Wayne State Big Data & Business Analytics Symposium  
March 23, 2017

Make Data Simple:  
IBM Watson Data Platform & Data Science Experience

Art Derico  
Analytics Client Leader
Ryan Kather  
Open Source Analytics Engineer
David West  
Data Scientist
Chris Williams  
Open Source Analytics Technical Leader
Tomorrow’s disruptors will be organizations that can converge digital business with a new level of digital intelligence.

Digital is the wires, but digital intelligence, or artificial intelligence as some people call it, is about much more than that. This next decade is about how you combine those and become a cognitive business.” — Ginni Rometty
Cognitive systems are generally defined by the ability to understand, reason, learn, and interact.

**Definition of Cognitive Technologies**

**UNDERSTAND**
Cognitive systems can understand unstructured information the same way humans do.

**REASON**
They can reason, grasp underlying concepts, form hypotheses, and infer to extract ideas.

**LEARN**
Each data point, interaction and outcome helps to continuously sharpen expertise.

**INTERACT**
With abilities to see, talk and hear, cognitive systems interact with humans in a natural way.

Watson is the platform for cognitive business

[https://youtu.be/kFY4ZGvq9IA](https://youtu.be/kFY4ZGvq9IA)
Introducing IBM Watson Data Platform
The first data and analytics platform for the Cognitive Business

http://ibm.co/makedatasimple
IBM Watson Data Platform
Experience New Ways To Put Data To Work

Data Engineering
Data Science
Business Analysis
App Development

Experiences
task-specific, collaborative

Data and Analytics Services
comprehensive

open • intelligent • hybrid
IBM Watson Data Platform
A closer look at what makes up the platform

- **Data Engineering**
  - Bluemix Data Connect

- **Data Science**
  - Data Science Experience

- **Business Analysis**
  - Watson Analytics

- **App Development**
  - Bluemix Platform

**Ingest**
- IIG Catalog
- Bluemix Data Connect
- Bluemix Lift
- Streaming Analytics

**Persist**
- BigInsights on Cloud
  - IBM Cloudant
  - Compose / Graph
  - DashDB
  - Object Store

**Analyze**
- Data Science Experience
  - Watson Analytics, Cognos Analytics
  - SPSS, Decision Optimization

**Deploy**
- Spark-as-a-Service
  - IBM Machine Learning
  - Streaming Analytics

**Govern**
- Open Metadata Services
  - (coming soon)

- IBM Machine Learning
- IIG Catalog

- Bluemix DevOps Platform
- IBM Machine Learning
Tailored Experiences and User Collaboration

**OUTPUT**
- Deliver and Deploy Model
- Understand Problem and Domain
- Ingest Data
- Explore and Understand Data
- Communicate Results
- Create and Build Model
- Evaluate
- Transform: Clean
- Transform: Shape

**INPUT**
- 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 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
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

http://datascience.ibm.com
IBM is positioned as a Leader in the 2017 Gartner Magic Quadrant for Data Science Platforms, based on its ability to execute and completeness of vision.

IBM is positioned highest in execution and furthest in vision within the Leaders Quadrant.

This is the fourth consecutive year for IBM to be recognized as a leader in this report.
Core Attributes of the Data Science Experience

Community

- Find tutorials and datasets
- Connect with Data Scientists
- Ask questions
- Read articles and papers
- Fork and share projects

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 - DOcplexcloud
- Projects and Version Control
- Managed Spark Service

* Closed beta

Powered by IBM Watson Data Platform
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.

datasceince.ibm.com
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
DSX / Spark Tutorial

HTTPS://GITHUB.COM/CARLOAPP2/SPARK-POT
APPENDIX
### Supported Data Sources for DSX via on-premises and cloud Connections

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

All of the supported targets are compatible with each source.
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

Wizard GUI
Data Scientist Workbench vs Data Science Experience

Demo1: Flight Delay Prediction

In this tutorial, we use the popular Flights Dataset to analyze and predict flight delays in airports based on past flight records. We show how you can use Jupyter Notebook and Local Spark to read, explore, analyze and visualize your results.

For this dataset, we will only look at the flights in 2007 - this is still 7 million flights!
Select Libraries to Meet Use-Case Challenges

- Unified data access: Query structured data sets with SQL or DataFrame APIs
- Fast, familiar query language across all of your enterprise data
- Use BI tools to connect and query via JDBC or ODBC drivers
Select Libraries to Meet Use-Case Challenges

- Micro-batch event processing for near-real time analytics
- e.g. Internet of Things (IoT) devices, Twitter feeds, Kafka (event hub), etc.
- Spark’s engine drives some action or outputs data in batches to various data stores
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”

\[ f(x) = \sum_{i=0}^{n} \alpha_i y_i x_i^T x + b \]
PySpark Notebook and Shiny App for Demo

- 23 commits
- 1 branch
- 0 releases
- 1 contributor

gfilla committed on GitHub update - readme environment details, manage files

- .ipynb_checkpoints
- R
- Screenshots
- data
- BlocPower with Sparkling.ipynb
- README.md

Latest commit 4d45748 12 days ago

- update notebook 26 days ago
- add script 22 days ago
- add environment screenshot 12 days ago
- updating data a month ago
- update with new notebook 13 days ago
- 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

10–100x faster than network & disk
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.

1http://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.

- 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.
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

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

505 Howard Street, San Francisco
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

Top 7 Contributing Companies to Spark 2.0.0

Contributions to Spark 2.0.0

- Databricks 40%
- IBM 22%
- Hortonworks 10%
- Cloudera 4%
- Intel 3%
- IVU Traffic Technologies 5%
- Tencent 4%
# Spark Infused Across IBM Analytics Portfolio

<table>
<thead>
<tr>
<th>Category</th>
<th>Components</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free and Open Data</td>
<td>- Analytics Exchange</td>
</tr>
<tr>
<td>Data Storage</td>
<td>- <strong>On-Premises:</strong> IBM Open Platform with Apache Hadoop (IOP), BigInsights, Neteeza, Cloudant, DB2, dashDB local and Informix</td>
</tr>
<tr>
<td></td>
<td>- <strong>On-Cloud:</strong> Cloudant, dashDB, Object Storage, SQL DB, BigInsights</td>
</tr>
<tr>
<td>Data Feeds, Load &amp; Refinement</td>
<td>- Watson Data Platform</td>
</tr>
<tr>
<td></td>
<td>- IBM Streams</td>
</tr>
<tr>
<td></td>
<td>- IBM Insights for Twitter</td>
</tr>
<tr>
<td></td>
<td>- IBM Insights for Weather</td>
</tr>
<tr>
<td>Analytics and Solutions</td>
<td>- IBM Analytics for Apache spark</td>
</tr>
<tr>
<td></td>
<td>- SPSS Modeler and Analytics Server</td>
</tr>
<tr>
<td></td>
<td>- Watson Analytics</td>
</tr>
<tr>
<td></td>
<td>- Watson Health</td>
</tr>
<tr>
<td></td>
<td>- IBM Commerce</td>
</tr>
<tr>
<td></td>
<td>- Data Science Experience</td>
</tr>
<tr>
<td>Learning Tools</td>
<td>- Big Data University</td>
</tr>
</tbody>
</table>
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. 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.