



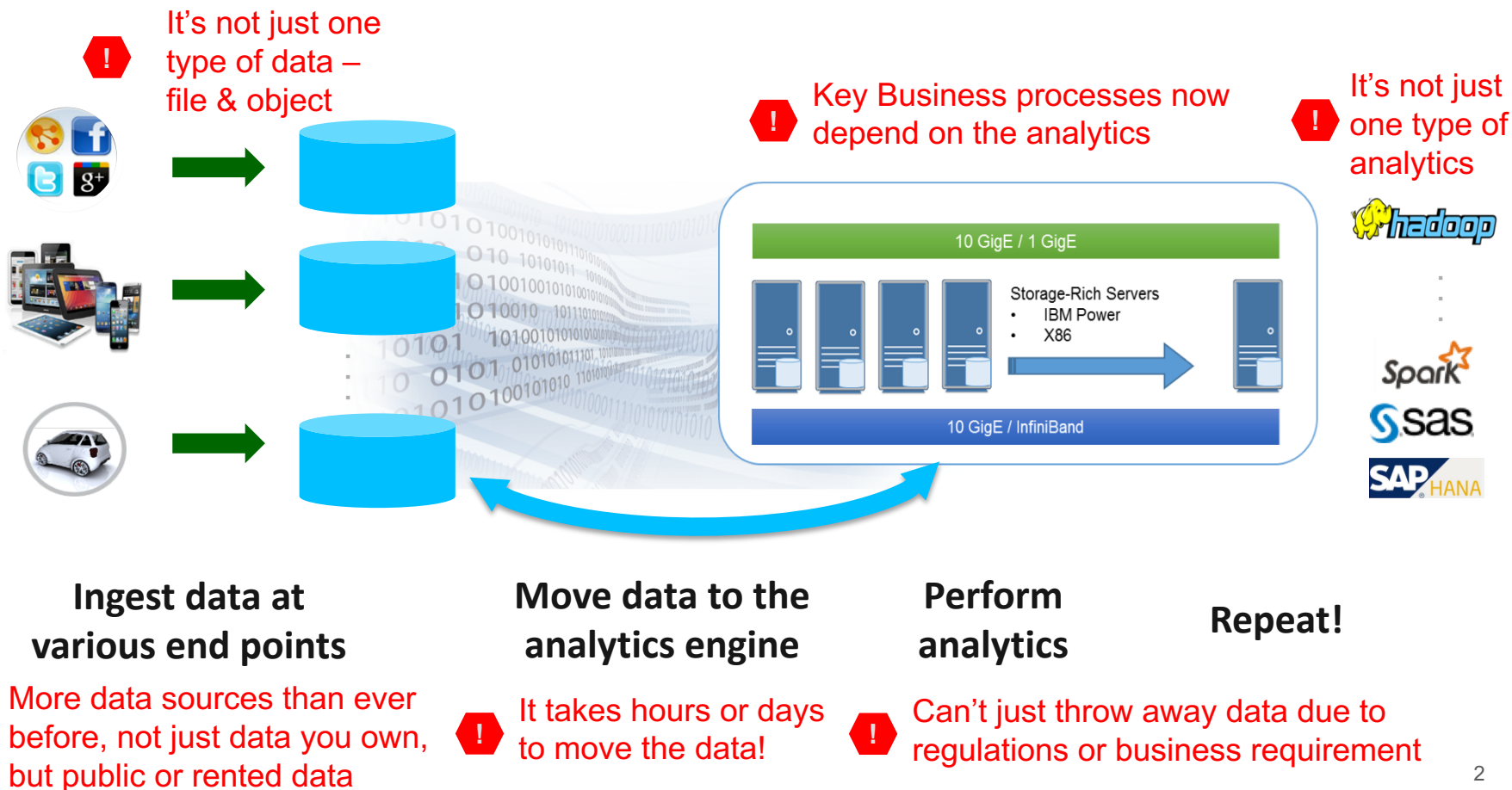
# Hortonworks HDP with IBM Spectrum Scale

*Chih-Feng Ku (cku@hortonworks.com)  
Sr Manager, Solution Engineering APAC*

*Par Hettinga ( par@nl.ibm.com)  
Program Director, Global SDI Enablement*

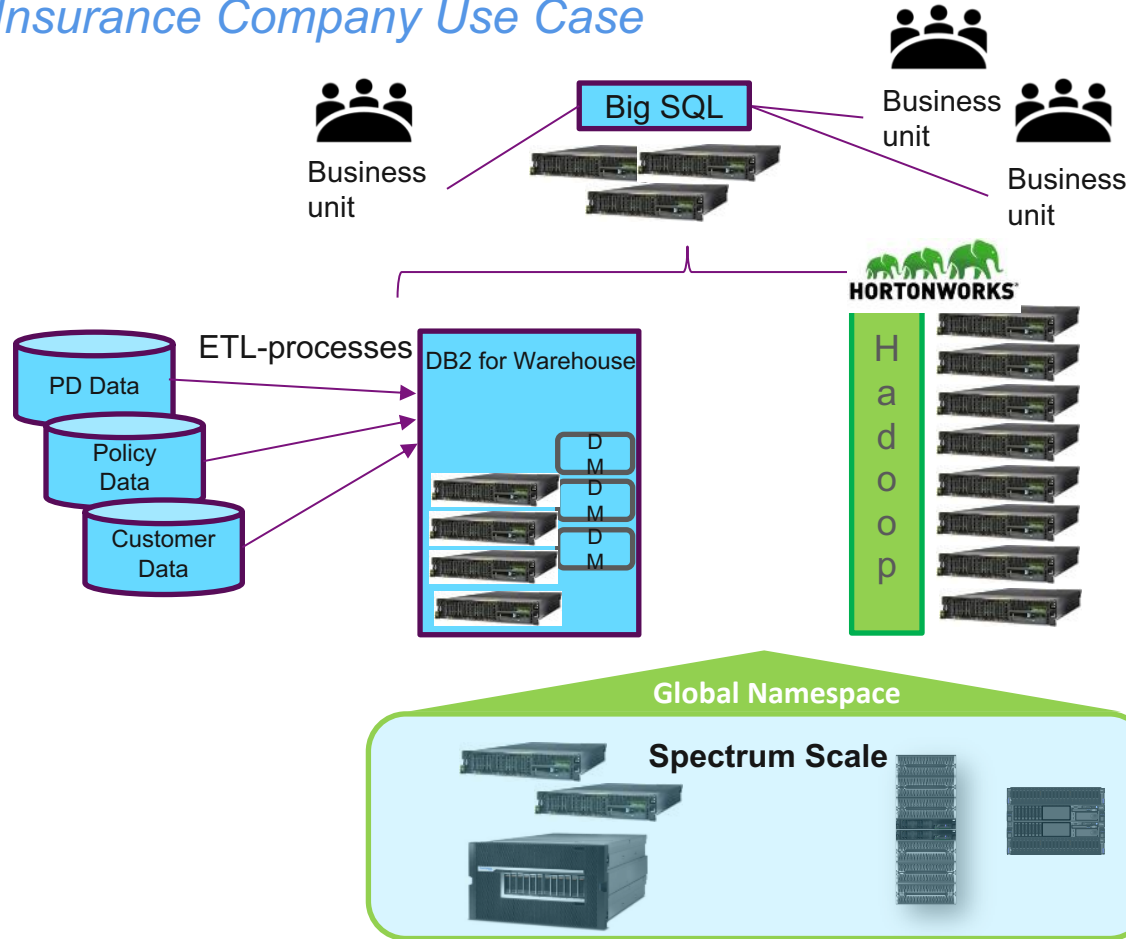


# Challenges with the Big Data Storage Models



# Modernizing and Integrating Data Lake Infrastructure IBM Storage & SDI

## Insurance Company Use Case



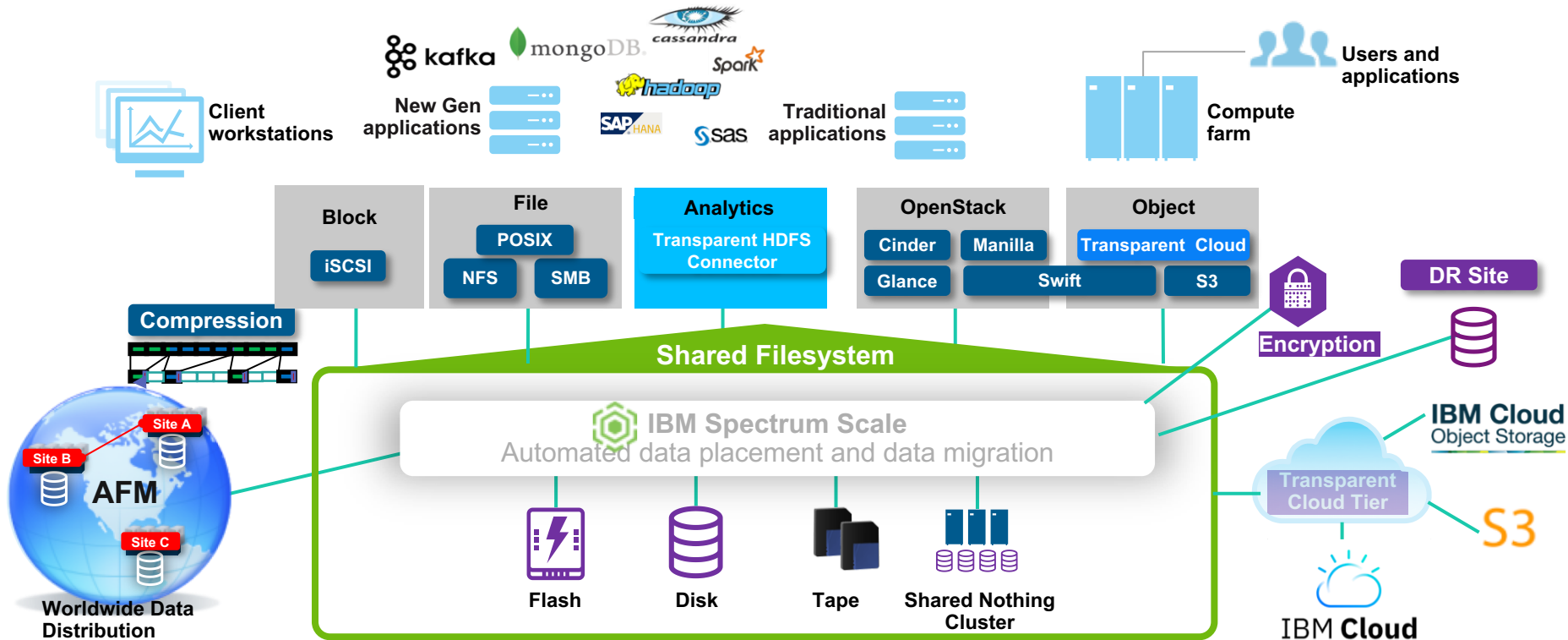
### Value of IBM Power Systems

- Performance and scalability
- High memory, I/O bandwidth
- Optimized server for analytics and integration of accelerators

### Value of IBM Spectrum Scale

- Separation of compute and storagePart
- In-place analytics (Posix conform)
- Integration of objects and Files
- HA / DR solutions
- Storage tiering (inkl. tape integration)
- Flexibility of data movement
- Future: Using of common data formats (f.e. external tables)

# Spectrum Scale: Unleash storage economics on a global scale

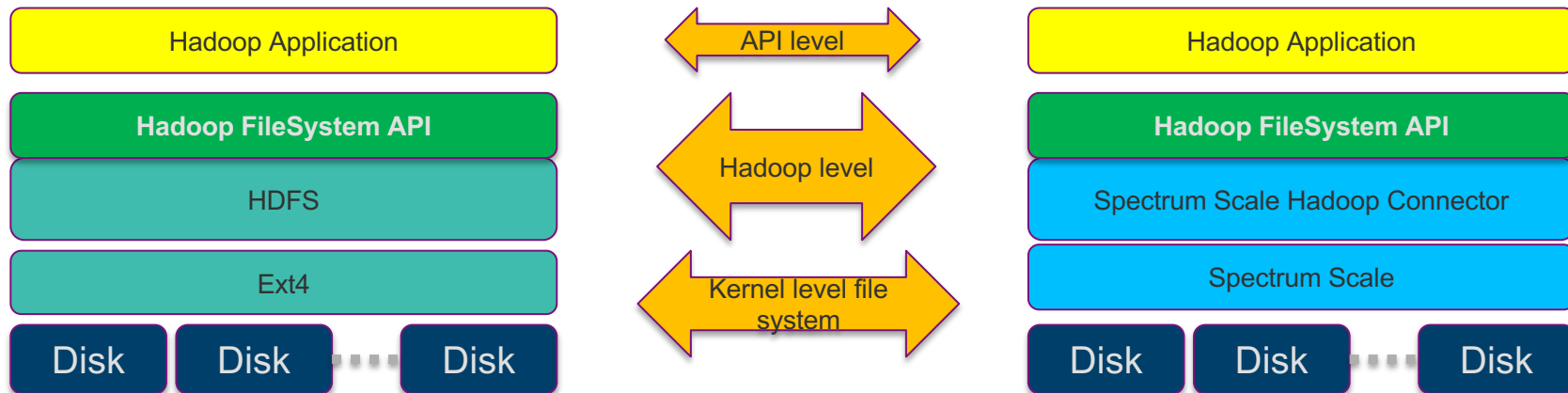


4000+ customers using Spectrum Scale as data plane for HPC and analytics workload

# Spectrum Scale Analytics – Hadoop Connector

IBM Storage & SDI

*Applications communicate with Hadoop using FileSystem API.  
Therefore, transparency is preserved.*

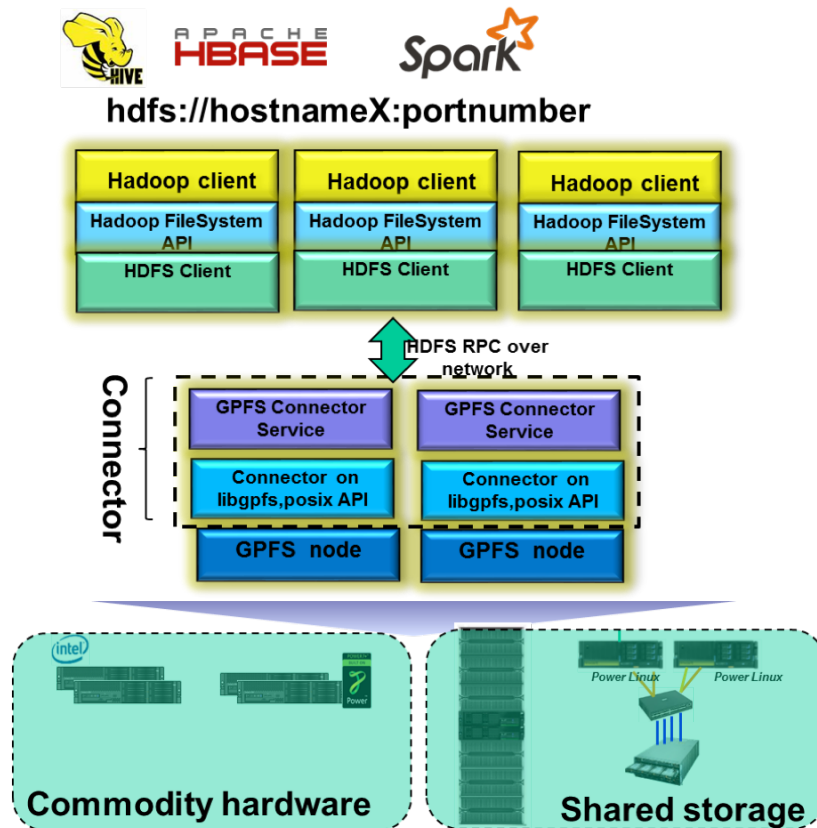


*“All user code that may potentially use the Hadoop Distributed File System should be written to use a **FileSystem** object.”*

**Source:** [hadoop.apache.org](http://hadoop.apache.org)

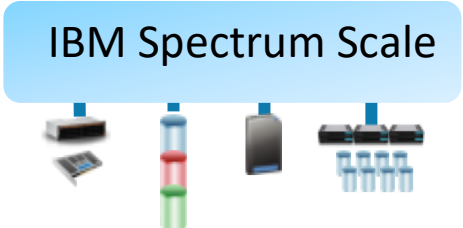
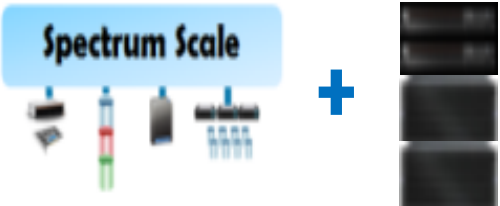

# Spectrum Scale HDFS Transparency Design

- HDFS Transparency connector integrates with HDFS instead of replacing it
  - Out of box support for Hadoop applications and ISVs
  - Easy to maintain compatibility
  - Leverage new open source enhancements
  - Speed up support for new versions
- Re-use HDFS client as-is
  - Faster startup
  - Faster failover
  - Integrate with HDFS centric tools (audit, compliance, etc)
- Stateless NameNode
  - Framework extended for Spectrum Scale Support
    - ✓ *Install Spectrum Scale Cluster*
    - ✓ *Create cluster with FPO configuration*
    - ✓ *Integrated Management of Spectrum Scale*





# Spectrum Scale Deployment model provides Choice

1	 <p>IBM Spectrum Scale</p>	Software Only	Software licenses: Standard or Data Management Edition
2	 <p>Spectrum Scale</p>	IBM Spectrum Scale SW, GUI, GNR, drives, services	IBM Elastic Storage Server GS, GL, GLxS and GSxS(all Flash) Models
3	 <p>IBM Spectrum Scale</p> <p>IBM Cloud</p>	Managed Service	Elastic Storage on Cloud

# Hortonworks HDP certification for Spectrum Scale

IBM Storage & SDI



- Certification is for Spectrum Scale SW and hence applies to all deployment models of Spectrum Scale: Elastic Storage Server(ESS) as well as SW only models.
- Meet in the channel play. No bundling of products.
- Hortonworks Data Flow (HDF - Hortonworks product for data in motion analysis) is certified and supported with this solution.
- Public references: <https://www-03.ibm.com/press/us/en/pressrelease/51562.wss> ,  
<https://hortonworks.com/partner/ibm/> ,  
Cube interview: <https://www.youtube.com/watch?v=kxpUrde99Nk>
- Solution reference guide: <http://www.redbooks.ibm.com/redpapers/pdfs/redp5448.pdf>



# Hortonworks Company Profile

ONLY

**100%**  
open source

Apache Hadoop data platform

Founded in 2011

**1<sup>ST</sup>** HADOOP  
provider to go public

IPO 4Q14 (NASDAQ: HDP)

**140%** Dollar-based  
net expansion rate  
(over trailing 4 quarters)

Support subscription  
Operating billings growth<sup>1</sup> **42%**  
(year-over-year in OCT 2017)

**~1,050**  
employees across

**17**  
countries

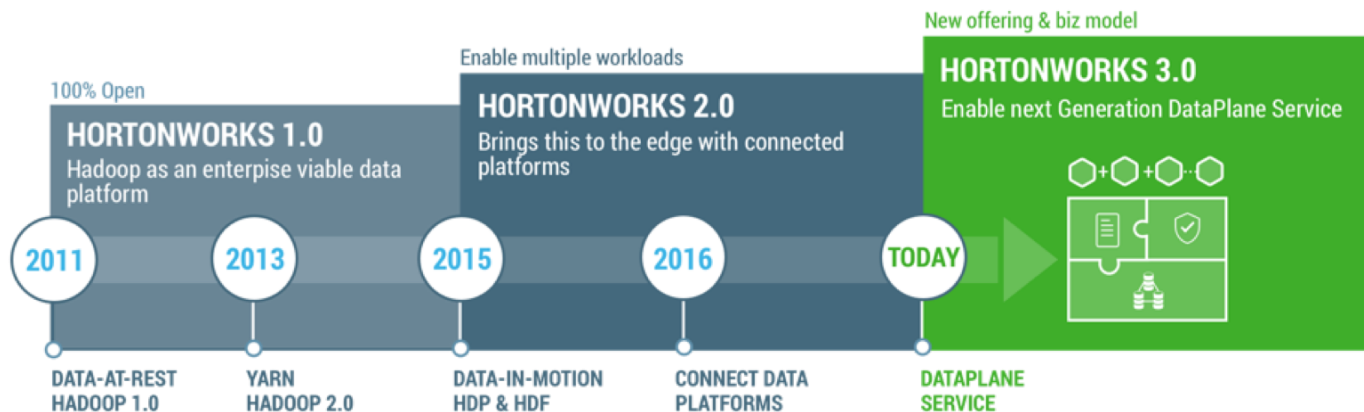
Operating billings is an operating measure defined as the aggregate value of all invoices sent to our customers in a given period

# Hortonworks: Enabling the Modern Data Architecture

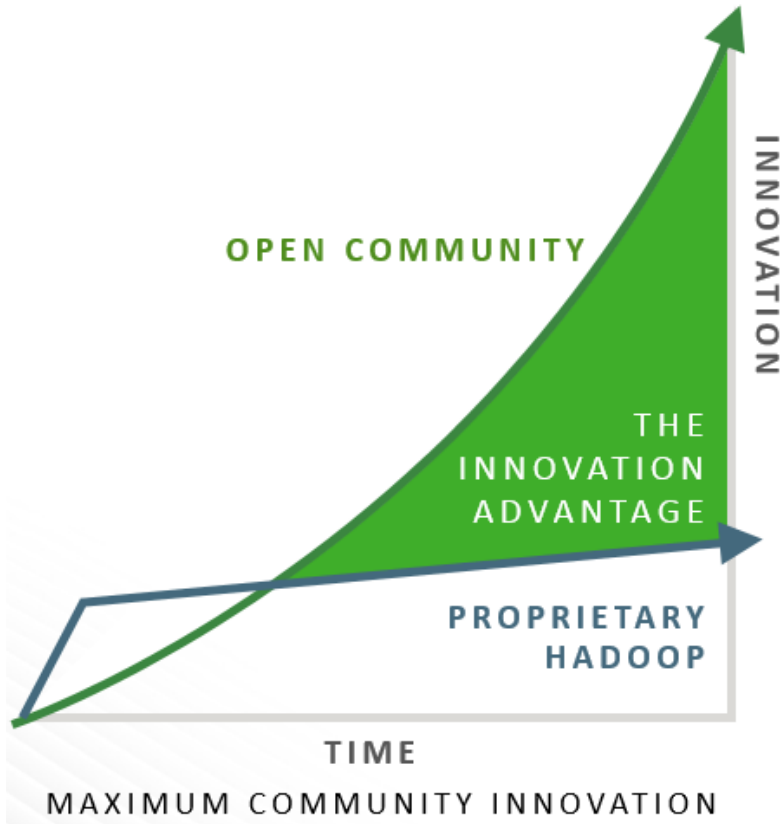
## Our Differentiators

- 100% Open Source driving innovation
- Enterprise Scale with robust security, governance and operation management
- Connected platform with end to end data in motion and data at rest

## Hortonworks consistent and continuous track record of innovation



# 100% Open Source Connected Data Platform



## Eliminates Risk

of vendor lock-in by delivering  
100% Apache open source technology

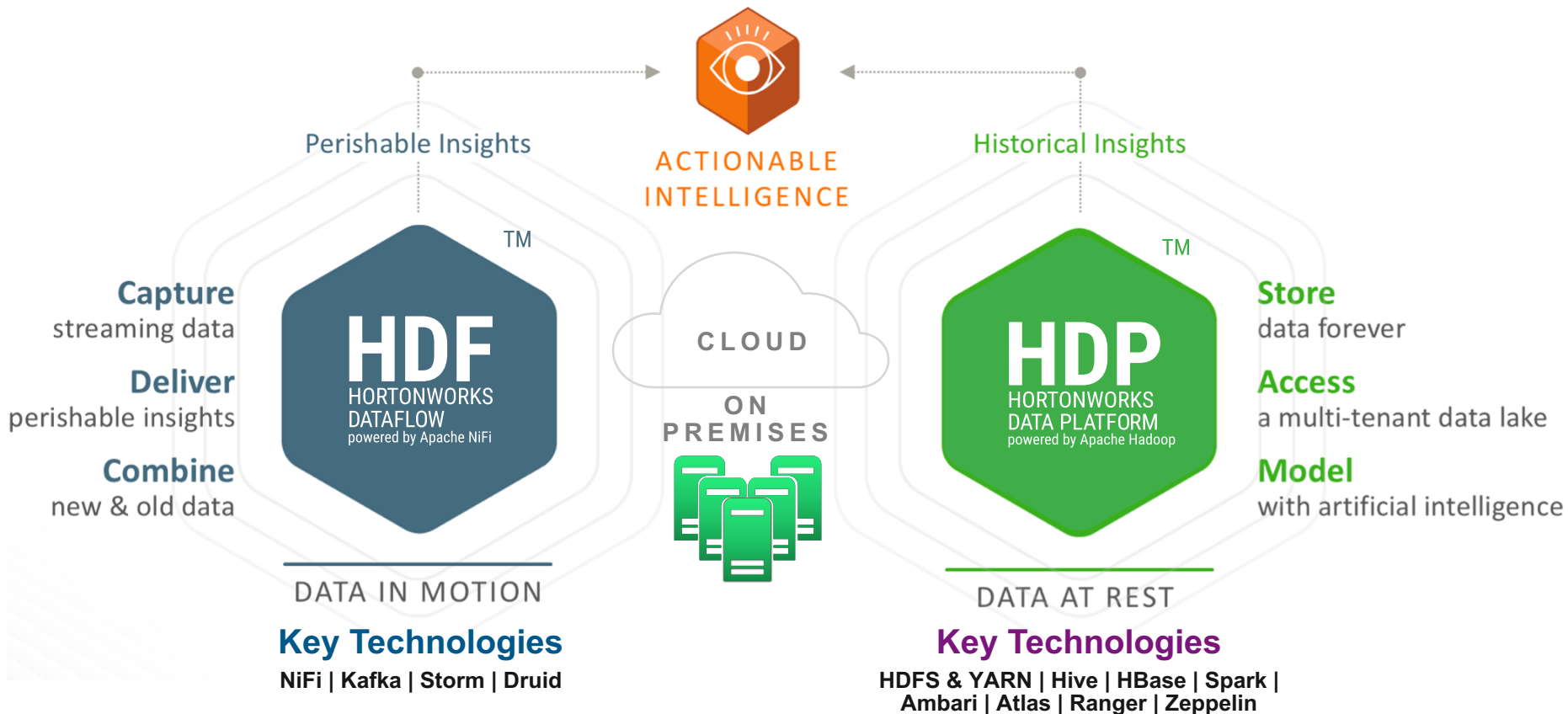
## Maximizes Community Innovation

with hundreds of developers across  
hundreds of companies

## Integrates Seamlessly

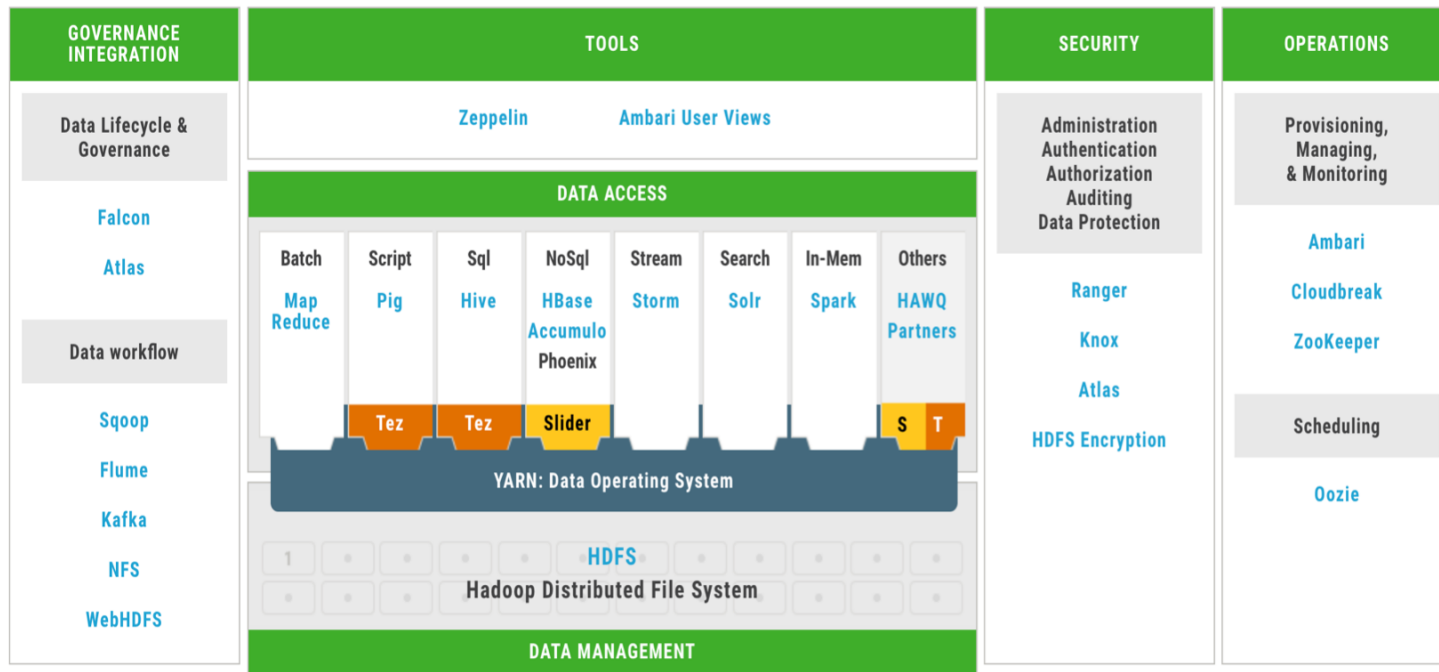
through committed co-engineering  
partnerships with other leading  
technologies

# The Next Generation Data Architecture Solves for All Data



# Hortonworks Data Platform (HDP)

Big data store and processing platform



## Industry's Leading Hadoop Platform

- Centralized Architecture (YARN)
- Enterprise-Ready
- Secure
- 100% Open Source

# Hortonworks Data Flow (HDF) Platform

## HDF 3.1 Data-In-Motion Platform

### Flow Management

Data acquisition and delivery  
Simple transformation and data routing  
Simple event processing  
End to end provenance  
Edge intelligence & bi-directional communication



C++  
Agent

NEW

Java  
Agent

### Stream Processing

Scalable data broker for streaming apps  
Scale out streaming computation engine



### Stream Analytics

Pattern Matching  
Prescriptive & Predictive Stream Analytics  
Complex Event Processing  
Continuous Insights



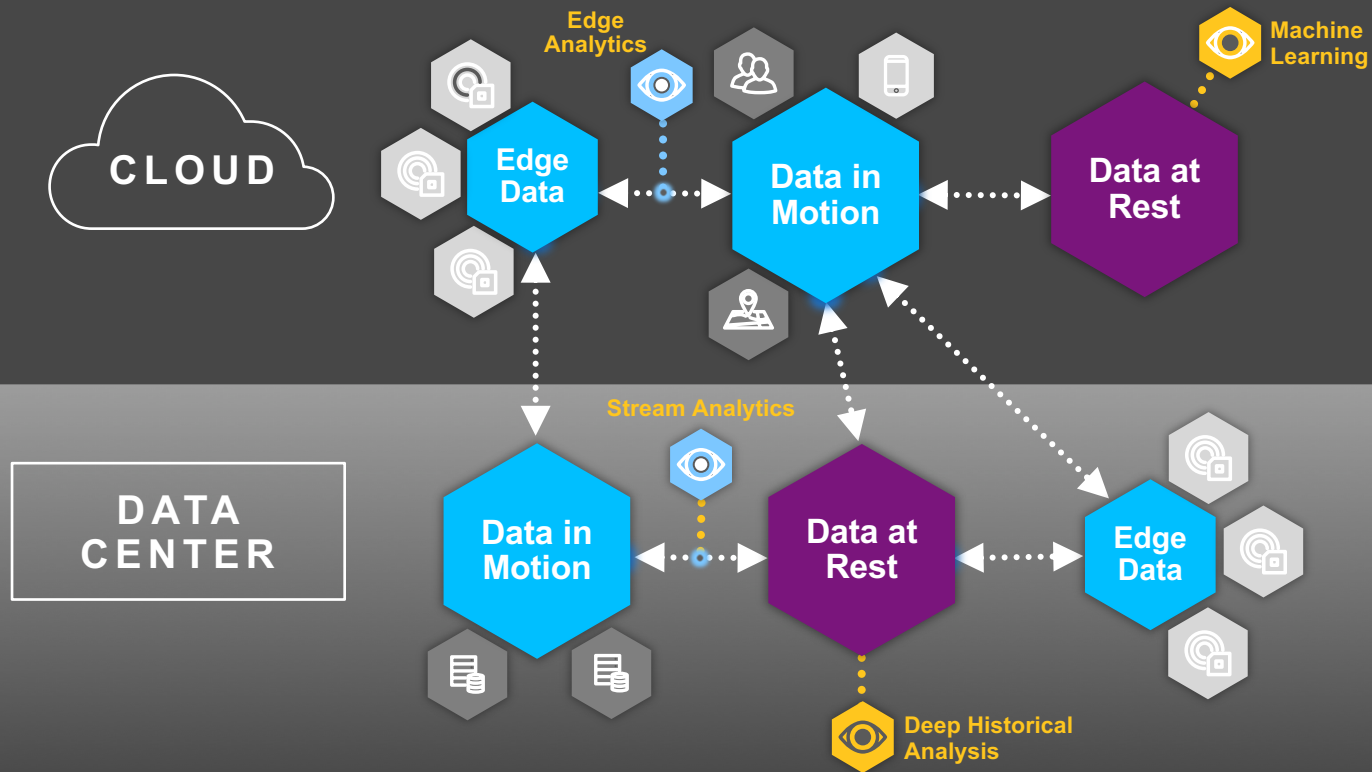
### Enterprise Services

Provisioning, Management, Monitoring, Security,  
Audit, Compliance, Governance, Multi-tenancy



Apache  
Ranger

# Transformational Applications Require Connected Data





EXPLORE

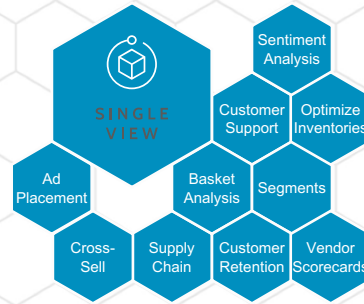
OPTIMIZE

TRANSFORM

INNOVATE



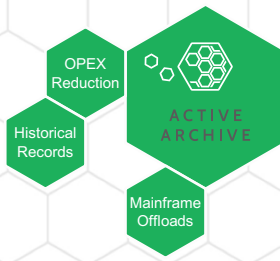
SINGLE VIEW



PREDICTIVE ANALYTICS



RENOVATE



Device Data Ingest

ETL ONBOARD



# DATA JOURNEY

Hortonworks® customers leverage our Connected Data Platforms to transform their industries – renovating their IT architectures and innovating with their Data in Motion or Data at Rest to deliver actionable intelligence through Modern Data Applications.

# Case Study: Persol Japan

Human Resource Services company offering services in temporary staffing, BPO and recruiting.



## Solution

Data lake on HDP on top of Spectrum Scale, where data from multiple source, including Relational database, application, access and search logs, are ingested and stored in HDP. They are building an AI, deep learning and analytics platform on top of this Data Lake using Spark and IBM Power System with GPUs. As a next phase, they will ingest streaming data into the system to build real-time recommendation of career matching for their users.

## Why Spectrum Scale

Spectrum Scale supports multiple protocols including NFS, HDFS and others, which makes injection easy due to multiple tenants with different skill sets. Spectrum scale offers extreme scalability with parallel file system architecture, which supports growth of data in Persol.

# Why Spectrum Scale for Hadoop HDFS

IBM Storage & SDI

*Unmatched Scalability with up to 70% less Storage Consumption*

- Reduce the datacenter footprint and cost
  - In-place analytics** for Hadoop data eliminating need for multiple copies of same data or large migrations of data between HDFS and the POSIX file system.
- Control cluster sprawl
  - Delivers **federation across clusters**, both Scale and HDFS while still maintaining needed separations
- Make HDFS access enterprise-ready
  - Adds **storage functions necessary to the enterprise** (e.g. Encryption, DR, SW RAID) to your Hadoop setup
- True Software Defined purchased as Software ONLY, ESS or on the Cloud
- Extreme Scalability with parallel file system architecture
- Global Name Space that can span geographies
  - Active - Active replica's of data for **real time global collaboration**
- Provides comprehensive Data Information Lifecycle Management
  - Automated data **placement** and data **migration**

## Scalability

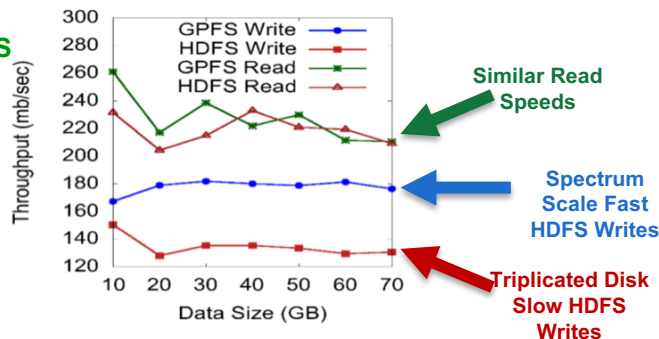
HDFS can scale up to **350 Million** files with a single name node due to scale-out architecture limitation. Name node becomes a bottleneck. Users have to use federation functionality to overcome this limitation.

Spectrum Scale has parallel file system architecture different from scale-out architecture of HDFS. No single metadata server in the architecture as a bottleneck. Metadata serving function is distributed across the cluster. Test limit for number of files per filesystem is **9 Billion**. We have Spectrum Scale customers running in production beyond this test limit.

## Performance

Performance depends on the underlying hardware configurations. But we claim comparable or better performance than HDFS on all equivalent hardware configurations. Here is an acm [publication](https://w3-connections.ibm.com/wikis/form/api/wiki/4cbe4ccf-e274-4830-8370-3ceb5d56bd06/page/71236e6b-a235-4a63-8827-a92a0c36d466/attachment/69835e1b-8a1f-4347-bf76-ecc0cdb8d7c1/media/Hadoop-SpectrumScale.pdf) that compares standard benchmark results for HDFS vs Spectrum Scale. <https://w3-connections.ibm.com/wikis/form/api/wiki/4cbe4ccf-e274-4830-8370-3ceb5d56bd06/page/71236e6b-a235-4a63-8827-a92a0c36d466/attachment/69835e1b-8a1f-4347-bf76-ecc0cdb8d7c1/media/Hadoop-SpectrumScale.pdf>

### Benchmarks for Spectrum Scale with HDFS IBM TJ Watson Research Center



### Results :

- ❖ Similar Spectrum Scale-HDFS read performance
- ❖ Faster Spectrum Scale-HDFS write performance

# Hortonworks HDP + Spectrum Scale WINS

## **American automaker is using HDP + Scale to store IoT data & perform Hadoop analytics**

Key value points were:

- In-place analytics to ingest data using POSIX
- Perform analytics directly on the data using HDFS APIs.

## **Japanese recruitment and HR company using Scale and HDP for SPARK, Analytics and AI**

Key value points were:

- Personnel data security
- Performance

## **IBM Spectrum Scale being used in Hadoop environments (non-Hortonworks)**

- NASA : <http://files.gpfsug.org/presentations/2016/SC16/06 - Carrie Spear - Spectrum Scale and HDFS.pdf>
- CRS4 : <https://wiki.apache.org/hadoop/PoweredBy>

# IBM Resources and Contacts

Name	Title	Contact Details
Sean Xiang	AP Senior Technical Architect	<a href="mailto:zhanx@sg.ibm.com">zhanx@sg.ibm.com</a>
Somas Balasubramanian	Chief Architect, Systems Storage	<a href="mailto:SOMAS@ae.ibm.com">SOMAS@ae.ibm.com</a>
JD Zeeman	Global Business Development Executive	<a href="mailto:jdzeeman@us.ibm.com">jdzeeman@us.ibm.com</a>
Kent Koeninger	Global Business Development Executive	<a href="mailto:rkkoenin@us.ibm.com">rkkoenin@us.ibm.com</a>
Chris Maestas	Global Senior Solution Architect	<a href="mailto:cdmaestas@us.ibm.com">cdmaestas@us.ibm.com</a>
Par Hettinga	Global SDI Enablement Leader	<a href="mailto:par@nl.ibm.com">par@nl.ibm.com</a>
Pallavi Galgali	Global Spectrum Storage Offering Manager	<a href="mailto:pgalgali@us.ibm.com">pgalgali@us.ibm.com</a>
Piyush Chaudhary	Senior Technical Staff Member	<a href="mailto:piyushc@us.ibm.com">piyushc@us.ibm.com</a>
Gary Tomchuck	Global Sales Executive Cognitive Systems	<a href="mailto:gtomchu@us.ibm.com">gtomchu@us.ibm.com</a>

- Hortonworks Resources
  - Hortonworks IBM Website - <https://ibm.biz/BdiUUa>
  - Hortonworks Youtube Channel - <https://www.youtube.com/user/Hortonworks/featured>
- Big SQL Resources
  - Big SQL Web Page - <https://ibm.biz/Bdi99F>
- Data Science Resources
  - Data Science Webpage - <https://ibm.biz/BdiU6N>
  - Data Science Cloud - <https://datascience.ibm.com/>
  - Data Science Download and Go - <https://datascience.ibm.com/local>
- Power Systems Resources
  - Hortonworks on IBM Power Webpage - <https://www.ibm.com/power/solutions/modern-data-platform-hortonworks>
- Spectrum Scale Resources
  - Spectrum Scale Webpage - <https://www.ibm.com/us-en/marketplace/scale-out-file-and-object-storage>
  - Spectrum Scale with Hortonworks - <https://hortonworks.com/partner/ibm/#spectrumscale>



**Thank You.**  
**IBM Storage & SDI**

A series of thick, blue diagonal stripes of varying lengths and orientations, creating a dynamic, abstract pattern in the bottom right corner of the slide.

# Legal notices

Copyright © 2016 by International Business Machines Corporation. All rights reserved.

No part of this document may be reproduced or transmitted in any form without written permission from IBM Corporation.

Product data has been reviewed for accuracy as of the date of initial publication. Product data is subject to change without notice. This document could include technical inaccuracies or typographical errors. IBM may make improvements and/or changes in the product(s) and/or program(s) described herein at any time without notice. Any statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business. Any reference to an IBM Program Product in this document is not intended to state or imply that only that program product may be used. Any functionally equivalent program, that does not infringe IBM's intellectually property rights, may be used instead.

THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER OR IMPLIED. IBM LY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NONINFRINGEMENT. IBM shall have no responsibility to update this information. IBM products are warranted, if at all, according to the terms and conditions of the agreements (e.g., IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided. Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products in connection with this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. IBM makes no representations or warranties, ed or implied, regarding non-IBM products and services.

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents or copyrights. Inquiries regarding patent or copyright licenses should be made, in writing, to:

IBM Director of Licensing  
IBM Corporation  
North Castle Drive  
Armonk, NY 1 0504- 785  
U.S.A.

# Information and trademarks

IBM, the IBM logo, ibm.com, IBM System Storage, IBM Spectrum Storage, IBM Spectrum Control, IBM Spectrum Protect, IBM Spectrum Archive, IBM Spectrum Virtualize, IBM Spectrum Scale, IBM Spectrum Accelerate, Softlayer, and XIV are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at <http://www.ibm.com/legal/copytrade.shtml>

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.

IT Infrastructure Library is a Registered Trade Mark of AXELOS Limited.

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 other countries.

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.

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.

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

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

ITIL is a Registered Trade Mark of AXELOS Limited.

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

\* All other products may be trademarks or registered trademarks of their respective 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.

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 presentation and the claims outlined in it were reviewed for compliance with US law. Adaptations of these claims for use in other geographies must be reviewed by the local country counsel for compliance with local laws.

# Special notices

This document was developed for IBM offerings in the United States as of the date of publication. IBM may not make these offerings available in other countries, and the information is subject to change without notice. Consult your local IBM business contact for information on the IBM offerings available in your area.

Information in this document concerning non-IBM products was obtained from the suppliers of these products or other public sources. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. Send license inquiries, in writing, to IBM Director of Licensing, IBM Corporation, New Castle Drive, Armonk, NY 10504-1785 USA.

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

The information contained in this document has not been submitted to any formal IBM test and is provided "AS IS" with no warranties or guarantees either expressed or implied.

All examples cited or described in this document are presented as illustrations of the manner in which some IBM products can be used and the results that may be achieved. Actual environmental costs and performance characteristics will vary depending on individual client configurations and conditions.

IBM Global Financing offerings are provided through IBM Credit Corporation in the United States and other IBM subsidiaries and divisions worldwide to qualified commercial and government clients. Rates are based on a client's credit rating, financing terms, offering type, equipment type and options, and may vary by country. Other restrictions may apply. Rates and offerings are subject to change, extension or withdrawal without notice.

IBM is not responsible for printing errors in this document that result in pricing or information inaccuracies.

All prices shown are IBM's United States suggested list prices and are subject to change without notice; reseller prices may vary.

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

Any performance data contained in this document was determined in a controlled environment. Actual results may vary significantly and are dependent on many factors including system hardware configuration and software design and configuration. Some measurements quoted in this document may have been made on development-level systems. There is no guarantee these measurements will be the same on generally-available systems. Some measurements quoted in this document may have been estimated through extrapolation. Users of this document should verify the applicable data for their specific environment.