Information Management

IBM

Session 11929

Big Data and Data Quality - Mutually Exclusive?

Tom Deutsch
tdeutsch@us.ibm.com
Program Director, Big Data

August 9, 2012

Abstract

- It is popular to think that Big Data technologies are so different as to change the need for information quality considerations. That is simply incorrect. This session will cover examples of how organizations got it wrong by not paying attention to well established best practices with an eye on making sure you do not.

Tom Deutsch
tdeutsch@us.ibm.com
Program Director, Big Data

Agenda

1. Why Big Data Is Both the Same and Different
2. The Role Of Analytics in Big Data Quality
3. Why Big Data Should Mean Better Quality
4. Some Closing Thoughts

Gartner - By 2014, 85% of currently deployed data warehouses will not scale appropriately to meet new information volume and complexity demands.

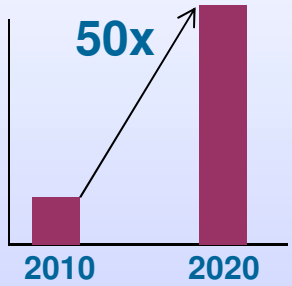
- Complex data types are already receiving more than experimental demand**. An additional complexity issue involves use-case pressure. End-user organizations now indicate that business intelligence already uses many sources other than the classic data warehouse (DW),** using enterprise content management repositories, XML repositories, streaming data, blogs and more**. This scope of end-user demand is further compounded by the increasing number of end-user making demands.**

Gartner Research Publication Date: 1 December 2010 ID Number: G00208101

Predicts 2011: Data Management Disciplines Elevate Business Criticality

The Characteristics Of Big Data

Cost efficiently processing the growing Volume



35 ZB

Responding to the increasing Velocity

30 Billion RFID sensors and counting

Collectively Analyzing the broadening Variety

80% of the worlds data is unstructured

Establishing the Veracity of big data sources

1 in 3 business leaders don't trust the information they use to make decisions

Hadoop Background

- Apache Hadoop **is a** software framework **that supports data-intensive applications under a free license. It enables applications to work with thousands of nodes and petabytes of data.** Hadoop was inspired by Google Map/Reduce **and** Google File System **papers.**
- **Hadoop is a top-level Apache project being built and used by a global community of contributors,** using the Java programming language**. Yahoo has been the largest contributor to the project, and uses Hadoop extensively across its businesses.**

Break It Down For Me Here…

- **Hadoop is a** platform and framework**, not a database**
  - It uses both the CPU and disc of single commodity boxes, or **node**
  - Boxes can be **combined into clusters**
  - New nodes can be **added as needed,** and added **without needing to change** the;
    - Data formats
    - How data is loaded
    - How jobs are written
    - The applications on top

So How Does It Do That?

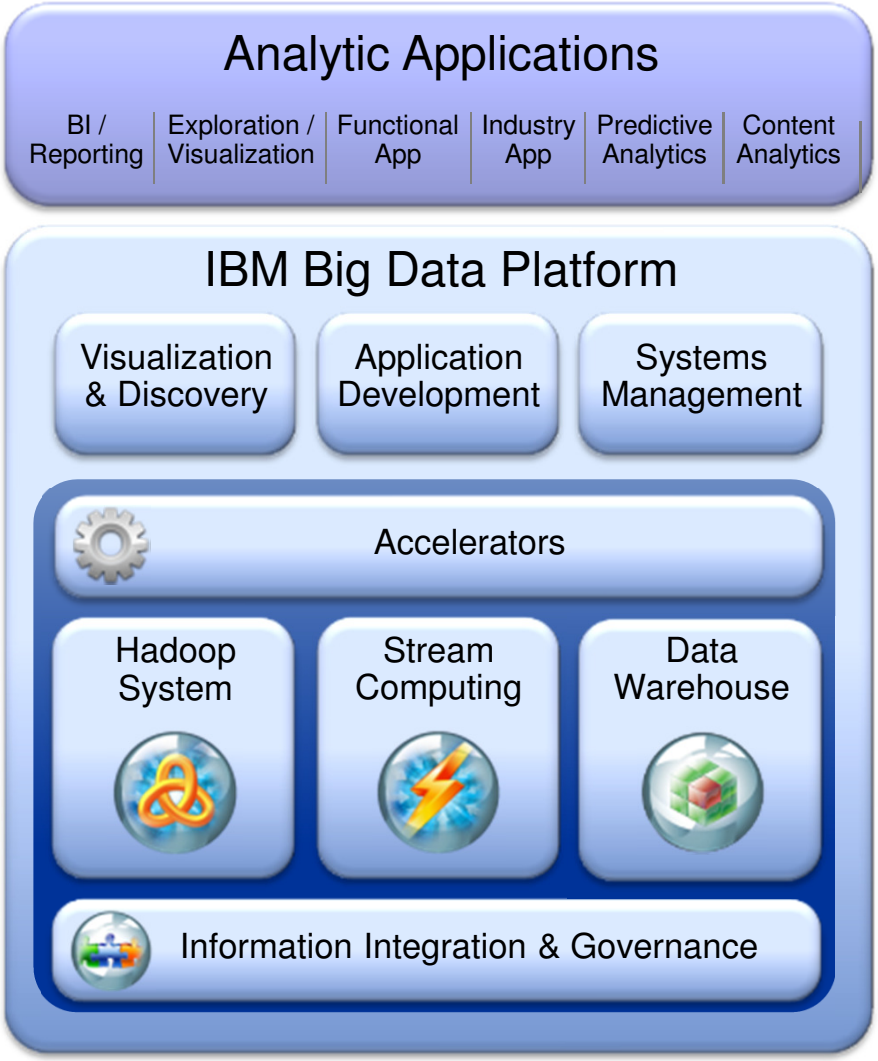
At its core, hadoop is made up of;

- **Map/Reduce**
  - How hadoop understands and assigns work to the nodes (machines)
- **Hadoop Distributed File System = HDFS**
  - Where hadoop stores data
  - A file system that's runs across the nodes in a hadoop cluster
  - It links together the file systems on many local nodes to make them into one big file system

IBM Big Data Strategy: Move the Analytics Closer to the Data

New analytic applications drive the requirements for a big data platform

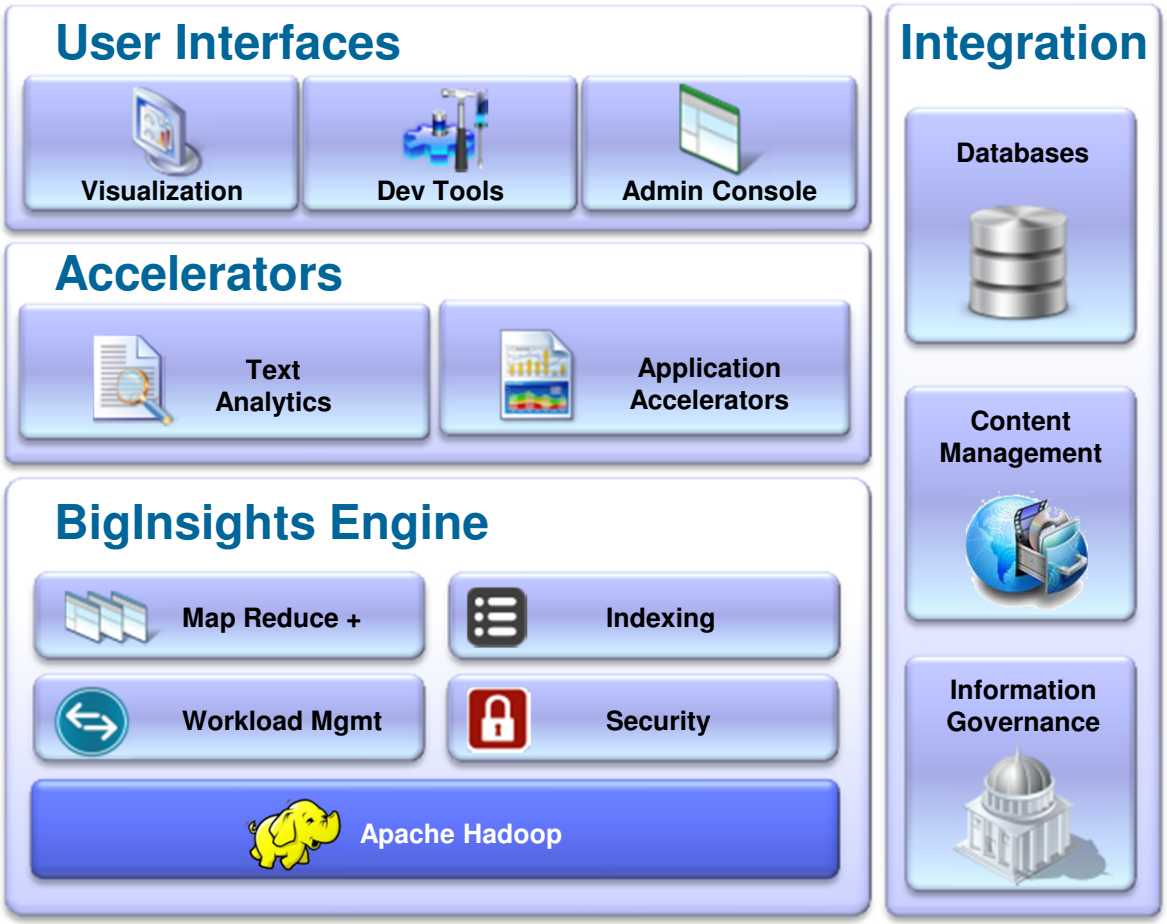
- **Integrate and manage the full variety, velocity and volume of data**
- **Apply advanced analytics to information in its native form**
- **Visualize all available data for ad-hoc analysis**
- **Development environment for building new analytic applications**
- **Workload optimization and scheduling**
- **Security and Governance**



InfoSphere BigInsights Brings Hadoop to the Enterprise

- **Manages a wide variety and huge volume of data**
- **Augments open source Hadoop with enterprise capabilities**
  - Performance Optimization
  - Development tooling
  - Enterprise integration
  - Analytic Accelerators
  - Application and industry accelerators
  - Visualization
  - Security
- **Provides Enterprise Grade Hadoop analytics**

InfoSphere BigInsights – A Closer Look



More Than Hadoop

- Performance & workload optimizations
- Unique text analytic engines
- Spreadsheet-style visualization for data discovery & exploration
- Built-in IDE & admin consoles
- Enterprise-class security
- High-speed connectors to integration with other systems
- Analytical accelerators

OK – Now Back To Data Quality

As my IBM Research colleague John McPherson stated

- "Keep in mind that often when we're talking about Big Data we are talking about using data that we haven't been able to exploit well in the past so we're typically trying to solve different problems."
- "We're not trying to figure out the profitability of each of our stores - that we should already be doing using high quality data from systems of record and doing the things we do to standardize and reshape as we put it into a data warehouse."

So What Is The Big Data | Data Quality Problem

- Many people are under the mistaken impression that there's an inherent trade-off between the volume of the data set and the quality of the data maintained within
- There are a lot of “we’ll fix it later” approaches being taken

A Quick Real Life Story

I was working with a large banking client and…

What Is The Same?

- The source of data quality problems in most organizations is usually at the source transactional systems--be they your customer relationship management (CRM) system, general ledger application, etc--which are usually in the terabytes range.

- When possible address data quality before landing the data

What Is Different

- A lot of Big Data initiatives are for deep analysis of aggregated data sources--such as social marketing intelligence, real-time sensor data feeds, data crawled from external resources, browser clickstream sessions, IT system logs, and the like--that you don't link to official reference data.
- You don't necessarily have to cleanse those because they don't feed into an official system of record

- Some of these – sensor, geolocation and machine generated data are typically *presumed* to be accurate

Agenda

1. Why Big Data Is Both the Same and Different
2. The Role Of Analytics in Big Data Quality
3. Why Big Data Should Mean Better Quality
4. Some Closing Thoughts

But As Boundaries Are Pushed On Engagement

Sources such as are becoming sources of insight;

- **social marketing intelligence**
- **real-time sensor data feeds**
- **data crawled from external resources**
- **Email / text**

And that means quality and accuracy matters.

So What Is Different?

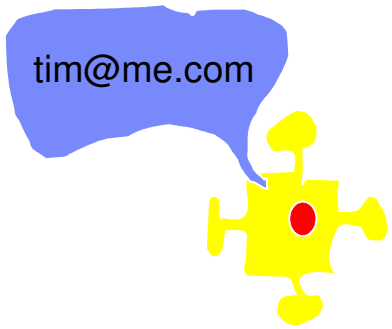
- Both the data types and methods will likely be different
- In the SQL world two different DB, same query, same results
- Not so in Big Data land…

This Movie Is A Bomb

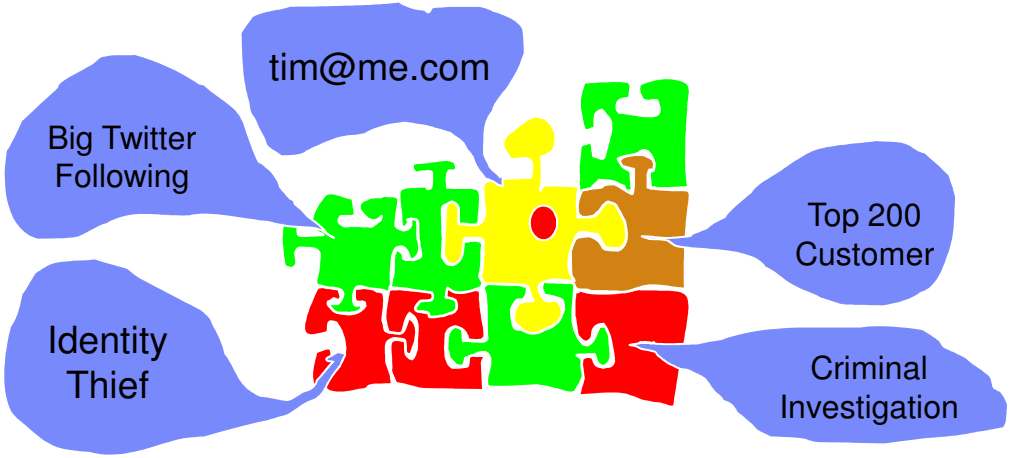
This Movie Is The Bomb

Context Matters

Context: Better understanding something by taking into account the things around it.



[Hardly actionable]

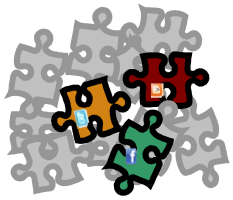


[Substantially more actionable]

Context Accumulation: The incremental process of integrating new observations with previous observations

Entity Integration Mechanics

All names and messages are fictitious



More than 50% of the profile attributes are populated from text.

Jane Doe's Twitter Profile:

Name: Jane Doe
Location: **Home of the Buccaneers**

@tonyc Sore after my morning yoga class !

Name: Jane Doe
Id: @jaydee
Address: Tampa, FL
Interests: running, yoga, football…
Network: Tony C.

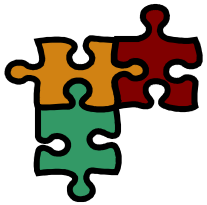
Jane's blog:

*Pinkberry fans, be sure to follow us also on twitter at **@jaydee** and be part of the Pinkberry fun!*

Name: Jane
Twitter ID:@jaydee
Blog Topic: food
Location: Tampa, FL

1. Social Media handles
2. Partial Name similarity
3. Partial Address

More than 100 rules for entity integration.



Name/Nick: Jane Doe
Address: Tampa, FL
Twitter: @jaydee
Blog Topic: food
Hobbies: running, yoga, …

Advanced Text Analytics – part of InfoSphere Streams and BigInsights

How it works

- **Parses unstructured text and detects meaning with annotators**
- **Understands the context in which the text is analyzed**
- **Hundreds of pre-built annotators for names, addresses, phone numbers, and others**
  - Parts of speech support for English, Spanish, French, German, Portuguese, Dutch, Japanese, Chinese
- **Distills structured info from unstructured text**
  - Sentiment analysis
  - Consumer behavior
  - Illegal or suspicious activities

…

Benefits

- **More precise and correct answers**
  - 2x vs. marketplace alternatives
- **50% faster than manual method**
  - Used to build world-class text analysis applications
- **Run faster text analysis**
  - 10x or more vs. marketplace alternatives

Unstructured text (document, email, etc)

Football World Cup 2010, one team distinguished themselves well, losing to the eventual champions 1-0 in the Final. Early in the second half, Netherlands' striker, Arjen Robben, had a breakaway, but the keeper for Spain, Iker Casillas made the save. Winger Andres Iniesta scored for Spain for the win.

Classification and Insight

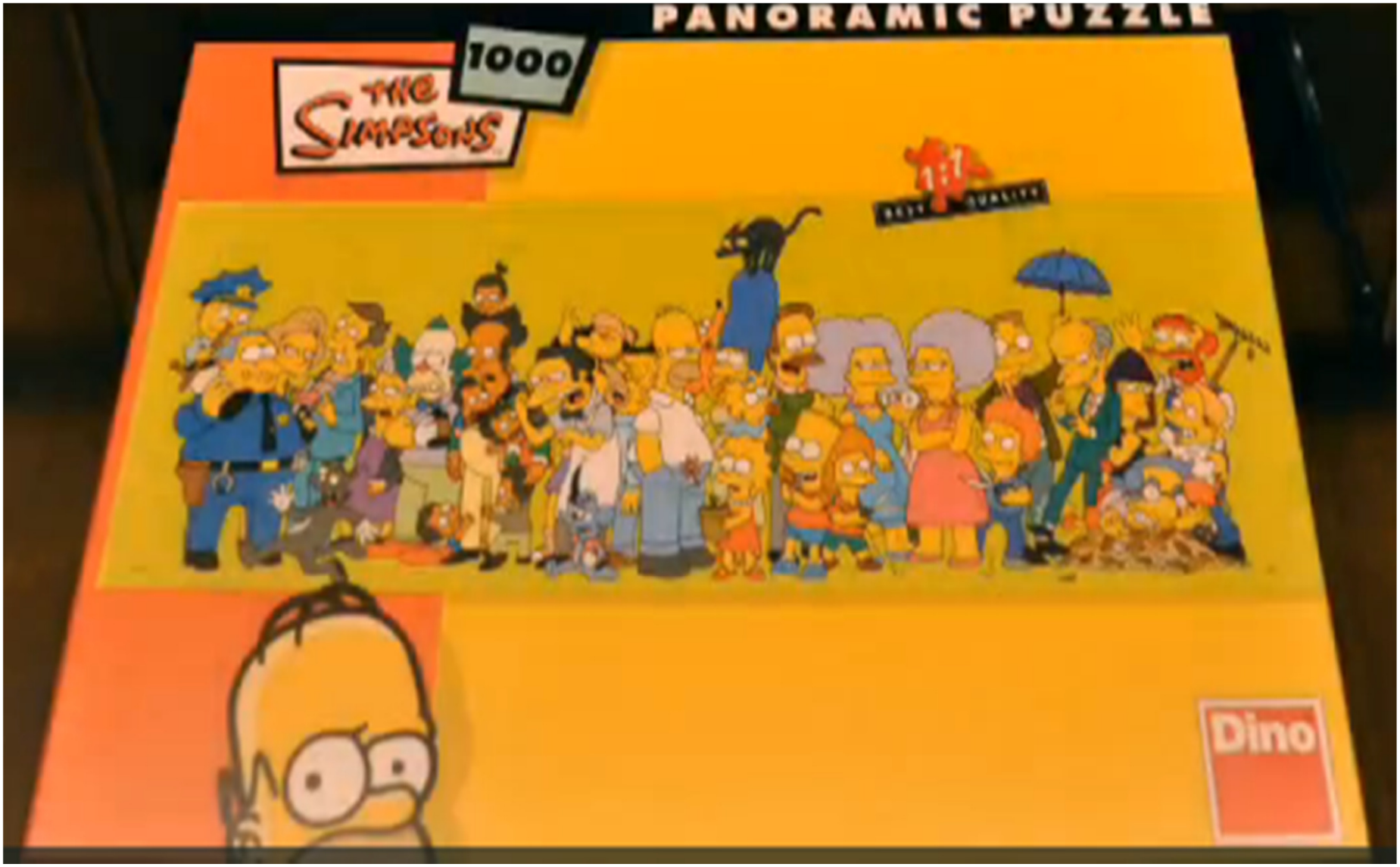
World Cup 2010 Highlights

Name	Position	Country
Arjen Robben	Striker	Netherlands
Iker Casillas	Keeper	Spain
Andres Iniesta	Winger	Spain

Agenda

1. Why Big Data Is Both the Same and Different
2. The Role Of Analytics in Big Data Quality
3. Why Big Data Should Mean Better Quality
4. Some Closing Thoughts

Well Let Me Explain With Some Help

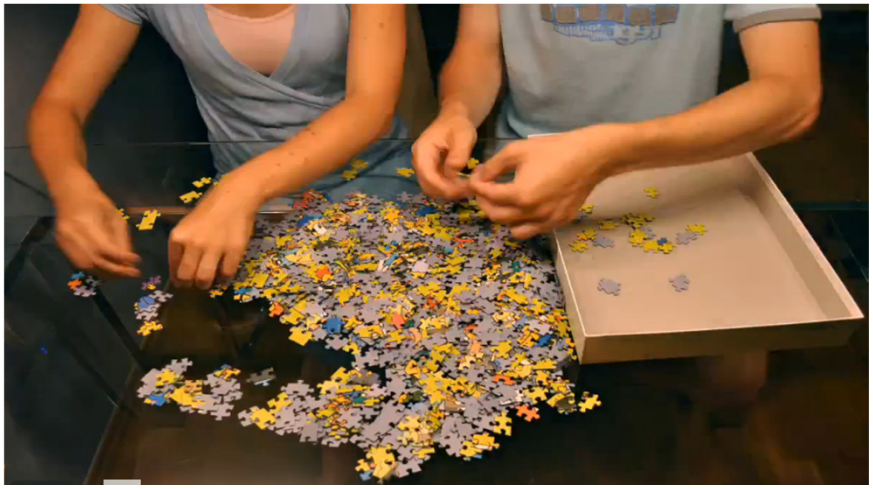


Imagine This Is Your Data

Reporting Handles The Edges, But What About The Middle?

Quick – Which Actually Contains More Data?

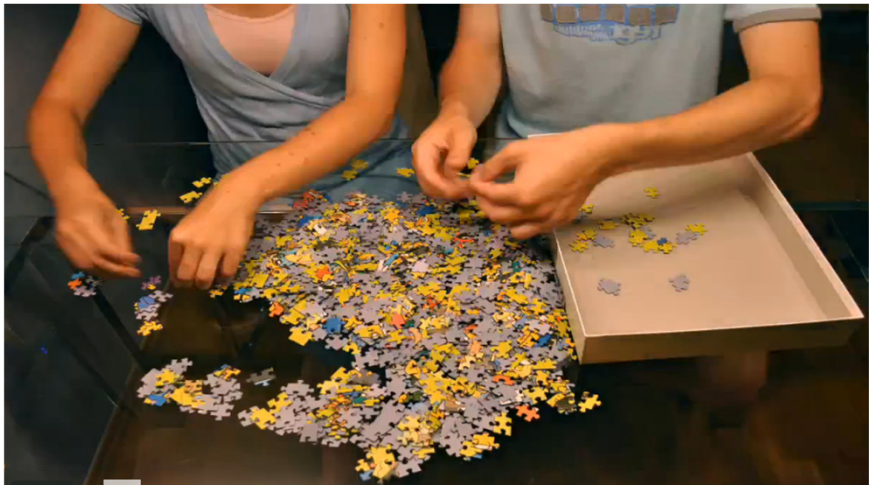
A



B

Quick - Which Of Those Would You Complete Faster?

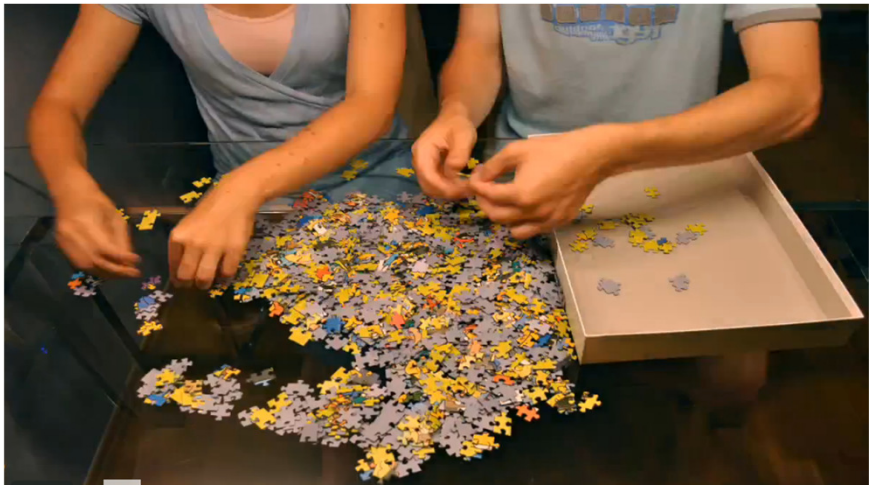
A



B

Quick – What If I Threw A Puzzle Piece From Another Puzzle In?

A



B

Two More Quick Thoughts….

Why We Use Machines Rather Than…

A Reminder Not To Skimp On HA and

Agenda

1. Why Big Data Is Both the Same and Different
2. The Role Of Analytics in Big Data Quality
3. Why Big Data Should Mean Better Quality
4. Some Closing Thoughts

Some Closing Thoughts

- **You cannot untangle quality of data and quality of analytics in the Big Data Space**
- **Try to extend your current tools to handle Big Data jobs as well**
  - Minimize moving parts
  - Leverage existing skills and governance
- **Strongly consider a COE approach**

Q&A and Discussion

Appendix

Hadoop is Well Suited for Handling Certain Types of Big Data Challenges

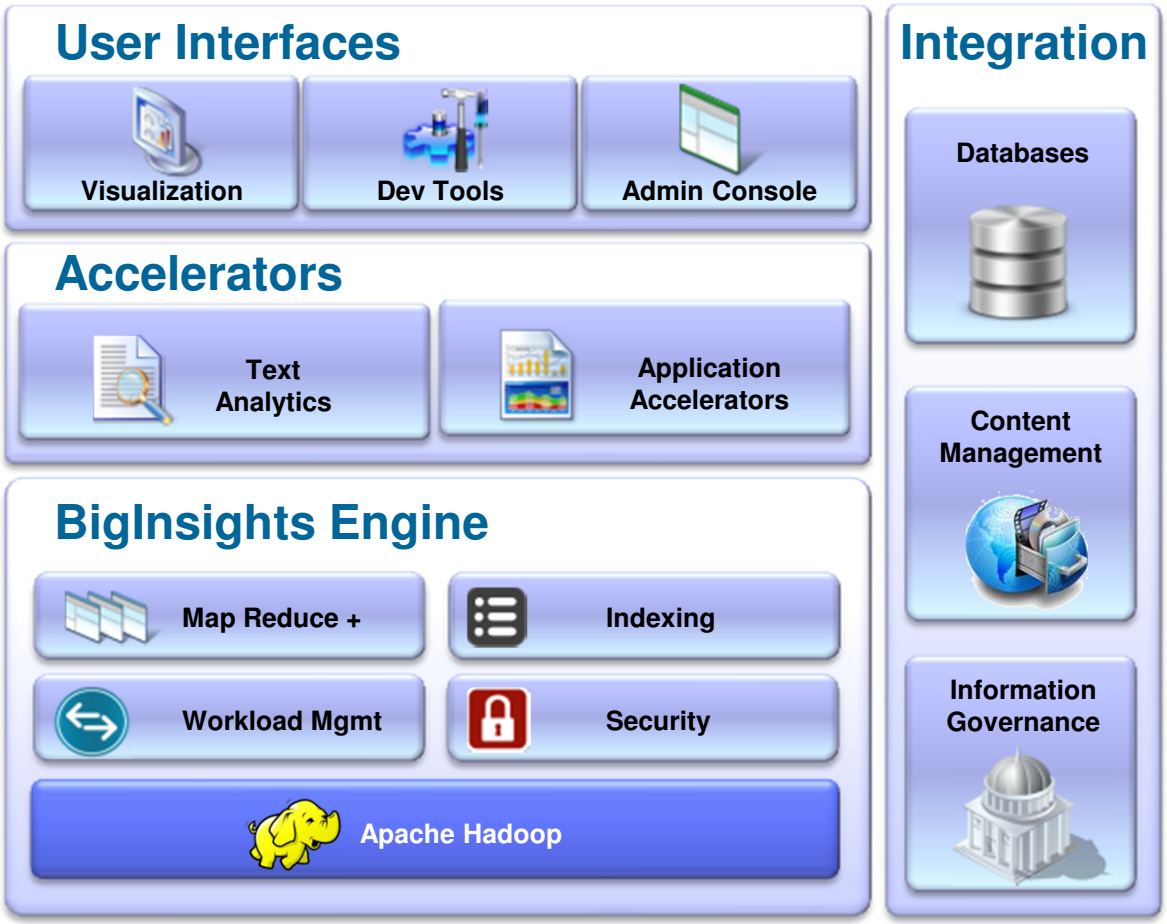
Analyzing larger volumes may provide better results

Deriving new insights from combinations of data types

Larger data volumes are cost prohibitive with existing technology

Exploring data – a sandbox for ad-hoc analytics

InfoSphere BigInsights – A Closer Look



More Than Hadoop

- Performance & workload optimizations
- Unique text analytic engines
- Spreadsheet-style visualization for data discovery & exploration
- Built-in IDE & admin consoles
- Enterprise-class security
- High-speed connectors to integration with other systems
- Analytical accelerators

Spreadsheet-Style Data Visualization

- **Ad-hoc analytics for LOB user**
- **Analyze a variety of data - unstructured and structured**
- **Browser-based**
- **Spreadsheet metaphor for exploring/ visualizing data**

Gather	Extract	Explore	Iterate
Crawl – gather statistically Adapter–gather dynamically	Document-level info Cleanse, normalize	Analyze, annotate, filter Visualize results	Iterate through any prior step

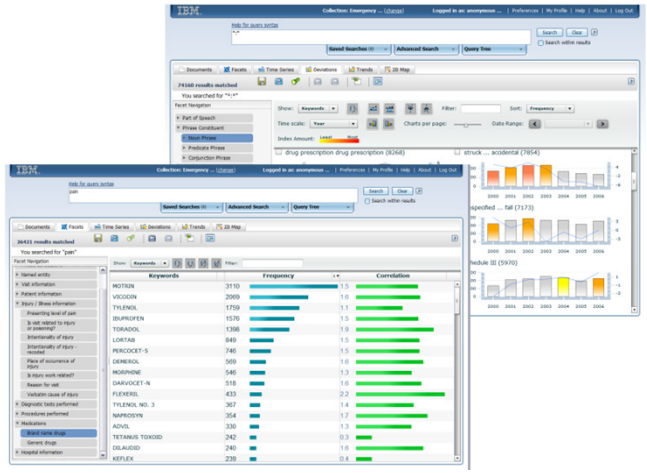
Text Analytics is the basis for Content Analytics

What is Text Analytics?

Text Analytics (NLP*) describes a set of linguistic, statistical, and machine learning techniques that allow text to be analyzed and key information extraction for business integration.

PC 143 (Hunter)
15 June 2006 23:47
Suspect identified himself as John Setsuko. Matched description given by night club doorman (IC1, Male, Ag 22-24 yrs, blue Everton shirt). Stopped whilst driving White Ford Mondeo, W563 WDL. Address given as 22 East Dene Ridge, Copdock, Ipswich. Searched at scene and found in possession of 1oz Cannabis Resin and lockable pocket knife.

Arresting_Officer	PC 143
Arrest_Date_Time	15/06/2006 : 23:47
Suspect_Forename	John
Suspect_Surname	Setsuko
Suspect_VRN	W563WDL
Suspect_Vehicle_Color	White
Suspect_Vehicle_Make	Ford Mondeo
Suspect_Addr_Street	22 East Dene Ridge
Suspect_Addr_Town	Ipswich
Evidence_1_Description	1 oz Cannabis Resin
Classification	Drug possession



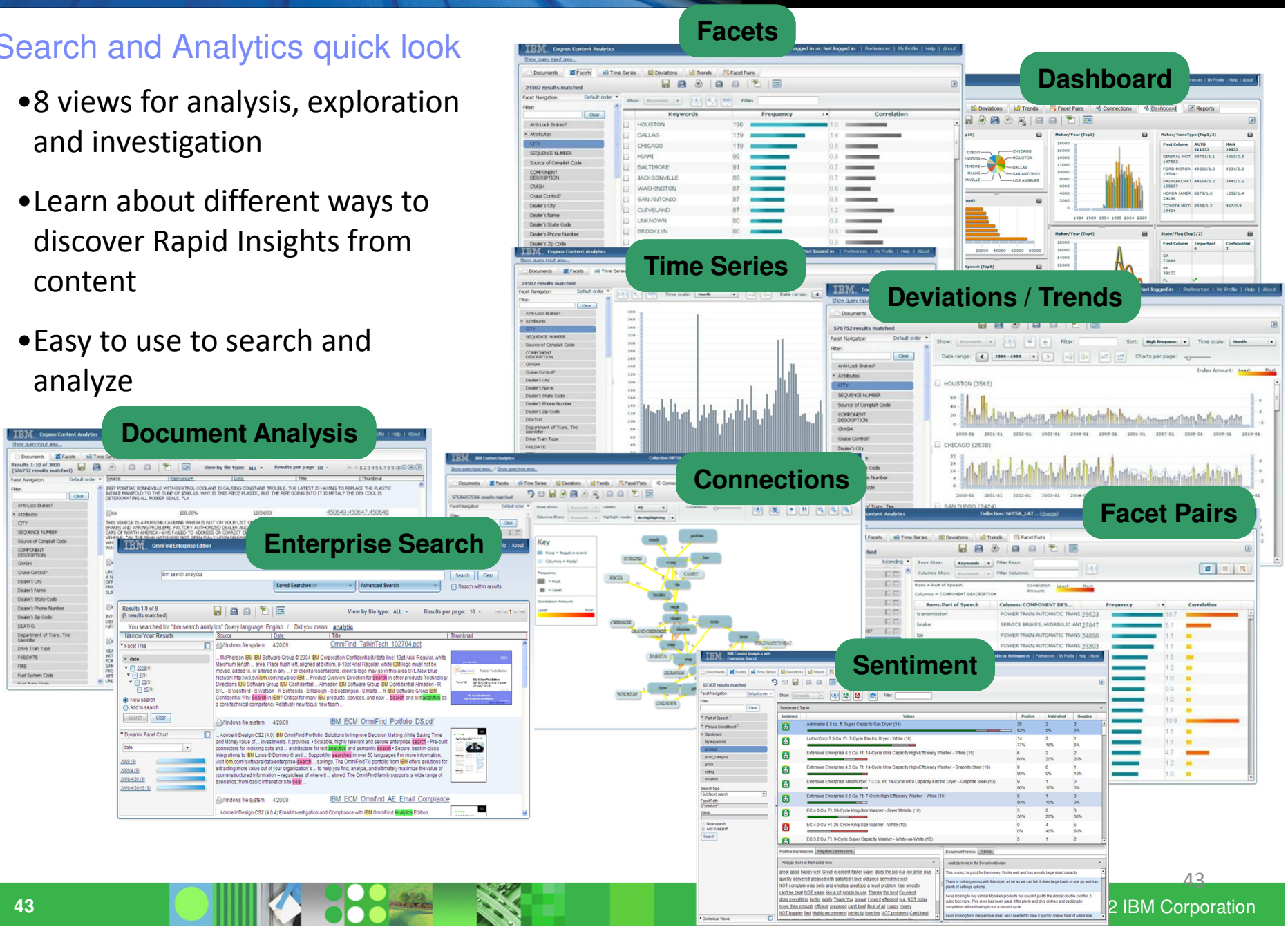
What is Content Analytics?

Content Analytics (Text Analytics + Mining) refers to the text analytics process plus the ability to visually identify and explore trends, patterns, and statistically relevant facts found in various types of content spread across internal and external content sources.

* Natural Language Processing

Search and Analytics quick look

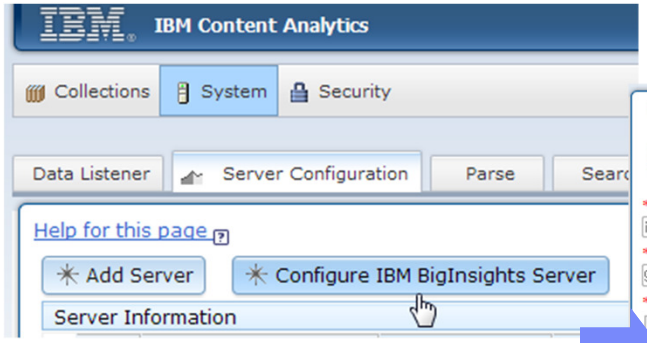
- 8 views for analysis, exploration and investigation
- Learn about different ways to discover Rapid Insights from content
- Easy to use to search and analyze



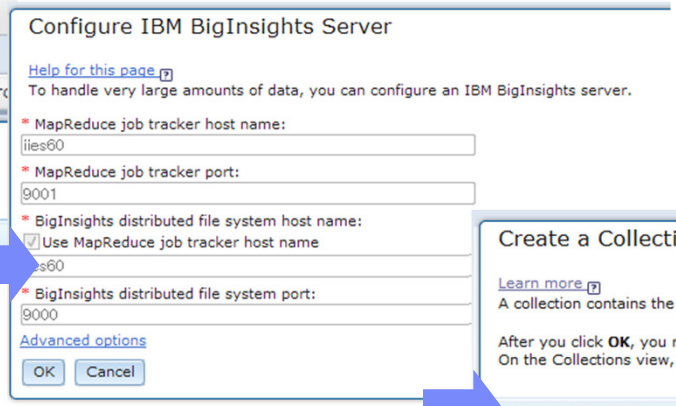
43

New in ICAwES v3.0 - Seamless Scale-out with BigInsights / Hadoop

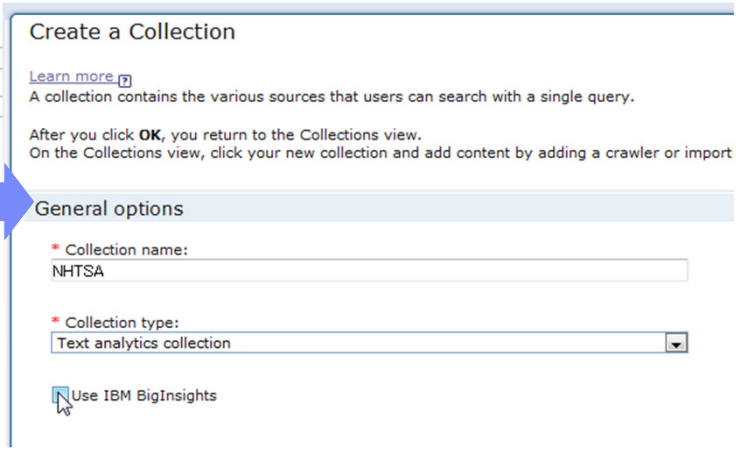
- **Select "Configure BigInsights Server"**



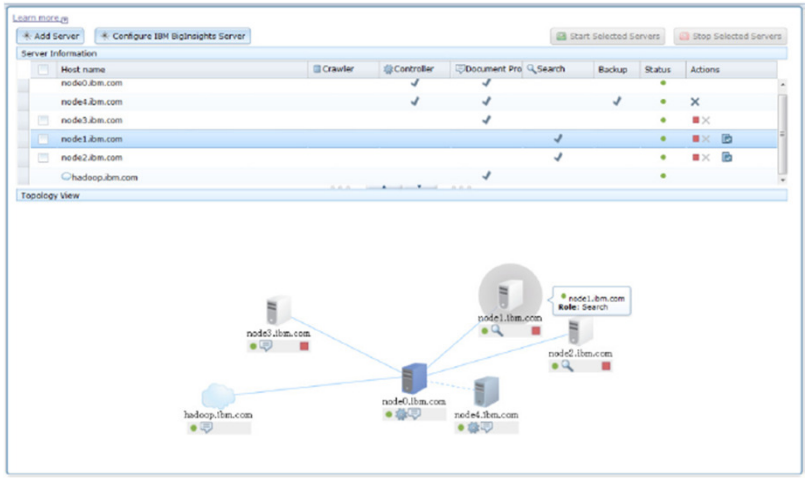
- **Enter the BigInsights Server Information**



- **Specify "Use IBM BigInsights" as a Collection setting**

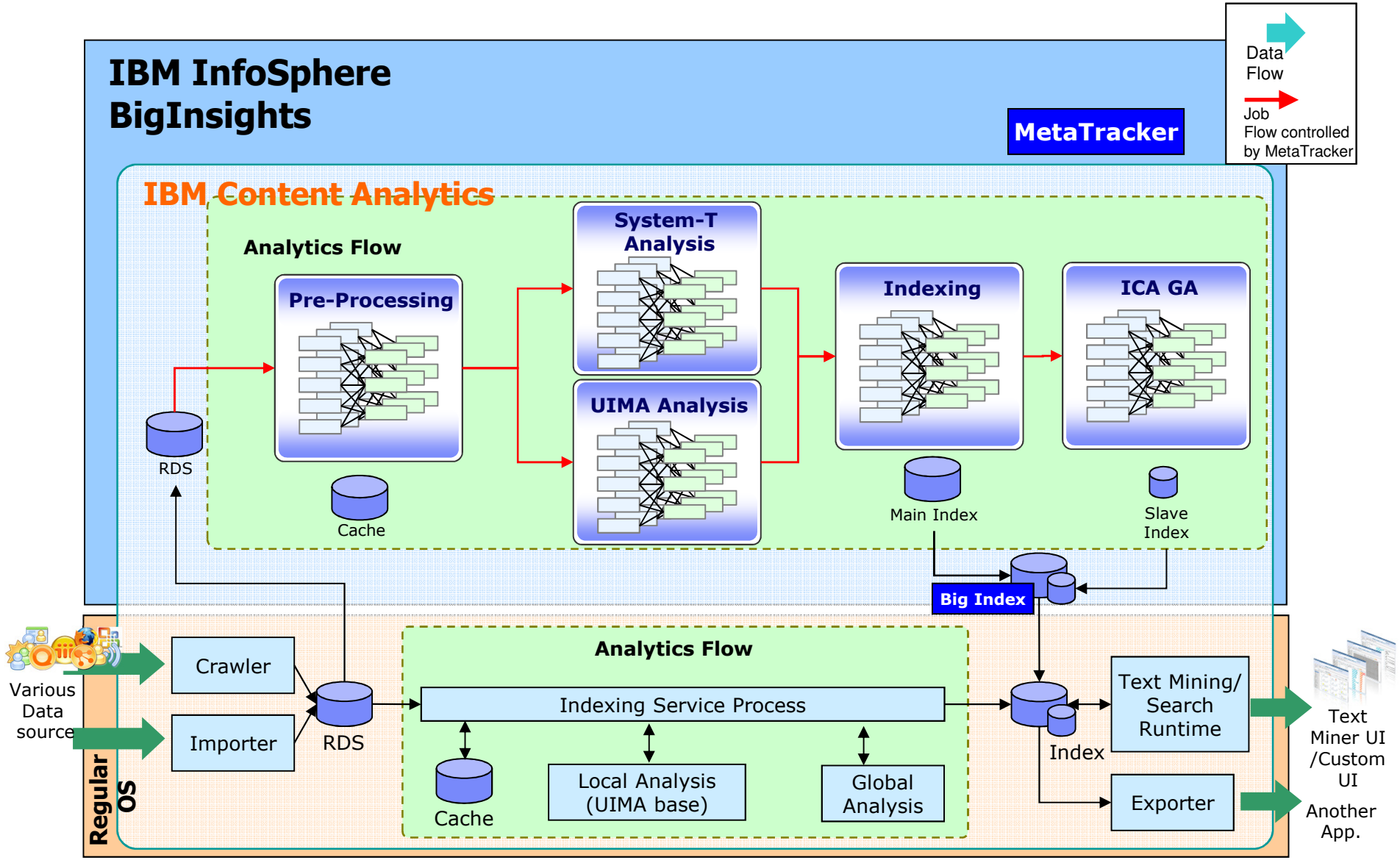


Admin user can confirm the setting on Topology View



…configuration files and ICA libraries, UIMA PEARs (including custom PEAR) and other required modules will be distributed to BIgInsights servers automatically for that collection

New in ICAwES v3.0 - Seamless Scale-out with BigInsights / Hadoop




THINK
BIG