

# Neues über Analytics auf z

Dr. Manfred Gnirß / Arwed Tschoeke  
Client Center Böblingen

© 2016 IBM Corporation

Linux on IBM z Systems



## Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.

BladeCenter*	FICON*	OMEGAMON*	RACF*	System z9*	zSecure
DB2*	GDPS*	Performance Toolkit for VM	Storwize*	System z10*	z/VM*
DS6000*	HyperSockets	Power*	System Storage*	Tivoli*	z Systems*
DS8000*	HyperSwap	PowerVM	System x*	zEnterprise*	
ECKD	IBM z13*	PR/SM	System z*	zOS*	

\* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

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.

Call Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, 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.

IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency which is now part of the Office of Government Commerce.

ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.

Java and all Java based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

OpenStack is a trademark of OpenStack LLC. The OpenStack trademark policy is available on the [OpenStack website](#).

TEALEAF is a registered trademark of Tealeaf, an IBM Company.

Windows Server and the Windows logo are trademarks of the Microsoft group of countries.

Worklight is a trademark or registered trademark of Worklight, an IBM Company.

UNIX is a registered trademark of The Open Group in the United States and other countries.

\* Other product and service names might be trademarks of IBM or other companies.

### Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon 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 throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

This information provides only general descriptions of the types and portions of workloads that are eligible for execution on Specialty Engines (e.g., zIPs, zAAPs, and IFLs) ("SEs"). IBM authorizes customers to use IBM SE only to execute the processing of Eligible Workloads of specific Programs expressly authorized by IBM as specified in the "Authorized Use Table for IBM Machines" provided at [www.ibm.com/systems/support/machine\\_warranties/machine\\_code/aut.html](#) ("AUT"). No other workload processing is authorized for execution on an SE. IBM offers SE at a lower price than General Processors/Central Processors because customers are authorized to use SEs only to process certain types and/or amounts of workloads as specified by IBM in the AUT.




Linux on IBM z Systems

SE  
STRADE WARD STRIVE  
Deutsche Region

IBM

# Watson Explorer



3

© 2016 IBM Corporation

Linux on IBM z Systems

SE  
STRADE WARD STRIVE  
Deutsche Region

IBM

## Product history

Lineage of WEX foundational components

- Vivisimo: founded by 3 Carnegie Mellon computer scientists in 2000
- Initial focus on clustering
- Focus shifted to Enterprise Search in 2005
- Acquired by IBM in 2012
  - Rebranded Infosphere Data Explorer by Information Management
- Merged with IBM Content Analytics in 2014 Watson reorg
  - Combined offering rebranded as Watson Explorer

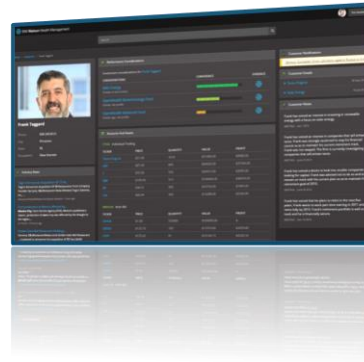
4

4

© 2016 IBM Corporation

# IBM Watson Explorer

- **Explore:** Securely connects and explores data, regardless of format or location
- **Analyze:** Mines unstructured content to reveal trends, patterns and insights
- **Interpret:** Delivers cognitive capabilities via the Watson Developer Cloud



**A single point of access for information and services**

- » On-premise enterprise systems
- » Public and private cloud
- » Public and private Internet sources

5

# Watson Explorer makes data from enterprise and non-enterprise silos easily accessible at “the point of impact” to people when they need it

## Unstructured Content

- Collaboration
- Web
- Email
- File Systems
- Content Management
- Cloud
- More...



## Structured Data

- Databases
- Data Warehouses
- Web Services
- Cloud
- More...

6

## Challenges

### Watson Explorer

**Information Access**  
Data, applications and services distributed on-premise and in cloud—employees struggle to get a complete view



**Explore**  
Provides a 360-degree view of information from ALL sources to enable better decisions

**Unstructured Content**  
80% of data is unstructured but only a small percentage leveraged for insights



**Analyze**  
Delivers insights from unstructured content

**Scaling Expertise**  
Pressure to increase performance and innovation—while doing more with less



**Interpret**  
Applies cognitive computing to scale human expertise

## Watson Explorer raises the bar with cognitive exploration

### Explore



**Watson Explorer**  
Search, visualize, and explore information across enterprise applications through 360° views of any topic



### Analyze



**Content Analytics**  
Analyze, visualize, and discover insight in structured and unstructured data through NLP and content mining



### Interpret

- Question Answer
- Relationship Extraction
- Concept Expansion
- User Modeling
- Language Identification
- Machine Translation
- Message Resonance
- ... more ...

**Watson Developer Cloud**  
Enhance, scale, and accelerate human expertise through user modeling, relationship extraction, and more



Linux on IBM z Systems

Watson Explorer Component View

**Watson Explorer Applications**  
Search • Analyze • Interpret

Mobile Security Collaboration Query Routing Application Builder Solution Gallery Content Miner Studio

**Management and application development**

Indexing Search Text analytics Content analytics

**Indexing, search and analytics**

File systems CMS Email DBMS Support Social SCM Wikis CRM External  
Big Data Cloud

**Connectivity**

**Watson Developer Cloud**  
Cognitive and information analysis services

- Question Answer
- Relationship Extraction
- Concept Expansion
- User Modeling
- Language Identification
- Machine Translation
- Message Resonance
- ... more ...

= available with Advanced Edition

9 © 2016 IBM Corporation

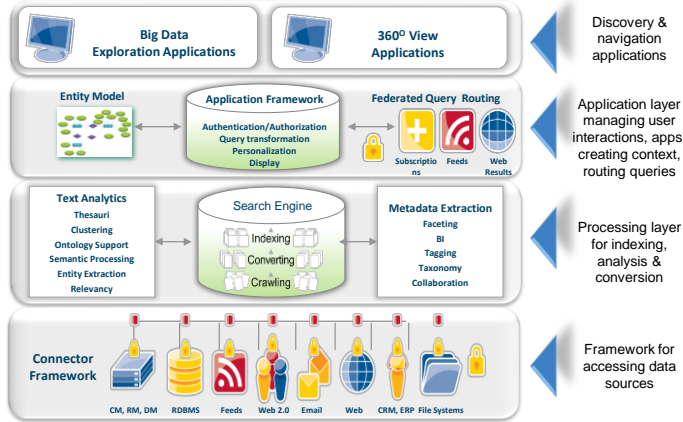
Linux on IBM z Systems

Why Watson Explorer is unique

- **Cognitive exploration** – unique combination of search, exploration computing
- Ability to **securely access and fuse structured and unstructured data** from many sources
- **Rapid deployment of 360-degree information applications** for complete, contextually-relevant view of information personalized to the user + collaborative and cognitive
- **Deep content mining** to unlock insights in unstructured *content*—without writing a single line of code
- Truly unique **cognitive computing** capabilities through Watson Developer Cloud integration
- Ability to do all of this **at scale** for organizations with TBs of data

10 © 2016 IBM Corporation

## Watson Explorer Application Architecture

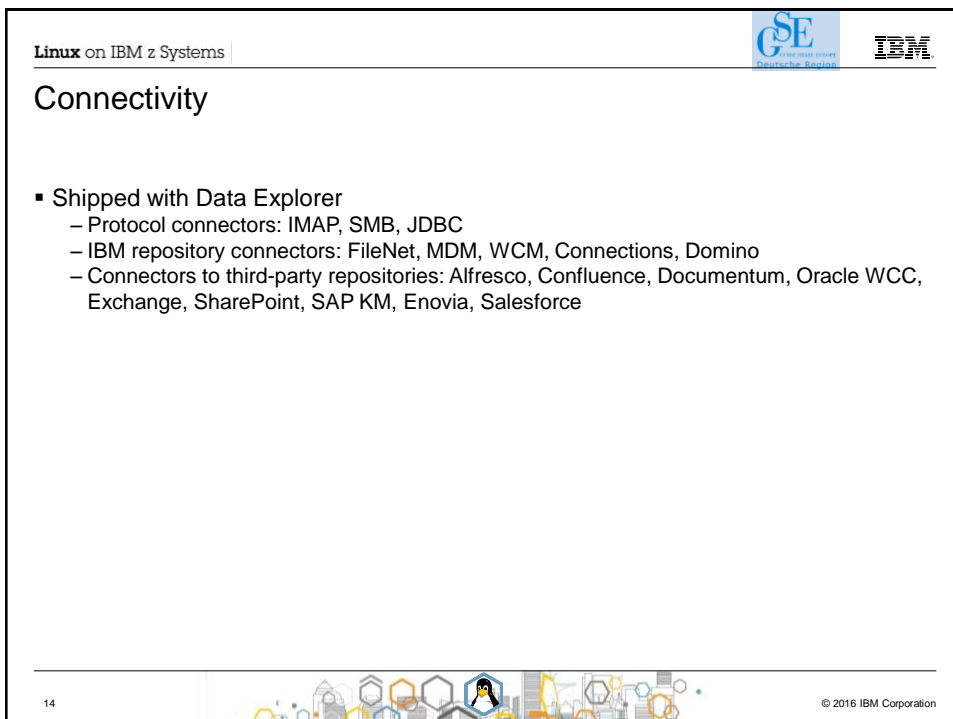
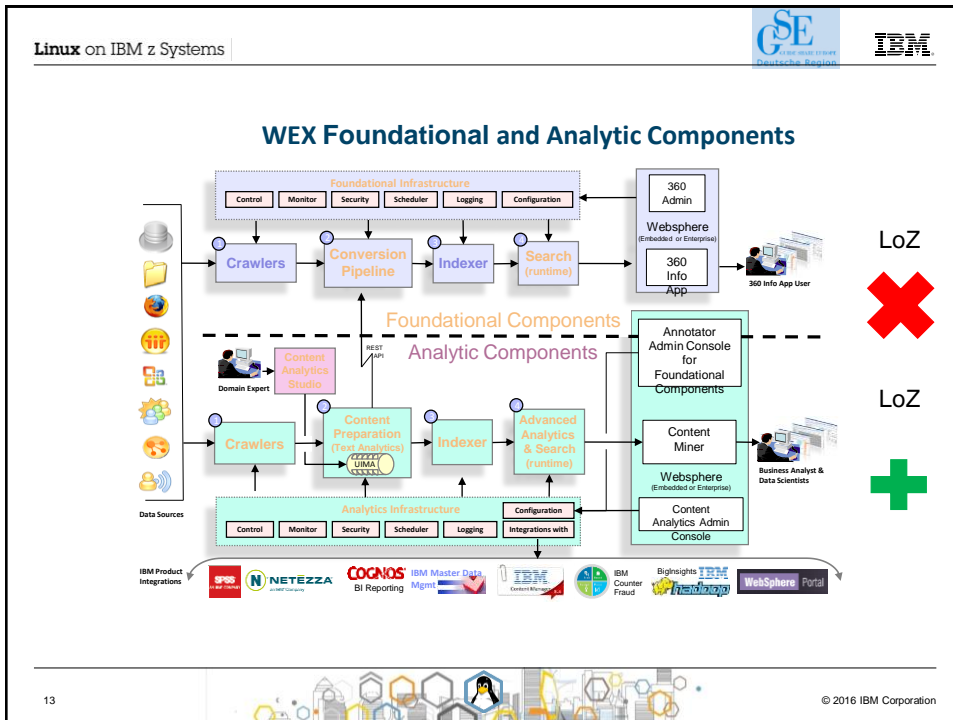


## WEX is available

Linux [Detailed system requirements](#) Filter

Operating System	Operating System Minimum	Hardware	Bitness	Components	Notes	Details
Red Hat Enterprise Linux (RHEL) Server 6	Update 4	IBM z Systems	64-Exploit	<input checked="" type="checkbox"/>	(6)	<a href="#">View</a>
Red Hat Enterprise Linux (RHEL) Server 7	Base	IBM z Systems	64-Exploit	<input checked="" type="checkbox"/>	(7)	<a href="#">View</a>
Red Hat Enterprise Linux (RHEL) Server 7	Base	POWER System - Big Endian	64-Exploit	<input checked="" type="checkbox"/>	(5)	<a href="#">View</a>
Red Hat Enterprise Linux (RHEL) Server 6	Varies	x86-64	64-Exploit	<input checked="" type="checkbox"/>	(3) (6)	<a href="#">View</a>
Red Hat Enterprise Linux (RHEL) Server 7	Base	x86-64	64-Exploit	<input checked="" type="checkbox"/>	(1)	<a href="#">View</a>
SUSE Linux Enterprise Server (SLES) 11	Base	IBM z Systems	64-Exploit	<input checked="" type="checkbox"/>	(2)	<a href="#">View</a>
SUSE Linux Enterprise Server (SLES) 12	Base	IBM z Systems	64-Exploit	<input checked="" type="checkbox"/>	(7)	<a href="#">View</a>
SUSE Linux Enterprise Server (SLES) 11	Base	x86-64	64-Exploit	<input checked="" type="checkbox"/>	(4)	<a href="#">View</a>
SUSE Linux Enterprise Server (SLES) 12	Base	x86-64	64-Exploit	<input checked="" type="checkbox"/>	(1)	<a href="#">View</a>

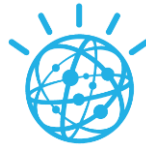
**Server**  Watson Explorer Analytical Components  Watson Explorer Foundational Components



## Analytics creates actionable data from unstructured content

### Unstructured content

Commander 4.0 Cu. Ft.  
 26-Cycle King-Size washer – white. I hate this machine. Have had 3 calls on machine. You can't wash **large items**, Won't clean in the middle. **Leaves dry spots** through the clothes, I can only do **½ basket** of clothes. Will **not clean** or **mix bleach** in with the water...

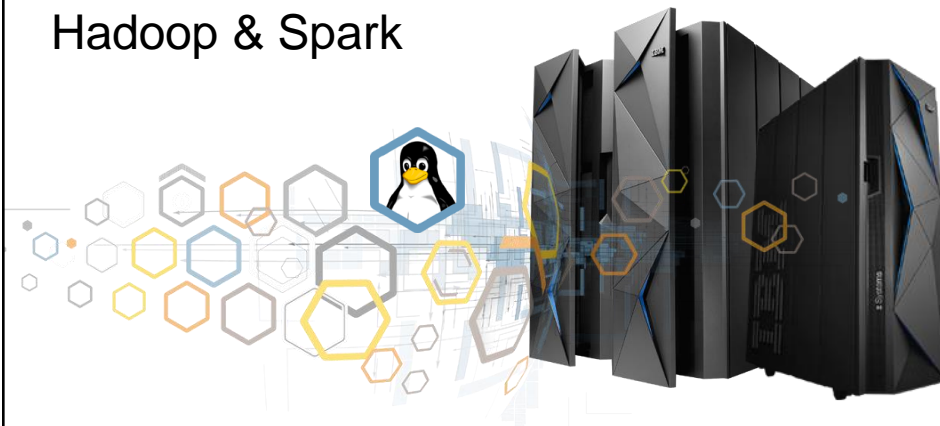


**Watson Explorer**  
 Deep natural-language analysis

### Structured data for analysis

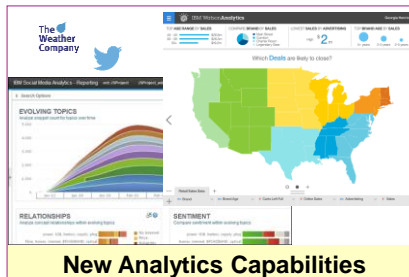
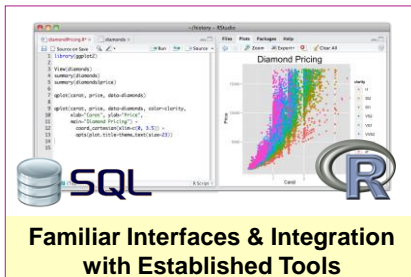
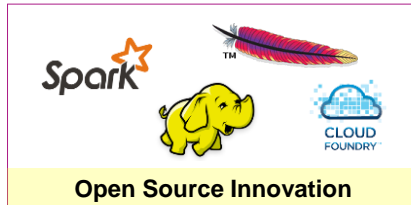
Product	Commander
Category	4.0 Cu. Ft.
Size	26-Cycle King-Size
Model	washer
Color	white
Issue	large items
Issue	leaves dry spots
Issue	½ basket
Issue	not clean
Issue	mix bleach

## Hadoop & Spark





## IBM Investing in Four Catalysts for Big Data Adoption



17

© 2016 IBM Corporation

## Our commitment to Spark

### Announcing:

Open Source SystemML

Educate one million data professionals

Establish Spark Technology Center

Founding Member of AMPLab

Contributing to the core

18

© 2016 IBM Corporation

## Hadoop Advantages

### Unlimited Scale

- Multiple data sources
- Multiple applications
- Multiple users

- Reliability
- Resiliency
- Security

### Enterprise Platform

### Wide Range of Data Formats

- Files
- Semi-structured
- Databases



## Hadoop MapReduce Challenges

- Need deep Java skills
- Few abstractions available for analysts

### Ease of Development

### In-Memory Performance

- No in-memory framework
- Application tasks write to disk with each cycle

- Only suitable for batch workloads
- Rigid processing model

### Combine Workflows



## Spark Advantages

- Easier APIs
- Python, Scala, Java

Ease of Development

In-Memory Performance

- Resilient Distributed Datasets
- Unify processing

- Batch
- Interactive
- Iterative algorithms
- Micro-batch

Combine Workflows



## Spark Libraries

Spark SQL

Spark Streaming

GraphX

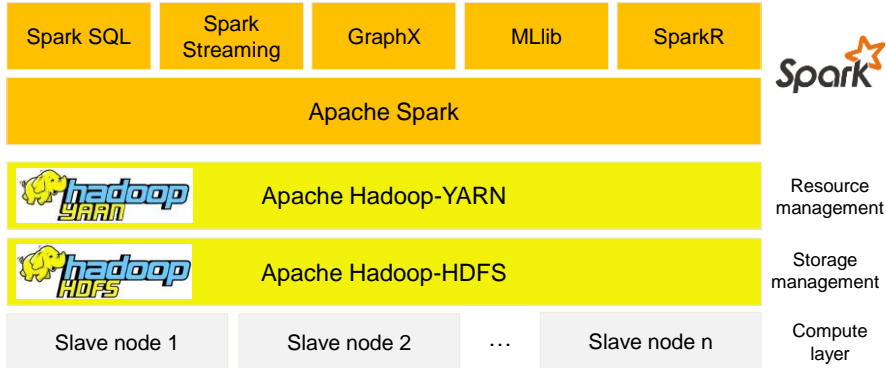
MLlib

SparkR

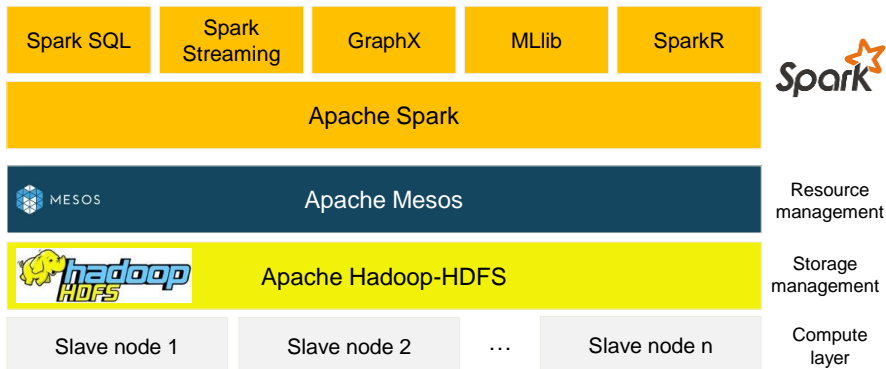
Apache Spark



## Spark on Hadoop






## Spark on Mesos



Linux on IBM z Systems GSE GLOBAL SERVICE CENTER  
Deutsche Region IBM

## Spark as a Service




Spark SQL	Spark Streaming	GraphX	MLlib	SparkR	
Apache Spark					
		Apache Hadoop-YARN			Resource management
Cloud Storage Management					Storage management
Node 1	Node 2	...	Node n		Compute layer

25 © 2016 IBM Corporation

Linux on IBM z Systems GSE GLOBAL SERVICE CENTER  
Deutsche Region IBM

## Spark Running in Standalone Mode

Sweetspot for LoZ (same for Hadoop)

Spark SQL	Spark Streaming	GraphX	MLlib	SparkR	
Apache Spark					
Single node, with local storage					Resource management
					Storage management
					Compute layer

26 © 2016 IBM Corporation

Linux on IBM z Systems

SE  
STRONG EARLY SUPPORT  
Deutsche Region

IBM

## Spark Resilient Distributed Datasets

The diagram illustrates the mapping between Spark RDD partitions and HDFS blocks across three slave nodes:

- Spark RDD (In-memory distribution):**
  - RDD1: partition3, partition1, partition2
  - RDD2: partition1, partition2, partition2
  - RDD3: partition2, partition3, partition1
- HDFS (On-disk distribution):**
  - Slave node 1: a<sup>2</sup> b<sup>1</sup>, c<sup>3</sup> d<sup>2</sup>
  - Slave node 2: a<sup>1</sup> b<sup>2</sup>, c<sup>2</sup> d<sup>1</sup>
  - Slave node 3: a<sup>3</sup> b<sup>3</sup>, c<sup>1</sup> d<sup>2</sup>

hadoop HDFS

27

© 2016 IBM Corporation

Linux on IBM z Systems

SE  
STRONG EARLY SUPPORT  
Deutsche Region

IBM

## The Combination: The Flexibility of Spark on a Stable Hadoop Platform

Unlimited Scale	Ease of Development
In-Memory Performance	Enterprise Platform
Wide Range of Data Formats	Combine Workflows

Spark

28

© 2016 IBM Corporation

## Key reasons for interest in Spark

Beware of the hype!

### Performant



- In-memory architecture greatly reduces disk I/O
- Anywhere from **20-100x faster** for common tasks

### Productive



- **Concise and expressive syntax**, especially compared to prior approaches
- **Single programming model** across a range of use cases and steps in data lifecycle
- **Integrated with common programming languages** – Java, Python, Scala
- **New tools** continually reduce skill barrier for access (e.g. SQL for analysts)

### Leverages existing investments



- Works well within **existing Hadoop ecosystem**

### Improves with age



- **Large and growing community** of contributors continuously improve full analytics stack and extend capabilities

29

© 2016 IBM Corporation

## 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**



30

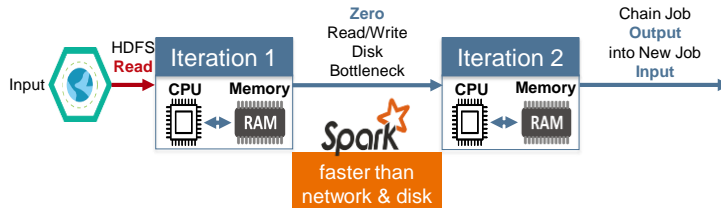
© 2016 IBM Corporation

## 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 more data in-memory with a new distributed execution engine



## Summary





## Summary

- Many new analytical methods are available on LoZ
- Potential to exploit the qualities of the z platform
- It is not always about extraction information from tweets and videos
  - typically it is important to interlink EXISTING data in a reliable, efficient and secure manner

