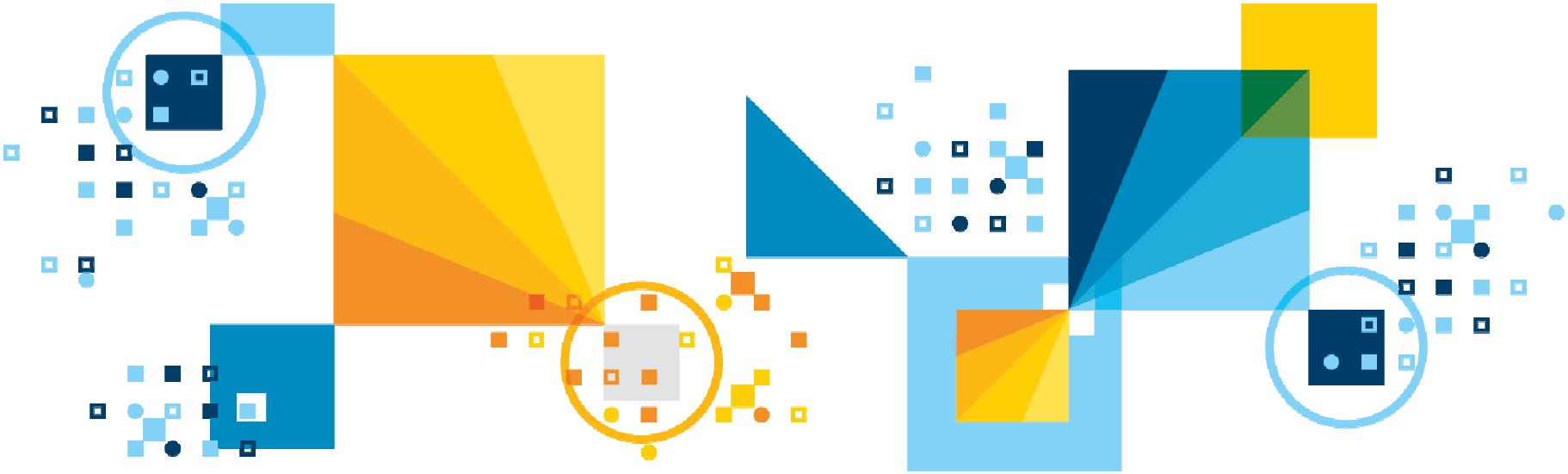
IBM Data Science Experience Overview

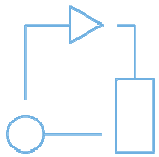


Agenda

- IBM Watson Data Platform
- Data Science Experience (DSX)
- Demo

Introducing IBM Watson Data Platform

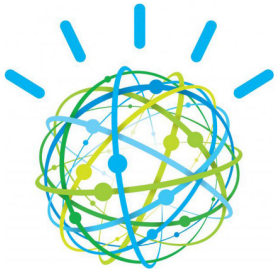
Mission: Make Data Simple and Accessible to All



Platform.

Method.

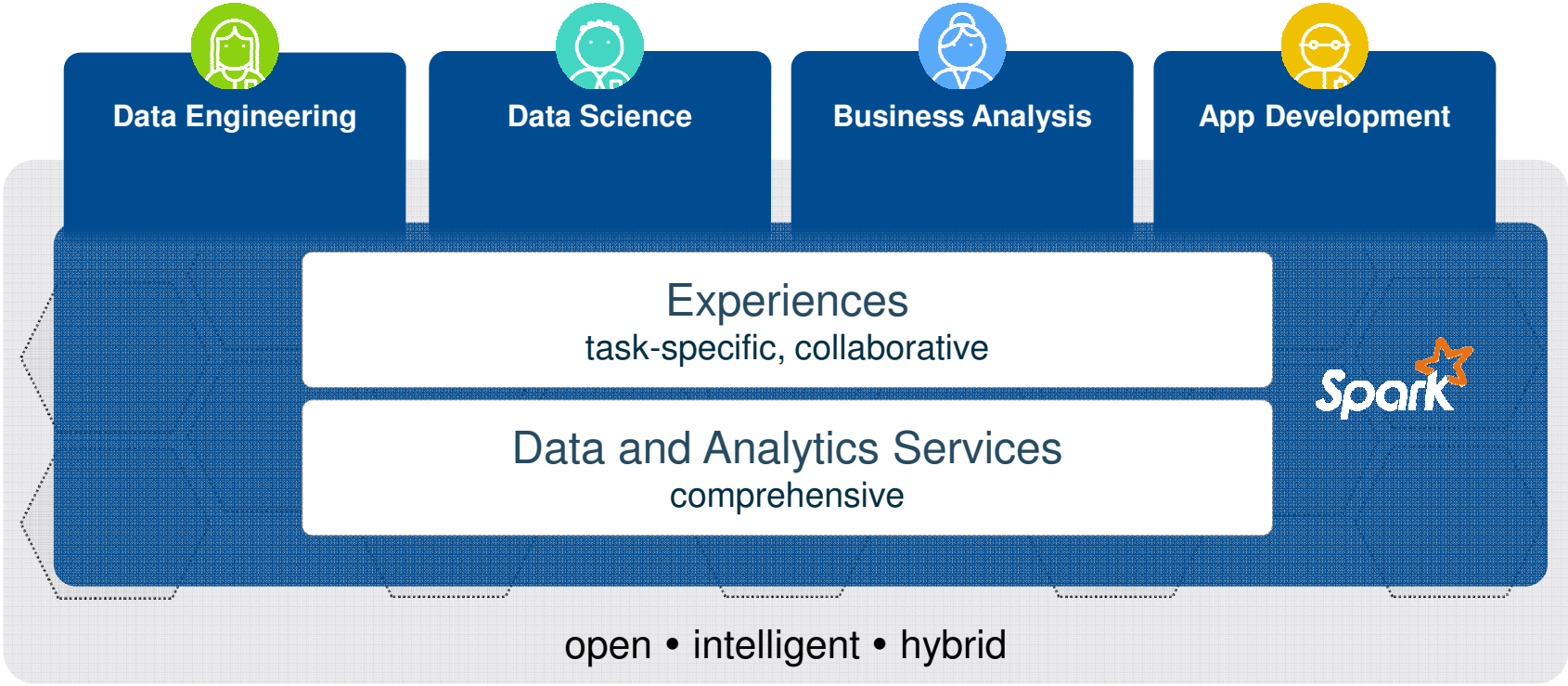
Ecosystem.



http://ibm.co/makedatasimple

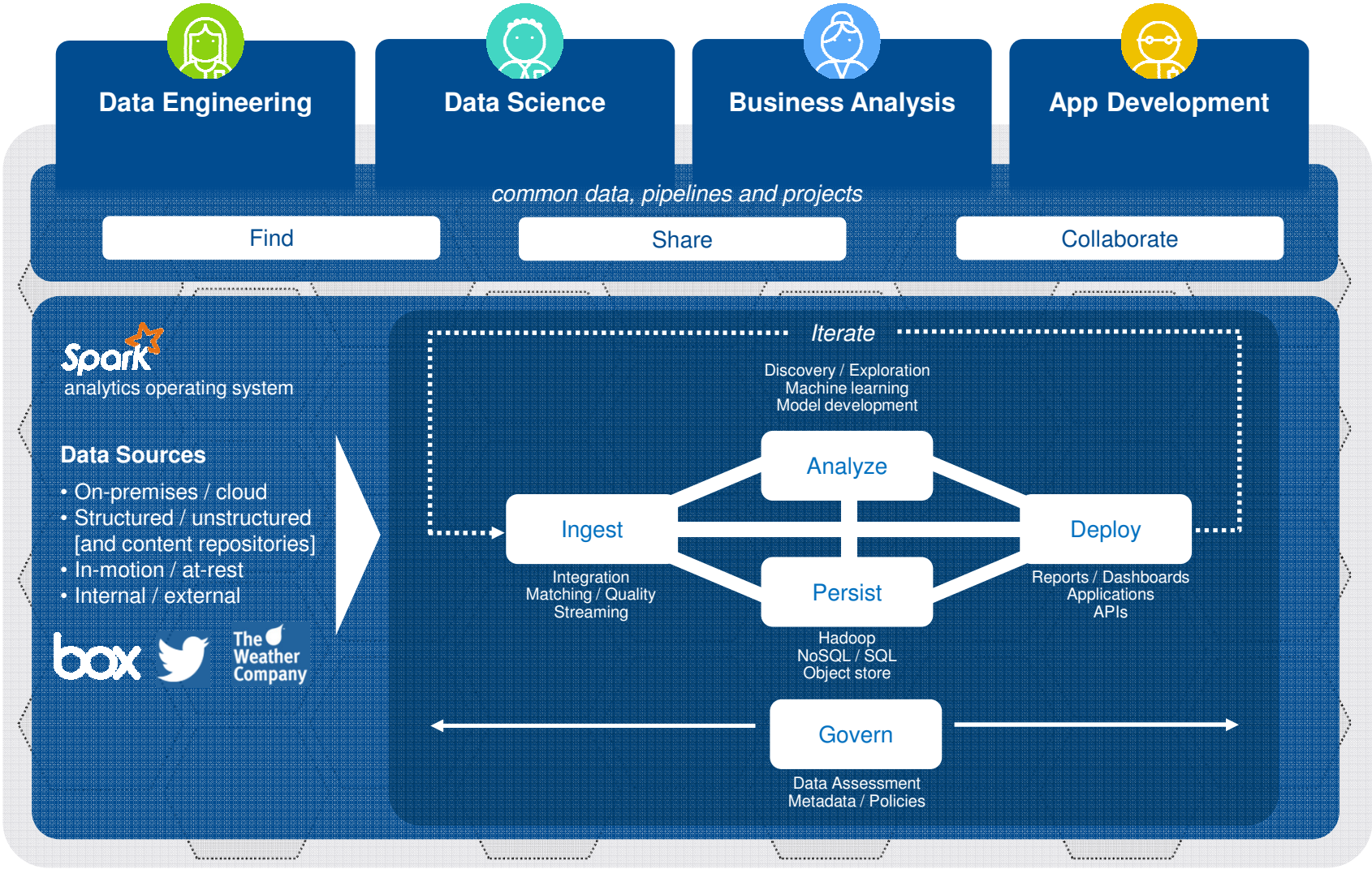
IBM Watson Data Platform

Experience New Ways To Put Data To Work



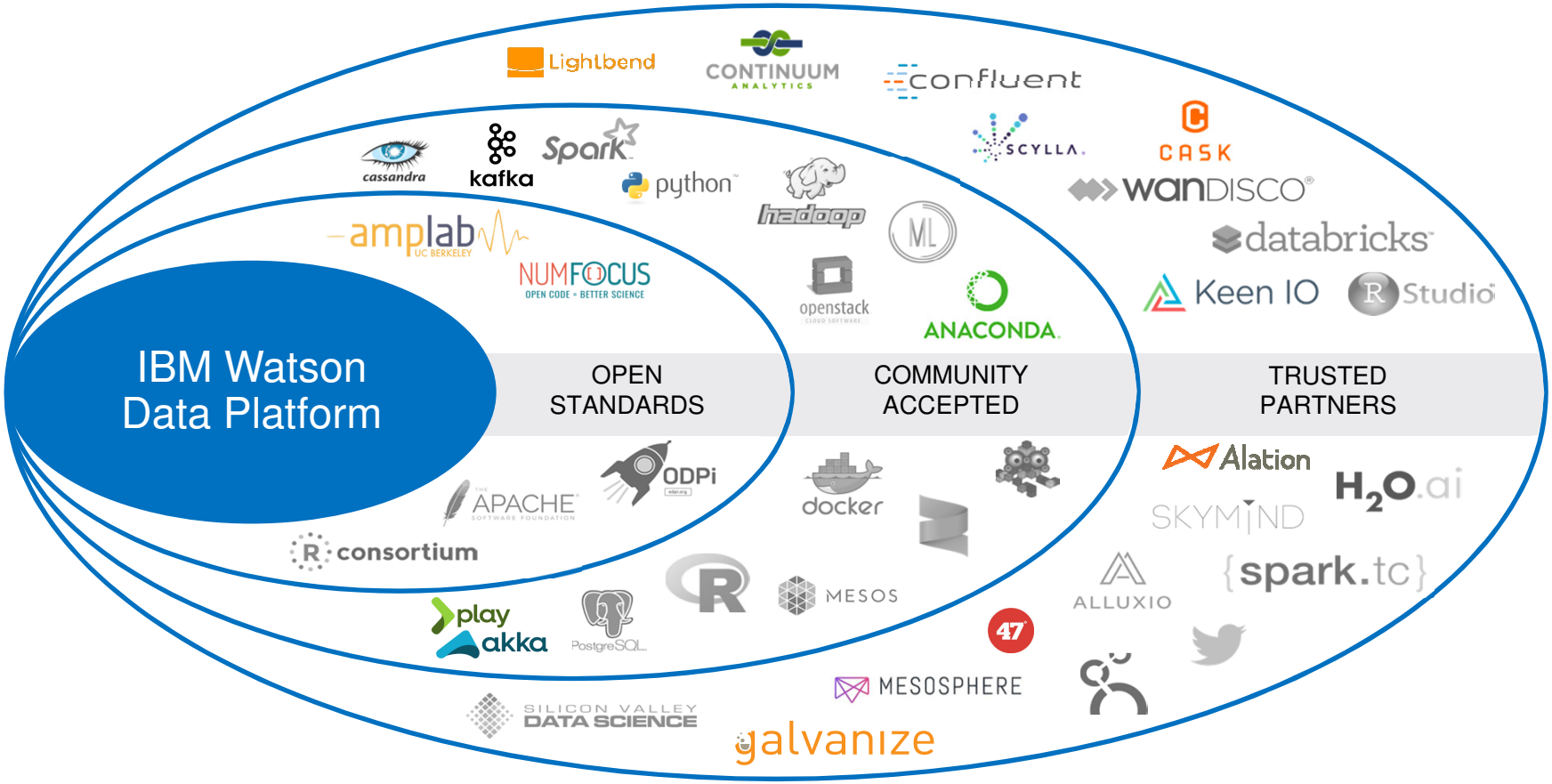
IBM Watson Data Platform

Connects Users to Data and Analytics



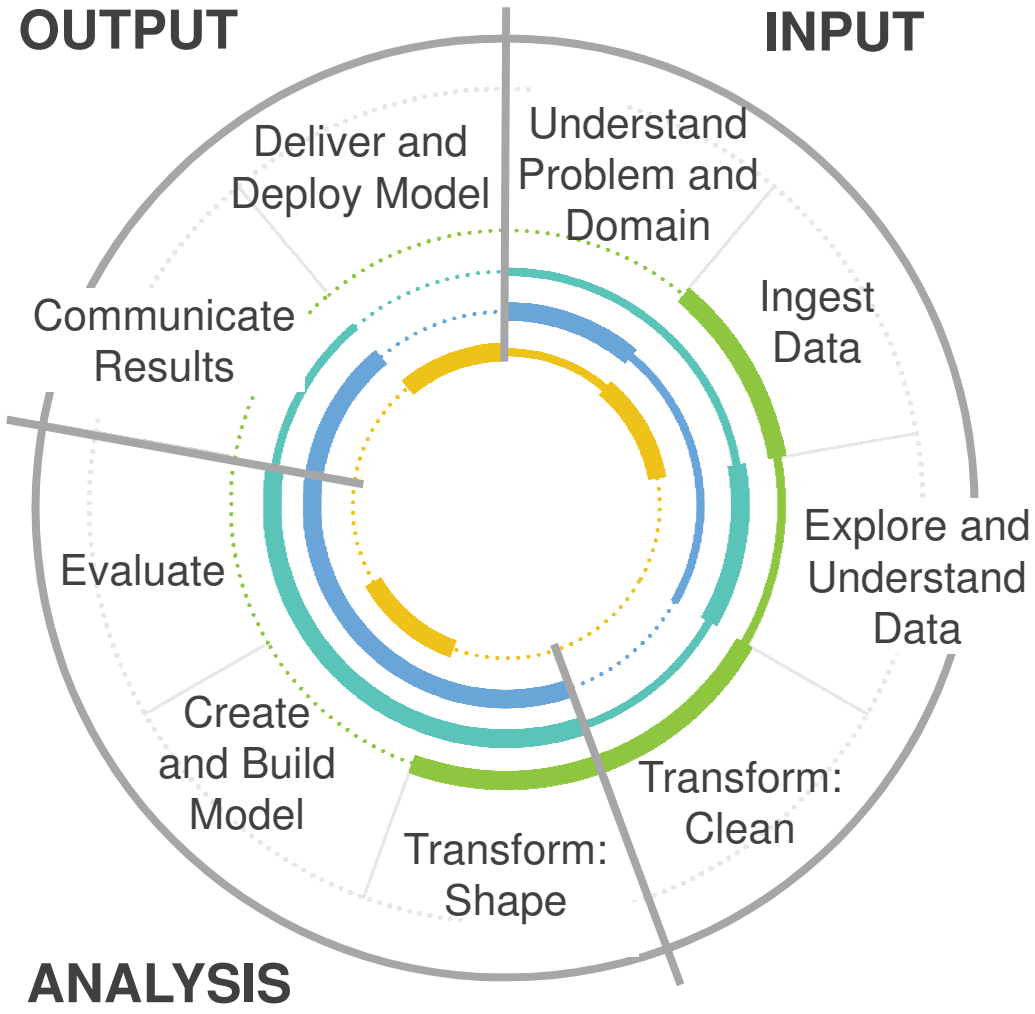
IBM Watson Data Platform Partner Ecosystem

The Open Community To Innovate Faster With Data



https://www.ibm.com/analytics/us/en/watson-data-platform/partners.html

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

DATA SCIENCE EXPERIENCE (DSX)

Data Scientist Persona - Current Challenges and Pain Points

- Keeping up with the many open source innovations
- Determining which programming language to use – standardize on one or allow the data scientists to choose.
- Difficult to collaborate across different tool sets and environments
- Lack of metadata about the data or knowledge of data lineage
- Difficult to share data products organizationally or deploy in operation.

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.

datascience.ibm.com

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

IBM Data Science Experience

Click here to watch video

Core Attributes of the Data Science Experience

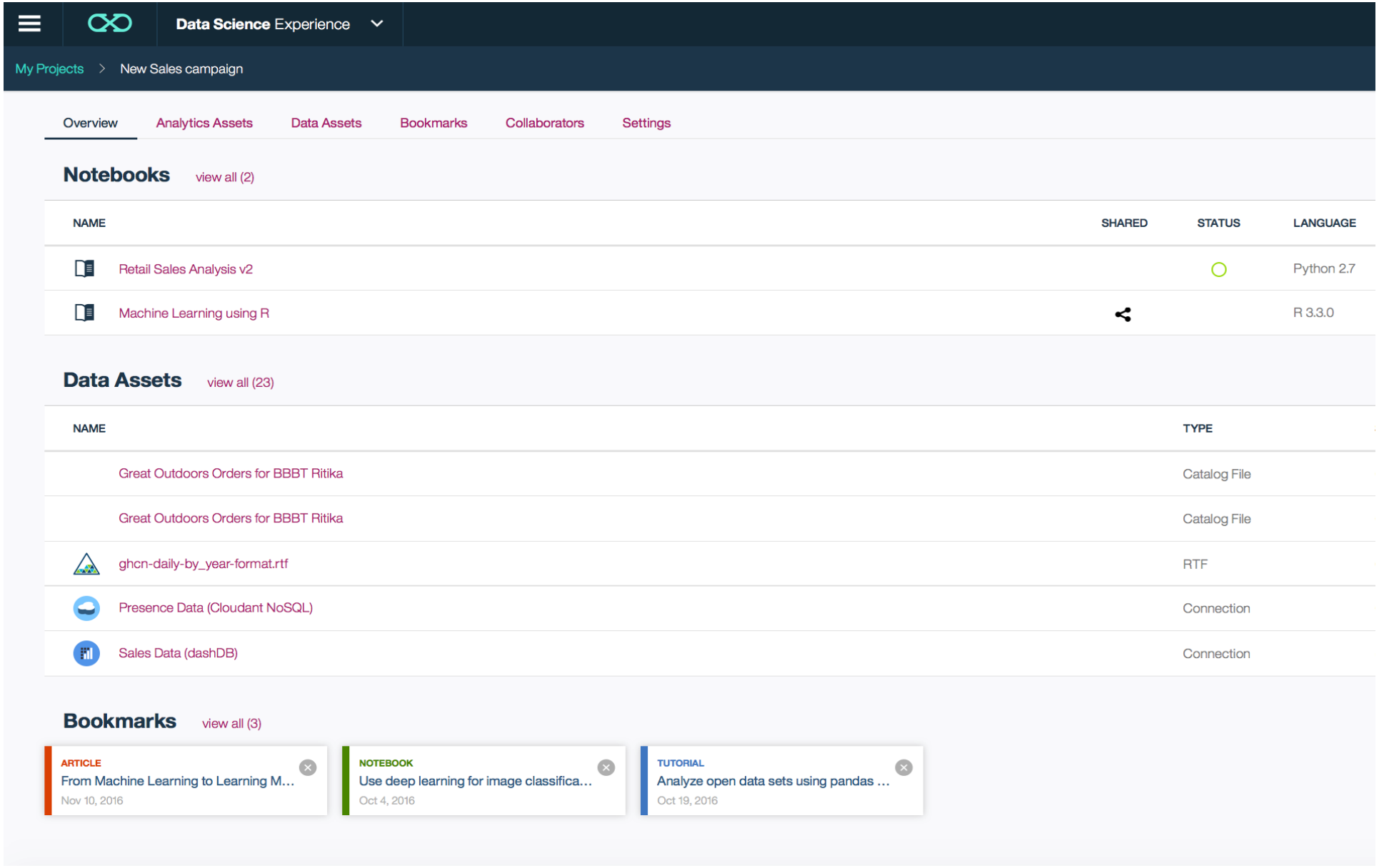
IBM Data Science Experience

Community	Open Source	IBM Added Value
• Find tutorials and datasets	• Code in Scala/Python/R/SQL	• IBM Machine Learning*
• Connect with Data Scientists	• Jupyter Notebooks	• SPSS Modeler Canvas*
• Share comments	• RStudio IDE and Shiny	• Prescriptive Analytics - DOcplexcloud
• Read articles and papers	• Apache Spark	• Projects and Version Control
• Fork and share projects	• Your favorite libraries	• Managed Spark Service

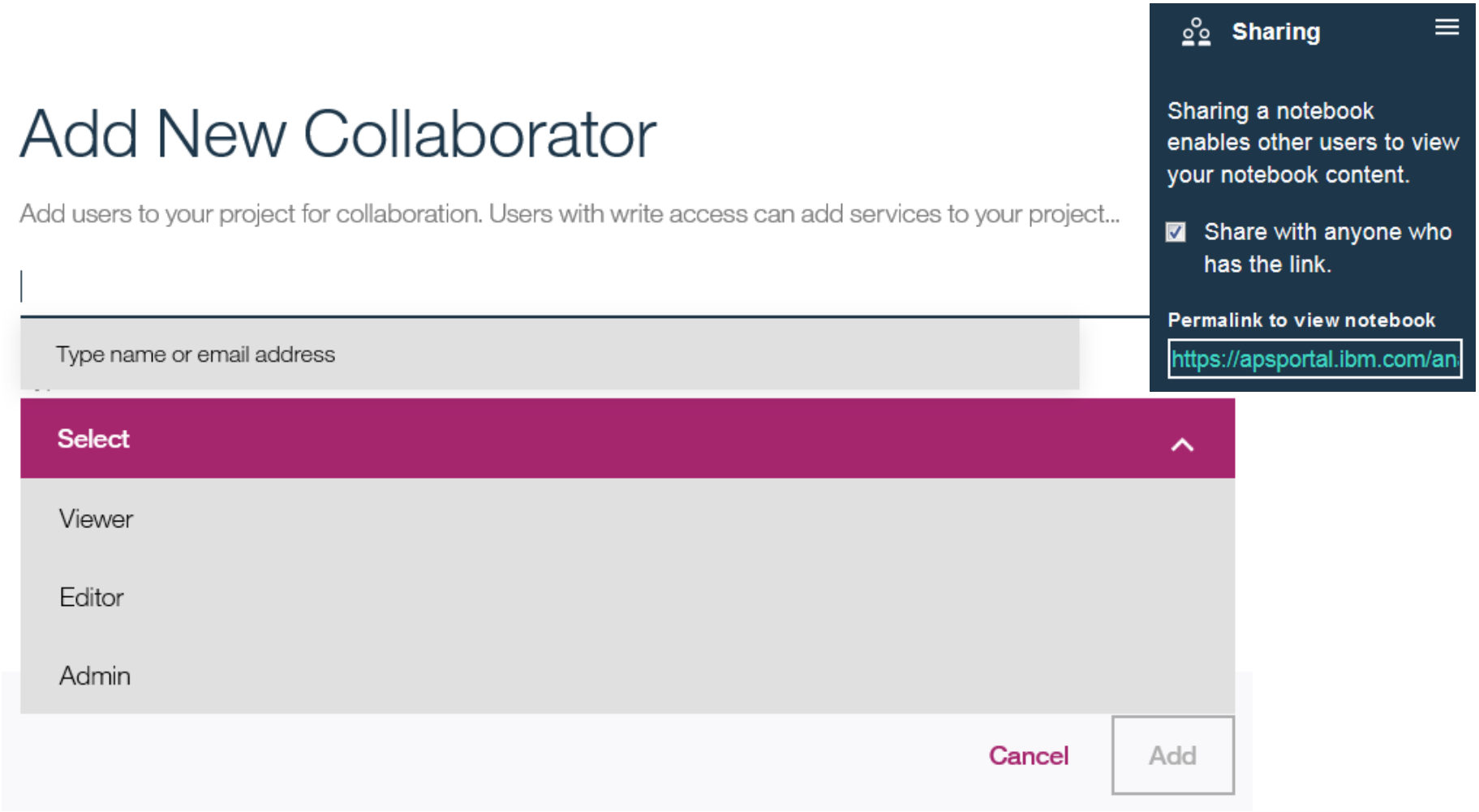
Powered by IBM **Watson Data Platform**

* Closed beta

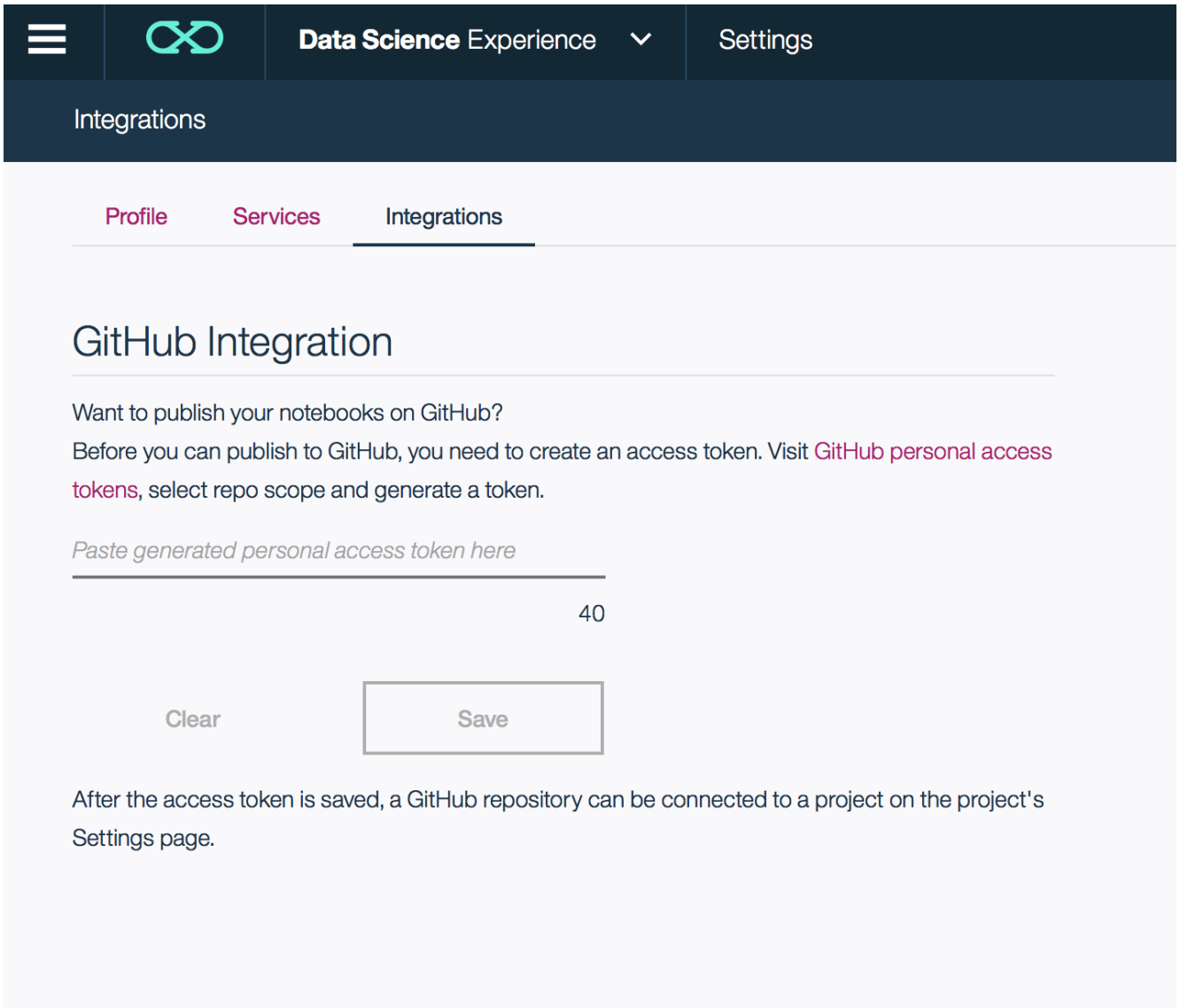
Collaborate Using Projects



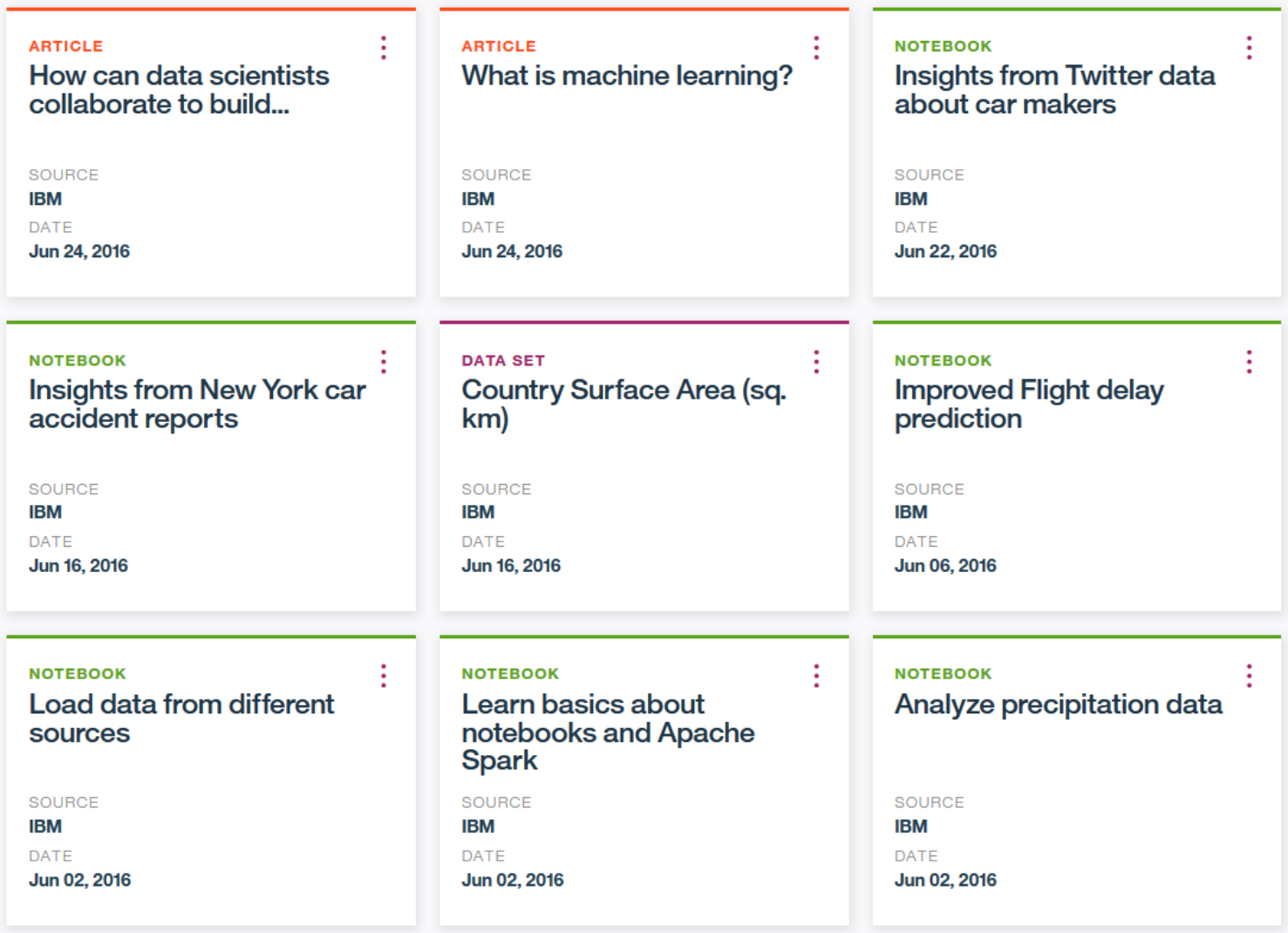
Features for sharing, forking, and reusing Project assets to increase your data science team's productivity



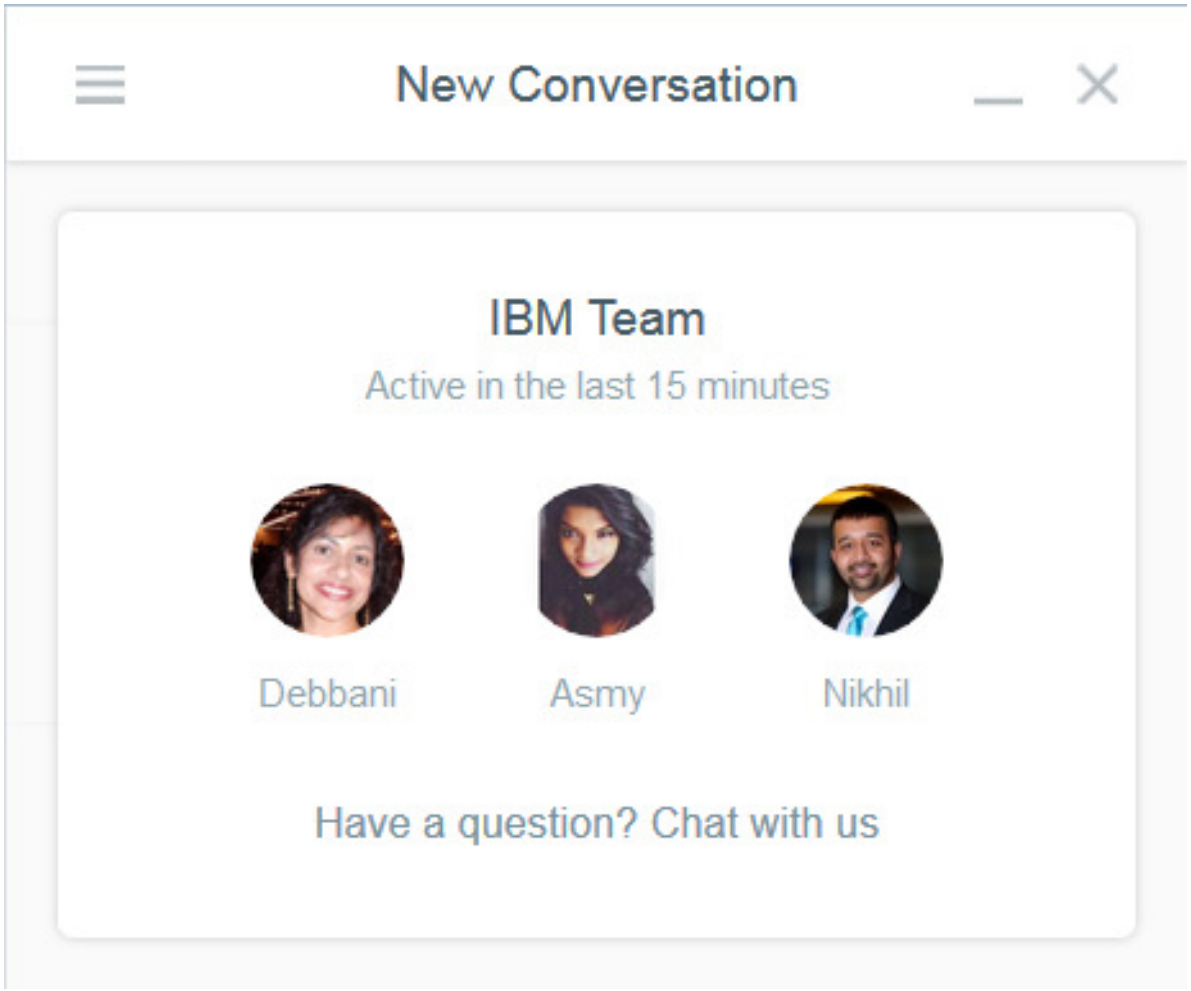
GitHub Integration



Community Cards provide in-context learning for users



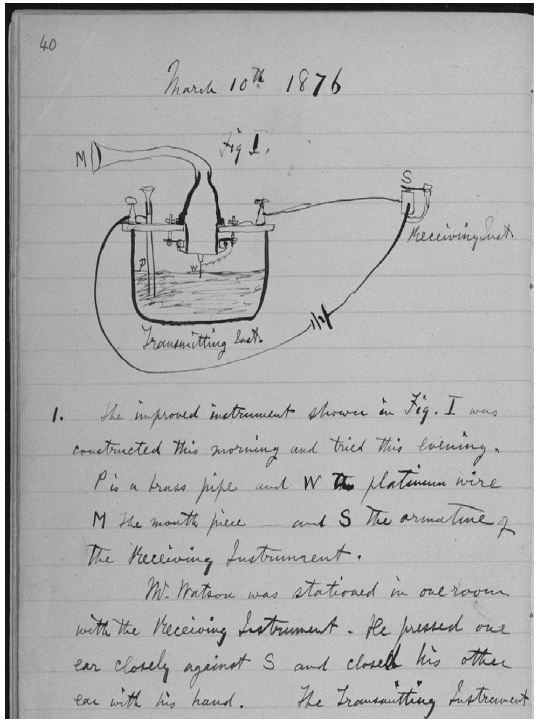
Live chat on Intercom for support from the IBM team and to provide your feedback on how we can improve DSX



What is a “Notebook”?

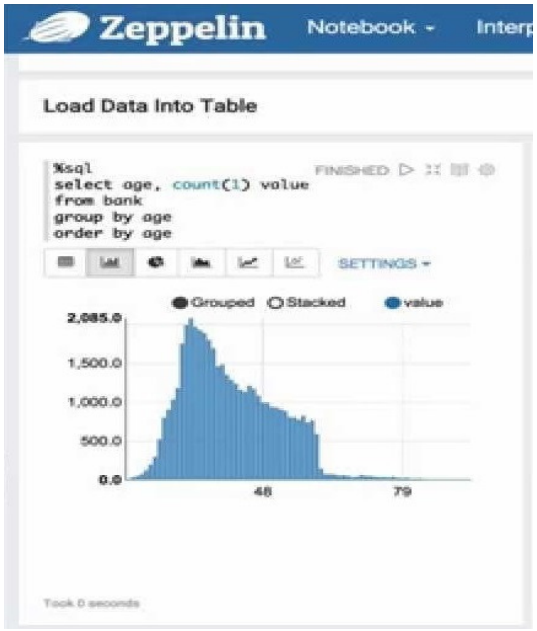
Pen and Paper

- **Pen and paper has long provided the rich experience that scientists need to document progress through notes and drawings:**
  - Expressive
  - Cumulative
  - Collaborative

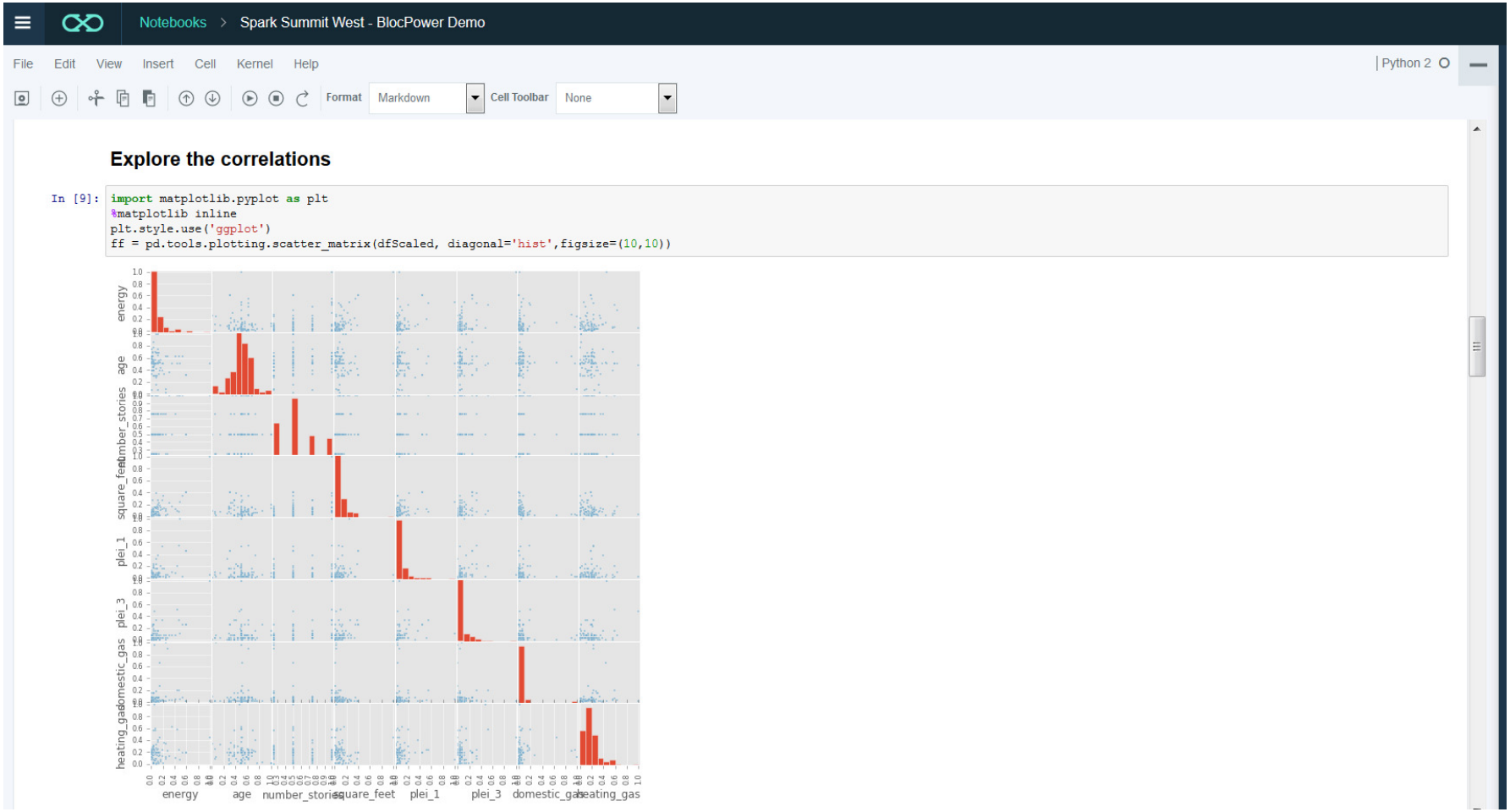


Notebooks

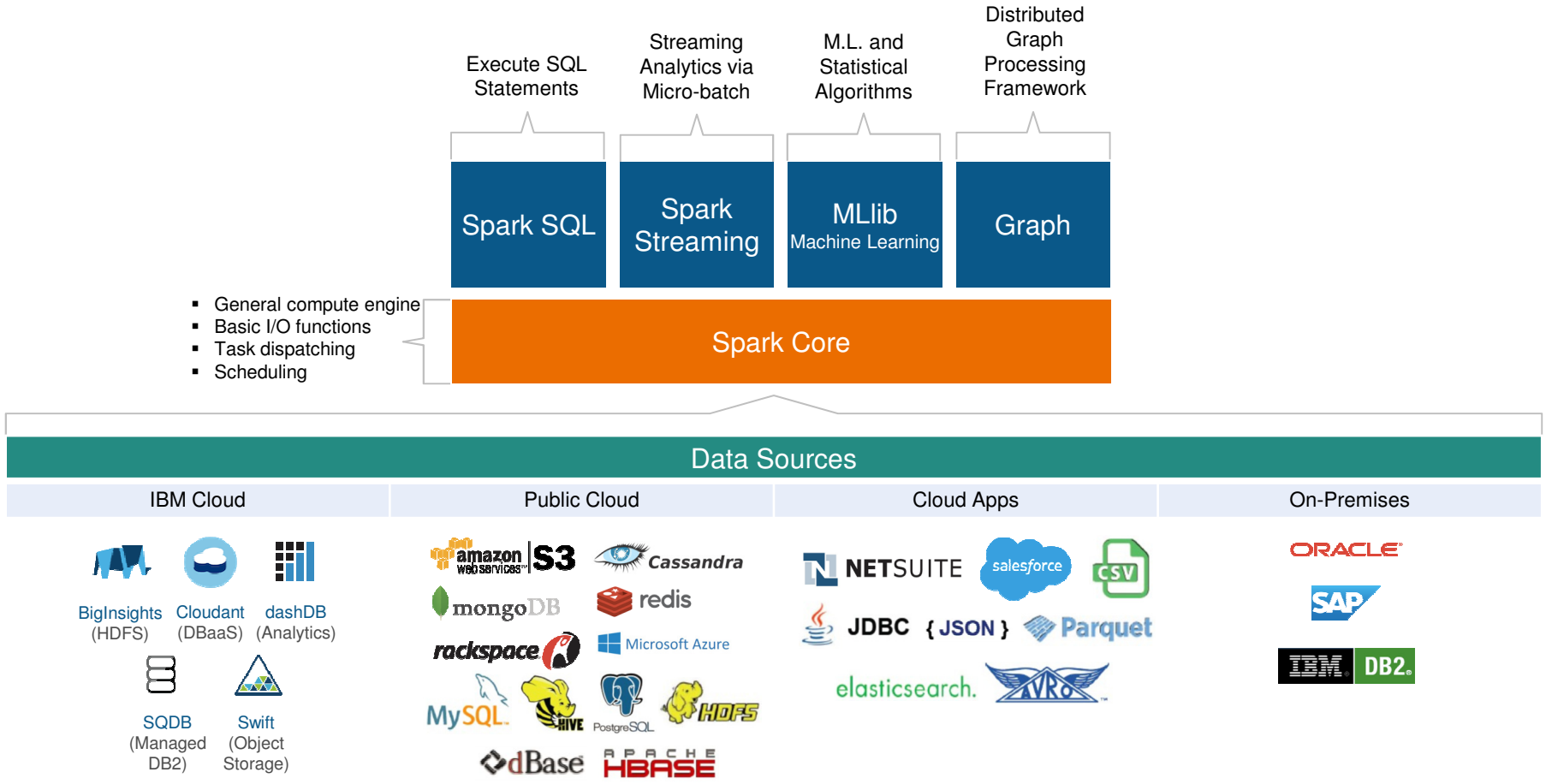
- **Notebooks are the digital equivalent of the “pen and paper” lab notebook, enabling data scientists to document reproducible analysis:**
  - Markdown and visualization
  - Iterative exploration
  - Easy to share



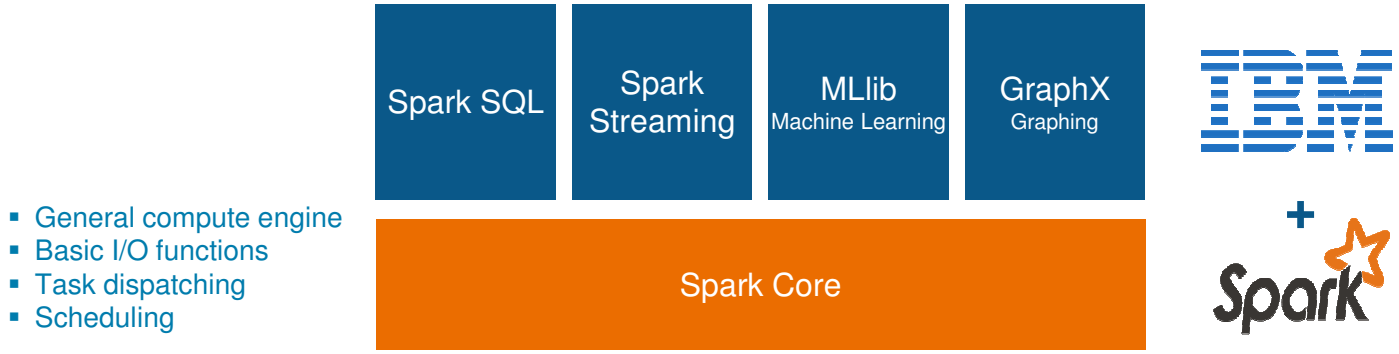
Integrated Jupyter Notebooks for interactive and collaborative development - seamless execution on Spark



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



Benefits of Spark for Data Science



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

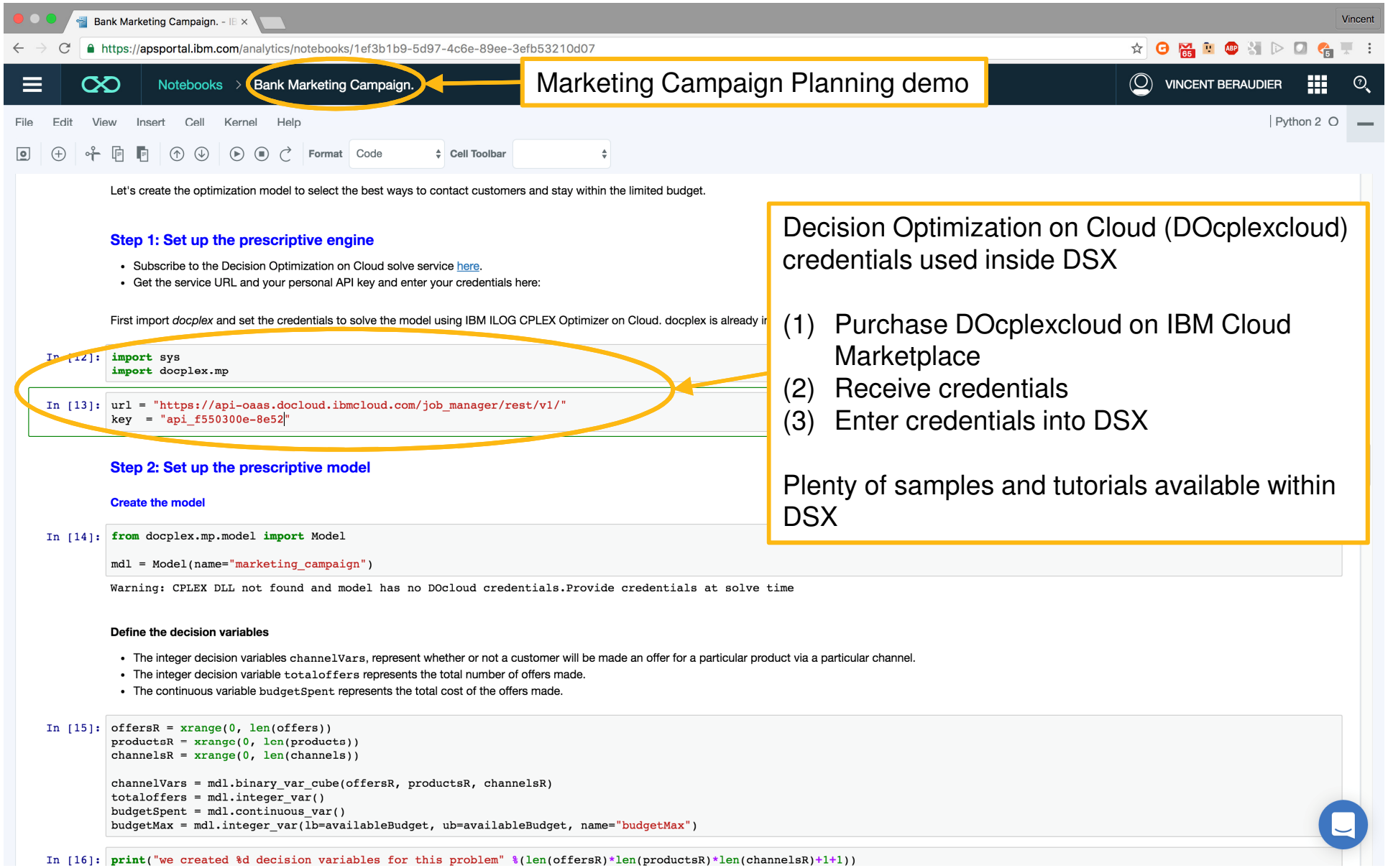
The Spark service uses Bluemix Object Storage as its preferred data store for building performant applications

- 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

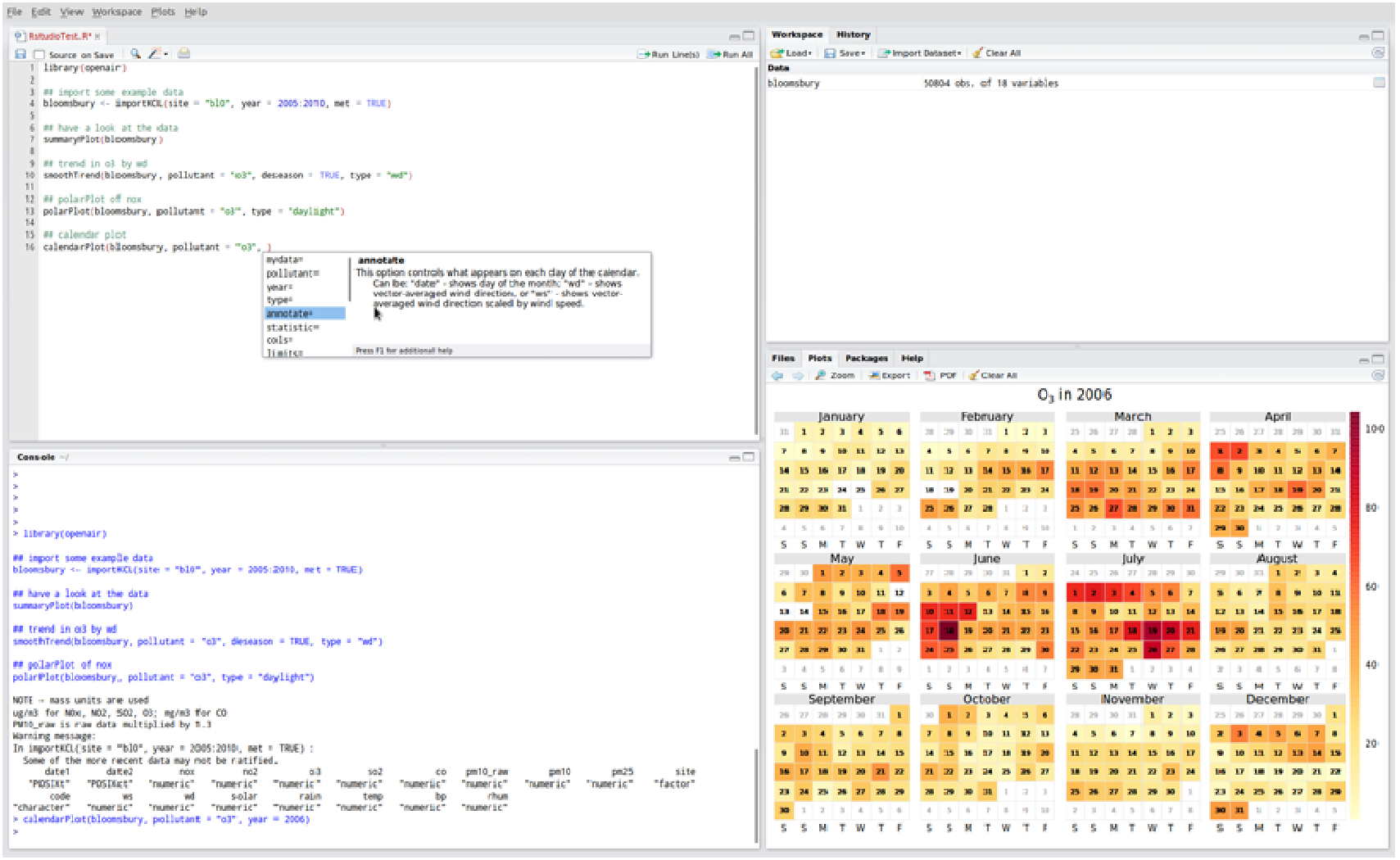
Object Storage

IBM

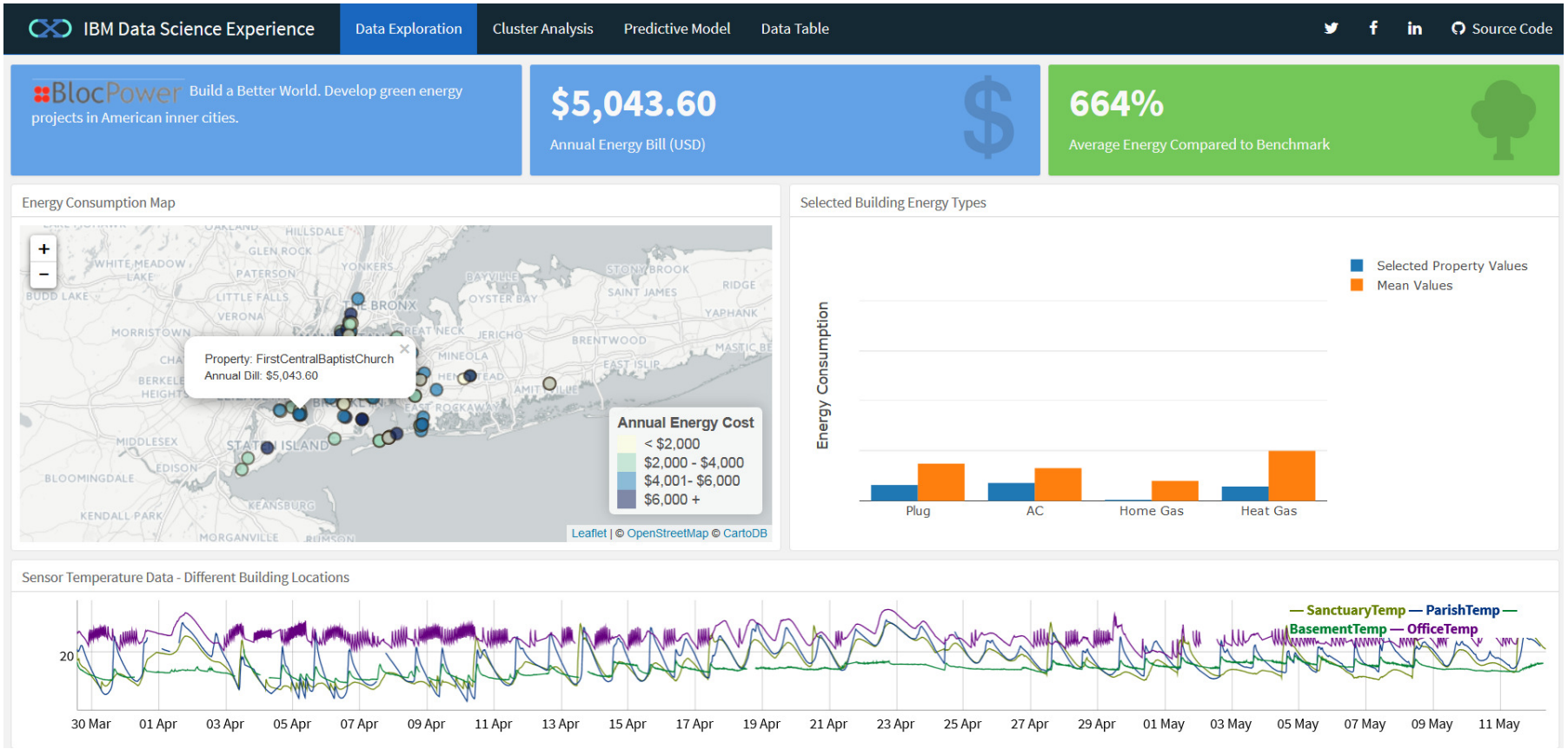
IBM Decision Optimization for DSX via API calls to DOcplexcloud



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



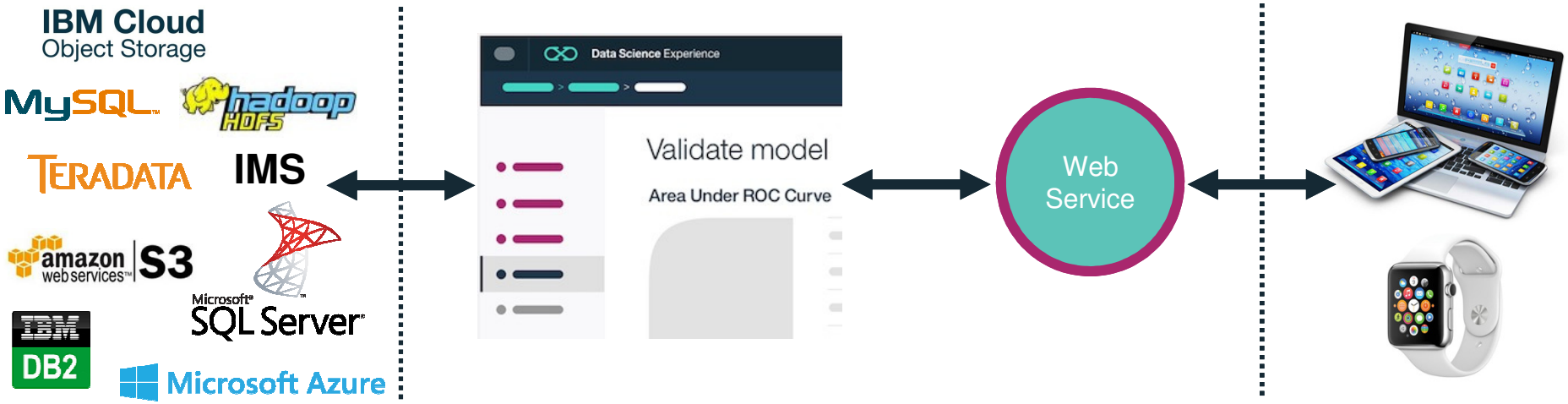
With RStudio you can create Shiny web applications to make your analysis accessible to the business



IBM Disclaimer

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion.
Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.
The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality.
Information about potential future products may not be incorporated into any contract.
The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

Operationalize insights with IBM Machine Learning (Closed Beta)



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:

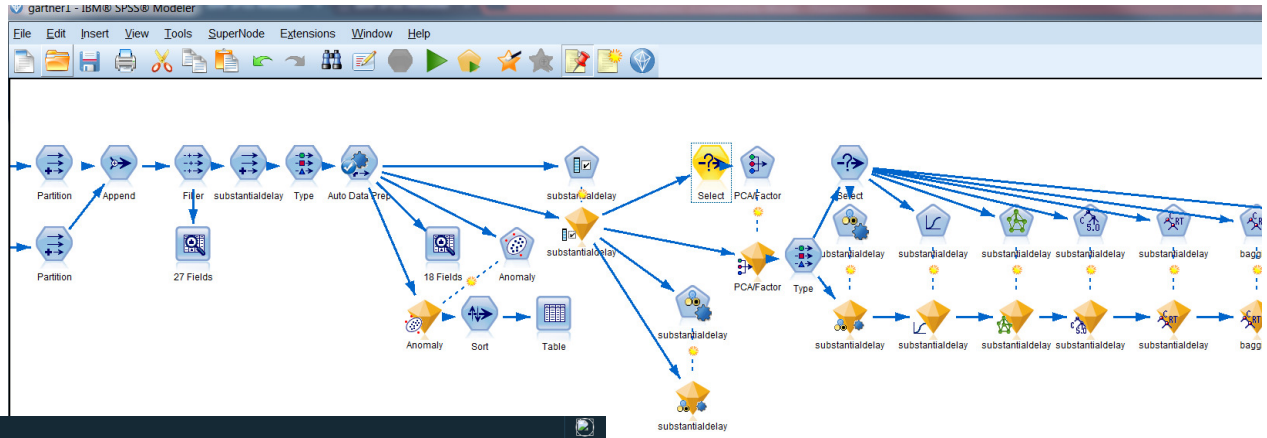
- Real-time, Streaming, and Batch Deployment
- Continuous Monitoring and Feedback Loop

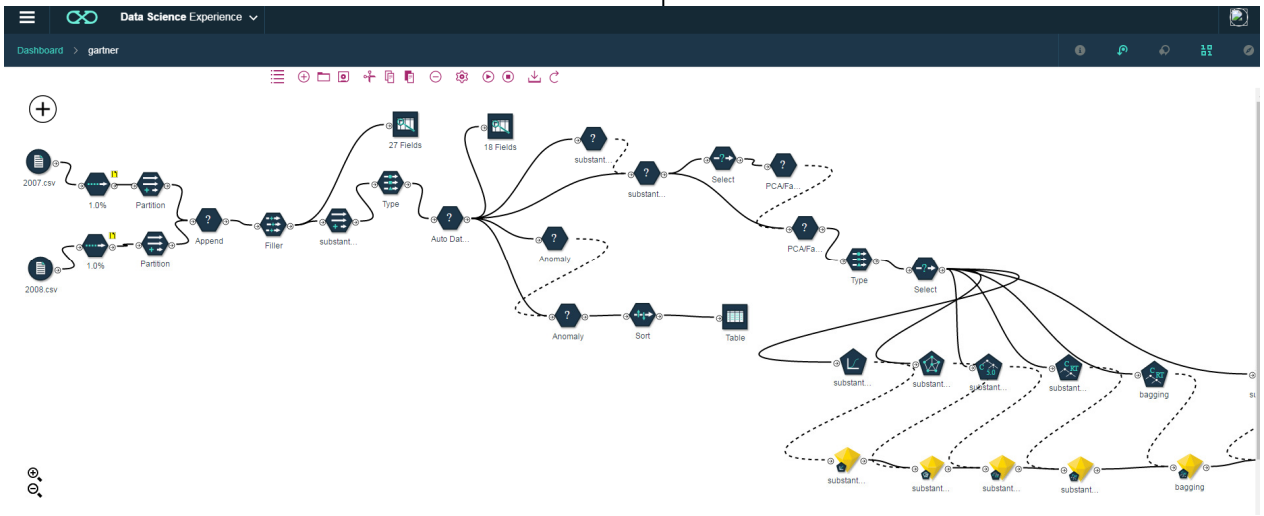
Intelligent Apps:

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

The DSX Canvas is planned to be compatible with existing SPSS Modeler Streams and be deployable via IBM Machine Learning (closed beta)

- DSX Canvas will have compatibility with legacy SPSS Modeler streams
- Multiple execution runtimes: SPSS Modeler(first), SparkML (later)
- Support for R/Python/SQL code in the pipeline





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

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
- **Sign up** to try it out at **datascience.ibm.com**

DEMO

Analytics Solution Center

THANK YOU!

Contact info: beekmanb@us.ibm.com

www.ibm.com/ASCdc

APPENDIX

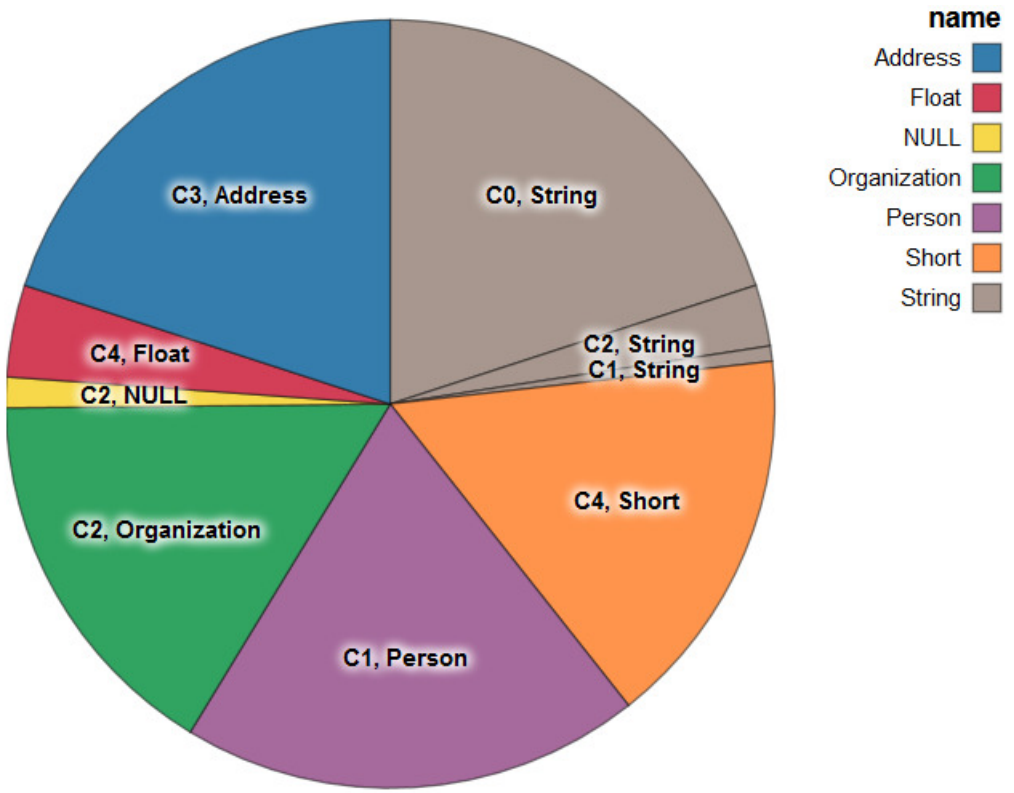
Supported Data Sources for DSX via on-premises and cloud Connections

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		

All of the supported targets are compatible with each source

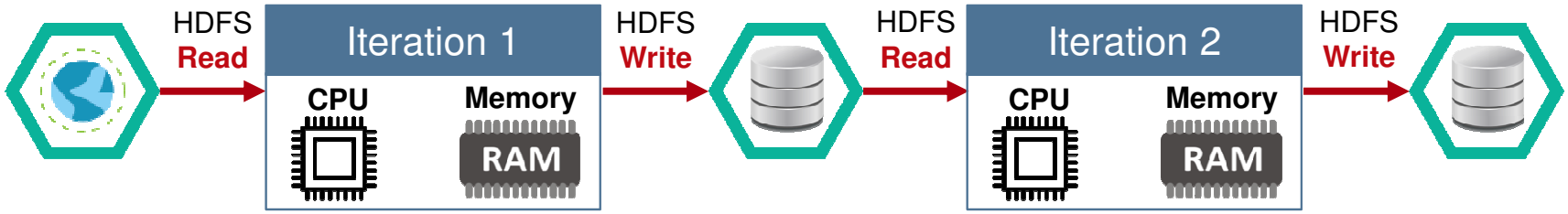
IBM Sparkling Data API enables users to perform scalable data cleaning and shaping within DSX Notebooks

- Sparkling Data Library provides tools to map, visualize and transform data for iterative analysis
- Outputs quality metrics, metadata and summary statistics for visceral sense of the data preparation progress



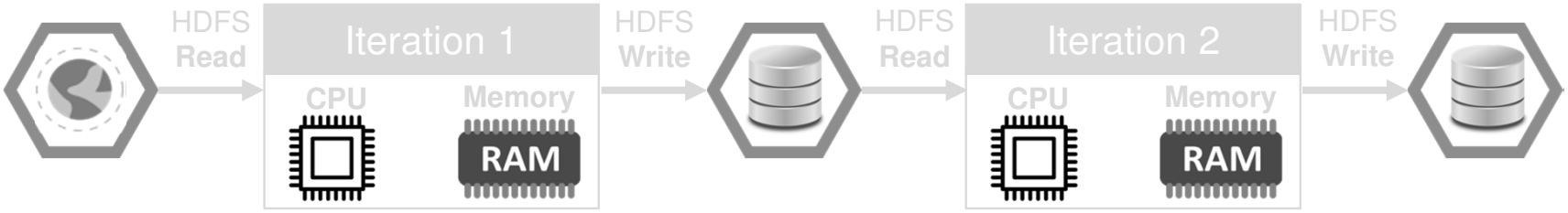
Motivation for Apache Spark

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

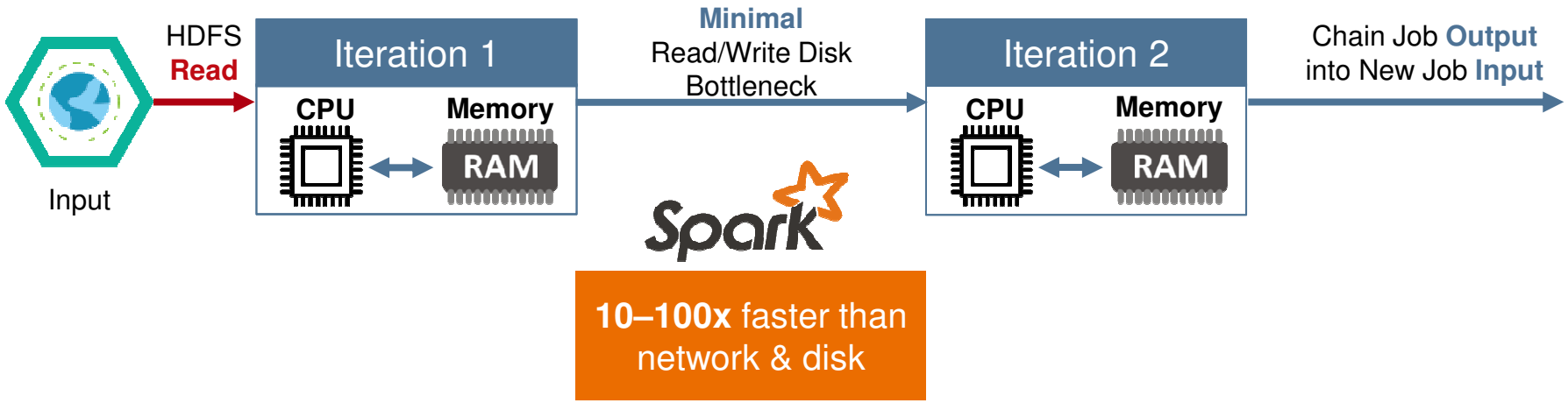


Motivation for Apache Spark

- **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**

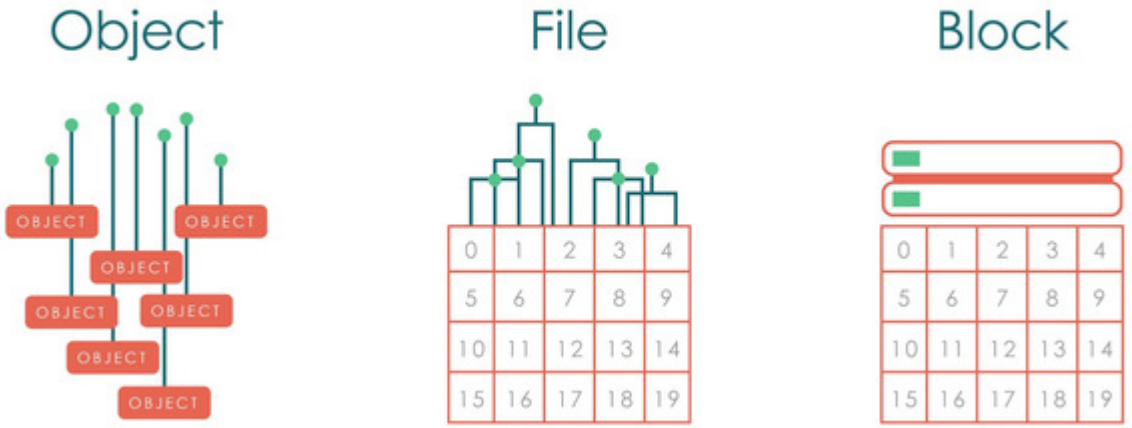


Object Storage

- 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

- 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

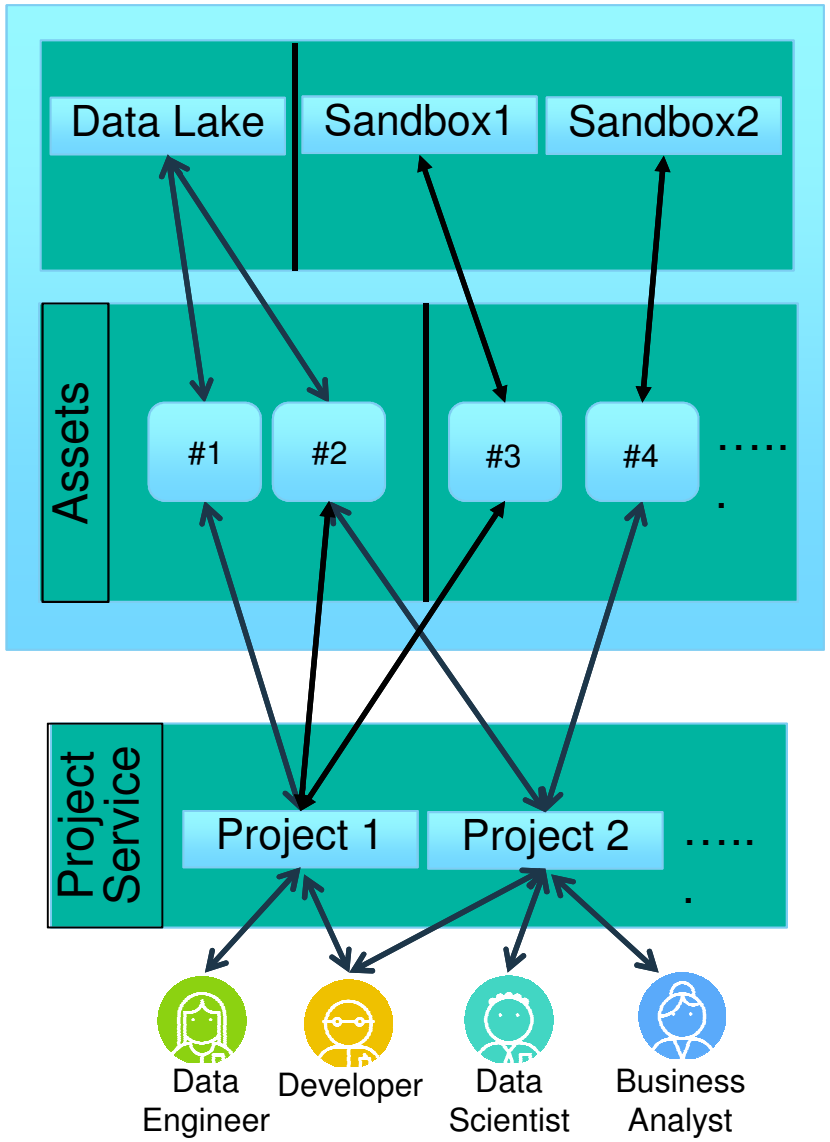
- Cloud Storage - store all your files (images, documents, and more) in the cloud. Use metadata to quickly tag and search your objects. Easily compose and bind to your object files from your Bluemix application.

- 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



IBM's Commitment to R

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.

IBM and RStudio Strategic Partnership

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.

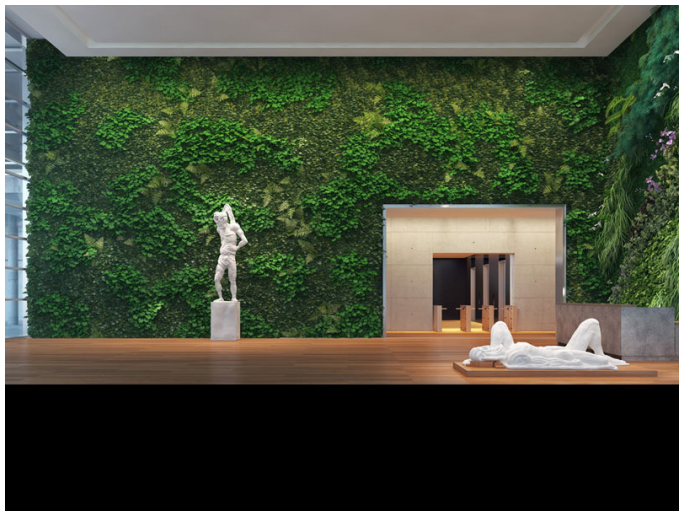
Overview of partnership
IBM and RStudio will collaborate to enable R packages to integrate with Apache Spark. R is a popular statistical programming language that offers a rich environment for statistical analysis and machine learning. The runtime for R is single-threaded, and by ensuring R code and 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

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

505 Howard Street, San Francisco



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

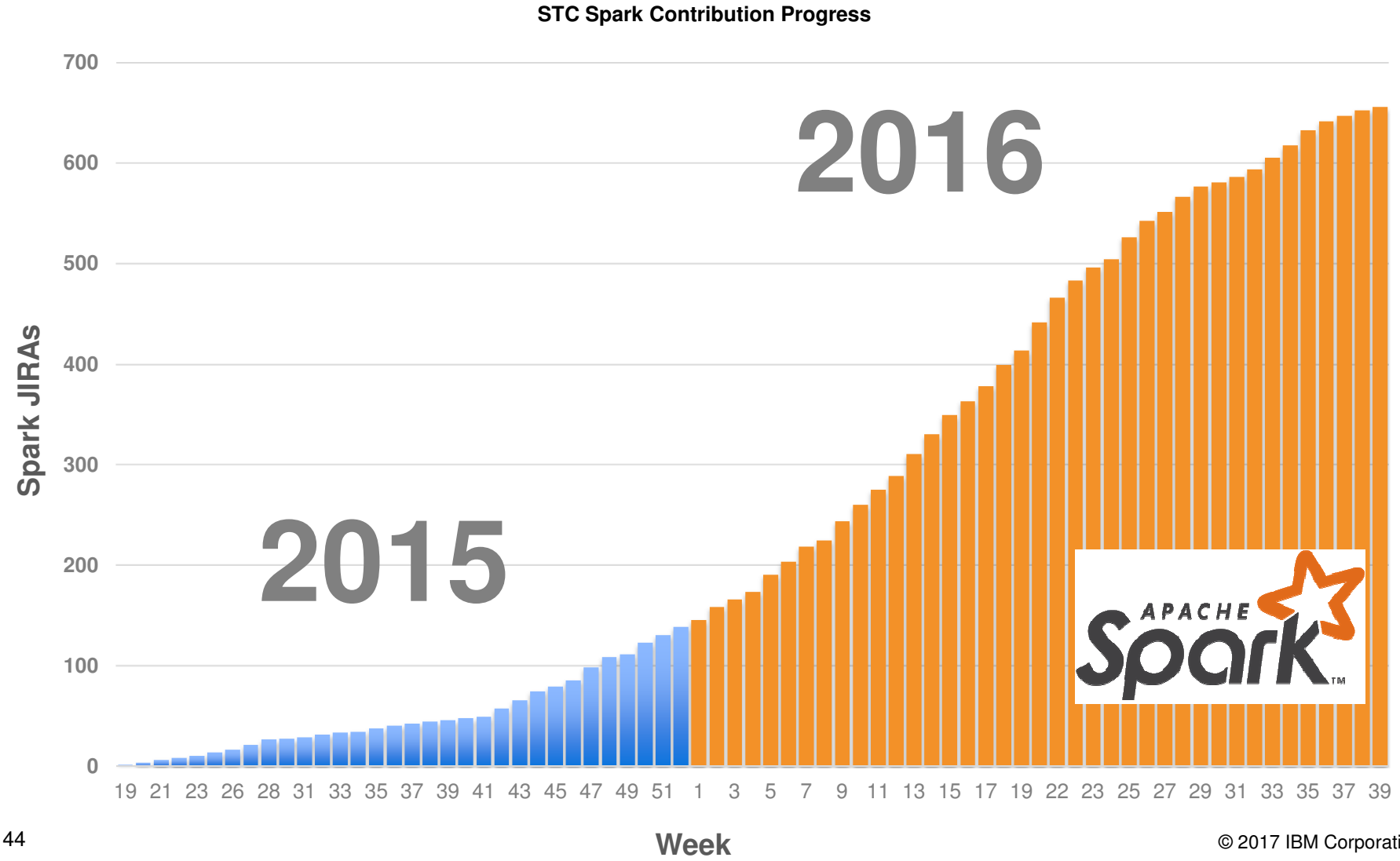
Founding member of AMPLab (and upcoming RISE Lab)

Member of R Consortium

Founding member of Scala Center

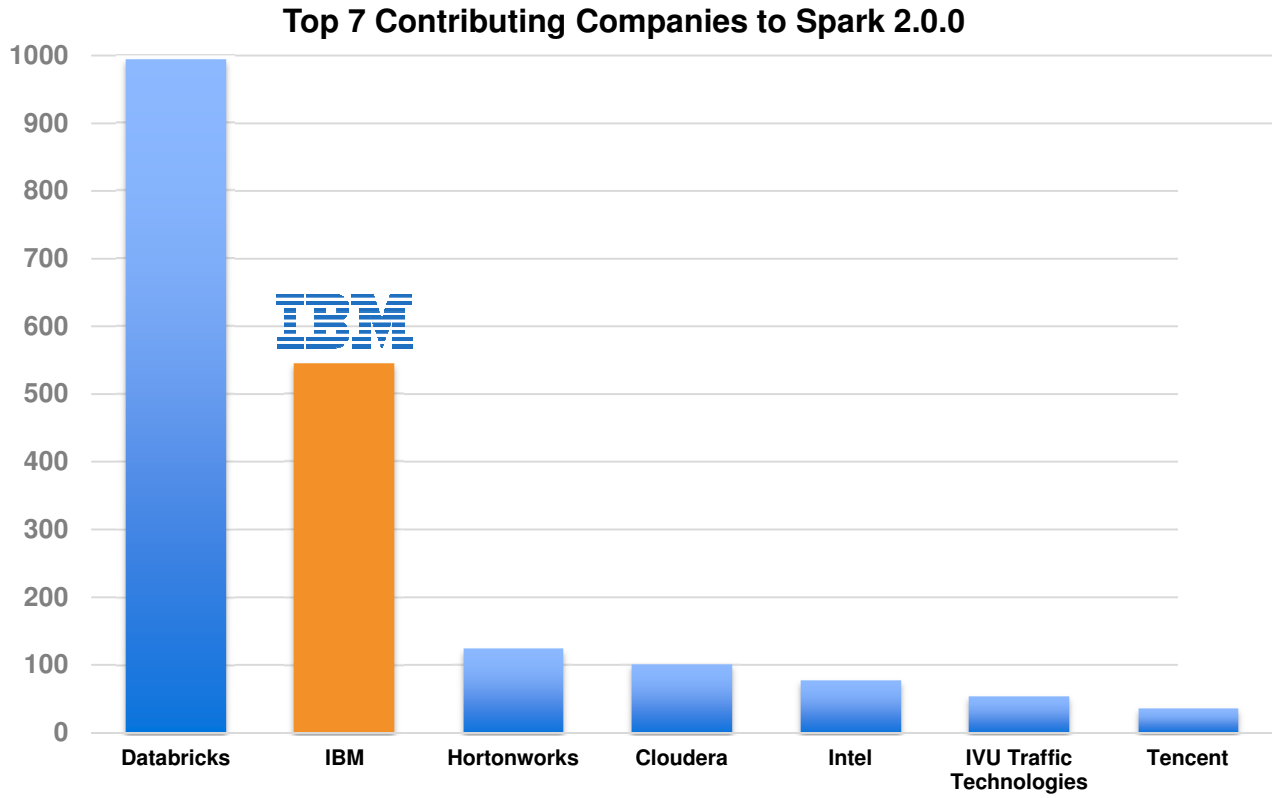
Partnerships in the ecosystem

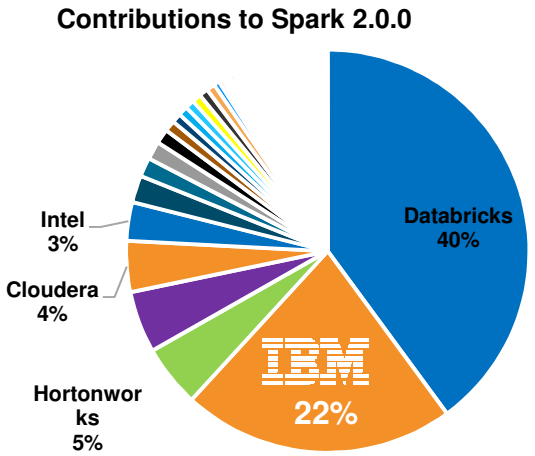
Spark Technology Center contributions have grown over 400% since start in June 2015



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	▪ Analytics Exchange
Data Storage	▪ **On-Premises:** IBM Open Platform with Apache Hadoop (IOP), BigInsights, Neteeza, Cloudant, DB2, dashDB local and Informix ▪ **On-Cloud:** Cloudant, dashDB, Object Storage, SQL DB, BigInsights
Data Feeds, Load & Refinement	▪ Watson Data Platform ▪ IBM Streams ▪ IBM Insights for Twitter ▪ IBM Insights for Weather
Analytics and Solutions	▪ IBM Analytics for Apache spark ▪ SPSS Modeler and Analytics Server ▪ Watson Analytics ▪ Watson Health ▪ IBM Commerce ▪ Data Science Experience
Learning Tools	▪ Big Data University

IBM Apache Spark References

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.