Course Guide Volume 1

IBM BigInsights Foundation v4.0

Course code DW613 ERC 1.0

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

IBM Training

June 2015 edition

NOTICES

This information was developed for products and services offered in the USA.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service. IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
United States of America

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you. 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 and cannot confirm the accuracy of 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.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

TRADEMARKS

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at www.ibm.com/legal/copytrade.shtml.

Adobe, and the Adobe logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

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

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

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

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

© Copyright International Business Machines Corporation 2015.

This document may not be reproduced in whole or in part without the prior written permission of IBM.

US Government Users Restricted Rights - Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Contents

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Course overview

Preface overview

This course is for those who want a foundation of IBM BigInsights. In the IBM BigInsights Overview part of this course, you will have an overview of IBM's big data strategy and review why it is important to understand and use big data. It will cover IBM BigInsights as a platform for managing and gaining insights from your big data. As such, you will see how the BigInsights have aligned their offerings to better suit your needs with the IBM Open Platform (IOP) along with the three specialized modules with value-add that sits on top of the IOP. You will also get an introduction to the BigInsights value-add including Big SQL, BigSheets, and Big R. In the IBM Open Platform with Apache Hadoop part of the course, you will review how IBM Open Platform (IOP) with Apache Hadoop is the collaborative platform to enable Big Data solutions to be developed on the common set of Apache Hadoop technologies. You will also have an in-depth introduction to the main components of the ODP core, namely Apache Hadoop (inclusive of HDFS, YARN, and MapReduce) and Apache Ambari, as well as providing a treatment of the main open-source components that are generally made available with the ODP core in a production Hadoop cluster. The participant will be engaged with the product through interactive exercises.

Intended audience

This course is for those who want a foundation of IBM BigInsights. This includes: Big data engineers, data scientists, developers or programmers, and administrators who are interested in learning about IBM's Open Platform with Apache Hadoop.

Topics covered

Topics covered in this course include:

IBM BigInsights Overview:

- Introduction to Big Data
- Introduction to IBM BigInsights
- IBM BigInsights for Analysts
- IBM BigInsights for Data Scientist
- IBM BigInsights for Enterprise Management

IBM Open Platform with Apache Hadoop

- IBM Open Platform with Apache Hadoop
- Apache Ambari
- Hadoop Distributed File System
- MapReduce and Yarn
- Apache Spark
- Coordination Management and Governance
- Data Movement
- Storing and Accessing Data
- Advanced Topics

Course prerequisites

Participants should have:

- None, however, knowledge of Linux would be beneficial.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Document conventions

Conventions used in this guide follow Microsoft Windows application standards, where applicable. As well, the following conventions are observed:

- **Bold**: Bold style is used in demonstration and exercise step-by-step solutions to indicate a user interface element that is actively selected or text that must be typed by the participant.
- *Italic*: Used to reference book titles.
- CAPITALIZATION: All file names, table names, column names, and folder names appear in this guide exactly as they appear in the application.
  To keep capitalization consistent with this guide, type text exactly as shown.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Additional training resources

- Visit IBM Analytics Product Training and Certification on the IBM website for details on:
  - Instructor-led training in a classroom or online
  - Self-paced training that fits your needs and schedule
  - Comprehensive curricula and training paths that help you identify the courses that are right for you
  - IBM Analytics Certification program
  - Other resources that will enhance your success with IBM Analytics Software
- For the URL relevant to your training requirements outlined above, bookmark:
  - Information Management portfolio: http://www-01.ibm.com/software/data/education/
  - Predictive and BI/Performance Management/Risk portfolio: http://www-01.ibm.com/software/analytics/training-and-certification/

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM product help

Help type	When to use	Location
Task-oriented	You are working in the product and you need specific task-oriented help.	*IBM Product* - Help link
Books for Printing (.pdf)	You want to use search engines to find information. You can then print out selected pages, a section, or the whole book. Use Step-by-Step online books (.pdf) if you want to know how to complete a task but prefer to read about it in a book. The Step-by-Step online books contain the same information as the online help, but the method of presentation is different.	Start/Programs/*IBM Product*/Documentation
IBM on the Web	You want to access any of the following:	
	• IBM - Training and Certification	• http://www-01.ibm.com/software/analytics/training-and-certification/
	• Online support	• http://www-947.ibm.com/support/entry/portal/Overview/Software
	• IBM Web site	• http://www.ibm.com

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

BigInsights Overview

- Introduction to Big Data
- Introduction to IBM BigInsights
- IBM BigInsights for Analysts
- IBM BigInsights for Data Scientist
- IBM BigInsights for Enterprise Management

© Copyright IBM Corporation 2015

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Introduction to Big Data

IBM BigInsights v4.0

© Copyright IBM Corporation 2015
Course materials may not be reproduced in whole or in part without the written permission of IBM.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Unit objectives

- Understand when and why you would use big data
- Explain the perception gap
- Explain the difference between data-at-rest and data-in-motion
- Describe the 3 Vs

© Copyright IBM Corporation 2015

Unit objectives

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

System of Units / Binary System of Units

International System of Units (SI)			Binary Usage (deprecated)
kilobyte	KB	10^3	2^{10}
megabyte	MB	10^6	2^{20}
gigabyte	GB	10^9	2^{30}
terabyte	TB	10^{12}	2^{40}
petabyte	PB	10^{15}	2^{50}
exabyte	EB	10^{18}	2^{60}
zettabyte	ZB	10^{21}	2^{70}
yottabyte	YB	10^{24}	2^{80}

International Electrotechnical Commission (IEC) - 1999		
kibibyte	KiB	2^{10}
mebibyte	MiB	2^{20}
gibibyte	GiB	2^{30}
tebibyte	TiB	2^{40}
pebibyte	PiB	2^{50}
exbibyte	EiB	2^{60}
zebibyte	ZiB	2^{70}
yobibyte	YiB	2^{80}

Source: Wikipedia, http://en.wikipedia.org/wiki/Kibibytes

Introduction to Big Data

© Copyright IBM Corporation 2015

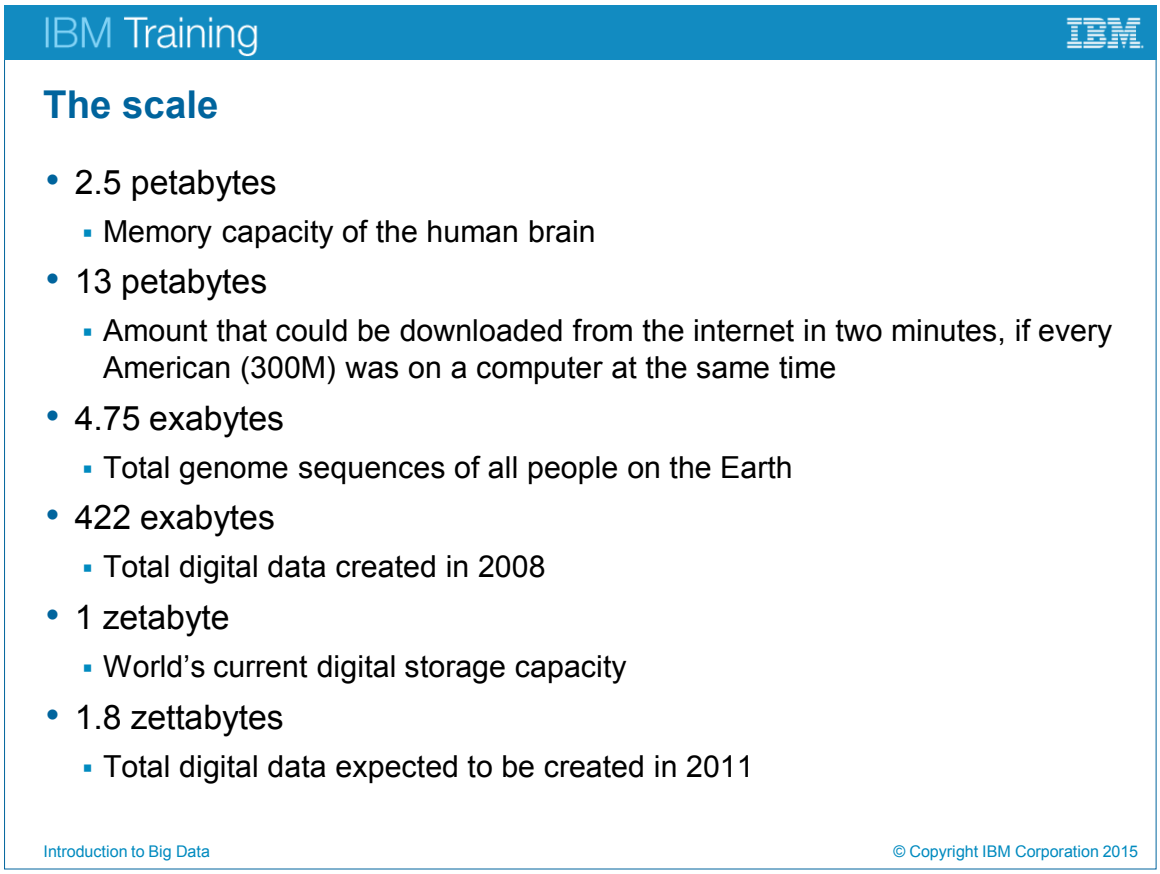
System of Units/Binary System of Units

When dealing with big data, we speak of numbers that are not part of our everyday conversations. Terms like kilobytes, megabytes, and gigabytes are commonly known. The term terabyte has been added to our discussions in the past couple of years. But to most people, the terms petabyte, exabyte, zettabyte, and yottabyte sound foreign. Like it or not, those terms are necessary when dealing with big data. Some of these terms will be used in this course. You should at least have a basic understanding of what they mean.

- kilobyte (KB) 10 to the 3rd power
- megabyte (MB) 10 to the 6th power
- gigabyte (GB) 10 to the 9th power
- terabyte (TB) 10 to the 12th power
- petabyte (PB) 10 to the 15th power
- exabyte (EB) 10 to the 18th power
- zettabyte (ZB) 10 to the 21st power
- yottabyte (YB) 10 to the 24th power

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



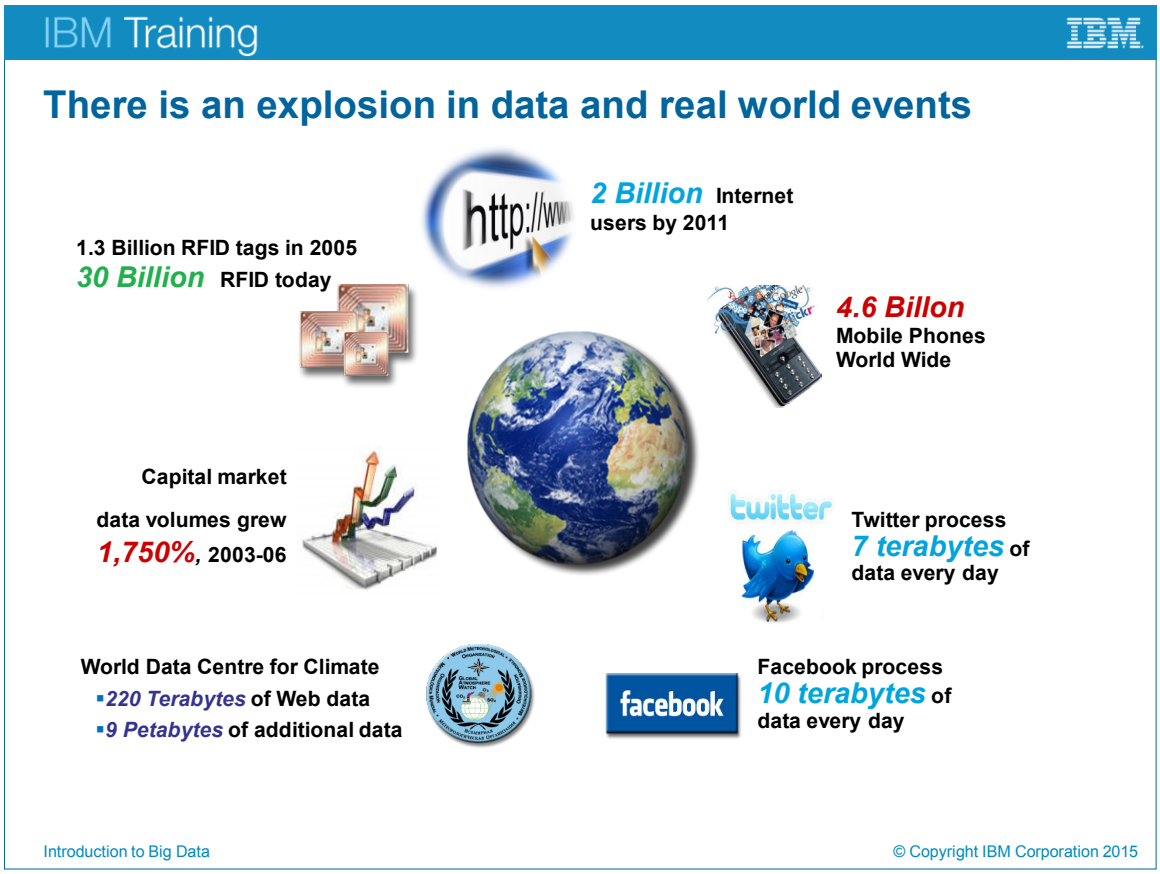
The scale

It is hard for most people to grasp the concept of how large a petabyte or an exabyte is. For a long time people thought that a billion was a large number. But as quickly as most governments spend a billion dollar or euros, obviously, it cannot be that large of a number. To better understand extremely large numbers, it is best to view them in comparison to something that you can understand. The capacity of the human brain is about 2.5 petabytes. (This is also the estimated size of Walmart databases that handle 1 million customer transactions a day.) The total genome sequences of all people on the Earth is 4.75 exabytes. The total amount of digital data created in 2008 was 422 exabytes. And the total that was expected to be created in 2011 was 1.8 zettabytes.

In 2000 the Sloan Digital Sky Survey began collecting astronomical data. In the first few weeks it amassed more data than was collected in the history of astronomy. And the total amount of data collected by the SDSS is the amount that its successor, the Large Synoptic Survey Telescope, is expected to collect every 5 days, when it comes online in 2016.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



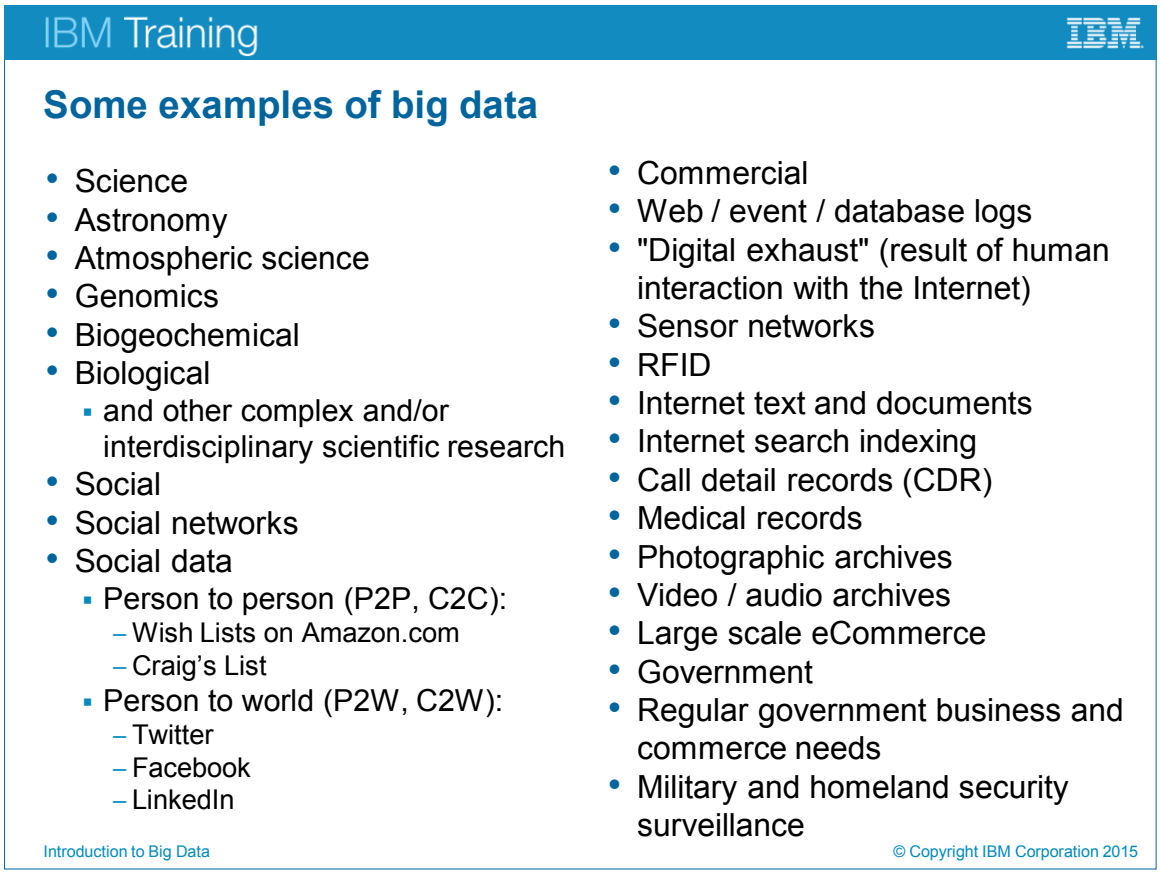
There is an explosion in data and real world events

The amount of data that gets created every day is at mind boggling proportions and will only continue to increase. Moore's law states that the speed of computer processing will double every two years. It seems that there is some sort of a corollary to this law when it comes to data as well. The problem that we have when dealing with so much data is that it becomes almost impossible to separate the important facts from the non-important facts. So we need computer programs to help us to distill the data. But a single program, working with terabytes of data, requires a lot of time to process that much data. And by the time the processing has been completed, the answer may no longer be relevant. For example knowing the traffic patterns of the last three days does not help you to determine how to cross a street at this instance.

The other thing that makes working with all of this data so difficult is the fact that most of the data is unstructured. Computer programs work well with structured data. If there is a data field that only has postal code values, then it is easy to search on that field and get all of the stores within a particular geographical area. But what if you are looking for some key words in recorded conversations? The data is there but accessing it becomes significantly harder.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

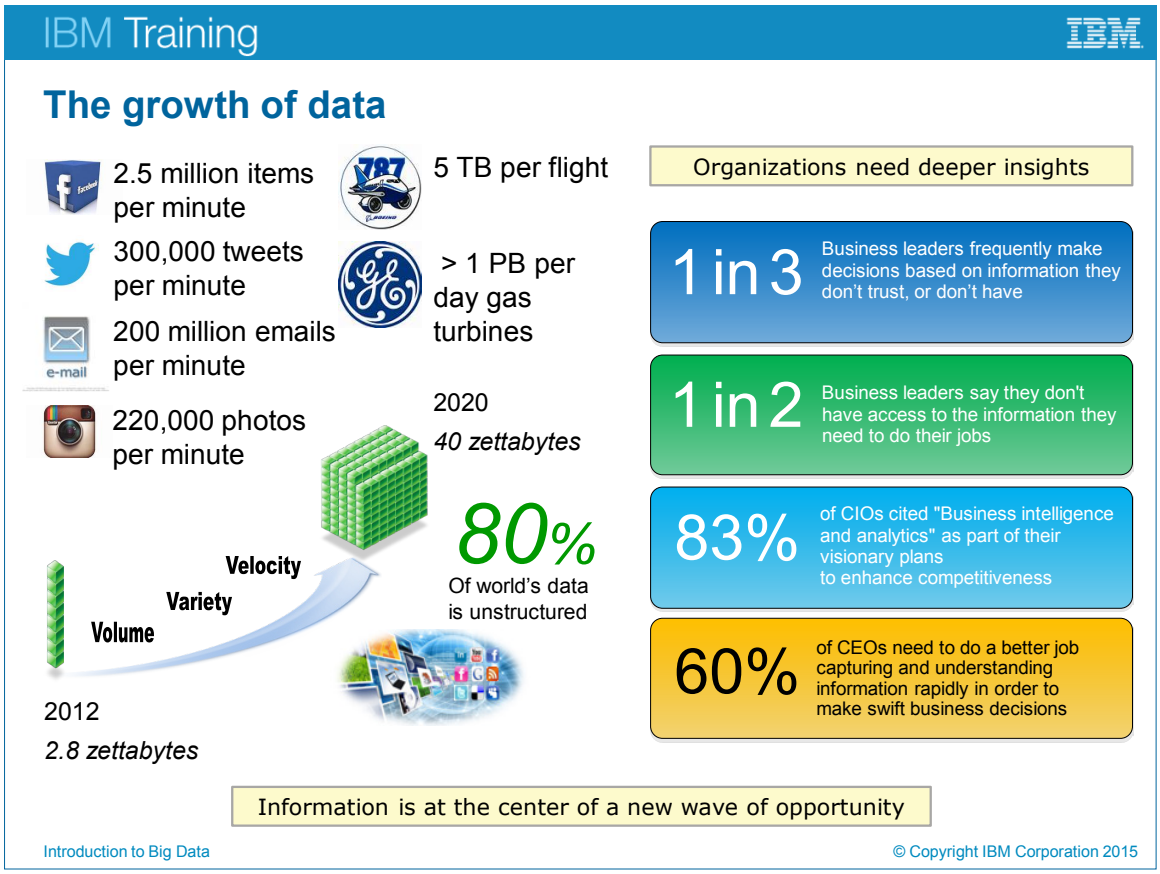


Some examples of big data

Some examples of big data are social networks, web logs, RFID information, video and audio archives, sensor data, military surveillance, astronomy, genomics and internet search indexing.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



The growth of data

The growth of data is staggering. Just look at some of the statistics shown in the diagram. The tremendous volume and variety of data being generated at an accelerated velocity creates tremendous opportunity for organizations. Unfortunately, organizations are struggling to gain deeper insights from this data. Business leaders continue to make decisions without access to the trusted information they need.

We all understand the well-organized structured data world. We've dealt with it for decades. It's at the very core of what information technology and the advancement of programmable computing has brought to us over the last 60 years. But a lot of data is unstructured or semi-structured. We're really just at the very beginning of in terms of what the possibilities are, how we can get at that information and what we can do with it. Think about all the information being generated by social networking sites (like Twitter, Facebook, LinkedIn, and so on), web logs, click streams, instant messages, emails, electronic sensor data, and so on. How would it change your business if you could efficiently filter through that data, aggregate the right bits and pieces, combine it with your operational data, and analyze it effectively?

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

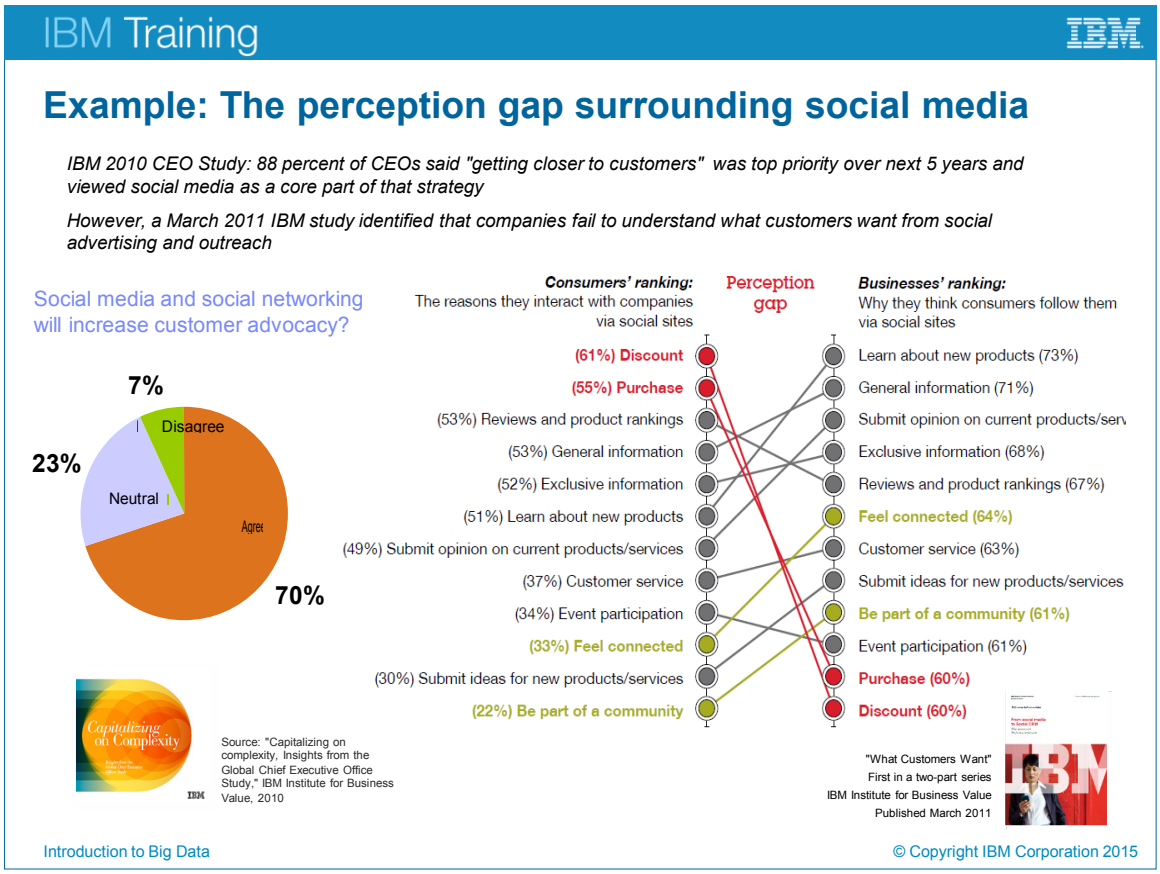
Sources:

- *The Guardian*, May 2010
- *IDC Digital Universe*, 2010
- *IBM Institute for Business Value*, 2009
- *IBM CIO Study 2010*
- *TDWI: Next Generation Data Warehouse Platforms* Q4 2009
- https://blog.kissmetrics.com/facebook-statistics/
- http://www.webopedia.com/quick_ref/just-how-much-data-is-out-there.html
- http://www.computerworlduk.com/news/infrastructure/3433595/boeing-787s-to-create-half-a-terabyte-of-data-per-flight-says-virgin-atlantic/
- http://www.webopedia.com/quick_ref/just-how-much-data-is-out-there.html
- http://www.forbes.com/sites/maribellopez/2013/05/10/ge-speaks-on-the-business-value-of-the-internet-of-things/
- http://www.idc.com/prodserv/4Pillars/bigdata;jsessionid=94A407E4522FB407627ECEBBAAA90A24
- http://www.digitalbuzzblog.com/infographic-24-hours-on-the-internet/
- ZB = 1 billion TB
- IDC reference:
    - http://idcdocserv.com/925
    - http://www.computer.org/portal/web/news/home/-/blogs/2613266;jsessionid=abbfded1402383e107abfa2641d6

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Example: The perception gap surrounding social media

Let's look into one facet of the big data challenge a bit further. An IBM CEO study revealed that social media was a core part of many firms strategy of getting closer to customers. However, a separate study showed that companies often don't understand what consumers really want from them on social media sites. Indeed, the top 2 consumer choices discounts and purchases didn't even make the top 10 list of what companies thought their consumers wanted. Wouldn't it be great if companies didn't have to guess? What if they could look at consumer behavior and sentiment expressed on social media sites and really know what people wanted. This is one aspect of the big data challenge.

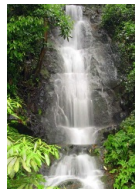
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Streams and oceans of information

- High speed information flowing in real-time, often transient
  - Information from sensors and instruments
  - Information flowing from real-time logs and activity monitors
  - Streaming content like audio and video
  - High speed transactions like tickers, trades, or traffic systems
- Information stored outside conventional systems. Data may originate from the Web or different internal different systems
  - Collection of what has streamed
  - Information from social media, logs, click streams, and emails
  - Unstructured or mixed schema documents like claims, forms, and desktop applications
  - Structured data from disparate systems



Information streams

Information oceans

Introduction to Big Data

© Copyright IBM Corporation 2015

Streams and oceans of information

There are two groups of big data. Some fall into the category of flowing in real time, for example, information coming from sensors or video feeds. Sometimes the real-time data can have very high volumes, like stock tickers or patient monitoring systems in a hospital. This type of data cannot use a "store and access" method. Knowing the volume of trades for a particular stock or a patient's vitals from two days ago does not help you make decision right now. IBM's InfoSphere Streams was developed to handle this type of data, which is referred to as being information streams.

On the other hand we can have massive amounts of stored data, like emails, web logs, and click streams, that need to be analyzed. This data can consist of both structured and unstructured data. The question then becomes how can we process this large amount of data in a timely manner? We refer to data of that type as being information oceans for which IBM BigInsights was designed to address.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Big data presents big opportunities

- Extract insight from a high volume, variety and velocity of data in a timely and cost-effective manner

Variety: Manage and benefit from diverse data types and data structures

Velocity: Analyze streaming data and large volumes of persistent data

Volume: Scale from terabytes to zettabytes

Introduction to Big Data

© Copyright IBM Corporation 2015

Big data presents big opportunities

ZB = 1 billion TB

We believe that big data presents organizations with a big opportunity to extract new insights that can improve their decision-making process and business plans. Massive volume, variety and velocity are defining characteristics of big data, and IBM has built its platform to address these characteristics.

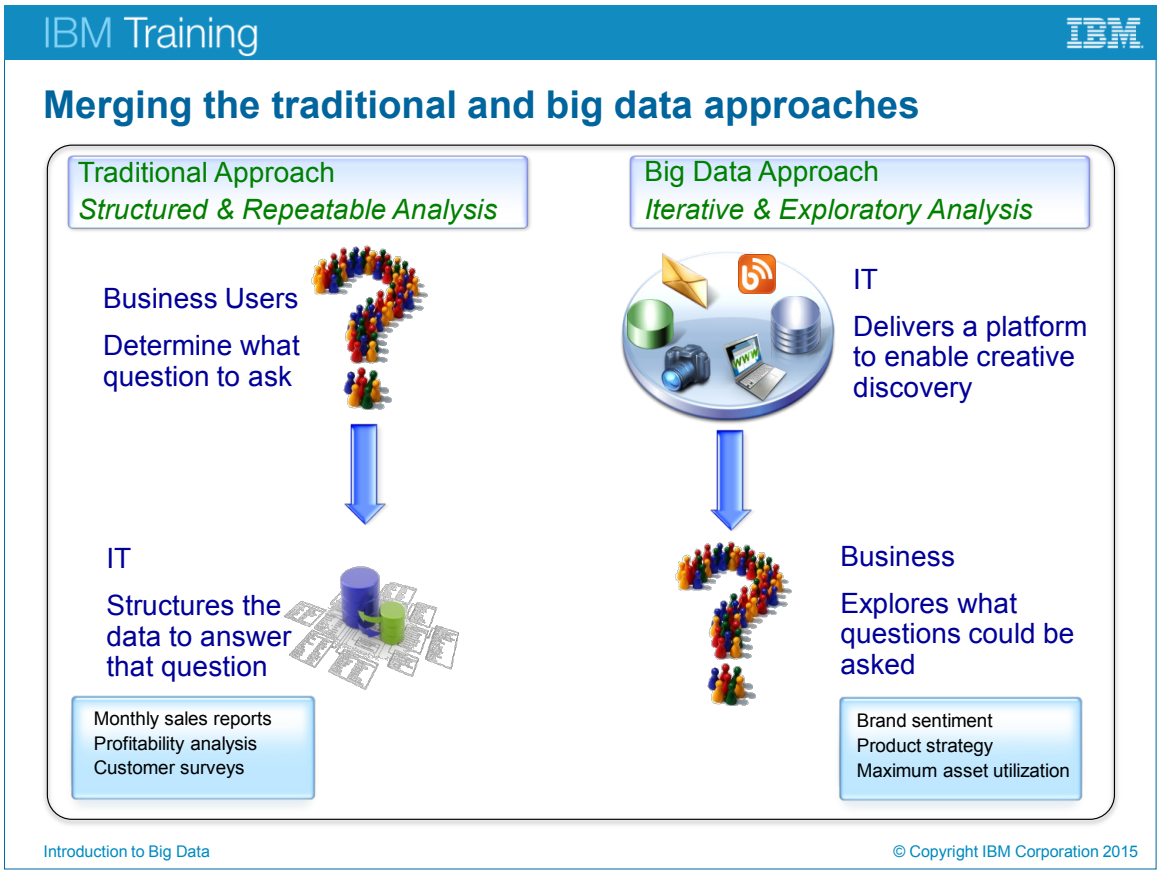
In order to capitalize on big data, firms must be able to analyze a wide variety of data, including text, sensor data, audio, video, transactional data, and others.

Sometimes, getting an edge over your competition can mean identifying a trend, problem or opportunity, seconds, or even microseconds before someone else. More and more of the data being produced today, has a very short half-life. Organizations must be able to analyze this data in real-time if they are to be able to find insights in this data.

And, as implied by the term big data, organization are facing massive volumes of data. Organizations that don't know how to manage this data are overwhelmed by it. But wouldn't it be great if the right technology was available to analyze all of the data so that you could gain a better understanding of your business, your customers, and the marketplace?

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Merging the traditional and big data approaches

The big data approach complements the traditional approach.

The traditional approach calls for business users to determine what questions to ask and IT structure the data to answer that question. This is well suited to many common business processes, such as monitoring sales by geography, product or channel; extract insight from customer surveys; cost and profitability analyses.

The big data approach is a bit different. With this approach, IT delivers a platform that consolidates data sources of interest and enables creative discovery. Then the business users use the platform to explore data for idea and questions to ask.

On the left, the traditional approach allows organization to answer questions that will be asked time and time again. On the right, users have the ability to explore their data in a more creative way. Before finding the answer, they must first define the question. Are my customers starting to change their preferences? What is the best way to measure brand health?

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

What we hear from customers

- Lots of potentially valuable data is dormant or discarded due to size/performance considerations
- Large volume of unstructured or semi-structured data is not worth integrating fully (such as Tweets, logs, etc.)
- Not clear what should be analyzed (exploratory, iterative)
- Information distributed across multiple systems and/or Internet
- Some information has a short useful lifespan
- Volumes can be extremely high
- Analysis needed in the context of existing information (not stand alone)

© Copyright IBM Corporation 2015

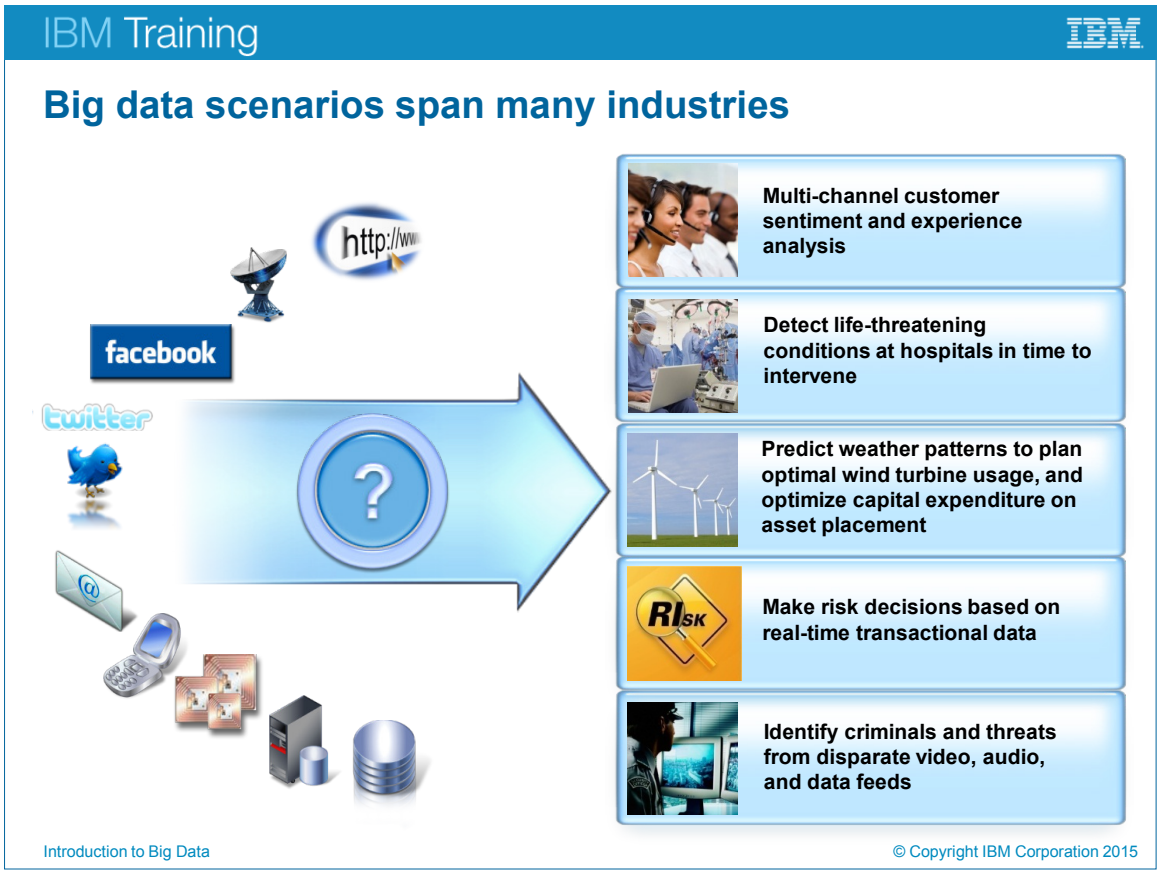
What we hear from customers

We've been working with a number of customers who tell us about the kinds of challenges they're facing with big data. In many cases, they're uncertain exactly what needs to be analyzed - that is, they need to explore the volumes of data they have to discover what might be of value. Quite often, large volumes of information are currently lying dormant in their firms or are discarded completely due to size or performance considerations.

In addition, potentially interesting business data is seldom in one place and much of the unstructured data that may contain useful tidbits of information isn't worth fully integrating into a data warehouse or in-house operational system. An example of this includes posts to social media sites, such as Facebook, Twitter, or Yelp. Some of the information has a short useful lifespan (sensor feeds, news feeds, and web logs are examples of this), and volumes can be extremely high. Firms are looking for an efficient, cost-effective way to address these issues within the context of their existing businesses.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Big data scenarios span many industries

The need to cope with and leverage big data spans many industries and application domains.

- Imagine if you could analyze all the tweets being created each day to figure out what people are saying about your products and who the key influencers are within your target demographics. Imagine being able to mine this data to identify new market opportunities.
- What if hospitals could take the thousands of sensor readings collected every hour per patients in ICUs to identify subtle indications that the patient is becoming unwell, days earlier that is allowed by traditional techniques.
- Imagine if a green energy company could use PBs of weather data along with massive volumes of operational data to optimize asset location and utilization, making these environmentally friendly energy sources more cost competitive with traditional sources.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

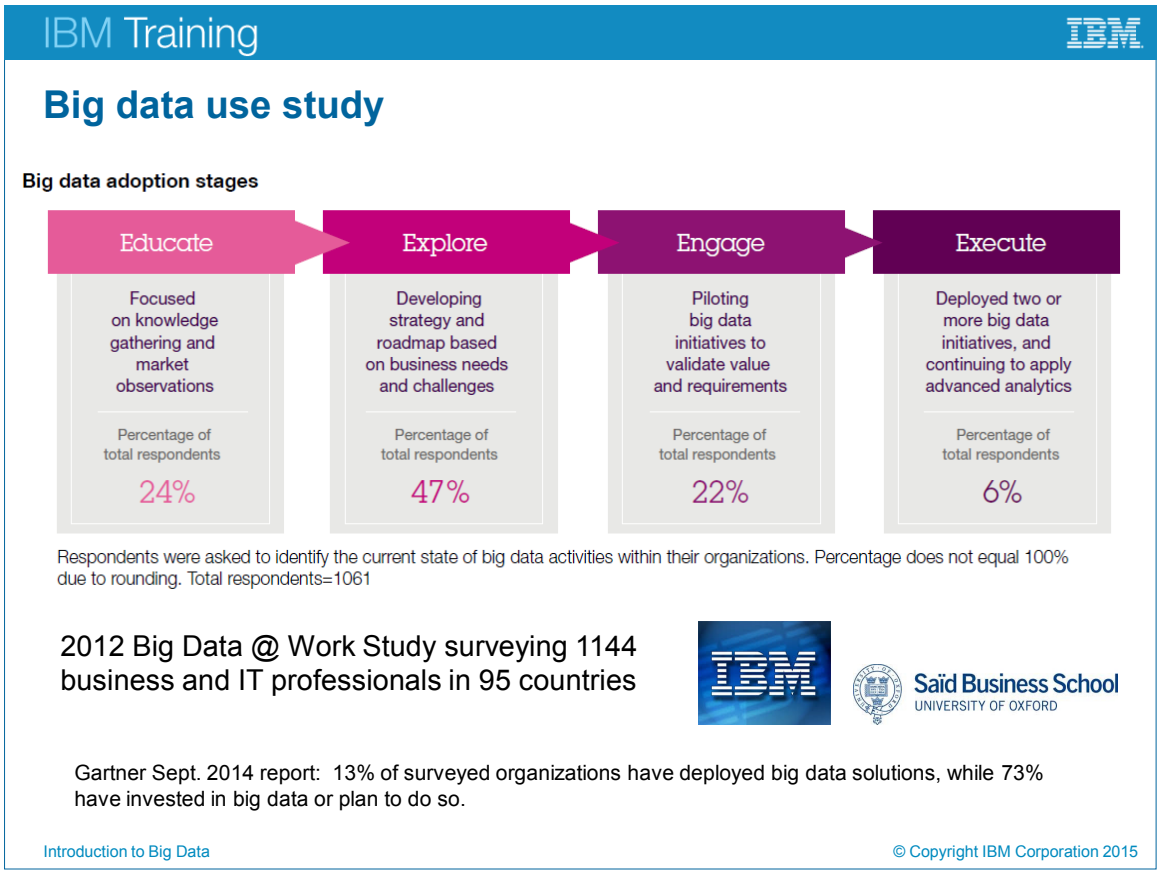
- Imagine if you could make risk decisions, such as whether or not someone qualifies for a mortgage, in minutes, by analyzing many sources of data, including real-time transactional data, while the client is still on the phone or in the office.
- Imagine if law enforcement agencies could analyze audio and video feeds in real-time without human intervention to identify suspicious activity.

As these new sources of data continue to grow in volume, variety and velocity, so too does the potential of this data to revolutionize the decision-making processes in every industry.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Big data use study

While some organizations have been very successful launching production big data projects, studies show that the majority of organizations are still in the early adoption stages, as shown in a study conducted by the University of Oxford and IBM in 2012.

Source: *Analytics: The real-world use of big data, How innovative enterprises extract value from uncertain data, IBM Institute for Business Value and Saïd Business School at the University of Oxford, 2012*

Link:
http://public.dhe.ibm.com/common/ssi/ecm/en/gbe03519usen/GBE03519USEN.PDF

Following are excerpts from this report related to the chart shown. Additionally, in September 2014 Gartner released a study showing similar results. IBM does not have the right to redistribute this report, titled *Major Myths About Big Data's Impact on Information Infrastructure, G00269433*.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

In the Educate stage, the primary focus is on awareness and knowledge development. Almost 25 percent of respondents indicated that they are not yet using big data within their organizations. While some remain relatively unaware of the topic of big data, our interviews suggest that most organizations in this stage are studying the potential benefits of big data technologies and analytics, and trying to better understand how big data can help address important business opportunities in their own industries or markets.

The focus of the Explore stage is to develop an organization's roadmap for big data development. Almost half of respondents reported formal, ongoing discussions within their organizations about how to use big data to solve important business challenges. Key objectives of these organizations include developing a quantifiable business case and creating a big data blueprint.

In the Engage stage, organizations begin to prove the business value of big data, as well as perform an assessment of their technologies and skills. More than one in five respondent organizations is currently developing proofs-of-concept (POCs) to validate the requirements associated with implementing big data initiatives, as well as to articulate the expected returns.

In the Execute stage, big data and analytics capabilities are more widely operationalized and implemented within the organization. However, only 6 percent of respondents reported that their organizations have implemented two or more big data solutions at scale, the threshold for advancing to this stage.

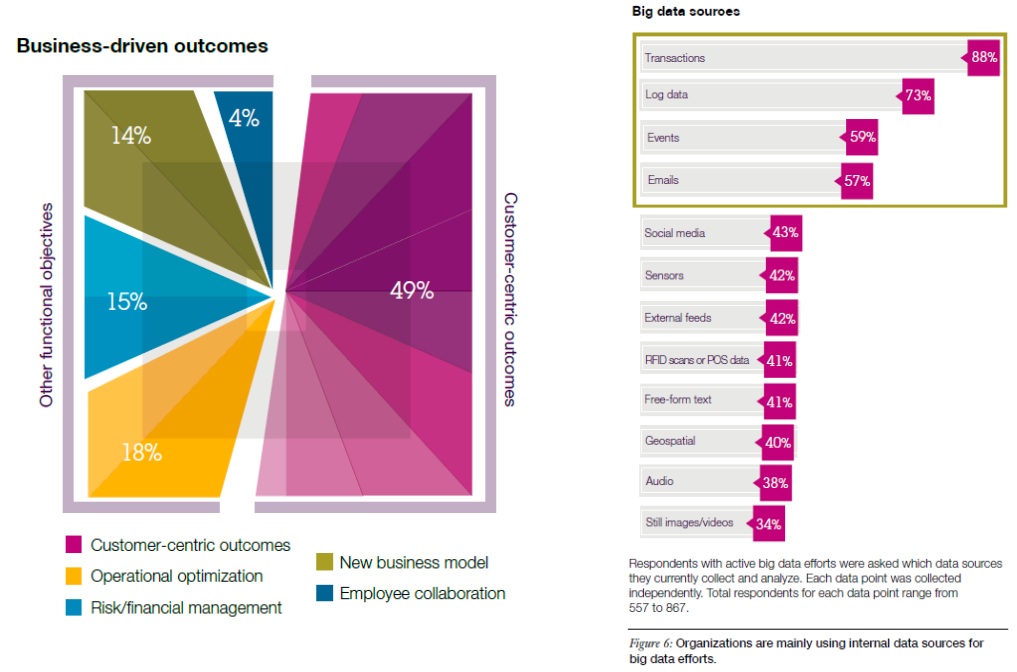
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Big data use: focus areas and data sources



Introduction to Big Data

© Copyright IBM Corporation 2015

Big data use: focus areas and data sources

When asked to rank their top three objectives for big data, nearly half of the respondents identified customer-centric objectives as their organization's top priority (see figure on the left). Companies clearly see big data as providing the ability to better understand and predict customer behaviors, and by doing so, improve the customer experience. Transactions, multi-channel interactions, social media, syndicated data through sources like loyalty cards, and other customer-related information have increased the ability of organizations to create a complete picture of customers' preferences and demands, a goal of marketing, sales and customer service for decades.

Any big data initiatives begin with untapped sources of internal information. In the figure on the right, you can also see that external data sources (such as social media) factor significantly in big data strategies as well.

Source: *Analytics: The real-world use of big data, How innovative enterprises extract value from uncertain data, IBM Institute for Business Value and Saïd Business School at the University of Oxford, 2012*

Link: http://public.dhe.ibm.com/common/ssi/ecm/en/gbe03519usen/GBE03519USEN.PDF

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Unit summary

- Understand when and why you would use big data
- Explain the perception gap
- Explain the difference between data-at-rest and data-in-motion
- Describe the 3 Vs

© Copyright IBM Corporation 2015

Unit summary

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Exercise 1

Setting up the lab environment

© Copyright IBM Corporation 2015

Exercise 1: Setting up the lab environment

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Exercise 1: Setting up the lab environment

Purpose:
You will set up your lab environment by starting the VMWare image, launching the Ambari console, and starting the required services. You will also learn about the file system and directory structures.

Estimated time: **30 minutes**
User/Password: **biadmin/biadmin**
root/dalvm3
Services Password: **ibm2blue**

Task 1. Configure your image.

As copies are made of the VMWare image, additional network devices get defined and the IP address changes. Configuration changes are required to get the Ambari console to work.

Note: Occasionally, when you suspend and resume the VM image, the network may assign a different IP address than the one you had configured. In these instances, the Ambari console and the services will not run. You will need to update /etc/hosts file with the newly assigned IP address to continue working with the image. No restart of the VM image is necessary, just give it a couple of minutes, at most. In some cases, you may need to restart the Ambari server, using *ambari-server restart* from the command line.

1. To open a new terminal, right-click the desktop, and then click **Open in Terminal**.
2. Type `ifconfig` to check for the current assigned IP address.
3. Take note of the IP address next to **inet**.

   You need to edit the /etc/hosts file to map the hostname to the IP address.
4. To switch to the root user, type `su -`.
5. When prompted for a password, type **dalvm3**.
6. To open the /etc/hosts file, type `gedit /etc/hosts`.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

7. Ensure that the contents of the file are similar to the following:

```
10.0.0.118 ibmclass.localdomain ibmclass
127.0.0.1 localhost.localdomain localhost
```

8. Update the IP address on the first line from step 3.
9. Save and exit the file, and then close the terminal.

Task 2. Start the BigInsights components.

You will start all the services via the Ambari console to ensure that everything is ready for the exercise. You may stop what you don't need later, but for now, you will start everything.

1. Launch **Firefox**, and then if necessary, navigate to the **Ambari** login page, **http://ibmclass.localdomain:8080**.
2. Log in to the **Ambari** console as **admin/admin**.

   On the left side of the browser are the statuses of all the services. If any are currently yellow, wait a couple of minutes for them to become red before proceeding.
3. Once all the statuses are red, at the bottom of the left side, click **Actions** and then click **Start All** to start the services.

   This will take several minutes to complete.
4. When the services have started successfully, click **OK**.

Task 3. Begin to explore Ambari.

This section will provide some basic Ambari administration and cluster management. The IBM Open Data Platform (IOP) with Apache Hadoop section of this course will cover Ambari administration in more detail.

1. Launch **Firefox**, and then if necessary, navigate to the **Ambari** login page, **http://ibmclass.localdomain:8080**.
2. Log in to the **Ambari** console as **admin/admin**.

   Once logged in, you will notice the statuses of the services on the left side. If everything is green, then all services are running.

   You can select any of the services to go to the details page for that service.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

3. Click the **HDFS** service.

   On the HDFS service page, you will see a Summary of the node(s) status as well as any Alerts or Health Checks.

   Alongside the Summary tab is the Configs tab. You can specify specific configurations from there. You will not do anything there in this exercise.

   At the far right of the screen, there is the Service Action dropdown. Here is where you can start, stop, restart, or any perform other actions specific to the selected service.

4. At the top right, beside the **Services** link, click the **Hosts** link.

   You can access the specific nodes of your cluster here.

5. Click **ibmclass.localdomain**.

   This will take you to the host summary page. From here, you can perform specific Host Actions, located on the right side of the page. Actions include starting, stopping, or restarting all of the components on the host.

   One final overview is the Background Operations. This is located next to the cluster name at the top. Currently, you should see 0 ops.

6. Click **0 ops**, to see current and past operations.

   0 ops (zero ops) means that there are no operations currently running in the background.

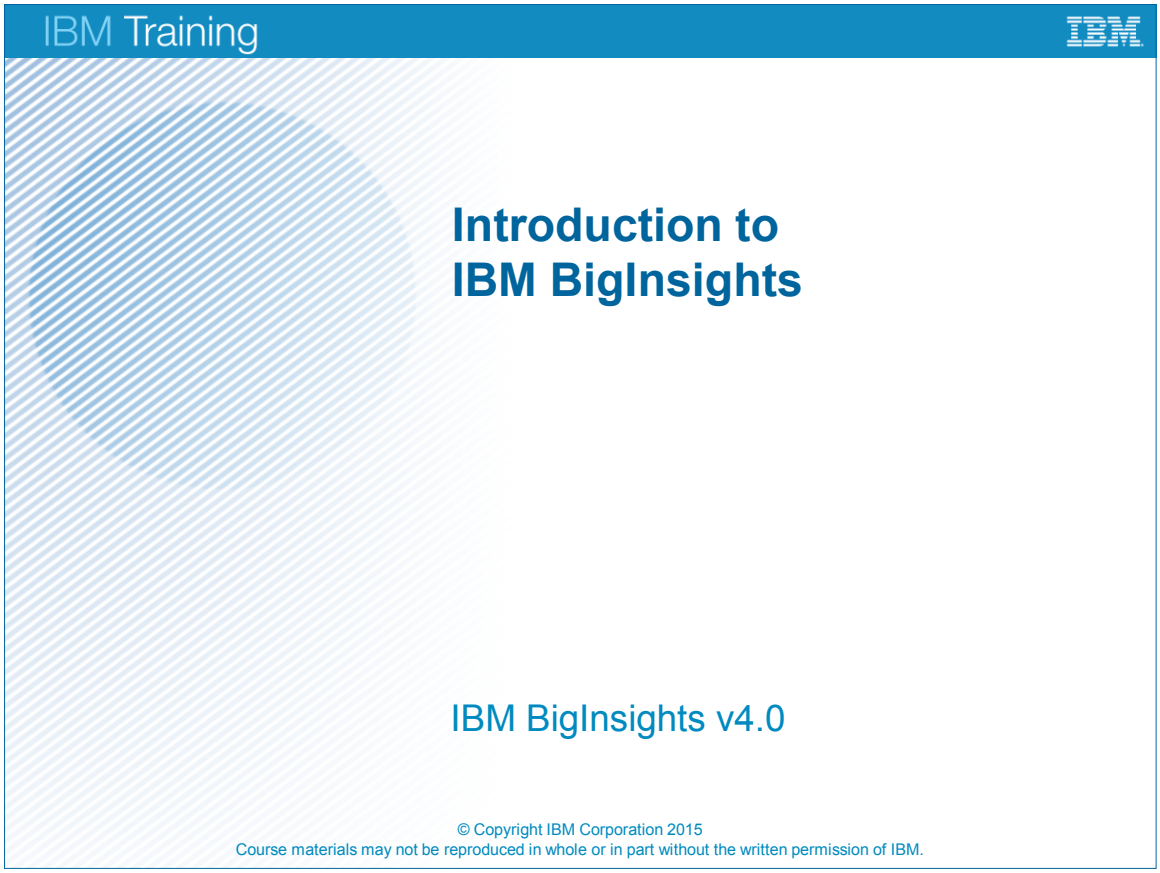
7. Click **OK**, and then close Firefox.

Results:
You have set up your lab environment by starting the VMWare image, launching the Ambari console, and starting the required services. You also learned about the file system and directory structures.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Unit objectives

- Describe the functions and features IBM BigInsights
- List the IBM value-add components that comes with BigInsights
- Give a brief description of the purpose of each of the value-add components

© Copyright IBM Corporation 2015

Unit objectives

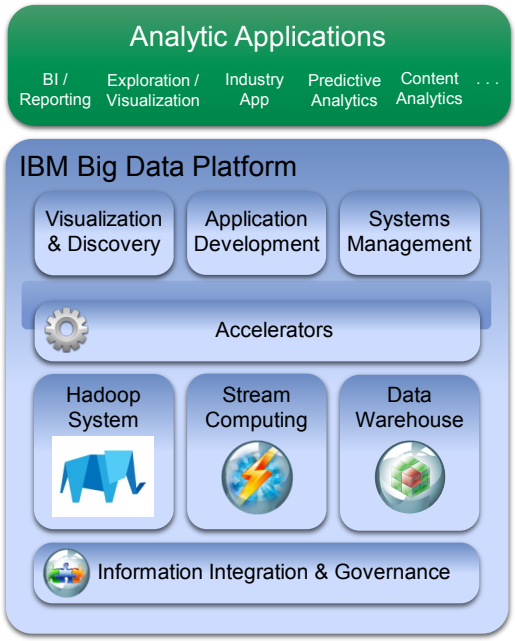
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

IBM big data strategy

- Integrate and manage the full variety, velocity and volume of big data
- Apply advanced analytics to information in its native form
- Visualize all available data for ad-hoc analysis
- Development environment for building new analytic applications
- Support workload optimization and scheduling
- Provide for security and governance
- Integrate with enterprise software



Introduction to IBM BigInsights

© Copyright IBM Corporation 2015

IBM big data strategy

Key Points

- Integrate - the point is to have one platform to manage all of the data - there's no point in having separate silos of data, each creating separate silos of insight. From the customer POV (a solution POV) big data has to be bigger than just one technology.
- Analyze - we see big data as a viable place to analyze and store data. New technology is not just a pre-processor to get data into a structured DW for analysis. Significant area of value add by IBM - and the game has changed - unlike DBs/SQL, the market is asking who gets the better answer and therefore sophistication and accuracy of the analytics matters.
- Visualization - need to bring big data to the users - spreadsheet metaphor is the key to doing so.
- Development - need sophisticated development tools for the engines and across them to enable the market to develop analytic applications.
- Workload optimization - improvements upon open source for efficient processing and storage.

Security and Governance - many are rushing into big data like the Wild West. But there is sensitive data that needs to be protected, retention policies need to be determined - all of the maturity of governance for the structured world can benefit the big data world.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

IBM BigInsights for Apache Hadoop

- Analytical platform for persistent big data
  - 100% open source core with add-on IBM technologies for data analysts, data scientists, and enterprise administrators
  - On premise installation or cloud offerings
- Distinguishing characteristics
  - Built-in analytics: Enhances business knowledge
  - Enterprise software integration: Complements and extends existing capabilities
  - Production-ready platform: Speeds time-to-value
- IBM advantage
  - Combination of software, hardware, services and advanced research

Introduction to IBM BigInsights

© Copyright IBM Corporation 2015

IBM BigInsights for Apache Hadoop

First, a quick summary of BigInsights now, and then you will dive into details in the next several diagrams. BigInsights is IBM's strategic platform for managing and analyzing persistent big data. As you will see, it is based on a 100% open source core with IBM-unique technologies for data analysts, data scientists, and system administrators. You choose what you want to install and tailor your system to your needs. In addition, you can install it on your cluster or use IBM's cloud offering.

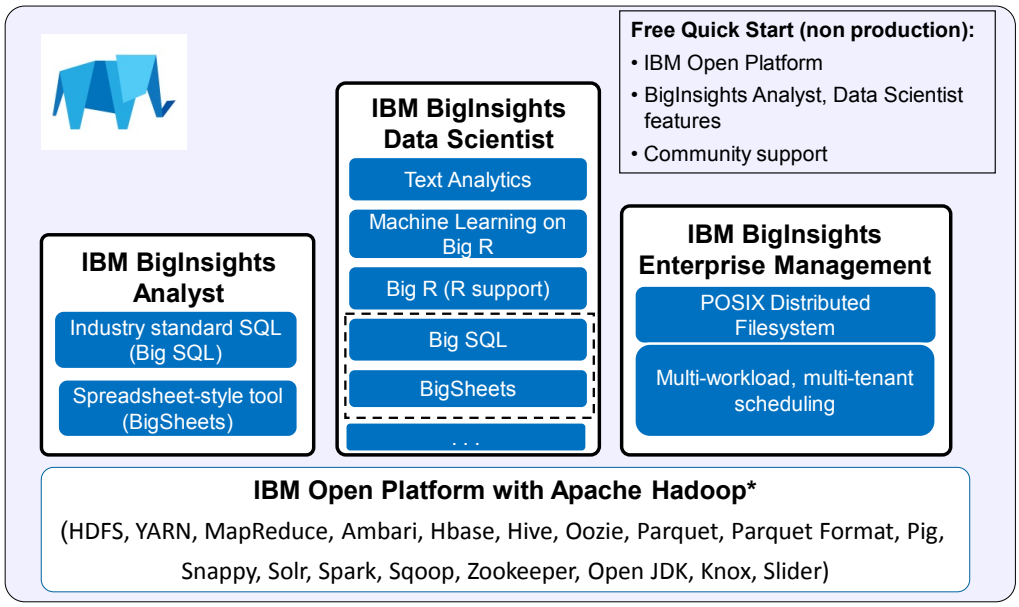
Some of the characteristics that distinguish BigInsights include its built-in support for analytics, its integration with other enterprise software, and its production readiness. You will see more about these topics later. Note that IBM is uniquely positioned to provide customers with the necessary software, hardware, services, and research advances in the world of big data.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Overview of BigInsights



*IBM Open Platform with Apache Hadoop is a 100% open source Apache Hadoop distribution. IBM will include the Open Data Platform common kernel once available.

© Copyright IBM Corporation 2015

Overview of BigInsights

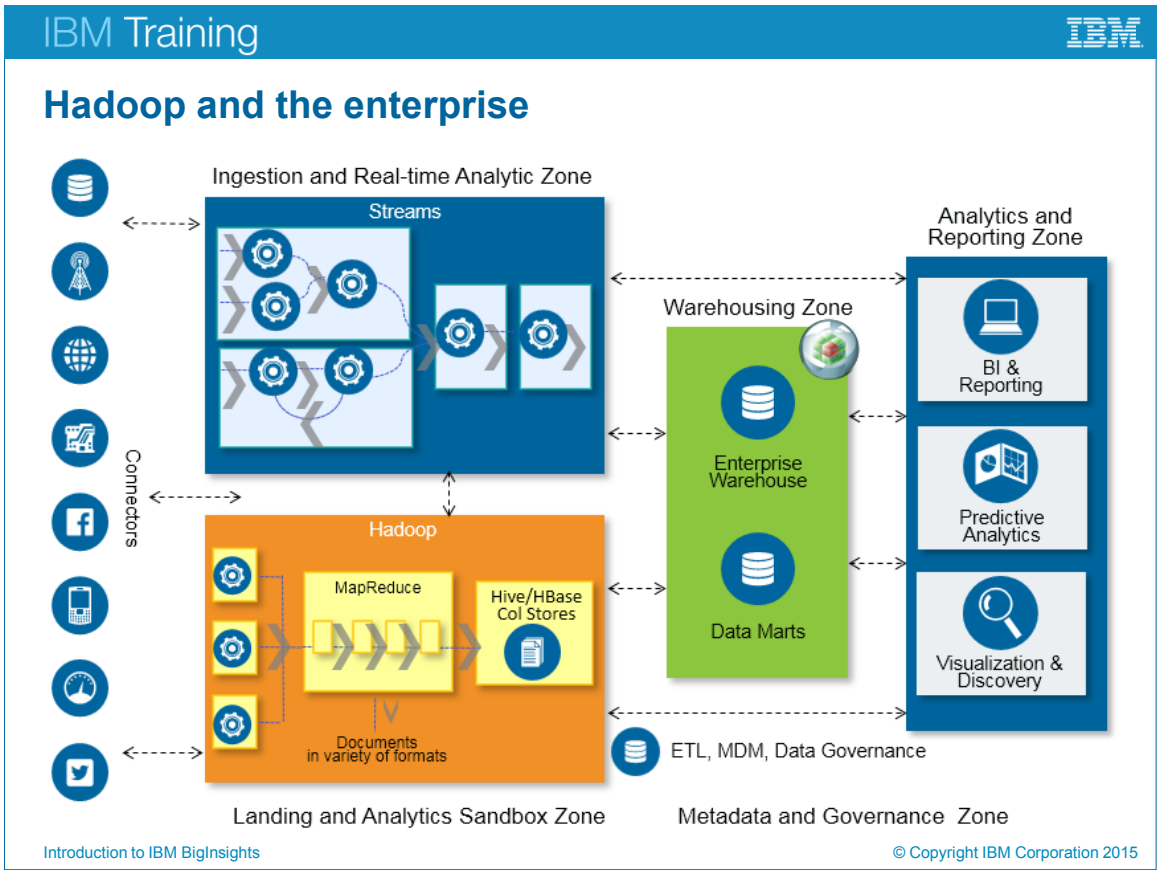
At the bottom, there's a 100% open source platform based on key Apache components, like HDFS, YARN, Spark, HBase, Hive, and others. IBM has joined an industry consortium (the Open Data Platform Initiative) whose mission is to define, test, and validate a common core of Hadoop components. When that core set becomes available, IBM will support it as part of its core platform. The goal here is to contribute to the open source community in a way that helps all organizations be active in the Hadoop space and mitigates concerns about "vendor lock in".

But BigInsights is more than that. Years of IBM research and development efforts have incorporated the results into 3 modules that you can add to the Open Platform stack. The BigInsights Analyst module includes a sophisticated SQL engine (Big SQL) and an easy-to-use spreadsheet-style tool (BigSheets) for exploring big data. BigInsights Data Scientist includes all the Analyst features and adds important analytical technologies for text, R integration, and machine learning. BigInsights Enterprise Management offers a robust, POSIX-compliant file system alternative to HDFS as well as key technologies for managing multiple workloads and multiple tenants on your cluster. These 3 offerings are fee based. However, if you want to get off to a quick start, we offer a free Quick Start edition for non-production use.

This will be covered in more detail later; now that you know a little about BigInsights, let's consider how this technology fits into a broader IT infrastructure.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Hadoop and the enterprise

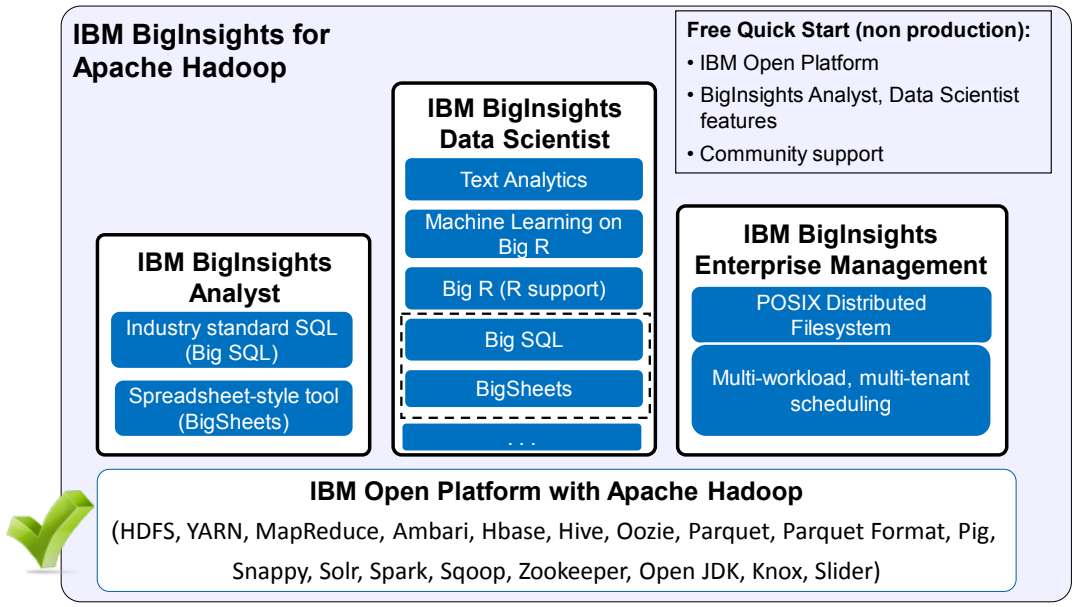
This diagram illustrates a pattern that has emerged in many organizations. The box at lower left shows Hadoop being used as a landing zone for capturing a variety of data in its native format. Built-in analytic and data management technologies allow developers and analysts to explore this raw data in a sandbox environment. Sources of data can include web pages, system logs, input from streaming engines (shown at upper left), and even reference data pulled from traditional sources, such as data warehouses or marts. Because BigInsights provides an industry-standard SQL interface, many traditional analytical and reporting tools used with relational DBMSs can work directly against BigInsights data, if needed.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Overview of BigInsights



Introduction to IBM BigInsights

© Copyright IBM Corporation 2015

Overview of BigInsights

You will start by exploring the base platform.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

About the IBM Open Platform for Apache Hadoop

- Flexible platform for processing large volumes of data
  - Includes Apache Hadoop and many popular open source projects in the Hadoop ecosystem
  - Supports wide variety of data
  - Supports variety of popular APIs (industry-standard SQL, MapReduce, etc.)
- Enables applications to work with thousands of nodes and petabytes of data in a highly parallel, cost effective manner
  - CPU + disks = "node"
  - Nodes can be combined into clusters
  - New nodes can be added as needed without changing
    - Data formats
    - How data is loaded
    - How jobs are written

© Copyright IBM Corporation 2015

About the IBM Open Platform for Apache Hadoop

BigInsights is a flexible, enterprise-ready platform for processing large volumes of data in a highly distributed and efficient manner. It includes Hadoop, an Apache project being built and used by a global community of contributors. As such, it can support a wide variety of data as well as many popular APIs.

Hadoop was originally designed to support batch-oriented, read-intensive applications. Recently, complementary projects have been broadening its usefulness for other types of applications. One of the key features of Hadoop is its ability to distribute and manage data across a large number of nodes and disks. Programmers can create applications that automatically take advantage of parallel processing by using the MapReduce API, certain scripting languages, and other technologies.

In Hadoop terminology, a computing node consists of one or more CPUs and attached disks. Such nodes can be combined into clusters, and new nodes can be added to a cluster without an administrator or programmer changing the format of the data, how the data was loaded, or how the jobs (programming logic) were written.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

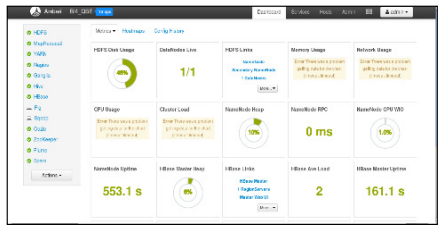
© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Open source currency

- Timely updates as new open source versions released
- Component levels as of March 2015:

Ambari	1.7	Apache License, Version 2.0
Avro	1.7.7	Apache License, Version 2.0
Flume	1.5.2	Apache License, Version 2.0
Hadoop	2.6.0	Apache License, Version 2.0
Hbase	0.98.8	Apache License, Version 2.0
Hive	0.14.0	Apache License, Version 2.0
Knox	0.5.0	Apache License, Version 2.0
Oozie	4.1.0	Apache License, Version 2.0
Parquet-MR / Format	1.5.0 / 2.1	Apache License, Version 2.0
Pig	0.14.0	Apache License, Version 2.0
Slider	0.60.0	Apache License, Version 2.0
Solr	4.10.3	Apache License, Version 2.0
Spark	1.2.1	Apache License, Version 2.0
Sqoop	1.4.5	Apache License, Version 2.0
Zookeeper	3.4.6	Apache License, Version 2.0



Apache Ambari

- Install only those components you need or want
- Ambari approach is expected to align with future direction of the Open Data Platform Initiative (the industry consortium dedicated to development of a common Hadoop core)

Introduction to IBM BigInsights © Copyright IBM Corporation 2015

Open source currency

Here is a closer look at what is in IBM's Open Platform for Apache Hadoop based on details available as of March 2015. IBM intends for its Open Platform to be as current as competitors in every package shipped.

Note that Apache Ambari is used as the installer. With Ambari, you download and install only those components that you want or need. Using this installation approach, should easily align with the direction of the Open Data Platform Initiative.

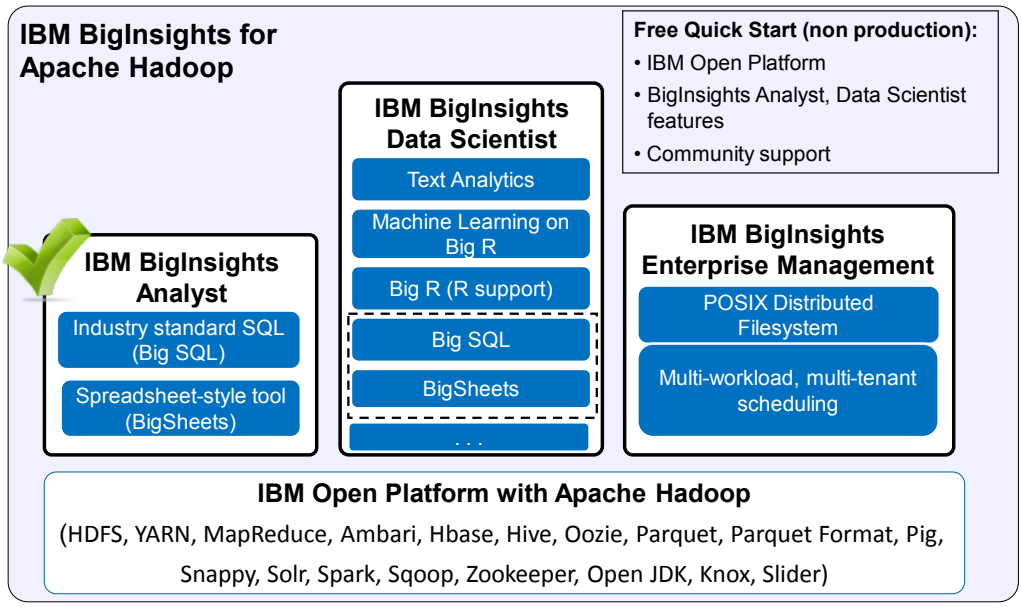
(In previous releases, IBM used their own installer and packaged everything together.)

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Overview of BigInsights



Introduction to IBM BigInsights © Copyright IBM Corporation 2015

Overview of BigInsights

Now that you understand the IBM Open Platform with Apache Hadoop, you will look at the IBM modules. To start, there is IBM BigInsights Analyst. BigSheets is one part of the Analyst offering, and Big SQL is another.

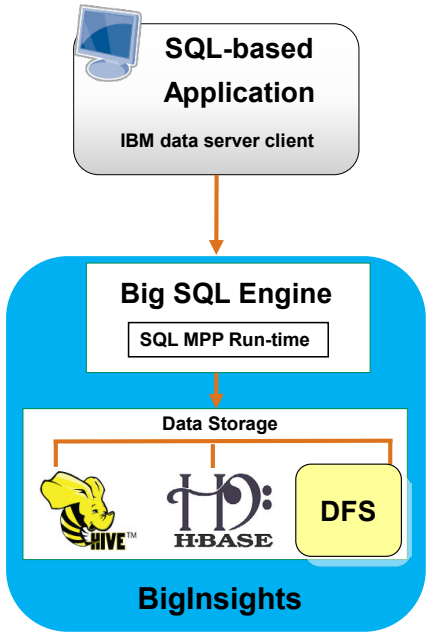
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

SQL for Hadoop (Big SQL)

- Comprehensive, standard SQL
  - SELECT: joins, unions, aggregates, subqueries . . .
  - GRANT/REVOKE, INSERT … INTO
  - Procedural logic in SQL
  - Stored procs, user-defined functions
  - IBM data server JDBC and ODBC drivers
- Optimization and performance
  - IBM MPP engine (C++) replaces Java MapReduce layer
  - Continuous running daemons (no start up latency)
  - Message passing allow data to flow between nodes without persisting intermediate results
  - In-memory operations with ability to spill to disk (useful for aggregations, sorts that exceed available RAM)
  - Cost-based query optimization with 140+ rewrite rules
- Various storage formats supported
  - Data persisted in DFS, Hive, HBase
  - No IBM proprietary format required
- Integration with RDBMSs via LOAD, query federation



Introduction to IBM BigInsights

© Copyright IBM Corporation 2015

SQL for Hadoop (Big SQL)

Big SQL is IBM's SQL processing engine for big data, and one of its great features is the comprehensive support for ISO SQL. This allows SQL professionals to model and work with Hadoop data in a familiar way. Big SQL uses Hadoop storage mechanisms for its tables, including HBase, DFS files, and the Hive warehouse.

Big SQL's runtime execution engine is all native code (C/C++), offering performance benefits over Java's MapReduce layer.

For common table formats a native I/O engine is utilized

- such as delimited, RC, SEQ, Parquet, …

For all others, a Java I/O engine is used

- maximizes compatibility with existing tables
- allows for custom file formats and SerDe's

All Big SQL built-in functions are native code.

Customer built UDx's can be developed in SQL, C++ or Java.

You can maximize performance without sacrificing extensibility.

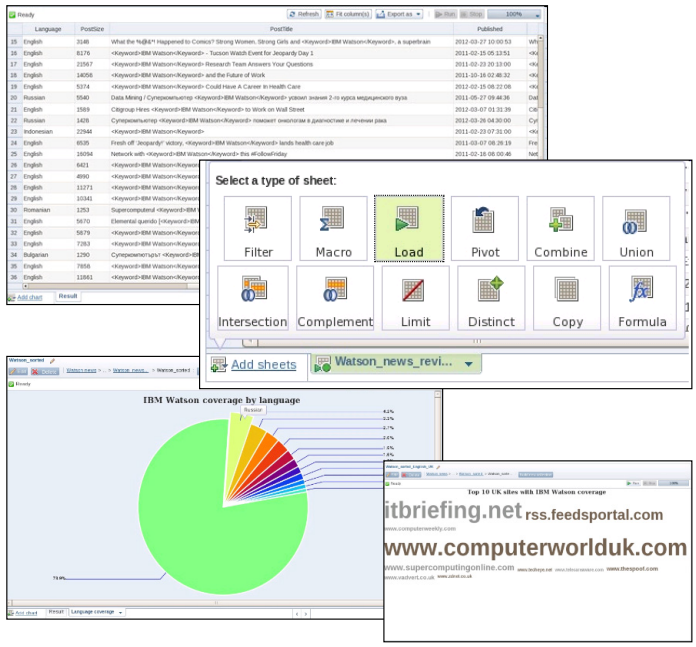
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Spreadsheet-style analysis (BigSheets)

- Web-based analysis and visualization
- Spreadsheet-like interface
  - Explore, manipulate data without writing code
  - Invoke pre-built functions
  - Generate charts
  - Export results of analysis
  - Create custom plug-ins
  - . . .



Introduction to IBM BigInsights

© Copyright IBM Corporation 2015

Spreadsheet-style analysis (BigSheets)

BigInsights Analyst provides a browser-based visualization and analysis tool designed to help non-programmers work with big data. The BigSheets tool contains many built-in functions to enable business users to filter, combine, manipulate, and explore their data. Optionally, users can generate charts from their workbooks or export their workbooks into popular formats, such as TSV, CSV, and others.

Characteristics of this tool:

- developed specifically for business intelligence and non-technical business users to facilitate data gathering and analysis
- able to work with structured and unstructured data
- able to combine data from different data sources so users can pinpoint opportunities and risks "hidden in the data"

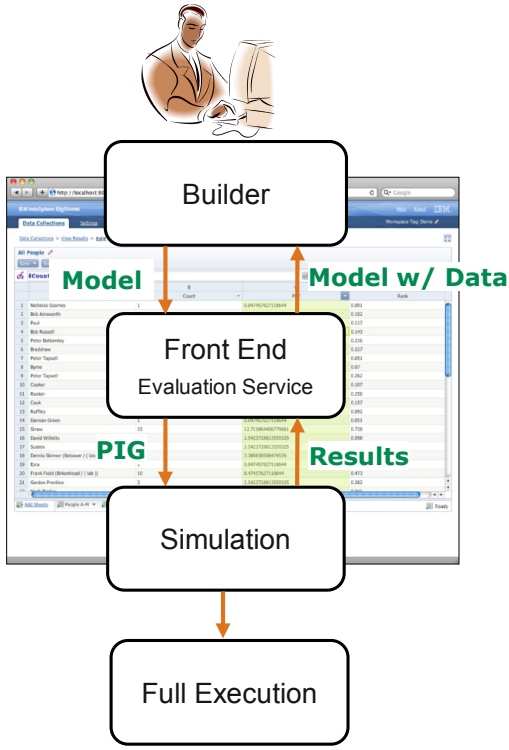
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Working with BigSheets

- Create workbook for data in DFS
- Customize workbook through graphical editor and built-in functions
  - Filter data
  - Apply functions / macros / formulas
  - Combine data from multiple workbooks
- "Run" workbook: apply work to full data set
- Explore results in spreadsheet format and/or create charts
- Optionally, export your data



© Copyright IBM Corporation 2015

Working with BigSheets

The BigSheets graphical interface enables users to create workbooks (tabular data models) and filter or transform the data as desired. Behind the scenes, the tool generates scripts as needed and executes the necessary work on a subset of the data (currently, a 50-row sample). This allows the user to iteratively explore various possibilities in an efficient manner. When satisfied, the user can "run" the workbook, which causes BigSheets to execute MapReduce jobs over the full set of data.

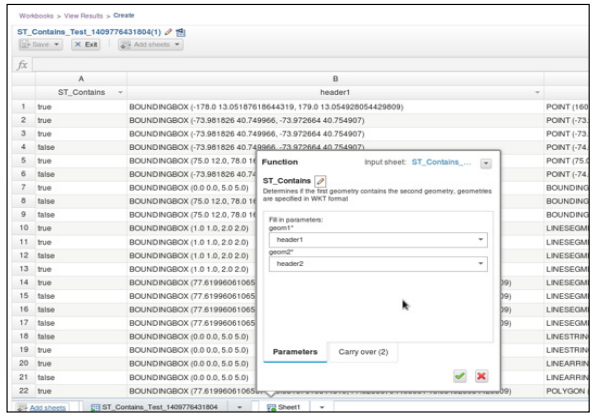
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

New geospatial capabilities in BigSheets

- Topological functions determine if regions contain, overlap, touch and more
- Metric functions measure area, distance and length
- Constructor and transformation functions convert data into different formats
- Helper functions validate data and count points, geometries and so forth



Introduction to IBM BigInsights

© Copyright IBM Corporation 2015

New geospatial capabilities in BigSheets

BigInsights v4 includes some new geospatial features in BigSheets.

Geospatial analytics adds another meaningful dimension to analysis. For example, it enables companies to analyze customer movement through a space, helps policemen to see patterns of crime locations, and helps municipalities to understand where people most often request taxicabs or other services.

Consistent with Streams: This is the same geospatial functionality that you can execute in your data on Streams.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

A partial list of built-functions includes:

- Point geometry: POINT (longitude latitude)
- LineSegment geometry: LINESEGMENT (start_longitude start_latitude, end_longitude end_latitude)
- LineString geometry: LINESTRING (p1_longitude p1_latitude, p2_longitude p2_latitude, ..., pn_longitude, pn_latitude)
- Polygon geometry (oriented polygon as per the left-hand side inclusion rule)
- ST_Area Computes the area of the provided geometry, is always 0.0 for 0-d and 1-d objects
- ST_Distance Computes the shortest distance in meters between given geometries in WKT format on the spherical earth model.
- ST_Equals Checks if the given geometries are equal
- ST_Intersects Computes if the two geometries intersect, resulting in either true or false as the output.

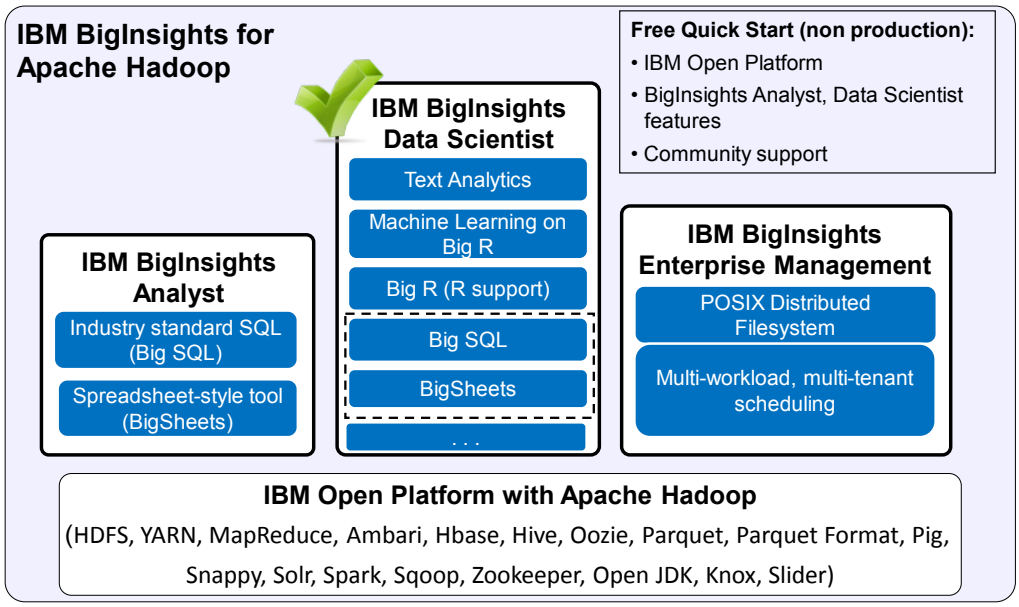
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Overview of BigInsights



Introduction to IBM BigInsights

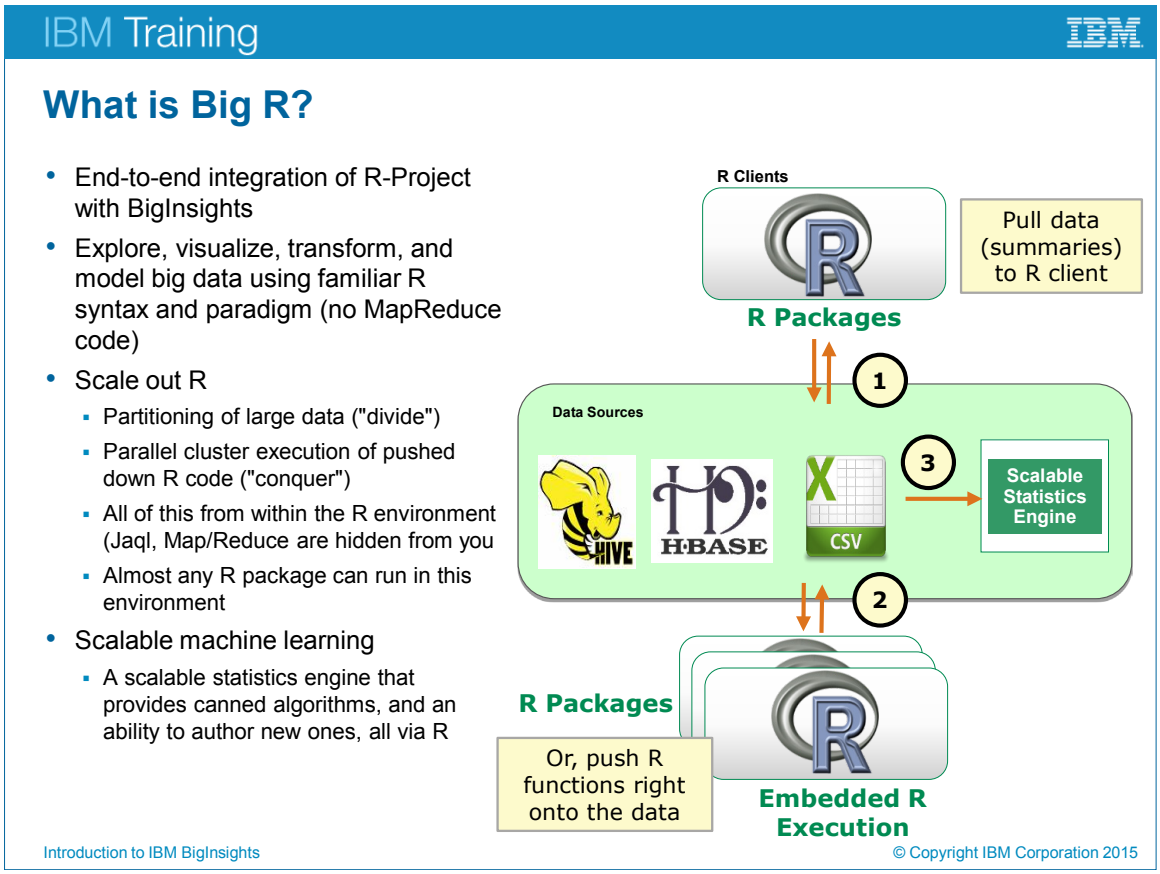
© Copyright IBM Corporation 2015

Overview of BigInsights

The IBM BigInsights Data Scientist module includes native support for the R programming language (Big R) and adds Machine Learning algorithms that are optimized for Hadoop. It also provides web-based tooling for text analysis. Each of these capabilities will be explored individually.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



What is Big R?

Big R is a library of functions that provide end-to-end integration with the R language and BigInsights. Big R can be used for comprehensive data analysis on your BigInsights cluster, hiding some of the complexity of manually writing MapReduce jobs.

Big R uses the open source R language to enable rich statistical analysis. You can use Big R to manipulate data by running a combination of R and Big R functions. Big R functions are similar to existing R functions but are designed specifically for analyzing big data.

Big R is:

- an R package: end-to-end integration of R into IBM BigInsights
- overloads a number of R primitives to work with big data

Big R's native support for open source R statistical computing helps clients leverage their existing R code or gain from more than 4,500 freely available statistics packages from the Open R community.

To learn more about R, create an account on the Big Data University website and take the course on R programming (http://bigdatauniversity.com/courses/course/view.php?id=522). The main R web site is http://www.r-project.org/.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

How Big R compares with other R solutions

- Other solutions: offer an R API for writing MapReduce from R.
- Example: Compute the mean departure delay for each airline on a monthly basis*.

RHadoop implementation

```
csvtextinputformat = function(line) keyval(NULL, unlist(strsplit(line, "\\,")))
deptdelay = function (input, output) {
    mapreduce(input = input,
        output = output,
        textinputformat = csvtextinputformat,
        map = function(k, fields) {
            # Skip header lines and bad records:
            if (!(identical(fields[[1]], "Year")) & length(fields) == 29) {
                deptDelay <- fields[[16]]
                # Skip records where departure delay is "NA"
                if (!(identical(deptDelay, "NA"))) {
                    # field[9] is carrier, field[1] is year, field[2] is month
                    keyval(c(fields[[9]], fields[[1]], fields[[2]]), deptDelay)
                }
            }
        },
        reduce = function(keySplit, vv) {
            keyval(keySplit[[2]], c(keySplit[[3]], length(vv), keySplit[[1]],
                mean(as.numeric(vv))))
        }
    )
}
from.dfs(deptdelay("/data/airline/1987.csv", "/dept-delay-month"))
```

Big R implementation

```
air <- bigr.frame(dataPath = "airline.csv",
                  dataSource = "DEL", na.string="NA")
summary(mean(DepDelay) ~ UniqueCarrier + Year + Month,
                  dataset = air)
```

RHIPE implementation

```
rhinit(TRUE, TRUE)
# Output from map is:
# "CARRIER|YEAR|MONTH \t DEPARTURE_DELAY"
map <- expression({
    # For each input record, parse out required fields and output new record:
    extractDeptDelays = function(line) {
        fields <- unlist(strsplit(line, "\\,"))
        # Skip header lines and bad records:
        if (!(identical(fields[[1]], "Year")) & length(fields) == 29) {
            deptDelay <- fields[[16]]
            # Skip records where departure delay is "NA":
            if (!(identical(deptDelay, "NA"))) {
                # field[9] is carrier, field[1] is year, field[2] is month:
                rhcollect(paste(fields[[9]], "|", fields[[1]], "|", fields[[2]],
                    sep=""),
                deptDelay)
            }
        }
    }
    # Process each record in map input:
    lapply(map.values, extractDeptDelays)
})
# Output from reduce is:
# YEAR \t MONTH \t RECORD_COUNT \t AIRLINE \t AVG_DEPT_DELAY
reduce <- expression(
    pre = {
        delays <- numeric(0)
    },
    reduce = {
        # Depending on size of input, reduce will get called multiple times
        # for each key, so accumulate intermediate values in delays vector:
        delays <- c(delays, as.numeric(reduce.values))
    },
    post = {
        # Process all the intermediate values for key:
        keySplit <- unlist(strsplit(reduce.key, "\\|"))
        count <- length(delays)
        avg <- mean(delays)
        rhcollect(keySplit[[2]],
        paste(keySplit[[3]], count, keySplit[[1]], avg, sep="\t"))
    }
)
inputPath <- "/data/airline/"
outputPath <- "/dept-delay-month"
# Create job object:
z <- rhmr(map=map, reduce=reduce,
    ifolder=inputPath, ofolder=outputPath,
    inout=c('text', 'text'), jobname='Avg Departure Delay By Month',
    mapred=list(mapred.reduce.tasks=2))
# Run it:
rhex(z)
```

*Dataset: "airline".

Scheduled flights in US 1987-2009.

Introduction to IBM BigInsights

© Copyright IBM Corporation 2015

How Big R compares with other R solutions

You might be wondering how Big R compares with other approaches to applying R to big data. Big R doesn't require programmers to write MapReduce applications to invoke R functions. As a result, it is compact and easy for R programmers to use. Some other implementations, such as those shown here, provide an R API for MapReduce. As you can see, such an approach typically requires more programming effort and skill.

RHIPE: www.datadr.org

Rhadoop: http://blog.revolutionanalytics.com/2011/09/mapreduce-hadoop-r.html

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Machine Learning with Big R

- Based on SystemML (IBM Almaden Research)
- Scalability for large data sets
- R API inspired by R's ML libraries

Big R functions	Inspired by R's	Algorithm
`bigr.lm()`	`lm()`	Linear regression
`bigr.glm()`	`glm()`	Generalized Linear Models
. . .	. . .	. . .
`bigr.kmeans()`	`kmeans()`	K-means clustering
`bigr.naive.bayes()`	`naiveBayes()`	Naïve Bayes classifier
`bigr.sample()`	`sample()`	Uniform sample by percentage, exact number of samples, or partitioned sampling.

© Copyright IBM Corporation 2015

Machine Learning with Big R

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Text Analytics

- Distills structured info from unstructured text
  - Sentiment analysis
  - Consumer behavior
  - Illegal or suspicious activities
  - …
- Parses text and detects meaning with annotators
- Understands the context in which the text is analyzed
- Features pre-built extractors for names, addresses, phone numbers, etc.

Introduction to IBM BigInsights © Copyright IBM Corporation 2015

Text analytics

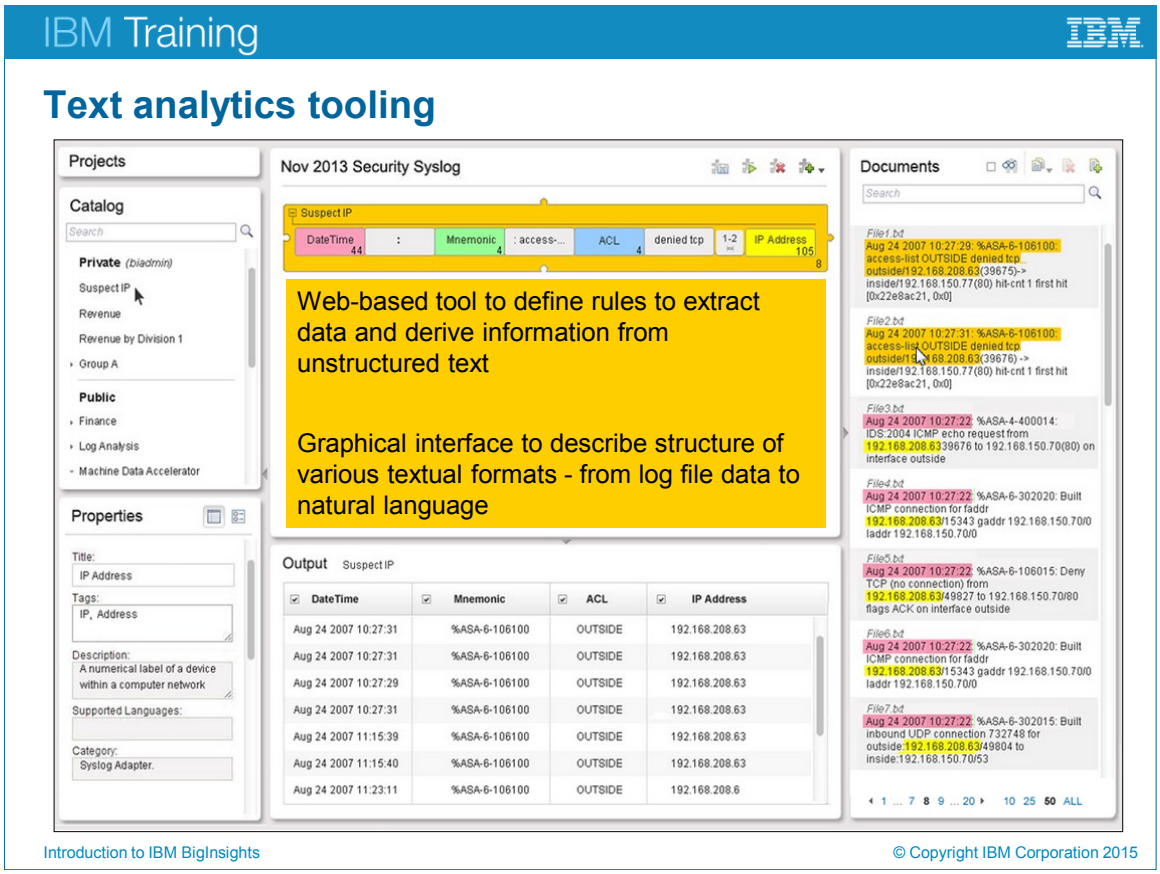
Of course, there's more to BigInsights Data Scientist than Big R. Text analytics is another capability included with this offering.

Key points:

- IBM research developed a sophisticated text analytics engine, similar technology to what was demonstrated in Watson and is now an integral part of BigInsights to identify meaning within unstructured text
- there are 100s of pre-built rules (annotators)
- the annotators are context sensitive and discover the relationship between terms even if they are separated by text
- it is built for top performance and has been optimized for BigInsights workloads

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Text analytics tooling

IBM's text analytics web tooling provides a highly interactive, easy to use interface that analysts and data scientists can use to quickly assemble text extractors. Its browser based and requires no programming, so you can be up and running in minutes.

The simple drag, drop and run interface includes result highlighting and grid views so you can quickly and iteratively assemble, run, review and refine the extractor.

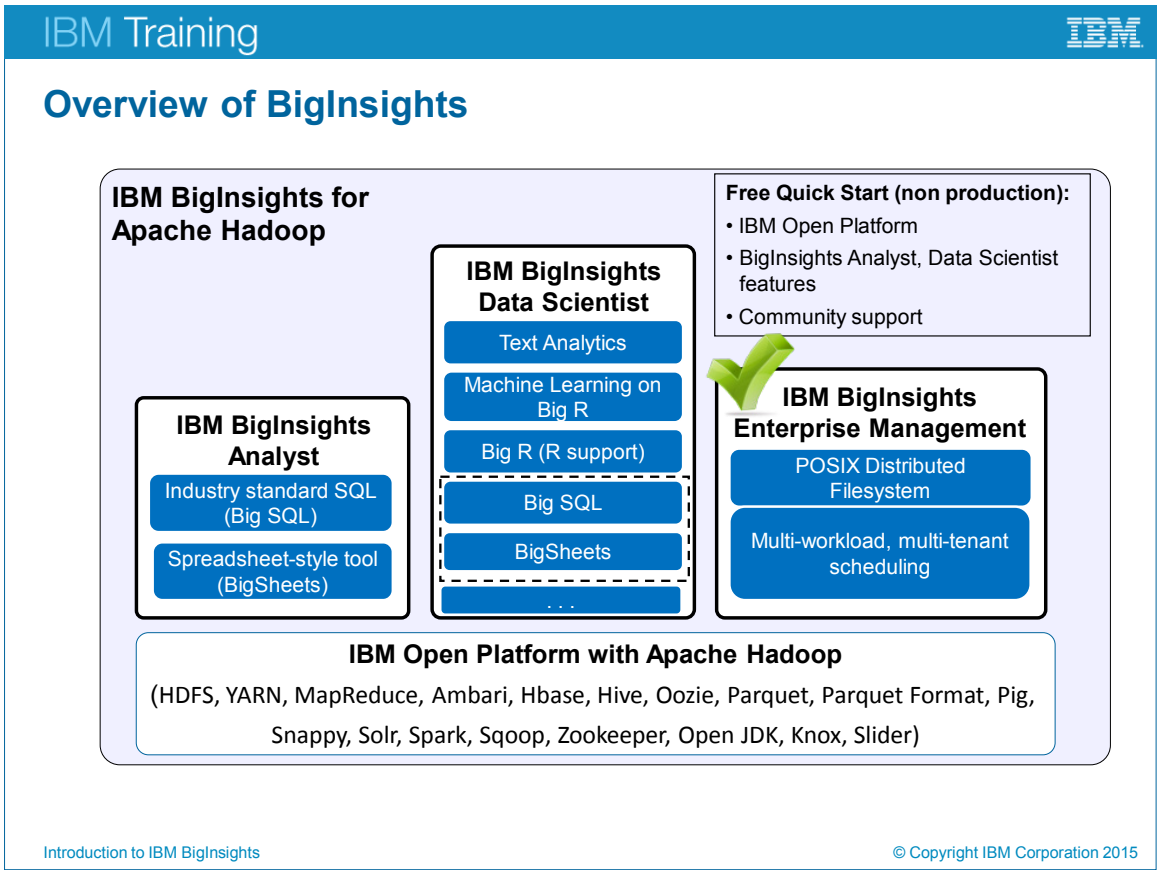
In addition to a small set of primitives (regular expression, dictionary and literal), the tool includes a large library of pre-built extractors. These cover the spectrum, including:

- simple extractors , such as numbers and capitalized words
- complex extractors, such as a "named entity" library for people, places, and organizations
- domain-specific extractors for financial services, machine data and sentiment.

The simple GUI leverages the power of IBM's tried and tested strategic information extraction technology, System T, to deliver a powerful, scalable, robust solution for information extraction.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Overview of BigInsights

The final module, IBM BigInsights Enterprise Management, helps administrators manage, monitor and secure their Hadoop distribution. BigInsights Enterprise Management introduces tools to allocate resources, monitor multiple clusters, and optimize workflows to increase performance.

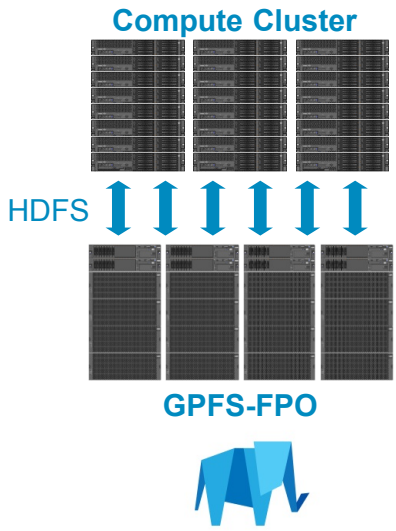
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

IBM GPFS: HDFS alternative

- Drop-in replacement for HDFS
- No need for dedicated analytics infrastructure
  - Cost savings
  - No need to move data in and out of an analytics dedicated silo
  - Software defined infrastructure for multi-tenancy
- Fast parallel file system
- POSIX & HDFS semantics
- Hadoop & non-Hadoop apps
- Built-in high availability



Introduction to IBM BigInsights

© Copyright IBM Corporation 2015

IBM GPFS: HDFS alternative

BigInsights Enterprise Management includes a POSIX-compliant, distributed file system for running Hadoop and non-Hadoop workloads. GPFS-FPO brings a proven distributed file system to the Hadoop environment. Its distributed metadata feature eliminates any single point of failure that can bring your analytics capabilities to a halt. POSIX enables regular file compliance. POSIX is a regular file system, like what you'd have on a Windows machine. Currently, Hadoop uses a file system that you can only access with Hadoop APIs. POSIX compliance support allows you to use all the commands and tools you would use normally in Windows/Unix.

GPFS is proven in dealing with large numbers of files and enables mixing of multiple storage types. Whether you are ingesting large unstructured social media feeds, or machine-to-machine real time data in smart grid systems, GPFS can easily manage this at speed.

GPFS also addresses one of the main concerns surrounding big data: security. It offers file and disk level access control so only those applications, or data itself can be isolated to privileged users, applications or even physical nodes in highly secured physical environments if needed.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

What is special about GPFS?

- filesystem built 12 years ago for HPC
  - HPC = High Performance Computing
- very highly regarded
- distributed file system
  - same data accessible from different computers
- runs on most common platforms (AIX, Linux, Windows)

What is expected from GPFS?

- high scale, high performance, high availability, data integrity
- POSIX semantics
- workload isolation
  - logical isolation: filesets are separate filesystems inside a filesystem
  - physical isolation: filesets can be put in separate storage pools
- enterprise features (quotas, security, ACLs, snapshots, etc.)

GPFS, a better file system:

- GPFS-FPO = GPFS + shared nothing support
- support for shared nothing clusters = local disks instead of a SAN

Why is GPFS highly available?

- metadata is replicated just like data
  - HDFS does not replicate metadata as such
  - HDFS makes a copy for the secondary NameNode

Data availability: replication

- same as HDFS

What is better than HDFS?:

- GPFS allows concurrent read and write by multiple programs
  - HDFS allows concurrent read but only one writer

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Because GPFS is a Posix-compliant filesystem, there is no need to have a separate local filesystem:

- HDFS forces one to decide up-front how much disk space to allocate to local and HDFS filesystems
- standard applications cannot use HDFS but they can use GPFS:
  - for example: Lucene, B-tree indexes, ftp, applications that need updating, etc.

No single point of failure at no extra cost or hassle:

- HDFS has a single point of failure in the NameNode.
- HDFS NameNode is typically run in a separate HA environment

Metadata doesn't need to get read into memory before the filesystem is available:

- big HDFS systems might take a long time to read all the metadata in memory
- if HDFS doesn't have enough memory, it will not start

Excerpt follows from a white paper:

Unlike HDFS, GPFS FPO is a true POSIX file system that can be directly read from and written to by any software running on the Linux operating environment. This makes it easier to manage and much more flexible since data can be shared by a more diverse set of applications. You also eliminate the need to employ specialized Hadoop commands because the distributed file system can be manipulated using standard OS level commands and utilities.

GPFS makes it possible to tailor data storage in a way that makes sense economically depending on the application and characteristics of data being stored. For example, one volume may be configured to use the file-placement option (FPO) with fast but relatively expensive n-way block replication in a shared-nothing environment. Other data sets may be shared using striping with more traditional parity schemes that require less overhead in terms of storage. With GPFS, organizations have the option of storing the data in the manner that best meets the needs of the business.

GPFS also addresses the archiving and backup challenges inherent in HDFS file systems. With advanced features such as policy-based archiving, data can be migrated automatically to appropriate tiers of storage based on access patterns and the relative cost and performance characteristics of storage. For example, users can set up a policy that says, "If nobody has touched this data in the last year, then automatically migrate the data to secondary storage; and if nobody has touched the data in three years, then archive it to tape." So as the data ages, it is automatically migrated to less-expensive storage, helping organizations automate the lifecycle management and sharing of data.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Unlike HDFS, which is optimized for large-block I/O, GPFS is flexible enough to support a variety of different access patterns, including applications with small or medium-size blocks as well as write-intensive applications. This means GPFS can provide better performance across a wider range of applications. As a specific example, a customer may be running an SAS workload, including a series of ETL-related steps to manipulate data on a shared GPFS file system. At a particular stage in the ETL workflow, a MapReduce program may be the most efficient way to process a specific data set. Because GPFS data can be accessed by both MapReduce and non-MapReduce workloads, the MapReduce job can be incorporated into the broader ETL workflow, executing against the same GPFS resident data, to avoid the time and cost associated with migrating data from one file system to another.

Also, the grid manager itself can support multiple workload patterns at the same time, which helps reduce the cost of infrastructure. Traditional batch workloads (using the IBM Platform LSF® batch scheduler) can coexist with service-oriented workloads and Hadoop MapReduce workloads (best implemented on Platform Symphony) across a common resource orchestration layer so that all workloads can share the same physical infrastructure.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

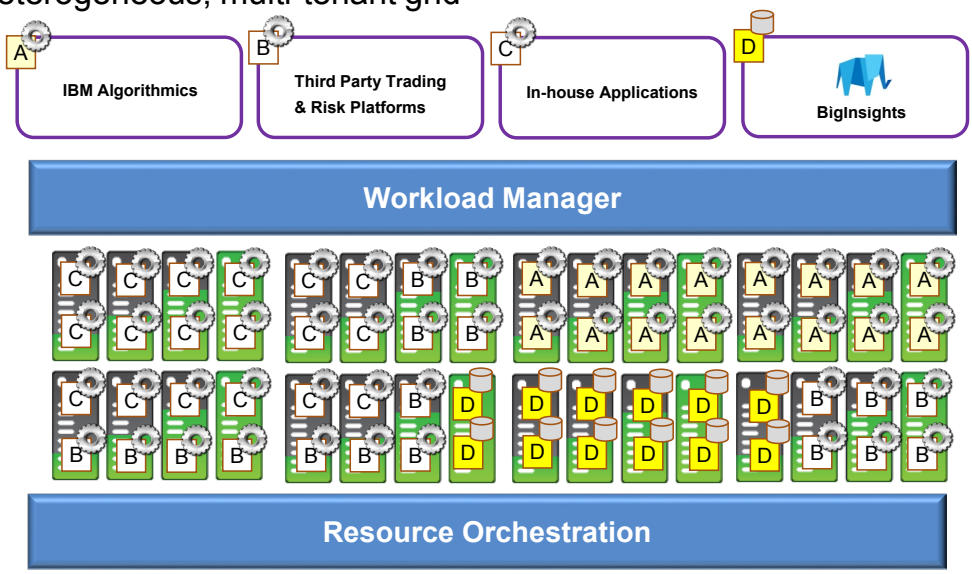
© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Platform Symphony

- Multiple users, applications and lines of business on a shared, heterogeneous, multi-tenant grid



Introduction to IBM BigInsights © Copyright IBM Corporation 2015

Platform Symphony

Platform Symphony (formerly known as Adaptive MapReduce) replaces the traditional MapReduce layer, allowing a computing cluster to support many different types of applications. It includes IBM-specific features for workload management, resource orchestration, and high performance.

- A heterogeneous grid management platform.
- A high-performance SOA middleware environment.
- Supports diverse compute & data intensive applications:
  - ISV applications
  - in-house developed applications (C/C++, C#/.NET, Java, Excel, R, etc.)
  - optimized low-latency Hadoop compatible run-time
  - can be used to launch, persist and manage non-grid aware application services
- Reacts instantly to time critical-requirements.
- A multi-tenant shared services platform with unique resource sharing capabilities.
- A limited-use run-time for Platform Symphony (called Adaptive MapReduce) included in BigInsights Enterprise Management.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Excerpt from white paper:

Using Platform Symphony to solve scheduling and multi-tenancy issues: By using Platform Symphony in conjunction with BigInsights, enterprises can consolidate clusters while increasing their utilization and performance. Providing a low-latency scheduling environment for heterogeneous workloads helps dramatically accelerate applications and enables IT groups to enhance the efficiency of existing resources. While it is possible to use general-purpose grid managers to support many of these workload patterns, workload management specialists know that to maximize efficiency and resource usage, workload managers need to be optimized to the specific workload pattern being supported. From a scheduling perspective, the series of steps that comprise a MapReduce workflow are just another type of distributed computing workload.

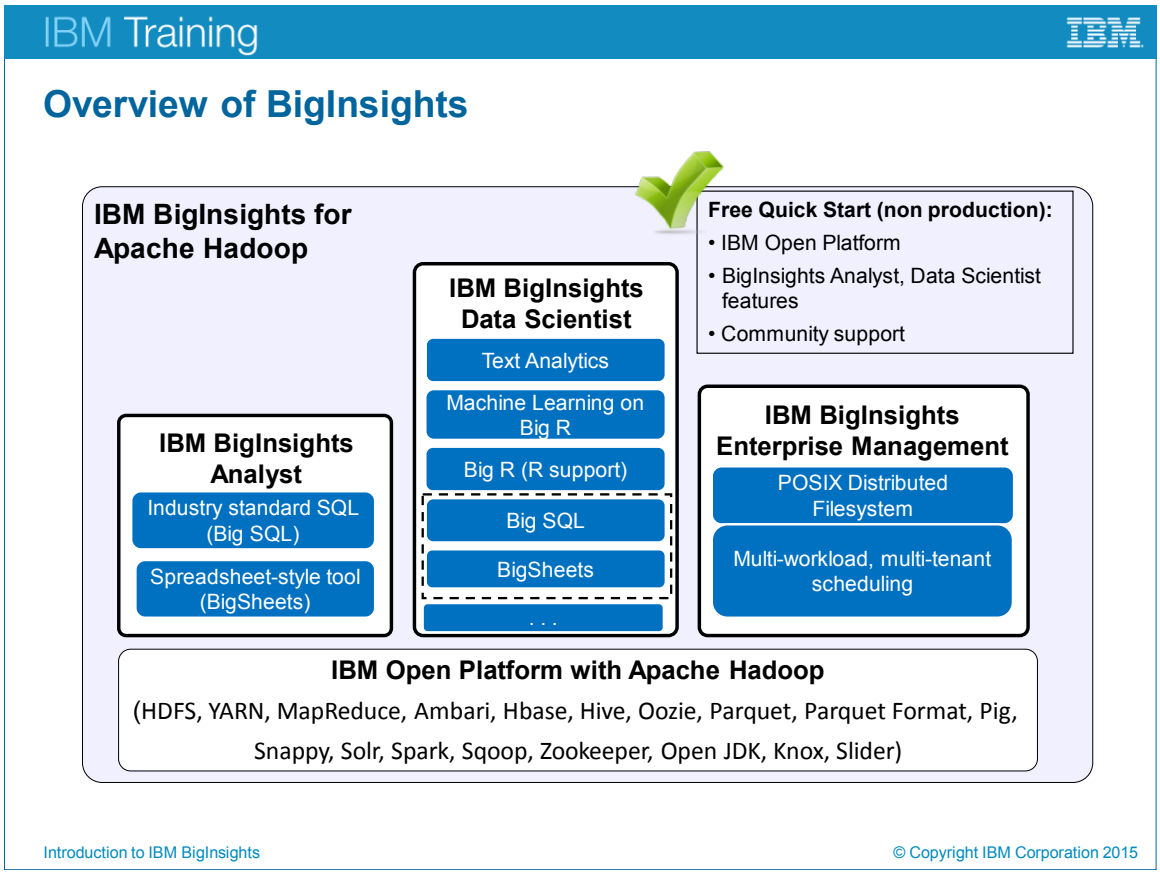
Fortunately, there are ample opportunities to optimize performance and efficiency in a way that is transparent to higher-level application frameworks that rely on distributed computing services. Platform Symphony was purpose-built as a low-latency scheduler, which means jobs can be started instantly with turnaround times in milliseconds. As a result, big data applications start faster and run faster on Platform Symphony, providing organizations with a critical competitive advantage.

In an audited benchmark conducted by a third-party testing firm, Platform Symphony - Advanced Edition was found to accelerate various social media workloads by factors ranging from 40 percent to 9.9 times, with the average workload running fully 7.3 times faster. The sophisticated MapReduce engine in Platform Symphony delivers higher performance compared with other Hadoop implementations for many different workloads, especially those requiring quick responses, as opposed to long-running batch workloads. While Hadoop clusters normally run one job at a time, Platform Symphony is designed for concurrency, allowing up to 300 job trackers to run on a single cluster at the same time with agile reallocation of resources based on real-time changes to job priorities. This means organizations can ensure that time-critical workloads get done quickly, with priorities being adjusted at runtime. Organizations with mixed workloads also can consolidate clusters running Hadoop and non-Hadoop workloads into a single multi-tenant cluster with Platform Symphony. This translates into better application performance, better utilization and an ability to respond quickly to business-critical demands.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

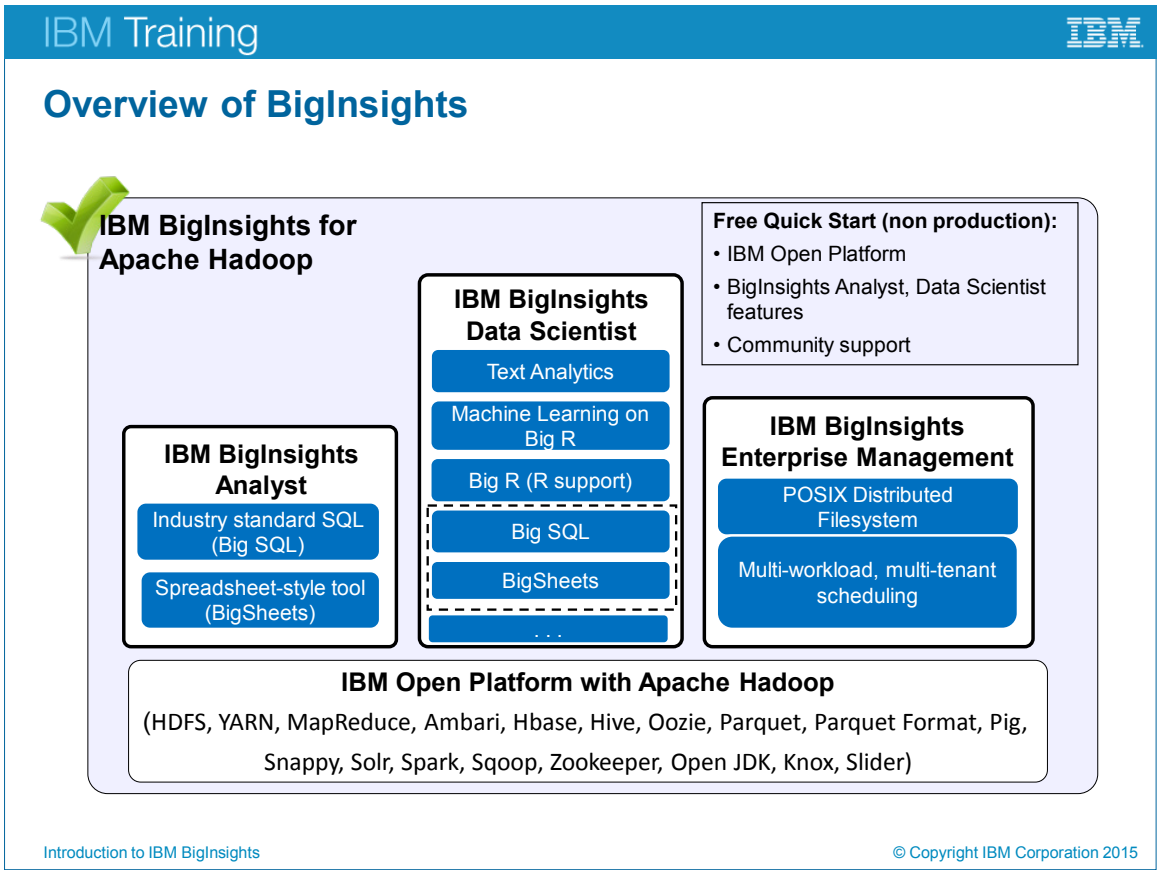


Overview of BigInsights

IBM also has a free offering for non-production use called the Quick Start edition. It includes the IBM Open Platform as well as most features of BigInsights Analyst and Data Scientist.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Overview of BigInsights

One last packaging aspect to review is the full production package called IBM BigInsights for Apache Hadoop. This includes all of the open source components, all of the IBM added value components in the Analyst, Data Scientist, and Enterprise Management modules, and a collection of limited use licenses to other IBM offerings, such as Cognos, Watson Explorer, InfoSphere Streams, and Data Click.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

IBM BigInsights for Apache Hadoop Offering Suite

IBM BigInsights v4	BigInsights Quick Start Edition	IBM Open Platform with Apache Hadoop	Elite Support for IBM Open Platform with Apache Hadoop	BigInsights Analyst Module	BigInsights Data Scientist Module	BigInsights Enterprise Management Module	BigInsights for Apache Hadoop
Apache Hadoop Stack: HDFS, YARN, MapReduce, Ambari, Hbase, Hive, Oozie, Parquet, Parquet Format, Pig, Snappy, Solr, Spark, Sqoop, Zookeeper, Open JDK, Knox, Slider	✔	✔	✔	*	*	*	✔
Big SQL - 100% ANSI compliant, high performant, secure SQL engine	✔			✔	✔		✔
BigSheets - spreadsheet-like interface for discovery & visualization	✔			✔	✔		✔
Big R - advanced statistical & data mining	✔				✔		✔
Machine Learning with Big R - machine learning algorithms apply to Hadoop data set	✔				✔		✔
Advanced Text Analytics - visual tooling to annotate automated text extraction	✔				✔		✔
Enterprise Mgmt - Enhanced cluster & resource mgmt & POSIX-compliant file systems						✔	✔
Governance Catalog				✔	✔		✔
Cognos BI, InfoSphere Streams, Watson Explorer, Data Click							✔

* Paid support for IBM Open Platform with Apache Hadoop required for BigInsights modules

Introduction to IBM BigInsights

© Copyright IBM Corporation 2015

IBM BigInsights for Apache Hadoop Offering Suite

Here's a summary of what IBM provides. The Quick Start edition includes the IBM Open Platform stack and nearly everything that you'll find in the Analyst and Data Scientist modules.This is a free offering for non-production use only.

The remaining offerings are available for production use. These include the IBM Open Platform (100% open source stack) as well as IBM Elite Support for this offering. These two items are shown in the yellow columns. The blue columns detail what you'll find in the Analyst, Data Scientist, and Enterprise Management modules. Finally, the dark blue column shows the full breadth of what you can buy with our broadest BigInsights package. That includes everything in the open source and IBM-specific modules, as well as additional limited use licenses for certain IBM products.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Pricing & licensing

- Community support via Hadoop Dev
  - Over 100,000 visitors since inception
  - Modeled after StackOverflow, most popular developer Q&A site on web

Products	BigInsights Quick Start	IBM Open Platform with Apache Hadoop	Elite Support for IBM Open Platform with Apache Hadoop	BigInsights Analyst Module	BigInsights Data Scientist Module	BigInsights Enterprise Management Module	BigInsights for Apache Hadoop
Pricing Terms	Free	Free	Yearly Subscription Only	Perpetual or Monthly License			
Support provided	Community	Community	IBM 24x7 support				
Usage License	Non-production, five node cap	Production Usage					
Pricing Model	Free	Free	Node based pricing				
Access via	ibm.com/hadoop		Passport Advantage				

© Copyright IBM Corporation 2015

Pricing and licensing

Here is a summary of the pricing and licensing model to review what is free and what is fee-based.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Cloud deployment options

Developer sandbox

Analytics for Hadoop

- Prototype, demo, trial in the cloud
- Empowers developers to rapidly drive insight from all data
- Adds Hadoop-based analytics to your application
- Enterprise features - BigSheets, Big SQL, Text analytics, HiveQL, HttpFS
- Delivered via IBM BlueMix

http://bluemix.net

Production environment

Enterprise Hadoop as a Service

- For Production, deployments at scale in the cloud
- Delivers flexibility and efficiency with subscription pricing
- Scales to meet spikes in demand without on-premise infrastructure
- Drives enterprise-class, complex analytics on big data sets
- Available via the IBM Cloud Marketplace and Bluemix

http://www.ibm.com/cloud

http://www.bluemix.net

Introduction to IBM BigInsights

© Copyright IBM Corporation 2015

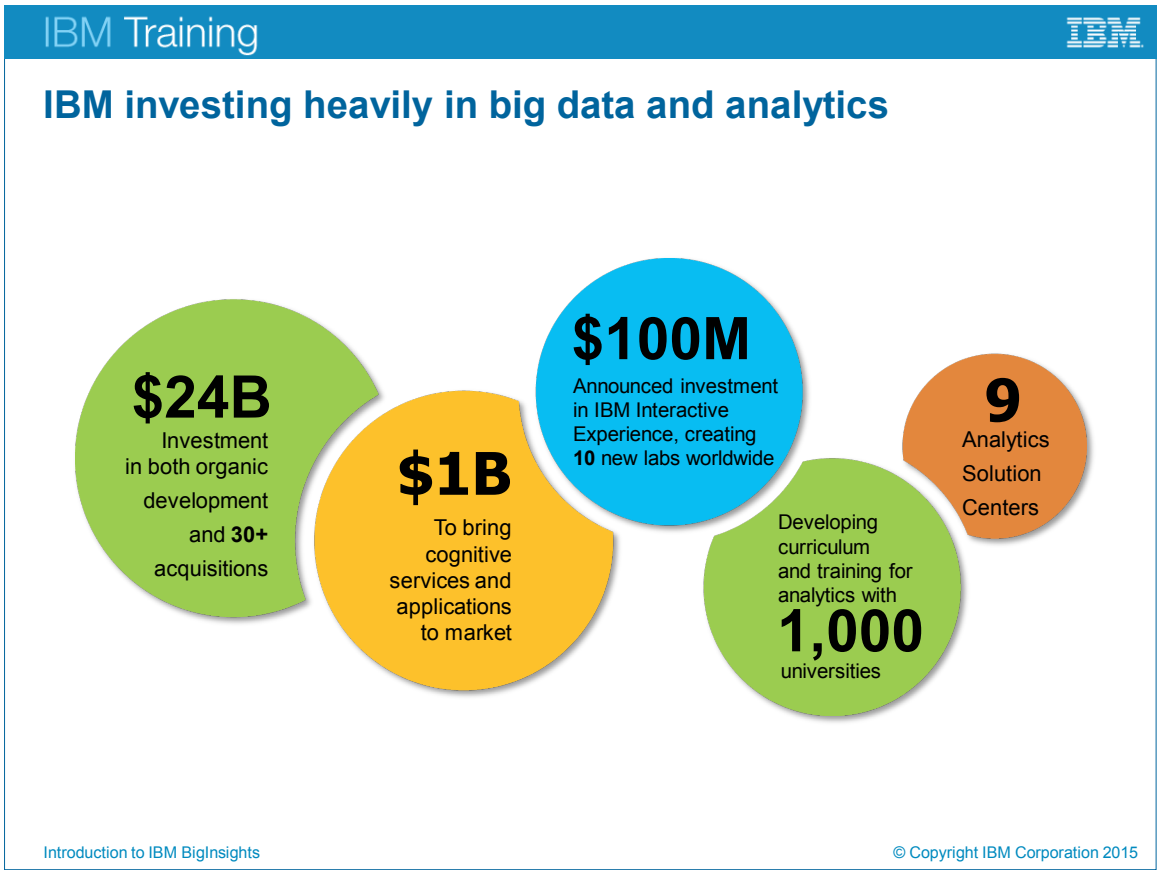
Cloud deployment options

BigInsights for the cloud allows organizations to respond faster to changing business environments by analyzing larger volumes of data more cost-effectively.

- BigInsights on IBM Cloud marketplace- BigInsights available with manual provisioning to SoftLayer (or other hosting platform) on PAYG basis, giving users the ability to pay for what they use while customizing the offering
- Analytics for Hadoop powered by BigInsights on IBM codename: BlueMix- BigSheets, BigSQL, HiveQL and HttpFSservices. The Analytics for Hadoop service adds Hadoop-based analytics to your application, with advanced enterprise capabilities. This service is powered by InfoSphere BigInsights for large-scale, complex analytics over big data sets.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



IBM investing heavily in big data and analytics

By now, you've probably realized that IBM is betting big on big data. If you want to know how big, look at the numbers on this chart.

$24 Billon investment in development and acquisitions.

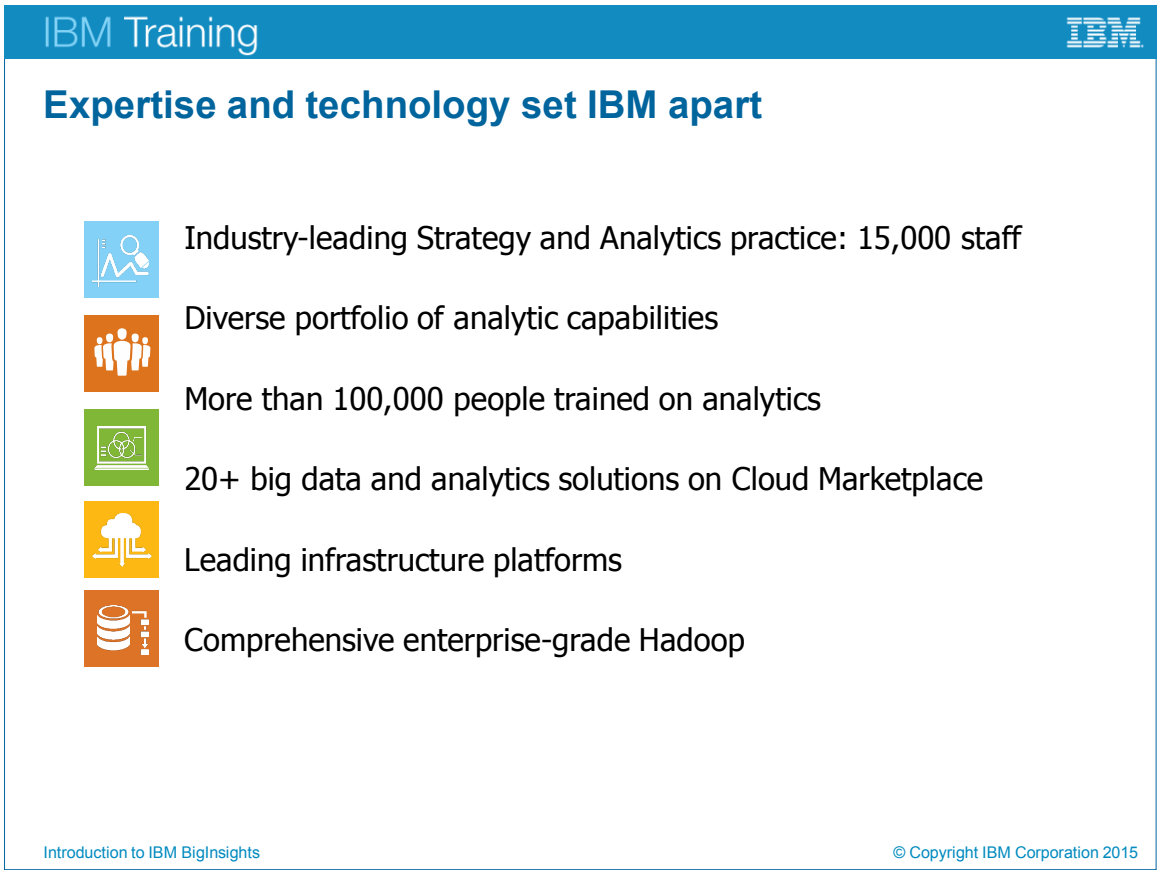
9 Analytics Solution Centers: IBM Business Partners can use the Analytic Solution Centers to conduct client meetings and request assistance from IBM with Proof of Concepts for their clients. IBM will frequently run client facing events in the Analytic Solution centers and invite Business Partners to participate. IBM Innovation centers are also available to Business Partners around the world: https://www-304.ibm.com/partnerworld/wps/servlet/ContentHandler/isv_com_tsp_iic_overview

$100M investment in IBM Interactive Experience is really a GBS momentum but business partners (BPs) are now networking with GBS for assistance with services they can't provide to their clients. We are seeing more networking with GBS by big data and analytics BPs.

Big data and analytics training is available through the Big Data University and is linked from Big Data & Analytics on PartnerWorld. Business Partners are building strategic relationships with universities to expand their skill base. Kingland and Serwise are two BPs who are actively working with universities on projects and to recruit new employees.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Expertise and technology set IBM apart

Key point: IBM's expertise and technology is rich and deep.

IBM has built differentiated strength that other vendors can only dream about. The strength of our expertise in 15k strong consulting, the capabilities in our technology (software, solutions, hardware), and our investment in cloud and in skills to benefit our clients is unmatched in market.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Background:

- When it comes to what sets IBM apart, they have rich offerings that other vendors can only dream about.
- From our expertise in consulting to our solutions, to our advances in technology that address context computing, streams, advanced analytics and cognitive, and infrastructure that addresses the compute intensive computing; there is not one vendor out there than can deliver the depth we can to our clients.
- The reason why we have invested in this breadth and depth of technology, expertise and reach is because our clients require it to be successful capitalizing on the competitive advantage of data.
- We understand the value of data as a new natural resource and the need for individuals and organizations to exploit it and put it to work.
- The market shifts we see creating the opportunity to realize the potential of the portfolio we have, and our investment areas establish us for the future.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Unit summary

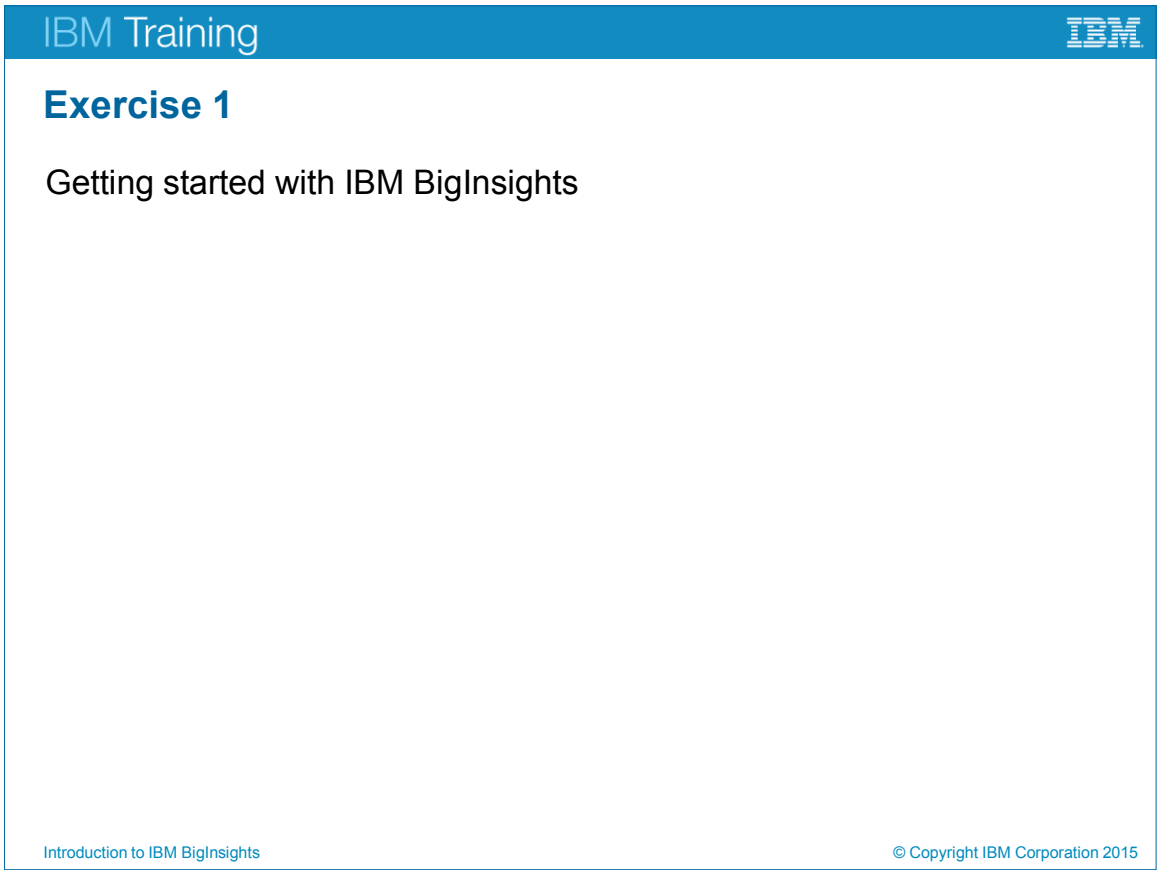
- Describe the functions and features IBM BigInsights
- List the IBM value-add components that comes with BigInsights
- Give a brief description of the purpose of each of the value-add components

© Copyright IBM Corporation 2015

Unit summary

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Exercise 1: Getting started with IBM BigInsights

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Exercise 1: Getting started with IBM BigInsights

Purpose:
You will learn more about the file system and directory structures of the IBM value-adds that are available with IBM BigInsights, and begin working with basic Hadoop commands.

User/Password: **biadmin/biadmin**
root/dalvm3
Services Password: **ibm2blue**

Important: Before doing this exercise, ensure that your access and services are configured and running. Check that:

- /etc/hosts displays your environment's IP address
- in the Ambari console, ensure that all BigInsights services are running

If you are unsure of the steps, please refer to Unit 1, Exercise 1 to ensure that your environment is ready to proceed. You should review the steps in Task 1 (Configure your image) and Task 2 (Start the BigInsights components).

Task 1. Navigate the file system.

In this task, you will get a brief overview of the file system.

1. To open a new terminal, right-click the desktop, and then click **Open in Terminal**.

   There are two main directories of where all of the BigInsights and the IBM Open Platform (IOP) components are installed.
2. To navigate to the IBM value-adds directory, type the following:
   `cd /usr/ibmpacks`
3. Type `ls` to see a listing of the components that are currently installed in the VM.

   Each of those directories contains specific functions related to it. There will be a directory for each of the components installed (such as bigr, bigsheets, etc.)
4. Navigate to **/usr/ibmpacks/bin**.

   This is where the scripts reside to remove the value-adds from the IOP stack. This is useful to know, if you no longer need any of the services and want to save space and memory.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

5. Navigate to **/usr/ibmpacks/current**.

   This directory links to the current releases of those value-add components.
6. To navigate to the **IOP directory**, where you can access the Apache stack containing the open source components, type the following:

   ```
   cd /usr/iop/current
   ```

   This is where you navigate to if you want to use the IOP components. You will use these components in the IOP section of this course.
7. Close the terminal.

Task 2. Working with basic Hadoop commands.

In this task, you will get a brief overview of the file system.

1. To open a new terminal, right-click the desktop, and then click **Open in Terminal**.
2. To switch to the root user, type **`su -`**, and then type the password **dalvm3**.
3. Create the **biadmin** folder on the **hdfs** under **/user**:

   ```
   su - hdfs -c "hdfs dfs -mkdir -p /user/biadmin/"
   ```
4. Change the ownership of the folder to **biadmin**:

   ```
   su - hdfs -c "hdfs dfs -chown -R biadmin /user/biadmin"
   ```
5. To log out of the root user, type **`exit.`**
6. Do a listing of the **/user** directory to see that the **biadmin** directory has been created.

   ```
   hdfs dfs -ls /user
   ```

   On your local system, in the /home/biadmin folder, is a labfiles directory. In this directory are some of the data files that you will be using throughout this exercise.
7. Navigate to the **/home/biadmin/labfiles** directory, and do a listing to see the files.
8. To upload the **Pride_and_Prejudice.txt** file in to the HDFS, type the following:

   ```
   hdfs dfs -put /home/biadmin/labfiles/Pride_and_Prejudice.txt
   /user/biadmin
   ```
9. To see a listing of the **/user/biadmin** directory, and the uploaded file on the hdfs, type the following:

   ```
   hdfs dfs -ls /user/biadmin
   ```

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

10. To view the contents of the file, type the following:

    ```
    hdfs dfs -cat /user/biadmin/Pride_and_Prejudice.txt
    ```

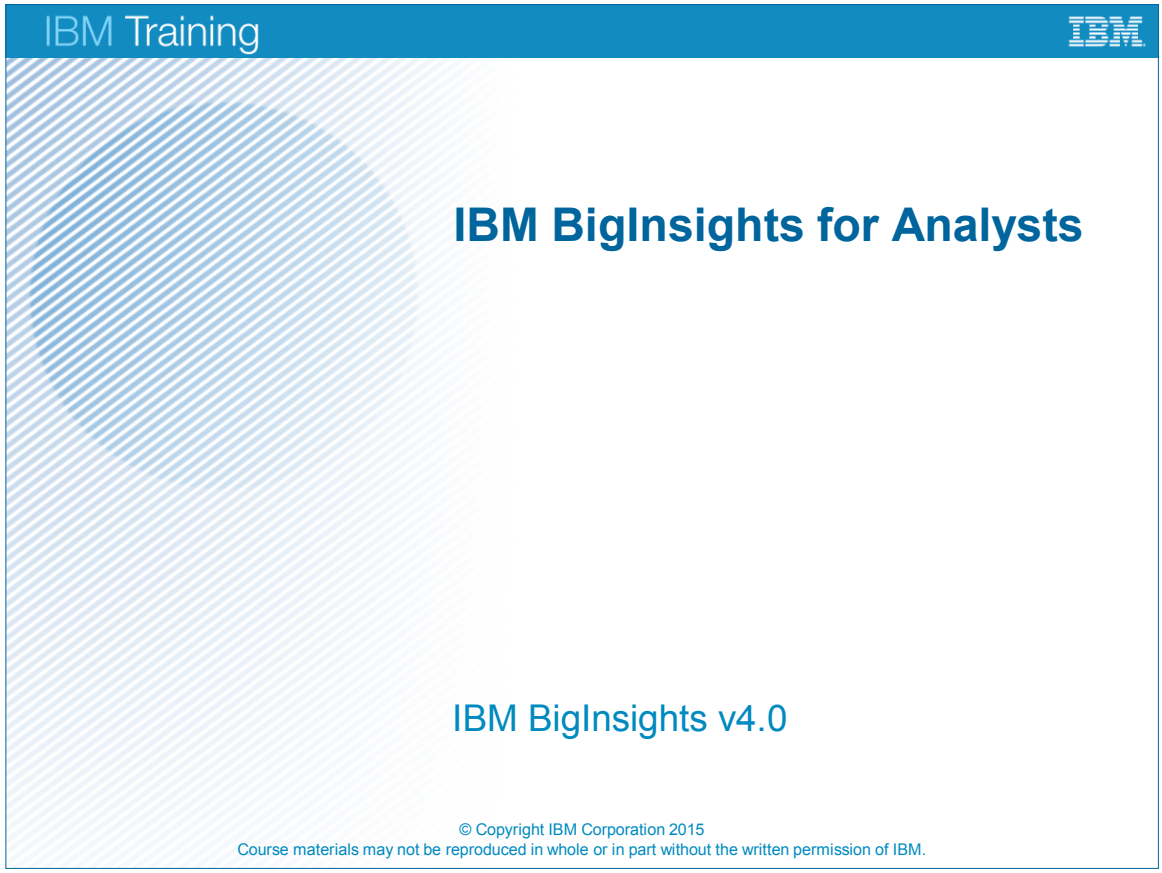
    You are not going to do anything else with that file now. The purpose of this exercise was to introduce you to some basic HDFS commands. They are similar, if not exactly the same as common Linux commands. You will work more with Hadoop commands in an upcoming exercise.

Results:
You have learned more about the file system and directory structures of the IBM value-adds that are available with IBM BigInsights, and you began working with basic Hadoop commands.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Unit objectives

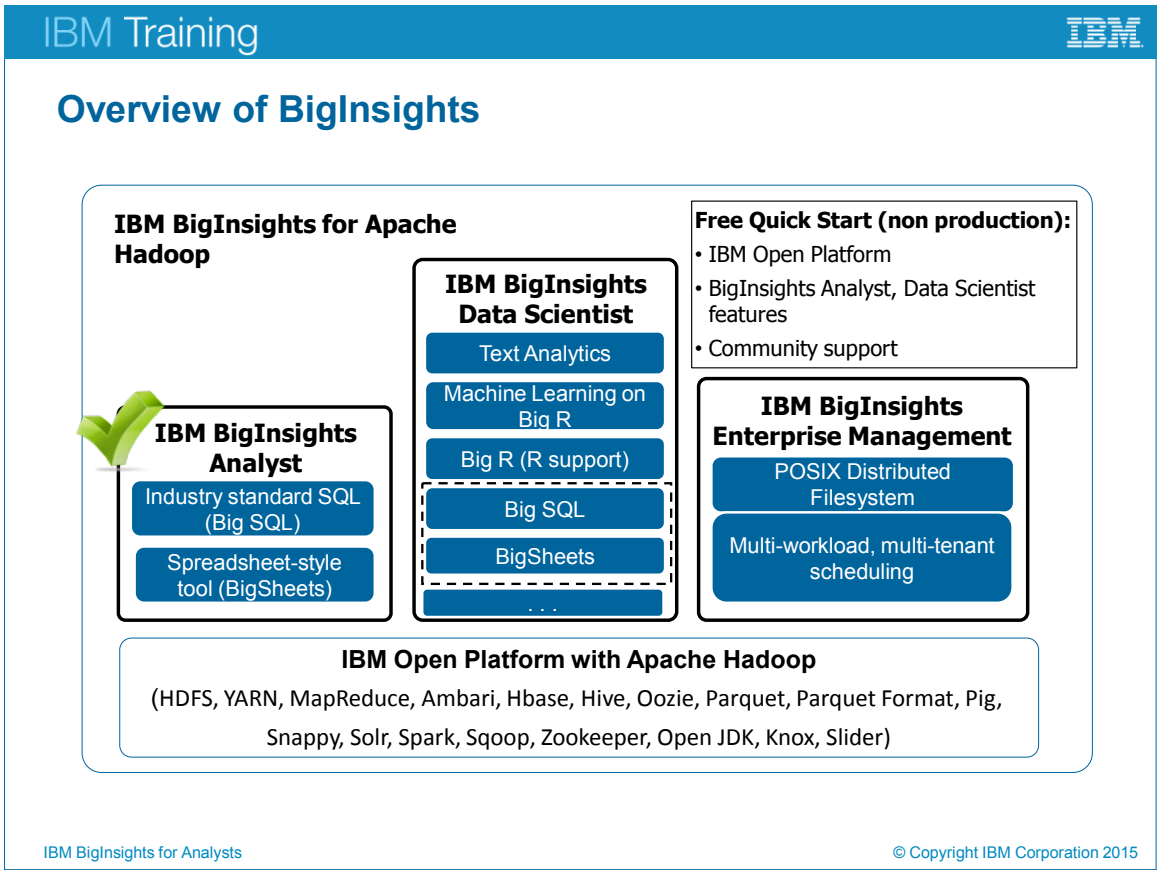
- Describe the components that come with the IBM BigInsights Analyst module
- Explain the benefits of using Big SQL for big data
- Understand the purpose of BigSheets

© Copyright IBM Corporation 2015

Unit objectives

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Overview of BigInsights

This unit will cover the IBM value-adds that comes with the IBM BigInsights Analysts module. A brief overview was provided earlier in the unit on BigInsights; more detail will be presented in this unit.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Executive Summary

- What is Big SQL?
  - Industry-standard SQL query interface for BigInsights data
  - New Hadoop query engine derived from decades of IBM R&D investment in RDBMS technology, including database parallelism and query optimization
- Why Big SQL?
  - Easy on-ramp to Hadoop for SQL professionals
  - Support familiar SQL tools / applications (via JDBC and ODBC drivers)
- What operations are supported?
  - Create tables / views. Store data in DFS, HBase, or Hive warehouse
  - Load data into tables (from local files, remote files, RDBMSs)
  - Query data (project, restrict, join, union, wide range of sub-queries, wide range of built-in functions, UDFs, etc.)
  - GRANT / REVOKE privileges, create roles, create column masks and row permissions
  - Transparently join / union data between Hadoop and RDBMSs in single query
  - Collect statistics and inspect detailed data access plan
  - Establish workload management controls
  - Monitor Big SQL usage
  - etc.

© Copyright IBM Corporation 2015

Executive Summary

In this presentation, you will be introduced to IBM's Big SQL technology in BigInsights Analyst and BigInsights Data Scientist. You'll learn what it can do and why IBM developed this technology. You'll see how you can create, populate, and query Big SQL tables. And you will be presented with some important concepts that relational DBMS experts should understand about Big SQL.

Although a number of vendors offer SQL-on-Hadoop implementations, IBM's vast SQL experience enabled IBM to deliver a range of SQL capabilities that you'll be hard-pressed to find in competing offerings today. For example, while most implementations have limited support for subqueries, perhaps not allowing them in SELECT lists, in the HAVING clause, or with certain quantifiers (such as SOME, ANY or ALL). IBM does not have comparable restrictions. In addition, IBM provides more than 200 built-in functions, including a wide range of OLAP functions, where other implementations have considerably fewer. IBM supports UDFs written in Java, C, and SQL. Most other implementations support only Java-based UDFs. Finally, IBM offers fine-grained access control (column masking and row-based permissions) as well as federated queries. Again, many competing offerings lack such capabilities.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Agenda

- Big SQL overview
  - motivation
  - architecture
  - distinguishing characteristics
- Using Big SQL: the basics
  - invocation options
  - creating tables and views
  - populating tables with data
  - querying data

© Copyright IBM Corporation 2015

Agenda

Here's a quick look at what will be reviewed. You will spend some time on the motivation and architecture of Big SQL, during which time distinguishing characteristics will be summarized. The bulk of this presentation will present how you can use Big SQL. First the basics will be reviewed, and then some advanced topics. Note that there is a lot more to Big SQL than can be covered in the scope of this introductory course.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Agenda

- Big SQL overview
  - motivation
  - architecture
  - distinguishing characteristics
- Using Big SQL: the basics
  - invocation options
  - creating tables and views
  - populating tables with data
  - querying data

© Copyright IBM Corporation 2015

Agenda

First, a quick overview of Big SQL.

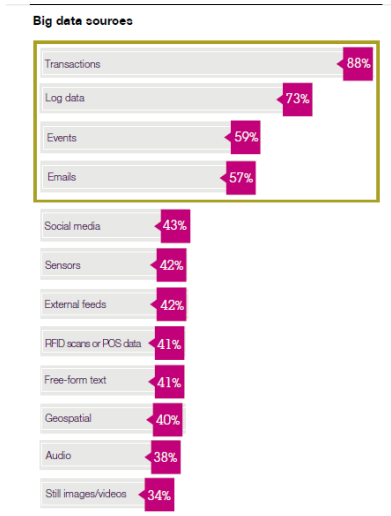
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

SQL access for Hadoop: why?

- Data warehouse modernization is a leading Hadoop use case
  - off-load "cold" warehouse data into query-ready Hadoop platform
  - explore / transform / analyze / aggregate social media data, log records, etc. and upload summary data to warehouse
- Limited availability of skills in MapReduce, Pig, etc.
- SQL opens the data to a much wider audience
  - familiar, widely known syntax
  - common catalog for identifying data and structure



2012 Big Data @ Work Study surveying 1144 business and IT professionals in 95 countries

© Copyright IBM Corporation 2015

SQL access for Hadoop: why?

Let's start with our motivation to build a new SQL engine for Hadoop.

There has been a massive amount of buzz about SQL access to Hadoop. The reasons are clear.

Hadoop is very appealing to businesses needing to store large volumes of potentially messy data in a cost effective way. It is quite easy to dump a lot of this data into a Hadoop cluster, and scale out inexpensively, however it is very difficult to get value out of that information. The expertise required to do Hadoop programming is in short supply, and is quite expensive. But the potential for Hadoop to extend the warehouse is so compelling that the pursuit of SQL access is being given a great deal of attention.

Just think of it: you have a flexible platform like Hadoop, and you have an entire industry with SQL skills and extensive tools that are based on SQL. If Hadoop clusters can have that native SQL interface, this will rapidly speed up the adoption of Hadoop, and make all sorts of analytics opportunities possible.

The chart at right of the slide shares the results of a big data study in 2012, and it shows that transactional data is the number one source of big data; think about how SQL has been run over transactional data for decades.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Source: *Analytics: The real-world use of big data, How innovative enterprises extract value from uncertain data,* IBM Institute for Business Value and Saïd Business School at the University of Oxford, 2012

Link:
http://public.dhe.ibm.com/common/ssi/ecm/en/gbe03519usen/GBE03519USEN.PDF

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

SQL-on-Hadoop landscape

- The SQL-on-Hadoop landscape changes constantly

- Being relatively new to the SQL game, they've generally had to compromise in one or more of these areas:
  - speed
  - robust SQL
  - enterprise features
  - interoperability
- Big SQL based on decades of IBM R&D investment in relational technology that addresses these areas

© Copyright IBM Corporation 2015

SQL-on-Hadoop landscape

Of course, IBM is not the only vendor to recognize the demand for SQL on Hadoop. Most of the major Hadoop vendors have jumped into the fray and delivered some level of SQL support. Features vary significantly, and vendors are rapidly evolving their offerings. But most players are pretty new to the world of SQL, and it's not quick or easy to build a truly enterprise-grade SQL engine. This means they were forced to compromise on some critical features. By contrast, Big SQL is based on decades of IBM's research and development investment in relational technology, affording IBM a better opportunity to deliver advanced technology to the Hadoop community today.

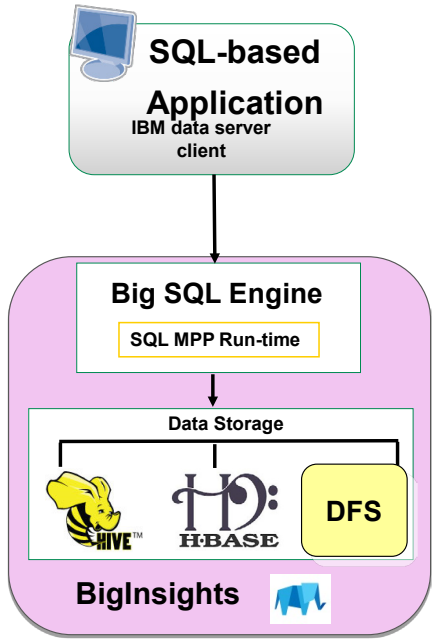
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

What is Big SQL?

- Comprehensive, standard SQL
  - SELECT: joins, unions, aggregates, subqueries
  - GRANT/REVOKE, INSERT … INTO
  - PL/SQL
  - Stored procs, user-defined functions
  - IBM data server JDBC and ODBC drivers
- Optimization and performance
  - IBM MPP engine (C++) replaces Java MapReduce layer
  - Continuous running daemons (no start up latency)
  - Message passing allow data to flow between nodes without persisting intermediate results
  - In-memory operations with ability to spill to disk (useful for aggregations, sorts that exceed available RAM)
  - Cost-based query optimization with 140+ rewrite rules
- Various storage formats supported
  - Text (delimited), Sequence, RCFile, ORC, Avro, Parquet
  - Data persisted in DFS, Hive, HBase
  - No IBM proprietary format required
- Integration with RDBMSs via LOAD, query federation



IBM BigInsights for Analysts

© Copyright IBM Corporation 2015

What is Big SQL?

Big SQL is designed to provide SQL developers with an easy on-ramp for querying data managed by Hadoop. It enables data administrators to create new tables for data stored in DFS, HBase, or the Hive warehouse. In addition, a LOAD command enables administrators to populate Big SQL tables with data from various sources. And Big SQL's JDBC and ODBC drivers enable many existing tools to use Big SQL to query this distributed data.

Big SQL's runtime execution engine is all native code (C/C++)

For common table formats a native I/O engine is utilized

- for example text (delimited), RC, SEQ, Parquet, Avro

For all others, a Java I/O engine is used

- maximizes compatibility with existing tables
- allows for custom file formats and SerDe's

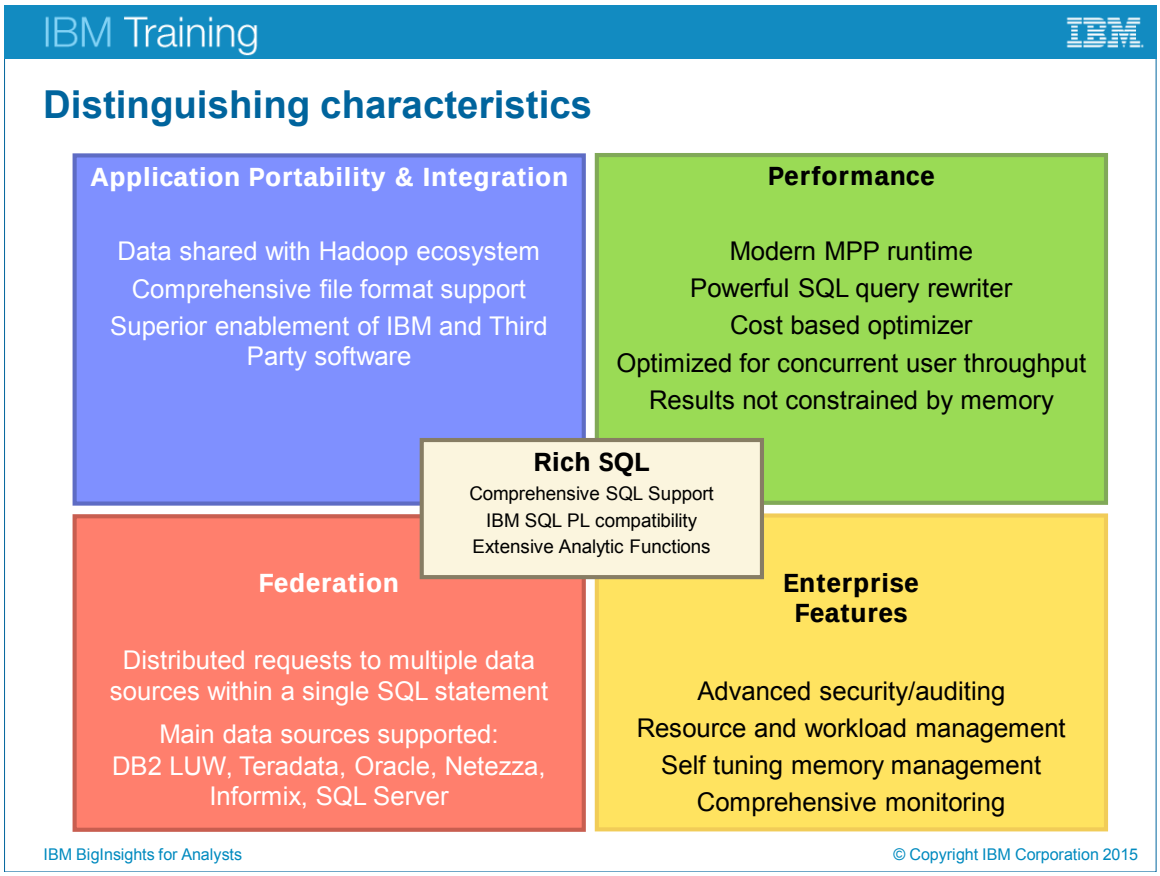
For details on supported file types, see the product documentation or this article:

- https://developer.ibm.com/hadoop/blog/2014/09/19/big-sql-3-0-file-formats-usage-performance/

Customer built UDx's can be developed in SQL, C++ or Java

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Distinguishing characteristics

What distinguishes Big SQL from other SQL-on-Hadoop offerings? Summarizing the key characteristics here, they have been categorized into four broad areas. Many of these features will be reviewed in greater detail later.

There is also a good white paper that summarizes IBM strengths: http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?subtype=WH&infotype=SA&appname=SWGE_SW_SW_USEN&htmlfid=SWW14019USEN&attachment=SWW14019USEN.PDF#loaded

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Agenda

- Big SQL overview
  - motivation
  - architecture
  - distinguishing characteristics
- Using Big SQL: the basics
  - invocation options
  - creating tables and views
  - populating tables with data
  - querying data

IBM BigInsights for Analysts

© Copyright IBM Corporation 2015

Agenda

With that background, it is time to look at Big SQL in action. Firstly, the basics: how to invoke Big SQL, how to create tables and views, how to populate tables with data, and how to query Big SQL tables. And in doing so, you will begin to understand that Big SQL provides an easy on-ramp to Hadoop for SQL professionals.

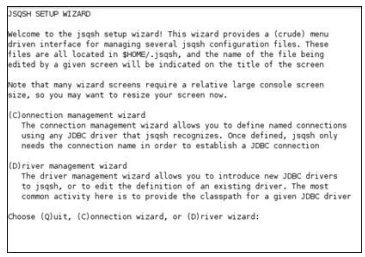
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

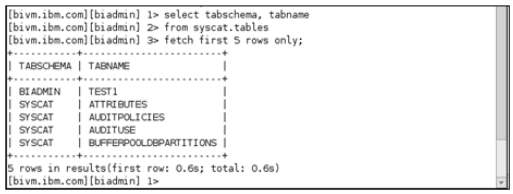
IBM Training

Invocation options

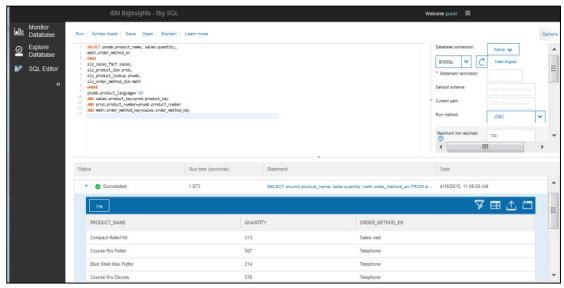
- Command-line interface: Java SQL Shell (JSqsh)



- Web tooling (Data Server Manager)



- Tools that support IBM JDBC/ODBC driver



IBM BigInsights for Analysts

© Copyright IBM Corporation 2015

Invocation options

Big SQL includes a command-line interface called JSqsh. JSqsh (pronounced J-skwish) is a short name for Java SQshell (pronounced s-q-shell). This is an open source database query tool featuring much of the functionality provided by a good shell, such as variables, redirection, history, command line editing, and so on. As displayed on this chart, it includes built-in help information and a wizard for establishing new database connections.

In addition, when Big SQL is installed, administrators can also install IBM Data Server Manager (DSM) on the Big SQL Head Node. This web-based tool includes a SQL editor that runs statements and returns results, as shown here. DSM also includes facilities for monitoring your Big SQL database.

For more on DSM, visit http://www-03.ibm.com/software/products/en/ibm-data-server-manager.

Tools that support IBM's JDBC / ODBC driver are also options.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Creating a Big SQL table

- Standard CREATE TABLE DDL with extensions

```
create hadoop table users
(
  id        int           not null primary key,
  office_id int           null,
  fname     varchar(30)   not null,
  lname     varchar(30)   not null)
row format delimited
   fields terminated by '|'
stored as textfile;
```

- Worth noting:
  - "Hadoop" keyword creates table in DFS
  - Row format delimited and textfile formats are default
  - Constraints not enforced (but useful for query optimization)
- Examples in these charts focus on DFS storage, both within or external to Hive warehouse. HBase examples provided separately.

© Copyright IBM Corporation 2015

Creating a Big SQL table

At the top of the slide is the syntax for creating a Big SQL table called **users** in the default schema (the user's ID). If you are familiar with SQL, most of this statement will look familiar to you. The items highlighted in the rectangle are Big SQL DDL extensions for Hadoop. For example, the Hadoop keyword indicates that the data is to be stored on the cluster in the distributed file system, not as a "local" table on the Big SQL head node only.

Other clauses specify the storage format of the data; in this case, a row-based text file, with fields delimited by a vertical bar ("|"). Other options and clauses are supported. However, what is important to take away from this example is that it is standard SQL plus some Hadoop-specific extensions.

In the slides, HDFS and Hive examples will be displayed, but you should be aware that Big SQL supports HBase too. HBase is a columnar data store and, as such, introduces additional data modeling considerations for SQL administrators.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Hadoop Keyword:

- Big SQL requires the HADOOP keyword
- Big SQL has internal traditional RDBMS table support
  - stored only at the head node
  - does not live on HDFS
  - supports full ACID capabilities
  - not usable for big data
- The HADOOP keyword identifies the table as living on HDFS

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Results from previous CREATE TABLE

- Data stored in subdirectory of Hive warehouse
  - . . . /hive/warehouse/myid.db/users
  - Default schema is user ID; can create new schemas
  - "Table" is just a subdirectory under schema.db
  - The table's data are files within table subdirectory
- Meta data collected (Big SQL & Hive)
  - SYSCAT.* and SYSHADOOP.* views
- Optionally, use LOCATION clause of CREATE TABLE to layer Big SQL schema over existing DFS directory contents
  - Useful if table contents already in DFS
  - Avoids need to LOAD data into Hive
  - Example provided later

© Copyright IBM Corporation 2015

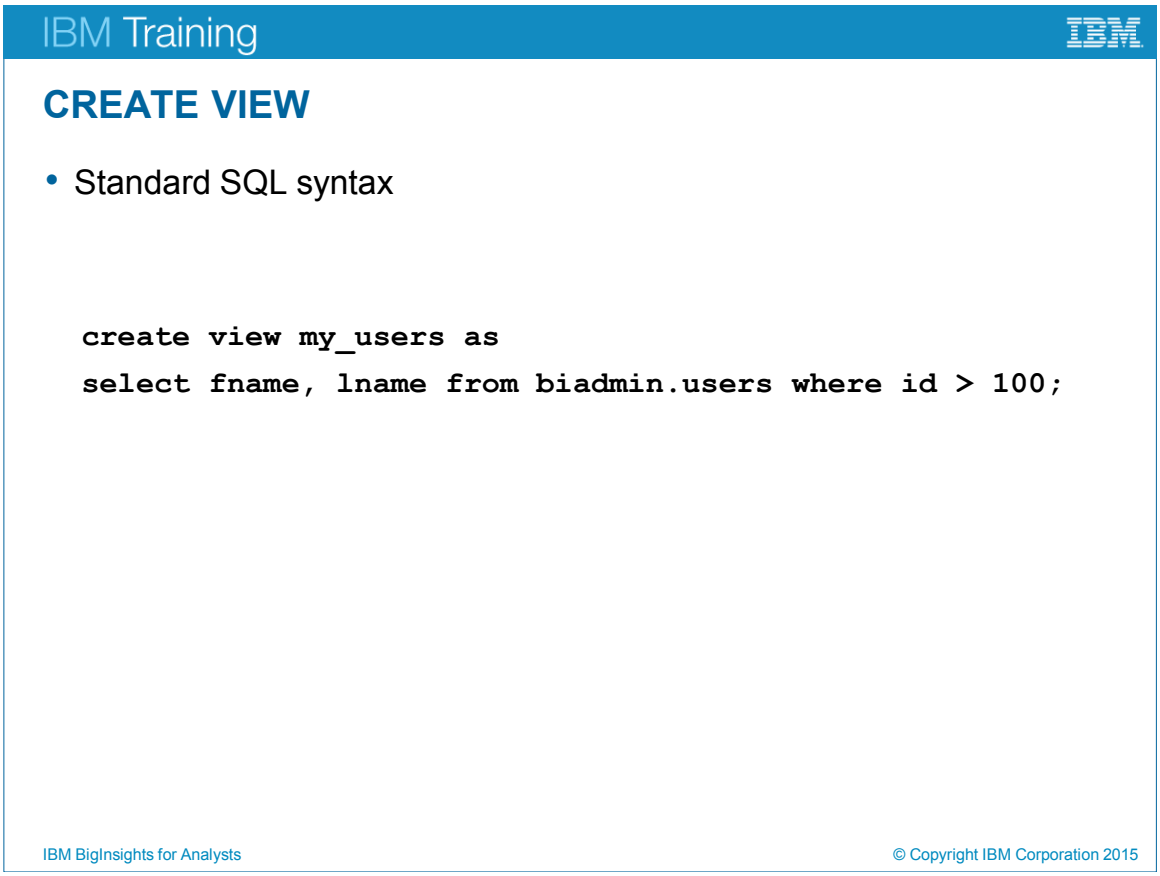
Results from a previous CREATE TABLE

Executing the previous CREATE TABLE statement creates a subdirectory in the Hive warehouse directory.

Since a schema name was not explicitly specified, Big SQL uses your login ID as a default; in this example, it is myid. Schemas are represented in Hive as folders, so your schema folder is myid.db. A table in Hive is just a subdirectory under the schema folder, so your USERS table is found in *. . . /hive/warehouse/myid.db/users*. When you load data into this table, files will be stored under this subdirectory.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



CREATE VIEW

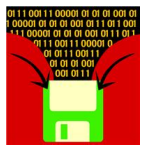
You can create Big SQL views in the same way that you would create a view in a relational DBMS. A simple example is shown here.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Populating tables via LOAD



- Typically best runtime performance
- Load data from local or remote file system

```
  load hadoop using file url
'sftp://myID:myPassword@myServer.ibm.com:22/install-
  dir/bigsql/samples/data/GOSALESDW.GO_REGION_DIM.txt' with SOURCE PROPERTIES
  ('field.delimiter'='\t')  INTO TABLE gosalesdw.GO_REGION_DIM overwrite;
```

- Loads data from RDBMS (DB2, Netezza, Teradata, Oracle, MS-SQL, Informix) via JDBC connection

```
load hadoop
using jdbc connection url 'jdbc:db2://some.host.com:portNum/sampledb'
with parameters (user='myID', password='myPassword')
from table MEDIA columns (ID, NAME)
where 'CONTACTDATE < ''2012-02-01'''
into table media_db2table_jan overwrite
with load properties ('num.map.tasks' = 10);
```

IBM BigInsights for Analysts © Copyright IBM Corporation 2015

Populating tables via LOAD

After you create a Big SQL table, you can use the LOAD command to populate it with data from files in a remote file system, files in your distributed file system, or data in the remote RDBMS servers listed. For RDBMS data, you specify JDBC URL properties and either enter a SQL query or a table name to identify the data to be retrieved. Behind the scenes, LOAD uses Sqoop connectors to retrieve the necessary data from the source.

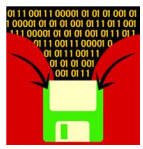
In general, LOAD typically offers the best runtime performance for populating tables with data. However, there are some other options that you may want to be aware of.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Populating tables via INSERT



- INSERT INTO . . . SELECT FROM . . .
  - Parallel read and write operations

```
CREATE HADOOP TABLE IF NOT EXISTS big_sales_parquet
( product_key INT NOT NULL, product_name VARCHAR(150),
Quantity INT, order_method_en VARCHAR(90) )
STORED AS parquetfile;
-- source tables do not need to be in Parquet format
insert into big_sales_parquet
SELECT sales.product_key, pnumb.product_name, sales.quantity, meth.order_method_en
FROM sls_sales_fact sales, sls_product_dim prod,sls_product_lookup pnumb,
sls_order_method_dim meth
WHERE
pnumb.product_language='EN'
AND sales.product_key=prod.product_key
AND prod.product_number=pnumb.product_number
AND meth.order_method_key=sales.order_method_key
and sales.quantity > 5500;
```

- INSERT INTO . . . VALUES(. . .)
  - Not parallelized. 1 file per INSERT. Not recommended except for quick tests

```
Create table foo (col1 int, col2 varchar(10));
insert into foo values (1, 'hello');
```

IBM BigInsights for Analysts © Copyright IBM Corporation 2015

Populating tables via INSERT

In some cases, it may be more convenient to populate a table with data using an INSERT statement. The INSERT INTO . . . SELECT FROM statement supports parallel read and write operations and can be a convenient way to populate a table with data retrieved from a query, particularly if you want to change the underlying storage format used for data.

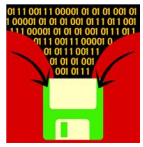
The traditional INSERT INTO . . . VALUES(…) format is also supported but not recommended for anything but simple test operations. Its work is never parallelized, and each INSERT results in a separate file.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

CREATE . . . TABLE . . . AS SELECT . . .



- Create a Big SQL table based on contents of other table(s)
- Source tables can be in different file formats or use different underlying storage mechanisms

```
 -- source tables in this example are external (just DFS files)
CREATE HADOOP TABLE IF NOT EXISTS sls_product_flat
 ( product_key INT NOT NULL
 , product_line_code INT NOT NULL
 , product_type_key INT NOT NULL
 , product_type_code INT NOT NULL
 , product_line_en VARCHAR(90)
 , product_line_de VARCHAR(90)
 )
as select product_key, d.product_line_code, product_type_key,
product_type_code, product_line_en, product_line_de
from extern.sls_product_dim d, extern.sls_product_line_lookup l
where d.product_line_code = l.product_line_code;
```

IBM BigInsights for Analysts © Copyright IBM Corporation 2015

CREATE … TABLE … AS SELECT …

In addition to the simple CREATE HBASE TABLE statement shown earlier, you can also create and populate new Big SQL tables based on the content of other Big SQL tables. This can be particularly convenient if you want to denormalize data stored in multiple tables to satisfy certain query workloads. In the example shown here, data is extracted from two different tables that were created as externally managed tables; that is to say that they are just stored as files in some DFS directories, and you are putting this data into a new Hadoop table (in the Hive warehouse). Specifically, the SELECT statement is taking two tables along the PRODUCT dimension of a data warehouse schema, and populating a new Big SQL table. Since the content for the source tables is stored in two separate files that each contain different sets of fields, you could not directly load them into our target table. Here, you can use Big SQL to help with that task.

The CREATE … TABLE .. AS SELECT statement can also be useful for populating Big SQL tables managed by HBase.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

SQL capability highlights

- Query operations
  - Projections, restrictions
  - UNION, INTERSECT, EXCEPT
  - Wide range of built-in functions (e.g. OLAP)
- Full support for subqueries
  - In SELECT, FROM, WHERE and HAVING clauses
  - Correlated and uncorrelated
  - Equality, non-equality subqueries
  - EXISTS, NOT EXISTS, IN, ANY, SOME, etc.
- All standard join operations
  - Standard and ANSI join syntax
  - Inner, outer, and full outer joins
  - Equality, non-equality, cross join support
  - Multi-value join
- Stored procedures, UDFs
  - DB2 compatible PL/SQL support
  - Cursors, flow of control (if/then/else, error handling, etc.)

```
SELECT
   s_name,
   count(*) AS numwait
FROM
   supplier,
   lineitem l1,
   orders,
   nation
WHERE
   s_suppkey = l1.l_suppkey
   AND o_orderkey = l1.l_orderkey
   AND o_orderstatus = 'F'
   AND l1.l_receiptdate > l1.l_commitdate
   AND EXISTS (
      SELECT
         *
      FROM
         lineitem l2
      WHERE
         l2.l_orderkey = l1.l_orderkey
         AND l2.l_suppkey <> l1.l_suppkey
   )
   AND NOT EXISTS (
      SELECT
         *
      FROM
         lineitem l3
      WHERE
         l3.l_orderkey = l1.l_orderkey
         AND l3.l_suppkey <> l1.l_suppkey
         AND l3.l_receiptdate >
               l3.l_commitdate
   )
   AND s_nationkey = n_nationkey
   AND n_name = ':1'
GROUP BY s_name
ORDER BY numwait desc, s_name;
```

IBM BigInsights for Analysts

© Copyright IBM Corporation 2015

SQL capability highlights

The basic relational operators associated with SQL are all supported by Big SQL, as shown here on this chart. Many more sophisticated query operations are supported, too, including various types of joins, correlated and uncorrelated subqueries, PL/SQL, OLAP functions, and more. Take a look at the query on the right; it contains a variety of SQL expressions that you would expect to find in any commercial RDBMS. But some SQL-on-Hadoop implementations cannot run this query because it includes SQL expressions that are not supported, such as non-equi joins, subqueries in the WHERE clause, etc.

Be aware that Big SQL is not case sensitive. The following statements are all equivalent:

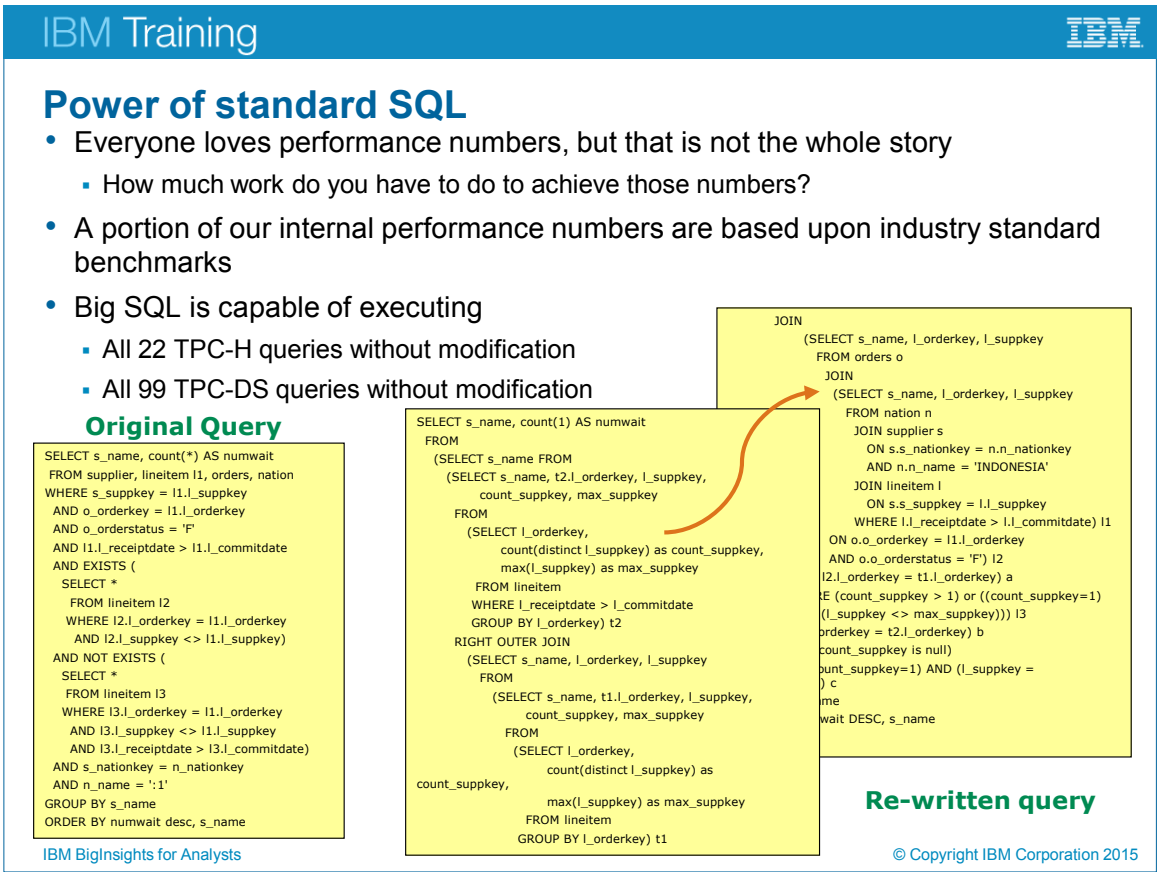
SELECT col1, col2 FROM t1;

select col1, col2 from t1;

SELECT COL1, COL2 FROM T1;

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



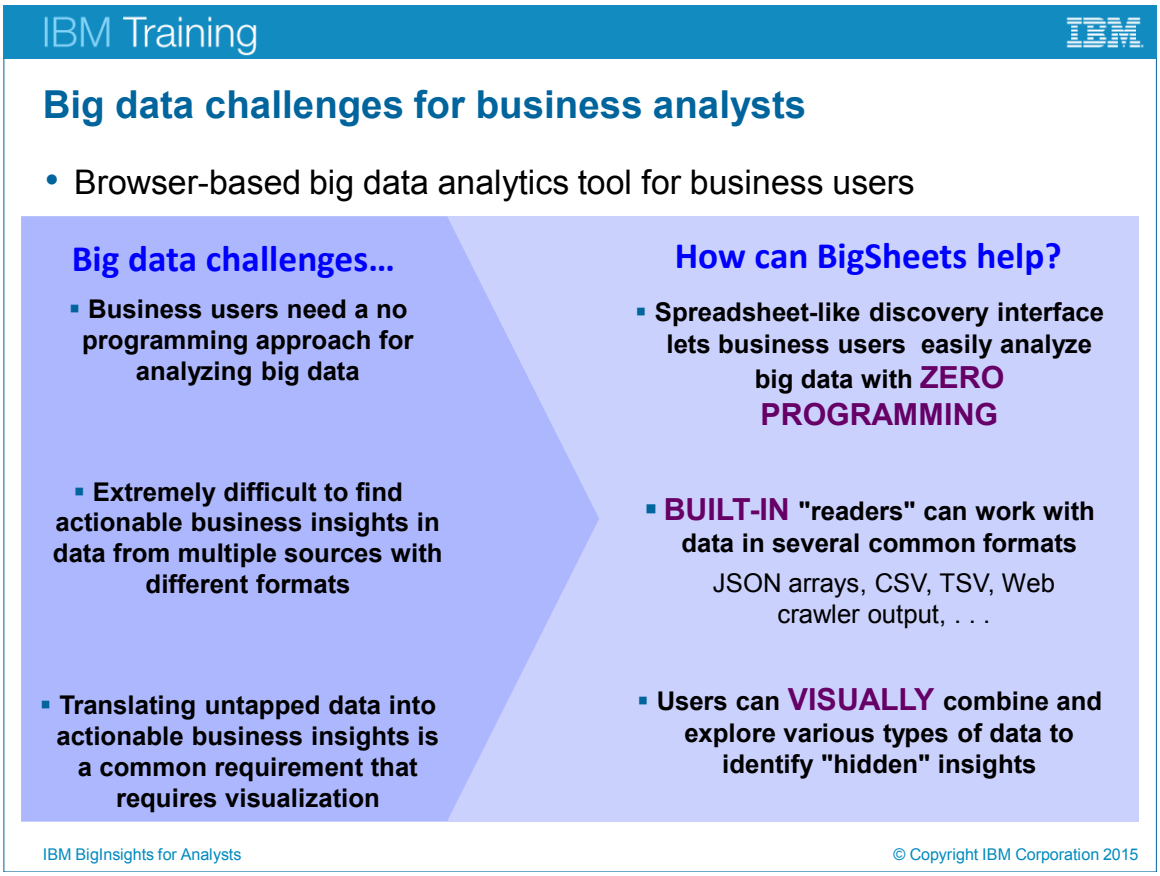
Power of Standard SQL

Another topic to consider: query workloads. Some of the more popular query workloads were defined by the Transaction Processing Council (TPC). These workloads have been used for years for performance benchmarking of RDBMSs. Indeed, IBM based internal tests on TPC-H and TPC-DS query workloads, and Big SQL can run all 22 TPC-H queries and all 99 TPC-DS queries without modification. Most other SQL-on-Hadoop implementations cannot do that. Certain queries had to be rewritten because support for certain standard SQL expressions is missing. The example shown is of a query that needed to be rewritten for a different SQL-on-Hadoop implementation. It is true that these implementations are changing all the time, however if you are going to consider a SQL offering other than Big SQL, it is worth asking if that implementation can run all TPC queries without significant modifications.

Before addressing performance, be aware that some of the numbers you see vendors touting do not convey the full story. In particular, it is important to understand the effort required to achieve those numbers, including the effort to rewrite standard SQL queries so that they can be run against the target platform.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Big data challenges for business analysts

One of the big data challenges is how to get analysts to go out and analyze this data with no programming. If you do not have such tooling, you create an unnatural dependency on development to code and build every piece of visualization and analysis. This is too expensive, inefficient, and time consuming. BigSheets gives you exactly this, with zero programming. Your analysts want to be able to visualize and analyze data in JSON, CVS and text file formats. They want a programming free crawler and more, all of which is included in BigSheets. To the person using BigSheets, it looks like a spreadsheet, but under behind the scenes, it generates PIG jobs to run on Hadoop.

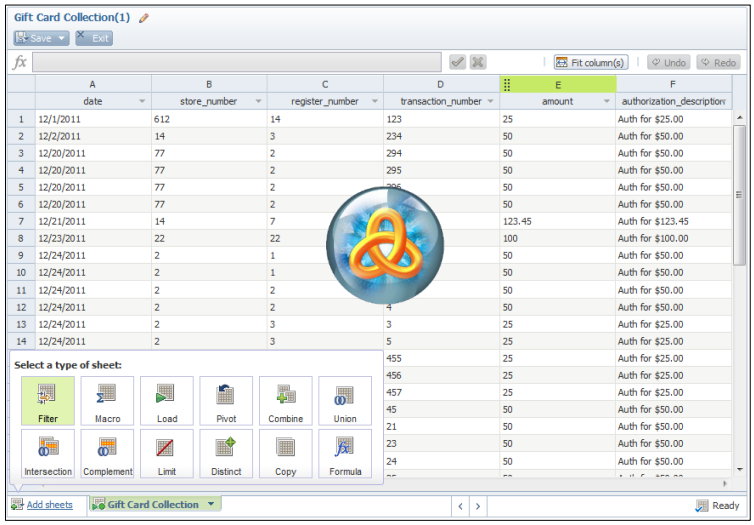
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

What is BigSheets?

- Browser-based analytics tool for business users
- Spreadsheet like interface for analyzing big data
- A component of the Analyst module of IBM BigInsights



IBM BigInsights for Analysts

© Copyright IBM Corporation 2015

What is BigSheets?

BigSheets is a browser-based analytic tool designed to work with big data. Unlike many other big data tools, it is designed to support business users and non-technical professionals. To do so, it presents a familiar, spreadsheet-like interface that allows users to gather, filter, combine, explore, and visualize data from various sources.

IBM chose the spreadsheet as the model for organizing data because most users are already familiar with such software. If users want to represent the data in more complex ways, the tool works with an IBM visualization tool called Many Eyes, and other visualization software.

As an important part of IBM's big data strategy, BigSheets is a feature of IBM BigInsights.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

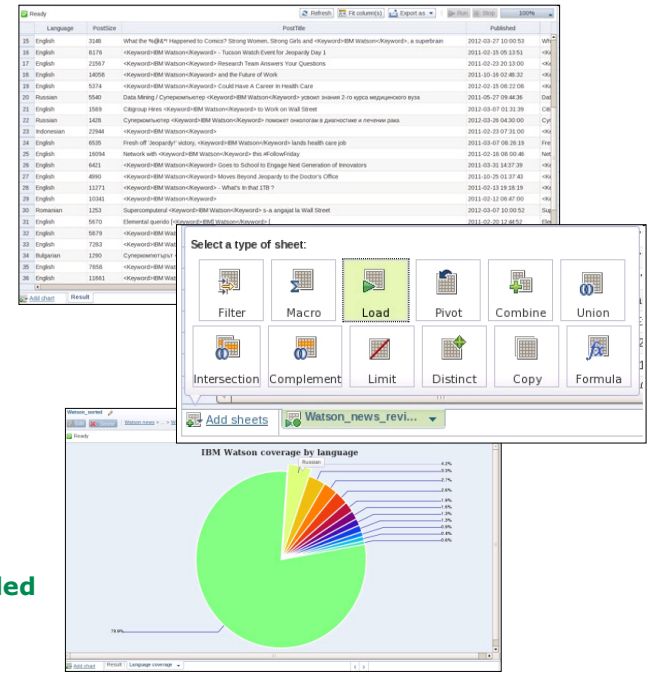
© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

What you can do with BigSheets?

- Model big data collected from various sources in spreadsheet-like structures
- Filter and enrich content with built-in functions
- Combine data in different workbooks
- Visualize results through spreadsheets, charts
- Export data into common formats (if desired)

No programming knowledge needed



IBM BigInsights for Analysts

© Copyright IBM Corporation 2015

What you can do with BigSheets?

BigSheets is a browser-based visualization and analysis tool designed to help non-programmers work with big data. It ships with BigInsights Quick Start, and the Analyst, Data Scientist, and the Apache Hadoop modules.

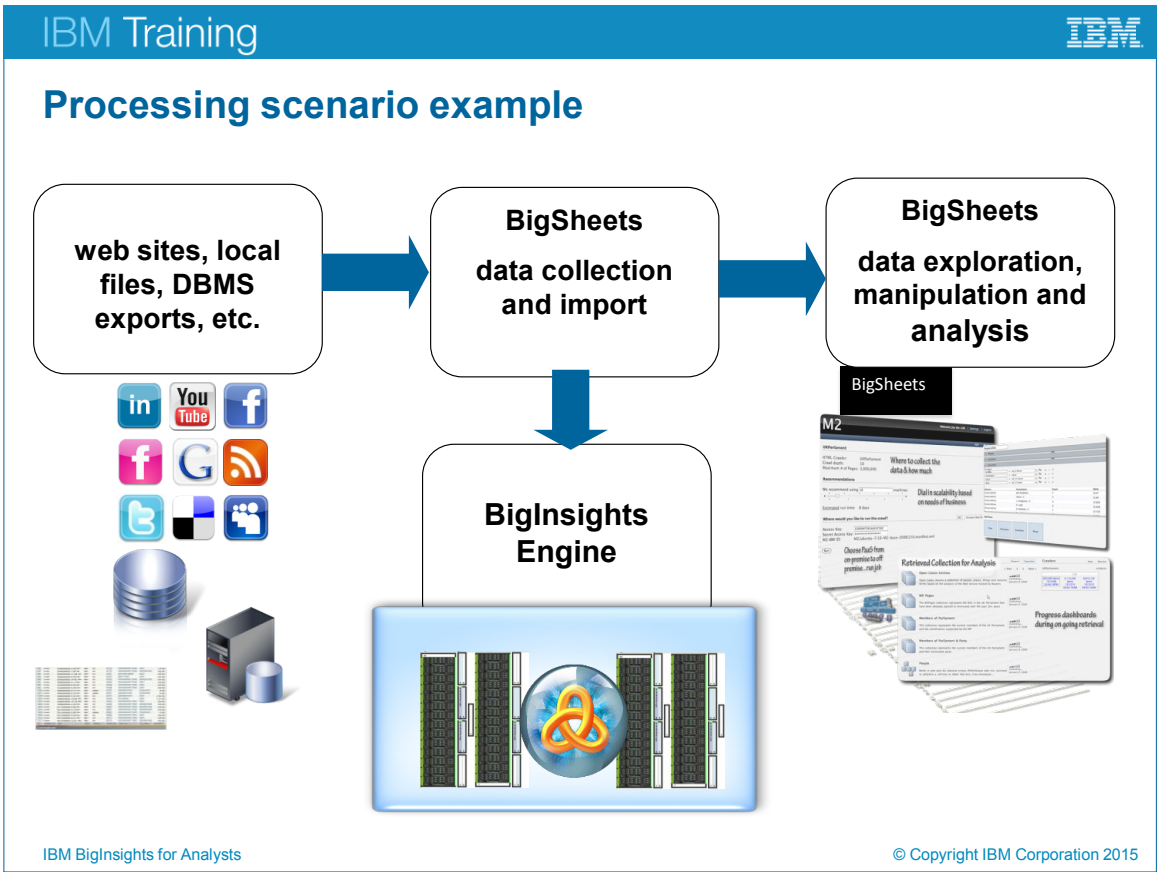
With this tool, users model their big data in workbooks, or familiar spreadsheet-like tabular data structures. Once data is represented in a workbook, business analysts can filter and enrich its content using built-in functions and macros. Furthermore, analysts can combine data residing in different workbooks as well as generate charts and new "sheets" (workbooks) to visualize their data. They can even export data into a variety of common formats with a click of a button.

Here are some of the distinguishing characteristics of BigSheets:

- It presents a user interface developed specifically for business intelligence and non-technical business users to facilitate data gathering and analysis.
- It can consume various kinds of data, such as CSV files produced by relational DBMSs and other applications or Web crawler data produced by a built-in application provided with BigInsights.
- It can combine data sources from different sources, potentially enabling users to identify trends, opportunities, and risks that are hidden in the data.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Processing scenario example

This is an example scenario that illustrates one way in which BigSheets can be used.

Firms can import data from web sites, local file systems, and other sources into BigSheets by using a simple graphical interface. Behind the scenes, BigSheets stores the data in BigInsights. Firms can then explore and manipulate the data using the BigSheets' simple spreadsheet interface and if desired, can generate charts to visualize specific results.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Unit summary

- Describe the components that come with the IBM BigInsights Analyst module
- Explain the benefits of using Big SQL for big data
- Understand the purpose of BigSheets

© Copyright IBM Corporation 2015

Unit summary

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Exercise 1

Working with BigSheets

© Copyright IBM Corporation 2015

Exercise 1: Working with BigSheets

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Exercise 1: Working with BigSheets

Purpose:
You will create a BigSheets workbook and derive a chart from it to visualize your data.

Estimated time: **30 minutes**
User/Password: **biadmin/biadmin**
root/dalvm3
Services Password: **ibm2blue**

Important: Before doing this exercise, ensure that your access and services are configured and running. Check that:

- /etc/hosts displays your environment's IP address
- in the Ambari console, ensure that all BigInsights services are running

If you are unsure of the steps, please refer to Unit 1, Exercise 1 to ensure that your environment is ready to proceed. You should review the steps in Task 1 (Configure your image) and Task 2 (Start the BigInsights components).

Task 1. Loading data into BigSheets.

BigSheets allows you to analyze the data residing on the HDFS. You can create master workbooks, apply various sheets types to refine and filter the data, and then create charts to visualize the data. This task will walk you through the start to the end from creating a workbook to visualizing the data with charts. More functions and features will be covered in the BigSheets specific module.

You will load in two set of data into the HDFS.

1. To open a new terminal, right-click **biadmin's Home**, and then click **Open in Terminal**.
2. Navigate to **/home/biadmin/labfiles/bigsheets** to see the files.
3. Upload **blogs-data.txt** and **news-data.txt** to **/user/biadmin/**.

```
hdfs dfs -put /home/biadmin/labfiles/bigsheets/blogs-data.txt /user/biadmin
hdfs dfs -put /home/biadmin/labfiles/bigsheets/news-data.txt /user/biadmin
```

Once the files are inside of the HDFS, you are ready to create the BigSheets workbook.

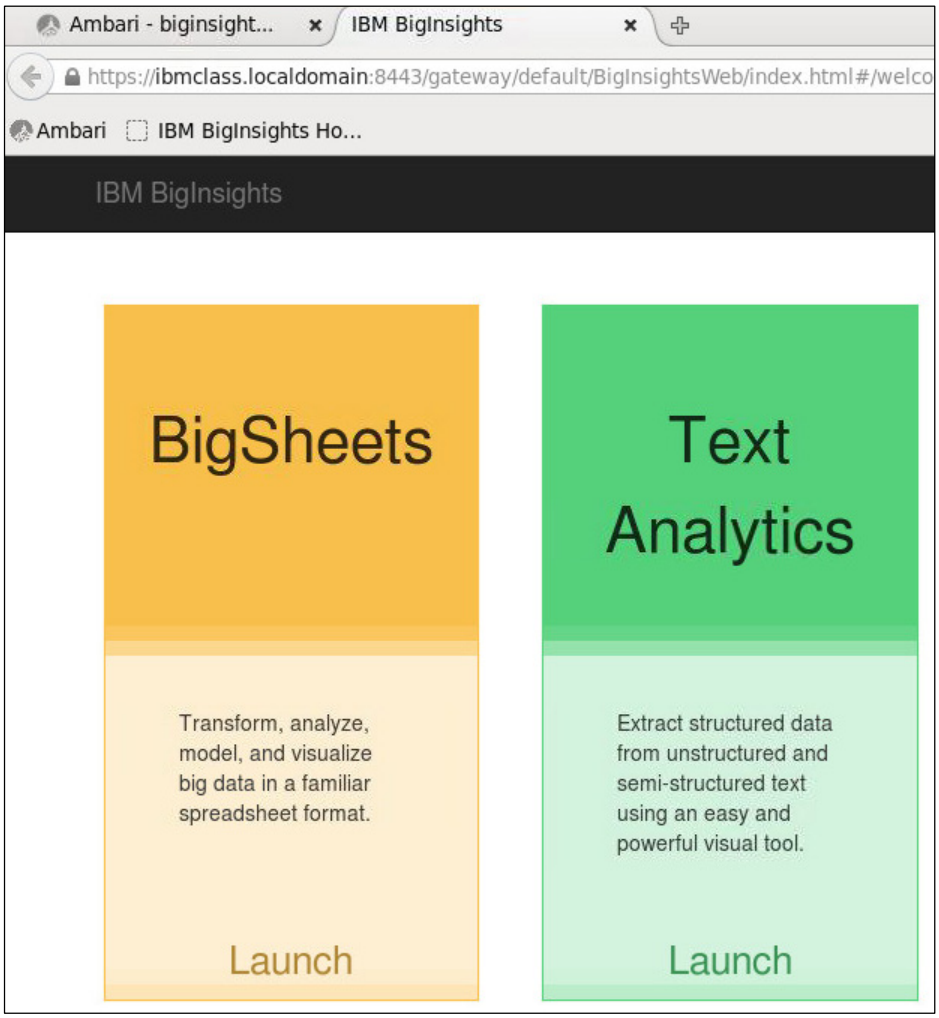
4. Launch **Firefox**, and then if necessary, navigate to the **Ambari** login page, **http://ibmclass.localdomain:8080**.
5. Log in to the **Ambari** console as **admin/admin**.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

6. Ensure that the **BigInsights - BigSheets** component is started.
   The BigInsights Home requires the LDAP server to be started.
7. Click the **Knox** component.
8. Under the **Service Actions** dropdown on the upper right, select **Start Demo LDAP**, and then click **OK** to close the confirmation window.
9. In Firefox, open a new tab, and navigate to the **Web UI (BigInsights Home)** page with the following URL.
   https://ibmclass.localdomain:8443/gateway/default/BigInsightsWeb/index.html
   If prompted for a login, use guest / guest-password. It should be saved in the Firefox browser so you can click OK to continue with the login.
   The results appear as follows:

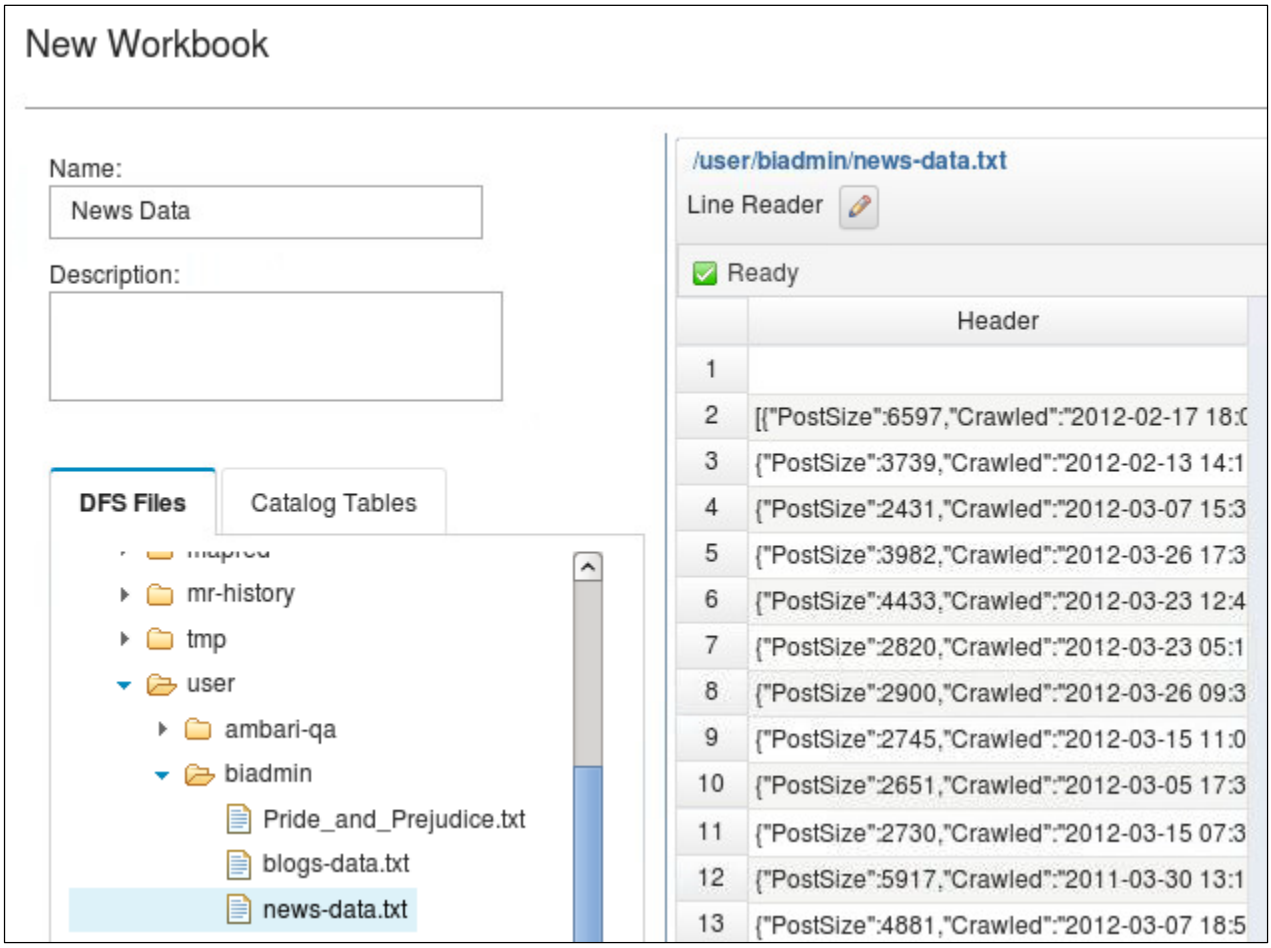


10. Click **BigSheets**, to launch BigSheets.
    In the next few steps, you will create two parent workbooks. One for the news-data.txt and one for news-blogs.txt residing on the HDFS.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

11. Click the **New Workbook** button.
12. On the New Workbook window, under **Name**, type **News Data.**
    You can leave the description field blank.
13. On the **DFS Files** tab, navigate to **/user/biadmin/** and select the **news-data.txt** file.
    The data will be previewed on the pane to the right.
    By default, it is using the Line Reader to parse the data. You will want to select the JSON Array reader so that the data is parsed correct.
    The results appear as follows:

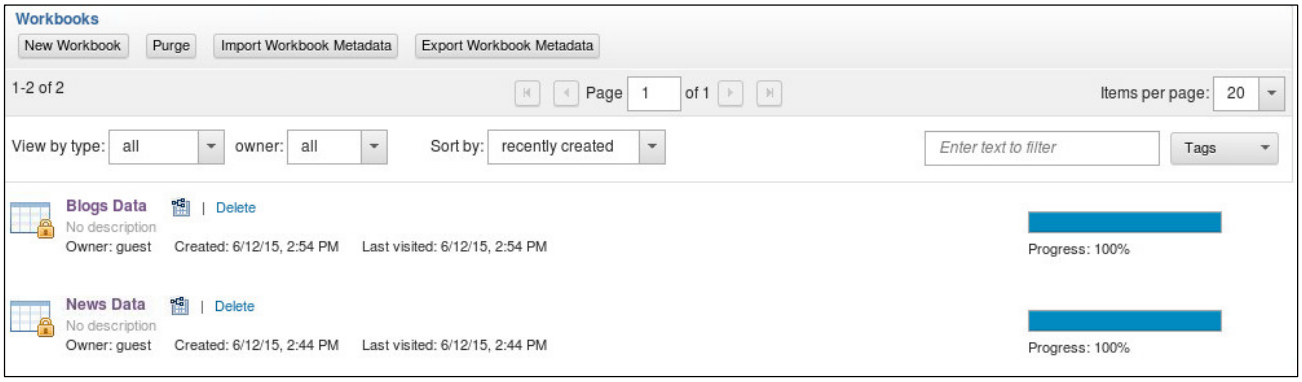


14. Beside **Line Reader**, click **Edit workbook reader**  .
15. In the **Select a reader** list, select **JSON Array**, and then click **Set reader**  .
    You can see now that the data is properly parsed.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

16. At the bottom of the New Workbook window, click **Save workbook**  .
    Using the same steps, you will create the Blog Data workbook.
17. Click the **Workbooks** link (breadcrumb) to go back to the BigSheets home page.
18. Click the **New Workbook** button.
19. On the New Workbook window, under **Name**, type **Blogs Data.**
    You can leave the description field blank.
20. On the **DFS Files** tab, navigate to **/user/biadmin/** and select the **blogs-data.txt** file.
21. Specify the **JSON Array** reader.
22. At the bottom of the New Workbook window, click **Save workbook**  .
23. Click the **Workbooks** link to return to the BigSheets home page.
    The results appear as follows:



Task 2. Creating and editing child workbooks.

1. Click **News Data** to open the workbook.
   You will create a child workbook.
2. Beside **News Data**, click the **Build new workbook** button.
   You will now remove unnecessary columns.
3. In the **IsAdult** column header, expand the dropdown menu, and then click **Remove**.
4. In any column header, expand the dropdown menu, and then click **Organize Columns**.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

5. Click **Remove item**  beside each of the following columns to delete them.
   - **Crawled**
   - **Inserted**
   - **MoreoverUrl**
   - **PostSize**
   - **URL**
6. Click **Apply settings**  to confirm your actions.
7. Beside **News Data(1)**, click **Edit workbook name**  , beside **Name**, type **NewsDataRevised**, and then click **Save Tag**  .
8. Expand the **Save** dropdown, and then click **Save and Exit**.
9. Note that you could also name the workbook in here. Since we had already named it, click the **Save** button.

   Part of BigSheets is the feature to view what you intend to do on the subset of the data. In order for the changes to take effect on the full dataset, you must run the workbook. When you save and exit from the workbook, you will be prompted to Run the workbook.
10. Click **Run** to run the workbook on the full set of data.
11. Click **Workbooks**.

    You will use the steps above to revise the Blogs Data workbook.
12. Click **Blogs Data**.
13. Beside **Blogs Data**, click the **Build new workbook** button.

    You will now remove unnecessary columns.
14. In the **IsAdult** column header, expand the dropdown menu, and then click **Remove**.
15. In any column header, expand the dropdown menu, and then click **Organize Columns**.
16. Click **Remove item**  beside each of the following columns to delete them.
    - **Crawled**
    - **Inserted**
    - **Url**
    - **PostSize**

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

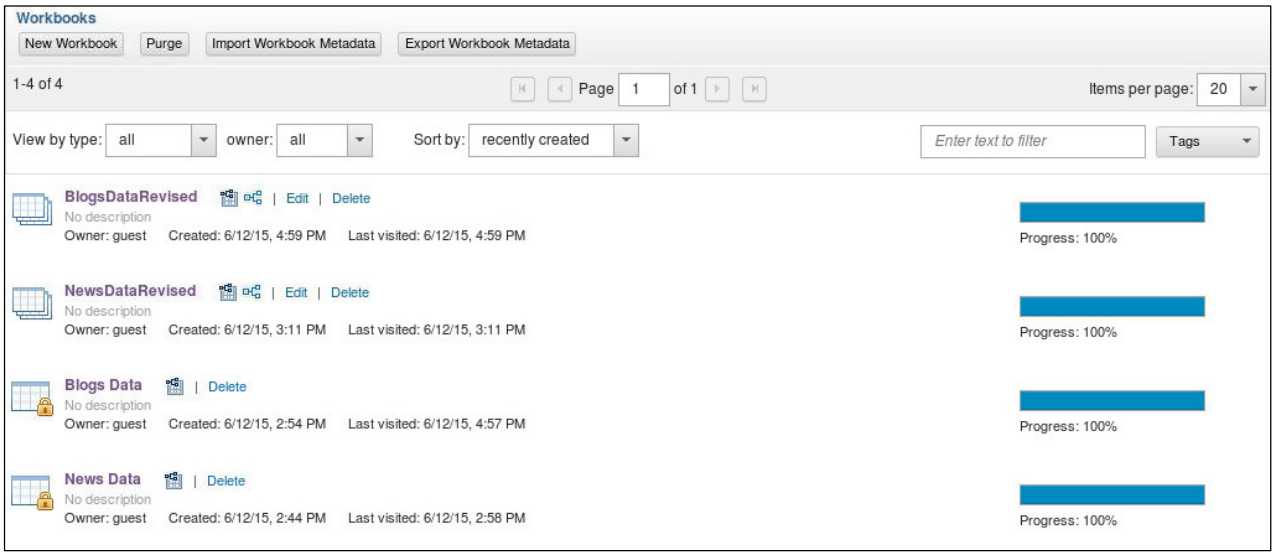
© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

17. Click **Apply settings**  to confirm your actions.
18. Beside **Blogs Data(1)**, click **Edit workbook name**  , beside **Name**, type **BlogsDataRevised**, and then click **Save Tag**  .
19. Expand the **Save** dropdown, and then click **Save and Exit**.
20. Note that you could also name the workbook in here. Since we had already named it, click the **Save** button.

    Part of BigSheets is the feature to view what you intend to do on the subset of the data. In order for the changes to take effect on the full dataset, you must run the workbook. When you save and exit from the workbook, you will be prompted to Run the workbook.
21. Click **Run** to run the workbook on the full set of data.

    The results appear as follows:



Leave The BigInsights - BigSheets window open for the next task.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Task 3. Combining workbooks.

In this task, you will be merging the two workbooks: NewsDataRevised and BlogsDataRevised with a union operation as a basis for exploring the data. To do so, both workbooks must have the same structure (or schema). In the last task, you modified the two workbooks to have the same columns so both workbooks are ready to be merged.

Before you can do a union operation, both sheets must be in the same workbook. You will open the NewsDataRevised and bring in the BlogsDataRevised sheet using the load operation.

1. Open the **NewsDataRevised** workbook.
2. Click the **Build new workbook** button.
3. Expand the **Add sheets** dropdown, and then click the **Load** sheet.
4. Under **Sheet Name**, type **BlogsDataRevised**.
5. Click the **BlogsDataRevised** workbook.
6. Click **Apply Settings**  to run the load operation.

   Once the operation completes, at the bottom left of the window, you will notice a new tab showing the Blog sheet that was just loaded.

   Now you are ready to create a union of these two sheets.
7. Click **Add sheets** and select the **Union** operation.
8. Name the sheet: **Union2Collection**.
9. From the **Select sheet** dropdown, add the **BlogsDataRevised** and the **NewsDataRevised** sheet to be used for the Union operation, and then click **Apply Settings**.

   You will see a new tab at the bottom when the operation completes.

   The results appear as follows:

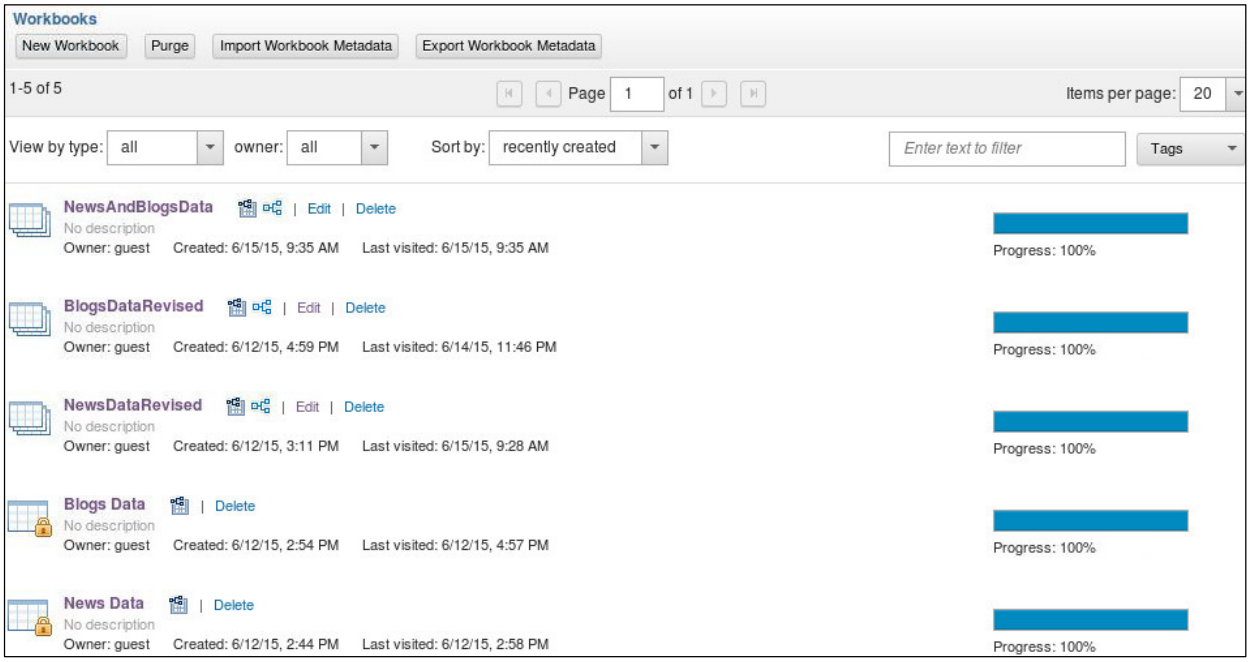
10. Click **Save**, click **Save & Exit**, in the **Name** box, type **NewsAndBlogsData**, and then click **Save**.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

11. Run the workbook.

    The results appear as follows:



Task 4. Sorting and creating charts.

1. Open the **NewsAndBlogsData** workbook.
2. To create a new child workbook, click **Build new workbook**.
3. From any column options menu, point to **Sort**, and then click **Advanced**.
4. In the **Add Columns to Sort** list, add the columns **Language** and **Type** to be sorted.

   Hint: Click  to add to the list.
5. Beside **Language**, select **Descending**, beside **Type** select **Ascending**, and then use the arrows to ensure that Language is the primary sort column.

   The results appear as follows:

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

6. Click **Apply Settings** to run the sort operation.
7. Save and run the workbook as **NewsAndBlogsDataSorted**.
   When the run completes, you will see more languages in the workbook.
   Now you will create a graph to visualize your results.
8. Within the **NewsAndBlogsDataSorted** workbook, expand the **Add chart** dropdown, click **Chart**, and then click **Pie**
9. Provide the following values for the Pie chart:
   - Chart Name: **Language coverage**
   - Title: **Watson coverage by language**
   - Value: **Language**
   - Count: **Count occurrences of X axis values**
   - Sort by: **Count**
   - Occurrences Order: **Descending**
   - Limit: **12**
   - Template: **Soda Cap**
   - Style: **Pie**
10. Click **Apply Settings** to create the chart, and then click **Run**.
    Once the run completes, you will see that English has the largest slice of pie. What is the second most appeared language?

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

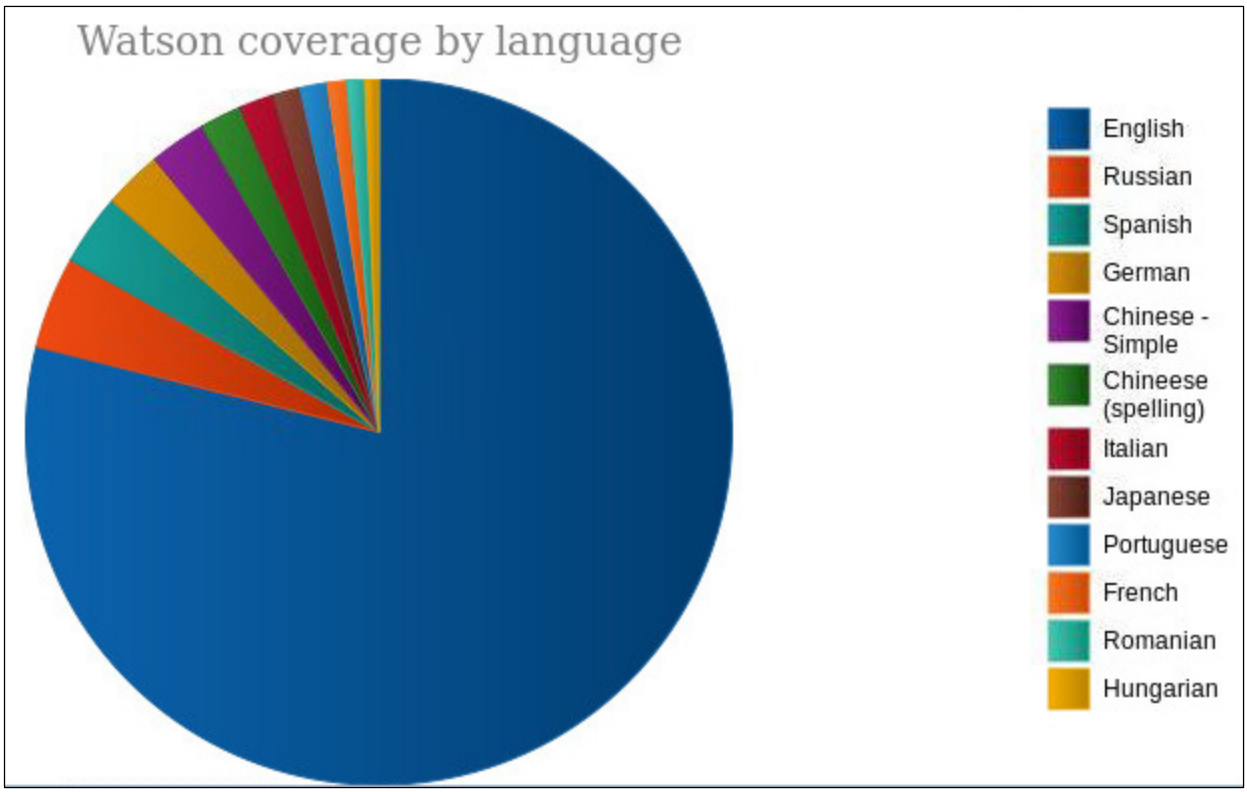
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

11. Point to the second largest piece of the pie.

    Russia has the second largest piece of the pie.

    Move the mouse pointer over the fifth and sixth largest slice and you will see that they're both Chinese. Chinese (Simplified) and Chinese (Spelling). This shows one of the common situations involving data from multiple sources where you may need to do additional refactoring of the data in order to get what you need.

    The results appear as follows:



    In this case, you have multiple entries that you need to treat as identical. For the purpose of our exercise, you will stop here, but if you have some time, you may play around with the different sheets and functions to see other types of operations you can perform on the data.

12. Close all open windows.

Results:
You have created a BigSheets workbook and a chart from it to visualize your data.

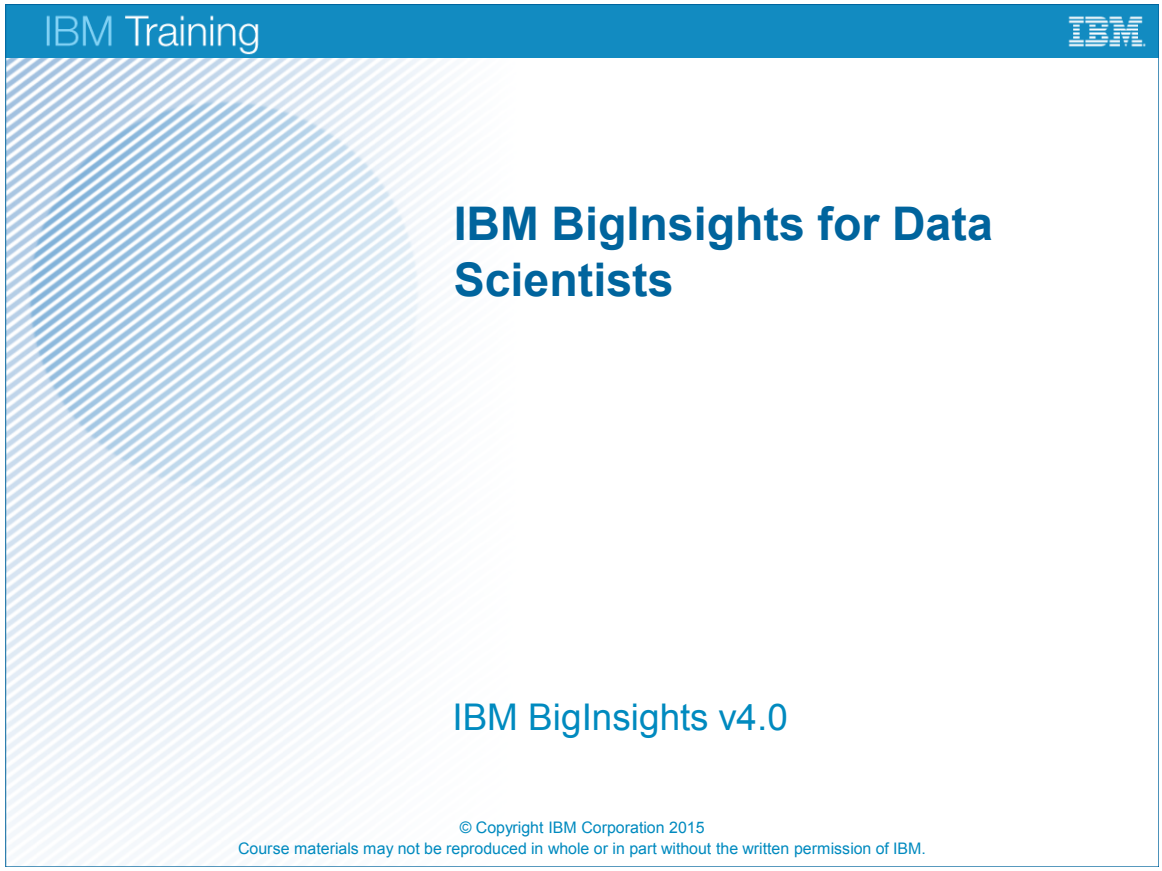
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Unit objectives

- Describe the components that come with the IBM BigInsights Data Scientist module
- Understand the benefits of using text analytics as opposed to coding MapReduce jobs
- Use R / Big R for statistical analysis on the Hadoop scale

© Copyright IBM Corporation 2015

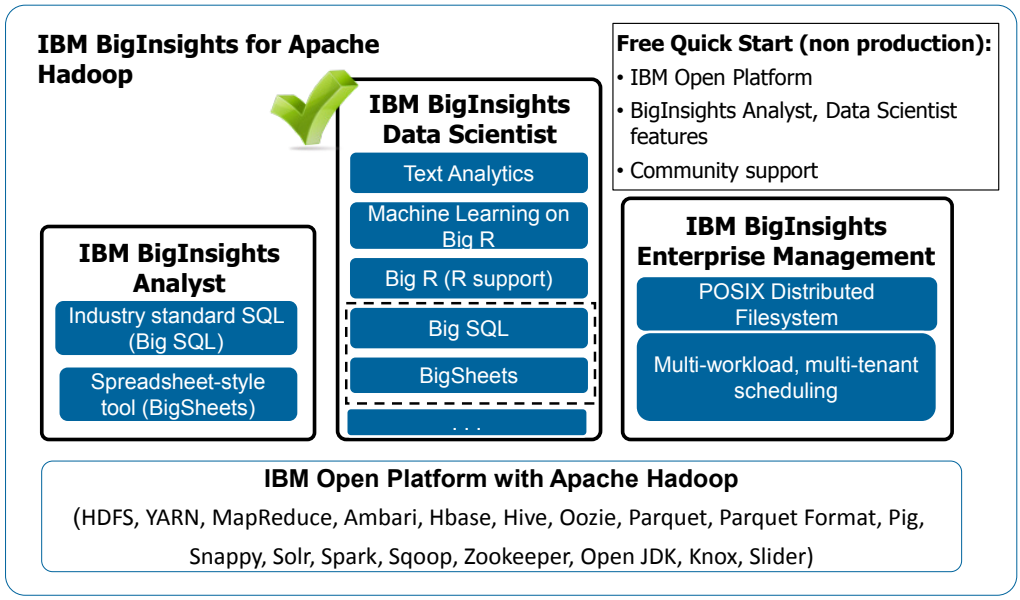
Unit objectives

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Overview of BigInsights



IBM BigInsights for Data Scientists

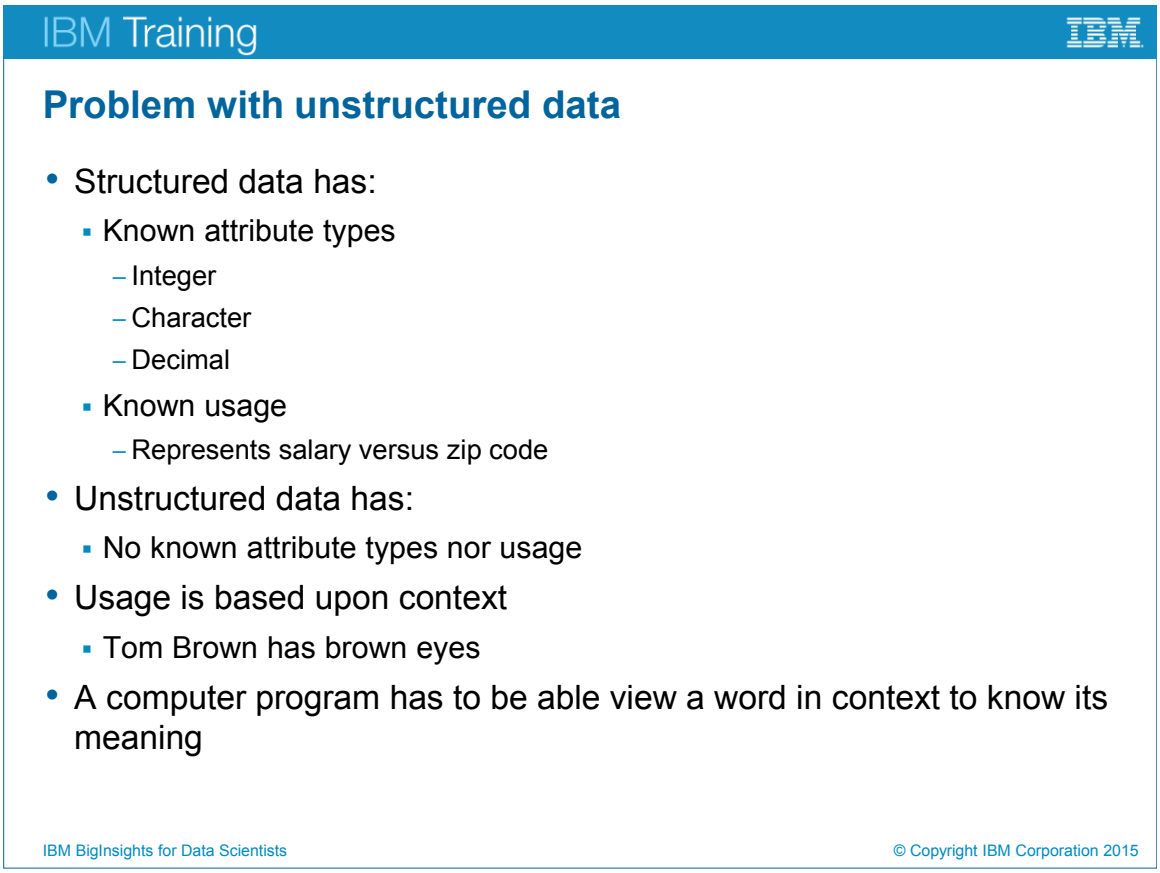
© Copyright IBM Corporation 2015

Overview of BigInsights

This unit will cover the IBM value-adds that comes with the IBM BigInsights Data Scientist module. You saw briefly what they are in the earlier unit on BigInsights. Remember that as part of the Data Scientist module, you will also get Big SQL and BigSheets, which is covered in the unit on the Analyst module.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



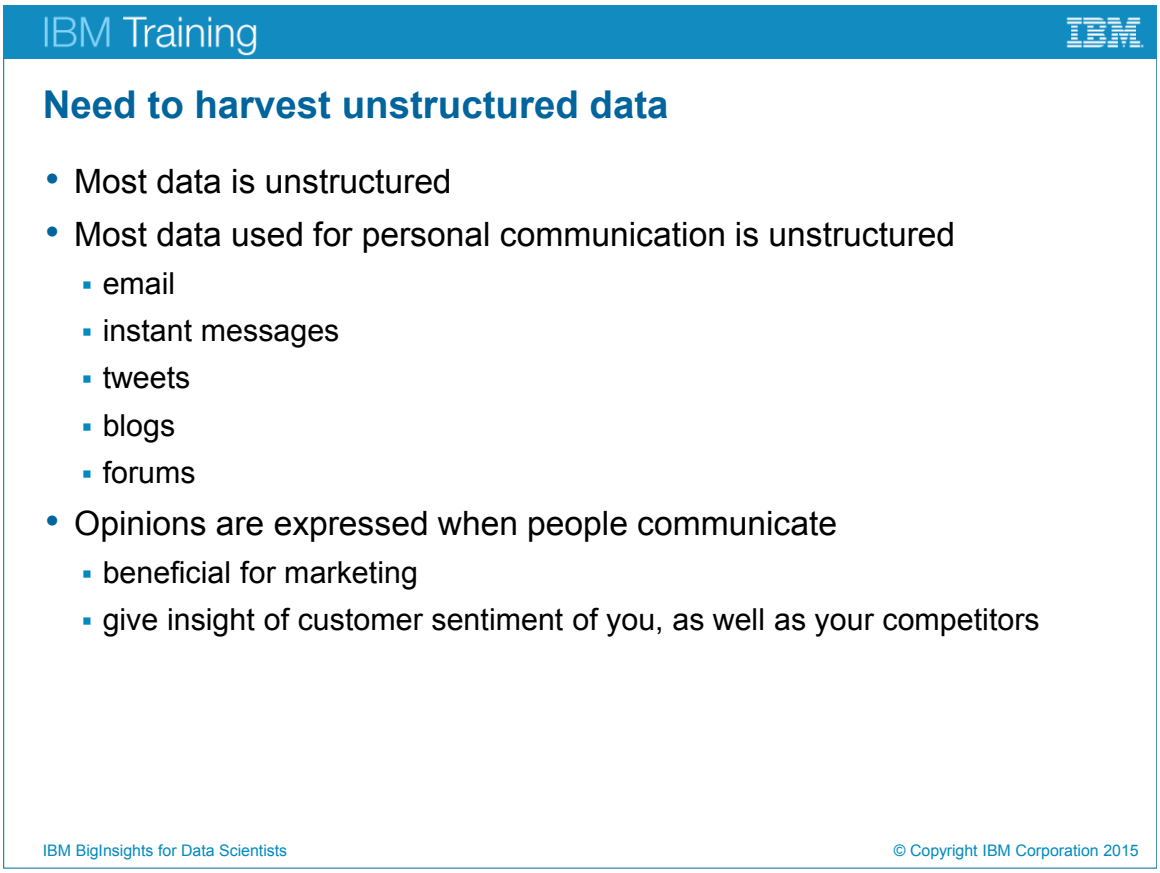
Problem with unstructured data

Computers have been working with structured data from the outset. With structured data you know its attribute type, integer, character, decimal, and you know its usage, such as if it represents a salary value or a zip code. With this knowledge, your program can process the data in a meaningful way.

But unstructured data, by its very nature, does not have known attribute types and data usage. The only way to discern that information is based upon the context usage of the data, which for human beings, usually does not present a problem. But writing a program to do that is extremely difficult. Take the phrase "Tom Brown has brown eyes." It is easy for us to understand that *Brown* is a proper noun and *brown* is an adjective. But context is very difficult for a program to understand. To a program the two words are essentially identical strings of characters.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Need to harvest unstructured data

Although computers have been working with structured data for close to sixty years, the amount of stored structured data is minute when compared to unstructured data. Just think about the number of emails, instant messages, tweets, and word processing documents that are created on a daily basis. Then think about the number of web pages that exist, the number of blogs and forums. And all of that data is unstructured or semi-structured, such as video and audio files.

If all of that unstructured data was meaningless, then even though there is a large amount of it, you would not bother with it. But it is not meaningless. The types of data listed are used for communication. People express opinions when they communicate. And opinions are very important when it comes to marketing. What if you were able to extract opinions about your products from all of this communication data? That would give you a good insight on what customers thought about your products and also how they compared your products with your competitor's.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Need for structured data

- Business intelligence tools work with structured data
  - OLAP
  - data mining
- To use unstructured data with business intelligence tools requires that structured data to be extracted from unstructured and semi-structured data
- IBM BigInsights provides a language, Annotation Query Language (AQL)
  - syntax is similar to that of Structured Query Language (SQL)
  - builds extractors to extract structured data from
    - unstructured data
    - semi-structured data

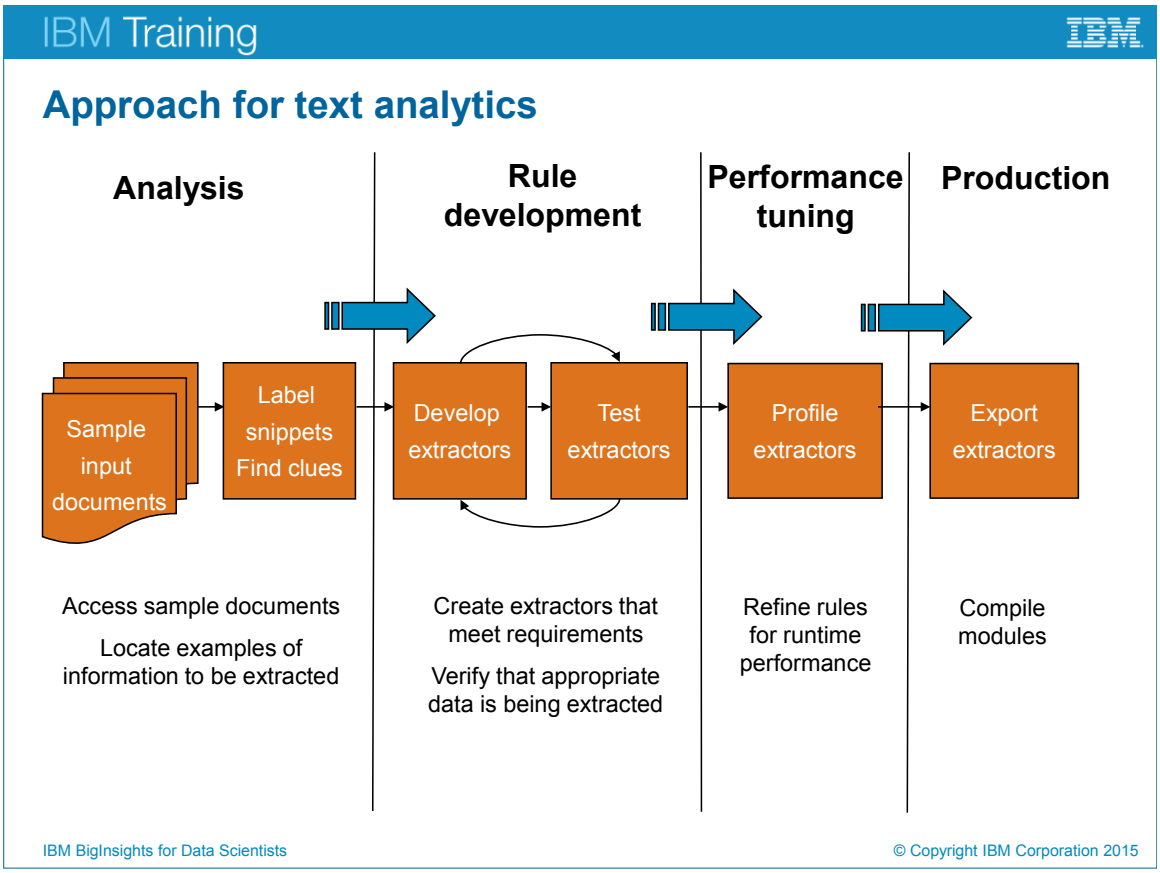
© Copyright IBM Corporation 2015

Need for structured data

The need is to extract structured data from unstructured and semi-structured data. Why? Because business intelligence tools that allow for data analysis, use structured data. This goes for OLAP (online analytical programming), data mining and even for simple spreadsheet analysis. IBM BigInsights provides a language called Annotation Query Language (AQL) that is designed to build extractors to extract structured data from both unstructured and semi-structured data. The syntax of the language was modeled after Structured Query Language (SQL) in order to lessen the learning curve.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

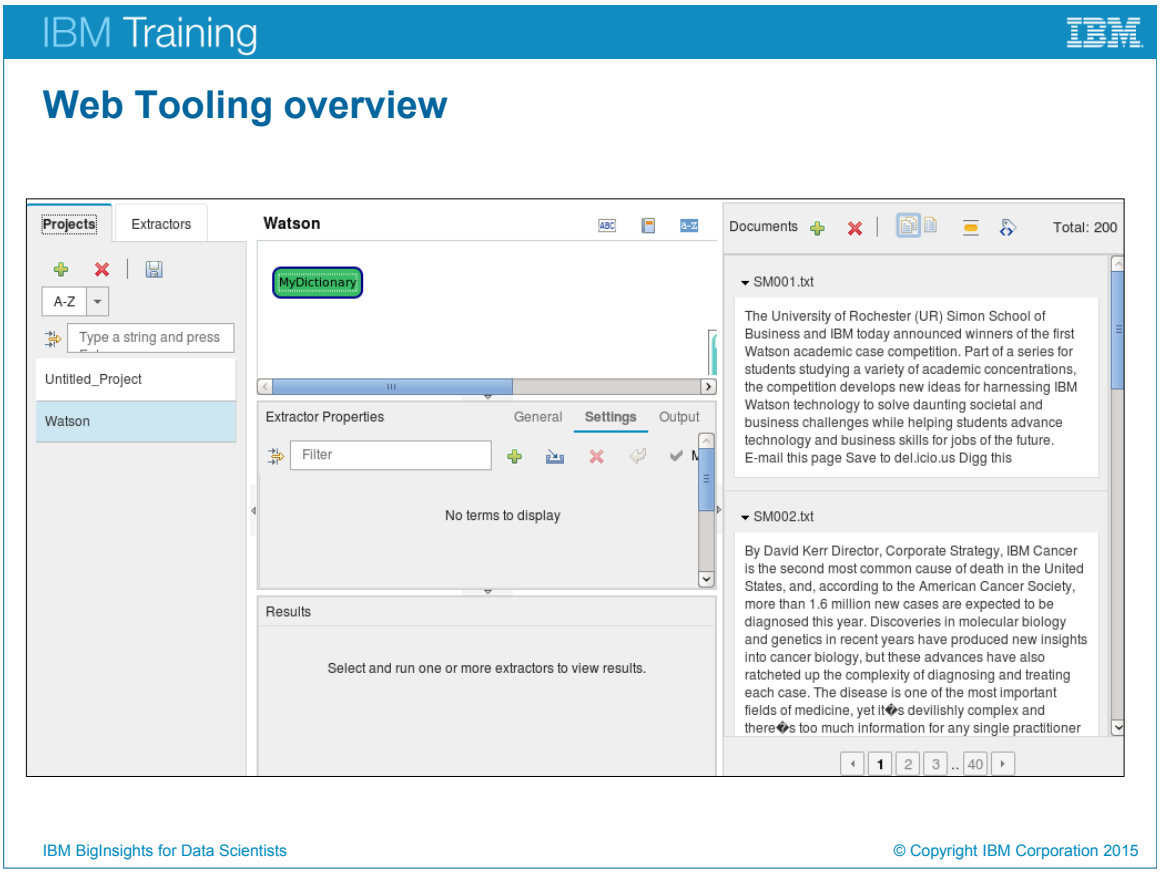


Approach for text analytics

The typical approach for text analytics is to first analyze your document sources, particularly a small subset of the data. You would label and identify keywords and information to be extracted. Then you would develop the extractors and rules that meet certain requirements. Additionally, you would verify that your data being extracted is appropriate to what you need to analyze. For example, if you wanted to extract a particular keyword, such as Watson, you must tell the extractors in which context you want the word Watson, such as in the computing and technology, and not a person's name. Once the extractors have been developed and tested, you would want to performance tune your extractors before exporting to production.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Web Tooling overview

The Web Tooling module that comes with the Data Scientist module allows you to build complex extractors from simple building blocks, such as dictionaries, regular expressions, etc.

There are prebuilt extractors for many use cases:

- sentiment analysis
- extracting financial data
- generic structures like phone numbers, addresses, people, etc.

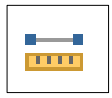
This is presented in a simple drag-and-drop canvas interface. You can also save and export extractors as AQL code.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Basic components of an extractor

- **Literal**
  - Match a single term 
- **Dictionary**
  - Match from a list of terms
  - Case sensitive or insensitive
  - Can be imported from a text file 
  - You can match multiple terms to the same term with a **mapping table**
    - Example: match personal names to common nicknames
- **Regular Expression** 
- **Proximity Rule**
  - Extract spans that occur within specified distance of each other
  - Distance measured in characters or tokens (words) 

IBM BigInsights for Data Scientists © Copyright IBM Corporation 2015

Basic components of an extractor

These are the basic components of building an extractor. A literal is used to match a single term from your documents. A dictionary is use to match a list of terms. You can define your own dictionary or you can import existing dictionaries from a text file. You can also map multiple terms to a single term with a mapping table. For example, match a person's name to nicknames or aliases. You can also create regular expression extractors to match specific patterns. Proximity rules allows you to specify the span within a certain distance of the matched tokens.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

What is open source R?

- R is a powerful programming language and environment for statistical computing and graphics.
- R offers a rich analytics ecosystem:
  - Full analytics life-cycle
    - Data exploration
    - Statistical analysis
    - Modeling, machine learning, simulations
    - Visualization
  - Highly extensible via user-submitted packages
    - Tap into innovation pipeline contributed to by highly-regarded statisticians
    - Currently 4700+ statistical packages in the repository
    - Easily accessible via CRAN, the Comprehensive R Archive Network
  - R is the fastest growing data analysis software
    - Deeply knowledgeable and supportive analytics community
    - The most popular software used in data analysis competitions
    - Gaining speed in corporate, government, and academic settings

© Copyright IBM Corporation 2015

What is open source R?

R is a GNU project.

CRAN is Comprehensive R Archive Network.

Open source R is the most popular and fastest growing software for data analysis. It is a powerful programming language and environment for statistical computing that is gaining speed across academia and industry. The R ecosystem offers thousands of freely available packages for data exploration, analysis, modelling, and visualization. Many of these packages are created by world-leading statisticians and mathematicians. All packages are publically available via CRAN, an online repository.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

The R appeal: what attracts users?

- R is an integrated suite of software facilities
  - Simple and effective programming language
  - Variety of open source GUIs for increased productivity
  - Publication-quality graphics capabilities
- Cutting edge algorithms
  - Statisticians usually first contribute their algorithms to R
  - Contributors are often working on today's most challenging data analysis
  - Algorithms developed for life sciences, finance, marketing, etc.
- Accessibility and education
  - Open source with many free online educational tools to help you learn R
  - Universities often teach data science skills with R
  - The network effect of R and its highly extensible packages system

© Copyright IBM Corporation 2015

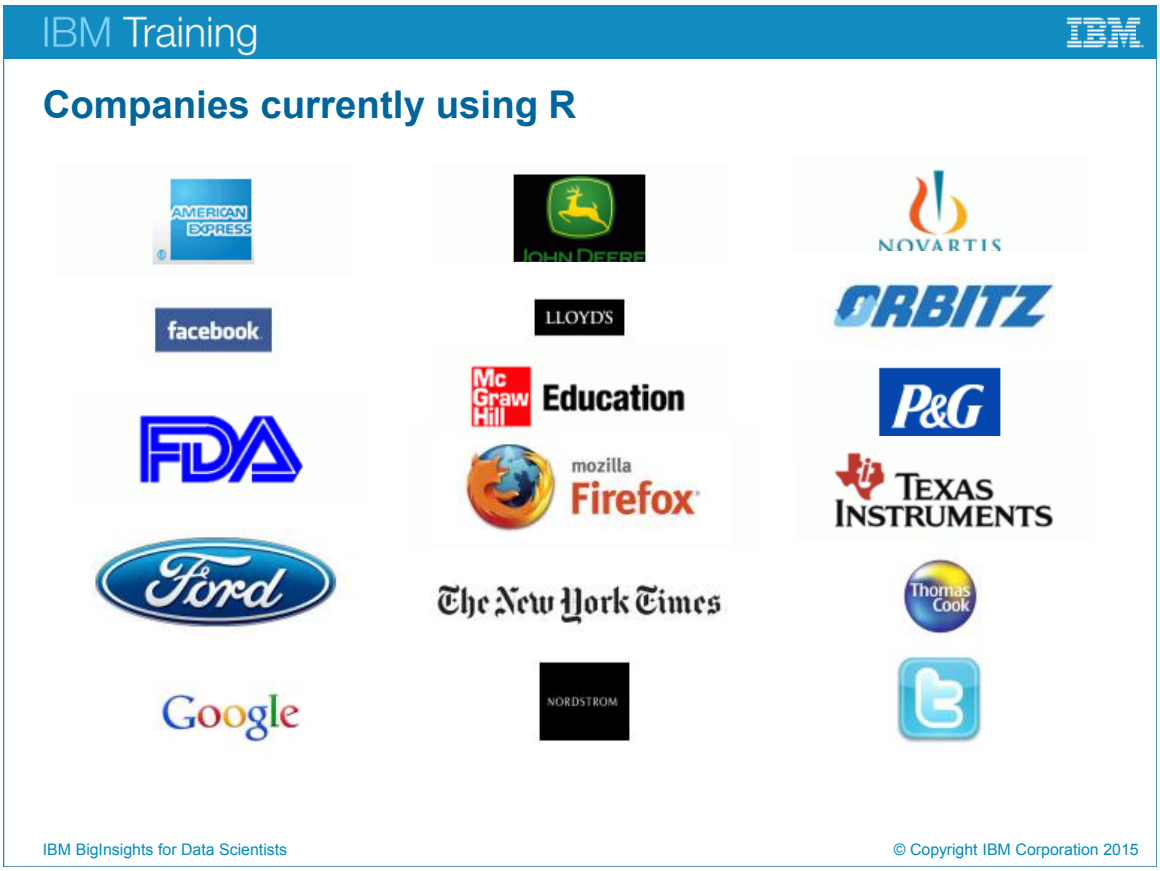
The R appeal: what attracts users?

Network effect: as number of R users increases, so does the value of the R ecosystem

Users are attracted to R for various reasons, such as its ease of access and that it is simple to use. R has an integrated suite of tools to facilitate software development, such as RStudio which is an open source IDE enabling you to access all of the development tools you need from an easy to use interface. R scripts are also very simple to write, using high-level language and high-quality existing functions that are freely available from the open source community. The lively and passionate R community continues to gain speed, which increasingly adds to the value of R. There is a broad sweeping open source movement in statistics and mathematics software that is changing the analytics industry. Users continue to flock to R to enable performing advanced analytics on their data assets.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Companies currently using R

There are many companies that are using open source R today within their research and operations. As exploiting data for competitive advantage is being increasingly important, we are seeing a wide range of industries adopting R. A non-exhaustive list of companies using open source R today includes:

- American Express
- Facebook
- US Food & Drug Administration
- Ford
- Genentech
- Google
- John Deer
- Lloyd's
- McGraw Hill
- Mozilla
- The New York Times
- Nordstrom
- Novartis
- Orbitz
- Procter & Gamble
- Texas Instruments
- Thomas Cook
- Twitter

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

What is the R programming language?

- Multi-paradigm programming language
  - Designed from the ground-up for statistical computation and graphics
  - http://cran.r-project.org/manuals.html
- Interactive, functional programming semantics
  - Allows "computing on the language"
  - Systematize repetitive work with functions, packages, scripts
- Simple expressions, with strong support for object orientation
  - Incremental and interactive addition of user-defined object orientation
  - More support for OOP than other statistics languages

© Copyright IBM Corporation 2015

What is the R programming language?

Computing on the language: makes it possible to write functions that take expressions as input.

R is an easy to use, high-level programming language that borrows techniques from a variety of languages. It is a domain-specific language (DSL) that was designed from the ground-up for statistical computation and graphics. It provides the user with ways to quickly develop and maintain their desired solution with support for functions, packages, and scripts. These scripts can be anything from simple expressions to complex object oriented code.

R is an interpretive language and includes interactive and functional programming semantics allowing for users to compute on the language. This makes is easy to use while also providing for more support for object-oriented programming (OOP) than other statistics language.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Limitations of open source R

- R was originally created as a single user tool
  - Not naturally designed for parallelism
  - Can not easily leverage modern multi-core CPUs
- Big data > RAM
  - R is designed to run in-memory on a shared memory machine
  - Constrains the size of the data that you can reasonably work with
- Memory capacity limitation
  - Forces R users to use smaller datasets
  - Sampling can lead to inaccurate or sub-optimal analysis

Key Take-Away
Open Source R is a powerful tool, however, it has limited functionality in terms of **parallelism and memory**, thereby bounding the ability to analyze big data.

© Copyright IBM Corporation 2015

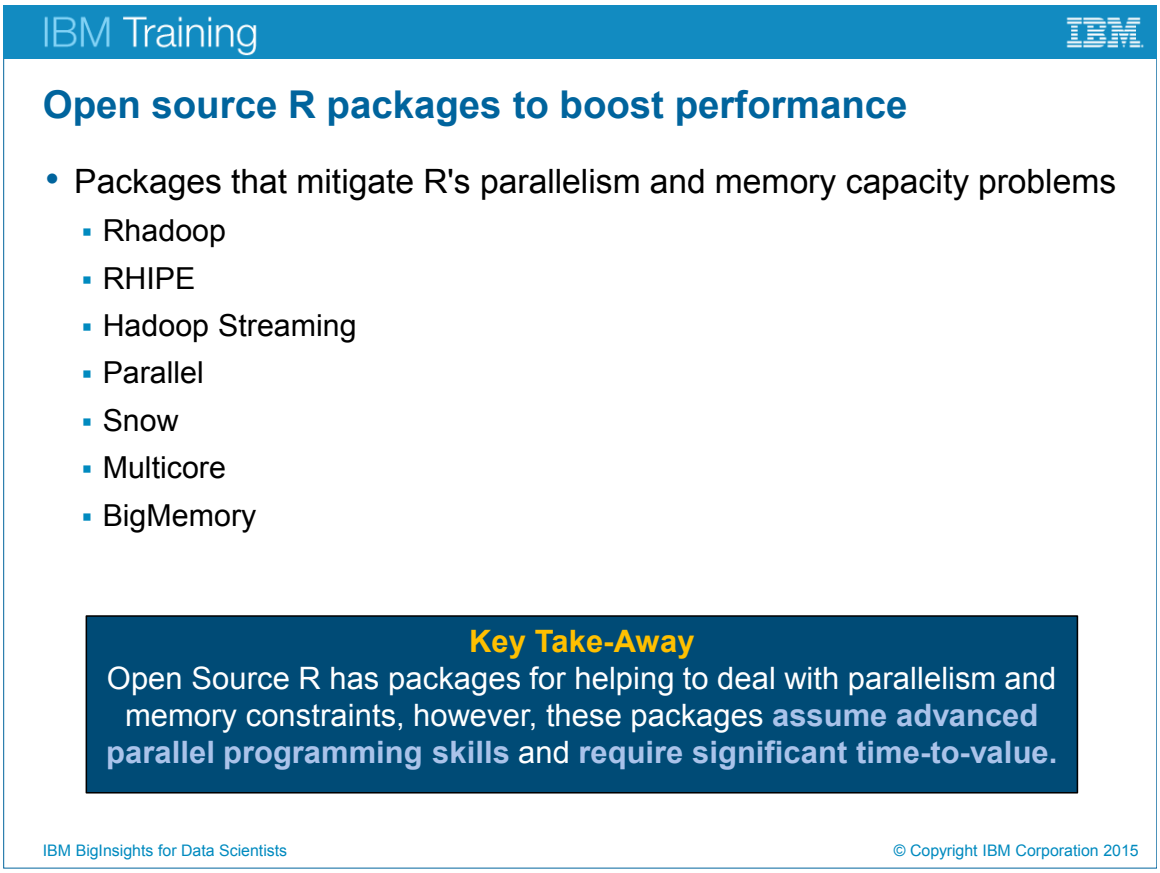
Limitations of Open Source R

RAM: Random-access memory, or commonly referred to as memory.

Although R offers the user a wide range of tools to tackle almost any statistics problem, open source R is constrained to running on small problem sizes. R was originally created as a single user tool that was to be used on something like a laptop and was not naturally designed for parallelism or distributed memory computation. Given the increasing size of today's analytics problems, the R user runs into a RAM barrier, constraining the size of the data that they can work with. This forces users to sample their data which leads to sub-optimal analysis.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

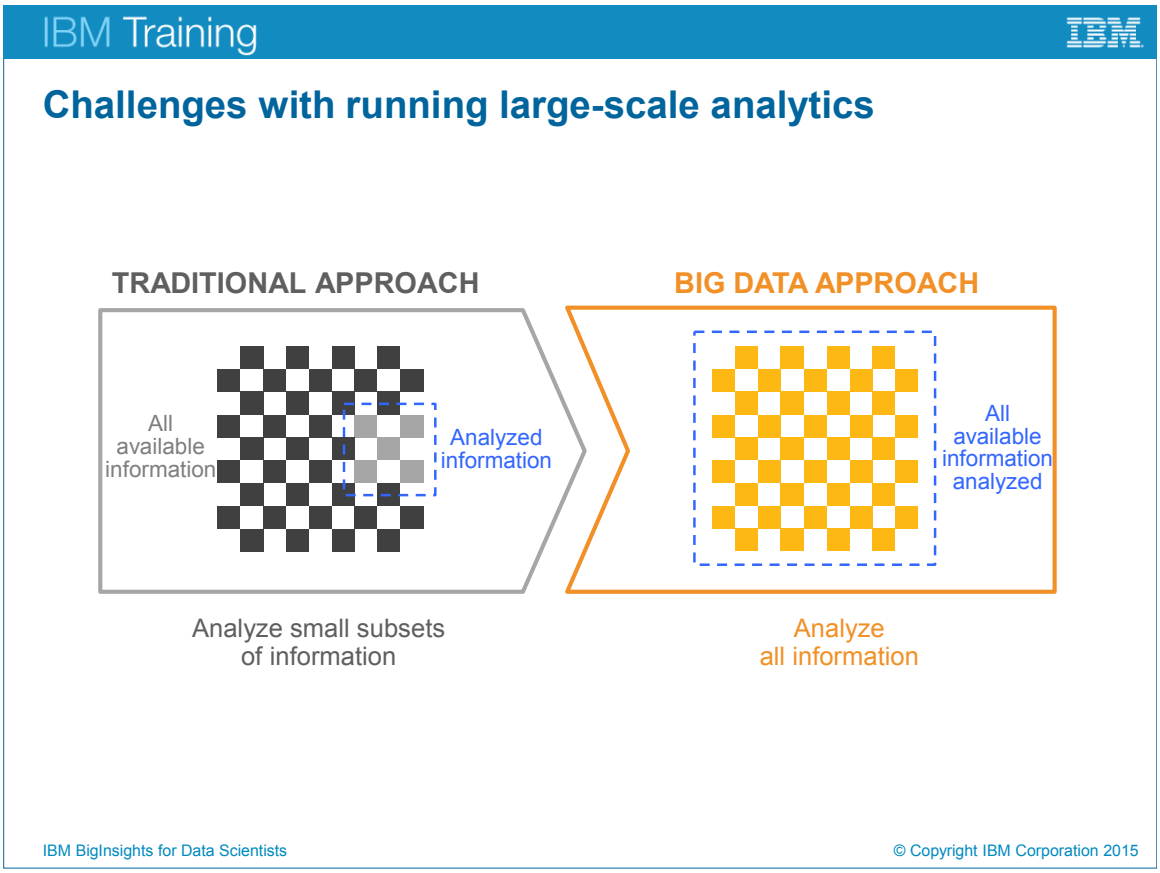


Open source R packages to boost performance

To mitigate R's lack of natural ability for dealing with parallelism and memory capacity constraints, there have been several open source packages developed for both shared-memory and distributed-memory systems. With these open source packages you will need to hand-code parallelism, requiring advanced parallel programming skills and this requires significant time-to-value.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Challenges with running large-scale analytics

There are many challenges in running large-scale analytics. Growing data volumes make it difficult to perform whole population analytics. Many types of software, such as R, are naturally built to run in memory on a single node on a single thread (of a CPU core). For data sets that are larger than what a system can reasonably load into memory or process using a single thread, this forces the data scientist to process only a sample from this large data set. This sample then becomes the basis of their analysis. This can lead to many problems. For example, the attribute set often contains skew. So, if the sample is not representative of the entire population, then the statistical analysis and models built by the data scientist might not apply to the whole population. Models that do not generalize well are of little use.

Another problem in running large-scale analytics that seems to surprise many people is that an algorithm might be numerically stable on a small dataset, but once you scale to big data, the algorithm might no longer might work well. In fact, the degree of error in the analysis is amplified as the volumes of data increase.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

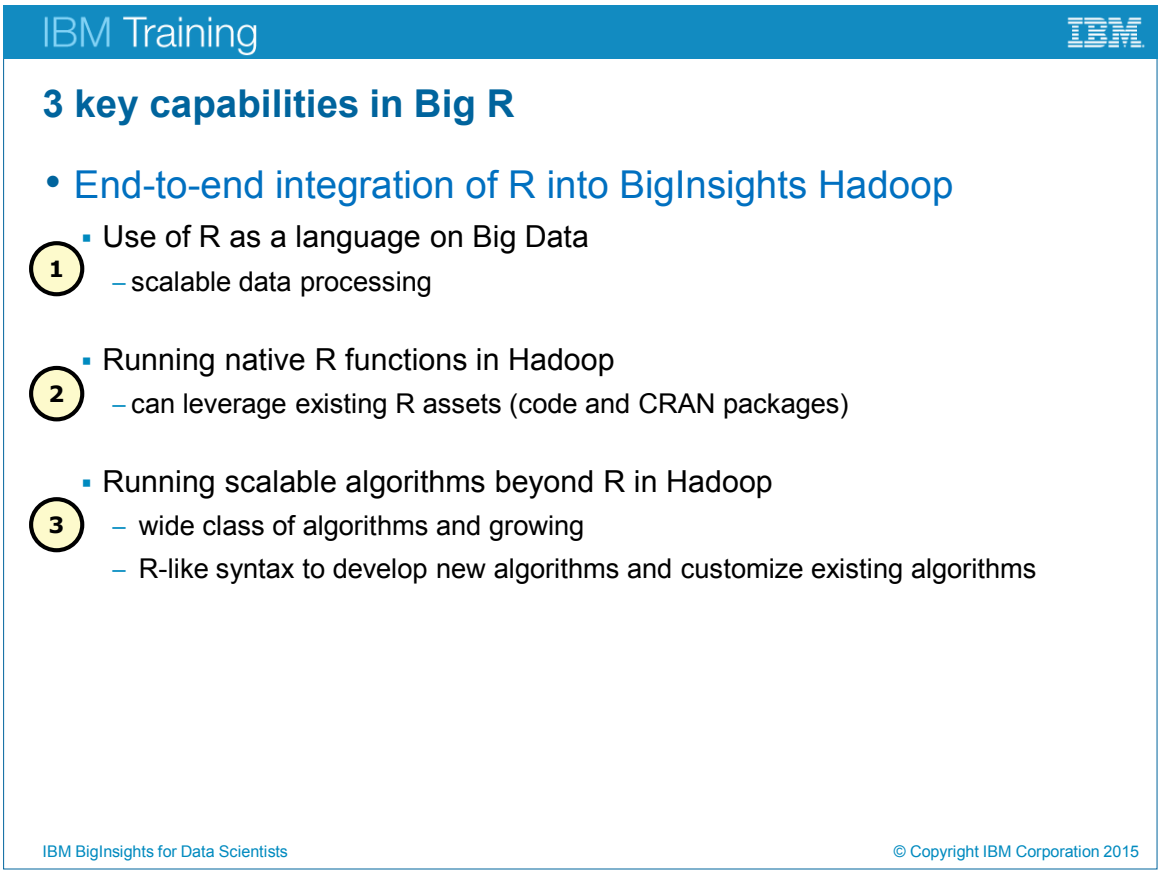
These types of problems with running large-scale analytics are well known to users of R; open source R is a statistical computing environment and is the most popular tool in data science today, with a vibrant and growing community. R is used across industries and academia for performing cost-effective statistical analysis. An important part of the open source R community is the CRAN repository. CRAN provides R users with over 4500+ freely available statistics packages. If you come up with your own improvements or your own algorithms, you can post these to CRAN for the world to use. These packages vary in quality, but it is important to note that many of the world's leading statisticians and researchers use R to develop their bleeding edge algorithms. R enables them to get their algorithms out into the hands of statisticians as quickly as possible. Many corporations actually have R packages that they are active contributors to and/or stewards of. R is an essential piece of many analytics workflows.

Although open source R has many, many benefits, it does have some restrictions. For example, R was originally conceived as a single user tool. It is naturally single threaded on a single node. Therefore, you can you only perform your analysis if both your dataset and the accompanying computational requirements will all fit into memory. In the context of big data, R and Hadoop are not naturally friends. The desire to be able to run R in Hadoop is the reason why IBM developed Big R for BigInsights. Big R **extends** open source R so that it can run with Hadoop.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



3 key capabilities in Big R

There are three main capabilities that the Big R package provides to the user. First, it enables the user to leverage R as a query language on Big Data. Data scientists spend approximately 70% of their time on data processing, and the first capability enables the user to seamlessly perform this analysis across the cluster, using the Big R package. It provides the required data connectivity to Hadoop and a set of scalable functions that enable data exploration, HDFS navigation, feature engineering, basic statistics, and more. Many of these functions are actually overloaded R functions.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

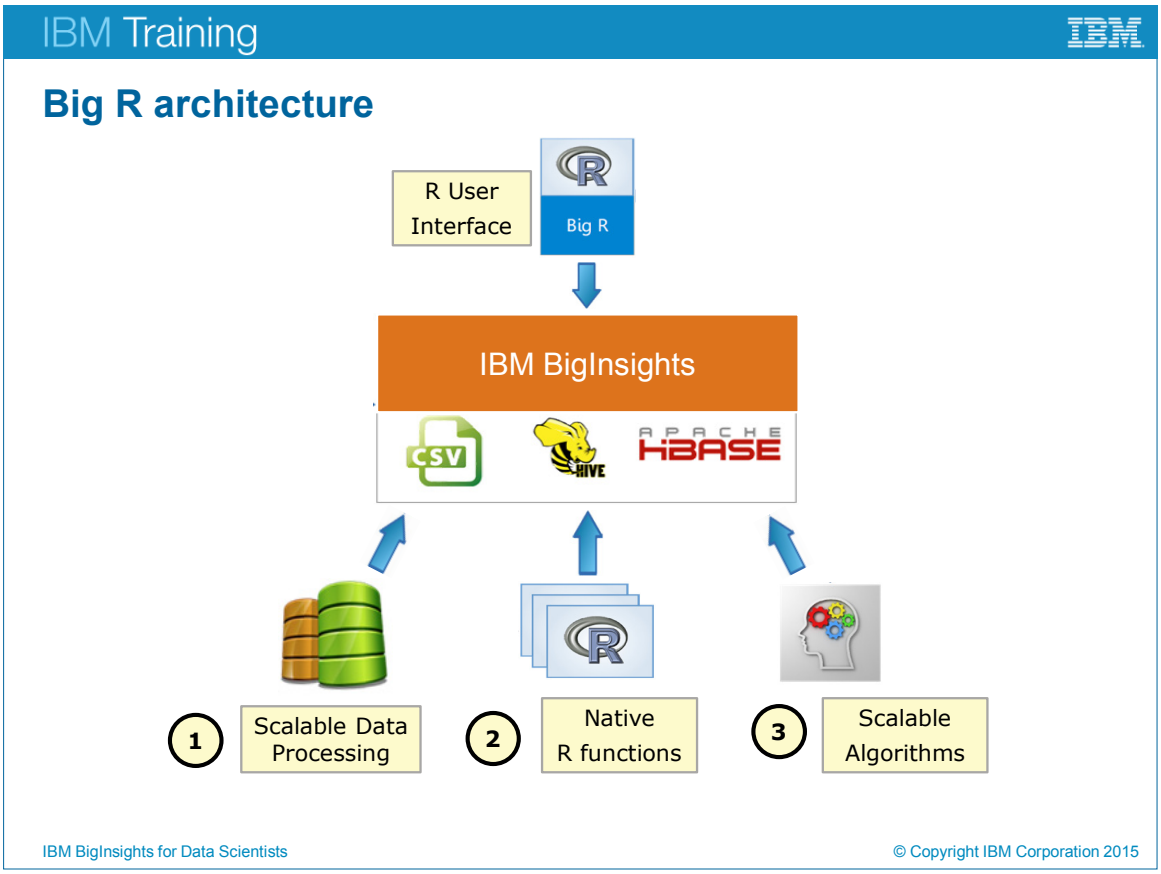
Big R also enables the user to leverage their existing R assets and push them to the cluster. You can take code snippets from your existing R code base and CRAN packages and push those to partitions of the data within the cluster for data parallel execution. This is called "partitioned execution", which follows the same "Apply approach" that is used in open source R. Through this method, Big R automatically stands up multiple instances of R in the Hadoop cluster, based on how you need to partition the data. Each of these instances of R will run your desired analysis, and then you can pull the analysis back to your client for further exploration or visualization. A common use case here is when the data scientist wants to build models on multiple subsets of the data. For example, you may build a decision tree model for each product category of interest. Beyond data parallelism, this flexible partitioned execution mechanism can also perform task parallelism. An example of task parallelism is when the data scientist wants to concurrent simulation of a given modelling technique (such as exploring the parameter space), and then use the best model that they find.

So although you can scale out your native R analysis across multiple nodes in your Hadoop cluster, through this approach you are still confined to the memory restrictions of R. In the situations where you need to run statistical analysis and machine learning beyond R in Hadoop, you can call Big R's wide array of scalable algorithms. Big R comes with a set of prepackaged scalable algorithms. These are written in an R-like declarative language under-the-hood, and can run optimally at any scale. Since they are declarative, they are compiled for automatic parallelization and optimization based on data characteristics and the Hadoop configuration. This means, automatic performance tuning, something that is very key when running analysis on Hadoop. In the future, IBM plans on opening up this R-like declarative language to the user so that they can tweak existing prepackaged algorithms, as well as provide the ability to write their own custom algorithms in a fairly R-like language that will automatically parallelize and optimize for the computation at hand. It seems that no one else is attempting to build anything anywhere close to this level of value and sophistication for the data scientist. Again, the value that these scalable algorithms bring is optimization for high performance, and flexibility to enable data scientists to customize these (or their own) algorithms. And you get all this capability from your favorite R tool!

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Big R architecture

The Big R architecture makes a lot of sense to the R user. R sits on your client using a normal R client, such as RStudio, as the IDE. This IDE will usually be on your data scientist's laptop. The Big R package itself will be installed on your client as well as the nodes on the cluster. From your client, Big R provides you with data connectivity to several different data sources within Hadoop, any delimited file (such as CSV), data sources cataloged by Hive, HBase data, or JSON files. Under the hood, Big R is simply opening a proxy to this entire dataset that is stored in Hadoop. It is not actually moving the data, which is obviously very important when dealing with large datasets. From the user's end, it will look and feel as if all of the data is sitting on their laptop, but obviously due to the data volume, that is not possible, as it is still sitting in Hadoop.

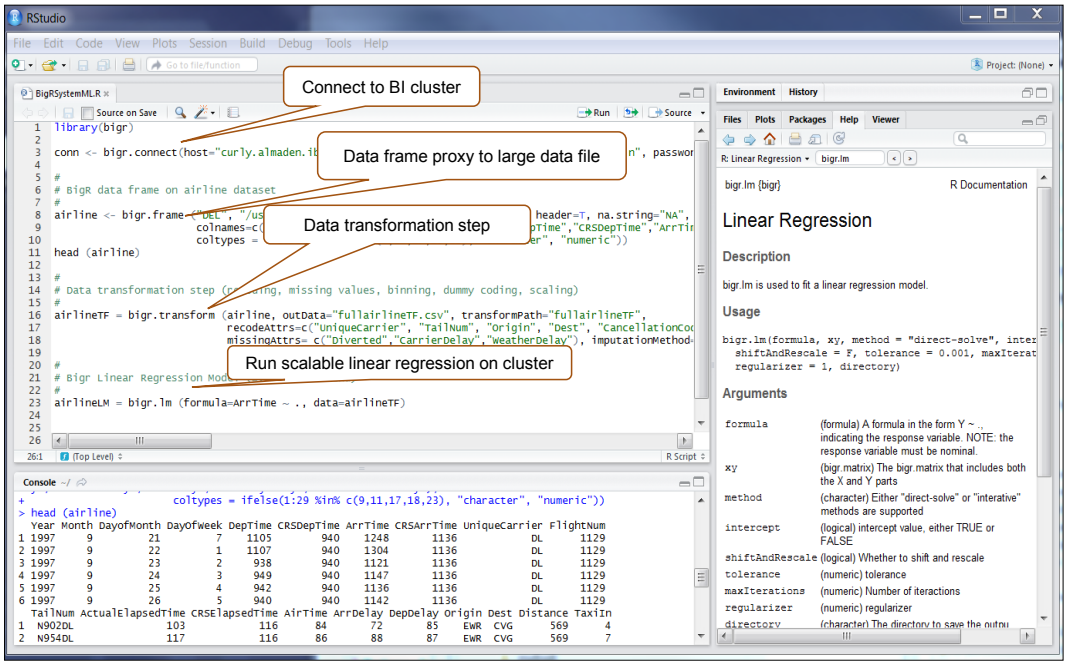
Big R moves the function in your R applications to the data in Hadoop. So those three key capabilities afforded by Big R will all be pushed down into the Hadoop cluster for scalable analysis. In short, you have the ability to perform scalable data processing, wrap up native R functions for parallel execution on the cluster, and run the scalable algorithms that seamless run across entire datasets to build machine learning models and descriptive statistics.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

User experience for Big R



IBM BigInsights for Data Scientists

© Copyright IBM Corporation 2015

User experience for Big R

The user experience for Big R is obviously the same as open source R, since Big R is simply a package that extends open source R to run on Hadoop. Here we see RStudio which is a free, and open source IDE, used by the majority of R users. In the typical working session, the user will need to connect to the BigInsights cluster and open a proxy to the dataset of interest. From there they can perform the three main categories of capabilities that Big R provides to the user.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

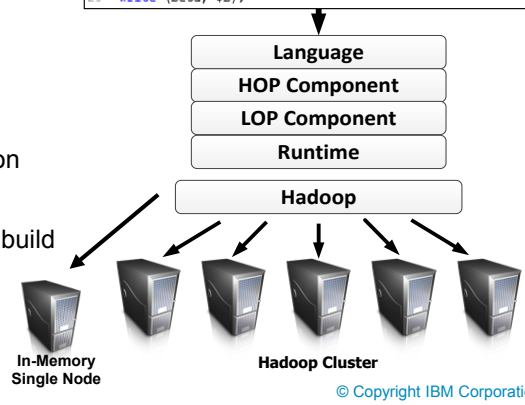
IBM Training

What's behind running Big R's scalable algorithms?

Declarative analytics:
1) Future-proof algorithm investment
2) Automatic performance tuning

- High-level declarative language with R-like syntax shields your algorithm investment from platform progression
- Cost-based compilation of algorithms to generate execution plans
  - Compilation and parallelization
    - Based on data characteristics
    - Based on cluster and machine characteristics
  - In-Memory single node and MR execution
- Enable algorithm developer productivity to build additional algorithms (scalability, numeric stability and optimizations)

```
# THIS SCRIPT SOLVES LINEAR REGRESSION USING A
#   DIRECT SOLVER FOR (X^T X + lambda) and beta = X^T y

X = read ($X);   # explanatory variables
y = read ($Y);   # predicted variables

n = nrow (X);
m = ncol (X);

# Rescale the columns of X if needed
scale_lambda = matrix (1, rows = 1, cols = m);
lambda = t(scale_lambda) * $reg;

# Construct and solve system of equations
A = t(X) %*% X + diag (lambda);
b = t(X) %*% y;

beta = solve (A, b);
...
write (beta, $B);
```



IBM BigInsights for Data Scientists

© Copyright IBM Corporation 2015

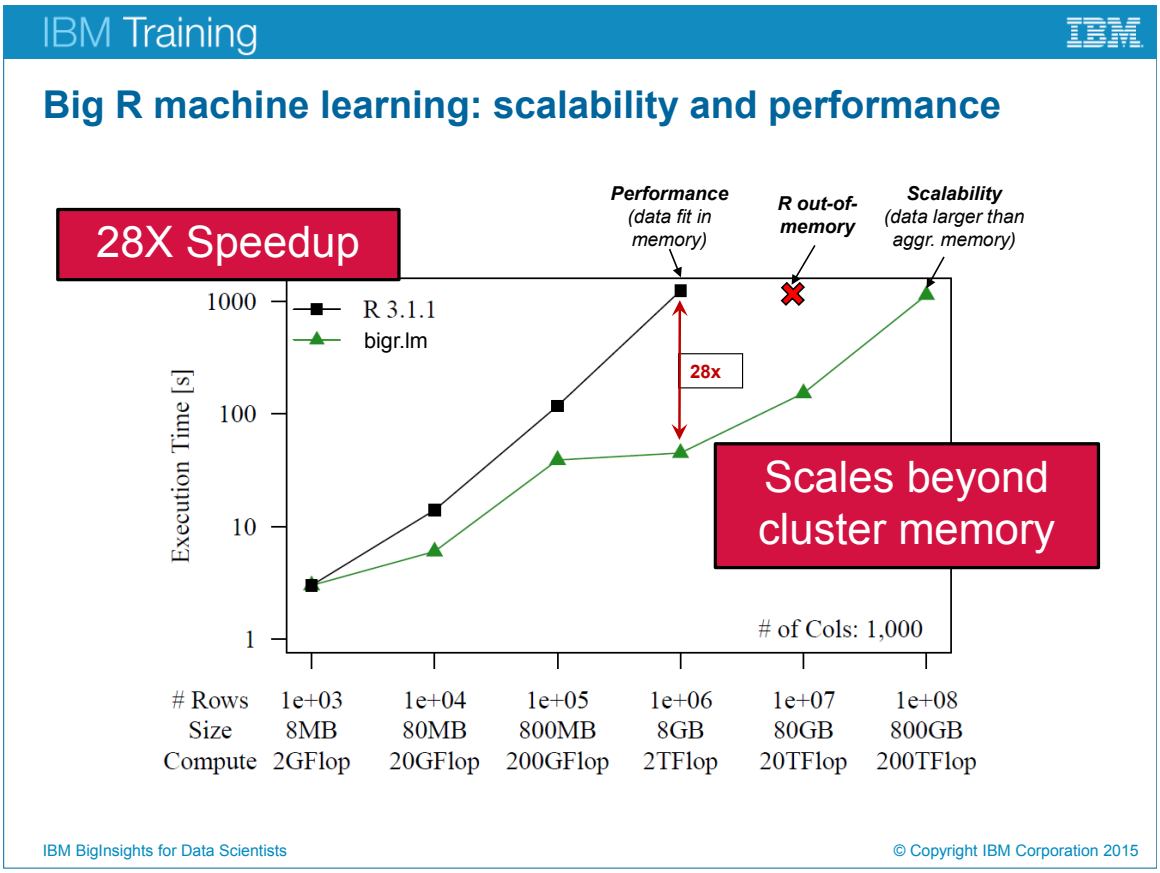
What's behind running Big R's scalable algorithms?

On the upper right in the slide is a linear regression algorithm in the high-level R-like declarative language that is running under the hood in Big R, which looks a lot like both R and MATLAB. This source code goes through a series of optimizations via the compiler and runtime to determine the optimal way to execute the algorithm for the given data characteristics and Hadoop configuration. This execution plan is determined by a cost-based optimizer, much like how queries are automatically generated and tuned in SQL. We really see scalable statistics and machine learning algorithms following the same path that SQL did. SQL started off many years ago with very simple and limited query sophistication. But over time, those query execution plans have become increasing more sophisticated and efficient. Vendors like IBM, Oracle, etc. provide those optimizations and performance tuning to the customer so that the customer can simply focus on using SQL's high-level statements to extract value from their data. Statistics and machine learning should be no different.

In addition to the automatic performance tuning, Big R's high-level language shields your algorithm investment from platform progression. Today Big R compiles down into MapReduce, but if Spark continues to mature, then it is also likely to compile to Spark, without the user having to change their code. This is very important in Hadoop's quickly evolving ecosystem; Big R enables a data scientist to write code today that will fully exploit the Hadoop of tomorrow.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Big R machine learning: scalability and performance

Here we are showing the relative performance of running a linear regression (for example, lm = linear model) using open source R's lm implementation and Big R's lm implementation. The open source R algorithm is the black line (with square markers) and Big R is in green (with triangle markers). Even when running on small data sets that easily fit into memory on a single node, the Big R implementation outperforms open source R. The real power of Big R is seen when scaling out the problem size.

Eventually, open source R will run out of memory; think about all the temporaries that are required throughout the algorithm's execution. The open source R implementation may be able to fit all of the data into memory to begin the computation, but along the way to computing the solution, the temporaries may require additional memory which will cause an out-of-memory error in open source R. You are probably familiar with this problem when using R's packages. Your analysis will have been running fine for a while, and then all of a sudden it crashes. As shown here, Big R's scalable algorithm's do not run into these types of problems; on the far right of the slide, the scalability exceeds the aggregate memory of available because Big R is built to spill to disk.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Simple Big R example

```
# Connect to BigInsights
> bigr.connect(host="192.168.153.219", user="bigr", password="bigr")

# Construct a bigr.frame to access large data set
> air <- bigr.frame(dataSource="DEL", dataPath="airline_demo.csv", …)

# Filter flights delayed by 15+ mins at departure or arrival
> airSubset <- air[air$Cancelled == 0
        & (air$DepDelay >= 15 | air$ArrDelay >= 15),
       c("UniqueCarrier", "Origin", "Dest",
        "DepDelay", "ArrDelay", "CRSElapsedTime")]
# What percentage of flights were delayed overall?
> nrow(airSubset) / nrow(air)
[1] 0.2269586
# What are the longest flights?
> bf <- sort(air, by = air$Distance, decreasing = T)
> bf <- bf[,c("Origin", "Dest", "Distance")]
> head(bf, 3)
Origin Dest Distance
1  HNL JFK   4983
2  EWR HNL   4962
3  HNL EWR   4962
```

IBM BigInsights for Data Scientists © Copyright IBM Corporation 2015

Simple Big R example

Want to know what it's like to work with Big R? In this brief example, you first connect to BigInsights and construct a Big R data frame (basically, an R table for storing data). That frame is loaded with data from one of your data sets, in this case data about flight schedules and departures. Next, you filter the frame for information about delayed flights and isolate the percentage of flights that were delayed. Finally, you determine the longest flights that were delayed.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Unit summary

- Describe the components that come with the IBM BigInsights Data Scientist module
- Understand the benefits of using text analytics as opposed to coding MapReduce jobs
- Use R / Big R for statistical analysis on the Hadoop scale

© Copyright IBM Corporation 2015

Unit summary

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Exercise 1

Working with Text Analytics and R / Big R

© Copyright IBM Corporation 2015

Exercise 1: Working with Text Analytics and R / Big R

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Exercise 1: Working with Text Analytics and R / Big R

Purpose:
You will create a new text analytics web tooling project and load documents to scan for certain keywords. You will start up the R console and run basic commands on it. You will also load the BigR libraries and run basic BigR operations.

Estimated time: **1 hour**
User/Password: **biadmin/biadmin**
root/dalvm3
Services Password: **ibm2blue**

Important: Before doing this exercise, ensure that your access and services are configured and running. Check that:

- /etc/hosts displays your environment's IP address
- in the Ambari console, ensure that all BigInsights services are running

If you are unsure of the steps, please refer to Unit 1, Exercise 1 to ensure that your environment is ready to proceed. You should review the steps in Task 1 (Configure your image) and Task 2 (Start the BigInsights components).

Task 1. Launching the text analytics Web Tooling module.

IBM BigInsights provides a Web Tooling module that makes text analytics easy. In this task, you will see how to use the Web Tooling module to create a project, and load some documents to start the analysis.

1. To open a new terminal, right-click the desktop, and then click **Open in Terminal**.

   You will review the set of files that you will be using for this exercise.
2. Navigate to **/home/biadmin/labfiles/ta/WatsonData/Data/**, and then type `ls` to see the files.

   These are sample blog files by IBM.
3. Launch **Firefox**, and then if necessary, navigate to the **Ambari** login page, **http://ibmclass.localdomain:8080**.
4. Log in to the **Ambari** console as **admin/admin**, and ensure that all of the components have started.
5. Click the **Knox** component, click **Service Actions**, and then click **Start Demo LDAP**.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

6. Click **OK**, and then when the **Start Demo LDAP** process is complete, click **OK** again.

   Leave the Ambari tab open in Firefox for Task 2 of this exercise.
7. To launch the BigInsights home page, open a new browser tab and type:

   https://ibmclass.localdomain:8443/gateway/default/BigInsightsWeb/index.html

   There is a bookmark saved on the toolbar. The id and password is guest / guest-password, but that is also saved for you in the lab envrionment.

   Note: You may need to wait for a minute before the two links display (BigSheets and Text Analytics).
8. Click the **Text Analytics** link to open up the Web Tooling module.

   You are going to create a project and load in some documents to do a text extraction for the Watson keyword.
9. On the **Projects** tab, click **New**  to create a new project.
10. Beside **Name**, type **Watson**, and then click **Create**.

    You will load in a set of documents.
11. In the **Documents** pane, click **Add Documents**  .
12. Click **Browse**, and then navigate to the **biadmin/labfiles/ta/WatsonData/Data/** directory.
13. Shift+click the first and the last document to select all of the documents, click **Open**, and then click **Add**.

    Notice the documents contain XML markup. You can easily remove all the tags at once.
14. Click the **Remove Tags**  button in the upper right corner.

    The goal is to find blogs about the Watson computer. The extractor should look for the word Watson.
15. Click the **New Literal**  button at the top of your Watson project pane.

    A textbox appears in the project pane.
16. Type **Watson** in the textbox, and then press **Enter** to confirm.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

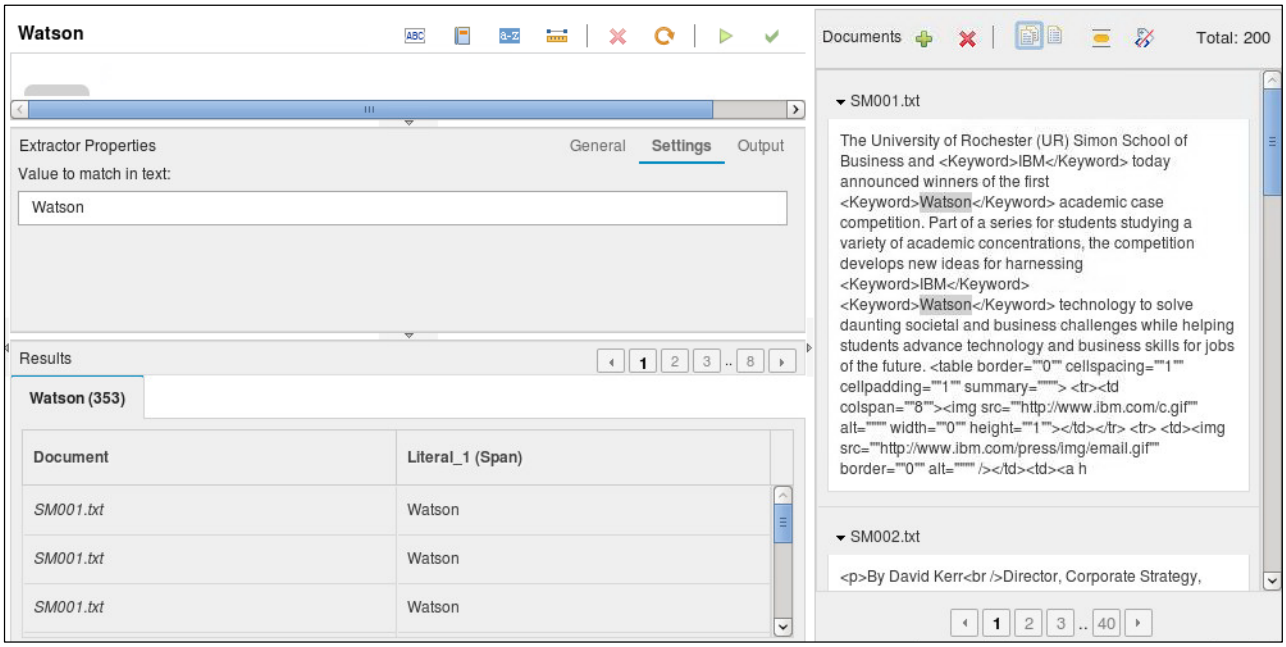
© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

17. Select the **Watson** literal, and then click the **Run Selected**  to run the extractor.

    Hint: You may need to scroll the project pane further to the right to see the rest of the icons.

    It will take a few seconds to run. All of the occurrences of the word Watson will be highlighted in the Documents pane (on the right). You can see details of each match in the Results pane (on the bottom).

    The results appear as follows:



    You will put some context around the word Watson, creating a dictionary of terms frequently associated with the Watson computer.

18. Click **New Dictionary**  in the middle pane, and then type **PositiveClues** in the text box that appears in the canvas.

    Now, you will want to peruse your documents for words. For example, Watson is commonly associated with IBM. You will want to add this to your PositiveClues dictionary.

19. Click on the **PositiveClues** extractor, and select **Settings** (this is located below your canvas).

20. Click **Add Term**  , and then type **IBM**.

    Hint: You can quickly add in words by typing and hitting enter.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

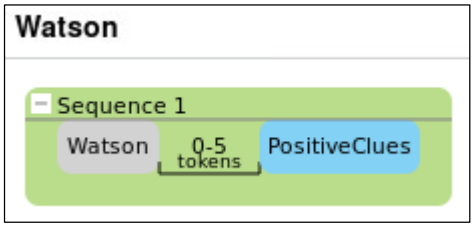
© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

21. Add in **computer**, **computing**, **solutions**, and **technology** as positive clues for the context around the Watson.
22. To run the **PositiveClues** extractor, select it in the canvas, and then click **Run Selected**.

    The words will be highlighted. Review the results pane for details of the search.

    You are going to combine the Watson extractor with the PositiveClues extractor with a proximity rule so that when the clues appear within five words, you know that the Watson keyword is of the correct context.
23. Click **New Proximity Rule**  , and then type **0-5**.
24. Drag the **proximity rule** to the Watson extractor and release it when you see a blue bar, indicating that they are joined together.
25. Drag and drop the **PositiveClues** to the right side of the rule as well.

    The results appear as follows:

    A new sequence is now created with all words of Watsons within five tokens of one of your dictionary terms.
26. Click the **Sequence 1**, and then in the **Extractor Properties** pane, click **Output**.
27. Rename **Sequence 1** to **WatsonSpan**.
28. Rename **Literal_1** to **Watson**.

    Do not rename PositiveClues.
29. Select the extractor, and then click **Run Selected**.
30. Click a few of the rows in the Results pane. You can see the words of Watson that comes with the dictionary words.

    This only extracts the word Watson with the clues that comes after. If you want to have the same context in front of the Watson word as well, what would you need to do?

    Similarly, you may have false positives, such as a person's name of Watson. You would not want to include that in your search. What would you need to do there?

    The last two are left as open exercises for you to try on your own. This exercise serves as an overview and just barely scratches the surface of text analytics with WebTooling.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Task 2. Running R commands and operations.

In Tasks 2 and 3, you will use R and Big R commands via the R Console. An additional component that you can install separately as an IDE for development of R / Big R projects, is RStudio. Open source R have been around, and is great for statistical analysis on small(er) datasets. With big data, Big R is needed to parallelize the operations across the Hadoop clusters in order to improve the accuracy and the model that is generated from the analysis. The goal of this section is to introduce you to Big R through the R Console.

1. Click the **Ambari** tab in Firefox.
2. Make sure that **R** and **Big R** has been started in the **Ambari** console; if not, start it now.
3. To open a new terminal, right-click the desktop, and then click **Open in Terminal**.
4. To start the R Console, type **`R`**.

   Once the console starts up, you should immediately see the R version, as well as the welcome message. You are now in the interactive mode.

   R comes with a base set of functionality, but it can be easily extended by installing R packages. To install R packages the `install.packages()` function is used, but you should not require any new packages during this lab.

   Note: The packages could either be made available within the Comprehensive R Archive Network (CRAN) or from a standard compressed file. For example, if you wanted to connect to a relational database then you could install the RJDBC package using the command `install.packages("RJDBC")`.

   A package is a collection or group of R objects. These functions may contain functions, data structures, links to other libraries, and documentation.
5. To list the install packages in R, as the ">" prompt, type **`library()`**.
6. To exit the *library()* command, type **`q`**.
7. Notice that the *bigr* is an installed package.

   To know more a bit about the packages, use the *help()* command for details on the contents of the packages, functions, and datasets. This *help()* function provides access to the embedded documentation and is very useful to use while learning about the libraries.
8. To learn more about the bigr package, type:

   ```
   help (package='bigr')
   ```

   The help should indicate the version of Big R, as well other information pertaining to the package.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

9. To quit using help, type `q`.

   Packages must be loaded into memory before you can use them.
10. To see which packages are currently loaded into memory, type `search()`.

    This returns an ordered list of packages that are currently in memory and available for our R scripts. Notice that the base package is the final package in the search path.
11. If you want to use the bigr package you would first need to load it into memory using the require() or library() function; to do this, type `require(bigr)`.

    You can use the help() command to get help on functions as well.
12. To fetch the help for the bigr.frame function, type `help(bigr.frame)`.
13. Quit the help function.
14. Another method of obtaining help is by appending a question mark (?) in front of the object. To get help for R's data.frame, type `?data.frame`.
15. Type `q` to quit.

    If you don't remember the exact name of the function you can use the double question mark (??) to search across all of the packages for a match
16. To look for possible histogram functions, type `??histogram`.
17. Type `q` to quit.

    Listing R objects (data structures / functions).
18. To list the current objects within your environment, type `ls()`.

    You will not see many objects if you run the ls() function now, as you have not created anything yet.

    When R is started, the current directory is considered the working area or working directory. Whenever you load or save files from within R, the location is relative to the current R working directory.
19. To verify the working directory, type `getwd()`.
20. To set the working directory to /home/biadmin/labfiles/bigr/labs-r, type:

    `setwd('/home/biadmin/labfiles/bigr/labs-r')`
21. To examine the new working directory, type `dir()`.

    You should see a list of files with various extensions. You can pass regular expressions into this command to only list the files that have the word "plot" in the name.
22. Type: `dir(pattern="*plot*")`.

    R programs or scripts are simple text files and they can be executed from inside the R Console. They can also be executed outside of the console.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

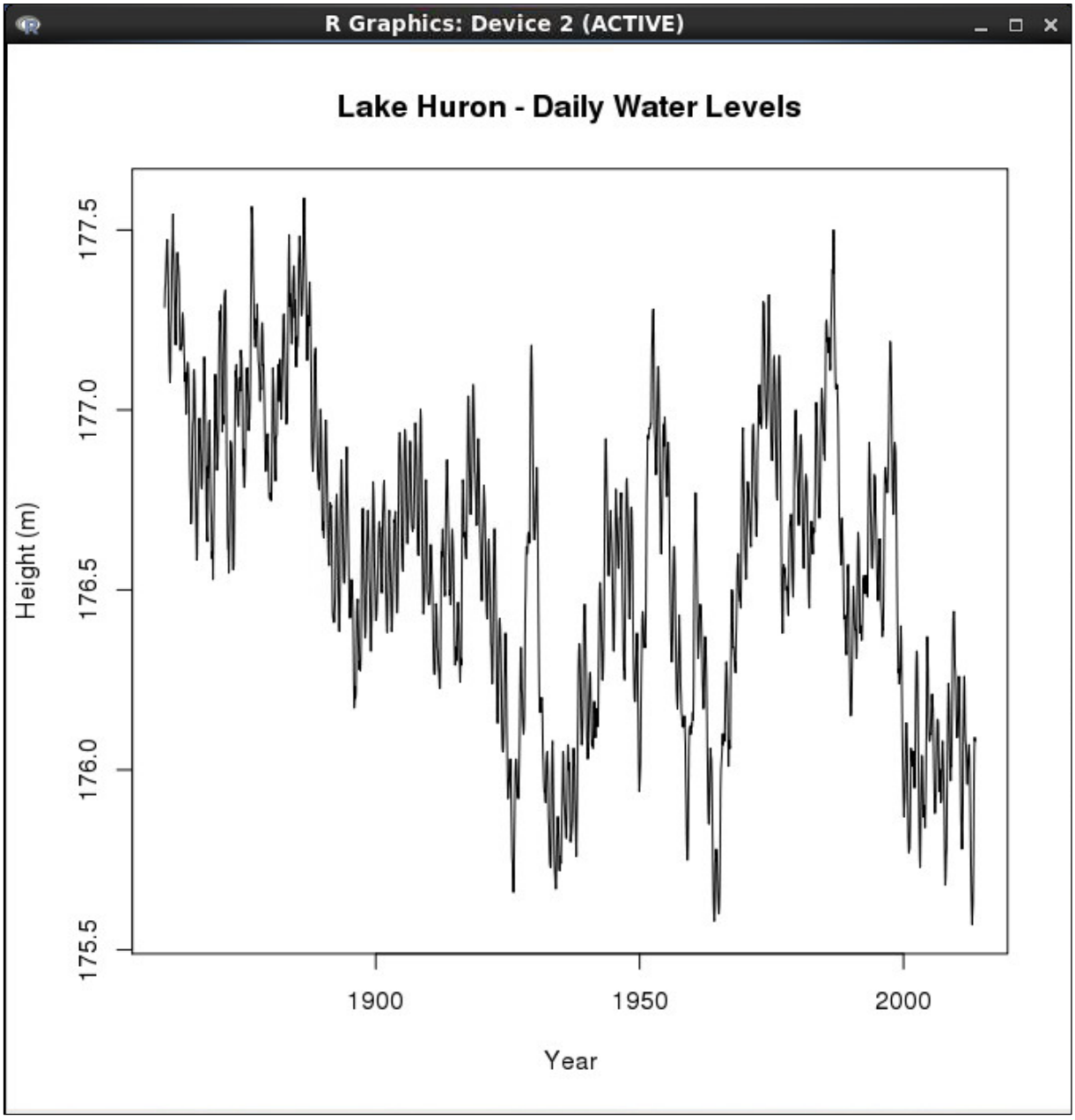
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

23. To will run the script "ex1_huron.R" from within the console using the source() function, type:

```
source("ex1_huron.R")
```

The ex1_huron.R script will generate output to the R console and it will create a graph of the water level of Lake Huron over many years.

The results appear as follows:



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

24. Close the graphics window and return to the R Console.

    The other scripts are provided just for reference if you wish to try them within this course. It will not be covered in this course.

25. To end an R Console session, the q() or quit() function is used; do this now.

    You will be prompted to save the workspace image [y/n/c].

26. At this time you do not want to save any changes, so type **`n`**.

Task 3. Running Big R commands and operations.

This task will get you up and running with Big R using the interactive R Console. RStudio can also be used if you download and install it yourself; that is beyond the scope of this course. Remember that Big R is designed to work with big data, like those stored on the HDFS.

1. Start the R Console, or resume from the previous section.
2. To set the working directory to /home/biadmin/labfiles/bigr/labs-bigr, type:

   ```
   setwd("/home/biadmin/labfiles/bigr/labs-bigr")
   ```

3. To clear the workspace, type **`rm(list = ls()).`**
4. To load the Big R package into your R session, type **`library(bigr).`**
5. To connect to BigInsights, type the following:

   ```
   bigr.connect("ibmclass.localdomain", "bigr", "ibm2blue")
   ```

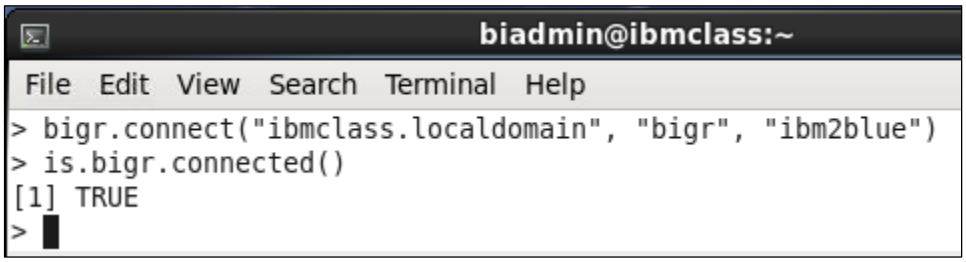
   where you supply the host, the user id, and the password.

   In the lab environment, LDAP has been turned off, so the user id and the password are ignored. Do not do this for the production environment.

   If you are unable to connect, make sure the Big R service is started. Try a restart if it still does not work.

6. To verify that the connection was successful, type **`is.bigr.connected().`**

   The results appear as follows:



Once connected, you will be able to browse the HDFS file system and examine datasets that have already been loaded onto the cluster.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

7. To list the files under root, type **`bigr.listfs()`**:
8. To list the files under /user/biadmin, type **`bigr.listfs("/user/biadmin/")`.**

   You will see some of the files that you uploaded to the hdfs from previous exercises that you have completed.
9. To upload the *airline_lab.csv* file onto the hfds under */user/biadmin/*, open a new terminal and type:

   ```
   hdfs dfs -put /home/biadmin/labfiles/bigr/labs-bigr/airline_lab.csv
   /user/biadmin/
   ```
10. Exit the terminal and return to the R Console.

    You want to connect to a big data set and do some exploration. The airline_lab.csv file is a comma-delimited file (type = "DEL"). You will create a bigr.frame over the dataset. A bigr.frame is an R object that mimics R's own data.frame. However, unlike R, a bigr.frame does not load that data in memory, as that would be impractical. The data stays in HDFS. However, you will still be able to explore this data using the Big R API.
11. Type:

    ```
    air <- bigr.frame(dataSource="DEL", header=T,
    dataPath="/user/biadmin/airline_lab.csv")
    ```
12. This creates an object air that is of bigr.frame; to check out the air object, type **`class(air).`**
13. Examine the structure of the dataset.

    Note that the output looks very similar to R's data.frames. The dataset has 29 variables (for example, columns). The first few values of each column are also shown.
14. To examine the columns and see what they may possibly represent, type **`str(air).`**

    Notice that the column types are all "character" (abbreviated as "chr"). Unless specified otherwise, Big R automatically assumes all data to be strings. However, only columns Year (1), Month (2), UniqueCarrier (9), TailNum (11), Origin (17), Dest (18), CancellationCode (23) are strings, while the rest are numbers. You will assign the correct column types.
15. To build a vector that holds the column types for all columns,type:

    ```
    ct <- ifelse(1:29 %in% c(1,2,9,11,17,18,23), "character", "integer")
    print (ct)
    ```

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

16. To assign the column types, type **`coltypes(air) <- ct`**.

    This data originally comes from US Department of Transportation (http://www.rita.dot.gov), and it provides us information on every US flight over the past couple of decades. The original data has approximately 125+ million rows. For this lab, you are only using a small sample; you will examine the dimensions of the dataset.

17. To get the number of rows (flights) in the data, type **`nrow(air)`**.

    Warning: Some of these commands may take a while to execute.

18. To get the number of columns (attributes) recorded for each flight, type: **`ncol(air)`**.

    You will summarize some key columns to gain further understanding of this data. You will see that the years range from 1987-2008.

19. Type:

    ```
    summary(air[, c("Year", "Month", "UniqueCarrier")])
    ```

    Summarizing columns (vectors) one by one will give you additional information. In some cases, you will also visualize the information. The following statement returns the distribution of flights by year. Again, you have 22 years worth of data. What you will see is a vector that has the "year" for the name, and the flight count for the values.

20. Type:

    ```
    summary(air$Year)
    ```

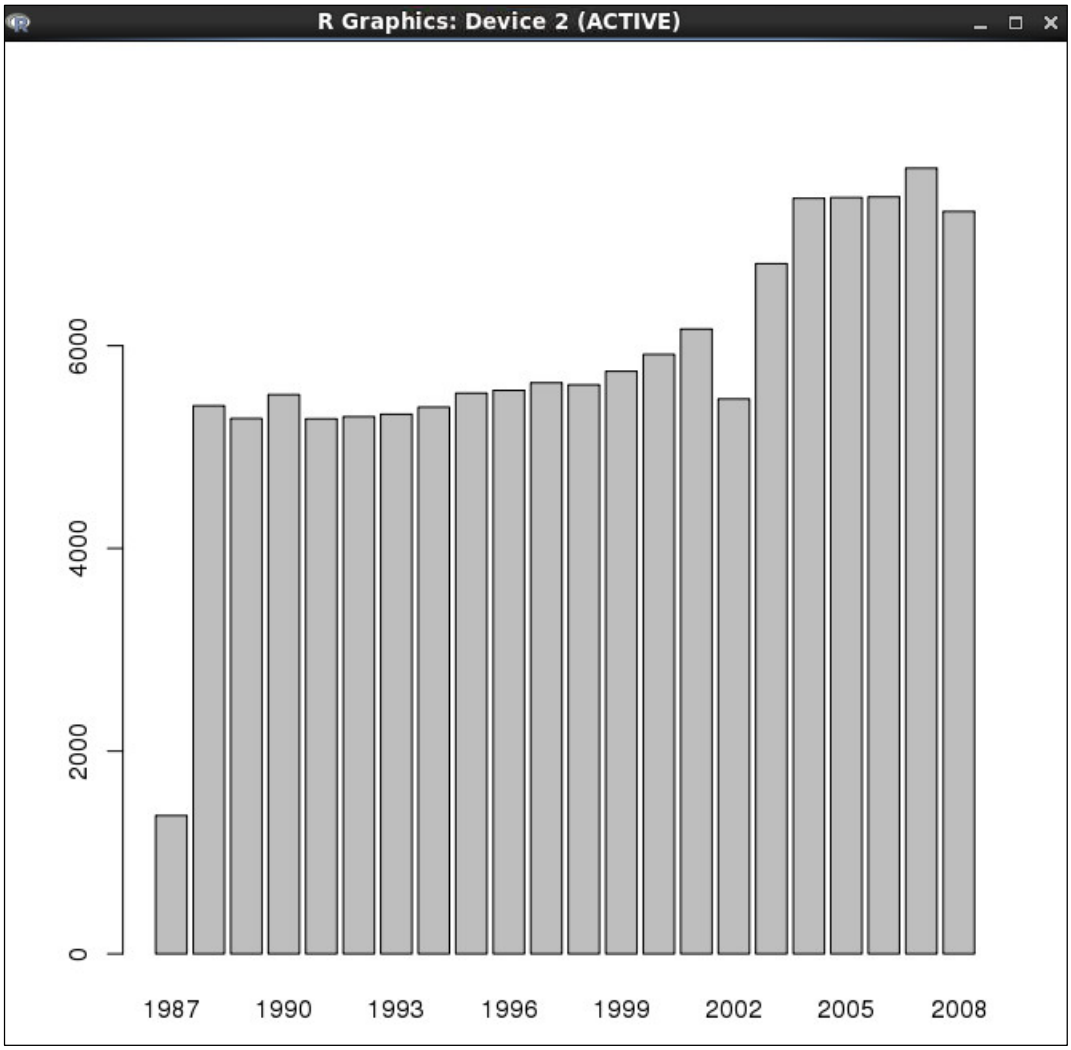
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

21. To visualize the data using some of R's visualization capabilities to see the same data distribution graphically, type **`barplot(summary(air$Year)).`**

    The results appear as follows:



22. Review and then close the graph.

    Similarly, you can also examine the distribution of flights by airline (UniqueCarrier). We have 29 airlines in this dataset, including United (UA), Delta (DL), and many others.

23. Type:

    `summary(air$UniqueCarrier)`

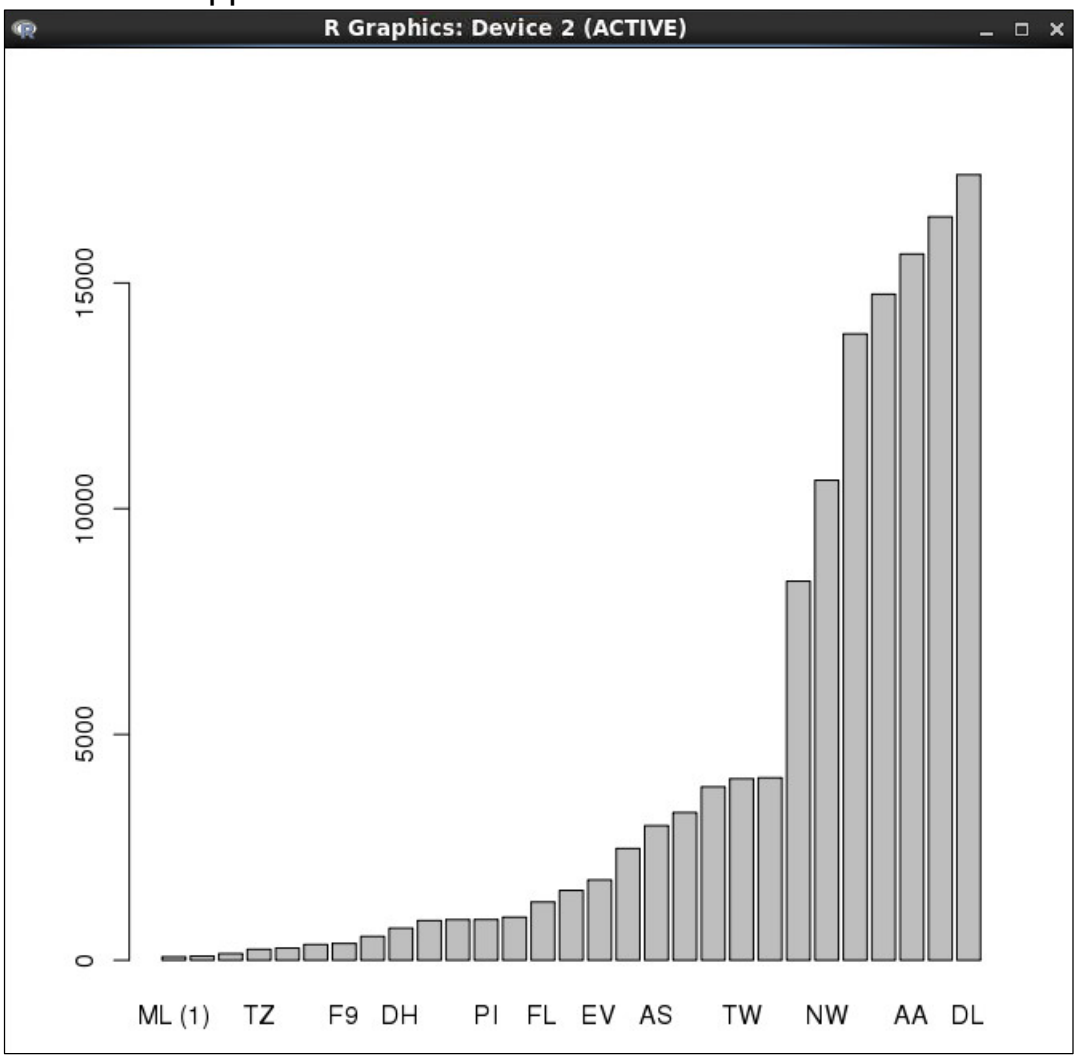
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

24. To visualize the data, type:

```
barplot(sort(summary(air$UniqueCarrier)))
```

The results appear as follows:



25. Close all open windows.

This concludes this exercise. You should now be able to get started with R and Big R exercises using IBM BigInsights. You also had an overview into the Web Tooling module for Text Analytics. These value-adds and more, are available with the IBM BigInsights Data Scientist module.

Results:
You have created a new Text Analytics Web Tooling project and loaded documents to scan for certain keywords. You started up the R console and ran basic commands on it. You also loaded the BigR libraries and ran basic BigR operations.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

IBM BigInsights for Enterprise Management

IBM BigInsights v4.0

© Copyright IBM Corporation 2015
Course materials may not be reproduced in whole or in part without the written permission of IBM.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Unit objectives

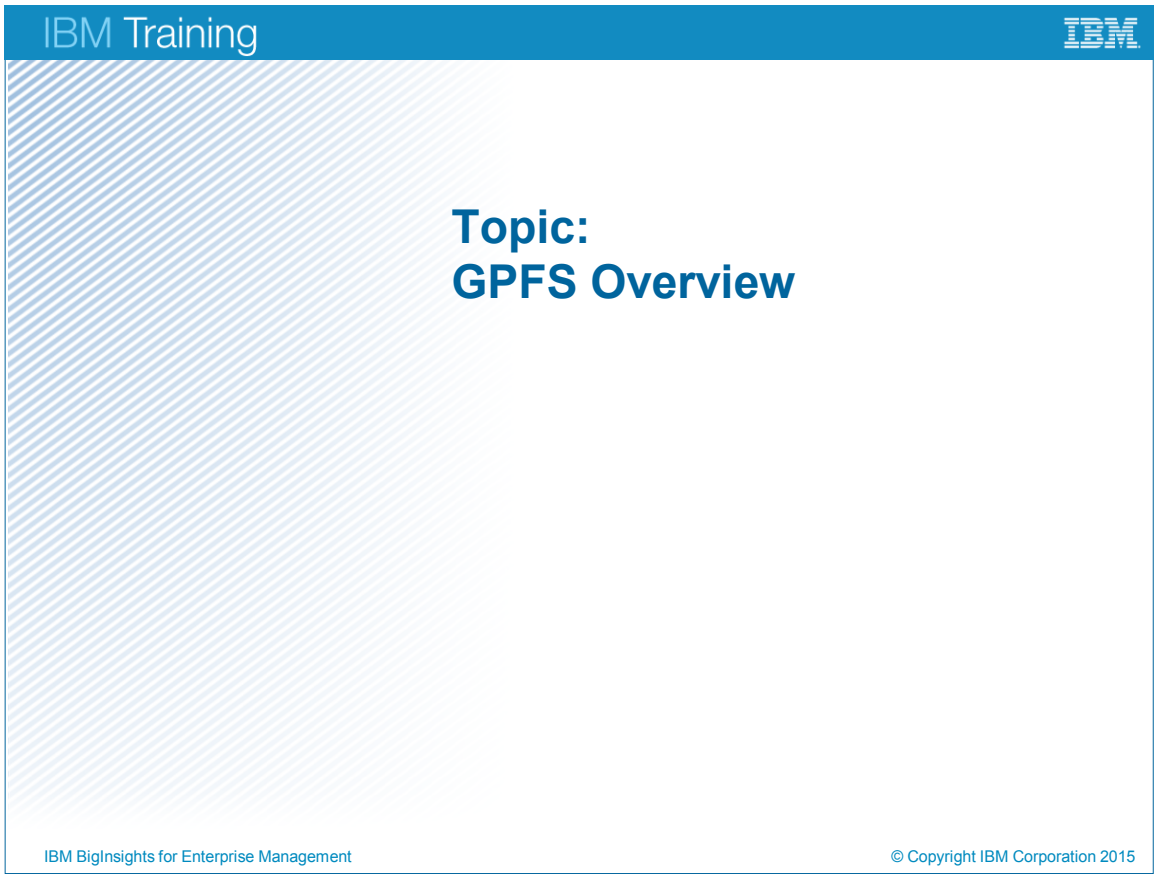
- List the advantages of using GPFS - FPO over HDFS
- Understand the benefits of the POSIX file system
- Describe the YARN architecture
- Understand the role of Platform Symphony

© Copyright IBM Corporation 2015

Unit objectives

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Topic: GPFS Overview

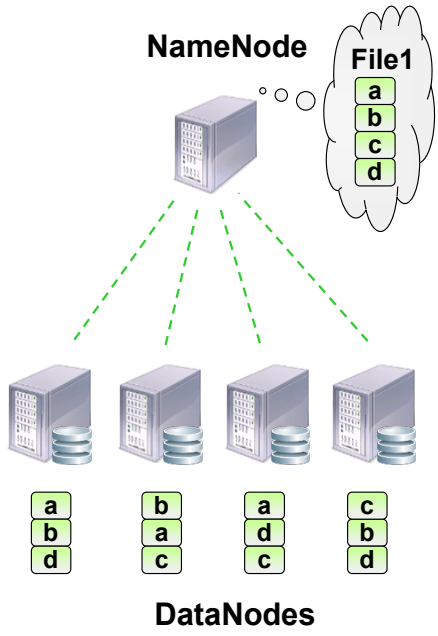
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

HDFS: architecture

- Master / Slave architecture
- NameNode
  - Manages the file system namespace and metadata
    - FsImage
    - EditLog
  - Regulates access by files by clients
- DataNode
  - Many DataNodes per cluster
  - Manages storage attached to the nodes
  - Periodically reports status to NameNode
  - Data is stored across multiple nodes
  - Nodes and components will fail, so for reliability data is replicated across multiple nodes



IBM BigInsights for Enterprise Management

© Copyright IBM Corporation 2015

HDFS: architecture

HDFS has a master/slave architecture. An HDFS cluster consists of a single NameNode, a master server that manages the file system namespace and regulates access to files by clients. In addition, there are a number of DataNodes, usually one per node in the cluster, which manage storage attached to the nodes that they run on. HDFS exposes a file system namespace and allows user data to be stored in files.

The HDFS namespace is stored by the NameNode. The NameNode uses a transaction log called the EditLog to persistently record every change that occurs to file system metadata. For example, creating a new file in HDFS causes the NameNode to insert a record into the EditLog indicating this. Similarly, changing the replication factor of a file causes a new record to be inserted into the EditLog. The NameNode uses a file in its local host OS file system to store the EditLog. The entire file system namespace, including the mapping of blocks to files and file system properties, is stored in a file called the FsImage. The FsImage is stored as a file in the NameNode's local file system too.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Internally, a file is split into one or more blocks and these blocks are stored in a set of DataNodes. In this example, File 1 is a large file that is divided into (a, b, c, d) chunks. Each chunk is replicated (default 3 times) to 3 different nodes so that there is data resiliency. If a node goes down, the blocks on that node are re-replicated to surviving nodes to re-establish the replication factor to 3. Having 3 copies for a block also allow Hadoop to run a calculation on 1 of 3 different servers - whichever is least busy.

The NameNode executes file system namespace operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to DataNodes. The DataNodes are responsible for serving read and write requests from the file system's clients. The DataNodes also perform block creation, deletion, and replication upon instruction from the NameNode.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

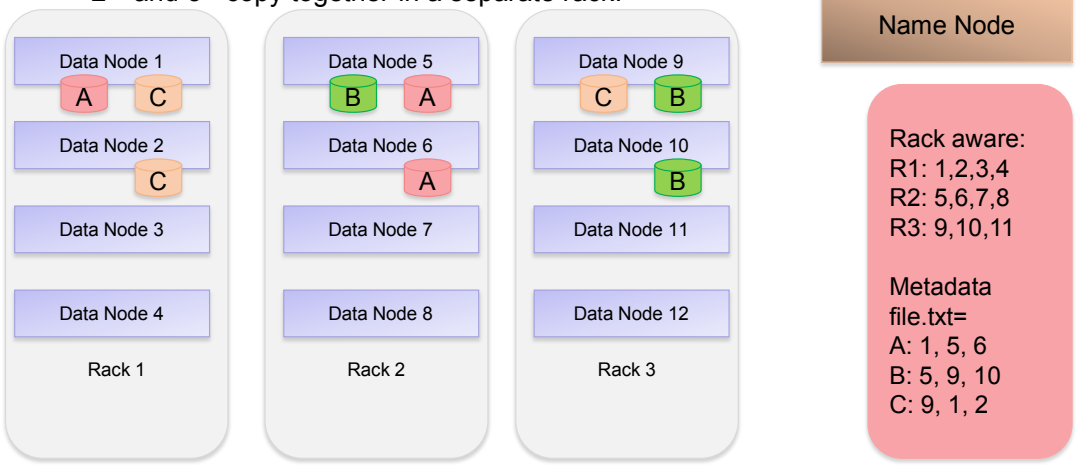
© Copyright IBM Corp. 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Replication of data and rack awareness

- Blocks of data are replicated to multiple nodes
  - Behavior is controlled by replication factor, configurable per file - Default is 3 replicas
- Replication is rack-aware to reduce inter-rack network hops/latency:
  - 1 copy in first rack
  - 2nd and 3rd copy together in a separate rack.



IBM BigInsights for Enterprise Management

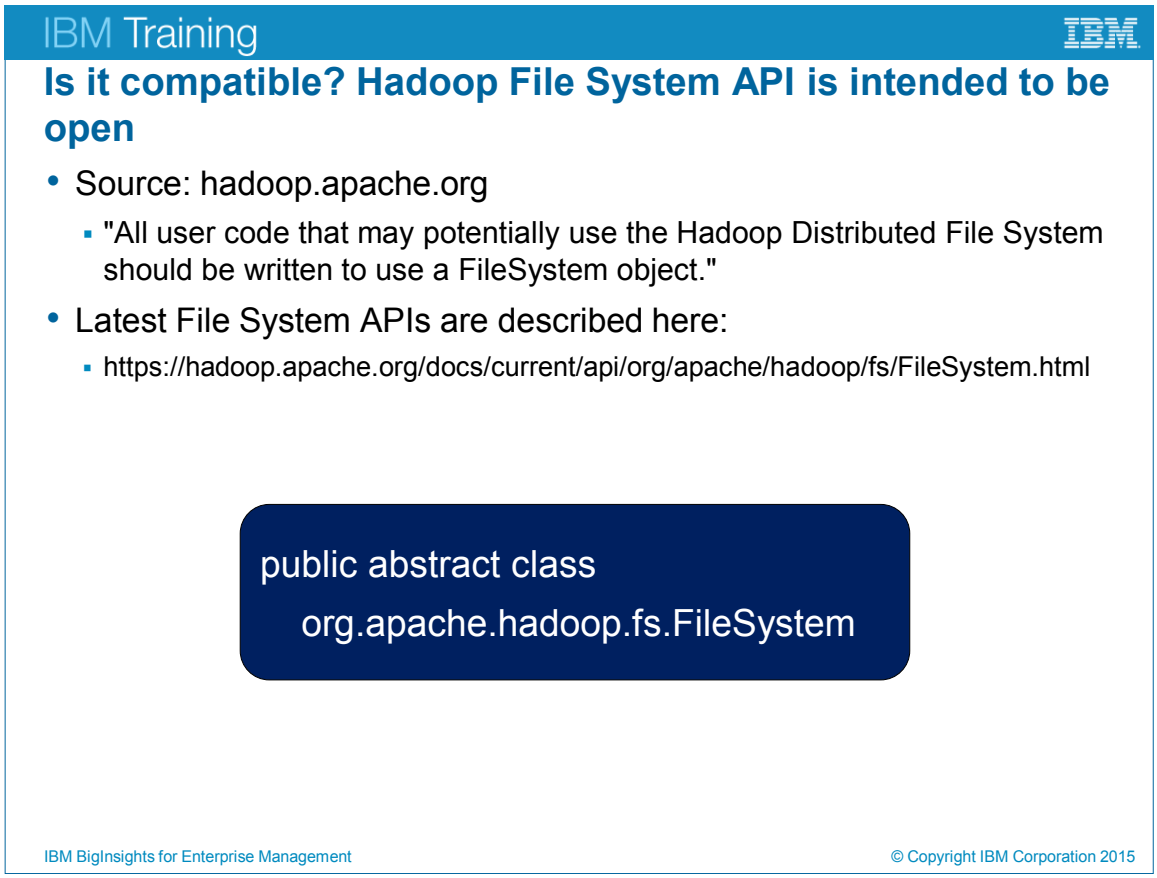
© Copyright IBM Corporation 2015

Replication of data and rack awareness

Rack awareness means that Hadoop namenode is able to track and provide information about where blocks physically reside across the cluster. Applications will be transparently scheduled to preferably run on the data node that actually stores the data to minimize network traffic. Recall there are 3 copies (by default) for any block, so the workload can run equally well on any of 3 locations. Further, if all 3 nodes that have the data are busy, work can be scheduled on a data node which shares the same rack as the data owning node to minimize network traffic.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Is it compatible? Hadoop File System API is intended to be open

While Hadoop distributions typically ship with the Hadoop Distributed FileSystem (HDFS), Hadoop can be configured to use alternate FileSystems by leveraging its pluggable file system architecture.

It is important to note that the ability to support different file systems has always been a design point of Hadoop.

Anyone can build a Hadoop compatible file system by implementing the abstract class org.apache.hadoop.fs.FileSystem. This class specifies, in a file system neutral manner, the operations that a Hadoop application may perform. How a file system actually performs the operation (such as open a file, read a file, write a file, relocate a file, etc.) is up to the file system, whether it is HDFS or something else.

Because all Hadoop applications are written to use FileSystem object, use of HDFS is not strictly required.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

File system for Hadoop designed to be extensible

- Each file system offers unique capabilities / advantages

All based on **org.apache.hadoop.fs.FileSystem API**

	Optimized for
HDFS	General Hadooop
GlusterFS	file-based scale-out NAS
OrangeFS	high end computing (HEC) systems
SwiftFS	write directly to containers in an OpenStack Swift object store
GridGain	In-Memory Data Fabric
Lustre	
MapR FileSystem	
Quantcast File System	
▪etc.	

GPFS is no different

Source: https://wiki.apache.org/hadoop/HCFS

© Copyright IBM Corporation 2015

File system for Hadoop designed to be extensible

Not only was Hadoop designed to allow multiple types of file systems, many HDFS alternatives exist today, including GPFS.

All of these implement org.apache.hadoop.fs.FileSystem interface, to remain compatible with HDFS.

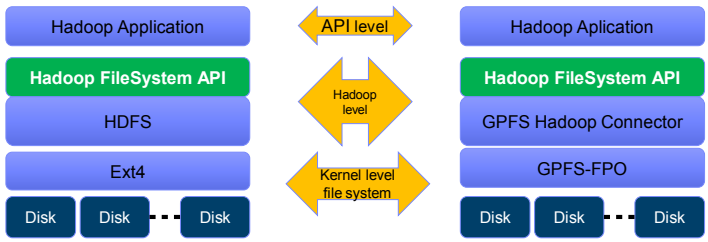
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Spectrum scale: connector architecture

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*

IBM BigInsights for Enterprise Management

© Copyright IBM Corporation 2015

Spectrum scale: connector architecture

This slide shows the general architecture of HDFS vs. GPFS for Hadoop. Applications work with the FileSystem API. Notice that the architecture is very comparable. HDFS implements FileSystem API and sits on Ext4 for its kernel level filesystem.

GPFS implements FileSystem API though GPFS Hadoop Connector, and GPFS-FPO is the kernel level filesystem.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Spectrum scale for Hadoop applications

Today	With GPFS*
• Setup GPFS client on Application Node: ▪ Install GPFS RPMs ▪ Add node to GPFS cluster ▪ Start GPFS connector	• Setup GPFS client on Application Node: ▪ Install GPFS RPMs ▪ Add node to GPFS cluster ▪ Start GPFS connector
• Application specifies gpfs:///hostname/	• Application specifies hdfs:///namenode:9001

*requires Enterprise Manager V4.1, which ships with GPFS V4.1.1 (July 2015)

IBM BigInsights for Enterprise Management © Copyright IBM Corporation 2015

Spectrum scale for Hadoop applications

In July, a significant step forward will be made to make GPFS more transparent. Applications will be able to specify HDFS style URI.

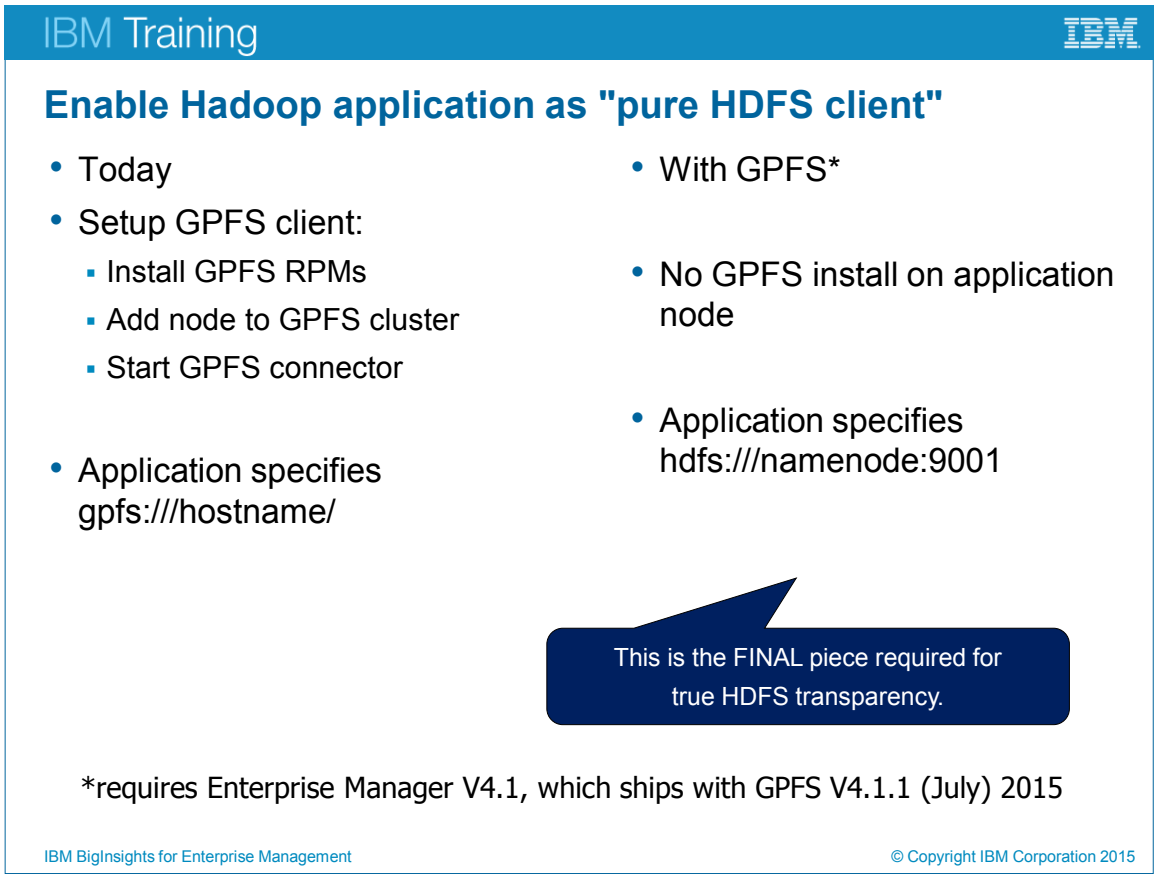
However, GPFS client will still need to be installed and configured because the GPFS connector needs to translate HDFS API calls to GPFS API calls.

The major benefit of this enhancement includes:

1. "comfort" for ISVs that they will be using an HDFS compatible file system
2. Customers CTPs can follow ISV application install/configuration instructions (once client GPFS is installed)

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



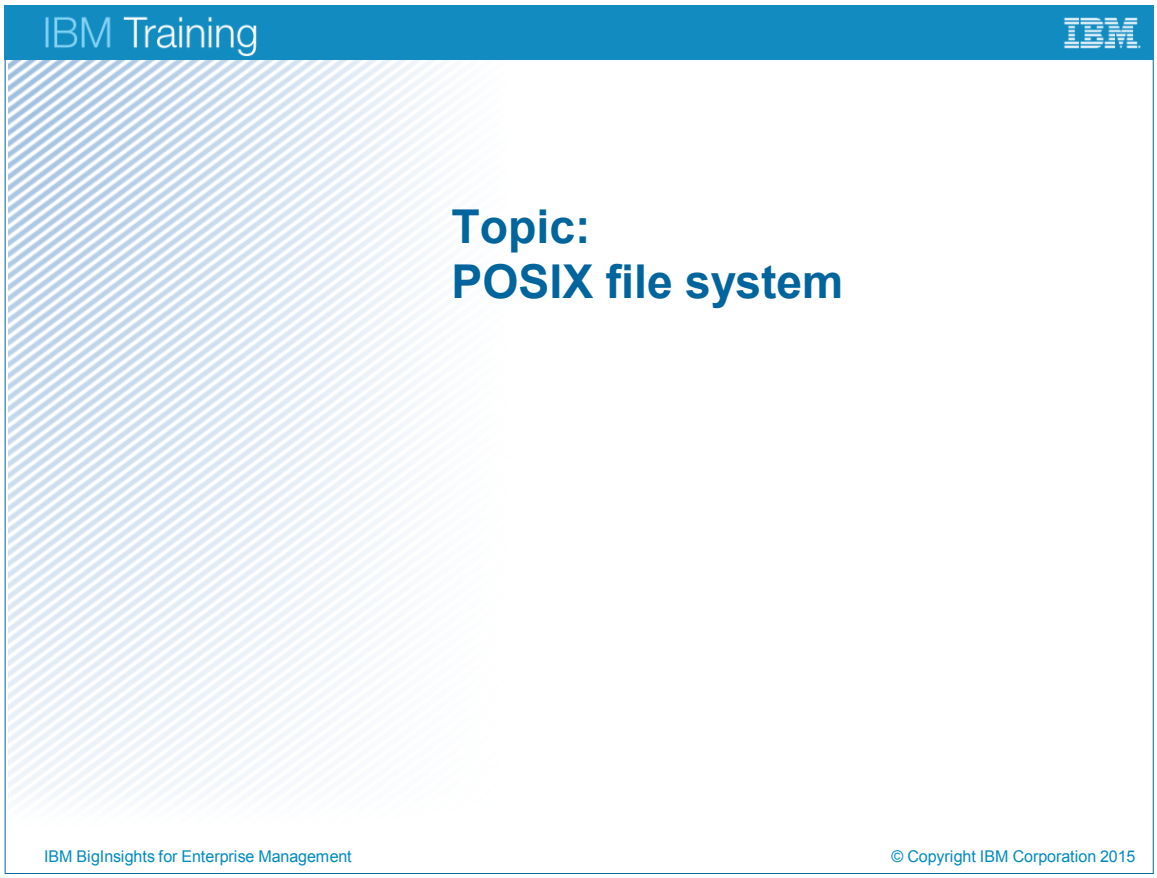
Enable Hadoop application as "pure HDFS client"

The plan to eventually make GPFS completely transparent to Hadoop applications is very exciting.

GPFS Connector will be "daemonized" to communicate like a name node. This will allow a simplified up and running experience.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Topic: POSIX file system

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

POSIX makes it easier! Example: Make a file available to Hadoop

HDFS:

```
hadoop fs -copyFromLocal /local/source/path /hdfs/target/path
```

GPFS/UNIX:

```
cp /source/path /target/path
```

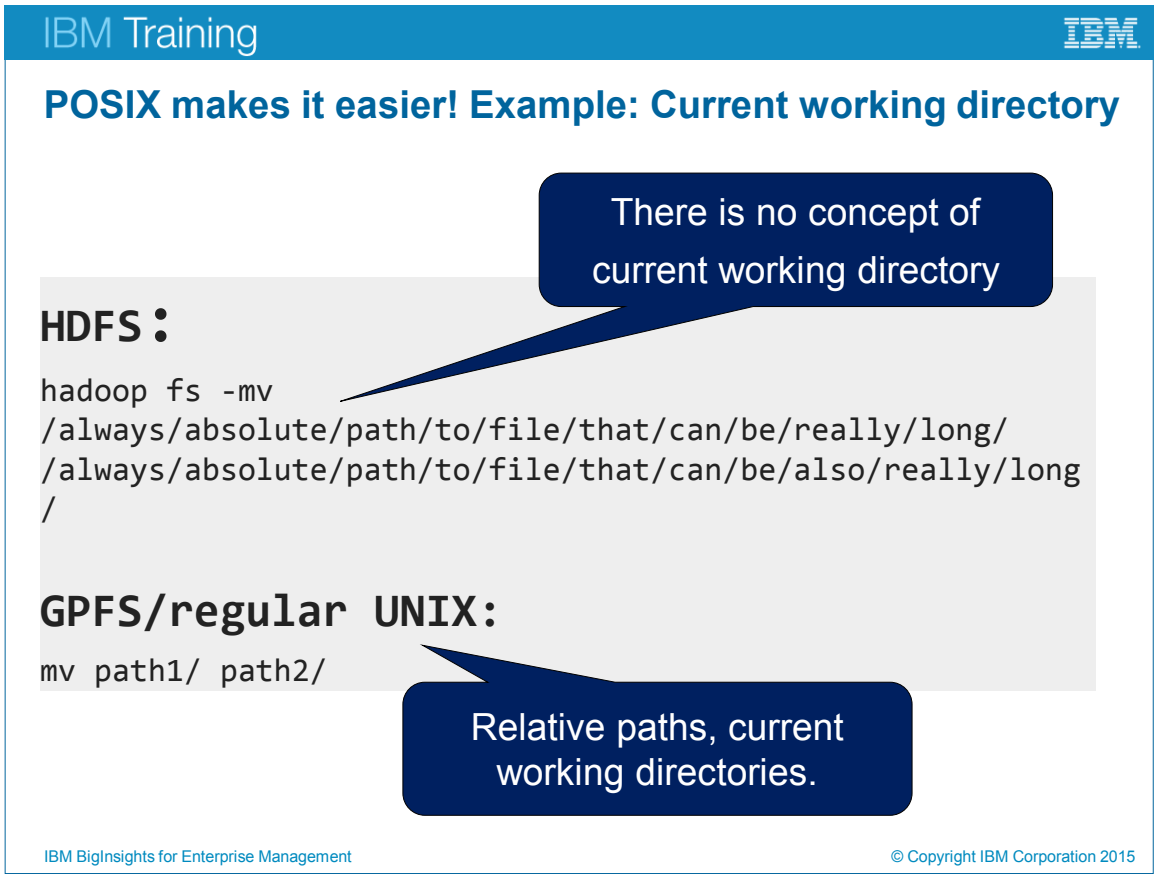
Any system, including non-Hadoop aware systems, can make a file visible to Hadoop

© Copyright IBM Corporation 2015

POSIX makes it easier! Example: Make a file available to Hadoop

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



POSIX makes it easier! Example: Current working directory

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

POSIX makes it easier! Example: Comparing two files

HDFS:

```
diff < (hadoop fs -cat /path/to/file) < (hadoop fs -cat
/path/to/file2)
```

GPFS/regular UNIX:

```
diff path/to/file1 path/to/file2
```

Hopefully these are not big files that are streaming from HDFS to UNIX

© Copyright IBM Corporation 2015

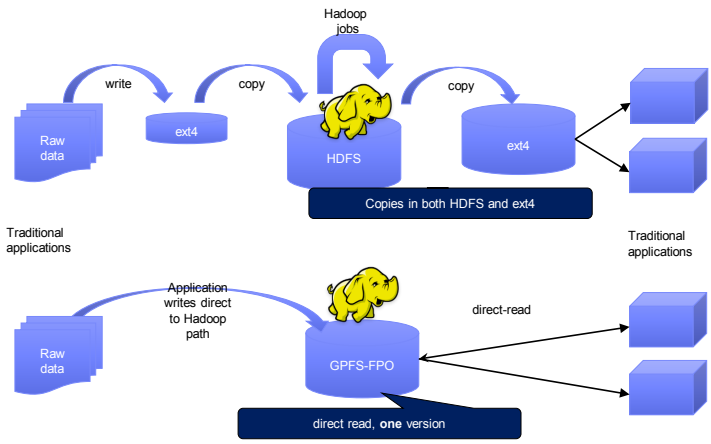
POSIX makes it easier! Example: Comparing two files

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

POSIX makes it easier! Hadoop processed output for other systems



© Copyright IBM Corporation 2015

POSIX makes it easier! Hadoop processed output for other systems

HDFS: 3 copies of the data (Raw in the app, ext4 on the node, HDFS). EXT4 on initial ingest node has to be large enough to hold the transitory data.

One could argue that the ext4 filesystem can be NFS exported, but it does not change the fact that once landed, the file has to be moved on to HDFS as an additional step. Further, the initial ext4 file system needs to be large enough to support the intake of the source file(s).

The same thing applies to ext4 file system on the right in the slide; it needs to be large enough to hold exported data, for a long enough period of time, for downstream applications to consume the data.

With GPFS: Only 1 copy of the data is required. Data producing applications can leverage the abundant disk pool that Hadoop owns and write directly to it. Hadoop applications can consume the data immediately. The outputs, without moving it out of Hadoop, can be consumed directly by traditional (non-Hadoop) applications.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

POSIX makes it easier! Existing operational processes extend naturally to Hadoop

Process	HDFS	POSIX
Backup to Tape	N?	Y
Incremental Backups to Tape*	N?	Y
Retrieval from Tape to Hadoop	N?	Y
File System Vulnerability Scanning**	N?	Y

There may be a some exotic/custom solutions, but will they integrate with your existing enterprise strategy?

*TSM incremental backups is integrated into GPFS snapshot capability

** ASK: has anyone left folders with weak permissions?

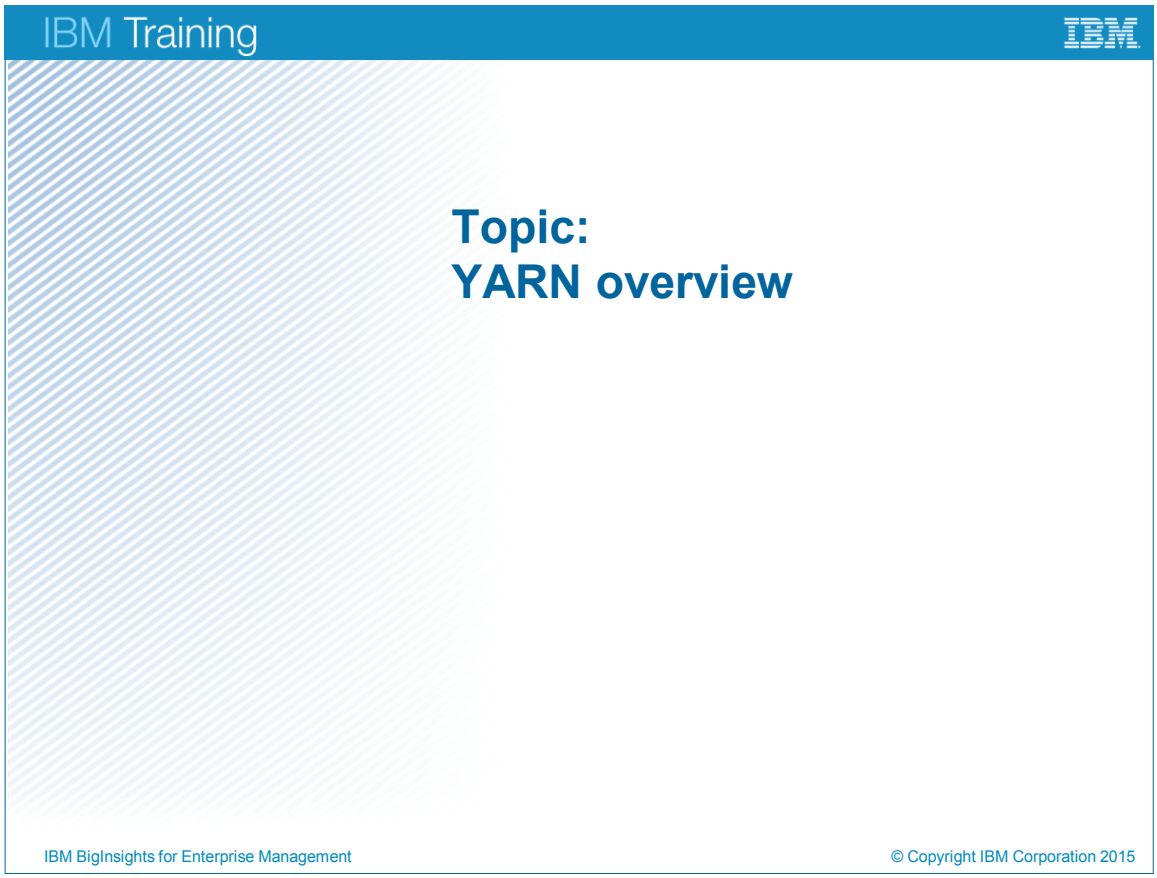
© Copyright IBM Corporation 2015

POSIX makes it easier! Existing operational processes extend naturally to Hadoop

As a POSIX filesystem, GPFS integrates with existing backup systems like Tivoli Storage Manager (TSM). TSM actually has some deep integration with GFPS. Enterprise customers also typically have vulnerability scanning of server file systems to ensure there are no weak permissions set. With HDFS, these vulnerabilities will be hidden from scans if the tool is not HDFS aware (most are not).

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



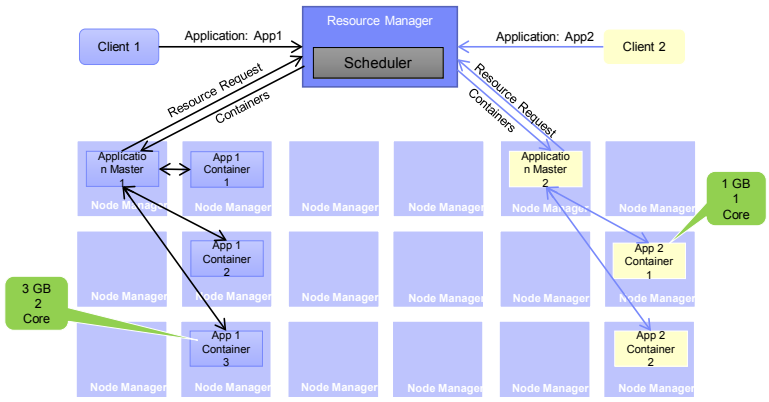
Topic: YARN overview

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

YARN architecture



IBM BigInsights for Enterprise Management © Copyright IBM Corporation 2015

YARN architecture

Yarn has 2 core components in the architecture:

- Resource manager: the master of all cluster resources and a central agent
- Node Manager: The enforcer or per node agent who do this like tasks trackers but more granularly

The resource manager which is essentially the master of all cluster resources and is like a Job tracker and the other one is the node manager is like task tracker.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Details

- Core services for YARN are via long-running daemons:
  - Resource Manager (one per cluster)
  - Node managers (one per node)
  - Timeline server (stores application history)
- Node Managers launch and monitor containers on behalf of Application Master
- A Container executes an application specific process within a constrained set of resources (memory, CPU).
  - IOP calculates container defaults based on cluster resources such as # of nodes, total memory, and number of cores available.
  - Virtual memory is supported within a container. Program permitted to exceed memory limits of container up to a (default) factor of 2.1x (or 210% of real memory in the container). For IOP, the default has been increased to 5x.
- For small jobs, once the Application Master is allocated additional containers may not be required to avoid unnecessary overheads.

© Copyright IBM Corporation 2015

Details

cgroups (abbreviated from control groups) is a Linux kernel feature that limits, accounts for and isolates the resource usage (CPU, memory, disk I/O, network, etc.) of a collection of processes.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

YARN application capabilities

- A YARN Application can
  - Request locality constraints. Request containers on a specific node or rack, or on a specific rack, or anywhere on the cluster (off rack).
  - Make additional resource requests at any time while its running.
    - Request all requests up front (for example, Spark currently uses this approach)
    - Request more resources dynamically as needed. (for example, Map Reduce requests map resources up front and the reduce tasks are not started until later)
- Application Life Span
  - One application per user job. Simplest form of Map Reduce.
  - One application per workflow/user session of (possibly unrelated) jobs.
    - More efficient since containers can be reused between jobs and there is potential to cache intermediate data between jobs. Spark uses this model.
  - Long-running application that is shared by different users.
    - Apache Slider uses this model (always ON). is a long-running application master for launching other applications on the cluster.
    - Llama (Low Latency Application Master), used by Impala, is another example.

© Copyright IBM Corporation 2015

YARN application capabilities

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

High availability

- Making Hadoop highly available has become a divide-and-conquer problem
  - Provide HA for the resource manager
  - Provide HA for each application (on a per-application basis)
- Hadoop 2 supports HA for both resource manager and AM for map reduce jobs.
- ResourceManager HA is similar to the NameNode QJM HA.
  - There is an active RM and a standby one
  - A group of Zookeeper nodes determine which is the active RM at any point in time
  - Since Hadoop 2.6.0, RM restart is work-preserving. Applications running in the cluster can keep running when the RM fails and a new instance takes the active role

© Copyright IBM Corporation 2015

High availability

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Topic: Platform Symphony YARN-Plugin

© Copyright IBM Corporation 2015

Topic: Platform Symphony YARN-Plugin

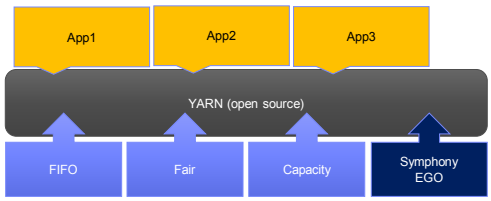
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Platform symphony integrates with open source Hadoop

- YARN uses a pluggable architecture for schedulers.
  - FIFO, Fair, and Capacity Schedulers implemented this way
  - Symphony EGO is also implemented this way.
- Therefore, scheduler is completely transparent to YARN applications.
- ISV Certification for Platform Symphony is not required.



Like other schedulers, queues and policies are defined in Platform Symphony EGO.

© Copyright IBM Corporation 2015

Platform symphony integrates with open source Hadoop

EGO: Enterprise Grid Orchestration (Platform Computing Technology).

Only one scheduler can be in force at any time, but the objective of this slide is to show that Symphony EGO scheduler is just leveraging the pluggable architecture of YARN.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

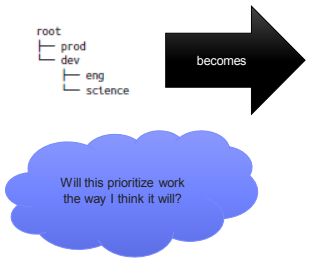
© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Capacity & fair scheduler policies are defined with XML

With XML approach, even simple hierarchies can be very **complex**.

"60/40 capacity scheduler with dev queue elasticity +15%"




```
<?xml version="1.0"?>
<configuration>
  <property>
    <name>yarn.scheduler.capacity.root.queues</name>
    <value>prod,dev</value>
  </property>
  <property>
    <name>yarn.scheduler.capacity.root.dev.queues</name>
    <value>eng,science</value>
  </property>
  <property>
    <name>yarn.scheduler.capacity.root.prod.capacity</name>
    <value>40</value>
  </property>
  <property>
    <name>yarn.scheduler.capacity.root.dev.capacity</name>
    <value>60</value>
  </property>
  <property>
    <name>yarn.scheduler.capacity.root.dev.maximum-capacity</name>
    <value>75</value>
  </property>
  <property>
    <name>yarn.scheduler.capacity.root.dev.eng.capacity</name>
    <value>50</value>
  </property>
  <property>
    <name>yarn.scheduler.capacity.root.dev.science.capacity</name>
    <value>50</value>
  </property>
</configuration>
```

IBM BigInsights for Enterprise Management

© Copyright IBM Corporation 2015

Capacity & fair scheduler policies are defined with XML

Configuring the resource plan for Capacity and Fair schedulers is done via scheduler specific parameters. Yarn.scheduler.capacity.* variable, for example, are specific to capacity scheduler only. Further the hierarchies and resource borrowing behaviors need to be expressed in XML.

In this example, notice that even a simple hierarchy requires a large number of parameters coded by XML which is complex to manage and understand.

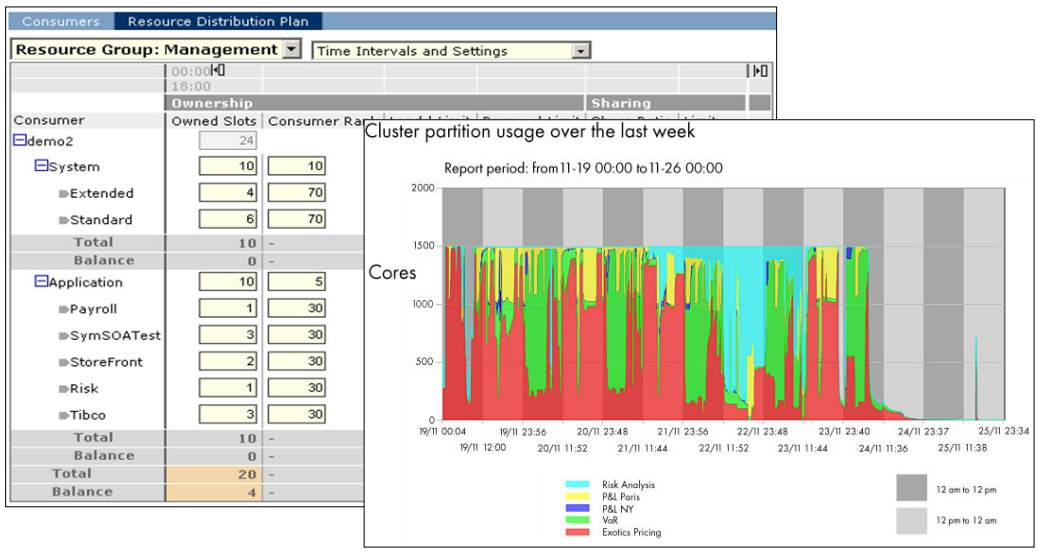
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Platform symphony makes life easier, simplifies queue management and configuration

- Simple interface to configure sharing policies
- Visualizations to validate queue/resource policies



IBM BigInsights for Enterprise Management © Copyright IBM Corporation 2015

Platform symphony makes life easier, simplifies queue management and configuration

With Platform Symphony, configuring queues and queue hierarchies is easier to define and manage with a web based GUI interface. Resource lending policies are also defined here.

Further, the clusters behavior can be visualized to confirm that the resource sharing policies are working as desired.

You can see, for example, that workload RED is allowed to use the whole cluster when nothing else is running. But, as soon as workload Green starts, RED should give up its resources.

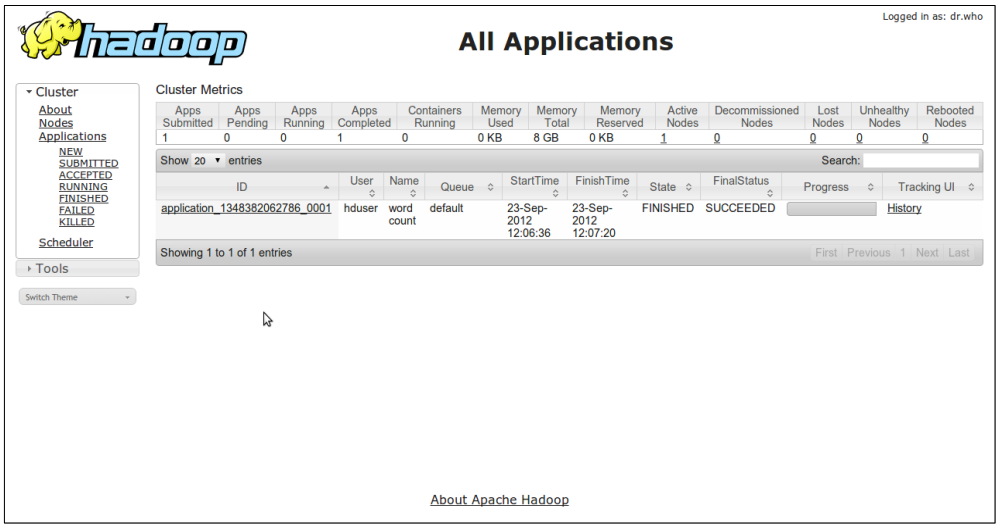
Further, when workload CYAN (blue) starts, both RED and GREEN need to give up their resources. As soon as CYAN is complete and nothing else is running, workload RED ramps up again. The result is very high cluster utilization.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

YARN Web Console: Basic view into containers and memory used



© Copyright IBM Corporation 2015

YARN Web Console: Basic view into containers and memory used

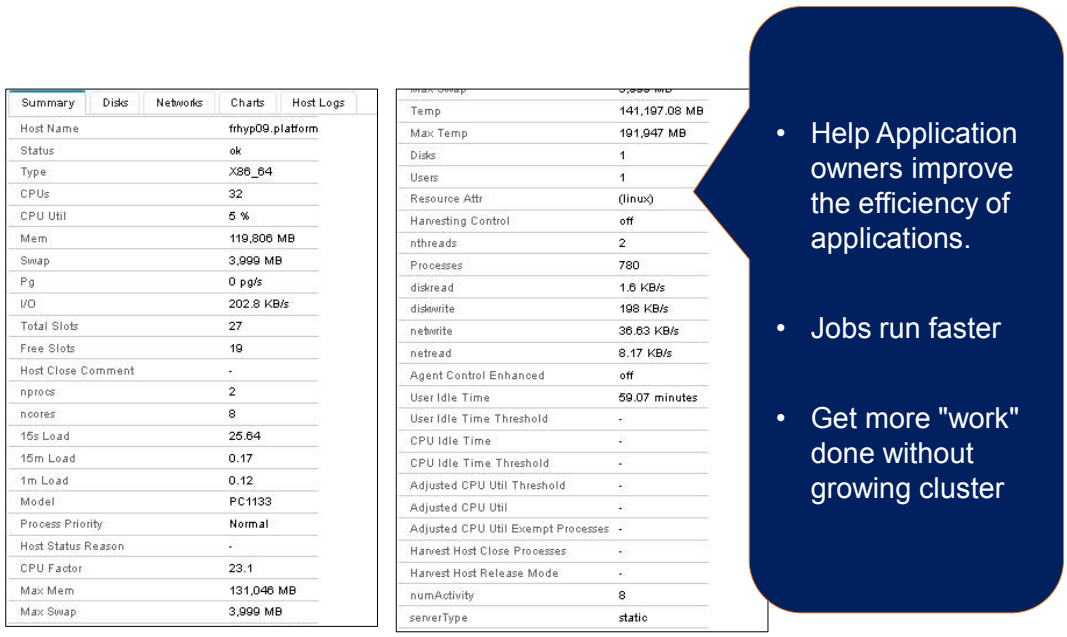
This is a screen capture of the YARN web interface. It is pretty basic, but important to show to help you understand what you get out of-the-box. Performance information largely relates to job status, memory usage, container usage, etc.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Platform Symphony: Deep Insights into Workloads (150 metrics)



IBM BigInsights for Enterprise Management

© Copyright IBM Corporation 2015

Platform Symphony - Deep Insights into Workloads (150 metrics)

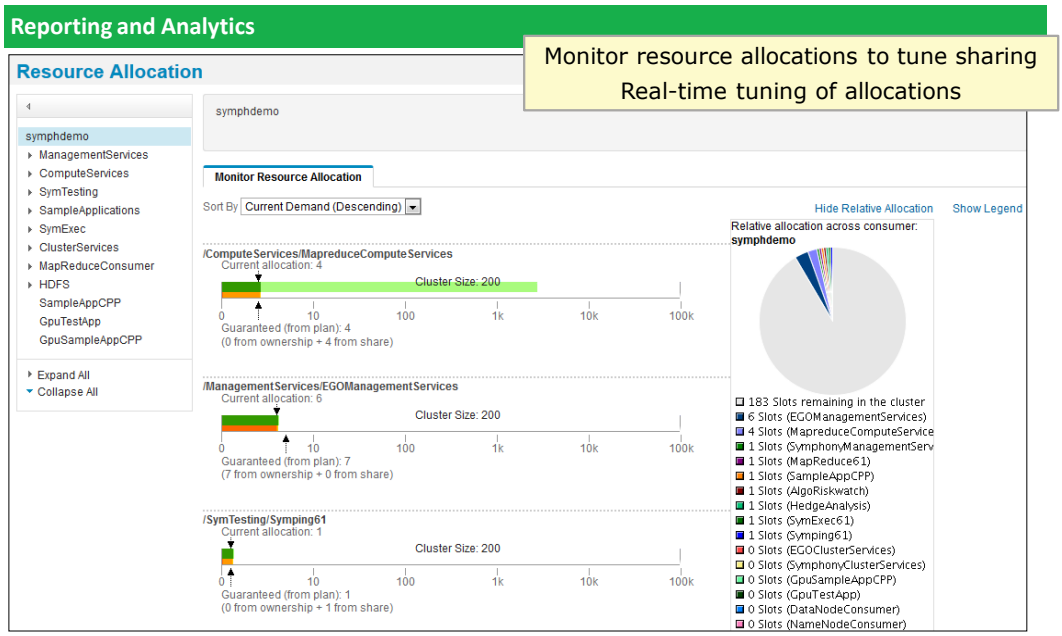
Platform Symphony provides 150 workload metrics that provide deep insight into what is happening. Remember, Platform Symphony has not displaced YARN. YARN is still the interface to applications. This information is available only because Symphony EGO is plugged into YARN.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Platform Symphony: Real time view of resource allocation

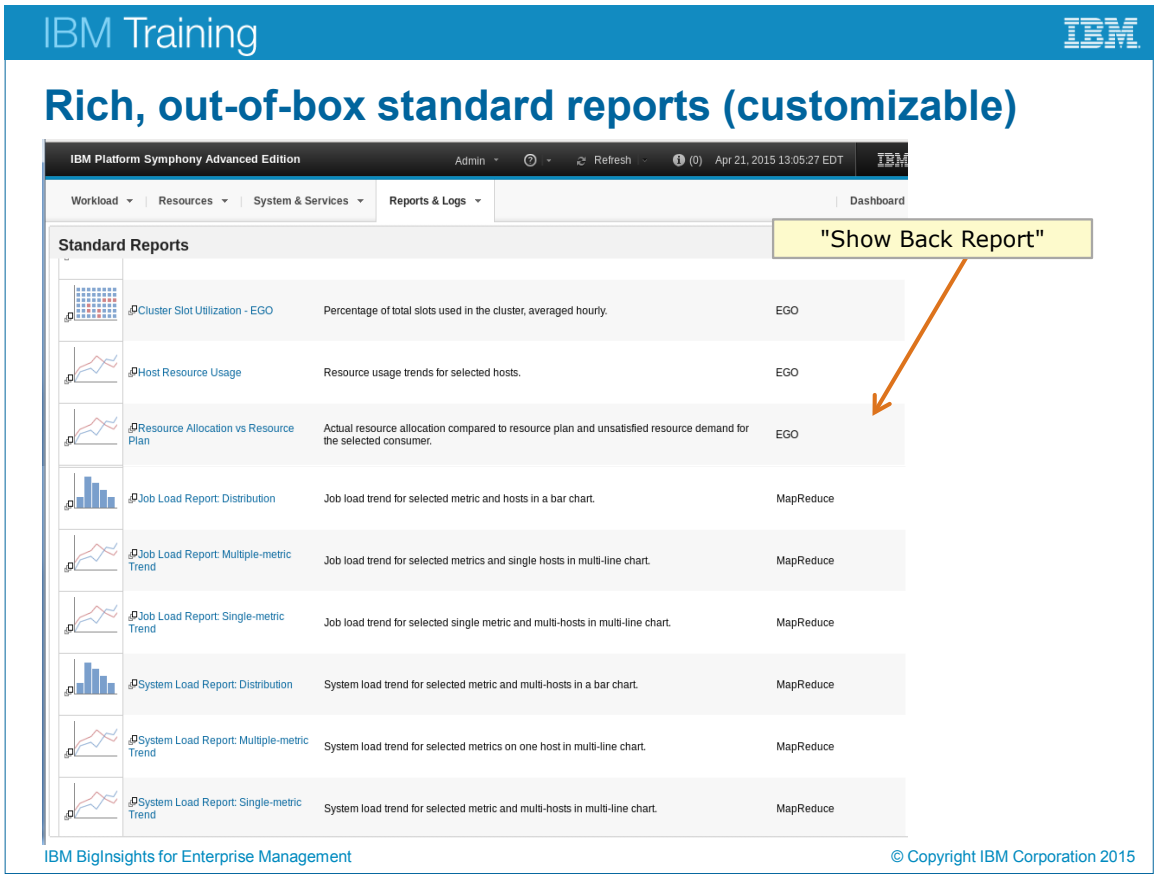


© Copyright IBM Corporation 2015

Platform Symphony: Real time view of resource allocation

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Rich, out-of-box standard reports (customizable)

Platform provides many standard reports out-of-the-box. Customers can start with these reports and further customize them. Of particular interest is the Show Back report.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Unit summary

- List the advantages of using GPFS - FPO over HDFS
- Understand the benefits of the POSIX file system
- Describe the YARN architecture
- Understand the role of Platform Symphony

© Copyright IBM Corporation 2015

Unit summary

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

IBM Training

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corporation 2015. All Rights Reserved.