Session 11413

Issues in Big Data: Analytics

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

Bob Foyle, bfoyle@us.ibm.com
Sr. Product Manager IBM Content Analytics with Enterprise Search

Abstract

- OK, so you’ve got all the data lying around and you’ve decided to make it work for you. Now, what you want to know is how to make sense of it. How do you do that? What do you get? Then, what do you do with it once you’ve gotten it? How do you know if the analysis is to be trusted?
- The speakers will address those issues, present IBM best practices, and discuss some real-world use cases.

Agenda

1 Big Data – Definitions & A Quick Story

2 Technology Primer

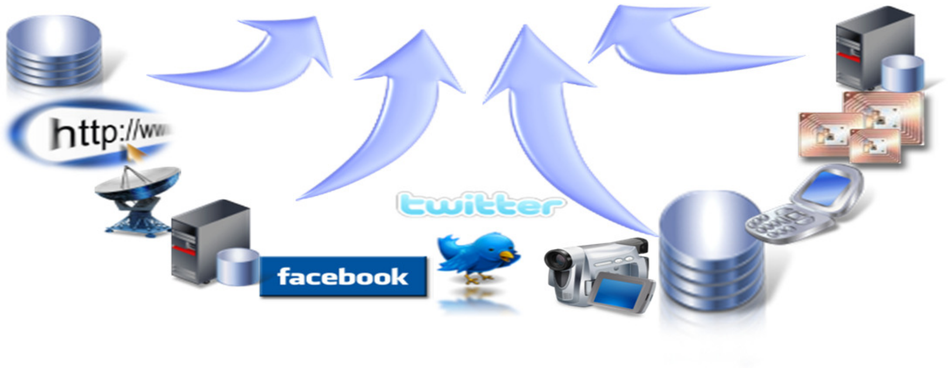
3 Dive on Content Analytics

4 Thoughts On Getting Started

Conventional Definition of “Big Data”

- Never before possible
- Solely defined by large volumes
- Only Unstructured Data
- Valuable insight, but difficult to extract
- Basically an ETL environment

This definition is <u>wrong</u>



Functional Big Data Definition

Dealing with information management **challenges** that don’t natively fit with traditional approaches to handling the problem

Functional Big Data Definition, Part 2

The technologies that deal with these problems are **broad and diverse, it is <u>not</u> just Hadoop**

Giving Rise to “Fit For Purpose Architectures”

The introduction of **purpose-built technology** focused on a **specific computing problem** that is compelling better than existing technologies

Giving Rise to “Fit For Purpose Architectures”

Match the compute problem to the best way to handle it **rather than assuming SQL by default**

…Which In-turn Enables Paradigm Shifts*

- **The Idea Of the Super-set**
- **Move from Sampling to “Absolute Knowledge"**
- **Combining Structured and Unstructured Analytics**
- **Importance of Augmented Decision Making**
- Streaming Computing Paradigm
- Lowering The Cost of Experimentation
- Changing Role Of Archiving

(* but the Laws of Gravity Still Apply!)

Agenda

1 Big Data – Definitions & A Quick Story

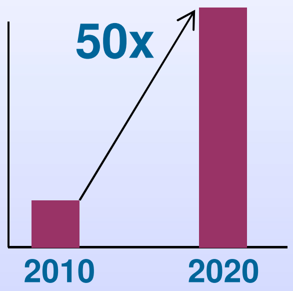
2 Technology Primer

3 Dive on Content Analytics

4 Thoughts On Getting Started

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 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

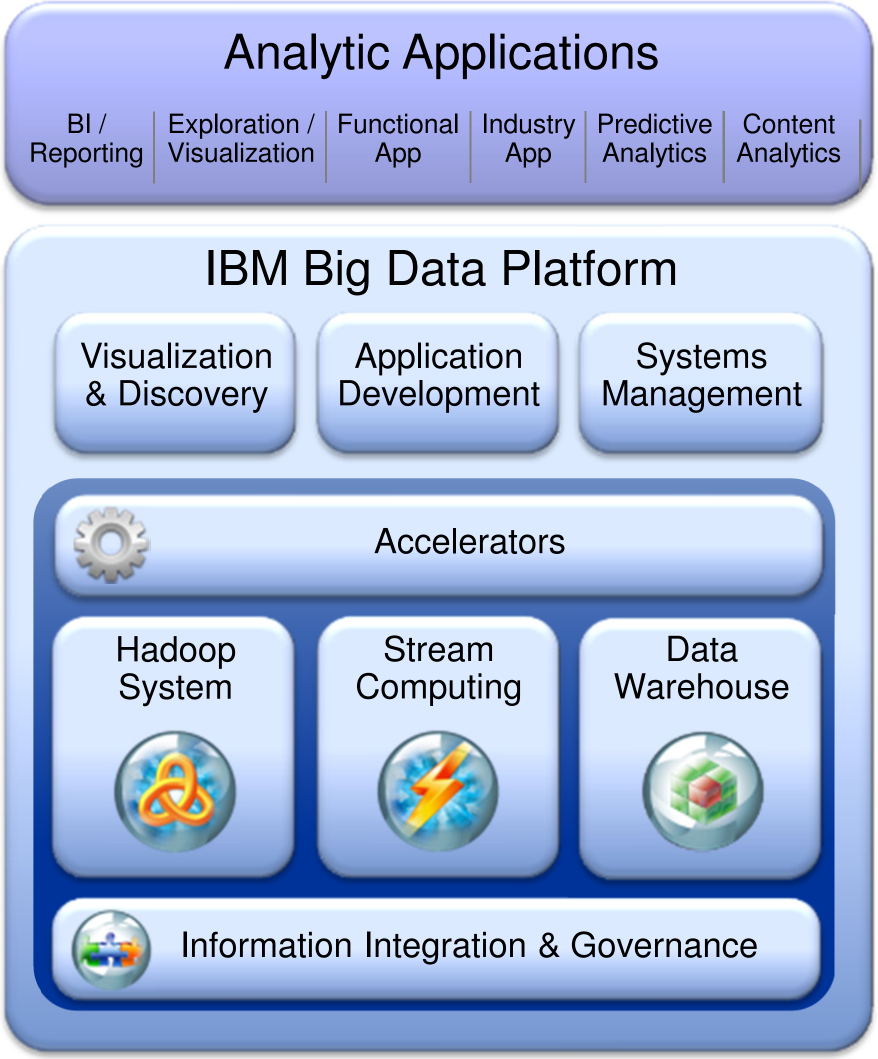
Exploring data –

a sandbox for ad-hoc analytics

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**


Analytic Applications

BI / Reporting | Exploration / Visualization | Functional App | Industry App | Predictive Analytics | Content Analytics

IBM Big Data Platform

Visualization & Discovery | Application Development | Systems Management

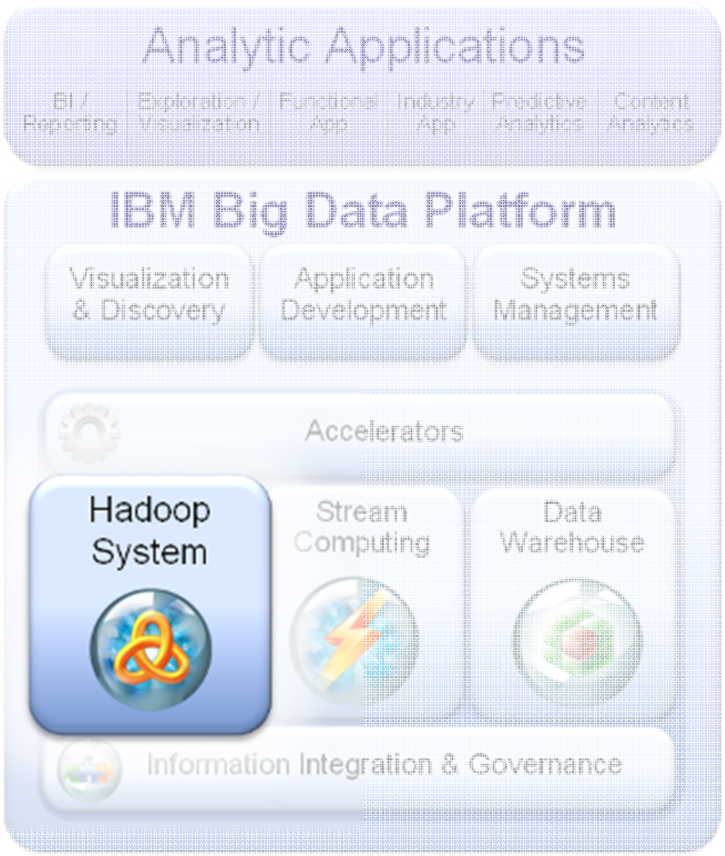
Accelerators

Hadoop System | Stream Computing | Data Warehouse

Information Integration & 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**



IBM Significantly Enhances Hadoop

hadoop

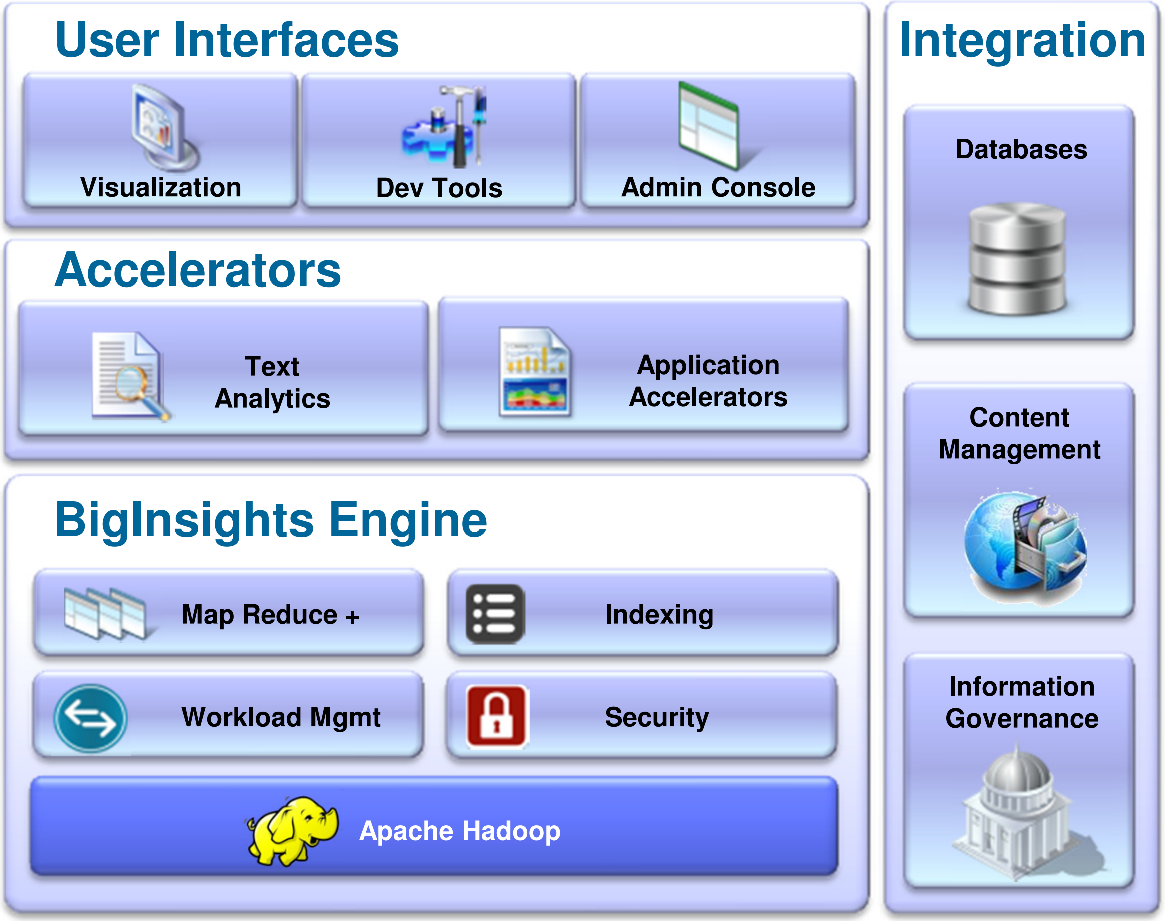
- **Scalable**
  - New nodes can be added on the fly.
- **Affordable**
  - Massively parallel computing on commodity servers
- **Flexible**
  - Hadoop is schema-less, and can absorb any type of data.
- **Fault Tolerant**
  - Through MapReduce software framework

+

IBM Innovation

- **Performance & reliability**
  - Adaptive MapReduce, Compression, Indexing, Flexible Scheduler
- **Analytic Accelerators**
- **Productivity Accelerators**
  - Web-based UIs
  - Tools to leverage existing skills
  - End-user visualization
- **Enterprise Integration**
  - To extend & enrich your information supply chain.

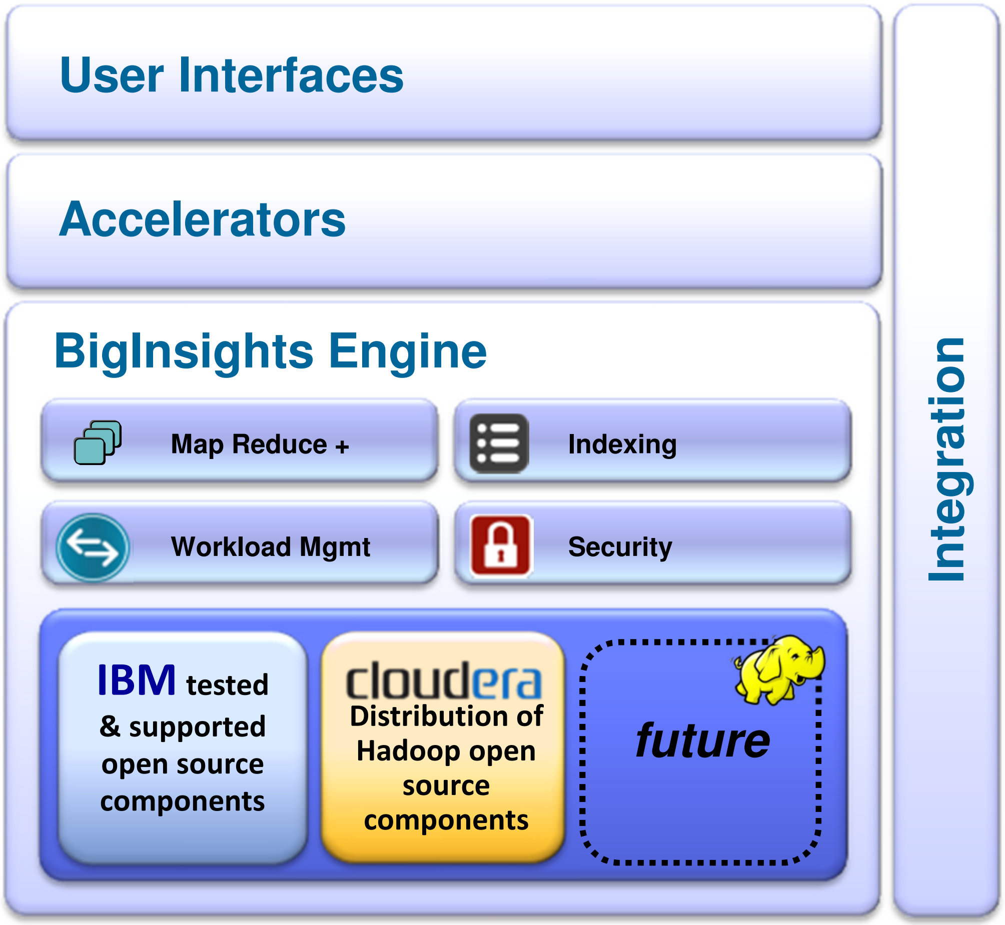
InfoSphere BigInsights – A Closer Look



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**

The Only Platform to Support Multiple Hadoop Distributions



- Provides a rich set of big data analytics and accelerators on top of open source
- Delivers a comprehensive big data platform on top of open source, that addresses all big data requirements.

Agenda

1 Big Data – Definitions & A Quick Story

2 Technology Primer

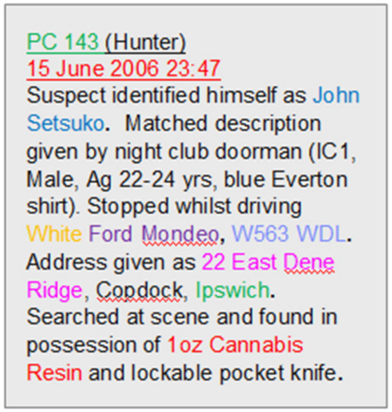
3 Dive on Content Analytics

4 Thoughts On Getting Started

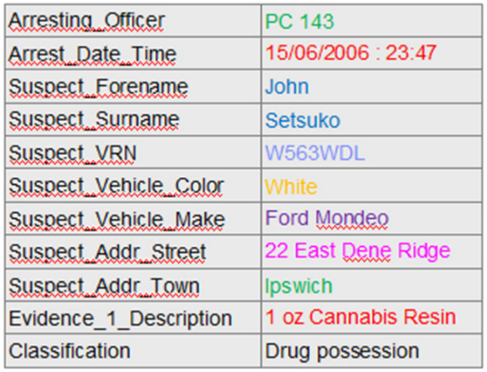
Text Analytics is the basis for Content Analytics

What is Text Analytics?

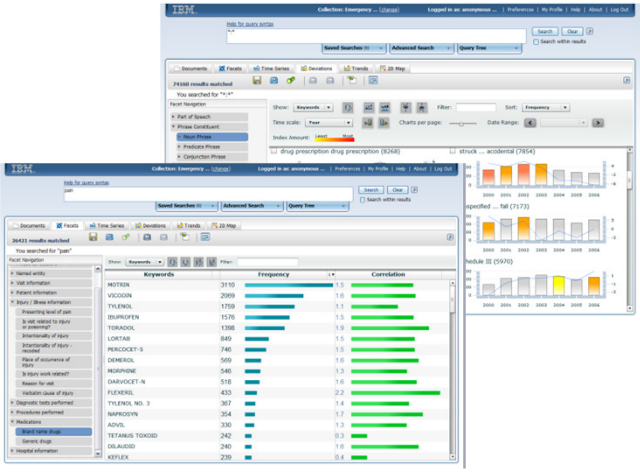
Text Analytics (NLP*) describes a set of linguistic, statistical, and machine learning techniques that allow text to be analyzed and key information extraction for business integration.



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



74160 results matched
You searched for "*:*"
drug prescription drug prescription (8268)

26421 results matched
You searched for "pain"

Keywords	Frequency	Correlation
MOTRIN	3110	1.5
VICODIN	2069	1.6
TYLENOL	1759	1.1
IBUPROFEN	1576	1.5
TORADOL	1398	1.9
LORTAB	849	1.5
PERCOCET-5	746	1.5
DEMEROL	569	1.6
MORPHINE	546	1.3
DARVOCET-N	518	1.6
FLEXERIL	433	2.2
TYLENOL NO. 3	367	1.4
NAPROSYN	354	1.7
ADVIL	330	1.3
TETANUS TOXOID	242	0.3
DILAUDID	240	1.6
KEFLEX	239	0.4

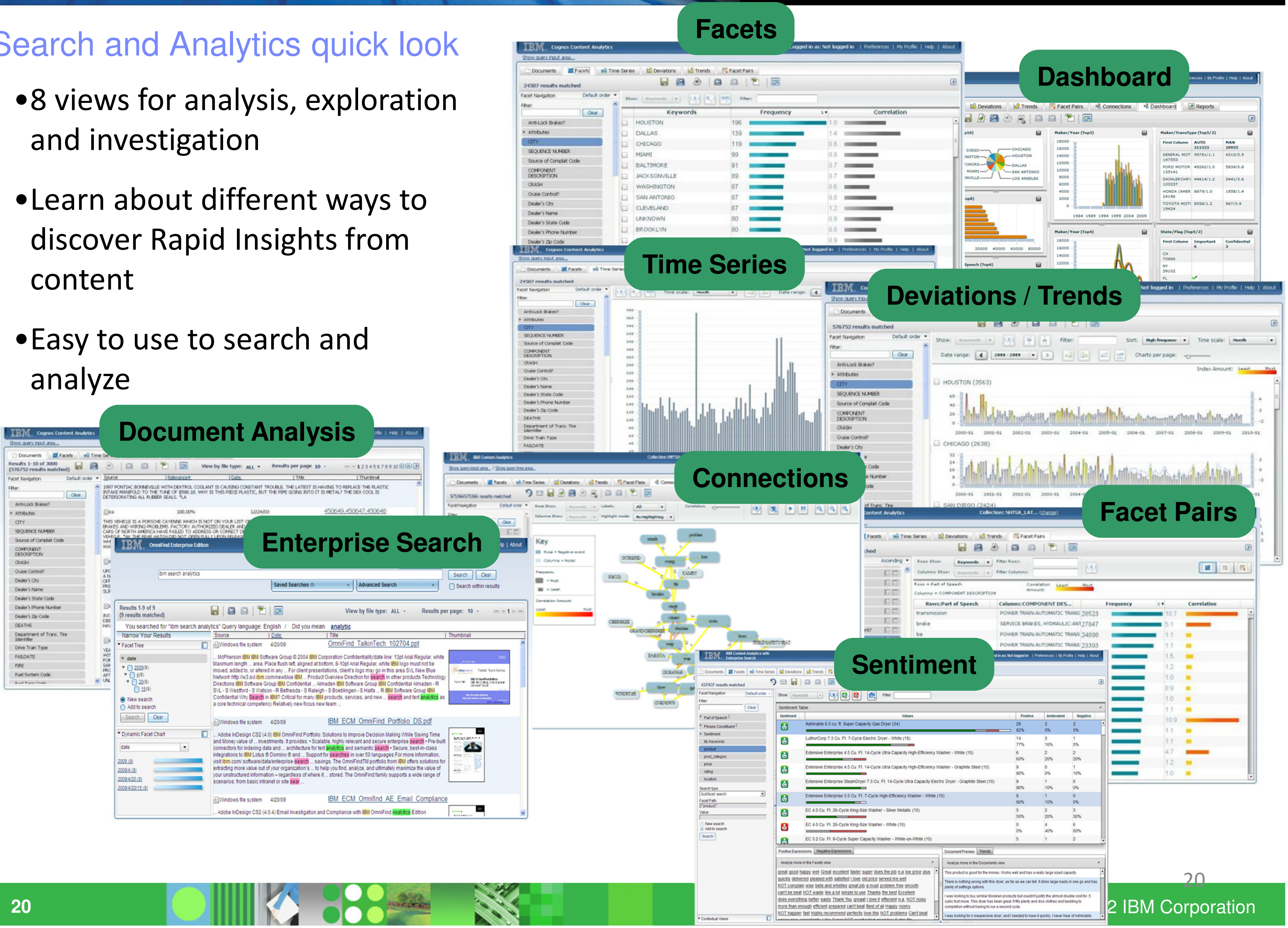
What is Content Analytics?

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

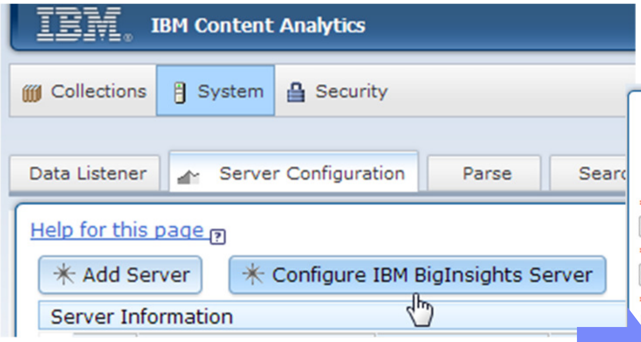
Search and Analytics quick look

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



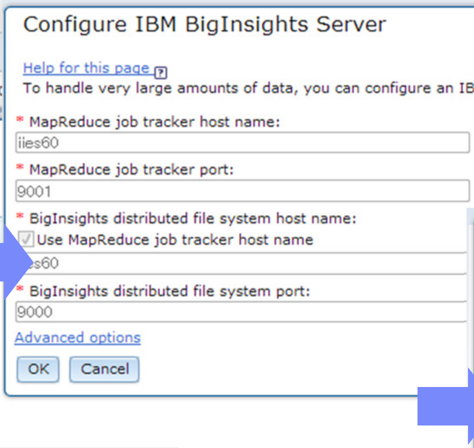
New in ICAwES v3.0 - Seamless Scale-out with BigInsights / Hadoop

- **Select “Configure BigInsights Server”**



IBM Content Analytics

- **Enter the BigInsights Server Information**



Configure IBM BigInsights Server

Help for this page

To handle very large amounts of data, you can configure an IBM BigInsights server.

* MapReduce job tracker host name: iies60

* MapReduce job tracker port: 9001

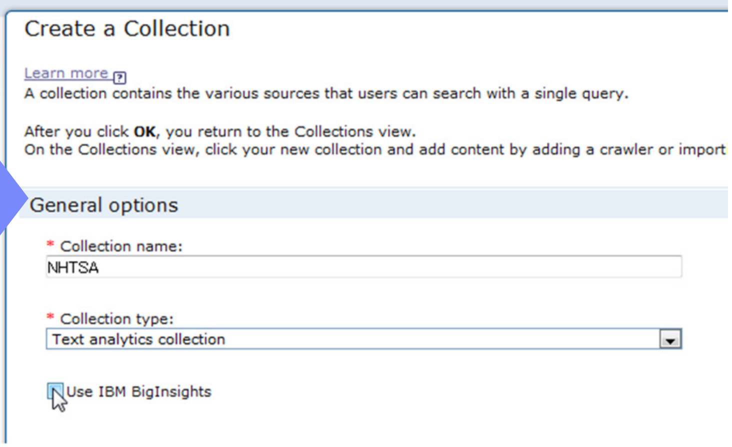
* BigInsights distributed file system host name:

☑ Use MapReduce job tracker host name

es60

* BigInsights distributed file system port: 9000

- **Specify “Use IBM BigInsights” as a Collection setting**



Create a Collection

Learn more

A collection contains the various sources that users can search with a single query.

After you click **OK**, you return to the Collections view.
On the Collections view, click your new collection and add content by adding a crawler or import

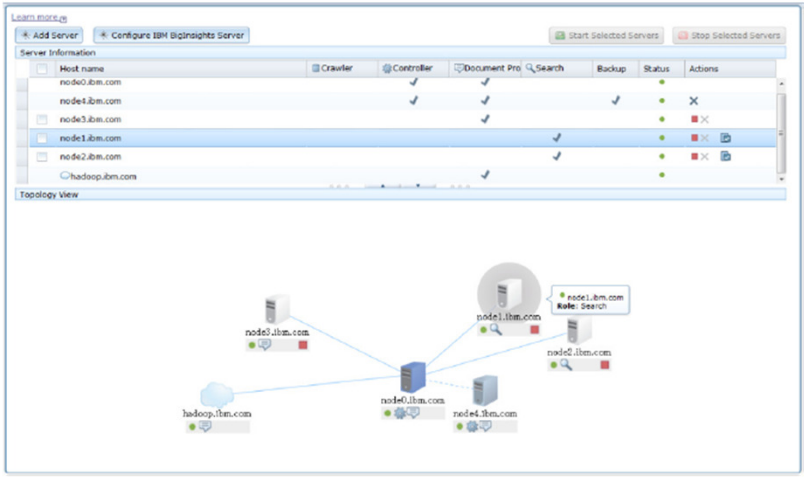
General options

* Collection name: NHTSA

* Collection type: Text analytics collection

□ Use IBM BigInsights

Admin user can confirm the setting on Topology View



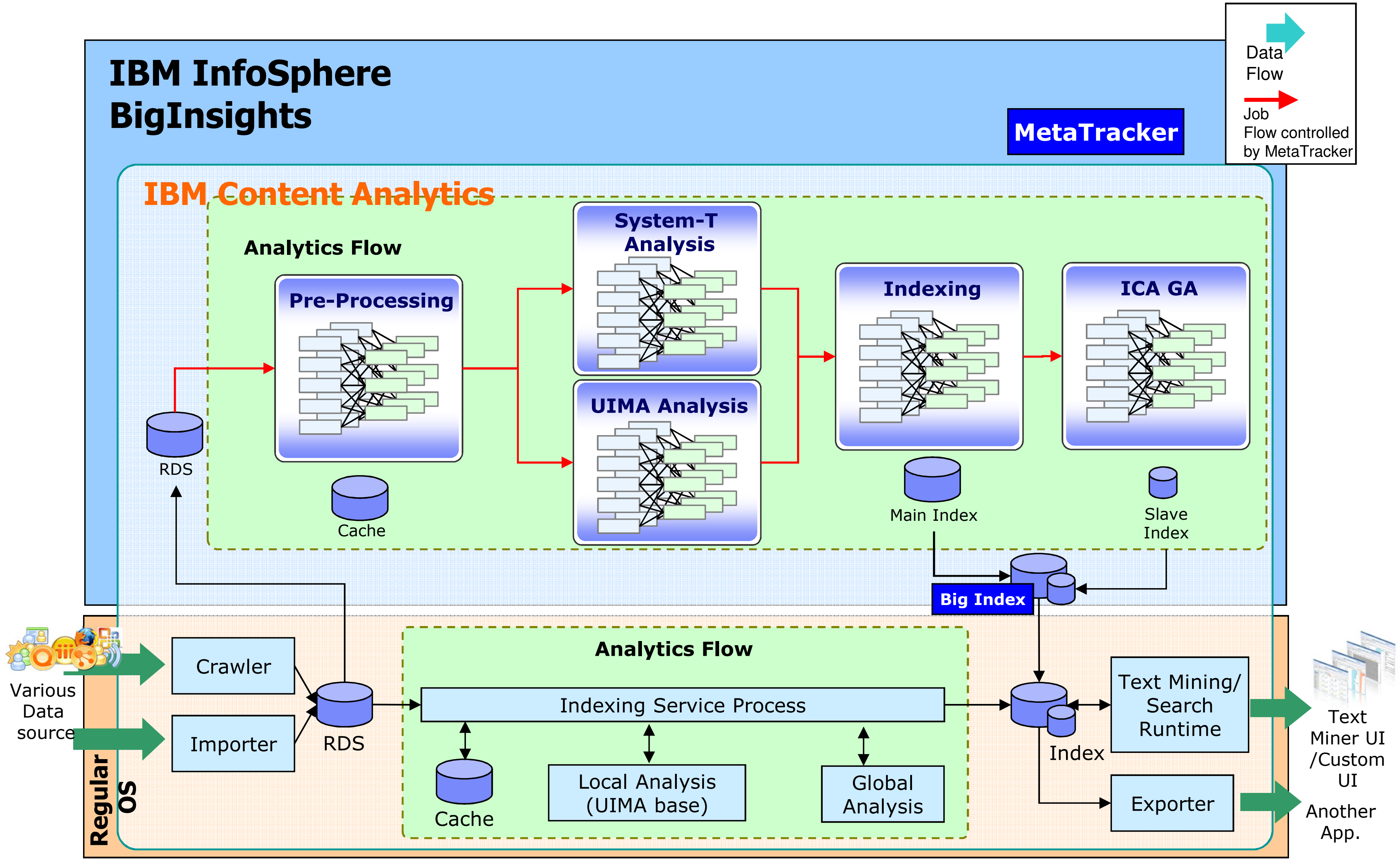
Server Information

Host name	Crawler	Controller	Document Pro	Search	Backup	Status	Actions
node0.ibm.com		✓	✓			●	
node4.ibm.com		✓	✓		✓	●	✕
node3.ibm.com			✓			●	■ ✕
node1.ibm.com				✓		●	■ ✕
node2.ibm.com				✓		●	■ ✕
hadoop.ibm.com			✓			●	

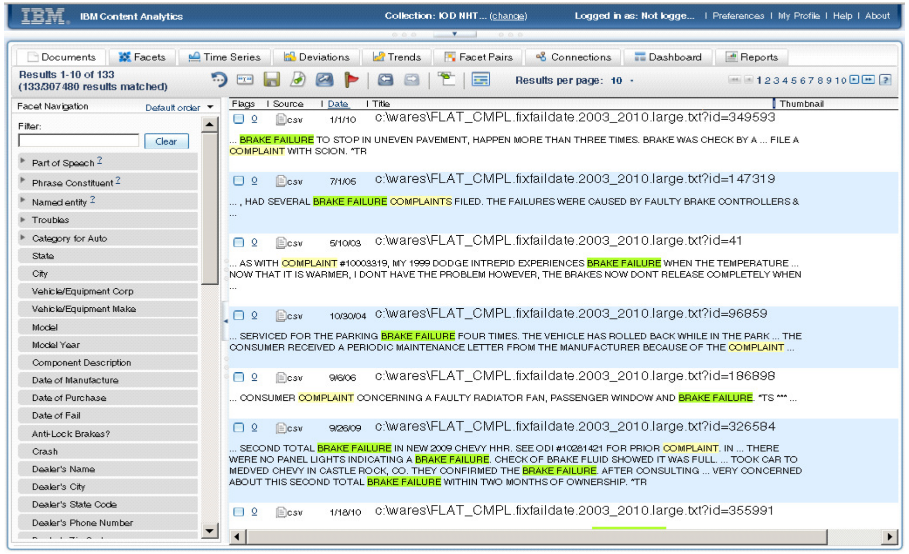
Topology View

…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



Documents View



Results 1-10 of 133
(133/307480 results matched)

Flags	Source	Date	Title	Thumbnail
□	csv	1/1/10	c:\wares\FLAT_CMPL.fixfaildate.2003_2010.large.txt?id=349593	
□	csv	7/1/05	c:\wares\FLAT_CMPL.fixfaildate.2003_2010.large.txt?id=147319	
□	csv	5/10/03	c:\wares\FLAT_CMPL.fixfaildate.2003_2010.large.txt?id=41	
□	csv	10/30/04	c:\wares\FLAT_CMPL.fixfaildate.2003_2010.large.txt?id=96859	
□	csv	9/6/06	c:\wares\FLAT_CMPL.fixfaildate.2003_2010.large.txt?id=186898	
□	csv	9/26/09	c:\wares\FLAT_CMPL.fixfaildate.2003_2010.large.txt?id=326584	
□	csv	1/18/10	c:\wares\FLAT_CMPL.fixfaildate.2003_2010.large.txt?id=355991	

id=349593: ... BRAKE FAILURE TO STOP IN UNEVEN PAVEMENT, HAPPEN MORE THAN THREE TIMES. BRAKE WAS CHECK BY A ... FILE A COMPLAINT WITH SCION. *TR

id=147319: ... , HAD SEVERAL BRAKE FAILURE COMPLAINTS FILED. THE FAILURES WERE CAUSED BY FAULTY BRAKE CONTROLLERS & ...

id=41: ... AS WITH COMPLAINT #10003319, MY 1999 DODGE INTREPID EXPERIENCES BRAKE FAILURE WHEN THE TEMPERATURE ... NOW THAT IT IS WARMER, I DONT HAVE THE PROBLEM HOWEVER, THE BRAKES NOW DONT RELEASE COMPLETELY WHEN ...

id=96859: ... SERVICED FOR THE PARKING BRAKE FAILURE FOUR TIMES. THE VEHICLE HAS ROLLED BACK WHILE IN THE PARK ... THE CONSUMER RECEIVED A PERIODIC MAINTENANCE LETTER FROM THE MANUFACTURER BECAUSE OF THE COMPLAINT ...

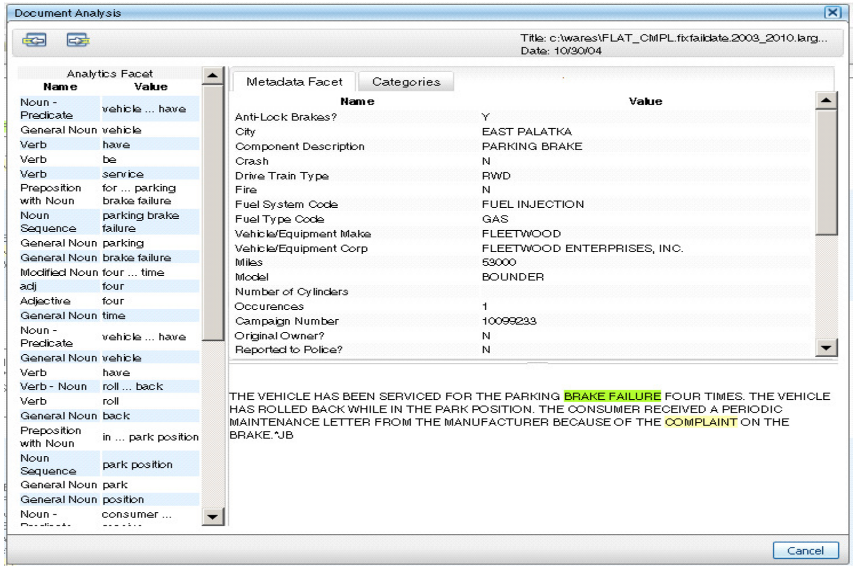
id=186898: ... CONSUMER COMPLAINT CONCERNING A FAULTY RADIATOR FAN, PASSENGER WINDOW AND BRAKE FAILURE. *TS *** ...

id=326584: ... SECOND TOTAL BRAKE FAILURE IN NEW 2009 CHEVY HHR. SEE ODI #10281421 FOR PRIOR COMPLAINT. IN ... THERE WERE NO PANEL LIGHTS INDICATING A BRAKE FAILURE. CHECK OF BRAKE FLUID SHOWED IT WAS FULL. ... TOOK CAR TO MEDVED CHEVY IN CASTLE ROCK, CO. THEY CONFIRMED THE BRAKE FAILURE. AFTER CONSULTING ... VERY CONCERNED ABOUT THIS SECOND TOTAL BRAKE FAILURE WITHIN TWO MONTHS OF OWNERSHIP. *TR

View the Documents based on current analysis criteria

- Flagging of documents
- Auto highlighting of key entities based on current query
- Export documents to be used in another system.

- **Document Analysis Dialog**
  - Preview the whole document
  - Display document metadata and annotated facets with highlighted texts



Title: c:\wares\FLAT_CMPL.fixfaildate.2003_2010.larg...
Date: 10/30/04

Analytics Facet

Name	Value
Noun - Predicate	vehicle ... have
General Noun	vehicle
Verb	have
Verb	be
Verb	service
Preposition with Noun	for ... parking brake failure
Noun Sequence	parking brake failure
General Noun	parking
General Noun	brake failure
Modified Noun	four ... time
adj	four
Adjective	four
General Noun	time
Noun - Predicate	vehicle ... have
General Noun	vehicle
Verb	have
Verb - Noun	roll ... back
Verb	roll
General Noun	back
Preposition with Noun	in ... park position
Noun Sequence	park position
General Noun	park
General Noun	position
Noun -	consumer ...

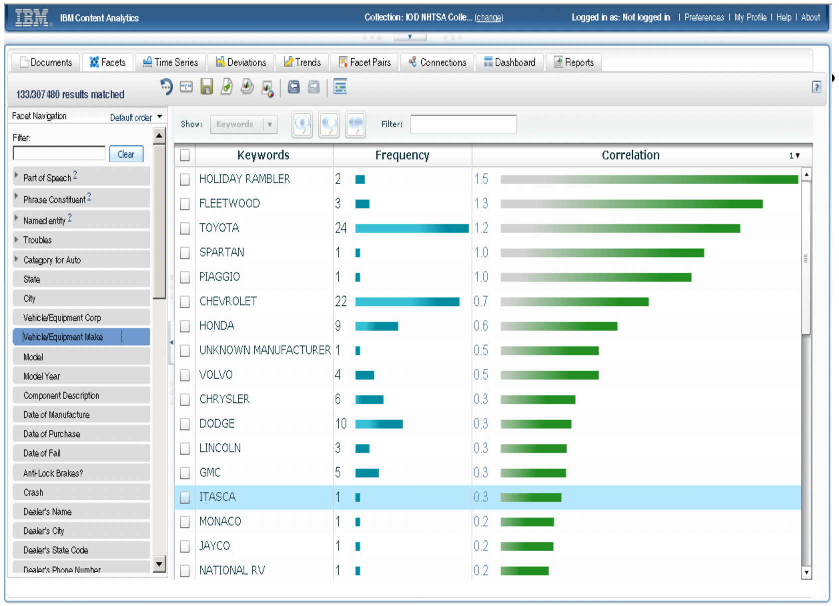
Name	Value
Anti-Lock Brakes?	Y
City	EAST PALATKA
Component Description	PARKING BRAKE
Crash	N
Drive Train Type	RWD
Fire	N
Fuel System Code	FUEL INJECTION
Fuel Type Code	GAS
Vehicle/Equipment Make	FLEETWOOD
Vehicle/Equipment Corp	FLEETWOOD ENTERPRISES, INC.
Miles	53000
Model	BOUNDER
Number of Cylinders	
Occurences	1
Campaign Number	10099233
Original Owner?	N
Reported to Police?	N

THE VEHICLE HAS BEEN SERVICED FOR THE PARKING BRAKE FAILURE FOUR TIMES. THE VEHICLE HAS ROLLED BACK WHILE IN THE PARK POSITION. THE CONSUMER RECEIVED A PERIODIC MAINTENANCE LETTER FROM THE MANUFACTURER BECAUSE OF THE COMPLAINT ON THE BRAKE.*JB

Facets View

Show the frequency and correlation indices of keywords belong to the selected facet

- *Add selected keyword to current search condition*



IBM IBM Content Analytics
Collection: IOD NHTSA Colle... (change)
Logged in as: Not logged in | Preferences | My Profile | Help | About
Documents
Facets
Time Series
Deviations
Trends
Facet Pairs
Connections
Dashboard
Reports

133/307 480 results matched

Facet Navigation — Default order

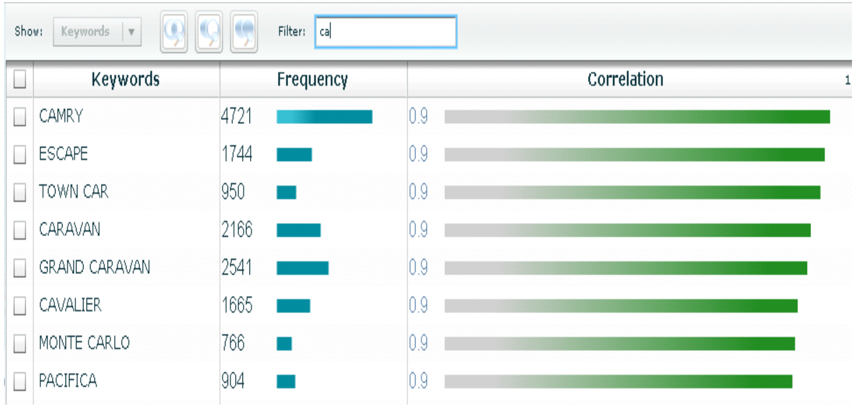
Filter: Clear

- Part of Speech ?
- Phrase Constituent ?
- Named entity ?
- Troubles
- Category for Auto
- State
- City
- Vehicle/Equipment Corp
- Vehicle/Equipment Make
- Model
- Model Year
- Component Description
- Date of Manufacture
- Date of Purchase
- Date of Fail
- Anti-Lock Brakes?
- Crash
- Dealer's Name
- Dealer's City
- Dealer's State Code
- Dealer's Phone Number

Show: Keywords Filter:

□	Keywords	Frequency	Correlation 1▼
□	HOLIDAY RAMBLER	2	1.5
□	FLEETWOOD	3	1.3
□	TOYOTA	24	1.2
□	SPARTAN	1	1.0
□	PIAGGIO	1	1.0
□	CHEVROLET	22	0.7
□	HONDA	9	0.6
□	UNKNOWN MANUFACTURER	1	0.5
□	VOLVO	4	0.5
□	CHRYSLER	6	0.3
□	DODGE	10	0.3
□	LINCOLN	3	0.3
□	GMC	5	0.3
□	ITASCA	1	0.3
□	MONACO	1	0.2
□	JAYCO	1	0.2
□	NATIONAL RV	1	0.2

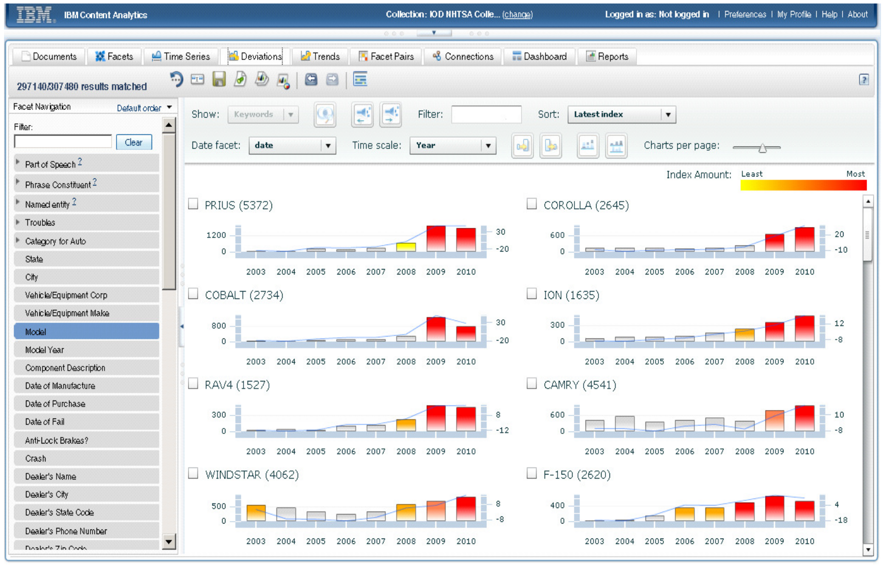
- *Quick filter*
- *Multi-column sort*



Show: Keywords Filter: ca

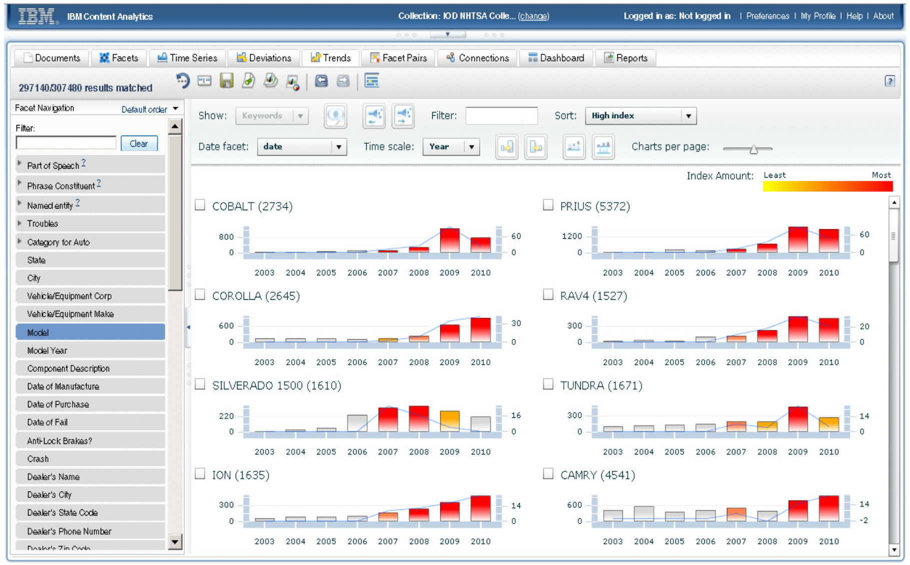
□	Keywords	Frequency	Correlation 1
□	CAMRY	4721	0.9
□	ESCAPE	1744	0.9
□	TOWN CAR	950	0.9
□	CARAVAN	2166	0.9
□	GRAND CARAVAN	2541	0.9
□	CAVALIER	1665	0.9
□	MONTE CARLO	766	0.9
□	PACIFICA	904	0.9

Deviations View and Trends View



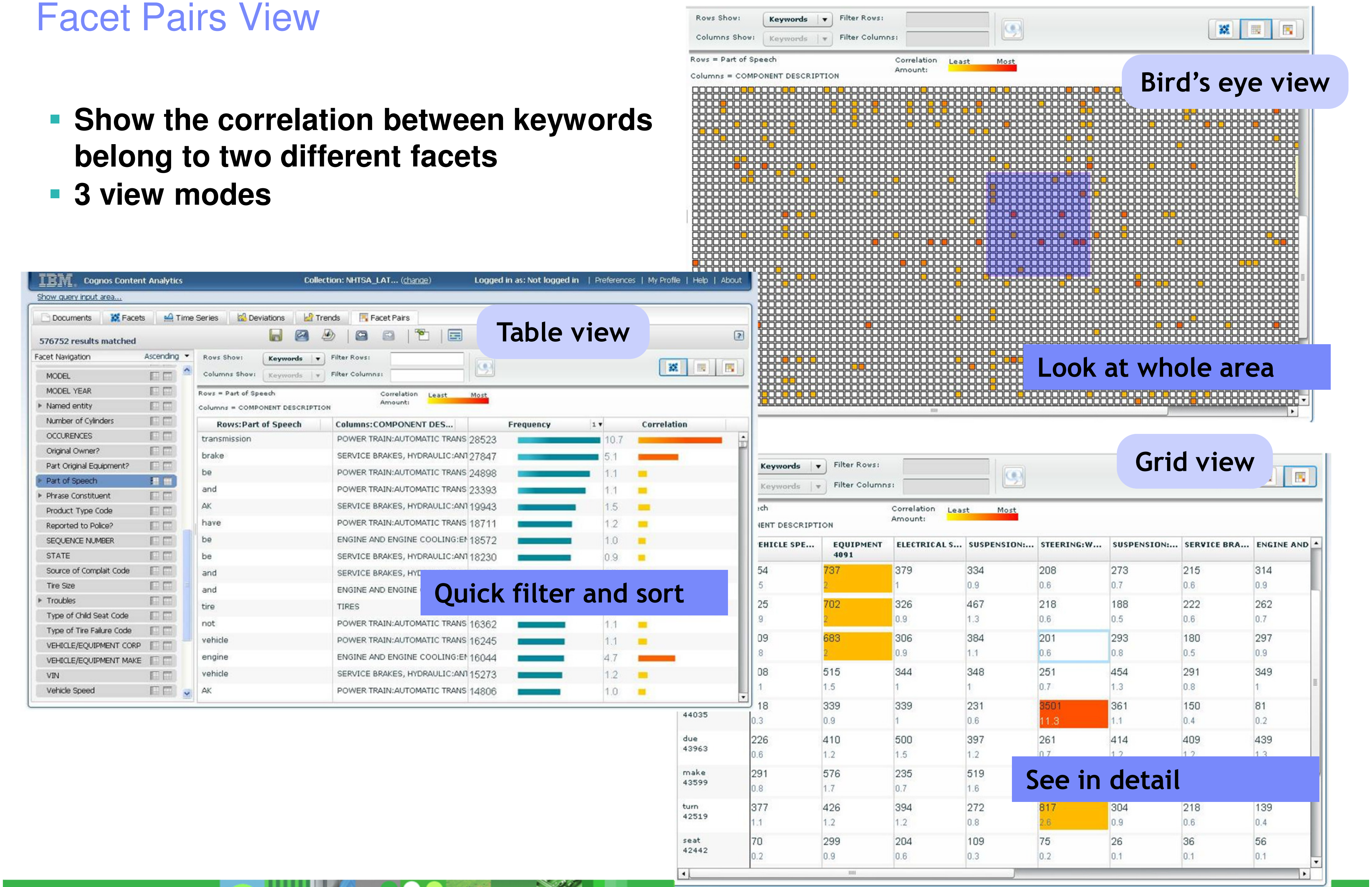
- *Deviations View*
  - *show how keyword deviates from average occurrence between other keywords.*
  - *Identify patterns that are cyclic.*

- *Trends View*
  - *show how keyword in each time period deviates from average occurrence (against the whole current matched documents)..*
  - *Predict pattern behaviors in future.*



Facet Pairs View

- Show the correlation between keywords belong to two different facets
- 3 view modes



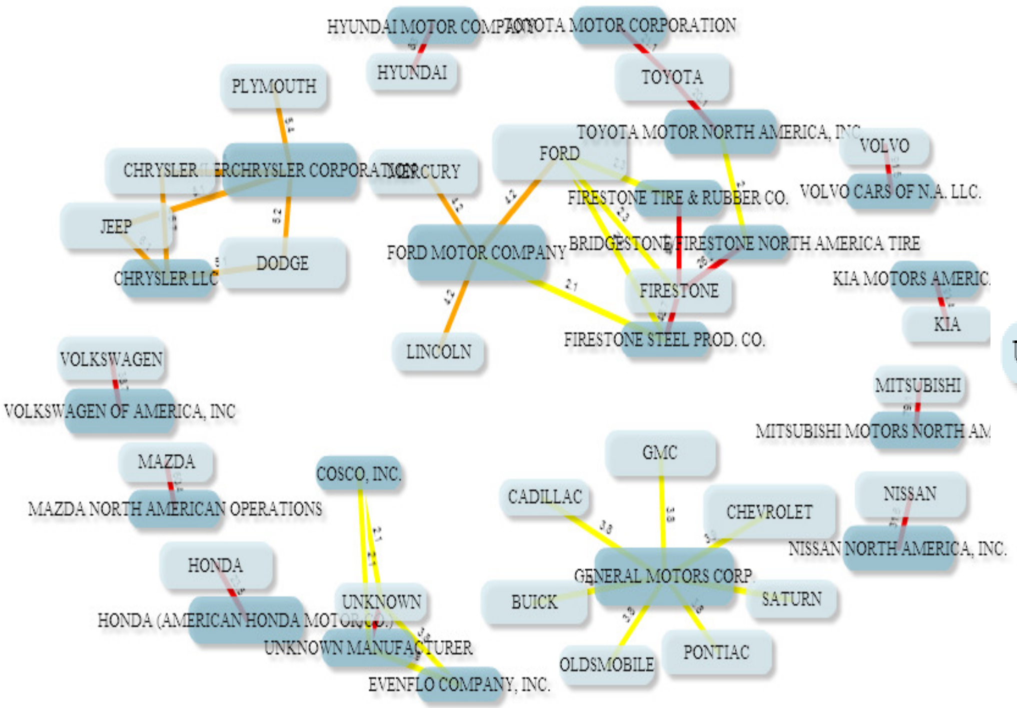
Rows Show: Keywords
Columns Show: Keywords
Filter Rows:
Filter Columns:
Rows = Part of Speech
Columns = COMPONENT DESCRIPTION

Rows:Part of Speech	Columns:COMPONENT DES...	Frequency	Correlation
transmission	POWER TRAIN:AUTOMATIC TRANS	28523	10.7
brake	SERVICE BRAKES, HYDRAULIC:ANT	27847	5.1
be	POWER TRAIN:AUTOMATIC TRANS	24898	1.1
and	POWER TRAIN:AUTOMATIC TRANS	23393	1.1
AK	SERVICE BRAKES, HYDRAULIC:ANT	19943	1.5
have	POWER TRAIN:AUTOMATIC TRANS	18711	1.2
be	ENGINE AND ENGINE COOLING:EN	18572	1.0
be	SERVICE BRAKES, HYDRAULIC:ANT	18230	0.9
and	SERVICE BRAKES, HYD		
and	ENGINE AND ENGINE		
tire	TIRES		
not	POWER TRAIN:AUTOMATIC TRANS	16362	1.1
vehicle	POWER TRAIN:AUTOMATIC TRANS	16245	1.1
engine	ENGINE AND ENGINE COOLING:EN	16044	4.7
vehicle	SERVICE BRAKES, HYDRAULIC:ANT	15273	1.2
AK	POWER TRAIN:AUTOMATIC TRANS	14806	1.0

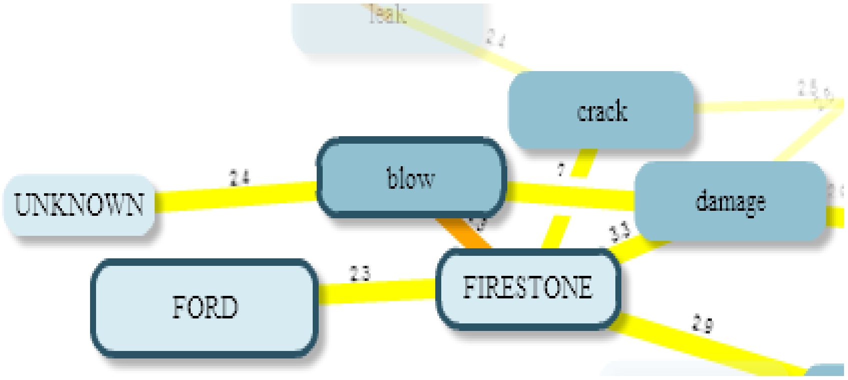
	EHICLE SPE...	EQUIPMENT 4091	ELECTRICAL S...	SUSPENSION:...	STEERING:W...	SUSPENSION:...	SERVICE BRA...	ENGINE AND
	54 5	737 2	379 1	334 0.9	208 0.6	273 0.7	215 0.6	314 0.9
	25 9	702 2	326 0.9	467 1.3	218 0.6	188 0.5	222 0.6	262 0.7
	09 8	683 2	306 0.9	384 1.1	201 0.6	293 0.8	180 0.5	297 0.9
	08 1	515 1.5	344 1	348 1	251 0.7	454 1.3	291 0.8	349 1
44035	18 0.3	339 0.9	339 1	231 0.6	3501 11.3	361 1.1	150 0.4	81 0.2
due 43963	226 0.6	410 1.2	500 1.5	397 1.2	261 0.7	414 1.2	409 1.2	439 1.3
make 43599	291 0.8	576 1.7	235 0.7	519 1.6				
turn 42519	377 1.1	426 1.2	394 1.2	272 0.8	817 2.6	304 0.9	218 0.6	139 0.4
seat 42442	70 0.2	299 0.9	204 0.6	109 0.3	75 0.2	26 0.1	36 0.1	56 0.1

Connections View links highly correlated terms

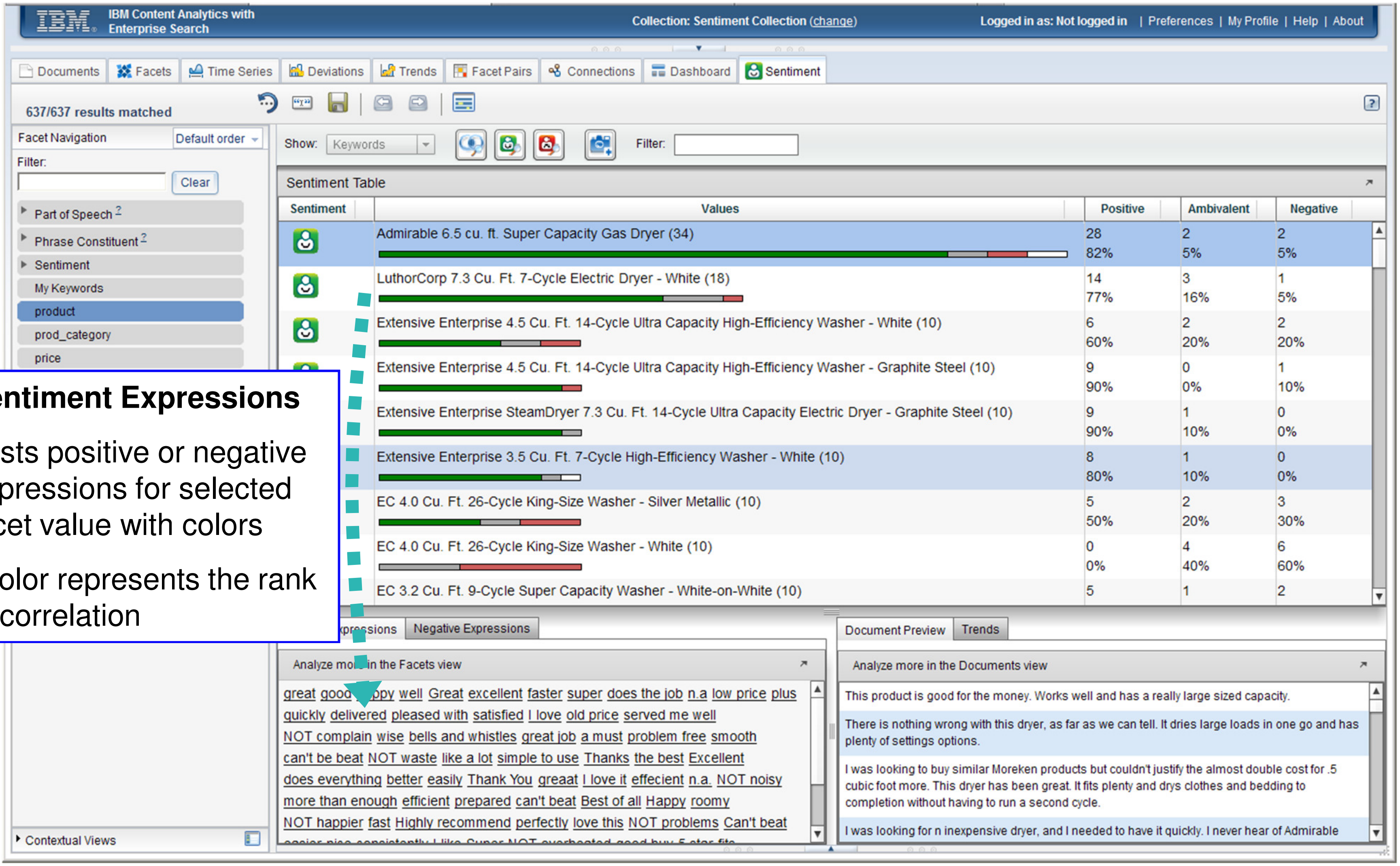
- Show relationship between multiple facet values
- Connections between nodes represents correlation between two facet values
- Color of line represents the importance of correlation index (red is the highest)



Identify relations between "FORD", "blow" and "FIRESTONE"



New in ICAwES v3.0 – Sentiment Analytics



637/637 results matched

Sentiment	Values	Positive	Ambivalent	Negative
☺	Admirable 6.5 cu. ft. Super Capacity Gas Dryer (34)	28 82%	2 5%	2 5%
☺	LuthorCorp 7.3 Cu. Ft. 7-Cycle Electric Dryer - White (18)	14 77%	3 16%	1 5%
☺	Extensive Enterprise 4.5 Cu. Ft. 14-Cycle Ultra Capacity High-Efficiency Washer - White (10)	6 60%	2 20%	2 20%
	Extensive Enterprise 4.5 Cu. Ft. 14-Cycle Ultra Capacity High-Efficiency Washer - Graphite Steel (10)	9 90%	0 0%	1 10%
	Extensive Enterprise SteamDryer 7.3 Cu. Ft. 14-Cycle Ultra Capacity Electric Dryer - Graphite Steel (10)	9 90%	1 10%	0 0%
	Extensive Enterprise 3.5 Cu. Ft. 7-Cycle High-Efficiency Washer - White (10)	8 80%	1 10%	0 0%
	EC 4.0 Cu. Ft. 26-Cycle King-Size Washer - Silver Metallic (10)	5 50%	2 20%	3 30%
	EC 4.0 Cu. Ft. 26-Cycle King-Size Washer - White (10)	0 0%	4 40%	6 60%
	EC 3.2 Cu. Ft. 9-Cycle Super Capacity Washer - White-on-White (10)	5	1	2

great good happy well Great excellent faster super does the job n.a low price plus quickly delivered pleased with satisfied I love old price served me well NOT complain wise bells and whistles great job a must problem free smooth can't be beat NOT waste like a lot simple to use Thanks the best Excellent does everything better easily Thank You greaat I love it effecient n.a. NOT noisy more than enough efficient prepared can't beat Best of all Happy roomy NOT happier fast Highly recommend perfectly love this NOT problems Can't beat easier nice consistently I like Super NOT overheated good buy 5 star fits

This product is good for the money. Works well and has a really large sized capacity.

There is nothing wrong with this dryer, as far as we can tell. It dries large loads in one go and has plenty of settings options.

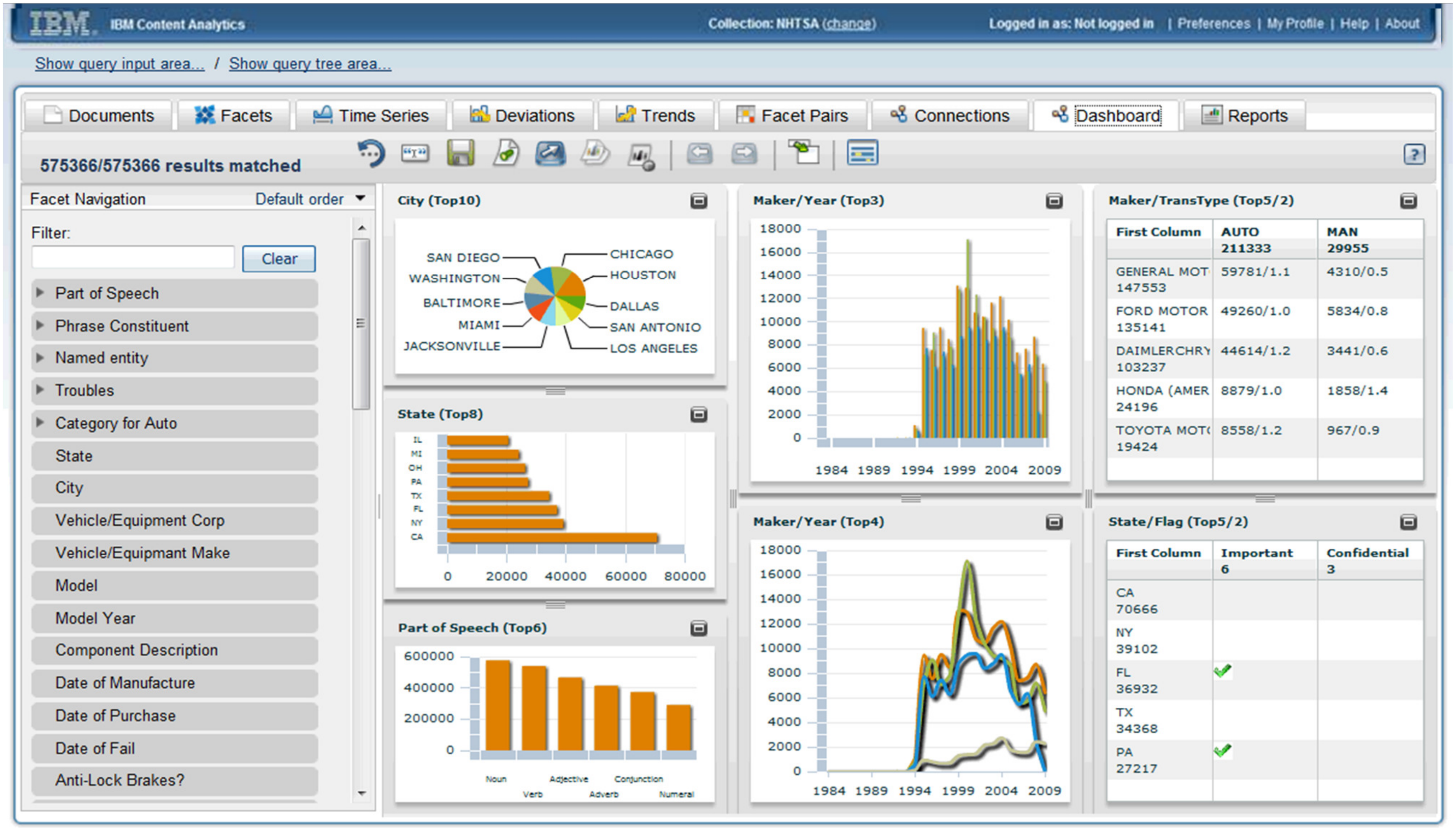
I was looking to buy similar Moreken products but couldn't justify the almost double cost for .5 cubic foot more. This dryer has been great. It fits plenty and drys clothes and bedding to completion without having to run a second cycle.

I was looking for n inexpensive dryer, and I needed to have it quickly. I never hear of Admirable

Sentiment Expressions

- Lists positive or negative expressions for selected facet value with colors
- Color represents the rank of correlation

Create Dashboard Views for Executive Summaries



575366/575366 results matched

Maker/TransType (Top5/2)

First Column	AUTO 211333	MAN 29955
GENERAL MOT 147553	59781/1.1	4310/0.5
FORD MOTOR 135141	49260/1.0	5834/0.8
DAIMLERCHRY 103237	44614/1.2	3441/0.6
HONDA (AMER 24196	8879/1.0	1858/1.4
TOYOTA MOT(19424	8558/1.2	967/0.9

State/Flag (Top5/2)

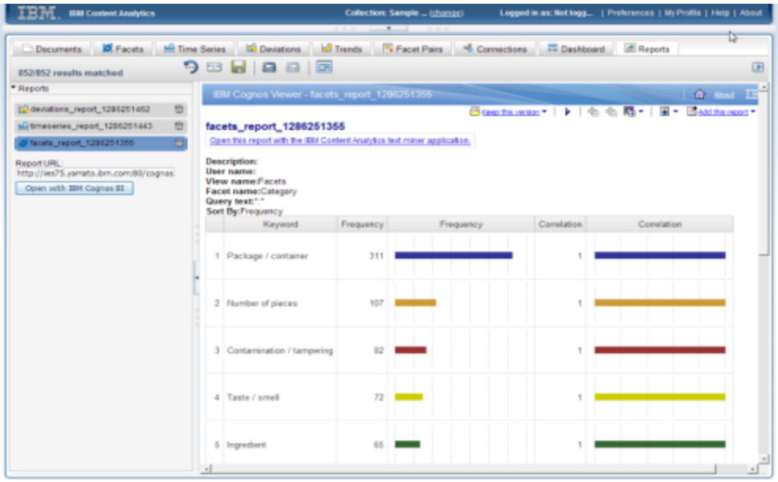
First Column	Important 6	Confidential 3
CA 70666		
NY 39102		
FL 36932	✓	
TX 34368		
PA 27217	✓	

Click here

ICA integrates with Cognos BI reports

From ICA Text Miner, a user can:

- Issue a request to create a report
- List the created reports
- Open the created report
- Delete the created report
- Cognos reports can link to and from Text Miner



852/852 results matched

Reports
- deviations_report_1286251462
- timeseries_report_1286251443
- facets_report_1286251355

Report URL: http://ies75.yamato.ibm.com:80/cognos

Open with IBM Cognos BI

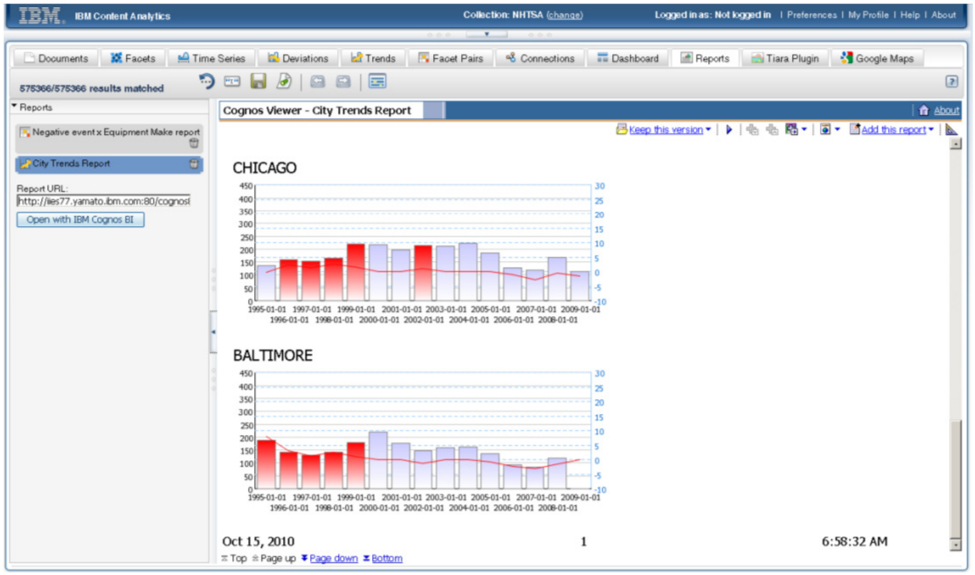
IBM Cognos Viewer - facets_report_1286251355

facets_report_1286251355

Open this report with the IBM Content Analytics text miner application.

Description:
User name:
View name: Facets
Facet name: Category
Query text:*:*
Sort By: Frequency

	Keyword	Frequency	Frequency	Correlation	Correlation
1	Package / container	311		1	
2	Number of pieces	107		1	
3	Contamination / tampering	82		1	
4	Taste / smell	72		1	
5	Ingredient	65		1	



575366/575366 results matched

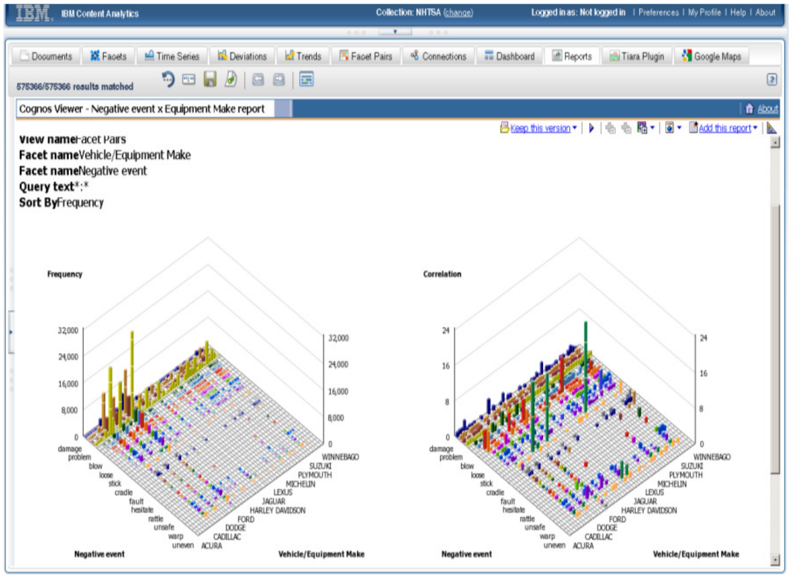
Reports
- Negative event x Equipment Make report
- City Trends Report

Report URL: http://ies77.yamato.ibm.com:80/cognos

Open with IBM Cognos BI

Cognos Viewer - City Trends Report

Oct 15, 2010 1 6:58:32 AM



575366/575366 results matched

Cognos Viewer - Negative event x Equipment Make report

View name Facet Pairs
Facet name Vehicle/Equipment Make
Facet name Negative event
Query text*:*
Sort By Frequency

IBM Content Analytics adds value to Big Data…

Healthcare Analytics

- ***Analyzing:*** *E-Medical records, hospital reports*
- ***For:*** *Clinical analysis; treatment protocol optimization*
- ***Benefits:*** *Better management of chronic diseases; optimized drug formularies; improved patient outcomes*

Customer Care

- ***Analyzing:*** *Call center logs, emails, online media*
- ***For:*** *Buyer Behavior, Churn prediction*
- ***Benefits:*** *Improve Customer satisfaction and retention, marketing campaigns, find new revenue opportunities*

Crime Analytics

- ***Analyzing:*** *Case files, police records, 911 calls…*
- ***For:*** *Rapid crime solving & crime trend analysis*
- ***Benefits:*** *Safer communities & optimized force deployment*

Insurance Fraud

- ***Analyzing:*** *Insurance claims*
- ***For:*** *Detecting Fraudulent activity & patterns*
- ***Benefits:*** *Reduced losses, faster detection, more efficient claims processes*

Automotive Quality Insight

- ***Analyzing:*** *Tech notes, call logs, online media*
- ***For:*** *Warranty Analysis, Quality Assurance*
- ***Benefits:*** *Reduce warranty costs, improve customer satisfaction, marketing campaigns*

Social Media for Marketing

Analyzing: *Call center notes, SharePoint, multiple content repositories*

For: *churn prediction, product/brand quality*

Benefits: *Improve consumer satisfaction, marketing campaigns, find new revenue opportunities or product/brand quality issues*

Interaction between Big Data and Content Analytics

- **IBM Content Analytics running on top of BigInsights for full analytics and rapid insights**

- **BigSheets and IBM Content Analytics – Filter with BigSheets then analyze in ICA for graphical view of analytic results**

- **IBM Content Analytics as a NLP engine to add define concepts and add depth to Cognos Consumer Insights**

Agenda

1 Big Data – Definitions & A Quick Story

2 Technology Primer

3 Dive on Content Analytics

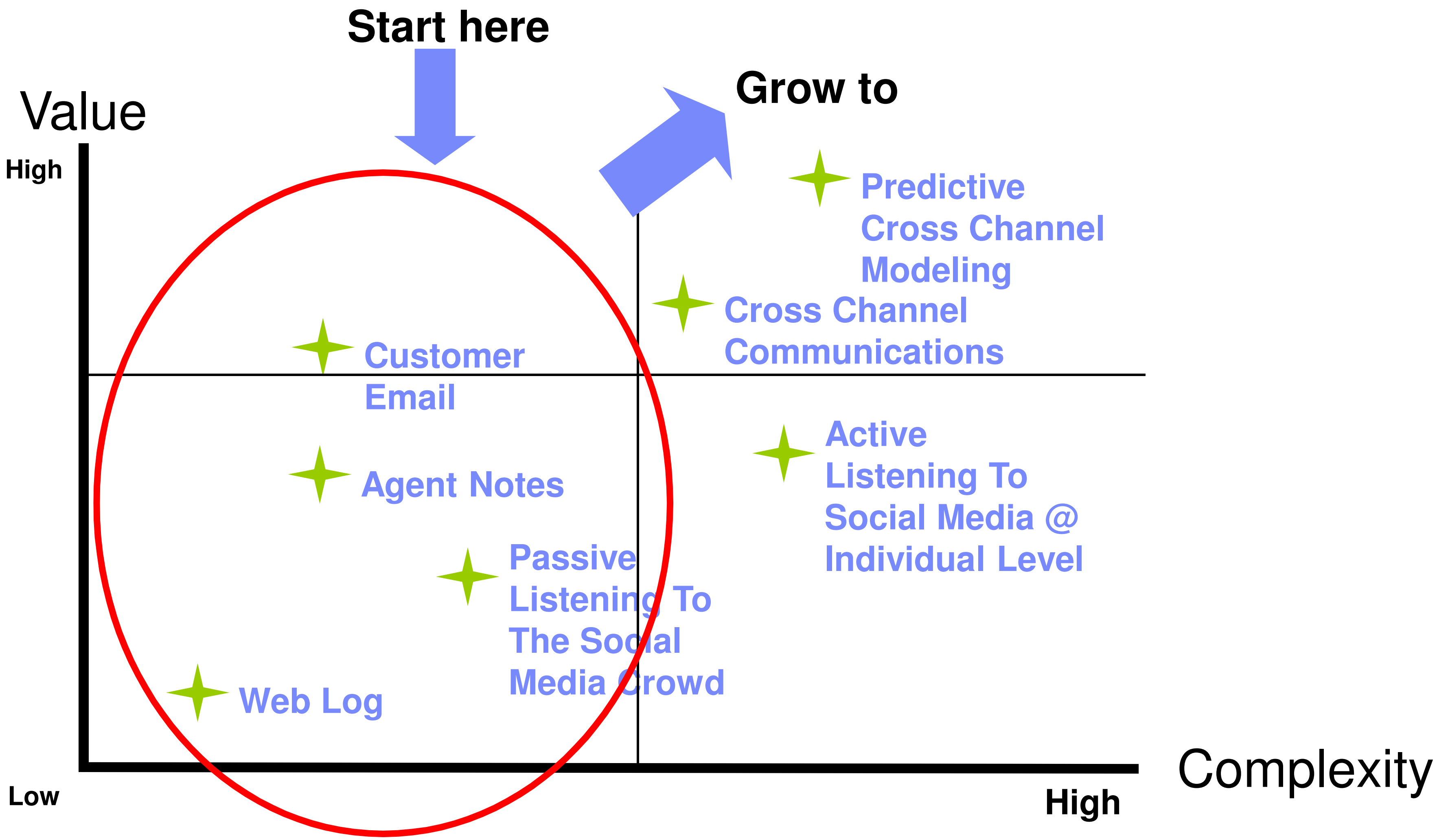
4 Thoughts On Getting Started

Thoughts On Getting Started

Use Case Selection Criteria

- Well proven path
- Can be done off-line/non-disruptive to existing systems
- Intuitive that there is low hanging fruit for additional insights
- Data set is already stored, but under-instrumented or overly summarized
- Initial findings can be arrived at 3 weeks or less
- Initial use cases have to be accretive to next set
- Initial use cases have to leverage common technology for next set of use cases

One Example Of How To Think About It


Start here
Grow to
Value
High
Low
Complexity
High
Predictive Cross Channel Modeling
Cross Channel Communications
Customer Email
Agent Notes
Active Listening To Social Media @ Individual Level
Passive Listening To The Social Media Crowd
Web Log

Use Case Starting Points

- **Mobile/Web Log**
  - Why: Reduce latency to under a day instead of weeks from 3rd parties
  - Why: Have the full data set for cross-correlation for other use cases
- **Agent Notes**
  - Why: Has proven to be an importance source of wisdom from the branches and Contact Centres
  - Why: An element of crowd sourcing
  - Why: Pulse of our internal teams whom we need to market to
- **Customer Email**
  - Why: They are often your best direct source of “pulse” of the business
  - Why: Lynchpin of crowd sourcing
  - Why: Six Sigma feedback component to cost take out
- **Passive Listening To The Social Media Crowd**
  - Why: Important early warning source of information
  - Why: You need the data – not summarization – for more advanced analytics
- **Cross Channel Communications**
- **Active Listening To Social Media @ Individual Level**
- **Predictive Cross Channel Modeling**

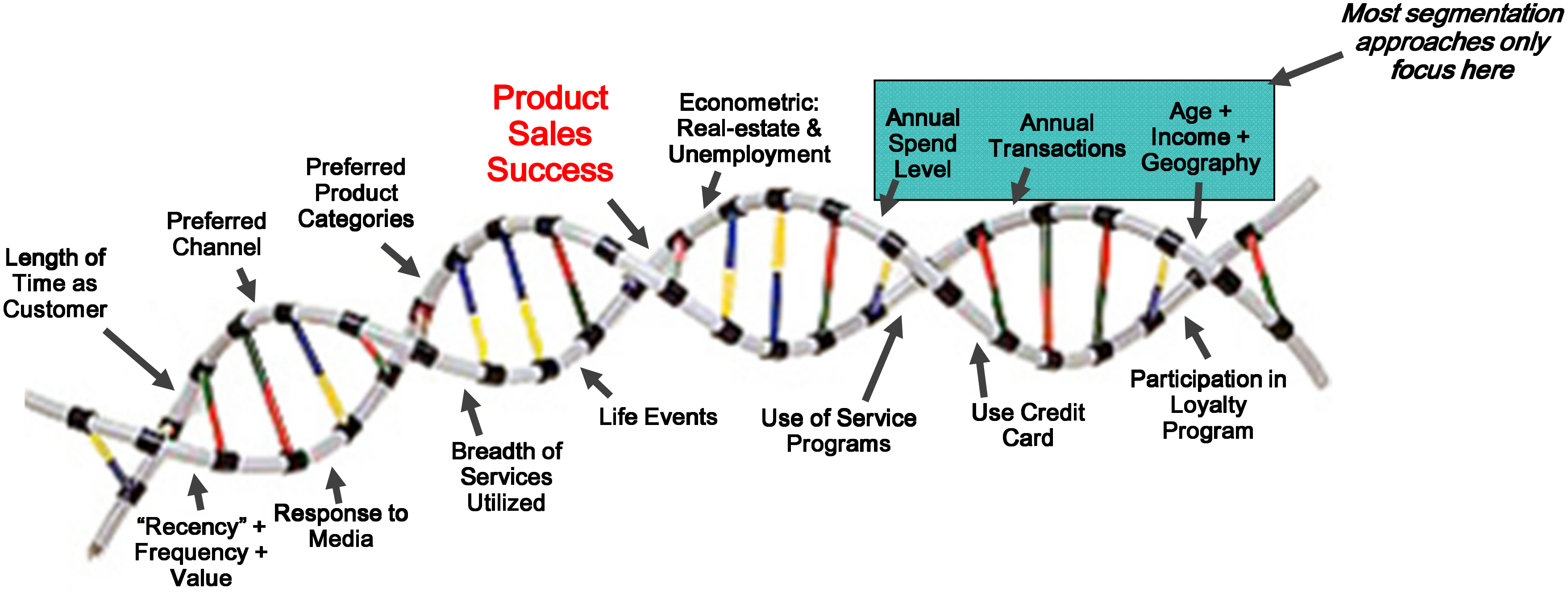
Examples Of Insight / Value Delivered

- **Web Log** *
  - Same day feedback on A|B testing new design (if sample is big enough)
  - Same day feedback on change in font/images/icons wording on behavior (if sample if big enough)
  - Input source into cross-channel behavior modeling
- **Agent Notes** *
  - Additional insight into customer “state” based on call handling
  - Feedback on features / policies not captured in formal system – nuance of customer opinion
  - Detecting of otherwise un-surfaced issues both at segment and individual customer level
  - Detection of staff dissatisfaction
  - Detection of otherwise hidden life events
- **Customer Email** *
  - Direct feedback on policy/fees
  - Notice of customers planning to attrition
  - Detecting of otherwise un-surfaced issues both at segment and individual customer level
  - (in private / brokerage situations early warning of law suits / advisor attrition)
  - Detection of otherwise hidden life events
  - Basis for cross-interaction behavioral / acceptance modeling
- **Passive Listening To The Social Media Crowd** *
  - Hour-by-hour feedback on announcements, offerings, policy changes
  - Taking the base reporting the rest of the org expects off the table so more important things can be focused on
  - Crowd feedback on competitor offerings
  - Path to SMARC

*** All are expanded “feature vectors”**

Case Study: Improvement Of Action Cluster Segmentation Approach

- Use of under-utilized data to expand modeled variables – “Feature Vectors” – The customers response to the firm’s value proposition
- Each feature vector is like a gene strand, which describes a facet, or set of customer behavior traits


Most segmentation approaches only focus here
Length of Time as Customer
Preferred Channel
Preferred Product Categories
Product Sales Success
Econometric: Real-estate & Unemployment
Annual Spend Level
Annual Transactions
Age + Income + Geography
“Recency” + Frequency + Value
Response to Media
Breadth of Services Utilized
Life Events
Use of Service Programs
Use Credit Card
Participation in Loyalty Program

Wait – How Do You Know If This Is Working?

- **Quick – how many of you were part of an experiment this week?**

- **Back testing**
- **A|B real world testing**
- **"Test Fast, Fail Fast, Correct Fast"**
- **Concept of MVP (minimum viable product)**

THINK

BIG