

CA Big Data Management: It's here, but what can it do for your business?

Mike Harer
CA Technologies

August 7, 2014
Session Number: 16256

Test link: www.SHARE.org

Insert
Custom
Session
QR if
Desired.

#SHAREorg



Big Data Management – Topics

- 1 Big Data Background
- 2 Big Data Management Challenges
- 3 Our Vision
- 4 Big Data Management Solution Overview
- 5 Target Personas
- 6 Solving the Important Management Pains

Big Data Means Different Things To Different People

- Customers tend to define Big Data in a broad sense
 - Any type of net new analytical processing that is different from the traditional data warehouse applications in place today
 - This is differentiated by (near) real time analytical capabilities
 - Defined by the types and speed of data being analyzed
-



"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

Big Data Definition

Large **Volumes** of a wide **Variety** of data collected from various sources across the enterprise

High-**Velocity** capture, discovery and/or analysis

Veracity – keeping the right, trusted data



Big Data – Growing Fast

CAPTURING AND MANAGING LOTS OF INFORMATION

- Data volumes are doubling every year
- Organizations are also storing three or more years of data

WORKING WITH MANY NEW TYPES OF DATA

- 80 percent of data is unstructured (such as images, audio, tweets, etc.)

EXPLOITING THESE MASSES OF INFORMATION AND NEW DATA TYPES WITH NEW STYLES OF APPLICATIONS

- New classes of analytic applications are reaching the market, all based on a next generation big data platform

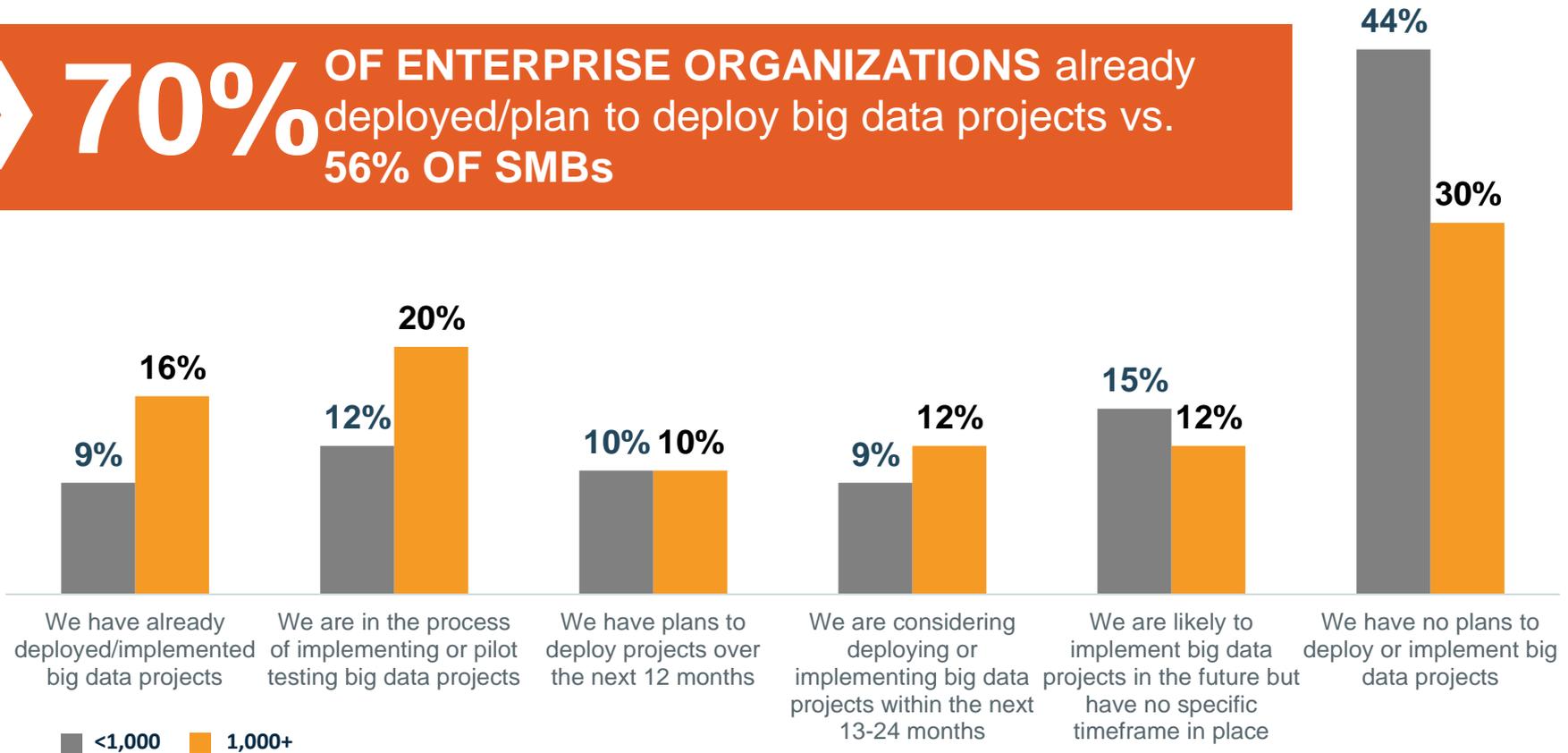


COMMODITIZED HARDWARE AND SOFTWARE

- The final piece of the big data puzzle is the low-cost hardware and software environments
- Capturing and exploiting big data would be much more difficult and costly without the contributions of these cost-effective advances

Enterprises Ahead in Big Data Initiatives

> 70% OF ENTERPRISE ORGANIZATIONS already deployed/plan to deploy big data projects vs. **56%** OF SMBs



Q. Is your company currently implementing, planning or considering projects (i.e. devising strategies and projects to generate more value from existing data)?

Going from the Science Project to Production

- The organization realizes that the analytics and insights coming out of a Big Data project are essential
- To keep costs down, you start with the basic Hadoop distribution from Apache
- Maybe a free tool or two and off you go
- Gain traction, under a huge amount of pressure to deliver or the business gets farther behind
- More tools and software and data sources are added

Hadoop Common – Common utilities
HDFS – Hadoop distributed file system
MapReduce – Parallel processing
Pig – High-level language for MapReduce
Streaming – Jobs based on any exe
Hive – Data warehouse
HBase – Non-relational dbase on HDFS
Nagios – Open Source Monitoring Tool
Ganglia – Open Source Monitoring Tool
Oozie – Workflow Engine / Scheduler
Cloudera Manager – Hadoop Management
Ambari – Hadoop Management Console
HCatalog, ZooKeeper, Sqoop...

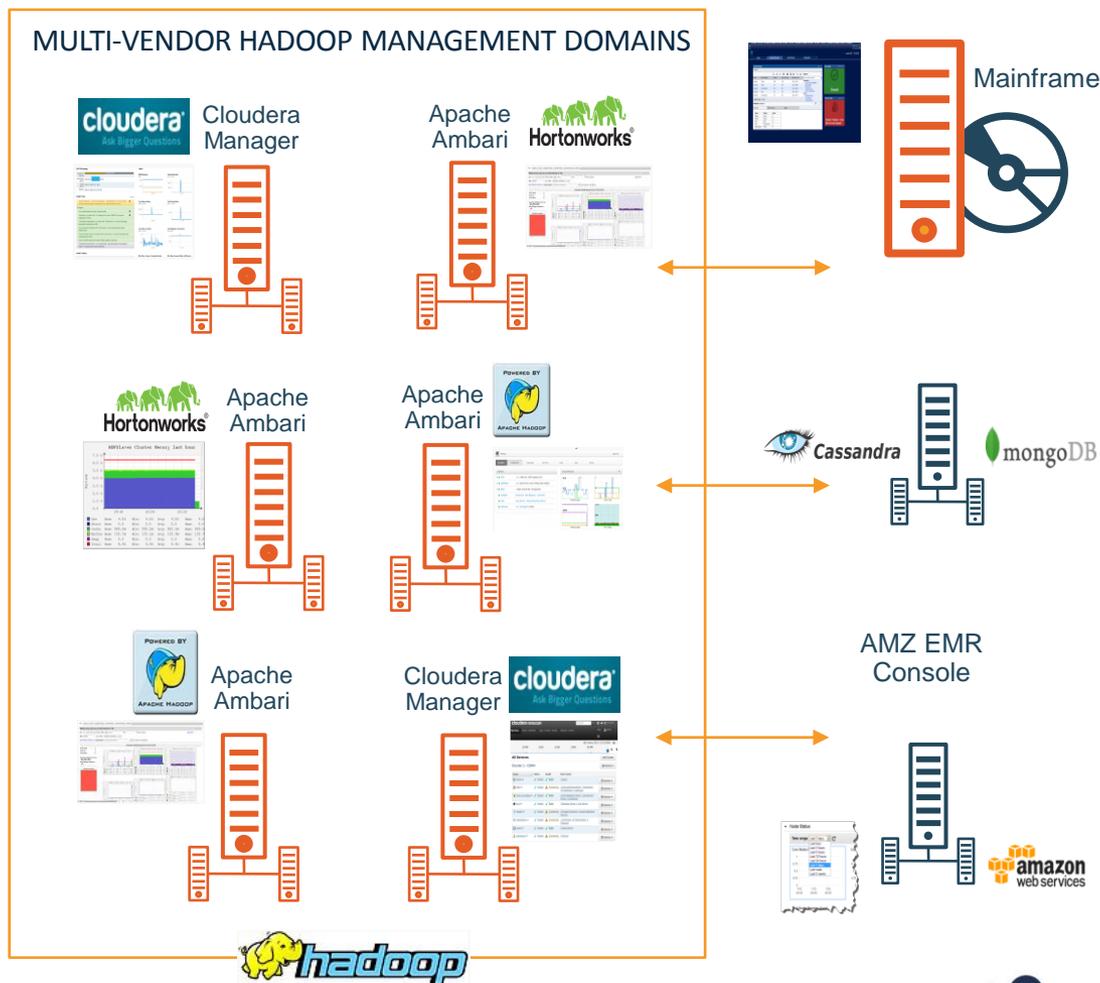


By the time you are ready to put this to productive use, you realize you have a huge number of moving parts, tools from many vendors and a ton of complexity has been created.

The “Big” Big Data Management Pains

The Need to Overcome Many Challenges

- Managing complex multi-vendor big data environments
- Finding Hadoop/Big Data experts
- Understanding capacity requirements for rapidly changing business needs
- Complexity increases, manual processes often required
- System problems hard to isolate, downtime increases
- Unique tools and shortcomings
- Multiple licensing agreements to manage
- Driving forces... acquisitions, department consolidations, driving need for greater operational efficiency





Ask Yourself...

- 1 How many people do you have to manage your Big Data infrastructure?
- 2 Do your Big Data administrators always know the health of the systems?
- 3 Can you detect most problems before significant system outages occur?
- 4 How many different monitoring tools do you have in place now?
- 5 How do you know if your capacity is optimized for cost and performance?
- 6 What was the financial impact of downtime over the past year?

CA Big Data Management

Our Vision

PROVIDE CUSTOMERS AN INNOVATIVE APPROACH TO MORE EFFICIENTLY MANAGE HETEROGENEOUS BIG DATA ENVIRONMENTS TO MAXIMIZE ENTERPRISE OPERATIONAL EFFICIENCY.

Traditional data management is changing rapidly. Evolving big data technologies offer a range of new business advantages when managing disparate environments that can be a mix of traditional, mainframe and distributed big data environments.

CA Technologies is the market leader in delivering world-class IT Management solutions. Delivering a world-class Big Data Management solution is a logical extension to our IT management portfolio and will continue to demonstrate our leadership and innovation in the Big Data market.

CRN 25 Big Data Infrastructure Companies

2
 3
 4

◀ Previous post Next post ▶

We examine the top 25 Big Data Infrastructure companies, part of CRN Big Data 100, which includes Amazon, IBM, and Microsoft.

By Grant Marshall, June 2014.

The CRN 25 Big Data Infrastructure companies includes notable companies who provide the tools for other companies to handle and analyze big data.

One interesting observation about the companies in this list compared to the companies in the rest of the Big Data 100, is that the average age of a big data infrastructure company (20) is older compared to the data management companies (8) and business intelligence companies (10).

This is also the only category containing companies founded before 1950 (IBM and HP). This would seem to imply that big data infrastructure rewards companies with more extensive experience in the field.

Here are the CRN 25 Big Data Infrastructure companies:

- Altiscale, claims its Data Cloud is the first cloud service purpose-built to run Hadoop, offering an on-demand, pay-as-you-go service based on the big data platform. Palo Alto, CA. Founded 2012.
- Amazon Web Services, includes Amazon DynamoDB NoSQL database service, Amazon Kinesis managed service for real-time processing and analysis of streaming big data, Amazon Redshift petabyte-scale data warehouse, Amazon Glacier for archival big data storage, and Amazon Elastic MapReduce providing the Hadoop framework through Amazon's Elastic Compute Cloud (EC2) service. Seattle, WA. Founded 1994.
- CA Technologies, provides a number of IT system capacity management tools to help businesses and service providers better predict and plan the IT resources needed to handle big data demands. Islandia, NY. Founded 1976.

CA Big Data Management – Solution Overview

(Release 1.0 -- CA World Launch Nov/2014)

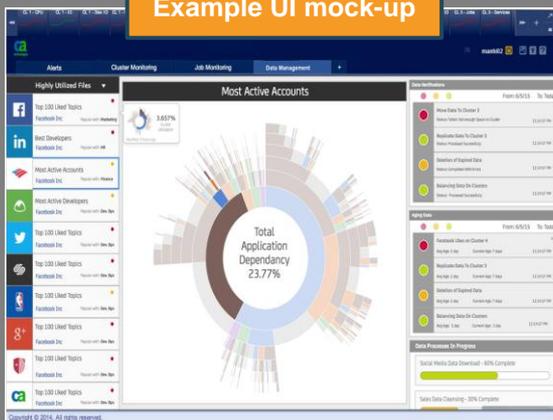
MANAGE ALL BIG DATA ENVIRONMENTS FROM ONE PLACE

SECURE ACCESS TO BIG DATA (HADOOP) ENVIRONMENTS

OPERATIONALIZE, MANAGE & SECURE HADOOP ENVIRONMENTS

SINGLE, CONSISTENT MANAGEMENT UI EXPERIENCE

Example UI mock-up



SIMPLIFIED HETEROGENEOUS ENVIRONMENT MANAGEMENT

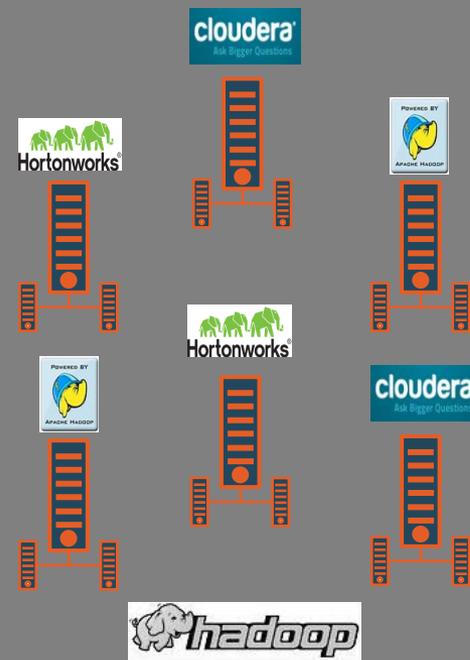
Big Data Management Server



Linux / x86



MULTI-VENDOR HADOOP MANAGEMENT DOMAINS



A New Role in the Organization is Born

Targeted User Personas

PERSONAS:

- **Big Data / Hadoop Administrator (Rel 1.0)**
Very different than the traditional DBA...
 - Big Data environment is much more complex than traditional database application even compared to large ERP systems
 - Many more moving parts
 - Volume, Velocity, Variety
- **Big Data / Hadoop Developer (Rel 1.x -2.0)**



Role / Responsibilities:

- Day-to-day operations & support of Hadoop infrastructure platforms
- Monitor/maintain existing clusters and provision new ones
- Integrate enterprise monitoring tools like Ganglia and OpenTSDB
- Analyze current workloads & enable subsequent capacity planning.
- Publish various production metrics to system owners & mgmt.
- Utilize enterprise automation tools as well as configuration management tools e.g. Puppet, Chef and/or Cloudera Mgr
- Perform regular back ups for replication as well as disaster recovery
- Educate existing SysOps team members on the Hadoop ecosystem

Solving the Important Management Pains

CA Big Data Management

Applying best practices to...

- Quickly and easily deploy new clusters
- Receive alerts before system degradation causes significant cluster/job failures (w/ all details)
- Monitor system capacity metrics and suggest corrective action(s)
- Optimize job performance based on watching heap size and auto-adjusting thru automation
- Detect misconfigured settings and send alert with suggested actions to remedy
- Schedule/monitor jobs thru a **single, consistent UI experience**

Hadoop Administrator Challenges



“We get alerts from zabbix saying workflow is slow or job failed. Some workflows automatically try and restart after the failure but if it doesn't start, we manually start it.”

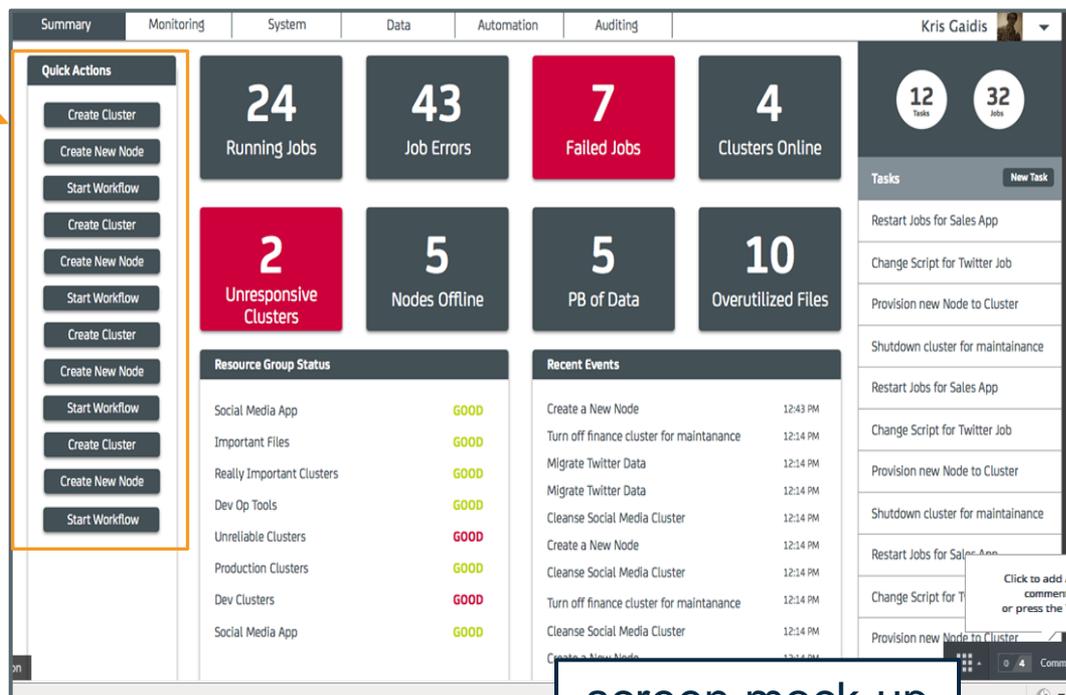
“We use one Cloudera Manager instance for every cluster so that we do not have a single point of failure. Only admins can login to CM, dev team use native Apache Hadoop UI”

Solving the Important Management Pains

Example Use Case

Heterogeneous System Management

- Ability to control system operations across all clusters in a heterogeneous big data environment from a **single, consistent UI experience**
- Quick access to system management operations
- Based on Best Practices, eliminates manual processes and decreases system errors/downtime



screen mock-up

Solving the Important Management Pains

Example Use Case

Resource Reporting

- Aggregated data enables Hadoop administrators to more easily identify the cause of capacity, performance and system disruptions.
 - Underperforming clusters
 - Job execution slowed or incomplete
- Visual graphing of time-series data coming from system metrics (CPU, Network IO, disk space, Map/Reduce slots, memory, swap space)
- Generate Historical Job Reports (over hours, days, weeks, months)

screen mock-up



Solving the Important Management Pains

Example Use Case

Activity (Job) Monitoring

- Assess integrity of all Hadoop jobs (running, killed, suspended, ...)
- Visually identify performance issues
- Unified view across all clusters, nodes within a mixed multi-vendor big data environment (e.g. Cloudera, Hortonworks, Apache)

screen mock-up

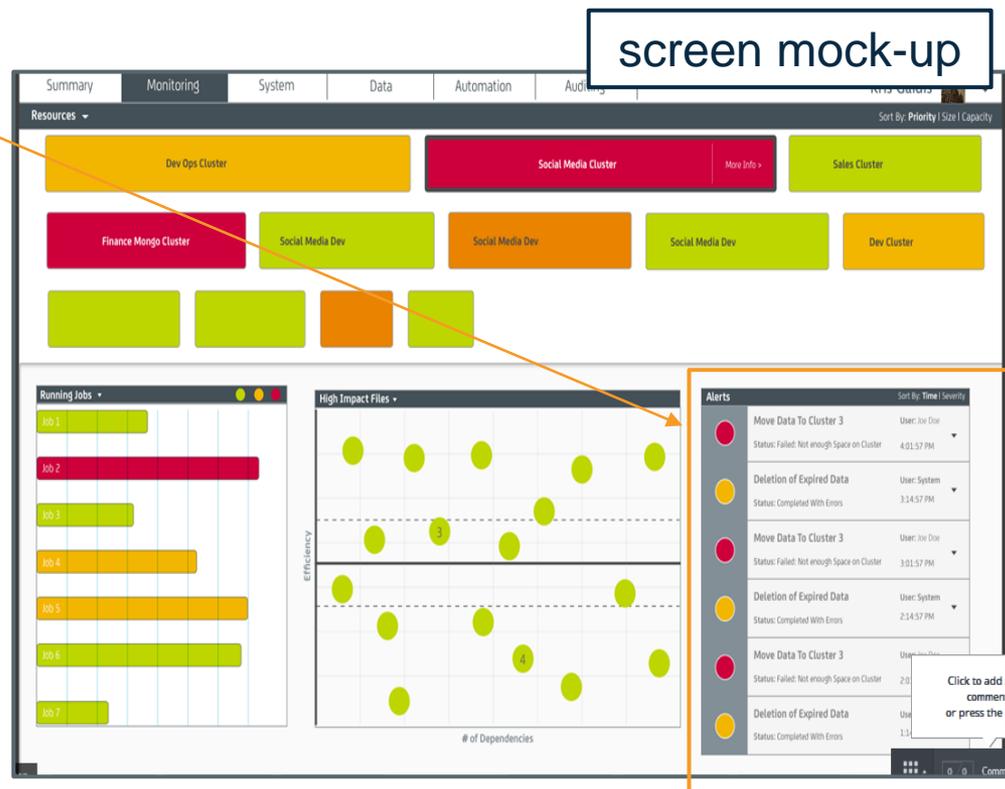


Solving the Important Management Pains

Example Use Case

Alert Management

- Tightly integrated with System Monitoring to send alerts when configured conditions are detected.
- Integrated with an Automation Framework
- Can be configured to trigger alerts based on absolute values and percentages. For example, generate alert:
 - based on metrics threshold violations
 - based on complex rules, like Event Frequency or patterns of Events
 - based on a Severity



Q & A



Mike Harer
Sr. Principal Product Manager
100 Staples Drive
Framingham, MA 01702
Office: +1-508-598-6547
Mobile: +1-978-424-5677
Michael.harer@ca.com

www.ca.com/bigdata

Disclaimer

Copyright © 2014 CA. All rights reserved. IBM, System z, zEnterprise and z/OS are trademarks of International Business Machines Corporation in the United States, other countries, or both. All trademarks, trade names, service marks and logos referenced herein belong to their respective companies.

This presentation was based on current information and resource allocations as of July 2014 and is subject to change or withdrawal by CA at any time without notice. Notwithstanding anything in this presentation to the contrary, this presentation shall not serve to (i) affect the rights and/or obligations of CA or its licensees under any existing or future written license agreement or services agreement relating to any CA software product; or (ii) amend any product documentation or specifications for any CA software product.

The development, release and timing of any features or functionality described in this presentation remain at CA's sole discretion. Notwithstanding anything in this presentation to the contrary, upon the general availability of any future CA product release referenced in this presentation, CA will make such release available (i) for sale to new licensees of such product; and (ii) to existing licensees of such product on a when and if-available basis as part of CA maintenance and support, and in the form of a regularly scheduled major product release. Such releases may be made available to current licensees of such product who are current subscribers to CA maintenance and support on a when and if-available basis. In the event of a conflict between the terms of this paragraph and any other information contained in this presentation, the terms of this paragraph shall govern.

Certain information in this presentation may outline CA's general product direction. All information in this presentation is for your informational purposes only and may not be incorporated into any contract. CA assumes no responsibility for the accuracy or completeness of the information. To the extent permitted by applicable law, CA provides this presentation "as is" without warranty of any kind, including without limitation, any implied warranties or merchantability, fitness for a particular purpose, or non-infringement. In no event will CA be liable for any loss or damage, direct or indirect, from the use of this document, including, without limitation, lost profits, lost investment, business interruption, goodwill, or lost data, even if CA is expressly advised in advance of the possibility of such damages. CA confidential and proprietary. No unauthorized copying or distribution permitted.