-box statistical and algorithmic models



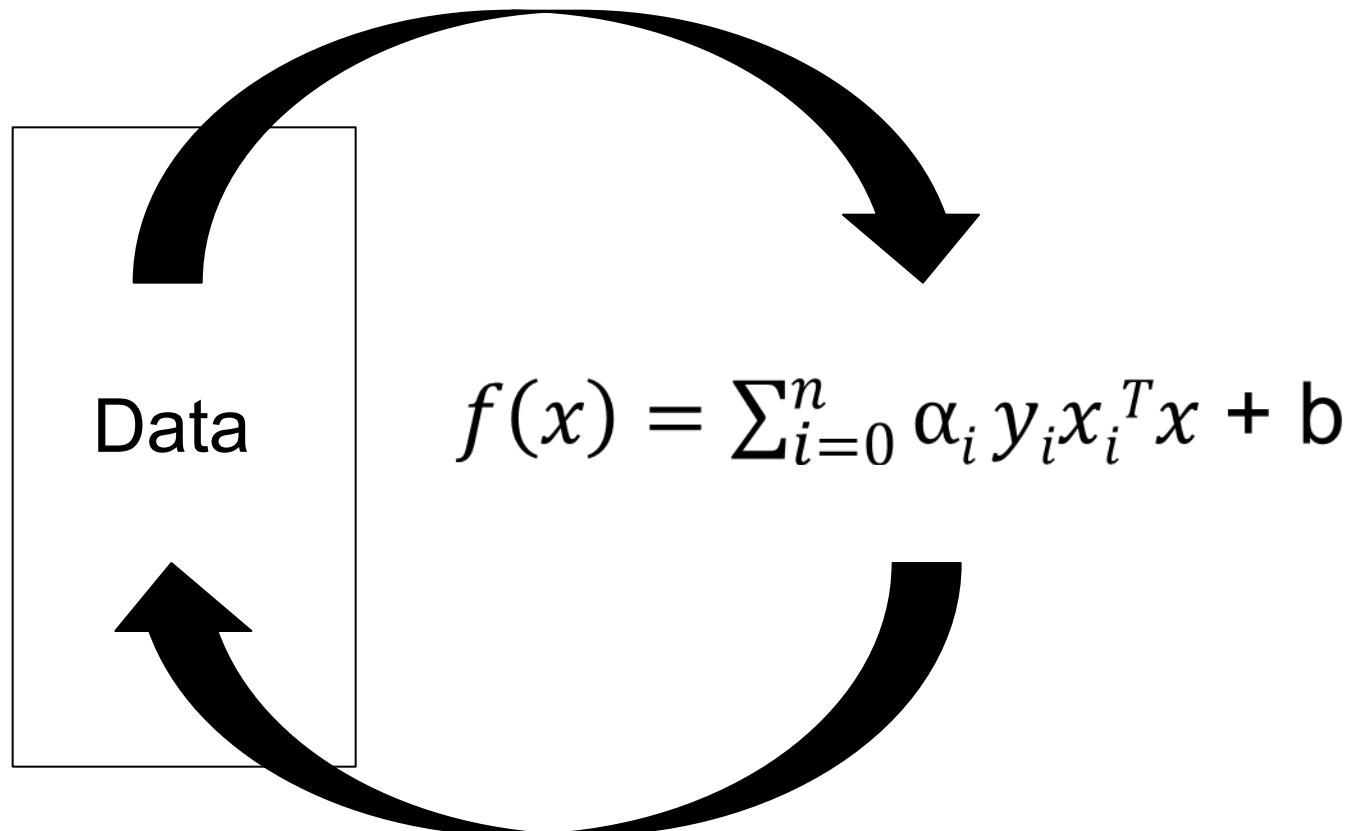
Select Libraries to Meet Use-Case Challenges

- Represent and analyze systems represented by graph nodes
- Trace interconnections between graph nodes
- Applicable to use cases in transportation, telecommunications, road networks, modeling personal relationships, social media, etc.

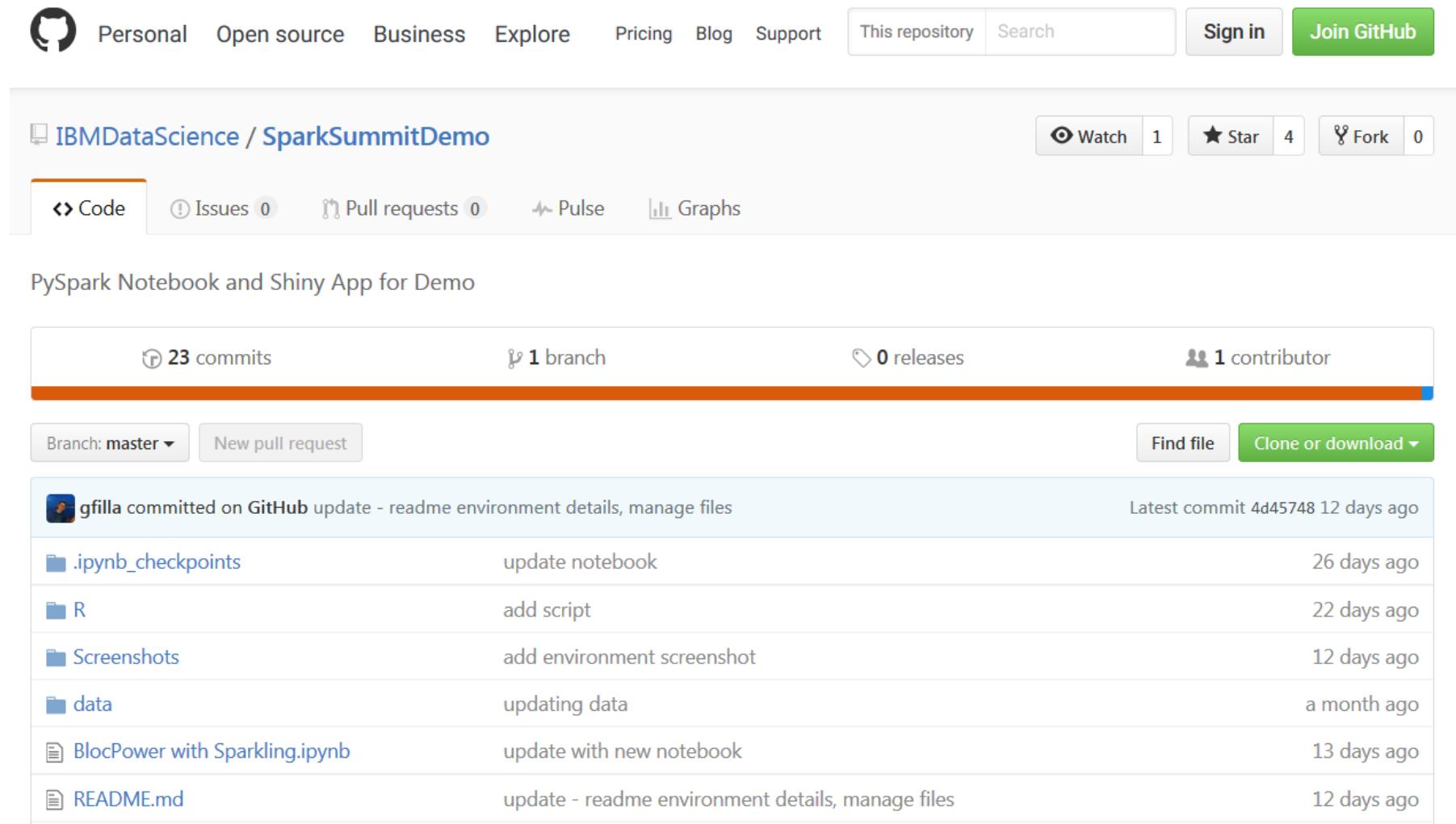


Machine Learning

- “The science of getting computers to act without being explicitly programmed”
- “Systems that can learn from data”



GitHub



The screenshot shows the GitHub repository page for `IBMDatascience / SparkSummitDemo`. The top navigation bar includes links for Personal, Open source, Business, Explore, Pricing, Blog, Support, and options to Sign in or Join GitHub. The repository header displays the name `IBMDatascience / SparkSummitDemo`, a Watch button (1), a Star button (4), and a Fork button (0). Below the header, there are tabs for Code, Issues (0), Pull requests (0), Pulse, and Graphs. A summary box shows 23 commits, 1 branch, 0 releases, and 1 contributor. The main content area lists recent commits by user `gfilla`, including updates to `ipynb_checkpoints`, `R`, `Screenshots`, `data`, and notebooks `BlocPower with Sparkling.ipynb` and `README.md`.

IBMDatascience / SparkSummitDemo

Code Issues 0 Pull requests 0 Pulse Graphs

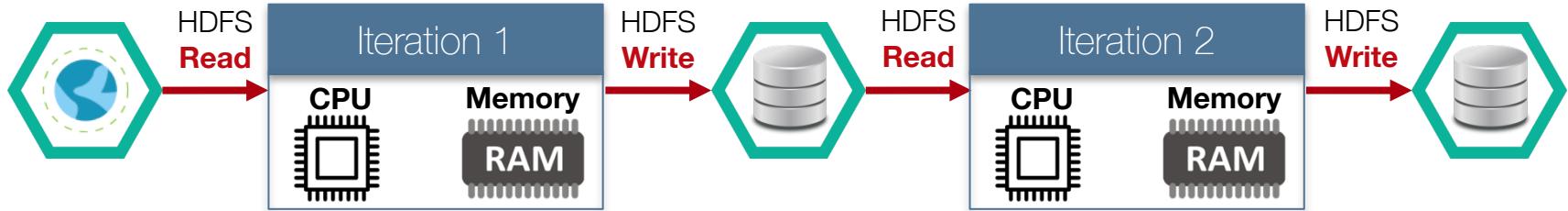
23 commits 1 branch 0 releases 1 contributor

Branch: master ▾ New pull request Find file Clone or download ▾

Commit	Message	Time
 gfilla	committed on GitHub update - readme environment details, manage files	Latest commit 4d45748 12 days ago
 .ipynb_checkpoints	update notebook	26 days ago
 R	add script	22 days ago
 Screenshots	add environment screenshot	12 days ago
 data	updating data	a month ago
 BlocPower with Sparkling.ipynb	update with new notebook	13 days ago
 README.md	update - readme environment details, manage files	12 days ago

Motivation for Apache Spark

- Traditional Approach: MapReduce jobs for complex jobs, interactive query, and online event-hub processing involves lots of (**slow**) disk I/O

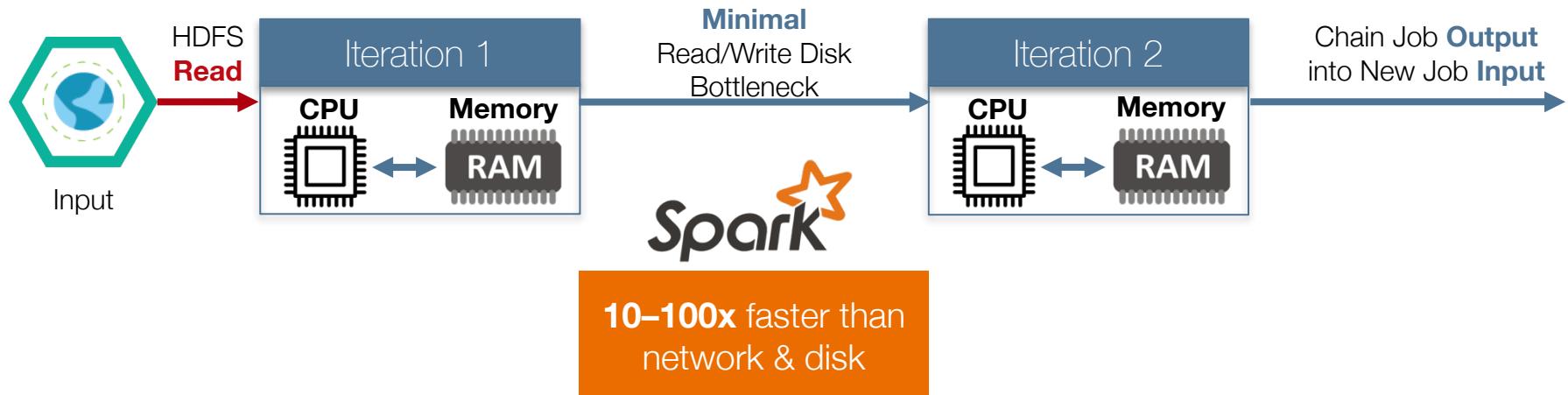


Motivation for Apache Spark

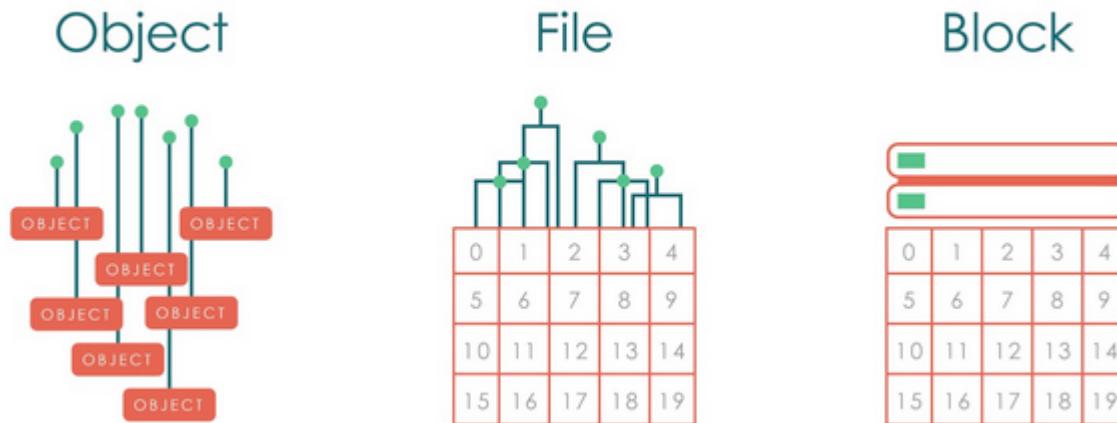
- Traditional Approach: MapReduce jobs for complex jobs, interactive query, and online event-hub processing involves lots of (**slow**) disk I/O



- Solution: Keep data **in-memory** with a new distributed execution engine



- Object storage, also called object-based storage, is a generic term that describes an approach to addressing and manipulating discrete units of storage called objects.
- Like files, objects contain data -- but unlike files, objects are not organized in a hierarchy. Every object exists at the same level in a flat address space called a storage pool and one object cannot be placed inside another object. Each object is assigned a unique identifier which allows a server or end user to retrieve the object without needing to know the physical location of the data.
- Object storage is often compared to valet parking at an upscale restaurant. When a customer uses valet parking, he exchanges his car keys for a receipt. The customer does not know where his car will be parked or how many times an attendant might move the car while the customer is dining. In this analogy, a storage object's unique identifier represents the customer's receipt.



¹<http://searchstorage.techtarget.com/definition/object-storage>

- Object storage systems allow relatively inexpensive, scalable and self-healing retention of massive amounts of unstructured data. Object storage is used for diverse purposes such as storing photos on Facebook, songs on Spotify, or files in online collaboration services, such as Dropbox.
- Object Storage provides an unstructured cloud data store to build and deliver cloud applications and services with lowered cost, reliability, and speed to market. Bluemix developers and users can access and store unstructured data content and can interactively compose and connect to applications and services. The Object Storage service also provides programmatic access via API, SDKs and a consumable UI for object management.



Object Storage

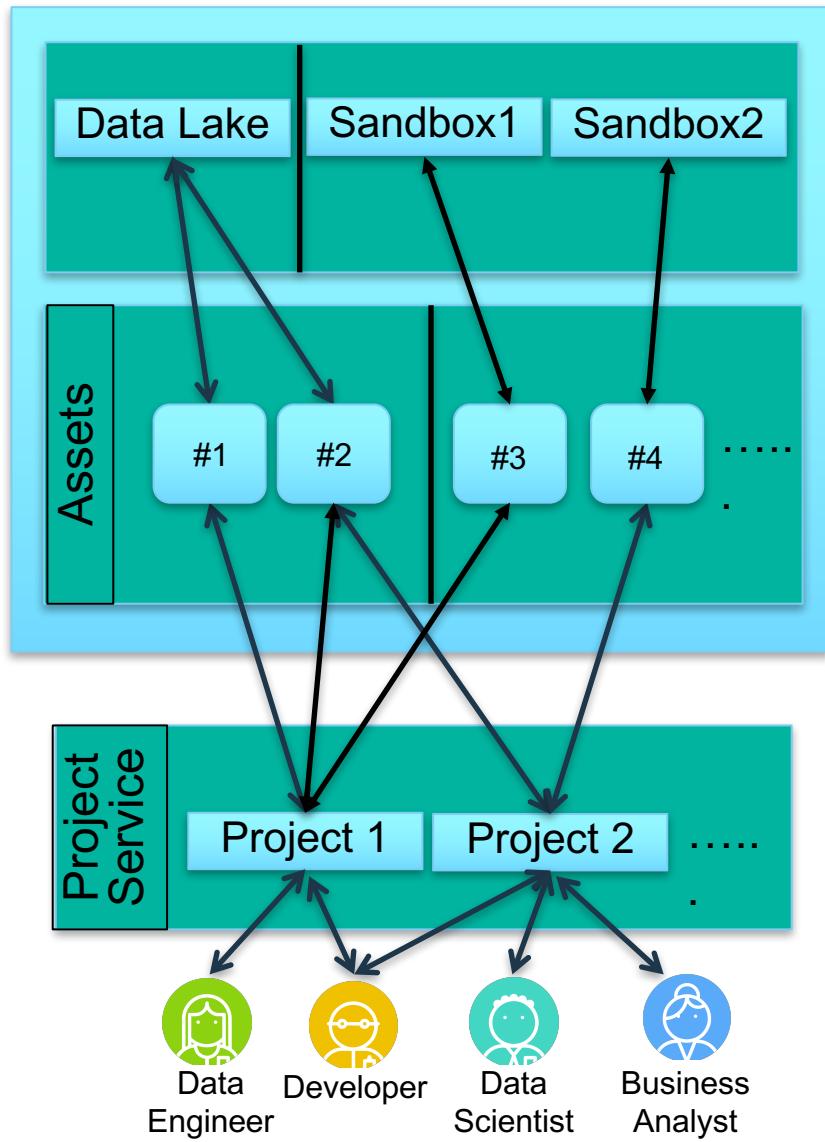
IBM

- Cloud Storage - store all your files (images, documents, and more) in the cloud. Use metadata to quickly tag and search your objects. Easily compose and bind to your object files from your Bluemix application.
- Easy Access - use drag and drop to quickly upload and manage your object store content or use industry adopted OpenStack Swift API and SDKs to access your object store programmatically.

DSX – Projects

Projects make collaboration easier

- Projects allow different users and personas to share a set of assets
- Projects enable you to collaborate and manage your notebooks, artifacts, etc.
- Projects have three levels of rights -- viewers, editors, and admin



What is the R Consortium?

The R Consortium is a group organized under an open source governance and foundation model to provide support to the R community, the R Foundation and groups and individuals, using, maintaining and distributing R software. The R Consortium aims to expand outreach and assistance of developers who are currently interested or using the R language. The R Consortium, while an independent organization, is a Collaborative Project of the Linux Foundation with the Linux Foundation providing operational support and guidance.

What is R?

The R language is an open source environment for statistical computing and graphics, and runs on a wide variety of computing platforms. The R language has enjoyed significant growth, and a broad range of industries have adopted the R language, including biotech, finance, and advertising industries. The R language is often integrated into third party analysis, visualization and reporting applications.

Why are we joining the R Consortium?

IBM is announcing its full support for the R community of 2M+ dedicated users who have traditionally been underserved. R users will benefit from IBM Analytics products that provide native support for R and deployment environments that support R. Membership in the R Consortium, whose mission is to “advance the worldwide promotion of and support for the R open source language”, demonstrates our commitment to R and the broader Data Science and Engineering community. It is important that IBM has a voice in the consortium to make sure our business interests are protected and also furthered as newer projects related to R are funded.

Overview of the membership

IBM will join Microsoft and RStudio as one of the three platinum members. A platinum membership gives IBM a seat on the board. Additionally, it also gives use a seat on the infrastructure steering committee that funds projects related to R. We will leverage the platform provided by the consortium to drive towards deeper integration of R into Spark. Also it helps to make sure IBM has a voice as changes to the R language and surrounding packages are made to for distributed computing, Spark, modern Big Data platforms, etc.

Who is RStudio?

RStudio is a provider of open-source and enterprise-ready commercial tools for the R community. Founded in 2008, it is headquartered in Boston, MA. Inspired by the innovations of R users in science, education, and industry, RStudio develops free and open tools for R and enterprise-ready professional products for teams to scale and share work.

Why are we partnering with RStudio?

A partnership with Rstudio bring together IBM's Big Data & Analytics technology depth and services breadth with RStudio's platform and expertise on R. We are mutually aligned on the goal to make R a first-class citizen in Spark.

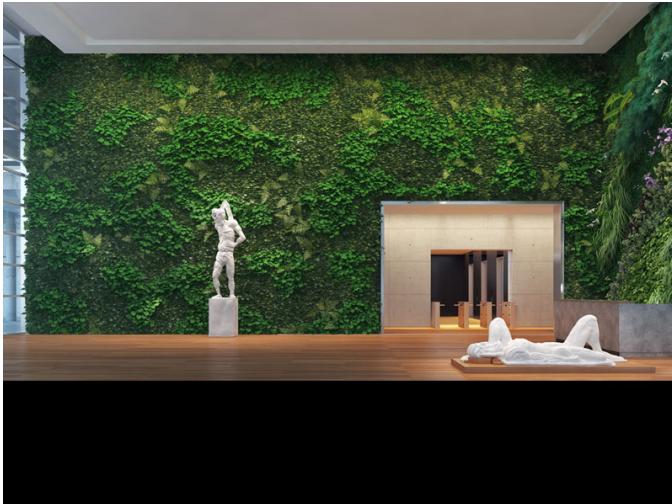
Overview of partnership

IBM and RStudio will collaborate to enable R packages to integrate with Apache Spark. R is a popular statistical programming language that offers a rich environment for statistical analysis and machine learning. The runtime for R is single-threaded, and by ensuring R code and packages work seamlessly with Spark, R users can leverage Spark's distributed computational engine to run large-scale data analysis from R.

IBM will integrate RStudio's open source offerings (RStudio Server and Shiny Server) in the Data Science Experience offering. RStudio Server is an Integrated Development Environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management. Shiny Server is an open source web application server that combines the power of R with the interactivity of the modern web providing the deployment needed for Shiny applications. Additionally, In the Data Science Experience offering, IBM will expose tutorials and news from RStudio blogs.

Spark Technology Center

505 Howard Street, San Francisco



IBM established Spark Technology Center to contribute to the Apache® Spark™ ecosystem – June 2015

IBM Spark Technology Center (STC)
San Francisco, USA

Growing pool of contributors
~50 world wide, and 3 committers

Apache SystemML now an official Apache Incubator project

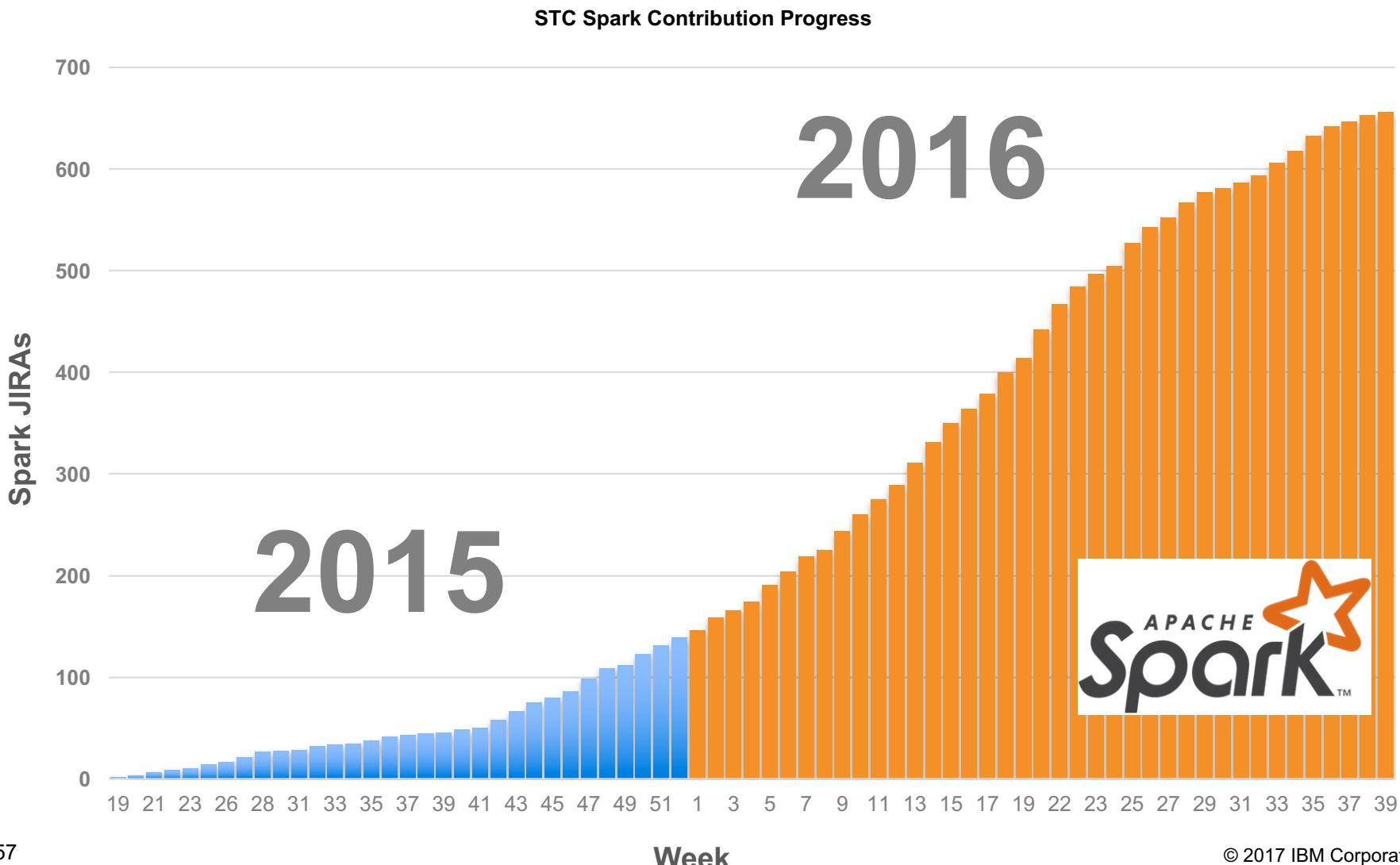
Founding member of AMPLab (and upcoming RISE Lab)

Member of R Consortium

Founding member of Scala Center

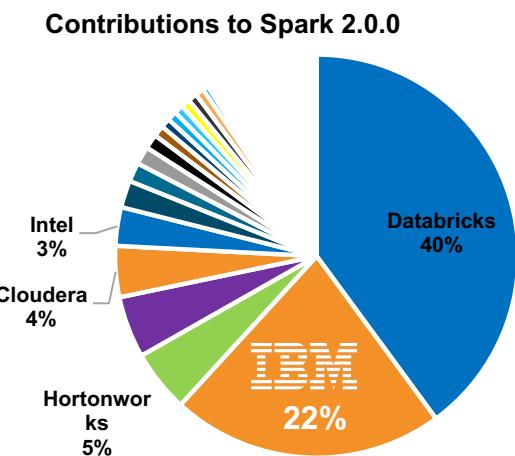
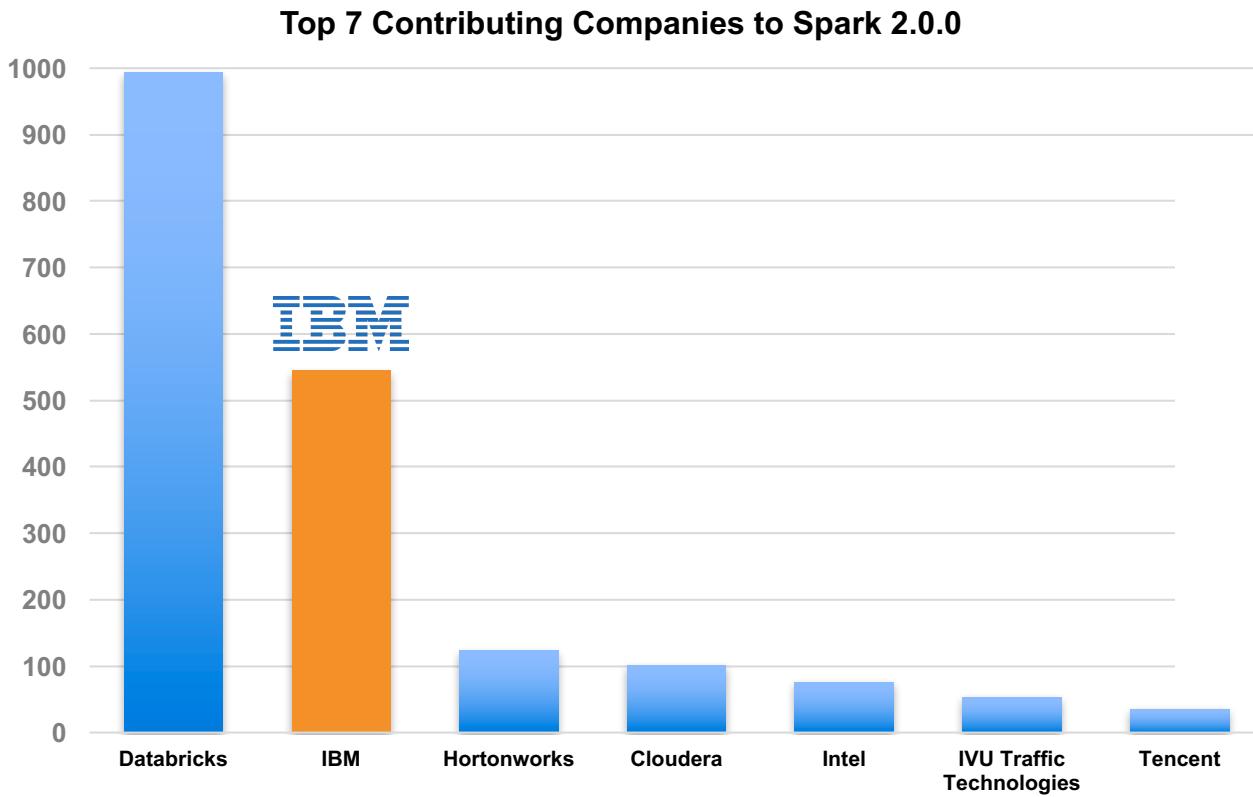
Partnerships in the ecosystem

Spark Technology Center contributions have grown over 400% since start in June 2015



IBM had a significant impact on Spark 2.0

- IBM is #2 contributor to Apache Spark
- IBM was the leading contributor in Spark 2.0 to SparkML, PySpark, and SparkR



Spark Infused Across IBM Analytics Portfolio

Free and Open Data	<ul style="list-style-type: none">Analytics Exchange
Data Storage	<ul style="list-style-type: none">On-Premises: IBM Open Platform with Apache Hadoop (IOP), BigInsights, Netezza, Cloudant, DB2, dashDB local and InformixOn-Cloud: Cloudant, dashDB, Object Storage, SQL DB, BigInsights
Data Feeds, Load & Refinement	<ul style="list-style-type: none">Watson Data PlatformIBM StreamsIBM Insights for TwitterIBM Insights for Weather
Analytics and Solutions	<ul style="list-style-type: none">IBM Analytics for Apache sparkSPSS Modeler and Analytics ServerWatson AnalyticsWatson HealthIBM CommerceData Science Experience
Learning Tools	<ul style="list-style-type: none">Big Data University

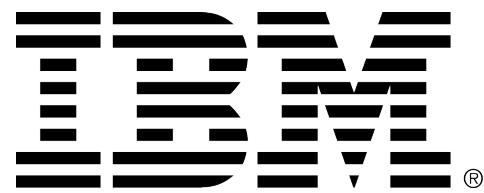
IBM Apache Spark References



RYERSON UNIVERSITY



BERNHARDT



Legal Disclaimer

- © IBM Corporation 2015. All Rights Reserved.
- The information contained in this publication is provided for informational purposes only. While efforts were made to verify the completeness and accuracy of the information contained in this publication, it is provided AS IS without warranty of any kind, express or implied. In addition, this information is based on IBM's current product plans and strategy, which are subject to change by IBM without notice. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, this publication or any other materials. Nothing contained in this publication is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers or licensors, or altering the terms and conditions of the applicable license agreement governing the use of IBM software.
- References in this presentation to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates. Product release dates and/or capabilities referenced in this presentation may change at any time at IBM's sole discretion based on market opportunities or other factors, and are not intended to be a commitment to future product or feature availability in any way. Nothing contained in these materials is intended to, nor shall have the effect of, stating or implying that any activities undertaken by you will result in any specific sales, revenue growth or other results.
- If the text contains performance statistics or references to benchmarks, insert the following language; otherwise delete:
Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.
- If the text includes any customer examples, please confirm we have prior written approval from such customer and insert the following language; otherwise delete:
All customer examples described are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer.
- Please review text for proper trademark attribution of IBM products. At first use, each product name must be the full name and include appropriate trademark symbols (e.g., IBM Lotus® Sametime® Unyte™). Subsequent references can drop "IBM" but should include the proper branding (e.g., Lotus Sametime Gateway, or WebSphere Application Server). Please refer to <http://www.ibm.com/legal/copytrade.shtml> for guidance on which trademarks require the ® or ™ symbol. Do not use abbreviations for IBM product names in your presentation. All product names must be used as adjectives rather than nouns. Please list all of the trademarks that you use in your presentation as follows; delete any not included in your presentation. IBM, the IBM logo, Lotus, Lotus Notes, Notes, Domino, Quickr, Sametime, WebSphere, UC2, PartnerWorld and Lotusphere are trademarks of International Business Machines Corporation in the United States, other countries, or both. Unyte is a trademark of WebDialogs, Inc., in the United States, other countries, or both.
- If you reference Adobe® in the text, please mark the first use and include the following; otherwise delete:
Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
- If you reference Java™ in the text, please mark the first use and include the following; otherwise delete:
Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.
- If you reference Microsoft® and/or Windows® in the text, please mark the first use and include the following, as applicable; otherwise delete:
Microsoft and Windows are trademarks of Microsoft Corporation in the United States, other countries, or both.
- If you reference Intel® and/or any of the following Intel products in the text, please mark the first use and include those that you use as follows; otherwise delete:
Intel, Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
- If you reference UNIX® in the text, please mark the first use and include the following; otherwise delete:
UNIX is a registered trademark of The Open Group in the United States and other countries.
- If you reference Linux® in your presentation, please mark the first use and include the following; otherwise delete:
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both. Other company, product, or service names may be trademarks or service marks of others.
- If the text/graphics include screenshots, no actual IBM employee names may be used (even your own), if your screenshots include fictitious company names (e.g., Renovations, Zeta Bank, Acme) please update and insert the following; otherwise delete: All references to [insert fictitious company name] refer to a fictitious company and are used for illustration purposes only.