# Lab 04: Text Analytics

Hands-On Lab



# **Table of Contents**

1.	Introduction	3
2.	Objectives	3
3.	Importing input documents and dictionaries	4
	3.1 Creating the text analytics project in Eclipse	4
	3.2 Copying input files and dictionaries	8
4.	Extraction Tasks - Step 1: Select Documents	. 14
5.	Extraction Tasks - Step 2: Label examples and clues	. 15
	5.1 Part a. Label Snippets of Interest	. 16
	5.2 Part b. Label extraction clues	. 18
6.	Writing and Testing Extractors for Basic Features	22
	6.1 Extraction Tasks - Step 3: Develop the Extractor	22
	6.2 Extraction Tasks - Step 4: Test the Extractor	26
	6.3 Writing the basic extractors for Number and Revenue	27
7.	Writing Extractors for Concepts based on Basic Features	. 31
	7.1 Create Extractor for AmountwithUnit	32
	7.2 Create Extractor for RevenueByDivision	. 34
	7.3 Extending the extractor for RevenueByDivision by including AmountwithUnit information	. 36
8.	Analyzing Extracted Results with Annotation Explorer	. 38
	8.1 Displaying Output View Table	. 38
	8.2 Filtering Annotation Explorer Rows	. 39
	8.3 Exporting an Output View	41
	8.4 Mouse-Over function to explain the annotated text	42
9.	Summary	43

# 1. Introduction

InfoSphere<sup>™</sup> BigInsights comes with sophisticated text analytics capabilities that allow users to easily specify rules to extract actionable insights from large amounts of text.

Annotation Query Language (AQL) is a language for building these "rules sets" or extractors that extract structured information from unstructured or semi-structured text. AQL is the primary method of creating new extractors in the InfoSphere BigInsights text analytics system.

In this lab we will analyze the press releases for IBM's quarterly earnings from the year 2006 to 2010 to extract the names of IBM divisions and their respective revenues.

# 2. Objectives

After completing this hands-on lab, you'll be able to:

- Explore and use the text analytics tooling environment.
- Run information extractors over a sample data set.
- Explore how to view, understand, and debug results from the extractor.

The Lab is organized in six sections following the six tasks shown in Figure 1.

![](_page_2_Figure_11.jpeg)

![](_page_2_Figure_12.jpeg)

#### Section 1: Importing input documents and dictionaries

This section guides you through the steps for setting up the extraction project in Eclipse IDE, importing the press release input files in the proper directory, and importing the dictionary of IBM division names in its proper location.

#### Section 2: Label Text / Clues

This section guides you through the steps involved in developing the 'Extraction Plan'. In this step we identify two concepts of interest and several basic features.

- First Concept: Reference to the revenue of a division
- Second Concept: The reported revenue for that division

Basic features are components of the above concepts, for example integers and decimal numbers are components of reported revenue. The occurrence of division names is a component of the first concept.

#### Section 3: Develop and Test Extractor (AQL)

In this section you go through the steps for writing the extraction logic (AQL code) for basic features. You will write extractors for three basic features: *division names, numbers* and reference to the term 'revenue'.

#### Section 4: Develop and Test Extractor (AQL) - continued

This part of the lab guides you through the steps for building upon the basic features of section 3 to extract the 'revenue of division' and the 'reported revenue' concepts in a single extracted unit.

#### Section 5: Analyze and Export the Results

In section 5 of the lab you will explore the tools for viewing the results from above test extraction tasks. The four tools are for:

- 1. Viewing an 'Output view in tabular form'.
- 2. Setting up filters for rows displayed in Annotation Explorer. They are covered in sections 5a and 5b respectively
- 3. Exporting all Output Views as an HTML and CSV files.
- 4. Viewing details of highlighted text in annotated document using mouse-over capability.

# 3. Importing input documents and dictionaries

Become acquainted with the BigInsights Text Analytics perspectives in the Eclipse tool, and the 'Extraction Task' and 'Extraction Plan' views in it.

- Become familiar with the steps for setting up Extraction projects.
- Importing input files and dictionaries.
- Examining the properties of the project and input files.

## 3.1 Creating the text analytics project in Eclipse

- 1. Login on to Virtual Machine using the following:
  - a. Username = **biadmin**
  - b. Password = password
- 2. On Desktop, Double click on Eclipse. And select the default workspace location provided.

Page 4 of 44

![](_page_4_Picture_1.jpeg)

Figure 2 - Launch Eclipse

3. Open the BigInsights perspective by clicking on the **Open perspective** button (top right corner) as shown in the figure below, and clicking on **BigInsights** 

![](_page_4_Picture_4.jpeg)

Figure 3 - Open Perspective

4. From the *Task Launcher for Big Data* select the **Tasks - Develop** as shown below. Alternatively, you can click on the **Develop** tab, also highlighted in the figure below.

![](_page_4_Picture_7.jpeg)

Figure 4 – Select Develop Tasks view in Eclipse

#### IBM Software Information Management

5. After navigating to the **Develop** tab of the *Task Launcher for Big Data,* select '**Create a text extractor**' from the tasks listed.

Overview   Accelerate   Design   Develop	Publish and run Preference
Tasks Create a text extractor Follow a step-by-step wizard to create a text extractor ( including generating regular expressions), visualize the results of running the extractor, and evalute extractor quality. Discover patterns in text, view differences in the results of two extraction runs, and interpret the lineage of extractor results.	Quick Links © Open Project Explorer Switch to the BigInsights perspective Create a new BigInsights project Open the BigInsights console
Create a BigInsights program Create a Jaql script or module, a BigSheets function or reader, a BigInsights Java program, a Java MapReduce program, a SQL script, or a Pig file.	<ul> <li>Import a published or deployed application</li> <li>Launch a shell (pig, Jaql, or HBase)</li> </ul>
Create a configuration and run a BigInsights program Create a configuration to run a Jaql, Pig, Java MapReduce, or Java program.	Learn More AQL reference
Monitor BigInsights jobs running in a cluster Open the BigInsights console to monitor jobs running in the cluster.	🤌 Jaql reference

Figure 5 – Create a text extractor

6. Enter the project name '*MyFirstProject*' in the popup window and click **Finish**.

		×
BigInsights Create a ne	s Project aw BigInsights project.	
<u>P</u> roject nam	ne: MyFirstProject	
☑ Use <u>d</u> ef	fault location	
Location: /	/home/biadmin/IBM/rationalsdp/workspace/MyFirstProject	owse
	ihoose file system: [default ]≎_]	
?	< Back Next > Cancel	Einish

Figure 6 - Define the project name

Now you will be switched automatically to another perspective called BigInsights Text Analytics Workflow

![](_page_5_Picture_8.jpeg)

Figure 7 - BigInsights perspective button

InfoSphere BigInsights – Text Analytics © Copyright IBM Corp. 2013. All rights reserved

Page 6 of 44

# • Note: The <u>Text Analytics Workflow</u> perspective could also have been selected directly using the '*Open Perspective*' button on the top right corner of the IDE.

Two key views are shown from this perspective:

- Extraction Tasks view, shown on the top left, consistent with the steps outlined in Figure 8.
- Extraction Plan shown on the right.

Currently the plan shows the top level project name (MyFirstProject) and it is empty.

![](_page_6_Picture_6.jpeg)

Figure 8 - BigInsights Text Analytics Perspective

#### laceble Note: Extraction Tasks view can also be open using Window menu ightarrow Show View ightarrow Extraction Tasks

A project structure is also created automatically. See the structure in the **Package Explorer** view, which is close to the *Extraction Tasks* view. If you cannot find it, you can also use the Eclipse menu, *Window -> Show View > Package Explorer*.

![](_page_6_Picture_10.jpeg)

Figure 9 - The default package view in BigInsights Text Analytics perspective.

7. To verify the project properties right click MyFirstProject in the Package Explorer view and select Properties.

InfoSphere BigInsights – Text Analytics © Copyright IBM Corp. 2013. All rights reserved

Page 7 of 44

8. In the *Properties* dialog, expand **BigInsights** and then choose **Text Analytics**.

Within this window, you can specify the source for this project or add associated projects for your current project.

9. Click on **Cancel** to close the properties dialog.

0	Properties for MyFirstProject	×
type filter text 🔪	Text Analytics	
▶ Resource ▼ BigInsights	General	
Text Analytics	Source directory	
Builders Java Build Path	textAnalytics/src	Browse
▶ Java Code Style		
<ul> <li>Java Compiler</li> <li>Java Editor</li> <li>Javadoc Location</li> </ul>	Build output directory textAnalytics/bin	Browse
<ul> <li>Pig</li> <li>Project Facets</li> <li>Project References</li> </ul>		
Run/Debug Settings	Restore Defaults	Apply
?	Cancel	ок

Figure 10 - Validating properties of a BigInsights Text Analytics project

## 3.2 Copying input files and dictionaries

1. Create a new folder named **data** under project **MyFirstProject** in the Package Explorer view by right-clicking on the project name and selecting **New** → **Folder**.

![](_page_7_Picture_8.jpeg)

Figure 11 - Create new folder

InfoSphere BigInsights – Text Analytics © Copyright IBM Corp. 2013. All rights reserved

Page 8 of 44

2. Input data as the name of the folder and select Finish. We will store the input files here.

0	New Folder	×		
Folder				
Create a new folder resource.				
Enter or select the parent folder:				
MyFirstProject				
🕨 🚰 MyFirstProject				
🗁 RemoteSystemsTempFiles				
Folder <u>n</u> ame: data				
<u>A</u> dvanced >>				
0	Cancel	L		
	Cancel			

Figure 12 - Input folder name

3. Right-click on the data folder, and choose Import  $\rightarrow$  General  $\rightarrow$  File System.

🖬 Import	
Select Import resources from the local file system into an existing pr	oject.
<u>S</u> elect an import source:	
type filter text	•
<ul> <li>✓ ➢ General</li> <li></li></ul>	E
<ul> <li>Preferences</li> <li>BigInsights</li> </ul>	~
<	Einish

Figure 13 – Select files to import from File System

InfoSphere BigInsights – Text Analytics © Copyright IBM Corp. 2013. All rights reserved

Page 9 of 44

- 4. Click Browse to navigate to /home/biadmin/bootcamp/input/lab04\_TextAnalytics folder and click OK.
- 5. Back in the Import dialog, expand the **lab04\_TextAnalytics** tab in the left pane and select **IBMQuarterlyReports** folder. It will auto-select all text files within the folder. Click **Finish**.

C Import 2	×			
File system     Import resources from the local file system.				
From directory:       /home/biadmin/bootcamp/input/lab04_TextAnalytics <ul> <li>Bgowse</li> <li>Bgowse</li> </ul> Image: State of the sta				
Filter Types     Select All        Filter Types     Select All        Into folder:        MyFirstProject/data				
Cancel Einish	]			

Figure 14 - Importing IBMQuarterlyReports into your project.

After the import, the IBMQuarterlyReports folder containing the actual reports should appear under the *data* folder as shown below.

![](_page_10_Picture_1.jpeg)

Figure 15 - Input files copied to the data folder.

- 6. Right-click MyFirstProject in Package Explorer and select Properties.
- 7. In the *Properties* dialog, ensure that the Text File Encoding is set to UTF-8 by selecting **Resource** → **Text file** encoding → Other → UTF-8. Click OK to close the *Properties* dialog.

٢	Properties for MyFirstProject	×
type filter text 🛛 🍗	Resource	↓ ↓ ↓
Resource     JAQL     Text Analytics     Builders     Java Build Path     Java Code Style     Java Code Style     Java Code Style     Javadoc Location     Pig     Project Facets     Project Facets     Project References	Path:       /MyFirstProject         Lype:       Project         Location:       /home/bladmin/workspace/MyFirstProject         Last modified:       July 2, 2013 11:22:09 AM         Text file encoding	
Kunnessug Setungs		Restore Defaults Apply
?		Cancel

Figure 16 - Select UTF-8 encoding

8. Repeat step 3 to import files into the MyFirstProject/textAnalytics/src folder. Right click the src folder in your project, and select Import.

Navigate to */home/biadmin/bootcamp/input/lab04\_TextAnalytics/* and check the dictionaries folder to select both the folder and its content. Verify that the boxes next to the dictionaries folder and division.dict file have check marks. Click **Finish**.

InfoSphere BigInsights – Text Analytics © Copyright IBM Corp. 2013. All rights reserved

0	lmp	ort		×	
File system Import resources from the local fi	le system.				
From directory: /home/biadmin.	/bootcamp/input/la	b04_TextAnalytic	s 🗸	B <u>r</u> owse	
Filter Types       Select All       Deselect All         Into folder       MyFirstProject/textAnalytics/src       Browse         Options					
?	< <u>B</u> ack	<u>N</u> ext >	Cancel	Einish	

Figure 17 – Dictionaries copied to the src folder

9. Finally, import **solution.aql** into a newly created folder within **MyFirstProject**, name the folder **aql**, this file contains the solution for this lab and might be useful for reference.

Right click folder aql, select Import. Browse to /home/biadmin/bootcamp/input/lab04\_TextAnalytics, click OK, In the Import dialog, select lab04\_TextAnalytics in the left pane (do not check the checkbox), then check solution.aql in the right pane. Verify that the box next to the lab04\_TextAnalytics folder has a horizontal line, and the box next to the solution.aql file has a checkmark. Click Finish.

0	lmp	ort		×
File system Import resources from the local file sys	item.			
From directory: //home/biadmin/bootc	amp/input/Ial	b04_TextAnalytic	s 🗸	B <u>r</u> owse
<ul> <li>Iab04_TextAnalytics</li> <li>I Implicationaries</li> <li>Implicational Implication</li> <li>Implication Implication</li> <li>Implication</li> <li>Implication</li></ul>		<ul> <li>□ → RevenueB</li> <li>✓ → solution.a</li> </ul>	yטועונוסn. aqı ql	
Filter Types) Select All De Into foIder: MyFirstProject/aql	eselect All			Bro <u>w</u> se
Options- Overwrite existing resources without warning Create complete folder structure Advanced >>				
?	< <u>B</u> ack	<u>N</u> ext >	Cancel	Einish

Figure 18 – Importing file solution.aql into your project

10. Your Package Explorer view should look like the following:

![](_page_13_Figure_1.jpeg)

#### Figure 19 – Project structure after importing required files

Now that we have created a project and imported the documents that require textual analysis, we will begin the text extraction process. The following sections outline the process of creating and running text extractors on a set of documents. We will cover the following steps (these are also outlined within the extraction tasks tab within the BigInsights perspective).

- Step 1: Select Documents
- Step 2: Label Examples and Clues
- Step 3: Develop the Extractor
- Step 4: Test the Extractor

# 4. Extraction Tasks - Step 1: Select Documents

- 1. Return to the Extraction Tasks view, and select the first step: Step 1: Select Documents → part a. Select Document Collection
- 2. Click Browse Workspace and select the folder /MyFirstProject/data/IBMQuarterlyReports. Click OK.

9	🖬 Choose your data 🛛 🗙
<u>Fi</u> le <u>E</u> dit <u>N</u> avigate Se <u>a</u> rch <u>P</u> roject Da <u>t</u> a <u>B</u> un <u>W</u> i	Select a file or directory
] [™ 🔛 🧐 🖆 ] 🐨 Task Navigator → ] 💒 ] 1	▼ 😂 MyFirstProject
Project Explor 🖀 Extraction Tas 🛛 🛤 Package B	🕨 🗁 .settings
▼ Step 1 : Select Documents	Þ 🗁 aql
a. Select Document Collection	😕 bin
Select a collection in one of the supported input coll	BMQuarterlyReports
/MyFirstProject/data/IBMQuarterlyReports	
Browse Workspace	> > > textAnalytics
Language en 😂	
b. Select the documents to work with	(?) Cancel OK

Figure 20 – Selecting input documents to be analyzed.

- 3. After clicking OK, the path will be added, and the documents in this folder will appear below the window. Select **en** for English from the **Language** drop-down menu.
- 4. In Step 1b of the Extraction Tasks view, click on /4Q2006 and click Open. The content of the document will open in the editor area.

🔁 Project Explorer 😫 Package Explore 🖺 BigInsights Serv 🎯 Extraction Tasks 🕴 🗸 🖱 🗖	🖄 Task Launcher for Big Data 🛛 🕹 /4Q2006.txt 🛛 🗧 🗖
Step 1 : Select Documents	IBM Reports 2006 Fourth-Quarter Results
a. Select Document Collection Select a collection in one of the supported input collection formats. [/myFirstProject/data/IBMQuarterlyReports Browse Workspace Clear Language en	Tab navigation Press release Related XML feeds Contact(s) information Related resources 
b. Select the documents to work with  D /4Q2006.bt  M /4Q2006.bt	Total revenues of \$26.3 billion, up 7 percent as reported; Diluted earnings of \$2.26 per share from continuing operations, u Services signings of \$17.8 billion, up 55 percent. IBM today announced fourth-quarter 2006 diluted earnings of \$2.2
i /4Q2008.txt	Fourth-quarter income from continuing operations was \$3.5 billion
0 /4Q2019.txt 0 /4Q2010.txt	Samuel J. Palmisano, IBM chairman, president and chief executive From a geographic perspective, the Americas fourth-quarter reven
Open See Example	Revenues from the Software segment were \$5.6 billion, an increase For the WebSphere family of software products, which facilitate v

Figure 21 – Selecting documents to be labeled

# 5. Extraction Tasks - Step 2: Label examples and clues

The objective of this section is to become acquainted with the process of creating an extraction plan by labeling snippets of interest and the clues contained in them. For example "*Revenues from Software were* \$3.9 *billion*" is a **snippet of interest**. Names of division ("*Software*") and revenue numbers ("3.9") are **clues within the snippets**.

Observe the automatic creation of an Extraction Plan and aql source folders as snippets of interest and clues are identified.

InfoSphere BigInsights – Text Analytics © Copyright IBM Corp. 2013. All rights reserved

Page 15 of 44

## 5.1 Part a. Label Snippets of Interest

1. In *step 2 part a. Label snippets of Interest* of the Extraction Task view; our goal is to extract the revenue generated by each business division from each of the quarterly reports. We start by labeling example snippets of revenue by division.

![](_page_15_Picture_3.jpeg)

Figure 22 – Labeling documents to identify snippets of interest.

2. In the /4Q2006.txt document that is open in the editor, highlight the portion of text that says "revenues from Global Technology Services increased 7 percent (...) to \$8.6 billion". (You may need to scroll to the right to find this text, or press CTRL+F to open a Search window where you can type the above text to search.) Right click the highlighted text, and select Add Example with New Label.

![](_page_15_Picture_6.jpeg)

Figure 23 – Selecting text snippets of interest

3. In the Add New Label dialog as shown below, type RevenueByDivision in the Label Name field and click Finish.

0	Add New Label	×
Add New Lab	el	
Label Name: Parent Label:	RevenueByDivision type filter text	<u> </u>
	- Leave parent empty to create a top-level Label - Double-click a label to select it as parent label	
?	Cancel	Einish

Figure 24 – Add New label

4. The **Extraction Plan** view is populated with the new label and example:

![](_page_16_Picture_4.jpeg)

Figure 25 – The extraction plan view in BigInsights text analytics perspective

 Go back to the /4Q2006.txt document in the editor and select one more snippet as an example to be labeled. Highlight the portion of text "*Revenues from the Systems and technology Group (S&TG) segment totaled* \$7.1 *billion*". Right click and select Label Example As → RevenueByDivision.

● Note the Label Example As option only appears after the first label, RevenueByDivision, (in this example), was created.

	From a geographic perspective, the Americas fourth-quarter revenues were	e \$11.1 billion, an increase o	of 6 percen <sup>.</sup>	a. Lab					
	Revenues from the Software segment were \$5.6 billion, an increase of 14 percent (11 percent, adjusting for curr								
	For the WebSphere family of software products, which facilitate customers' ability to manage a wide variety of I								
	For the Global Services business, segment revenues from Global Technology Services increased 7 percent (4 perce								
	Revenues from the Systems and Technology Group (S&TG) segment totaled \$7	1 billion for the quarter 1	In 3 nercen	0-1					
	🖻 Label Example As 🔰 🖏 RevenueE								
	Global Financing segment revenues increased 3 percent (flat, adjusting f	Add Example with New Label		extr					
	The company's total gross profit margin was 44.6 percent in the 2006 fou		. percent : 🗹						
l	< m	🔚 Add to Snippets	>	b. Lak					
ſ	🕈 Problems 🕱 🛛 📮 Console 🗮 Annotation Explorer	Input Methods >	~ - 8	Fror					
	) items			lexi					

Figure 26 – Adding additional examples to previously created label of a text snippet

6. The new example appears in the *Extraction Plan* view.

![](_page_17_Picture_4.jpeg)

Figure 27 – New label example appears in Extraction Plan

## 5.2 Part b. Label extraction clues

In this step we label *clues* within the snippets we selected as examples. These clues will help extract the information we need.

![](_page_18_Picture_1.jpeg)

Figure 28 – Labeling clues in snippets of interest

 Examine the examples labeled as *RevenueByDivision*. In the Extraction Plan view, click on the first example under RevenueByDivision → Examples. You will see the instances highlighted in yellow in the text. Repeat this process for the other example so you see how the other line is highlighted.

🖄 Task Launcher for Big Data	& /4Q2006.txt 13	- 0	🖉 Extraction Plan 🛛 🗖 🗖
Services signings of \$17 IBM today announced four	7.8 billion, up 55 percent. rth-quarter 2006 diluted earnings of \$2.26 per share from continuing operati	ons,	Image: Constraint of the state         Image:
Fourth-quarter income fr	rom continuing operations was \$3.5 billion compared with \$3.2 billion in the	fou	▼ 🖏 RevenueByDivision
Samuel J. Palmisano, IBM	M chairman, president and chief executive officer, said: "IBM had a terrific	qua	Examples revenues from Global Tech6 billion
From a geographic perspe	ective, the Americas fourth-quarter revenues were \$11.1 billion, an increase	of	Revenues from the Systems1 billion
Revenues from the Softwa	are segment were \$5.6 billion, an increase of 14 percent (11 percent, adjust	ing	
For the WebSphere family	y of software products, which facilitate customers' ability to manage a wide	var	
For the Global Services	business, segment <mark>revenues from Global Technology Services increased 7 perc</mark>	ent	
Revenues from the Syster	ms and Technology Group (S&TG) segment totaled \$7.1 billion for the quarter,	up 🗕	
Global Financing segment	t revenues increased 3 percent (flat, adjusting for currency) in the fourth	quar	
The company's total gros	ss profit margin was 44.6 percent in the 2006 fourth quarter compared with 4	4.1	
Total expense and other	income increased 11 percent to $6.9$ billion compared with the prior-year pe	riod	
IBM's effective tax rate	e in the fourth-quarter 2006 was 28.0 percent compared with 29.5 percent in	the	
For total operations, ne	et income for the fourth-quarter 2006 was \$3.5 billion, or \$2.31 per diluted	sha	
Share repurchases totale	ed approximately \$1.4 billion in the fourth quarter. The weighted-average nu	mber	
Full-Year 2006 Results		~	
< III		>	

Figure 29 – Reviewing labeled snippets of interest

2. Examine these snippets of text. What types of clues do you think would be useful for extraction?

revenues from Global Technology Services increased 7 percent (...) to \$8.6 billion

#### Revenues from the Systems and Technology Group (S&TG) segment totaled \$7.1 billion

Here are a few clues that are useful for this example, and what we are trying to extract:

- The word "revenues"
- Division names such as Global technology Services, Systems and Technology Group (S&TG)
- Amounts such as \$8.6 billion, \$7.1 billion

We record these clues as examples of **labels** in the *Extraction Plan*. The process of recording is very similar to recording full examples. Let's record the clue for "*revenues*" first.

3. In the /4Q2006.txt file, select "revenues" in "revenues from Global Technology Services ..." right-click and select Add Example with New Label.

🖄 Task Launcher for Big Data 🛛 💫 /4Q2006.txt 🕴	- 6
Services signings of \$17.8 billion, up 55 percent. IBM today announced fourth-quarter 2006 diluted earnings of \$2.26 per share from continuing	operations,
Fourth-quarter income from continuing operations was \$3.5 billion compared with \$3.2 billion	n in the fou
Samuel J. Palmisano, IBM chairman, president and chief executive officer, said: "IBM had a $\cdot$	terrific qua
From a geographic perspective, the Americas fourth-quarter revenues were \$11.1 billion, an :	increase of
Revenues from the Software segment were \$5.6 billion, an increase of 14 percent (11 percent	, adjusting
For the WebSphere family of software products, which facilitate customers' ability to manage	e a wide var
For the Global Services business, segment revenues from Global Technology Services increase	d 7 percent
Revenues from the Systems and Technology Group ( Add Example As )	quarter, up
Add to Snippets	Tourin quar
The company's total gross profit margin was 44.6uarter compared	d with 44.1
Total expense and other income increased 11 perceInput Methods	-year period
IBM's effective tax rate in the fourth-quarter 2006 was 28.0 percent compared with 29.5 perc	cent in the
For total operations, net income for the fourth-quarter 2006 was \$3.5 billion, or \$2.31 per	diluted sha
Share repurchases totaled approximately \$1.4 billion in the fourth quarter. The weighted-ave	erage number
Full-Year 2006 Results	~

Figure 30 – Identifying clues in the labeled snippets of interest

4. Fill in *Revenue* in the *Label Name* field and double-click on *RevenueByDivision* to add it to the *Parent Label* field. As shown in the figure below.

dd New Lab	Add New Label	
Label Name:	Revenue	
Parent Label:	RevenueByDivision	8
	🚯 RevenueByDivision	
	- Leave parent empty to create a top-level Label - Double-click a label to select it as parent label	
?	Cancel	h

Figure 31 – Adding clues to the extraction plan.

5. The new clue is recorded in the Extraction Plan view, under RevenueByDivision > Labels

![](_page_20_Picture_4.jpeg)

Figure 32 – View of the extraction plan after clue (label) is added

- 6. Repeat this process to label clues for two other 'Basic Features': *Amount* and *Division*. You would have to create *two label/clue examples for Amount*, and *two for Division*:
  - i) Highlight *Global Technology Services* and right click on it  $\rightarrow$  select **Add Example with New Label**  $\rightarrow$  enter **Division** for Label Name, and **RevenueByDivision** for Parent Label  $\rightarrow$  click **Finish**.
  - ii) Highlight Systems and Technology Group (ST&G) and right click on it  $\rightarrow$  select Label Example As  $\rightarrow$  Division.
  - iii) Highlight **\$8.6 billion** and right click on it  $\rightarrow$  select **Add Example with New Label**  $\rightarrow$  enter **Amount** for Label Name, and **RevenueByDivision** for Parent Label  $\rightarrow$  click **Finish**.
  - iv) Highlight \$7.1 billion and right click on it  $\rightarrow$  select Label Example As  $\rightarrow$  Amount.

Upon completing this step, you should have the two labels, *Division* and *Amount*, with two examples each under *RevenueByDivision*, as shown below.

![](_page_21_Picture_1.jpeg)

Figure 33 – The Extraction Plan

# 6. Writing and Testing Extractors for Basic Features

In this section we will write extractors for three basic features: division names, numbers and reference to the term 'revenue'.

- You will get familiar with two tools for extracting the basic features:
  - Use of regular expressions in the aql Extract statement. We will locate the numeric value of the revenue of a division, and occurrence of the clue 'Revenue' in a statement using regular expressions.
  - Use of dictionaries in the Extract statement, both inline dictionaries and dictionary files. We will use a
    dictionary to identify names of the IBM divisions

## 6.1 Extraction Tasks - Step 3: Develop the Extractor

In Step 3 of the *Extraction Task* view, we will create the extractor for 'division' using the dictionary *division.dict*, which we copied into the AQL folder during setup.

![](_page_22_Picture_1.jpeg)

![](_page_22_Figure_2.jpeg)

 Add a rule to extract division names. In the Extraction Plan view, navigate to RevenueByDivision → Labels → Division. Right-click on the Division label and select New AQL Statement, and choose Basic Feature AQL Statement.

![](_page_22_Figure_4.jpeg)

Figure 35 – Selecting the extractor to be written for basic features of the labeled clues in the extraction plan

In the *Create AQL Statement* dialog, use *Division* as the view name, and then for the AQL statement type, choose *Dictionary* since we have the dictionary *division.dict* that we loaded earlier in the lab and specify the aql script name as Division\_basic (this names the aql file containing the extractor code). Select Output view, to be able to see results. Click OK.

InfoSphere BigInsights – Text Analytics © Copyright IBM Corp. 2013. All rights reserved

Page 23 of 44

			2	Division	ew Name:	ew
on_BasicFeatures	icF	ion_E	eByDivi	Revenu	2L Module:	QL
			basic	Division	≬L script:	QL
			ary	Diction	pe:	/pe
			ut view rt view	Outp		
			ut view rt view	Outp		

Figure 36 – Specifying the extractor name and extractor type

The file *RevenueByDivision/Division\_basic.aql* opens in the editor and a template for the *extract dictionary* AQL statement is automatically added (do not worry about the parse error, you will fix that shortly):

Î	Task Launcher for Big Data	💫 /4Q2006.txt	🔂 Division_basic.aql 🛛
	<b>module</b> RevenueByDivisio	on_BasicFeature	5;
0	create dictionary Divis from file ' <path to="" you<br="">with language as 'en';</path>	sionDict ur dictionary he	ere>'
	create view Division as	5 	
~	extract dictionary Div	/isionDict	
w	on R. <1nput column:	> as match	
	from <input view=""/> R;		
0	output view Division;		

![](_page_23_Figure_5.jpeg)

A few things to notice about the template:

- It contains a CREATE DICTIONARY statement defining a dictionary DivisionDict from an external file you will need to fill in the template for *<path to your dictionary here>* with the path to the dictionary file in your project.
- The language used for matching the dictionary is 'en' (for English) (the same language was specified in Step 1.a of the Extraction Tasks view)
- It contains a template for an *extract dictionary* statement for matching the previously defined *DivisionDict* dictionary. You will need to fill in the templates for *<input column>* and *<input view>*. When dealing with text documents from the file system we use the input view keyword '*Document*' to reference all input documents and the input column keyword '*text*' to reference the text of the input documents. *Document.text* is the correct way to refer to the content of any input document.

The corrected template should therefore look like:

Note: The file path reference stated above has to match the path to division.dict and is relative to the root of the module in which the create dictionary statement is issued.

Dictionaries are usually stored in files to keep them separate from the aql code. The dictionary for 'Division' is shown below.

![](_page_24_Picture_5.jpeg)

![](_page_24_Figure_6.jpeg)

3. Save the aql files by selecting **File**  $\rightarrow$  **Save** from the top menu bar in Eclipse, or press or ctrl+s on your keyboard.

## 6.2 Extraction Tasks - Step 4: Test the Extractor

The results of the extraction are displayed in the Annotation Explorer as shown below.

1. Now that we have created an extraction rule for one basic feature - 'Division', it is time to test it.

Go to Step 4: Test the Extractor in the Extraction Task view. Choose the last option, **Run the extraction plan on the set** of documents that are labeled, as shown in the figure below. Click on the **Step 1** icon beside it.

The results of the extraction are displayed in the *Annotation Explorer* view as shown below.

🔁 Proj 🔰 Pac 🖺 Bigl 🎬 Extr 🕱 🖛	' 🗆	🖄 Task Launch	er for Big Data 🛛 💫 /4Q2006	.txt 🔰 Division_basic.aql 🛿 🔋 divi	sion.dict	- 8
<ul> <li>Step 4 : Test the Extractor</li> <li>a. Run your extractor</li> <li>You can run on the entire data collection that you specified in Step 1, only run on</li> </ul>		module Re create di	venueByDivision_BasicFe .ctionary DivisionDict	aatures;		
the documents that you selected in Step 1.b, or only run on the documents that you labeled in Step 2. Run the extraction plan on the entire data collection		from file with lang create vi extract d on R.	<pre>: '/dictionaries/divis uage as 'en'; .ew Division as lictionary 'DivisionDic1 text as match upper D</pre>	sion.dict'		
Run the extraction plan on the set of selected documents Run the extraction plan on the set of decuments that are labeled		output vi	ment K; ew Division;			
The extraction results are displayed in the Annotation Explorer.		Problems	Console 📰 Annotation Explo	prer Σ Showing page 1 of 1	4 🕨 🏢	. ⊪× \$× 4 2 ° 0
b. Identify mistakes in the extracted		Input Documen	t Left Context	Span Attribute Value	Right Context	Span Attribute Name
The result of your extractor might contain		4Q2006.txt	and segment revenues from	Global Business Services [4213-4237]	increased 6 percent (3 percent, a	RevenueByDivision_Basic
mistakes: false positives (results		4Q2006.txt	Transformation Outsourcing,	Global Business Services [4525-4549]	, Integrated Technology Services	RevenueByDivision_Basic
incorrectly identified by the extractor) and		4Q2006.txt	rith 2005. Revenues from the	Global Business Services [10192-10216]	segment were \$16.0 billion, flat (	RevenueByDivision_Basic
not identified by the extractor)		4Q2006.txt	prage increased 9 percent.	Global Financing [5371-5387]	segment revenues increased 3 p	RevenueByDivision_Basic
Use the Provenance Viewer to find out		4Q2006.txt	on, up \$2.2 billion, excluding	Global Financing [7927-7943]	receivables. Diluted earnings pe	RevenueByDivision_Basic
the rules that cause false positives.		4Q2006.txt	cent, adjusting for currency).	Global Financing [10394-10410]	revenues totaled \$2.4 billion, a d	RevenueByDivision_Basic
Go back to Step 2 to identify additional		4Q2006.txt	g the year-to-year change in	Global Financing [11077-11093]	receivables, was \$15.3 billion - a	RevenueByDivision_Basic
		-				
clues, or go back to Step 3 to create new statements or fix existing statements		4Q2006.txt	itstanding. Debt, including	Global Financing [12507-12523]	, totaled \$22.7 billion, compared v	RevenueByDivision_Basic
clues, or go back to Step 3 to create new statements or fix existing statements.		4Q2006.txt 4Q2006.txt	utstanding. Debt, including cent at the end of 2006, and	Global Financing [12507-12523] Global Financing [12720-12736]	, totaled \$22.7 billion, compared debt increased \$1.8 billion from y	RevenueByDivision_Basic RevenueByDivision_Basic

Figure 39 – Fourth step of the task flow, testing the extractor and the Annotation Explorer view

2. The Span Attribute column in the middle of Annotation Explorer shows the *division* names picked up by the extractor. You will notice towards the bottom in the explorer view that "*software*" is being picked up incorrectly as a division name. This can be fixed by modifying the Extract Dictionary statement to use the '*Exact*' flag to ensure that the text string matches the dictionary entry exactly, including case.

Modify the create view statement for Division\_basic.aql as shown below, *save the new aql* and *run the extraction plan again* in the same way as before.

![](_page_25_Figure_10.jpeg)

3. **Double-click** on one of the rows for **Systems and Technology Group** in **Annotation Explorer** to see the position of the extracted text in the document. It is highlighted in blue in the edit area. An annotation tree view pops up on the

InfoSphere BigInsights – Text Analytics © Copyright IBM Corp. 2013. All rights reserved

Page 26 of 44

right side. Check the *division annotator box*, and all division entities extracted from the document will be highlighted in the document as shown in the figure below.

Î	Task Launche	er for Big Data	💫 /4Q2006.t	xt 🎦 Division_basic.aql	🗐 4Q2006.txt - Document.tex	t 🕱	- 0	🖉 Extractio 📴 4Q2006.t 🕱 🗖 🗖
	From a geo	ographic persp	pective, th	ne Americas fourth-quart	er revenues were \$11.1	billion, an increase	of	〕 〕 ↑ ↓ ↓ 수 수 2 2
i	Revenues -	from the <mark>Soft</mark>	<mark>ware</mark> segmer	nt were \$5.6 billion, an	increase of 14 percent	(11 percent, adjust	inç	type filter text
	For the W	abenhana famil	ly of cofts	and products which fac	ilitate customers, shil	ity to manage a vide		
	For the we	ebophere railit	Ly OF SOLL	vare products, which had	Ititate customers abit	ity to manage a wide	V c	✓ □ RevenueByDivision_BasicFeature:
i	For the <mark>G</mark>	lobal Services	s business,	segment revenues from	Global Technology Servi	<mark>ces</mark> increased 7 perc	en1	✓ ✓ match (SPAN)
	Povopuos	from the Svet	ome and Too	theology Group (SETG) co	amont totaled #7 1 bill	ion for the quarter	1.07	✓ Software [2595-2603]
1	Revenues	From the Syste	enis and rec	nnotogy Group (Sard) se	gment totated \$7.1 bitt	ion for the quarter,	ut	✓ Global Services [4039-4054]
i	Global Fin	nancing segmer	nt revenues	s increased 3 percent (f	lat, adjusting for curr	ency) in the fourth	qua	Global Technology Services
	<b>The energy</b>				in the poor faunth and	ماطني ليحتجم محمد	=	✓ Global Business Services [4
	The compar	ny s totat gro	uss prorite	margin was 44.0 percent	In the 2006 rounth qua	rter compared with 4	4	✓ Global Business Services [4
	Total exp	ense and other	r income ir	ncreased ll percent to \$	6.9 billion compared wi	th the prior-year pe	ric	Systems and Technology Gr
	TOMIN off	antiva tau an	to in the d	sunth sugator 2005 use	20. 0 percent compand	ith 20 E paraget in	the lat	✓ Global Financing [5371-5387
	TBM.S ELL	ective tax ra	te in the i	ourth-quarter 2006 was	28.0 percent compared w	ith 29.5 percent in		✓ Global Financing [7927-7943]
	For total	operations, r	net income	for the fourth-quarter	2006 was \$3.5 billion,	or \$2.31 per diluted	sł	✓ Software [9898-9906]
	Share repu	urchases total	led approxi	mately \$1.4 billion in	the fourth quarter. The	weighted-average nu	mbe	Global Technology Services
		_		-				✓ Global Business Services [1
	Full-Year	2006 Results						✓ S&TG segment [10290-1030
	Total reve	enue of \$91.4	billion, u	up 4 percent excluding t	he divested PC business	;		✓ Global Financing [10394-104
	Income fro	om continuing	operations	of \$9.4 billion, up 18	percent as reported, o	r up 9 percent exclu	dir	✓ Global Financing [11077-110
	Diluted ea	arnıngs of \$6. from operation	.06 per sha	are from continuing oper billion up #2 2 billi	ations, up 23 percent a	s reported, or up 14	p€	✓ Global Financing [12507-125]
1	Diluted ea	arnings per sh	nare from o	continuing operations we	re \$6.06 compared with	\$4.91 per diluted sh	are	Global Einancing [12720-127
								Global Einancing [15082-150
_								
	Problems 🖳	Console 🏦 Ann	otation Explor	er 🛙	4	▷ 📗 IBY 🐎 Y 🖒 🖁		
Tex	xt analytics res	sult, Number of rov	vs: 17/17	Showing page 1 of 1				
In	put Document	L	eft Context	Span Attribu	ite Value	Right Context Span A	ttribu ^	
4	Q2006.txt	irrency). Revenue	s from the	Global Technology Sen	vices [10035-10061]	segment totaled § Revenu	eByD	
4	Q2006.txt	ss, segment reve	nues from	Global Technology Se	rvices [4087-4113]	increased 7 perce Revenu	eByD	
4	Q2006.txt	ent, adjusting for	currency).	S&TG segment [	10290-10302]	revenues were \$2 Revenu	eByD	
4	Q2006.txt	cent compared wit	th 2005.	Software [98	<del>co coco</del> ]	segment revenue: Revenu	eByD	
4	Q2006.txt	Jarter. Revenue	s from the	Software [25	95-2603]	segment were \$5 Revenu	eByD 🗧	
4	Q2006.txt	eriod. Revenue	s from tile	Systems and Technology Group	(S&TG) segment [4691-4734]	totaled \$7.1 billior Revenu	eByD 🧹	
<							>	< III >>

Figure 40 – Visual examination of the extracted basic features (Division names)

4. Back in the Extraction Tasks view, click the icon beside step 4.a "Run the extraction plan on the entire data collection". This will display the results extracted from all the documents in your collection (Step 1.a of the Extraction Tasks), not just the document 4Q2006.txt that you have labeled.

![](_page_26_Picture_5.jpeg)

The button "*Run the extraction plan on the entire data collection*" of Step 4.a of Extraction Tasks performs the same task as if you would configure a Text Analytics run configuration manually (in Package Explorer, right click on project and select Run Configurations. In the wizard, create a new Text Analytics configuration and specify the input data collection as the collection selected in Step 1.a of Extraction Tasks view.)

# 6.3 Writing the basic extractors for Number and Revenue

We will now create extractors for two more *basic features* – Number and Revenue. We are looking for numbers - 12, 12.5, 27.2. Later we will use this basic feature in a *pattern* to identify instances of an amount with a unit - \$1.2 billion or \$12.5 million. This is a powerful feature of AQL that lets us identify basic features based on a regular expression or a dictionary and then use a pattern to put the basic features into context.

InfoSphere BigInsights – Text Analytics © Copyright IBM Corp. 2013. All rights reserved

Page 27 of 44

D Note: To refer to the complete code for these extractors, you can refer to the solution provided in the solution.aql file within the aql folder.

- I. Regular expression rule to identify 'Number'
- 1. In the Extraction Plan view, navigate to *RevenueByDivision* → *Labels* → *Amount*. Right-click the label Amount and select *New AQL Statement*. Select Basic Features AQL statement.

![](_page_27_Picture_4.jpeg)

Figure 41 – Create new AQL statement for an extractor for Amount

2. In the Create AQL Statement dialog enter "Number" as the view name and select Regular Expression as the type. As we plan to use this view later in a pattern, check the box 'Export view'. The export view statement will be generated in the template for the aql script. This makes the view available outside the RevenueByDivision\_BasicFeatures AQL module. Name the AQL script Number\_basic. After you have entered all that information click the OK button.

Division_BasicFeatures
sic Yeression Contract of the second
ew
ew
BM
ew
<u>sw</u>

Figure 42 – Create new AQL statement for an extractor for Amount

You will get the following template in the Number\_basic file within the aql editor pane. It will show certain errors.

![](_page_27_Figure_10.jpeg)

InfoSphere BigInsights – Text Analytics © Copyright IBM Corp. 2013. All rights reserved

Page 28 of 44

- 3. The following errors need to be fixed:
  - 1) the regular expression to look for a contiguous sequence of digits followed by an optional decimal point together with another sequence of digits,
  - 2) the <input column> for input files is text (as we saw before), and
  - 3) <input view> name for input files is always Document.
  - Save the aql files by selecting File → Save from the top menu bar in Eclipse , or press or ctrl+s on your keyboard

```
module RevenueByDivision_BasicFeatures;
create view Number as
extract regex /\d+(\.\d+)?/
        on R.text as match
from Document R;
export view Number;
```

Regular expression  $/\b\d+(\,\d{3}) * (\.\d+) ?\b/$  is more robust to use for numbers than the simple one  $/d+(\.\d+) ?/$  given above, though the simple regular expression suffices for this exercise.

#### II. Regular expression rule to identify Revenue:

Now we create another regular expression based extract statement called "*Revenue*" to identify Revenue clues. Basically, the mention of the term revenue along or the mention of the term revenue along with mention of first/second/third or fourth quarter, is a clue that we may be talking about the revenue of a division if the division also happens to be mentioned in close proximity of the revenue clue.

- 1. Navigate to **RevenueByDivision** → **Labels** → **Revenue.** Right-click the label **Revenue** and select **New AQL** Statement. Select Basic Features AQL statement.
- 2. Follow the steps stated in the previous section and fill in the pop-up box for 'Create AQL Statement'.
  - Enter "*Revenue*" as the *view name* and choose *Regular Expression* as the *type*.
  - Enter the aql script name as *Revenue\_basic* and check the box 'Output view'. Click the OK button.

4. The <input view> and <input column> have to be filled in as we did before for the *number* view. An implementation for the *regex* is suggested below. The final extractor would look as follows.

InfoSphere BigInsights – Text Analytics © Copyright IBM Corp. 2013. All rights reserved

Page 29 of 44

```
module RevenueByDivision_BasicFeatures;
create view Revenue as
extract regex /((.\s*)?(first|second|third|fourth)\s*-\s*quarter)?\s*(revenues?)/
    with flags 'CASE_INSENSITIVE'
    on between 1 and 5 tokens in R.text
    return
        group 0 as match
from Document R;
output view Revenue;
```

5. At this point, the Extraction Plan view should look as follows:

![](_page_29_Figure_3.jpeg)

Figure 43 – Extraction plan view after all extractors for basic features have been written.

To see the extractions of the three basic features (Division, Number and Revenue), ensure that all aql files have been saved. Navigate to *File* → *Save* (or ctrl+s) for each file.

Then navigate to the **Extraction Tasks** view, as shown in section 6.2. Select **Step 4: Test the Extractor** and choose the last option, **Run the extraction plan on the set of documents that are labeled** and click the **is icon** beside it.

7. The results of the extraction are displayed in the *Annotation Explorer* view. Double clicking on one of the rows in the annotation explorer will open the labeled document in the edit window. The extracted basic features can be highlighted by checking the corresponding boxes in the annotations window on the right.

2	Division_basic.aql 😰 Number_basic.aql 😰 Revenue_basic.aql 📳 4Q2006.txt - Documen 🕱 🔭 🗖 🗖	🖉 Extraction Plan 📴 4Q2006.txt - Document.text 🕱
i	Diluted earnings of \$6.06 per share from continuing operations, up 23 percent as reported, or up	D D G + + + + + + + + + + + + + + + + +
1	Diluted earnings per share from continuing operations were \$6.06 compared with \$4.91 per diluted	type filter text
	Income from continuing operations for the year ended December 31, 2006 was \$9.4 billion, compare	▼ □ Annotations
i	Revenues from continuing operations for 2006 totaled \$91.4 billion, essentially flat as reported	🕨 🗹 match (SPAN)
i	From a geographic perspective, the Americas full-year revenues were \$39.5 billion, an increase (	▼ □ RevenueByDivision_BasicFeatures.Revenue
i	Software segment revenues in 2006 totaled \$18.2 billion, an increase of 8 percent (7 percent, at a	match (SPAN)
	For total operations, net income for 2006 was \$9.5 billion, or \$6.11 per diluted share, which ir	

Figure 44 – Visualization of all the basic features extracted

① Note: Annotation matches for "Number" view are not displayed because we asked to export the view and not output.

The aql code for the extractors for the three basic features can be found under the folder **RevenueByDivision\_BasicFeatures** in Package Explorer. The code for all the files is listed below:

```
create dictionary DivisionDict
from file '../dictionaries/division.dict'
with language as 'en';
create view Division as
extract dictionary 'DivisionDict'
   with flags 'Exact'
   on R.text as match
from Document R;
output view Division;
create view Number as
extract regex /\d+(\.\d+)?/
   on R.text as match
from Document R;
export view Number;
create view Revenue as
extract
regex /((.\s*)?(first|second|third|fourth)\s*-\s*quarter)?\s*(revenues?)/
      with flags 'CASE INSENSITIVE'
      on between 1 and 5 tokens in R.text
      return
            group 0 as match
from Document R;
output view Revenue;
```

# 7. Writing Extractors for Concepts based on Basic Features

In this section we will extract two concepts:

InfoSphere BigInsights – Text Analytics © Copyright IBM Corp. 2013. All rights reserved

Page 31 of 44

- 1. Mention of revenue of an IBM division and
- 2. The reported revenue of that division.

We will make use of two aql constructs; the Extract Pattern statement, and the Select statement. The three basic features outlined in the previous section will be used as inputs.

- The reported revenue is a simple pattern a '\$' symbol followed by a number, followed by a revenue unit. We will use the extract pattern statement together with the view **Number** we previously created and a list of revenue units.
- The mention of revenue of a division has two different patterns: 1) mention of division name followed by some variant of the word 'revenue', 2) or the word 'revenue(s)' followed by mention of the division name. We will identify both patterns and use the 'union all' construct of aql to combine them into a single view.

#### 7.1 Create Extractor for AmountwithUnit

We will add a candidate generation rule for Amount. This uses Sequence Pattern from the Extract statement.

1. In the Extraction Plan view, navigate to *RevenueByDivision*  $\rightarrow$ *Labels*  $\rightarrow$  *Amount*, and right-click on the Amount Label. Select *New AQL statement* and the option *Candidate Generation AQL Statement*.

💶 📋 😳 🖓 Labels	🕟 New Label	>
Rasic Feature AQL Statement	🛄 New AQL Statement	$\rightarrow$
🛄 Candidate Generation AQL Statement	🗸 Mark Completed	
E Filter and Consolidate AQL Statement	💢 Delete	
💷 Final AQL statement	📄 Сору	
	of Cut	

Figure 45 – New AQL Statement to create extractor for candidate generation

 Fill in the dialog as shown in the following figure, with view name "AmountwithUnit", AQL script "AmountwithUnit\_concept", and for Type select "Pattern". Also, remember to check "Output view". Then click the OK button.

View Name:	AmountwithUnit	_
AQL Module:	RevenueByDivision_CandidateGeneration	\$
AQL script:	AmountwithUnit_concept	~
Туре:	Pattern	0
	Output view	
	Export view	

Figure 46 – Assigning name and extractor type to extractor for candidate generation

#### IBM Software Information Management

3. The new view appears in the Extraction Plan, and the following template is added to *AmountwithUnit\_concept.aql* in the *aql editor pane.* 

![](_page_32_Picture_2.jpeg)

① Note: Label names are case-sensitive, ensure that names are used correctly in later sections.

- 4. Here we will use the *basic feature* that we previously created: **Number**.
- 5. Add the following lines of code under the module definition for **RevenueByDivision\_CandidateGeneration**, to import the required view into the current module.

```
module RevenueByDivision_CandidateGeneration;
--<add the following line>
import view Number from module RevenueByDivision_BasicFeatures as Num;
```

● Note: An import statement puts objects in the context of the current module. The import statement is only used to import objects from other modules, not from the current module. Remember that an object that is declared in an AQL file of the module is visible to any other AQL file in that same module.

6. After the import, we fix the statement so it extracts the pattern we want. Your final aql file should look similar to the one below. **Save** the newly modified file.

```
module RevenueByDivision_CandidateGeneration;
import view Number from module RevenueByDivision_BasicFeatures as Num;
create view AmountwithUnit as
extract pattern '$'<N.match> ('million'|'billion'|'trillion')
return group 0 as match
from Num N, Document R;
output view AmountwithUnit;
```

InfoSphere BigInsights – Text Analytics © Copyright IBM Corp. 2013. All rights reserved

Page 33 of 44

7. Run the extractor by navigating to the Extraction Tasks view and selecting *Step 4.a* in the *Extraction Tasks view* (3<sup>rd</sup> option "Run the extraction plan on the set of documents that are labeled"), will produce the following AmountwithUnit results. *Double-click* on a row in Annotation Explorer to see the panel on the right, and check the box for *AmountwithUnit*.

🕒 Proj 📕 Pac 🐁 Bigl 🎯 Extr 🕱 🛛 🗖	🞦 Number_basic.aql 🛛 🔁 Revenue_basic.aql 🎦 AmountwithUnit_conce 🗐 4Q2006.txt - Documen 🕱 🦄 👘	- 0	🖉 Extractio 😉 4Q2006.t 🛿 🗖 🗖
<ul> <li>▷ Step 1 : Select Documents</li> <li>▷ Step 2 : Label Examples and Clues</li> <li>▷ Step 3 : Develop the Extractor</li> <li>▽ Step 4 : Test the Extractor</li> <li>a. Bun your extractor</li> </ul>			Image: Constraint of the second s
You can run on the entire data			match (SPAN)
collection that you specified in Step 1, only run on the documents that you selected in Step 1.b, or only run on the documents that you labeled in Step 2			RevenueByDivision_BasicFeature:      match (SPAN)      RevenueByDivision_ConditionGor      RevenueByDivision_ConditionGor
Run the extraction plan on the entire data collection	Y		▶ ✓ match (SPAN)
Run the extraction plan on the set of selected documents	i i hanges. Diluted earnings per share for the fourth-quarter 2006 grew 7 percent compared with	וו	
Run the extraction plan on the set of documents that are labeled	i ourth quarter of 2006 of \$26.3 billion in reased 7 percent (4 percent, adjusting for currer ients and shareholders."		
The extraction results are displayed in the Annotation Explorer.	i ncy) to <b>\$4.8 billion</b> . OEM revenues were <b>\$1.0 billion</b> , down 3 percent compared with the 2000	5	
b. Identify mistakes in the extracted results	i ems revenues decreased 2 percent to \$642 million compared with the prior-year quarter. Reve evenues from <u>Tivoli</u> software, infrastructure software that enables customers to centrally percent and the prior of the second s	n	
The result of your extractor might contain mistakes: false positives (results incorrectly identified by the	${f i}$ d ended the full year with an estimated services backlog, including Strategic <u>Outsourcing</u> ,		
extractor) and false negatives (correct mentions that are not identified by the extractor)	i ercent. Revenues from the System b LNTX server products increased 4 percent compared with 3		
Use the <u>Provenance Viewer</u> to find out the rules that cause false positives.	🖹 Problems 📮 Console III: Annotation Explorer 😫 👘 🕹 📳 💀 Y 🍰 🎬 Y 🎲 Y 📫 🎬 Text analytics result. Number of rows: 677/677 Showing page 1 of 1		
Go back to Step 2 to identify additional	Input Document Left Context Span Attribute Value Right Context Span A	.ttr 🦳	
clues, or go back to Step 3 to create new statements or fix existing	402006 tyt billion, increased its liabilities by \$0.2 billion [11072.11094] and reduced stockholders' equity Revenue	σE	
statements.	4Q2006.txt \$4.8 billion. OEM revenues were \$1.0 billion [2506-2518] , down 3 percent compared with t Revenues	еE	
Repeat Steps 2 to 4 until you are	4Q2006.txt arges of \$1.7 billion, offset by the \$1.1 billion [8756-8768] gain on the sale of the Personal Reven	еE	
satisfied with the results of your extractor.	4Q2006.txt purchases totaled approximately \$1.4 billion [7260-7272] in the fourth quarter. The weighte Revenue	еE	
	4Q2006.txt remental restructuring charges of \$1.7 billion [8728-8740] , offset by the \$1.1 billion gain on Revenue	IEE	
See Example	402006.txt   Global Financing debt increased   \$1.8 billion [12752-12764]   from year-end 2005 to a total of \$	IEE	

Figure 48 – Results for AmountwithUnit extractor, sequential occurrence of a number followed by a unit

# 7.2 Create Extractor for RevenueByDivision

At this point we need to add a Candidate Generations rule for RevenueByDivision to identify occurrence of a division name followed by a reference to revenue, with at most one token in between. For example "**Global Financing segment revenues** increased 3 percent" or "**Software segment revenues** in 2006 totaled \$18.2 billion" For this section, the AQL script was already created. You will need to import the aql file.

- 1. From **Package Explorer** view, under MyFirstProject, navigate to **textAnalytics** → **src** and right click **RevenueByDivision\_CandidateGeneration**.
- 2. Select Import → File System and Next. Browse to /home/biadmin/bootcamp/input/lab04\_TextAnalytics
- 3. Select **RevenueByDivision.aql.** Verify that the box next to **lab04\_TextAnalytics** folder has a horizontal line and the box next to **RevenueByDivision** file has a **checkbox**. Click **Finish.**

0	Import	×
File system Import resources from the loca	file system.	
From directory //home/biadm	n/bootcomp/input/lab04_ToxtApolydics	Browse
Iab04_TextAnalytics	V 🎬 RevenueByDivi	ision.aql
Filter Types	Deselect All	
Into folder: MyFirstProject/tex	tAnalytics/src/RevenueByDivision_CandidateC	Generation Bro <u>w</u> se
Options           Options           Overwrite existing resource	es without warning	
Create complete folder st	ucture	
Advanced >>		
?	< <u>B</u> ack Next >	Cancel <u>Einish</u>

Figure 49 Import RevenueByDivision.aql

- 4. In order to use the two views from other AQL module, they have to be available. Change the definition for *Revenue* and *Division* to *export* the view rather than use as output. Edit **Division\_basic.aql** and change "*output view Division*" statement to "export view Division". Same for Revenue\_basic.aql.
- 5. Before running the extraction plan, save all the scripts.
- 6. In the Extraction Tasks view, select Step 4: Test the Extractor, choose the last option, Run the extraction plan on the set of documents that are labeled and click the *icon* beside it. You will see mentions of division names reference with revenue, and their revenues as shown below.

() Note: Select the match spans for Division and Revenue in the Right Pane to highlight the extracted text

🞦 Division_basic.aql  🖹 Revenue_basic.aql 🕼 4Q2006.txt - Documen 🕄 🍡	- 0	🖉 Extractio 🔚 4Q2006.t 🔉 🗖 🗖
	^	D D fr 4- 4 5 5
		type filter text
1 i		
i		RevenueByDivision_CandidateGer
		P ✓ match (SPAN)
i hange and incremental restructuring charges of \$1.7 billion, offset by the \$1.1 billion		<ul> <li>RevenueByDMston_CandidateGer</li> <li>match (SPAN)</li> </ul>
i period.		▷ division (SPAN)
i adjusting for currency and PCs) to $\mathbf{x}_{17}$ 6 billion OEM revenues were $\mathbf{x}_{3}$ 9 billion up 18	-	
adjusting for currency marries in arrest of the off. Off Perences where pass bittion, up is		
i djusting for currency). S&TG segment revenues were \$22.0 billion an increase of 5 perce	•	
i ffect of the FASB Interpretation No. 47 accounting change of \$36 million.		
i	_	
i financial position (or balance sheet). The funded status is measured as the difference b		
i		
1 1.		
the company's failure to continue to develop and market new and innovative products and	~	
(         III         >		
🖹 Problems 📮 Console 🏦 Annotation Explorer 🛿 🛛 🔹 👘 🗸 🚔		
Text analytics result, Number of rows: 475/475 Showing page 1 of 1		
Input Document Left Context Span Attribute Value Right Context Sp	an 🔒	
4Q2006.txt billion, increased its liabilities by \$0.3 billion [11972-11984] and reduced stockholders' equity Re	ve 🗖	
4Q2006 tut \$4.9 billion, OEM revenues were \$1.0 billion [2606 2619]		
4Q2006.txt arges of \$1.7 billion, offset by the \$1.1 billion [8756-8768] gain on the sale of the Personal Re	ve	
422006 tut purchasses totaled approximately \$1.4 billion [7260-7272] in the fourth quarter. The weighte Pa		
4Q2006.txt remental restructuring charges of \$1.7 billion [8728-8740] , offset by the \$1.1 billion gain on Re	ve	
402006.txt   Global Hinancing debt increased   \$1.8 billion [12752-12764]   from year-end 2005 to a total of \$ Re	ve 🕑	

Figure 50 – Visualizing the RevenueByDivision and AmountwithUnit entities extracted

The two red rectangles in the figure highlight the missing association between the division name and its revenue, which we will establish in the next exercise. Key to this association is that the revenue amount follows the reference to revenue of a division within a few tokens.

# 7.3 Extending the extractor for RevenueByDivision by including AmountwithUnit information

In this section we will create one view for the reference to the revenue of a division and the reported revenue of that division. This will be done by using the '*select*' statement of AQL to combine the RevenueDivision and AmountwithUnit views. The pattern we will look for is revenue of Division followed by AmountwithUnit with a maximum of 15 intervening tokens.

 Add a Candidate Generation rule for DivRevenueWithAmount by navigating to Extraction Plan and right-clicking on RevenueByDivision. Select New AQL Statement → Candidate Generation AQL statement. Fill in the fields as follows: the view name "DivRevenueWithAmount", AQL script "DivRevenueWithAmount", and type "Select". Also, remember to check "Output view". Then click the OK button:

View Name:	DivRevenueWithAmount			
AQL Module:	RevenueByDivision_CandidateGeneration	\$		
AQL script:	DivRevenueWithAmount			
Type:	Select	\$		
	Output view			

Figure 51 – Creating view DivRevenueWithAmount

2. List the two views needed as the input in the 'from' clause, RevenueByDivision and AmountwithUnit. In the 'where' clause, we specify that a maximum of 15 intervening tokens are allowed between the reference to a division and the revenue number. Finally in the select clause we select not only the spans of interest, but also the values using the GetText() function. The final code for this view is:

```
module RevenueByDivision_CandidateGeneration;
create view DivRevenueWithAmount as
select CombineSpans(R.match, A.match)
    as match,
    R.division as division,
    A.match as revenue,
    GetText (R.division) as divName,
    GetText (A.match) as revText
from RevenueByDivision R, AmountwithUnit A
where FollowsTok(R.match, A.match, 0,15);
output view DivRevenueWithAmount;
```

3. Save the file and in Extraction Tasks, select Step 4: Test the Extractor, choose the last option, Run the extraction plan on the set of documents that are labeled and click the *icon* beside it.

Double Click one of the entries inside the Annotation Explorer.

You will see the **DivRevenueWithAmount** view on the panel on the right hand side. Within that panel, toggle the checkboxes for *match, division*, and *revenue* fields to see the corresponding extracted entities as shown in the figure below.

		🖀 Extraction 📴 4Q2006.t 😫 📃 🗖
	^	▶ ₽ ☆ ↓ ↓ 수 수 완 달
		type filter text
i roe:		
		match (SPAN)      RevenueRvDivicion_CandidateGong
\$2.01 per share in the fourth quarter of 2005; the fourth-quarter 2005 diluted earnings include \$0.10 per		✓ match (SPAN)
i rations for the fourth quarter grew 2 percent compared with the fourth-quarter 2005 income from continuing		division (SPAN)
l as record <u>payouts</u> to shareholders. We are well-po <u>sitioned in the g</u> rowth areas of a changing IT industry,		revenue (SPAN)
i eriod. Revenues from Europe/Middle East/Africa were \$9.3 billion, up 11 percent (3 percent, adjusting for c		
		match (SPAN)
1 's <u>middleware</u> brands, which include WebSphere, Inf <del>ormation Manageme</del> nt, <u>Tivoli</u> , Lotus and Rational products,		division (SPAN)
ons, data and operating systems, revenues increased 22 percent. Revenues for Information Management softwar		
i <mark>from Global Business Services</mark> increased 6 percent (3 percent, adjusting for currency) to <b>\$4.2 billion</b> . IBM		
i yes products increased 5 percent compared with the year-age period. Total delivery of System 2 computing po		
1		
i d 9 percent compared with the year-ago period. Intellectual property and custom development income increase		
i d 9 percent compared with the year-ago period. Intellectual property and custom development income increase	~	
i d 9 percent compared with the year-ago period. Intellectual property and custom development income increase yorship affect of saveral itams in the guarter including the retroactive reinstatement of the U.S. research (	~	
i d 9 percent compared with the year-ago period. Intellectual property and custom development income increase vorshie affect of saveral itams in the guarter, including the retroactive reinstatement of the U.S. receases (C) = Console # Annotation Explorer © Problems ■ Console # Annotation Explorer ©		
i d 9 percent compared with the year-ago period. Intellectual property and custom development income increase yorshie affect of saveral items in the guarter including the retroactive reinstatement of the U.S. research () 2 Problems  Console  I Annotation Explorer  Console		
i d 9 percent compared with the year-ago period. Intellectual property and custom development income increase vorable affact of exercl items in the quarter including the retroactive reinstatement of the U.S. recearch ()          Image: Console # Annotation Explore IS       Image: Console # Annotation Explore IS         Image: Console # Annotation Explore IS       Image: Console # Annotation Explore IS         Image: Console # Annotation Explore IS       Image: Console # Annotation Explore IS         Image: Console # Annotation Explore IS       Image: Console # Annotation Explore IS         Image: Console # Annotation Explore IS       Image: Console # Annotation Explore IS         Image: Console # Annotation Explore IS       Image: Console # Annotation Explore IS         Image: Console # Annotation Explore IS       Image: Console # Annotation Explore IS         Image: Console # Annotation Explore IS       Image: Console III Image: Console IIII Image: Context         Image: Console III Image: Im		
1       d 9 percent compared with the year-ago period. Intellectual property and custom development income increase variable affect of several items in the quarter including the retroactive reinstatement of the U.S. researce (C)         (C)       (C)       (C)		
i d 9 percent compared with the year-ago period. Intellectual property and custom development income increase (a) a ffact of several items in the quarter including the retroactive reinstatement of the U.S. research (c) a ffact of several items in the quarter including the retroactive reinstatement of the U.S. research (c) a ffact of several items in the quarter including the retroactive reinstatement of the U.S. research (c) a ffact of several items in the quarter including the retroactive reinstatement of the U.S. research (c) a ffact of several items in the quarter including the retroactive reinstatement of the U.S. research (c) a ffact of several items in the quarter including the retroactive reinstatement of the U.S. research (c) a ffact of several temperature in the quarter including the retroactive reinstatement of the U.S. research (c) a ffact of several temperature in the quarter including temperature including temperature including temperature including temperater in the retroactive retroactive temperature including temperatere including temperature including temperature including temper		
i d 9 percent compared with the year-ago period. Intellectual property and custom development income increase worship affact of saveral items in the guarter including the retroactive reinstatement of the U.S. researce (C) Problems Console II: Annotation Explore II Problems Console II: Annotation Explore II Input Document Left Context Span Attribute Value Right Context Span Attribute Name 100005 tit billion increased its Habilities by 60.2 billion [1007] 10041 and reduced testibules and the RevenueByDMision_Candidate 402006 tit As Buillion offset by the \$1.0 billion [2560-2518] gain on the sale of the Personal RevenueByDMision_Candidate 402006.tit purchases totaled approximately \$1.4 billion [2560-2727] in the fourth quarter. The weighted RevenueByDMision_Candidate		
i d 9 percent compared with the year-ago period. Intellectual property and custom development income increases worshie affact of saveral items in the quarter including the retroactive reinstatement of the U.S. researce (C) Problems Console II: Annotation Explore II Problems Console II: Annotation Explore III Problems Console II: Annotation Explore III Problems Console II: Annotation Explore III Input Document Left Context Span Attribute Value Right Context Span Attribute Name 402006 txt \$48 billion Offset by the \$1.0 billion [2506-2518], down 3 percent compared with t RevenueByDMsion_Candidate 402006 txt arges of \$1.7 billion, offset by the \$1.1 billion [2506-2722] In the fourth quarter. The weighter RevenueByDMsion_Candidate 402006 txt purchases totaled approximately \$1.4 billion [250-2720] In the fourth quarter. The weighter RevenueByDMsion_Candidate 402006 txt purchases totaled approximately \$1.4 billion [250-2720] In the fourth quarter. The weighter RevenueByDMsion_Candidate 402006 txt purchases totaled approximately \$1.4 billion [250-2720] In the fourth quarter. The weighter RevenueByDMsion_Candidate 402006 txt purchases totaled approximately \$1.4 billion [250-2720] In the fourth quarter. The weighter RevenueByDMsion_Candidate 402006 txt purchases totaled approximately \$1.7 billion [8728-8740] , offset by the \$1.7 billion gain on RevenueByDMsion_Candidate 402006 txt purchases totaled approximately \$1.4 billion [250-2522] In the fourth quarter. The weighter RevenueByDMsion_Candidate 402006 txt purchases totaled approximately \$1.4 billion [8728-8740] , offset by the \$1.7 billion gain on RevenueByDMsion_Candidate 402006 txt purchases totaled approximately \$1.4 billion [8728-8740] , offset by the \$1.1 billion gain on RevenueByDMsion_Candidate 402006 txt purchases totaled approximately \$1.4 billion [8728-8740] , offset by the \$1.1 billion gain on RevenueByDMsion_Candidate 402006 txt purchases totaled approximately 40 billion [8728-8740] , offset by the \$1.1 billion [8728-8740] , of		

Figure 52 – Division and revenue extracted as a single concept

# 8. Analyzing Extracted Results with Annotation Explorer

The Annotation Explorer has a row entry for every span identified in every output view. In this section we will first display one of the output view tables defined in the previous sections. Then we will use filters to display a selected subset of these rows.

# 8.1 Displaying Output View Table

 To display an output view table, click the down arrow in the highlighted area in Figure 53 below, and select *'RevenueByDivision\_CandidateGeneration.DivRevenueWithAmount*' view (in case if you do not find the icon, on the top menu bar, select *Window* → *Reset Perspective* first). This will generate the output view table as shown in figure 54. Each row in Figure 53 is a relation, or tuple, of the output view. The columns of the table are the fields in the tuple. For fields of type span, the text corresponding to SPAN is also included.

Annotation	Explorer 🛿			∢ ▶ 📳	⊪× ≱× 4 ≌ □ □
Text analytics result, Number of rows: 123/123 Showing page 1 of 1				RevenueRyDivision. CandidateGeneration AmountWithUnit	
Input Docume	Left Context	Span Attribute Value	Right Context	Span Attribute Name	RevenueByDivision_CandidateGeneration.DivRevenueWithAmount
/4Q2006.txt	ncreased its liabilities by	0.3 billion [11972-11984	and reduced stock	RevenueByDivision_C	RevenueByDivision_CandidateGeneration.RevenueByDivision
/4Q2006.txt	ion. OEM revenues were	\$1.0 billion [2506-2518]	, down 3 percent co	RevenueByDivision_C	andidateGeneration.Ar
/4Q2006.txt	\$1.7 billion, offset by the	\$1.1 billion [8756-8768]	gain on the sale of	RevenueByDivision_C	andidateGeneration.Ar
/4Q2006.txt	s totaled approximately	\$1.4 billion [7260-7272]	in the fourth quart	RevenueByDivision_C	andidateGeneration.Ar
/4Q2006.txt	restructuring charges of	\$1.7 billion [8728-8740]	, offset by the \$1.1	RevenueByDivision_C	andidateGeneration.Ar
/4Q2006.txt	inancing debt increased	1.8 billion [12752-12764	from year-end 200	RevenueByDivision_C	andidateGeneration.Ar

Figure 53 – Setting up display of an Output View

🖶 RevenueByD	ivision_Candidat	eGeneration.Di	vRevenueWithAr	mount 🕱		4 ▷ 🗏 🔤 🗸 🗳
Text analytics re	sult, Number of	rows: 10	Showing pag	Showing page 1 of 1		
match (SPAN)	division (SPAN	revenue (SPA)	divName (TEX	revText (TEXT)	Input Docume	Double-click this column to explain a tuple !
Revenues from	Software [259	\$5.6 billion [26	Software	\$5.6 billion	/4Q2006.txt	Explain
revenues from	Global Technol	\$8.6 billion [41	Global Technol	\$8.6 billion	/4Q2006.txt	Explain
revenues from	Global Busines	\$4.2 billion [42	Global Busines	\$4.2 billion	/4Q2006.txt	Explain
Revenues from	Systems and 1	\$7.1 billion [47	Systems and 1	\$7.1 billion	/4Q2006.txt	Explain
Global Financin	Global Financir	\$620 million [5	Global Financir	\$620 million	/4Q2006.txt	Explain
Software segm	Software [989	\$18.2 billion [9	Software	\$18.2 billion	/4Q2006.txt	Explain
Revenues from	Global Technol	\$32.3 billion [1	Global Technol	\$32.3 billion	/4Q2006.txt	Explain
Revenues from	Global Busines	\$16.0 billion [1	Global Busines	\$16.0 billion	/4Q2006.t×t	Explain
S&TG segmen	S&TG segmen	\$22.0 billion [1	S&TG segmen	\$22.0 billion	/4Q2006.txt	Explain
Global Financir	Global Financir	\$2.4 billion [10	Global Financir	\$2.4 billion	/4Q2006.txt	Explain

Figure 54 – Output View Table

## 8.2 Filtering Annotation Explorer Rows

1. Click on the *Select Filters for the Annotation Explorer* drop down button (highlighted in the figure below), and select *Show hide filter view*.

🗄 Annotatio	n Explorer 없		< > 🖩 💀 🗸	
Text analytics	result, Number of rows: 123/123			Show hide filter view
Input Docun	ne Left Context	Spantext	Span Attribute Name	Clear all filters
/4Q2006.txt	illion, increased its liabilities by	:0.3 biluced stock	${\tt RevenueByDivision\_CandidateGeneration.AmountWithUnit.mat}$	tch
/4Q2006.txt	4.8 billion. OEM revenues were	\$1.0 bpercent co	RevenueByDivision_CandidateGeneration.AmountWithUnit.mat	tch
/4Q2006.txt	jes of \$1.7 billion, offset by the	\$1.1 bthe sale of	RevenueByDivision_CandidateGeneration.AmountWithUnit.mat	tch
/4Q2006.txt	rchases totaled approximately	\$1.4 burth quart	RevenueByDivision_CandidateGeneration.AmountWithUnit.mat	tch

Figure 55 – Setting up filters for information displayed in Annotation Explorer

2. In the drop-down menu for *Filter Criteria* shown in below, check the boxes for '**Span Attribute Name Filter**' and '**Input Document Filter**' as shown in the figure below.

Filter Criteria	
🗹 Span Attribute Name Filter	
Not Configured	2
🗌 Span Attribute Value Filter	
Not Configured	2
🗹 Input Document Filter	
Not Configured	
🗌 Left Context Filter	
Not Configured	
🗌 Right Context Filter	
Not Configured	2

Figure 56 – Selecting the filter Criteria

#### IBM Software Information Management

Then click the wrench icon to bring up the wizard for configuring the '*Span Attribute Name Filter*' shown in figure 57.

3. In the *Configuration Filter* wizard shown below, select the span name

"RevenueByDivision\_CandidateGeneration.DivRevenueWithAmount.match" as shown, click Add and then click OK. This sets up the permitted values of Span Attribute Name, which is the last column in the Annotation Explorer view, for the rows displayed in Annotation Explorer.

0	Span Attribute Name Selection X
Selec	t those span columns you would like to see in the Annotation Explorer
Rev	enueByDivision_CandidateGeneration.AmountwithUnit.match
Rev	enueBvDivision CandidateGeneration.DivRevenueWithAmount.division
Rev	enueByDivision_CandidateGeneration.DivRevenueWithAmount.match
Rev	
Rev	enueByDivision_CandidateGeneration_RevenueByDivision_match
1.01	
3	
~ .	Add
Selec	
Reve	nueByDivision_CandidateGeneration.DivRevenueWithAmount.match
	Remove
0	
0	Cancel OK

Figure 57 – Configuring each Filter Criteria

4. Next we click on the wrench icon for *'Input Document Filter'* and add */4Q2006* to the selected entries, then click *OK* to exit. This sets up the permitted values of Input Document, which is the *first* column in the Annotation Explorer view. Once the configurations for the two filters are set, we will see the 10 annotations in annotation explorer, as shown below.

😰 Problems 📮 Console 🗱 Annotation Explorer 😫 🔹 🕨 🍰 🙄						
Text analytics res	sult, Number of rows: 10/123	Showing page 1 of 1				
Input Document	Left Context	Span Attribute Value	Right Context	Span Attribute Name	Filter Criteria	
4Q2006.txt	· percent, adjusting for currency).	bal Financing revenues totaled \$2.4 billion [10394-104	, a decrease of 2 percent (2 perce	RevenueByDivision_Candid	🗹 Span Attribute Name Filter	
4Q2006.txt	n Storage increased 9 percent.	reased 3 percent (flat, adjusting for currency) in the fo	. The company's total gross pro	RevenueByDivision_Candid	RevenueByDivision_CandidateGenerati 🔗 📝	
4Q2006.txt	or currency) compared with 2005.	1e Global Business Services segment were \$16.0 billi	, flat (up 1 percent, adjusting for c	RevenueByDivision_Candid	on.DivRevenueWithAmount.match	
4Q2006.txt	' percent, adjusting for currency).	Global Technology Services segment totaled \$32.3 b	, an increase of 3 percent (2 perc	RevenueByDivision_Candid	Span Attribute Value Filter	
4Q2006.txt	ed with the 2005 fourth quarter.	ies from the Software segment were \$5.6 billion [257;	, an increase of 14 percent (11 pe	RevenueByDivision_Candid	Not Configured 🛛 🖉 💌	
4Q2006.txt	illion from the prior-year period.	tems and Technology Group (S&TG) segment totaled	for the quarter, up 3 percent (flat,	RevenueByDivision_Candid		
4Q2006.txt	percent, adjusting for currency).	TG segment revenues were \$22.0 billion [10290-103]	, an increase of 5 percent (4 perc	RevenueByDivision_Candid	☑ Input Document Filter	
4Q2006.txt	B percent compared with 2005.	e segment revenues in 2006 totaled \$18.2 billion [989	, an increase of 8 percent (7 perc	RevenueByDivision_Candid	4Q2006.txt	
4Q2006.txt	ncy) to \$8.6 billion, and segment	vices increased 6 percent (3 percent, adjusting for c	. IBM signed services contracts to	RevenueByDivision_Candid	<b></b>	
4Q2006.txt	obal Services business, segment	ervices increased 7 percent (4 percent, adjusting for (	, and segment revenues from Glo	RevenueByDivision_Candid	Left Context Filter	
					Not Configured 🕢 📝	
					Right Context Filter	
					Not Configured	

Figure 58 – Results of Filters applied – fewer rows displayed in Annotation Explorer, those matching the filter criteria

The filters for the Annotation Explorer view can be reset by clicking on the *Clear all filters* in the *Select Filters for the Annotation Explorer* drop down menu; this would reset the default view of Annotation Explorer. In general, one can choose any subset of filter criteria; and for each filter criteria, any subset of values.

5. By clicking on the filter icon again you can get rid of the drop down menu for Filter Criteria.

### 8.3 Exporting an Output View

1. A left-click on the '*Export Results*' icon highlighted in the figure below, will bring in the Export Results dialog box. Enter /*home/biadmin/Desktop* for the name of the output directory. Click *Finish* to close the dialog box.

_	
	Export Results X
)3 in 16 \$ ioi : t	Export Results All results will be exported. Applied filters will be ignored.
	Path for the exported results           Browse File System
	Cancel Einish

Figure 59 – Exporting extracted views from Annotation Explorer tools bar

2. Two folders named 'csv' and 'html' are created in the directory specified. Each folder contains a file for each output view. The *DivRevenueWithAmount.html* file is shown below .

RevenueByDivision_Candid	ateGeneration DivRevenueWithAmount - Mozilia Fir	elox	and the second state of th	-	•
Eile Edit View Higtory Bookmarks Tools Help					
RevenueByDivision_CandidateGene.					
🕘 🕘 file://home/biadmin/Desktop/html/RevenueByDwision_CandidateGeneration.DwRevenueWithAmount.html		10 × 60 🛽	¥ Google	Q 4	2
Eginsights Console 😡 Data Explorer					
Document '4Q2006.txt'					
match (SPAN)	division (SPAN)	revenue (SPAN)	divName (TEXT)	revText	
Revenues from the Software segment were \$5.6 billion [2577 - 2629]	Software [2595 - 2603]	\$5.6 billion [2617 - 2629]	Software	\$5.6 billion	24
revenues from Global Technology Services increased 7 percent (4 percent, adjusting for currency) to \$8.6 billion [4073 - 4185]	Global Technology Services [4087 - 4113]	\$8.6 billion [4173 - 4185]	Global Technology Services	\$8.6 billion	
revenues from Global Bunness Services increased 6 percent (3 percent, adjusting for currency) to \$4.2 billion [4199 - 4309]	Global Business Services [4213 - 4237]	\$4.2 billion [4297 - 4309]	Global Business Services	\$4.2 billion	
Revenues from the Systems and Technology Group (S&TG) segment totaled \$7.1 billion [4673 - 4755]	Systems and Technology Group (S&TG) segment [4691 - 4734]	\$7.1 billion [4743 - 4755]	Systems and Technology Group (S&TG) segment	\$7.1 billion	
Global Financing segment revenues increased 3 percent (flat, adjusting for currency) in the fourth quarter to \$620 million [5371 - 5493]	Global Financing [5371 - 5387]	\$620 million [5481 - 5493]	Global Financing	\$620 million	
Software segment revenues in 2006 totaled \$18.2 billion [9898 - 9953]	Software [9898 - 9906]	\$18.2 billion [9940 - 9953]	Software	\$18.2 billion	
Revenues from the Global Technology Services segment totaled \$32.3 billion [10017 - 10091]	Global Technology Services [10035 - 10061]	\$32.3 billion [10078 - 10091]	Global Technology Services	\$32.3 billion	
Revenues from the Global Buriness Services segment were \$16.0 billion [10174 - 10343]	Global Business Services [10192 - 10216]	\$16.0 bilion [10230 - 10243]	Global Business Services	\$16.0 billion	
S&TG segment revenues were \$22.0 billion [10290 - 10330]	S&TG segment [10290 - 10302]	\$22.0 billion [10317 - 10330]	S&TG segment	\$22.0 billion	
Global Financing revenues totaled \$2.4 billion [10394 - 10440]	Global Financing [10394 - 10410]	\$2.4 billion [10428 - 10440]	Global Financing	\$2.4 billion	

-----

Figure 60 – Output html document for the view DivRevenueWithAmount

## 8.4 Mouse-Over function to explain the annotated text

1. Left-clicking on the icon *Disable Span Tooltip* shown below toggles the mouse-over function on the annotated text document. When the mouse-over function is disabled, there is a red diagonal line over the icon as shown in the figure.

	ons, data	and operating systems, re	evenues increased 22 percent. Revenues	for Information Managemen	nt software,		type filter text
1	i from <mark>Glo</mark> k	al Business Services incr	reased 6 percent (3 percent, adjusting	for currency) to \$4.2 bil	lion. IBM 🖞		
	i ver produc	ts increased 5 percent co	ompared with the year-ago period. Tota	l delivery of System z com	puting powe		▼ 🗌 RevenueByD
	i				_		🕨 🗹 match (Sf
							▼ 🗹 RevenueByD
							🕨 🗹 match (Sf
🖹 Problems 🖳 Console 🗄 Annotation Explorer 🖇 🕴 👘 🗸 📅 🗖							Þ 🗹 division (S
Text analytics result, Number of rows: 123/123 Showing page 1 of 1						🕨 🗹 revenue (S	
L	Input Document	Left Context	Span Attribute Value	Right Context	Span Attribute		▼ 🗌 RevenueByD
	4Q2006.txt	billion, increased its liabilities by	\$0.3 billion [11972-11984]	and reduced stockholders' equit	y RevenueByDiv		👂 📃 match (Sf
I	4Q2006.txt	\$4.8 billion. OEM revenues were	\$1.0 billion [2506-2518]	, down 3 percent compared with	t RevenueByDiv		division (S
	4Q2006.txt	arges of \$1.7 billion, offset by the	\$1.1 billion [8756-8768]	gain on the sale of the Personal	RevenueByDiv		
			1				

Figure 61 – Clicking on the icon in red rectangle enables/disables the mouse-over function in Annotation Explorer

2. When the mouse-over function is enabled, moving the mouse over an entry in Annotation Explorer view causes a pop-up. The pop-up contains the name of the top level view for the annotated text, the annotated text, and the span for the annotated text.

Problems	Console 🔹 Annotation Explorer 🕱			4 ▶ ≣ ₽× \$>× 4 ⊠	- 0
Text analytics res	sult, Number of rows: 123/123	Showing page 1 of 1			
Input Document	Left Context	Span Attribute Value	Right Context	Span Attribute Name	^
4Q2006.txt	omputing (PC) business, and the	\$775 million [8835-8847]	legal settlement received from M	RevenueByDivision_CandidateGeneration.Amo	nuo
4Q2006.txt	5 was \$9.4 billion, compared with	\$8.0 billion [8416-8428]	in the year-ago period, or up 18 p	RevenueByDivision_CandidateGeneration.Am	nuo
4Q2006.txt	purchases totaled approximately	\$8.0 billion [12	Division Condidate Convertion Am	sion_CandidateGeneration.Am	nuo
4Q2006.txt	ercent, adjusting for currency) to	\$8.6 billion [4 match: \$8.0	billion [12251,12263]	sion_CandidateGeneration.Amo	our 🔤
4Q2006.txt	ercent, adjusting for currency) to	\$8.6 billion [4173-4185]	, and segment revenues from Glo	RevenueByDivision_CandidateGeneration.DivF	Rev
4Q2006.txt	educed the company's assets by	\$9.2 billion [11929-11941]	, increased its liabilities by \$0.3 b	RevenueByDivision_CandidateGeneration.Am	nuo
4Q2006.txt	1 Europe/Middle East/Africa were	\$9.3 billion [2328-2340]	, up 11 percent (3 percent, adjust	RevenueByDivision_CandidateGeneration.Amo	nuo
402006 tyt	mo from continuing operations of	\$0.4 billion [7616 7629]	up 19 percept as reported, or up	RevenueRyDivision_CandidateGeneration Am	~ ~

Figure 62 – Illustration of mouse-over function

# 9. Summary

This lab explored how to use the Text Analytics feature of BigInsights Enterprise. Working mainly with Eclipse with the BigInsights Eclipse Tooling plug-in installed, you were introduced to the ideas of basic features, concepts, patterns, regular expressions, and more. Following the steps of the Extraction Task view simplified the task of developing AQL to extract the required information. The Extraction Plan view was useful to keep track of all basic features, concepts, candidate generation, and so on. You can combine what you have learned here with other Big Data tools such as BigSheets, JAQL, and Streams.

![](_page_43_Picture_1.jpeg)

© Copyright IBM Corporation 2013 All Rights Reserved.

IBM Canada 8200 Warden Avenue Markham, ON L6G 1C7 Canada

IBM, IBM (logo), and DB2 are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both.

 $\mbox{Linux}$  is a trademark of Linus Torvalds in the United States, other countries, or both.

Windows is a trademark of Microsoft Corporation in the United States, other countries, or both.

 $\mathsf{VMware}$  is a trademark or  $\mathsf{VMware}$  Inc. in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.

References in this publication to IBM products or services do not imply that IBM intends to make them available in all countries in which IBM operates. 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 performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurement may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

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.

The information in this publication is provided AS IS without warranty. Such information was obtained from publicly available sources, is current as of July 2010, and is subject to change. Any performance data included in the paper was obtained in the specific operating environment and is provided as an illustration. Performance in other operating environments may vary. More specific information about the capabilities of products described should be obtained from the suppliers of those products.

InfoSphere BigInsights – Text Analytics © Copyright IBM Corp. 2013. All rights reserved

Page 44 of 44