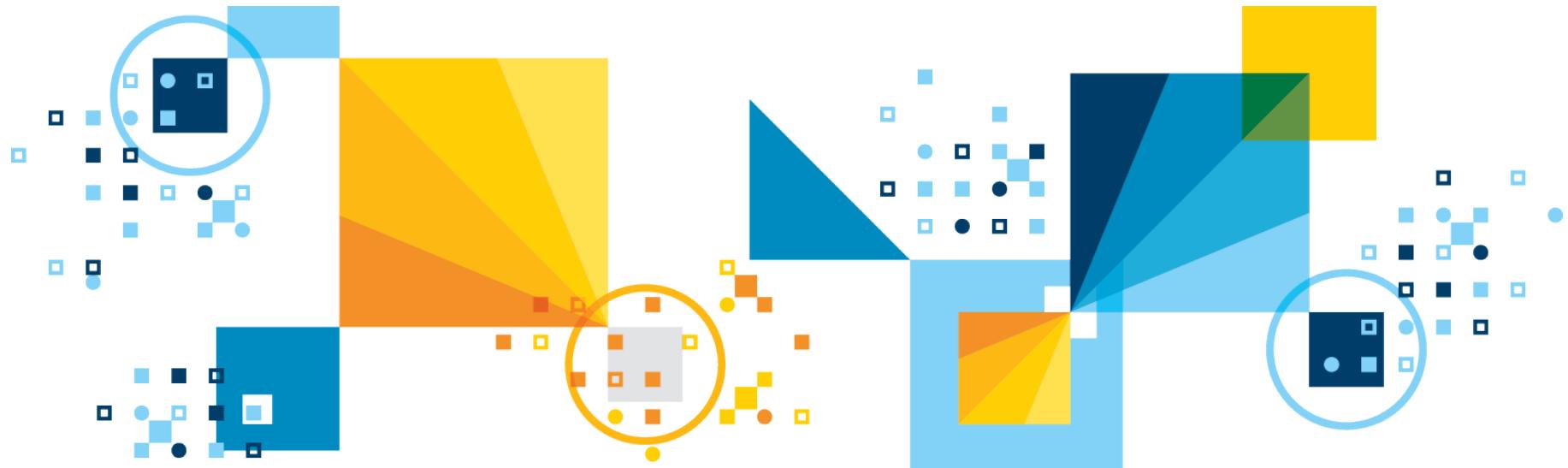
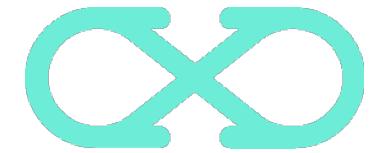


IBM Data Science Experience Overview



Introducing the Data Science Experience



Learn

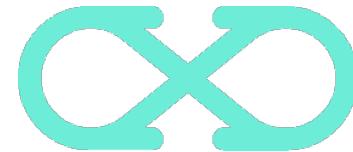
Built-in learning to get started or go the distance with advanced tutorials

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>

Data Scientist Persona - Current Challenges and Pain Points

▪ 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



IBM Data Science Experience
[Click here to watch video](#)



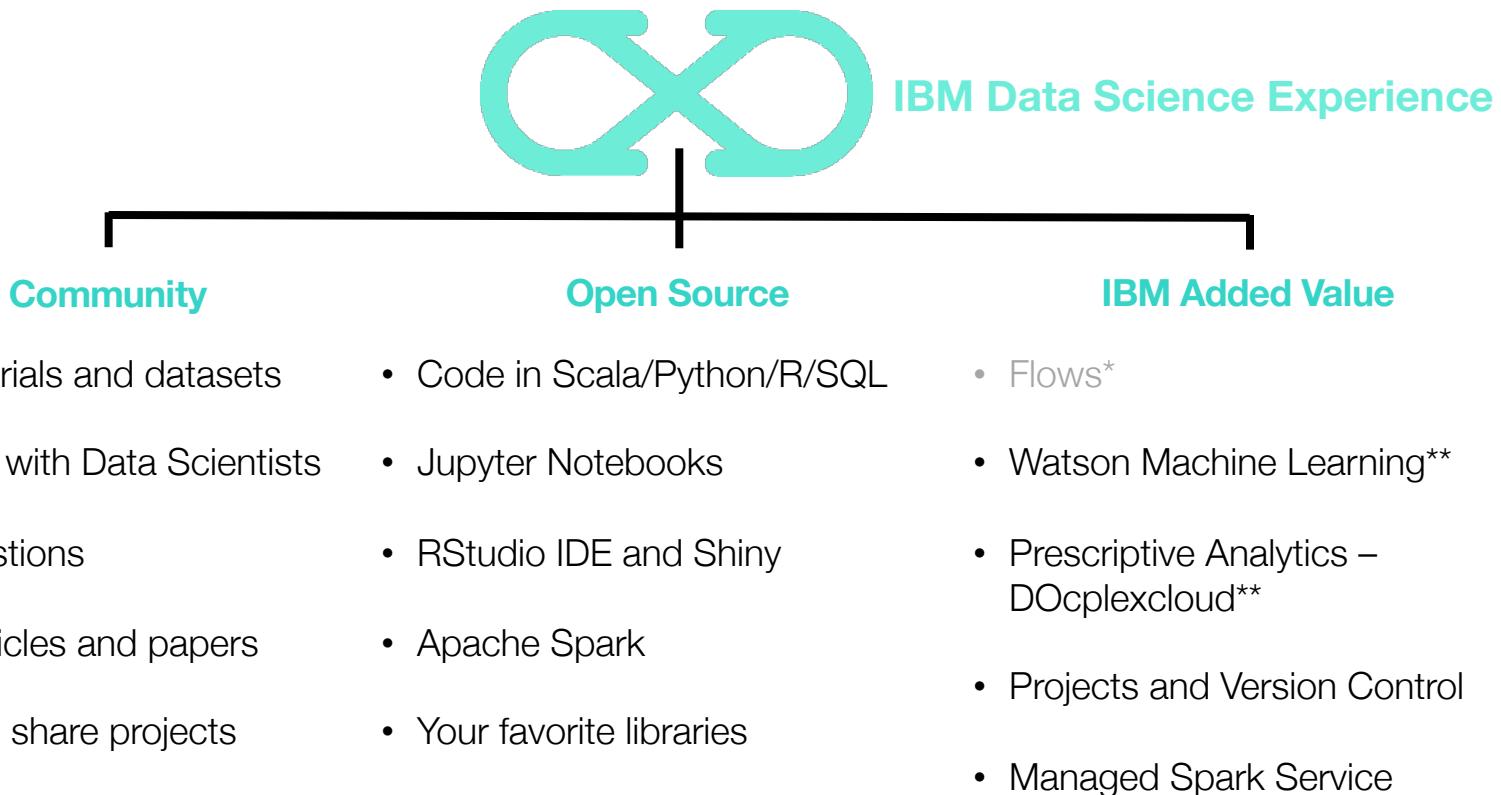
IBM Data Science Experience

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

** Integrated with DSX, but separate part numbers need to be purchased to use this capability

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
ghcn-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

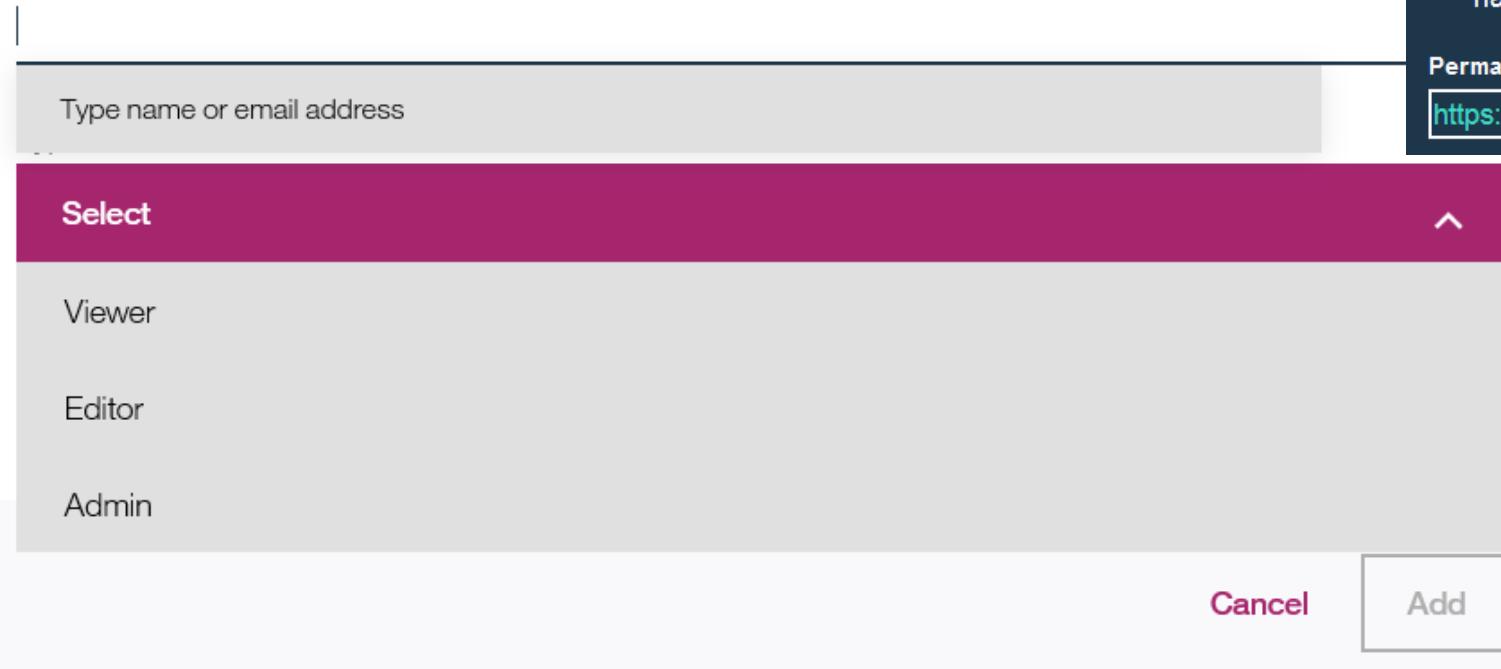
Analyze open data sets using pandas ...

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 a new collaborator to a project. At the top right, there is a 'Sharing' section with a subtitle: 'Sharing a notebook enables other users to view your notebook content.' A checked checkbox says 'Share with anyone who has the link.' Below this is a 'Permalink to view notebook' with the URL <https://apsportal.ibm.com/an>. The main area is titled 'Select' and contains three options: 'Viewer', 'Editor', and 'Admin'. At the bottom right of this area are 'Cancel' and 'Add' buttons.

Type name or email address

Select

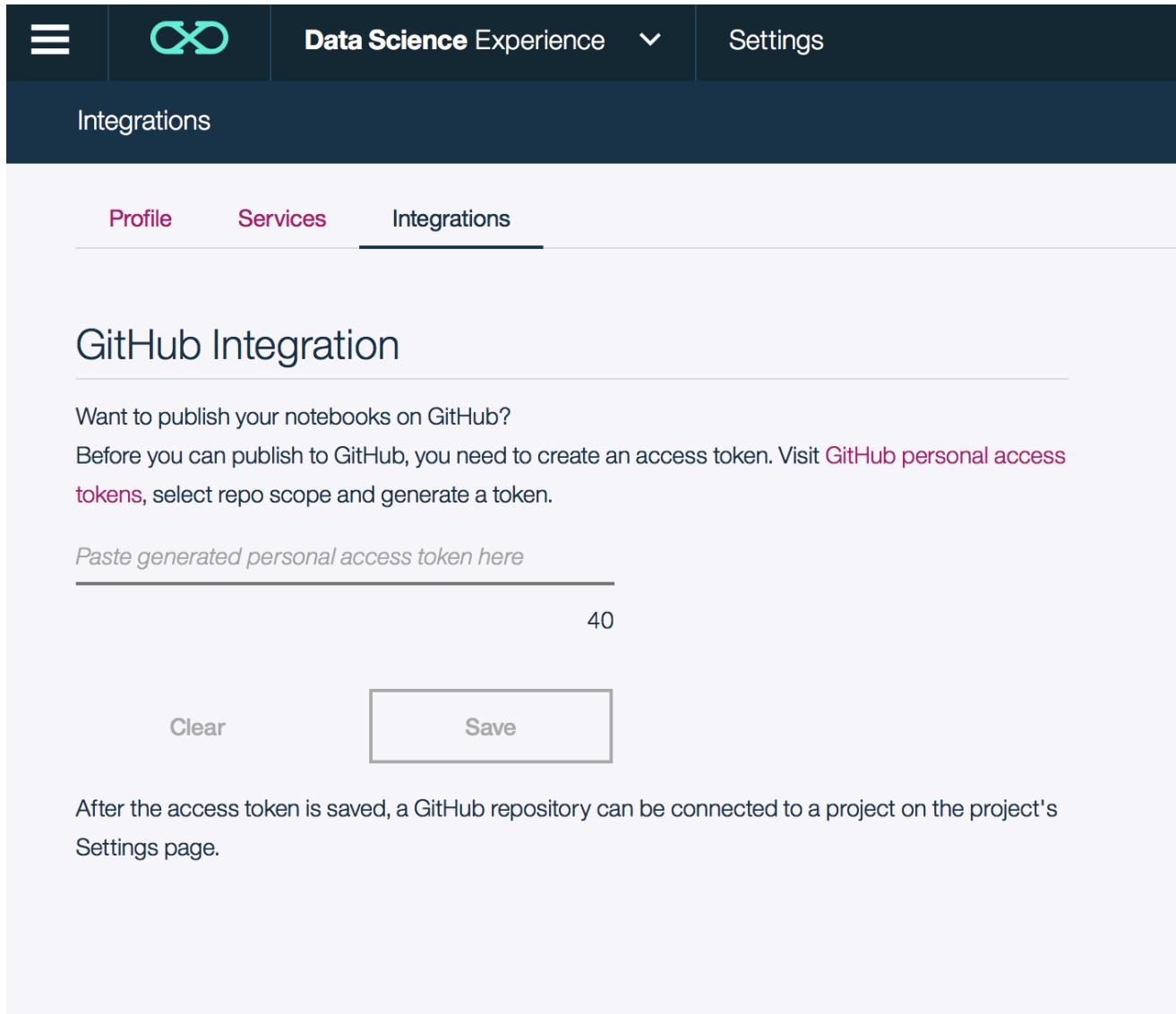
Viewer

Editor

Admin

Cancel Add

GitHub Integration



The screenshot shows the IBM Data Science Experience interface with the "Integrations" tab selected. The main section is titled "GitHub Integration" and contains instructions for publishing notebooks to GitHub, mentioning the need for a personal access token. A text input field is provided for pasting the token, with a character count indicator "40". Below the input 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

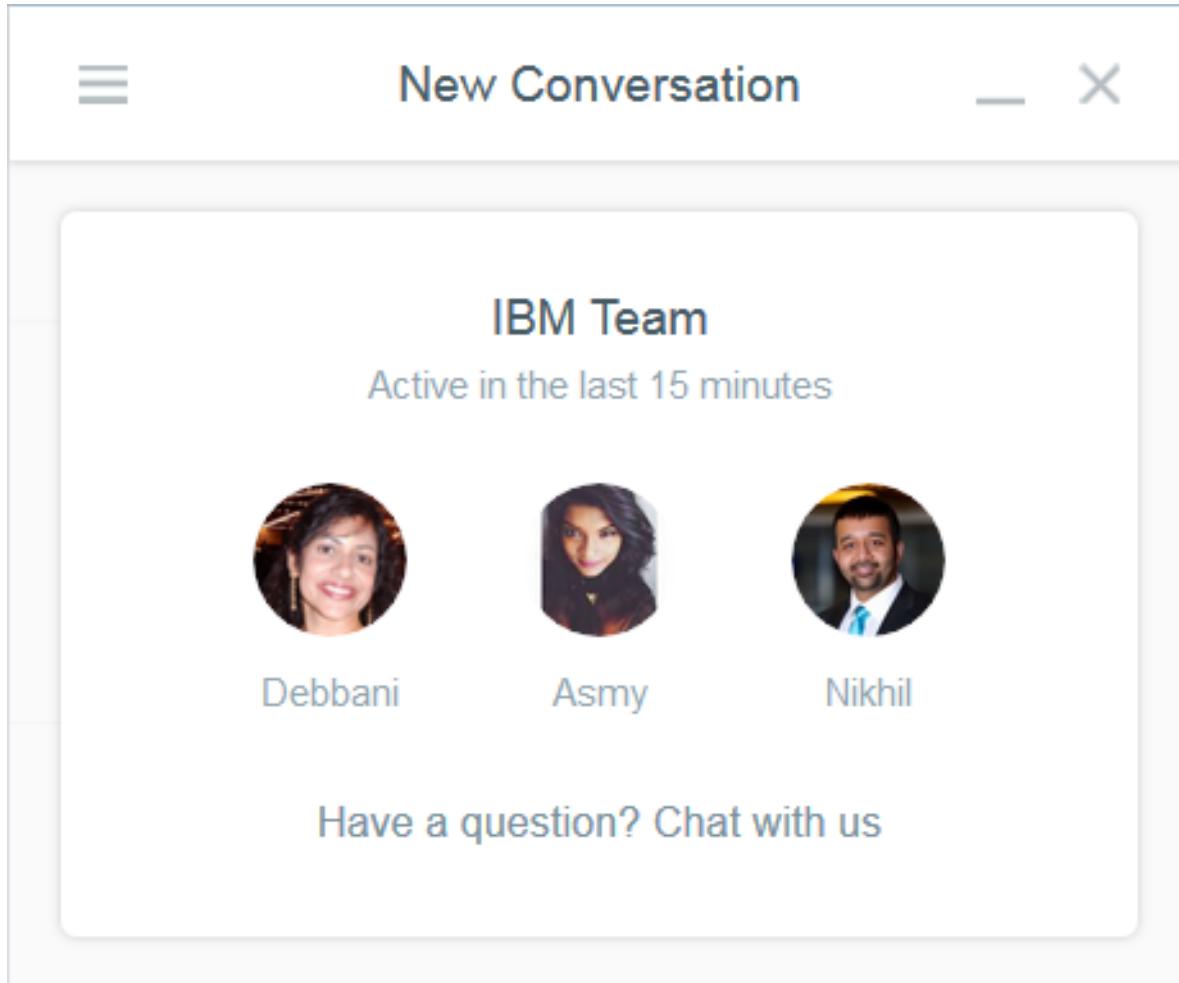
Clear Save

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>

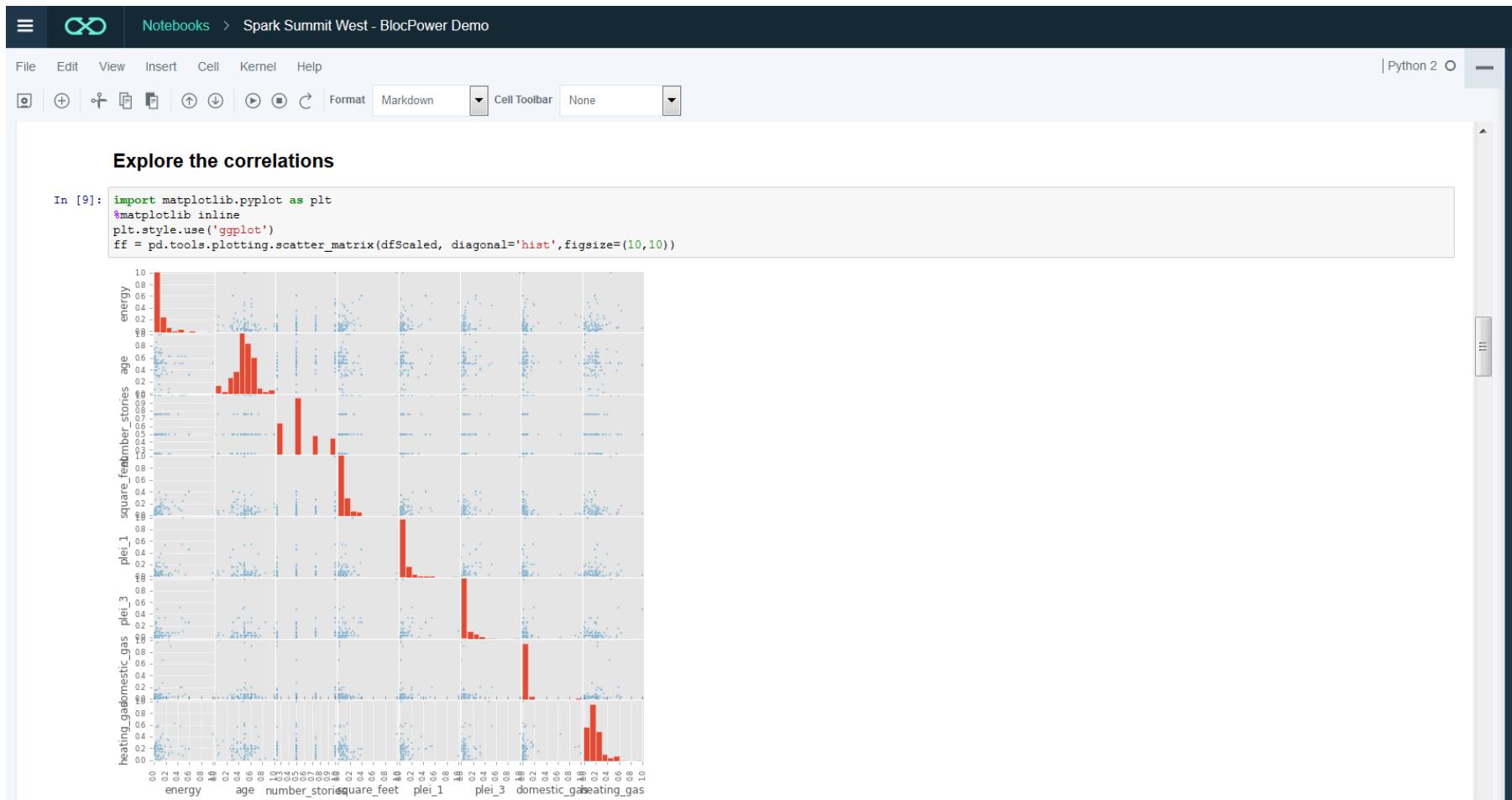
Live chat on Intercom for support from the IBM team and to provide your feedback on how we can improve DSX



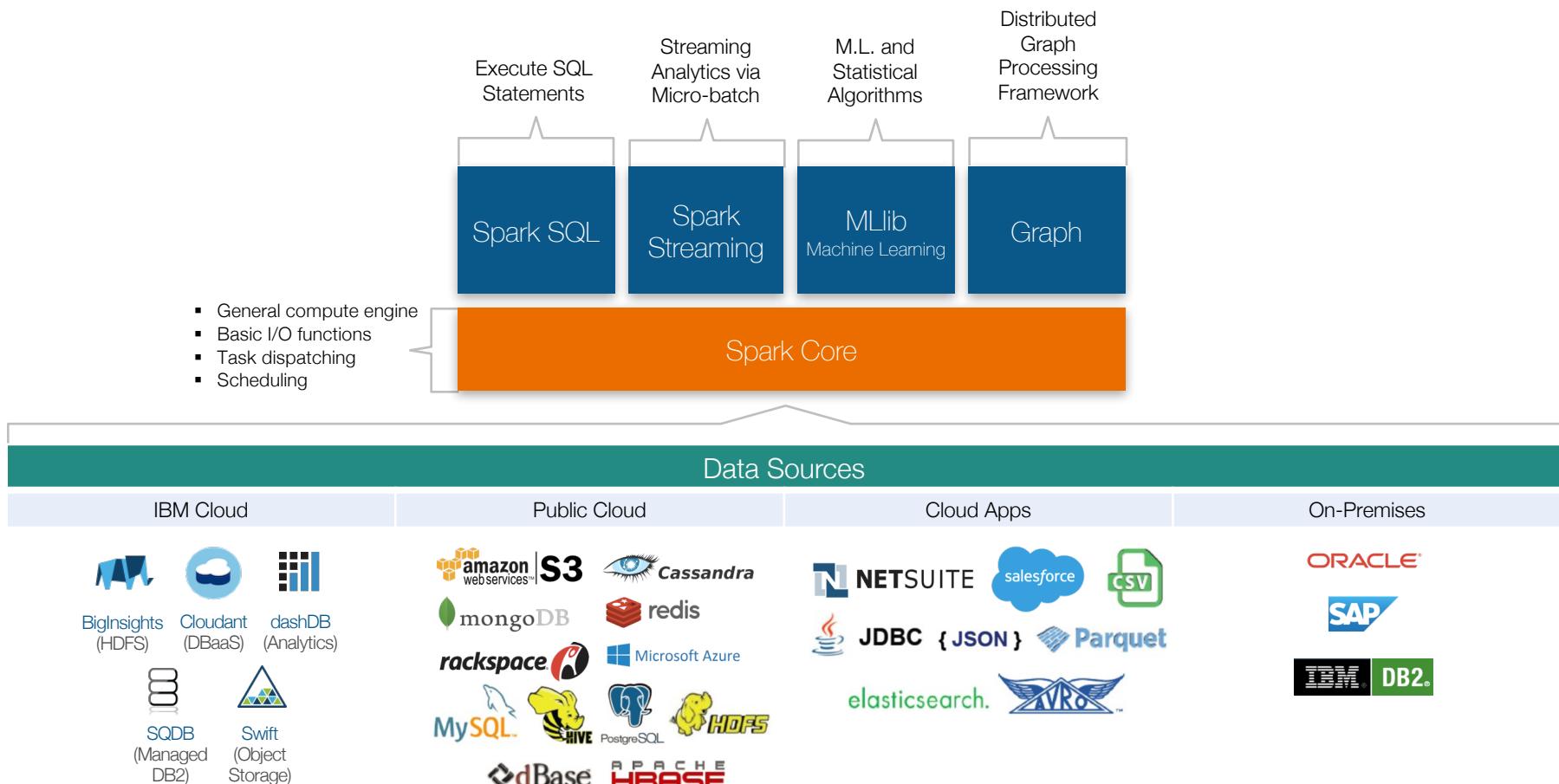
The screenshot shows the 'New Conversation' interface of Intercom. At the top, it says 'New Conversation' with a minimize and close button. Below that, it displays 'IBM Team' which was 'Active in the last 15 minutes'. Three team members are listed with their names and profile pictures: Debbani, Asmy, 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



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		

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

The screenshot displays the RStudio IDE interface. On the left, the code editor shows R script code for data import, summary plots, and calendar plots. A tooltip is visible over the 'annotate' argument in a 'calendarPlot' call, 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 console window below shows the execution of the R code. To the right, a large calendar heatmap for the year 2006 is displayed, with each cell's color representing the value of the 'o3' pollutant for that specific date. The color scale ranges from light yellow (low values) to dark red (high values), with a legend on the far right indicating values from 20 to 100.

```

File Edit View Workspace Plots Help
RStudioTest.R* | Run Line(s) Run All
Source on Save
library(openair)
## import some example data
bloomsbury <- importKCL(site = "blo0", 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
Data
bloomsbury      50804 obs. of 18 variables
  
```

Console

```

>
>
>
>
> library(openair)
## import some example data
bloomsbury <- importKCL(site = "blo0", 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; mg/m3 for CO
PM10_raw is raw data multiplied by 1.3
Warning message:
In importKCL(site = "blo0", 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"   "lp"      "rhum"   "numeric" "numeric"
  "character" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
> calendarPlot(bloomsbury, pollutant = "o3", year = 2006)
  
```

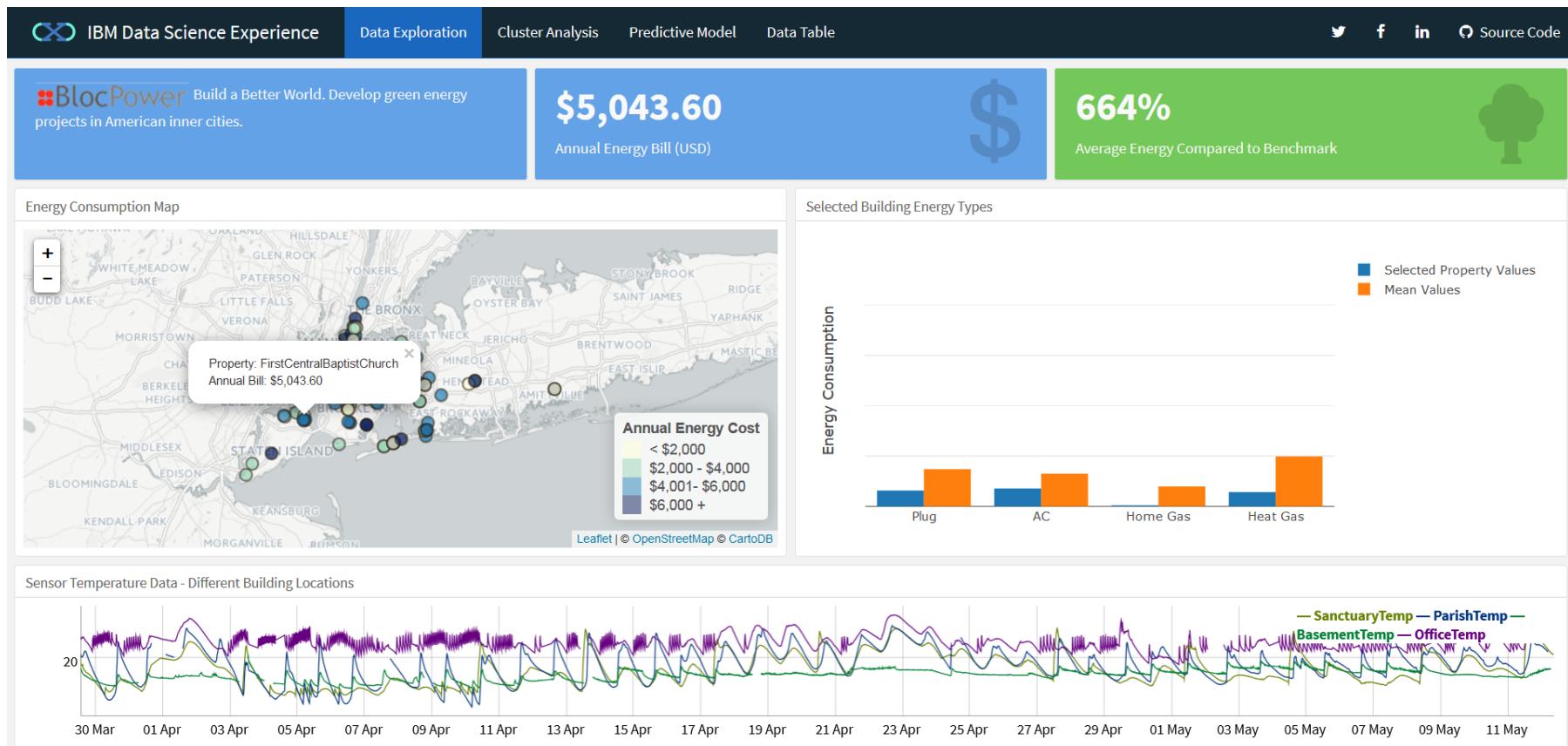
Workspace History

January February March April May June July August September October November December

O₃ in 2006

100
80
60
40
20

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



Interactively explore the analysis of your data science team

BlocPower Demo v3 - Blue x **Data & Analytics Portal** x **IBM Data Science Experience** x

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

<https://apsnginxrstudio.stage1.mybluemix.net/node1/rstudio40008/p/4456/shinyDemo.Rmd#cluster-analysis>

Greg

IBM Data Science Experience Data Exploration Cluster Analysis Predictive Model Data Table

Twitter **f** **in** **Source**

Energy Consumption Map

The clustering model help us identify buildings.

- Note in the above figure that most buildings are
- Buildings that are part of the brown, yellow and light blue ones that are part of the purple cluster
- Labels are re-coded into a binary variable where 1 means building belongs to the purple 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')
```

To [22]: from sklearn.metrics import confusion_matrix

Clusters by Heating and Plug Consumption

CommunityBaptistChurch

Cluster Labels

Efficient

Inefficient

Leaflet | © OpenStreetMap © CartoDB

Adjust parameters on-the-fly and visualize model predictions

BlocPower Demo v3 - Blue

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

Linear Regression Model.

Hypothesis: energy usage (kWh) can be predicted by the following 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

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
	0.105810

Data & Analytics Portal

IBM Data Science Experience

Predict Energy Use and Cost for New Property

Enter Age of Property:

Enter number of Stories:

Enter Property Square Footage:

Enter Age of Property:

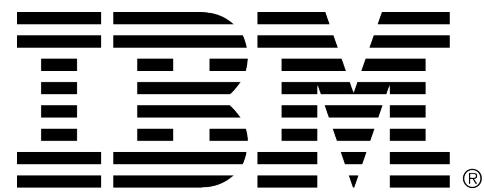
Predicted Annual Energy Bill (USD @ \$0.18/kwh)

43.7K\$

Predicted Annual Energy (kwh)

242.8K

Regression Results: Actual Vs. Predicted Values



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.