

Hands on Introduction to Data Science and IBM's Data Science Experience



Power of data. Simplicity of
design. Speed of innovation.

Joel Patterson
Piotr Mierzejewski

Hands on Introduction to Data Science Experience

Agenda

9:00 – 9:30 - Kick off

Overview of Data Science Experience (DSX), DSX Local and DSX Desktop

9:30 - 10:30 - Lab 1 – Learning Data Science Experience / Bluemix

Notebook basics, connecting to external sources

10:30 – 11:30 - Lab 2 – Machine Learning for Classification

Reading from external sources, versioning, scheduling

11:30 – 12:30 - Lab 3 – R, Shiny and GUI Interfaces

RStudio, Shiny

12:30 – 1:30 - Lunch**1:30 - 2:30 - Optional Labs**

Decision Optimization

Visualization

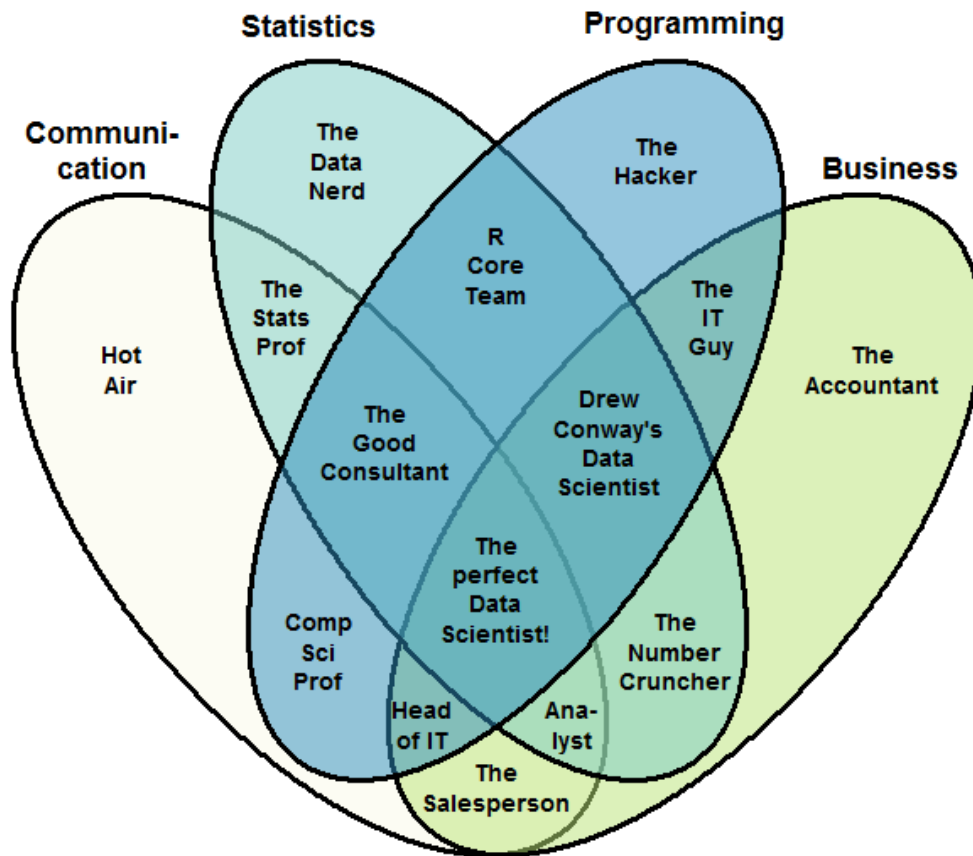
Data connections

Scoring

2:30 - 3:30 – DSX Local presentation and demo**3:30 - 4:00 – Questions and Wrap-up**

The perfect Data Science Team

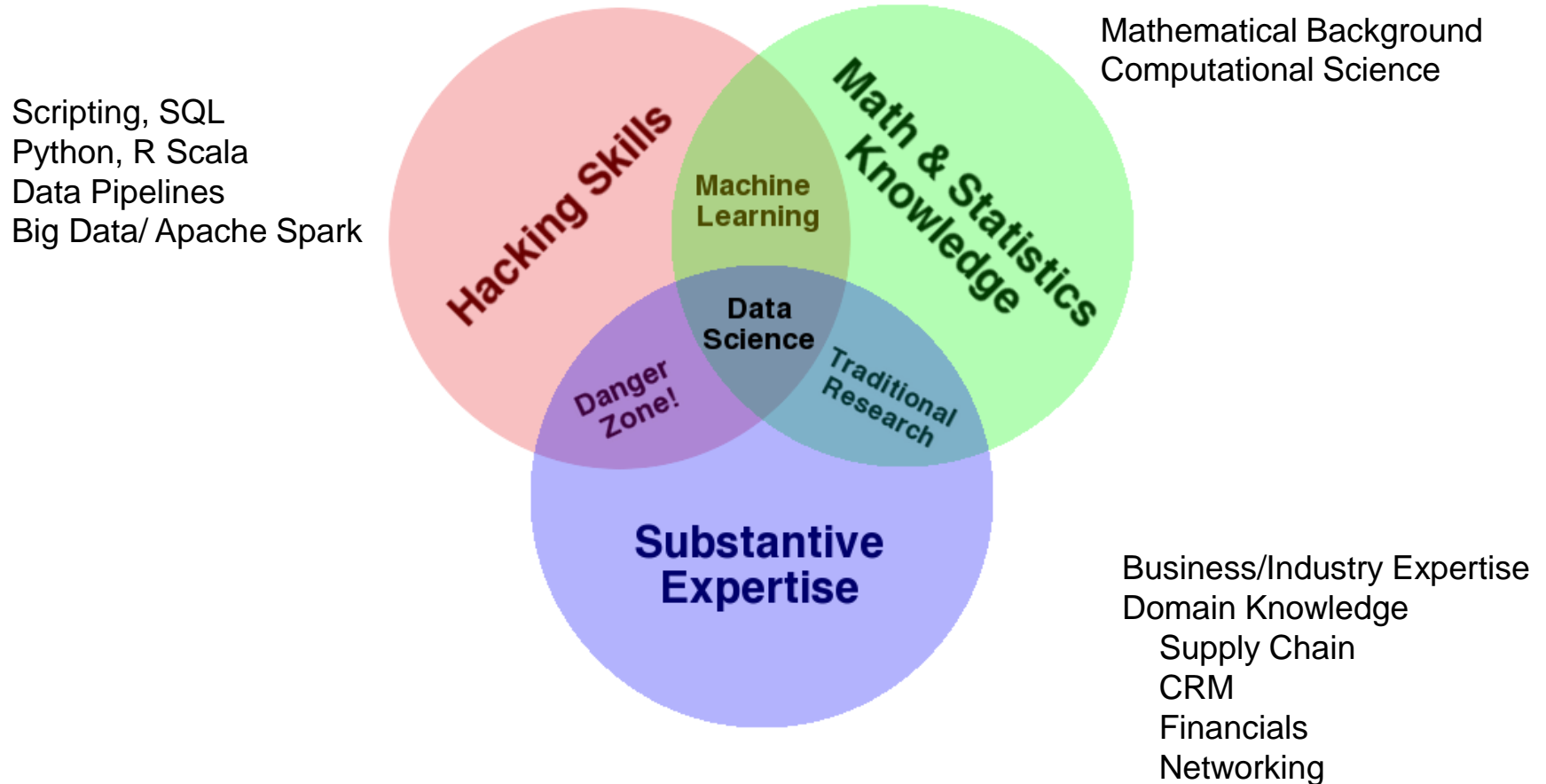
The Data Scientist Venn Diagram



Normally not all the skills are in one single person but rather in a data science team

In IBM Data Science Experience we include tools to make the perfect Data Science Team All in a collaborative, cloud environment that scales in demand

What is the Data Scientist?



Drew Conway's Data Science Venn Diagram

Data Scientist Issues

▪ Rigid toolset

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

▪ Fragmented and time consuming

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

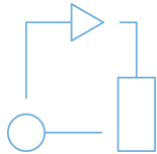
▪ Analytical Silo

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



IBM Watson Data Platform

Mission: Make Data Simple and Accessible to All



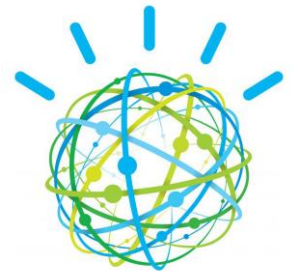
Platform.



Method.



Ecosystem.



<http://ibm.co/makedatasimple>

Data Science Experience

Brings together popular Data Science **Open Source** tools with IBM value-add functionalities coupled with **community and social** features



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>

Core Attributes of the Data Science Experience



IBM Data Science Experience

Community

- Find tutorials and datasets
- Read articles and papers
- Connect with Data Scientists
- Share comments
- Copy and share notebooks

Open Source

- Code in Scala/Python/R/SQL
- Jupyter Notebooks
- RStudio IDE and Shiny
- Apache Spark
- Your favorite libraries

IBM Added Value

- IBM Machine Learning
- SPSS Modeler Canvas
- Prescriptive Analytics - DOpnexcloud
- Projects and Version Control
- Managed Spark Service

Powered by IBM **Watson Data Platform**

* Closed beta

DSX Architecture

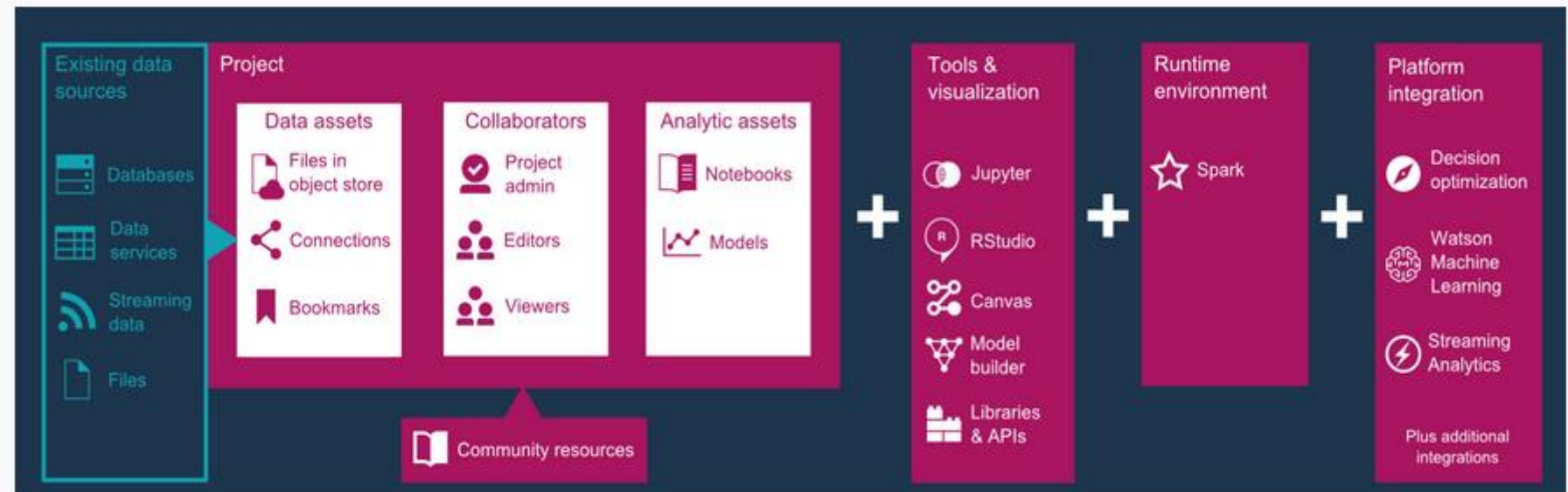
DSX architecture

Last updated: June 27, 2017

 Search this document



DSX provides you with the environment and tools to solve your business problems by collaboratively analyzing data. This illustration shows how the architecture of DSX is centered around the project. A project is how you organize your resources for solving a business problem.



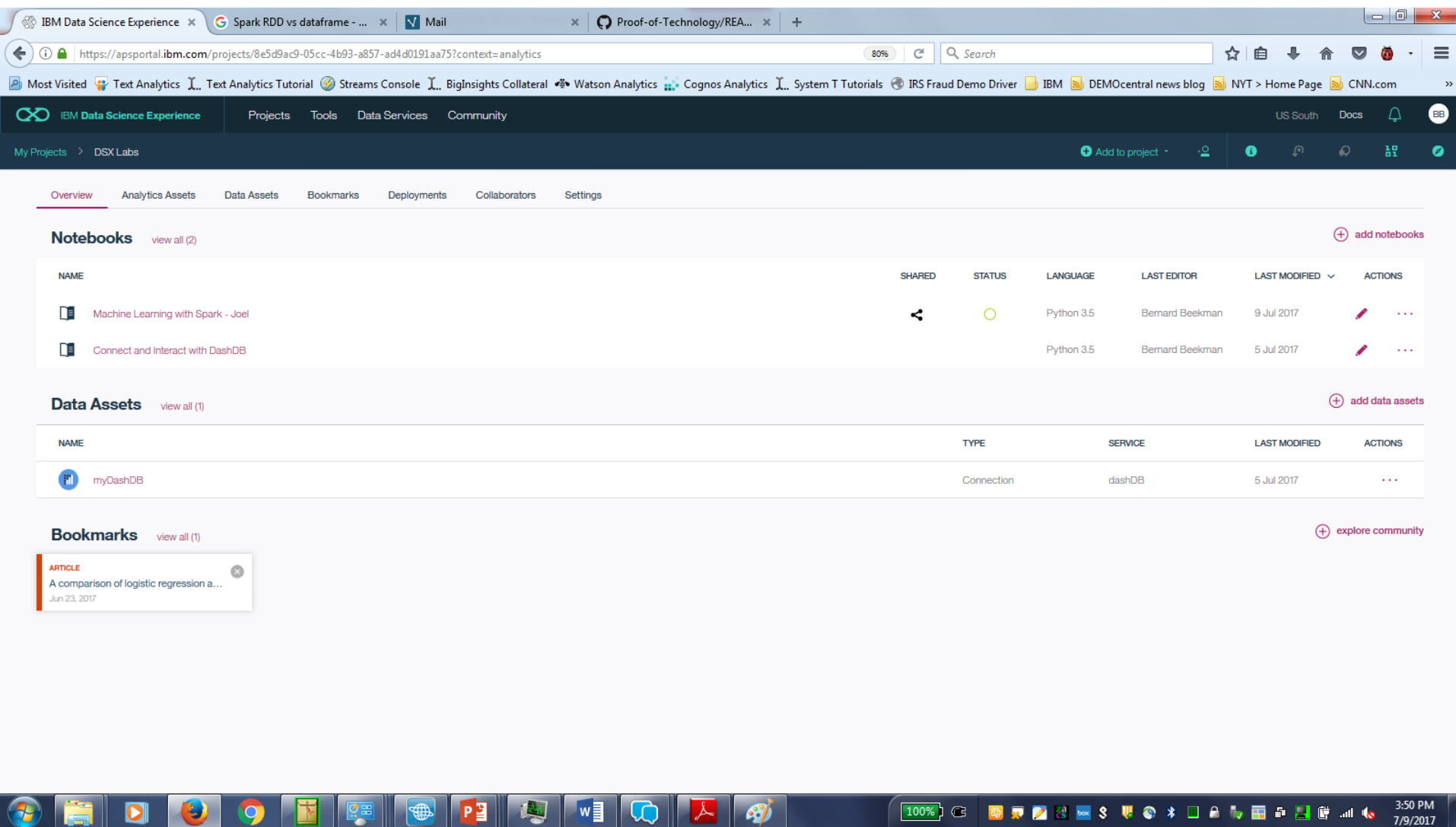
Community Cards provide in-context learning for users

The screenshot displays the IBM Data Science Experience interface, specifically the Community Cards section. The top navigation bar includes the IBM Data Science Experience logo, links to Projects, Tools, Data Services, and Community, and user information (US South, Docs, a bell icon, and a profile icon labeled BB).

The main content area is divided into four sections, each with a "View All >" link:









- Articles:** Displays four article cards. Each card includes a title, author, date, topic, and format. The first article is "Web Picks by DataMiningApps" by DataMiningApps, dated Jul 06, 2017, on the topic of Data Science, in Web page format, with 1 heart. The second is "Using Deep Learning to Reconstruct..." by Jeffrey Hetherly, dated Jun 28, 2017, on the topic of Deep Learning, in Web page format, with 1 heart. The third is "A comparison of logistic regression and..." by Andrew Y. Ng & Michael L. Jordan, dated Jun 23, 2017, on the topic of Data Science, in PDF format, with 3 hearts. The fourth is "A Dynamic Duo – Inside Machine learning ..." by John Thomas, dated Jun 14, 2017, on the topic of Machine Learning, in Video format, with 3 hearts.
- Data Sets:** Displays four data set cards. Each card includes a title, author, date, topic, and format. The first is "Breast Cancer Wisconsin (Diagnostic) Data Set" by IBM, dated Jun 26, 2017, on the topic of Health, with 0 hearts. The second is "IBM Watson Facebook posts for 2015" by IBM, dated Jun 28, 2017, on the topic of Economy & Business, with 3 hearts. The third is "Uncertain demand per store" by IBM, dated Jun 06, 2017, on the topic of Economy & Business, with 4 hearts. The fourth is "Demand per store" by IBM, dated Jun 06, 2017, on the topic of Economy & Business, with 2 hearts.
- Notebooks:** Displays four notebook cards. Each card includes a title, author, date, topic, and format. The first is "Access MySQL with R" by IBM, dated Jul 06, 2017, on the topic of Transportation, with 5 hearts. The second is "Classify tumors with machine learning" by IBM, dated Jun 28, 2017, on the topic of Health, with 3 hearts. The third is "Analyze Facebook Data Using IBM Watson and..." by IBM, dated Jun 28, 2017, on the topic of Economy & Business, with 4 hearts. The fourth is "Predicting churn with the SPSS random tree..." by IBM, dated Jun 28, 2017, on the topic of Communications, with 2 hearts.
- Tutorials:** This section is partially visible at the bottom of the screenshot, with a "View All >" link.

Collaborate Using Projects





The screenshot shows the IBM Data Science Experience (DSX) interface. The browser address bar displays the URL: <https://apportal.ibm.com/projects/8e5d9ac9-05cc-4b93-a857-ad4d0191aa75?context=analytics>. The page title is "IBM Data Science Experience". The navigation bar includes "Projects", "Tools", "Data Services", and "Community". The main content area is titled "My Projects > DSX Labs" and features a "Add to project" button. Below the navigation bar, there are tabs for "Overview", "Analytics Assets", "Data Assets", "Bookmarks", "Deployments", "Collaborators", and "Settings". The "Overview" tab is active, showing a list of "Notebooks" and "Data Assets".

Notebooks [view all \(2\)](#) [+ add notebooks](#)

NAME	SHARED	STATUS	LANGUAGE	LAST EDITOR	LAST MODIFIED	ACTIONS
 Machine Learning with Spark - Joel			Python 3.5	Bernard Beekman	9 Jul 2017	 
 Connect and Interact with DashDB			Python 3.5	Bernard Beekman	5 Jul 2017	 

Data Assets [view all \(1\)](#) [+ add data assets](#)

NAME	TYPE	SERVICE	LAST MODIFIED	ACTIONS
 myDashDB	Connection	dashDB	5 Jul 2017	

Bookmarks [view all \(1\)](#) [+ explore community](#)

ARTICLE

A comparison of logistic regression a...

Jun 23, 2017

The bottom of the screenshot shows the Windows taskbar with various application icons and the system clock displaying 3:50 PM on 7/9/2017.

Add Collaborators to a Project

Add New Collaborator

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

Type name or email address

Select



Viewer



Editor


Admin

Cancel

Add

GitHub Integration



Data Science Experience 

Settings

Integrations

Profile

Services

Integrations

GitHub Integration

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

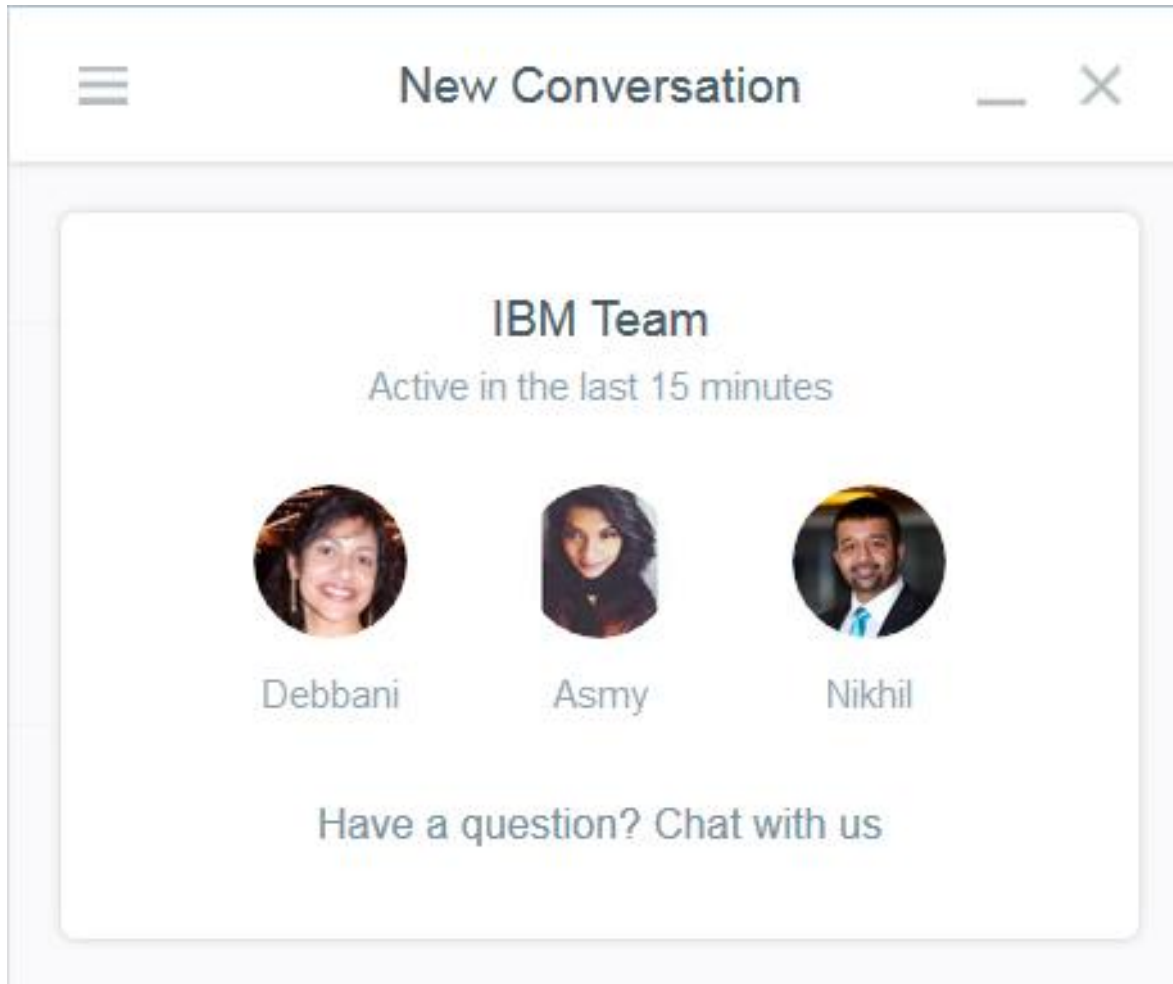
40

Clear

Save

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

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



Docs, Forums, Blogs and Ideas

- Online documentation for DSX, DSX Local and DSX Desktop
- DSX discussion forum on Stack Overflow
- Blog posts from IBM Developers
- Give feedback on DSX to IBM for new features

Helpful links

Docs



Find the information you need. Watch videos of key tasks.

Discussion forum



Stack Overflow is a community of 6.9 million programmers just like you, helping each other. Join the conversation on Data Science Experience.

Blog



Read and follow our blog to keep up with the latest updates about Data Science Experience.

Got ideas?



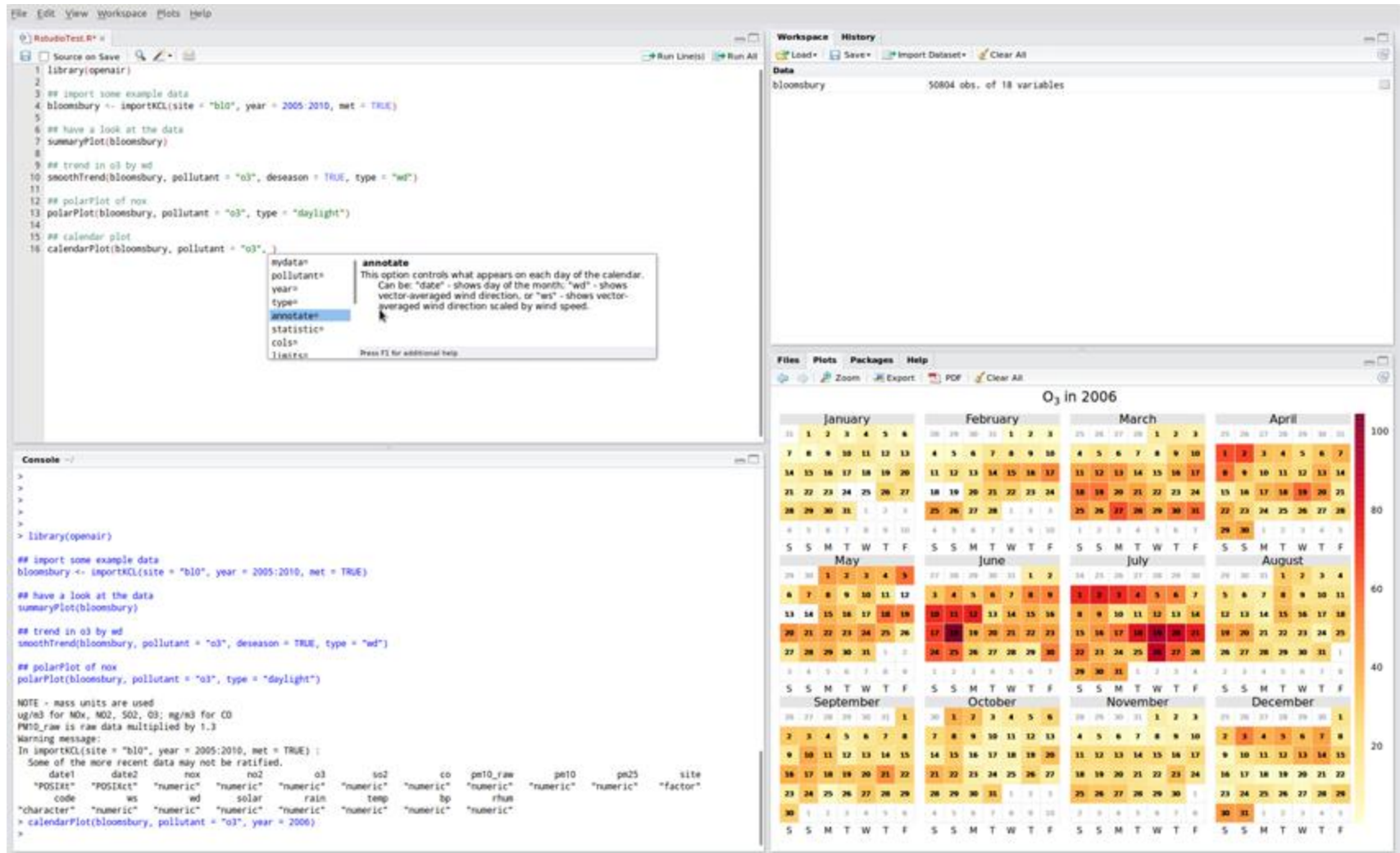
Have feedback on Data Science Experience? Submit your ideas in our product forum or vote on ideas submitted by others.

Supported Data Sources/Targets for DSX via on- premises and cloud **Connectors**

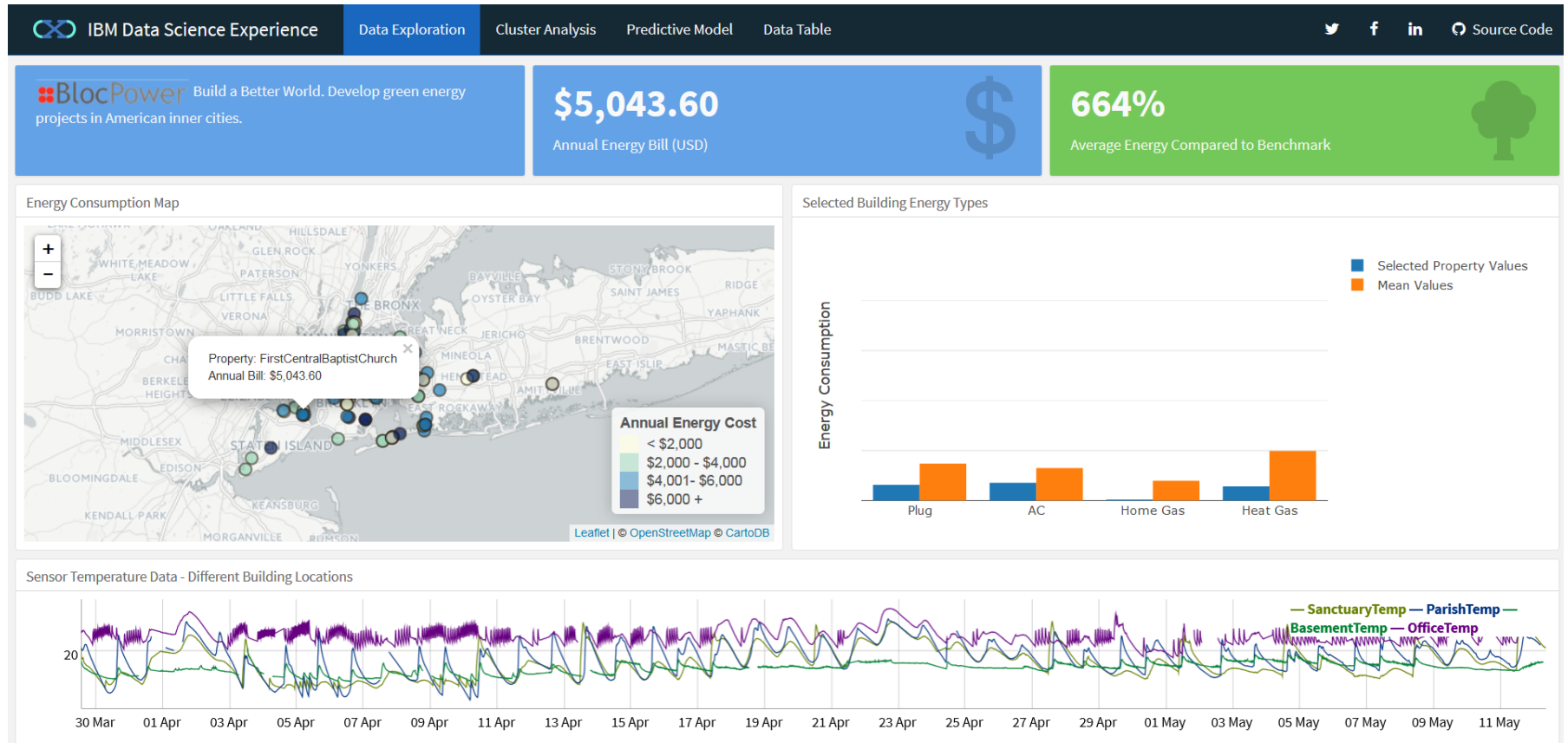


Cloud Sources	On-Premises Sources	Cloud Targets	On-Premises Targets
Amazon Redshift	Apache Hive	Amazon S3	IBM DB2® LUW
Amazon S3	Cloudera Impala	Bluemix Object Storage	IBM Pure Data for Analytics®
Apache Hive	IBM DB2® LUW	IBM Cloudant™	Teradata
Bluemix Object Storage	IBM Informix®	IBM dashDB	
IBM BigInsights™ 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

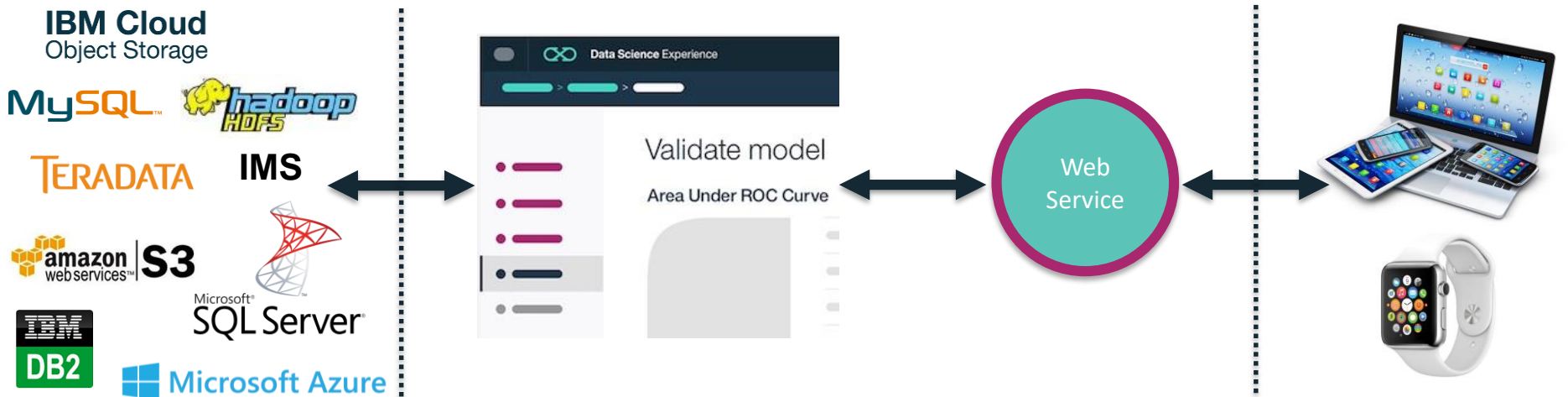


With RStudio you can create Shiny web applications to make your analysis accessible to the business



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-Modeling 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

DSX Local

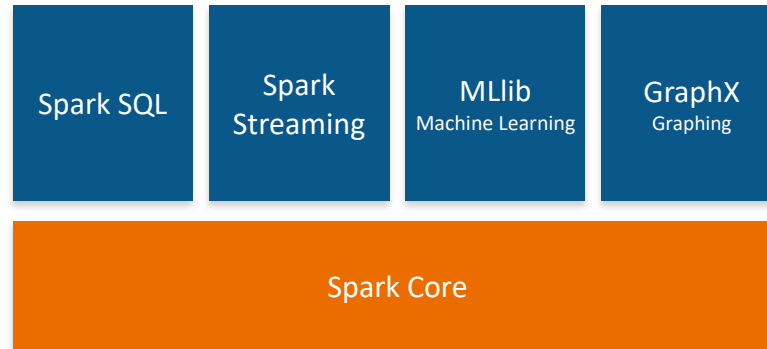
- **Very similar to the public cloud version of DSX**
- **Runs on hardware that is provided by the customer**
 - The DSX Local software and hardware are managed by the customer
- **DSX Local comes with all the software it needs to run, although it can integrate with existing customer systems such as**
 - Databases and HDFS storage
 - LDAP servers for authentication

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



Benefits of Spark for Data Science

- General compute engine
- Basic I/O functions
- Task dispatching
- Scheduling



- **Allows Data Scientists to code at scale**
 - In-Memory processing that scales in a distributed architecture
- **Supports multiple programming interfaces (Scala, Python, Java and R)**
- **Provides unified APIs (SQL, Streaming, Machine Learning, etc.)**

IBM is all-in on Spark

Contribute to the Core

Launch Spark Technology Cluster (STC), 300 engineers

Open source SystemML

Partner with Databricks

Foster Community

Educate 1M+ data scientists and engineers via online courses

Sponsor AMPLab, creators and evangelists of Spark

Infuse the Portfolio

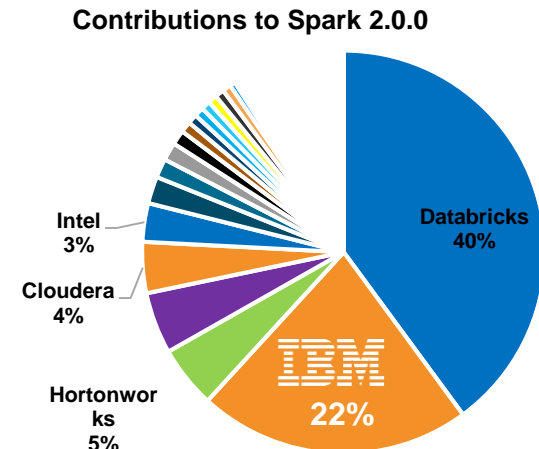
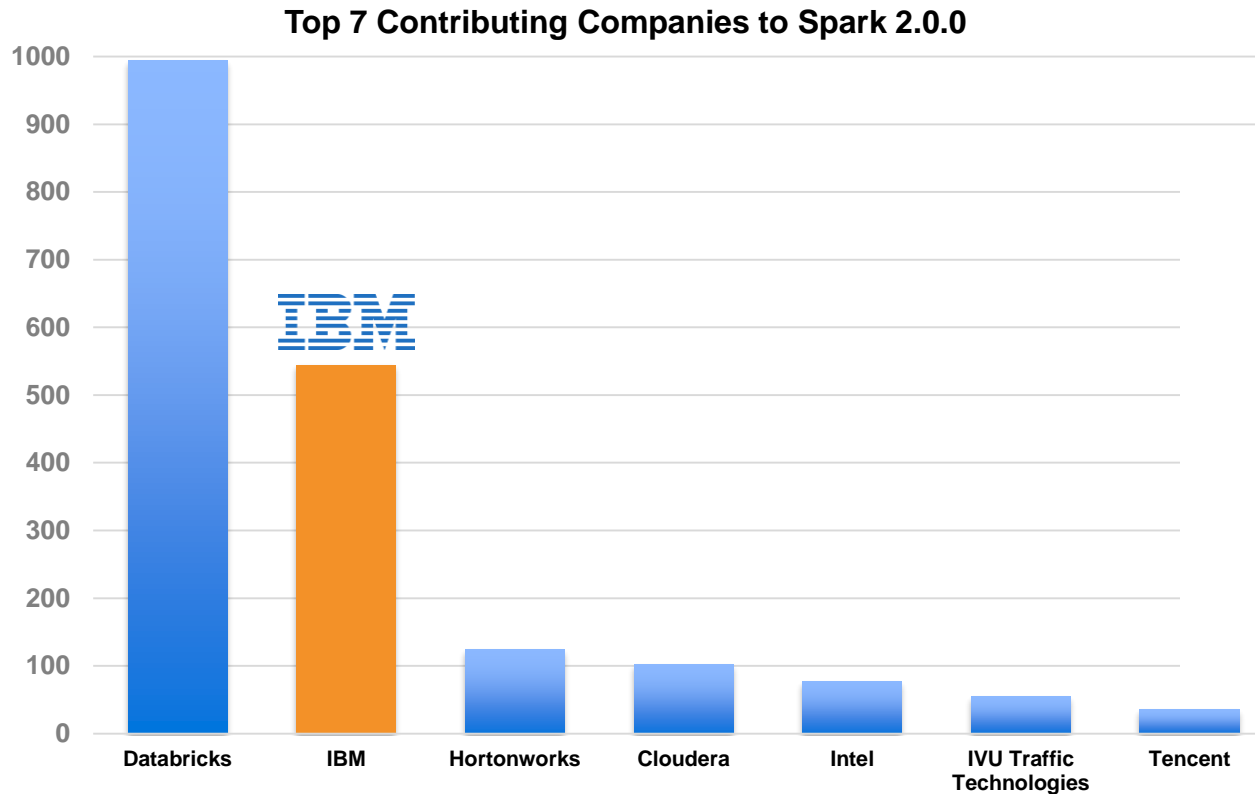
Integrate Spark throughout portfolio

3,500 employees working on Spark-related topics

Spark however customers want it – standalone, platform or products

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



Lab Exercise – Female Human Trafficking


▪ Input


- Generated fake travel records based on incoming custom forms.
- Subset of records were vetted as “high”, “medium”, or “low” risk for Female Human Trafficking by an analyst.

- **Goal is to train a model on the vetted data to be able to score the unvetted travel records into high, medium, or low categories.**

Demo Data

Field	Description
UUID	Hash-based unique identifier
VETTING_LEVEL	Analyst vetting status : 100- PENDING, 10 – HIGH, 20 – MED, 10 - LOW
NAME	Person name
GENDER	Person Gender
AGE	Person age at time of travel
BIRTH_DATE	Person birth date
BIRTH_COUNTRY	Person full birth country
BIRTH_COUNTRY_CODE	Person ISO 2 country
OCCUPATION	Person occupation as declared on form
ADDRESS	Person US address
SSN	Person Social Security Number
PASSPORT_NUMBER	Person Passport Number
PASSPORT_COUNTRY	Person Passport Issuing Country
PASSPORT_COUNTRY_CODE	Person Passport Issuing Country ISO 2 Code
COUNTRYIES_VISITED	The countries visited as declared on form
COUNTRIES_VISITED_COUNT	The number of countries visited as declared on form
ARRIVAL_AIRPORT_COUNTRY_CODE	ARRIVAL Airport country code ISO2
AIRPORT_ARRIVAL_IATA	ARRIVAL Airport 3 character code
AIRPORT_ARRIVAL_MUNICIPALITY	ARRIVAL Airport Municipality Derived from Code
ARRIVAL_AIRPORT_REGION	ARRIVAL Airport Region Derived from Code
DEPARTURE_AIRPORT_COUNTRY_CODE	DEPARTURE Airport Country code ISO2
DEPARTURE_AIRPORT_IATA	DEPARTURE Airport 3 character code
DEPARTURE_AIRPORT_MUNICIPALITY	DEPARTURE Airport Municipality Derived from Code.

 Target

 Features

Demo Flow

- **Read in dataset as a DataFrame from DB2 Warehouse**
 - Connect to DB2 Warehouse
 - Read in the data
- **Identify Labels**
 - Label the data (“VETTING_LEVEL”)
 - Select features
- **Feature Engineering (Transformation)**
 - StringIndexer (occupation, country, gender, birth year variables)
 - VectorAssembler
 - Normalizer
- **Define Model and Setup Pipeline**
 - Naïve Bayes
- **Train the Model**
 - Split input data into Training (70%) and Test (30%) DataFrames
 - Cache the resulting DataFrames
 - Fit the Pipeline to the Training data set



Demo Flow (continued)

- **Evaluate the resulting predictions**
 - Area under the ROC curve

- **Tune the model (hyperparameters)**
 - Build Parameter Grid
 - Cross-evaluate to find the best model

- **Score the unvetted records**
 - Use Best Model to Score unvetted records (VETTING LEVEL == 100)
 - Write results into DB2 Warehouse table



Classification - Naïve Bayes

- **Two or more outcomes.**
- **Assumes independence among explanatory variables, which is rarely true (thus “naïve”).**
- **Despite its simplicity, often performs very well... widely used.**
- **Significant use cases:**
 - Text categorization (spam vs. legitimate, sports or politics, etc.) using word frequencies as the features
 - Medical diagnosis (e.g., automatic screening)

Get Started with Data Science Experience Today!

Calling all Data Scientists!

- Our mission is to win the **hearts and minds** of Data Scientists
- IBM Data Science Experience is a **freemium model** with value-add features, pricing and up-sell in development
- **Sign up** and encourage your colleagues to do so at **datascience.ibm.com**

Optional Labs

- Watson Machine Learning
 - Lab-4
- Decision Optimization
 - Linear Programming / Beyond Linear Programming
 - Docs > Analyze Data > Decision Optimization in DSX > Decision Optimization Tutorial
 - Community
 - Sudoku
 - Finding optimal locations of new store using Decision Optimization
- Visualization (PixieDust) [latest version 1.1]
 - Community
 - Welcome to PixieDust
 - Twitter Sentiment with Watson and PixieDust (Python 2/Spark 1.6)
 - Analyze traffic data using PixieDust & Spark (*must* use PixieDust 1.1)

Optional Labs

- SPSS
 - Community
 - Load SPSS predictive models to Bluemix & score data
 - Model bike sharing data with SPSS
 - Predicting churn with the SPSS random tree algorithm
- Amazon EMR
 - Community
 - Analyze accident reports on Amazon EMR Spark
- Connection Examples
 - Shares
 - Hana: https://apsportal.ibm.com/analytics/notebooks/ab1612fc-0c6e-4f23-83bd-d6ccd3c363ed/view?access_token=0e9c6e16abefc0f52199c62d826b1acbdab58373c6cfa781e28bf0a0db082bdd
 - Cloudera: https://apsportal.ibm.com/analytics/notebooks/e8a3e5bc-ed3-462b-b4bb-ced4a279fa3d/view?access_token=b97c6c8faf9467adc796a444fe2608b797a251b243c17f3cd1faf1937923521b



IBM Data Science Experience
<https://www.youtube.com/watch?v=1HjzkLRdP5k&t=29s>