Rules! Mining! Scoring! Custom Analytics! IBM InfoSphere Streams Does It All with Microsecond Latencies

Sreekanth Iyer
Product Manager, InfoSphere Streams
IBM India Software Labs

InformationOnDemandIndia2011

The Premier Conference for Information Management
Manage. Analyze. Govern.

February 2, 2011
Hyatt Regency I Mumbai, India

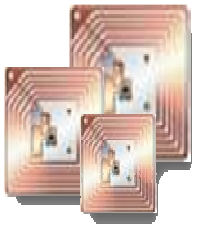
IBM Disclaimer

Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

Data and Real World Events are growing …

2 Billion **Internet users by 2011**

1.3 Billion RFID tags in 2005
30 Billion **RFID tags by 2010**



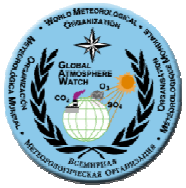
4.6 Billon
Mobile Phones World Wide

Capital market data volumes grew ***1,750%*****, 2003-06**

twitter

Twitter process ***7 terabytes*** **of data every day**

World Data Centre for Climate
- ***220 Terabytes*** **of Web data**
- ***9 Petabytes*** **of additional data**



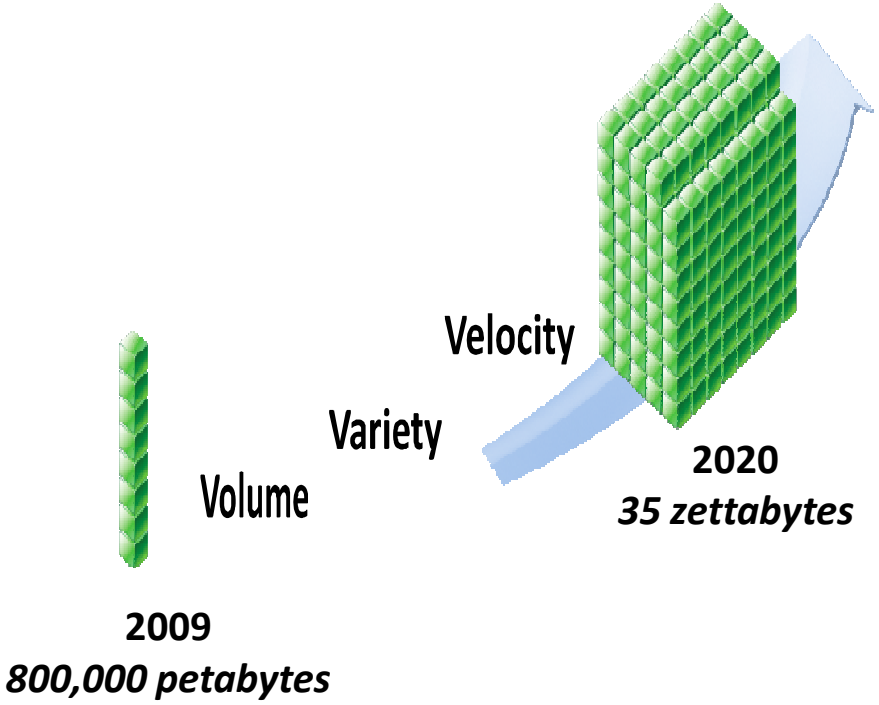
facebook

Facebook process ***10 terabytes*** **of data every day**

Information is Exploding…

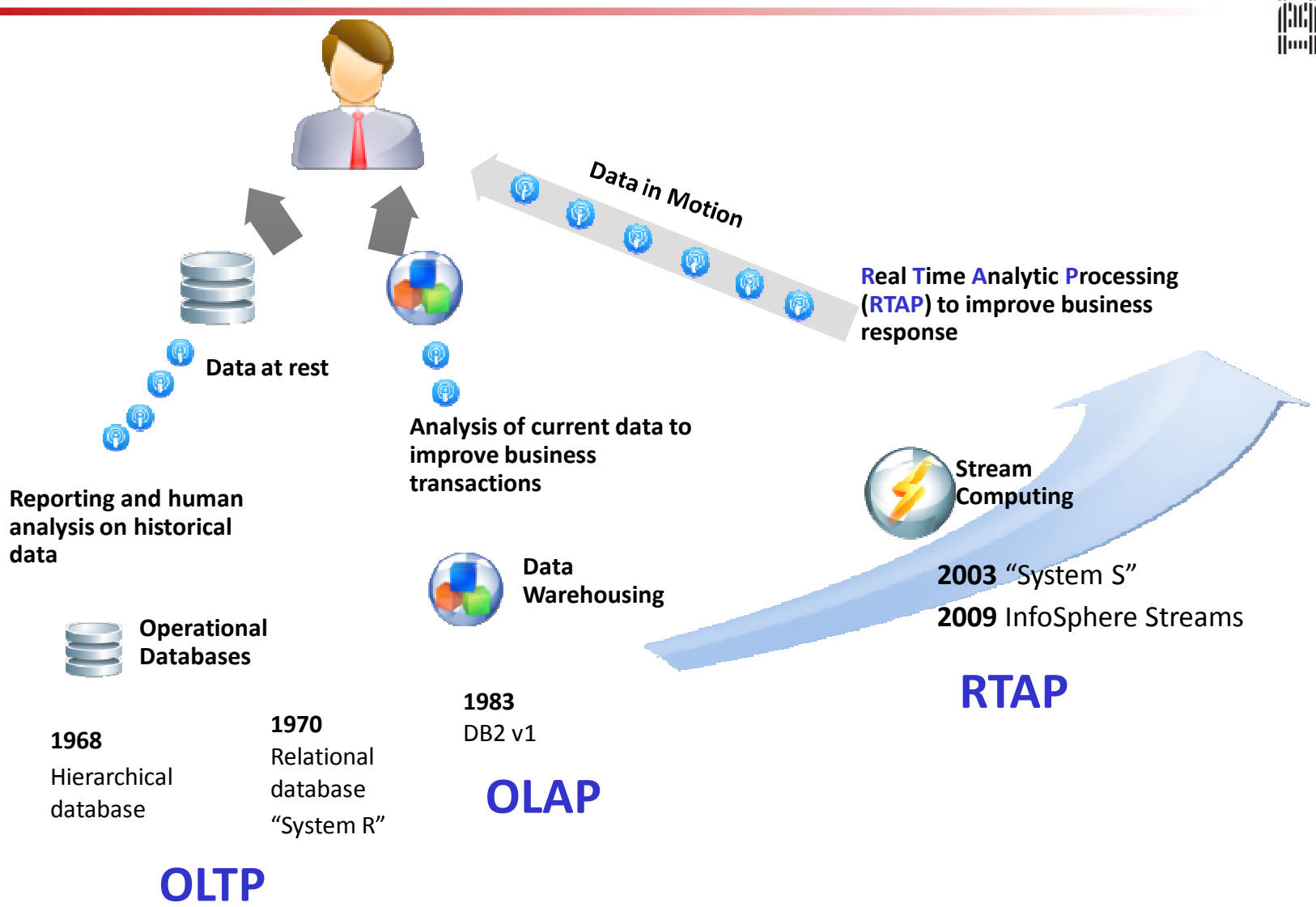
44x as much Data and Content Over Coming Decade

80% Of world's data is unstructured

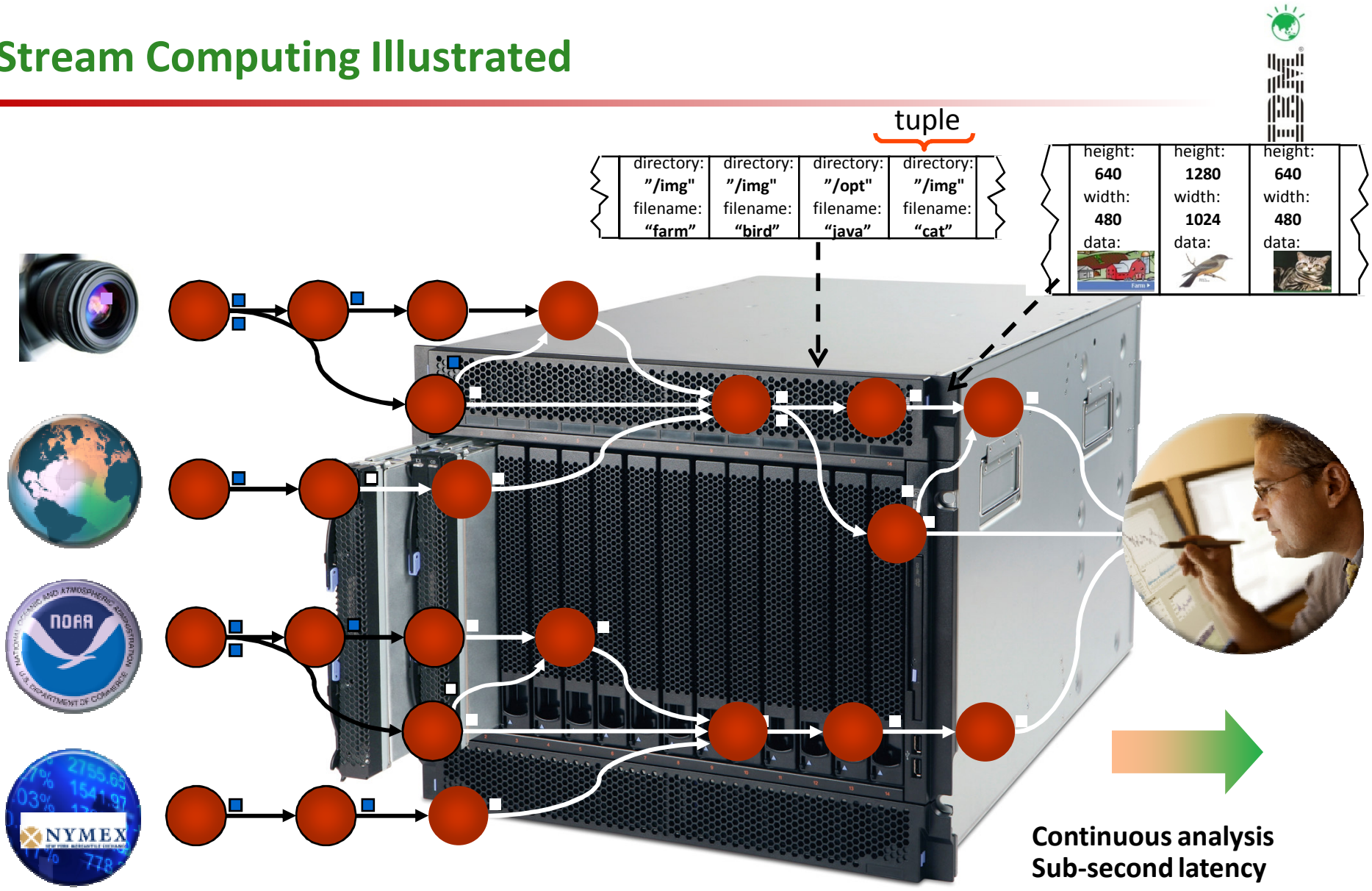


Source: IDC, The Digital Universe Decade – Are You Ready?, May 2010

InfoSphere Streams represents a paradigm shift



Stream Computing Illustrated



Streams delivers key capabilities:

Volume, Variety, Velocity, Analytics, and Agility

Language built for Streaming applications:

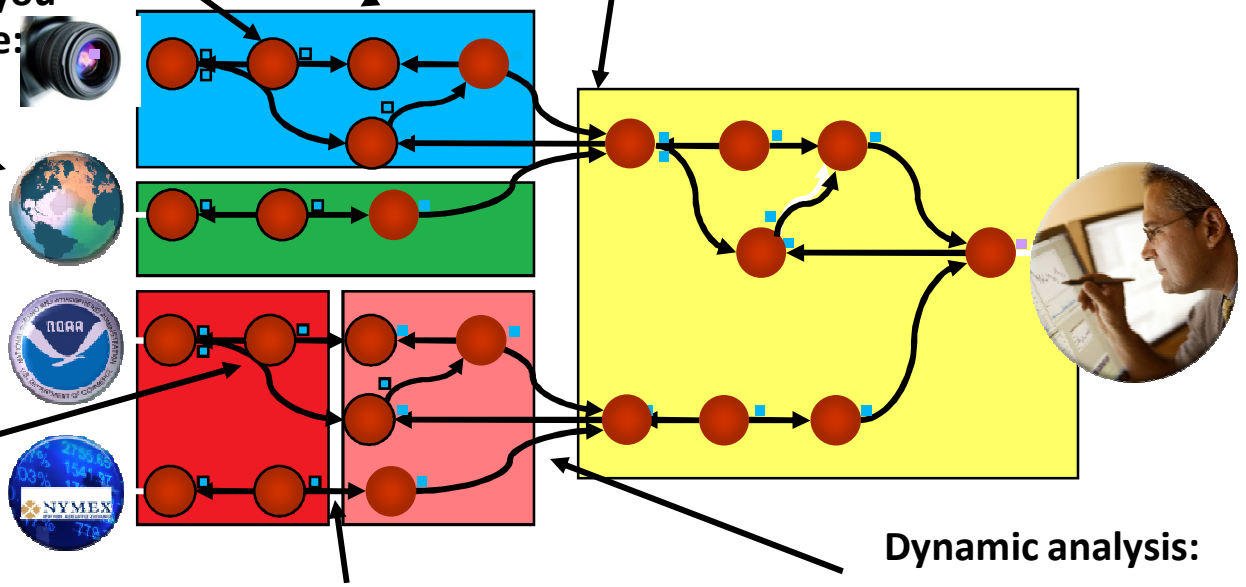
Reusable operators
Rapid application development
Continuous "pipeline" processing

Compile groups of operators into single processes:

Efficient use of cores
Distributed execution
Very fast data exchange
Can be automatic or tuned
Scaled with push of a button

Use the data that gives you a competitive advantage:

Can handle virtually any data type
Use data that is too expensive and time sensitive for traditional approaches



Easy to extend:

Built in adaptors
Users add capability with familiar C++ and Java

Dynamic analysis:

Programmatically change topology at runtime
Create new subscriptions
Create new port properties

Easy to manage:

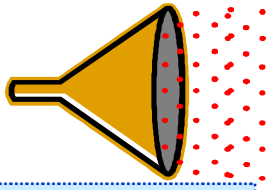
Automatic placement
Extend applications incrementally without downtime
Multi-user / multiple applications

Flexible and high performance transport:

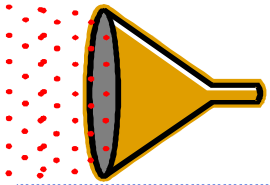
Very low latency
High data rates

Streams has many core analytic capabilities

Stream-Relational Built-in Operators



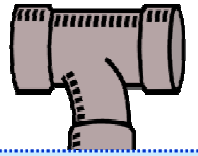
The **Split** operator is used for dividing incoming tuples into separate streams for parallel processing



The **Bundle** operator is used for merge many streams after parallel processing from a **Split** operator

The **Aggregate** operator is used for grouping and summarization of incoming tuples

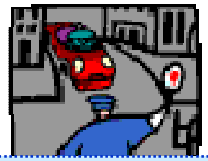
The **Functor** operator is used for performing tuple-level manipulations



The **Join** operator is used for correlating two streams

The **Punctor** operator is for inserting punctuation marks in streams

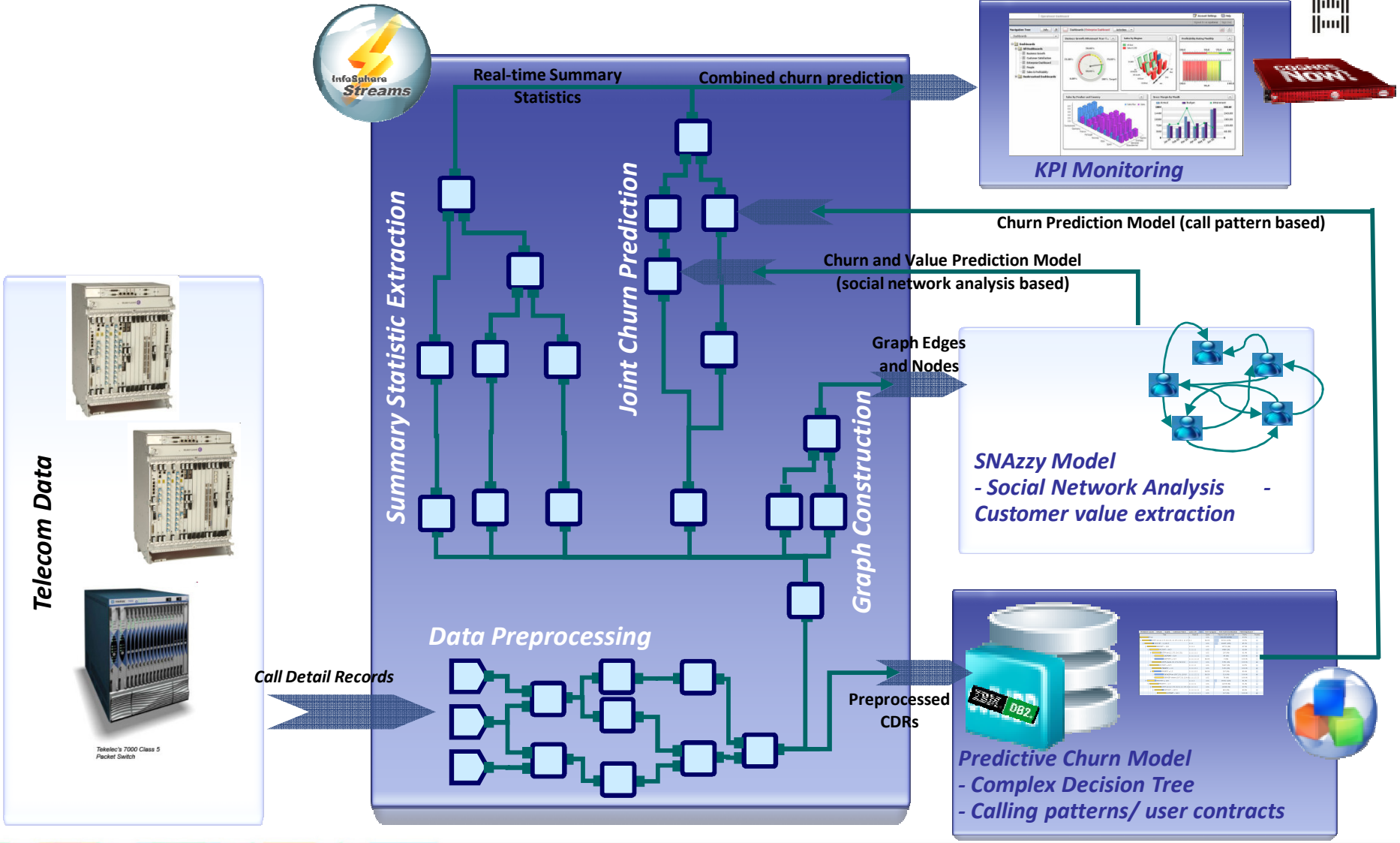
The **Sort** operator is used for imposing an order on incoming tuples in a stream



The **Barrier** operator is used as a synchronization point

The **Delay** operator is used to “artificially” slowdown a stream

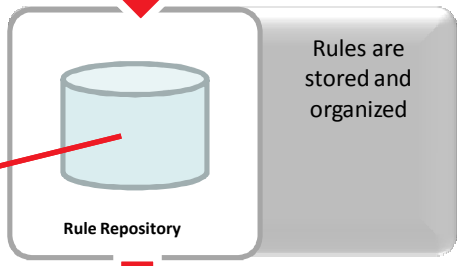
Example Application – CDR Churn Prediction

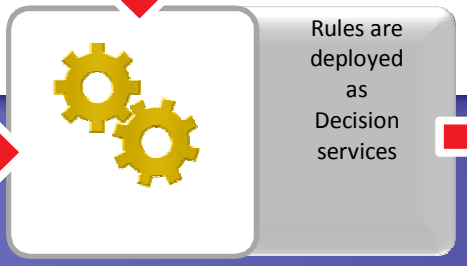


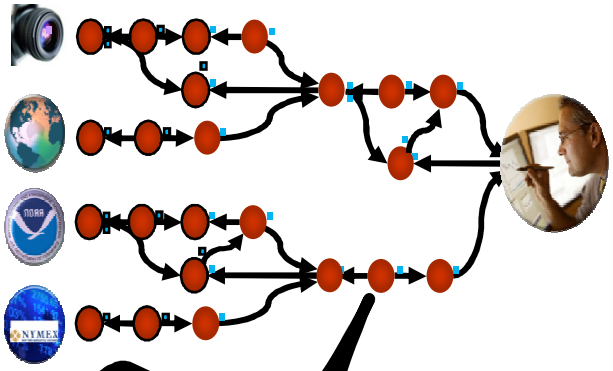
Business rules integration

WebSphere ILog -- A technology for creating, maintaining and deploying rule based decision services

```
if the yearly repayment of 'the loan' is more than
     the yearly income of 'the borrower' * 0.3
then
  add "Too big Debt-To-Income ratio" to the
     messages of 'the loan' ;
  reject 'the loan' ;
```







Is this customer eligible for a loan?

Incorporated into a Streams operator

No, for the following reason:
Too big Debt-To-Income ratio

Streams Mining Toolkit v1.2 Goals

- Data Mining has been a valuable technique for many years
  - Extracting relevant information or intelligence from large data sets
  - Traditionally done on stored data from some time in the past

- Sometimes, analysis of data *after events occur* has limited value:
  - Cyber security
  - Fraud detection
  - Market trade monitoring
  - Law enforcement
  - Disease detection

- Combines benefits of traditional mining of data at rest with streaming analysis of data in motion

Enables predictive and reactive business intelligence on time-sensitive data

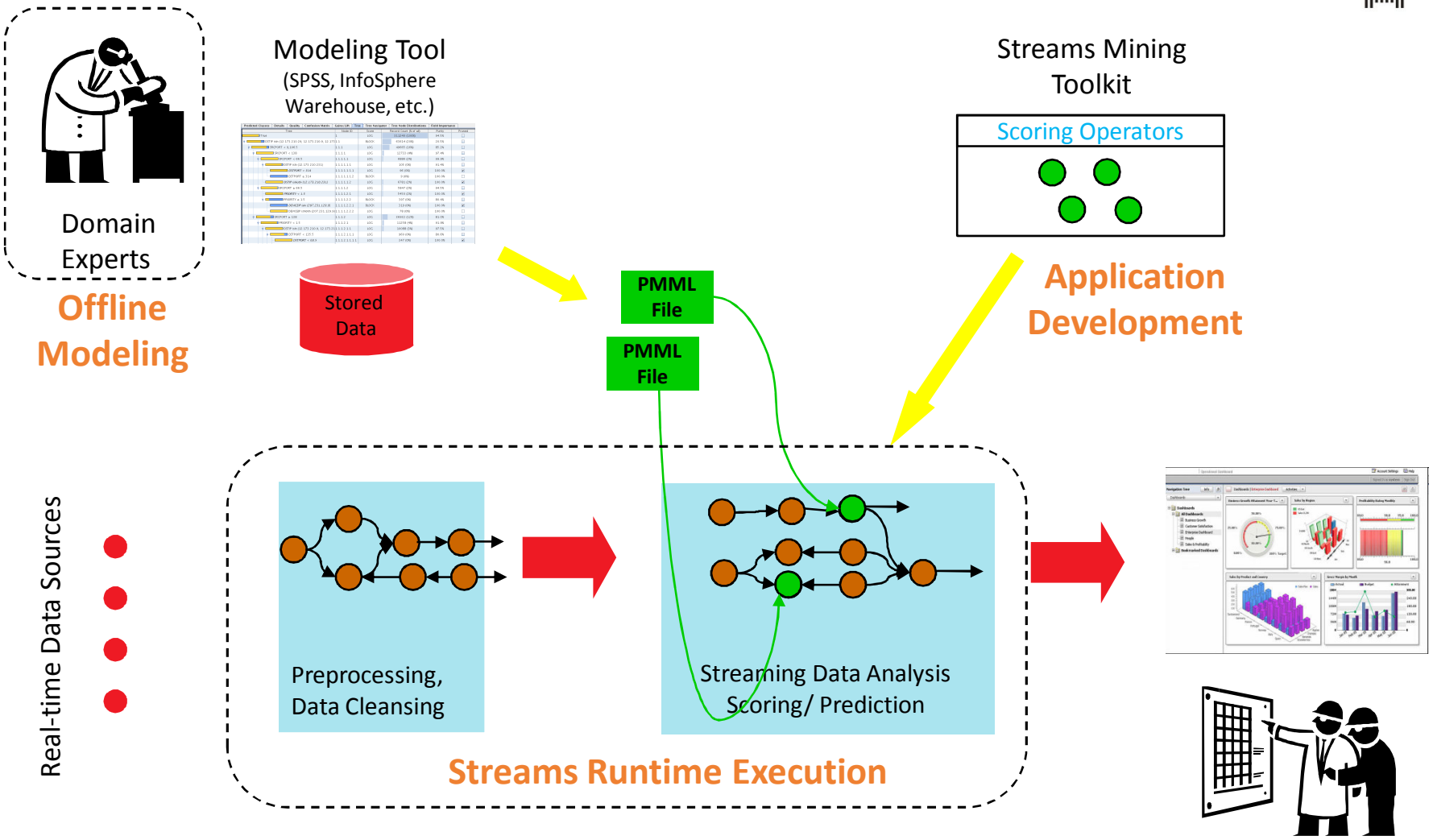
Streams Mining Toolkit v1.2

- Enables Mining in Microseconds in Streams applications
  - Scoring is performed against a predefined model
  - Supports a variety of model types and scoring algorithms

- Models described in Predictive Model Markup Language (PMML)
  - Standard for statistical and data mining models
  - XML Representation
  - http://www.dmg.org/

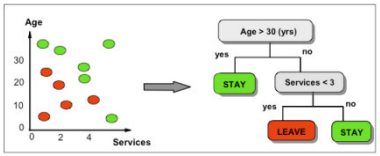
- Toolkit provides four new SPL operators to enable scoring
  - Classification
  - Clustering
  - Regression
  - Associations

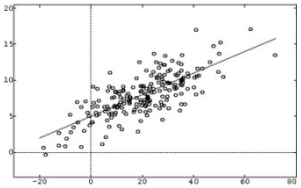
Basic Use Case – Real-time Scoring

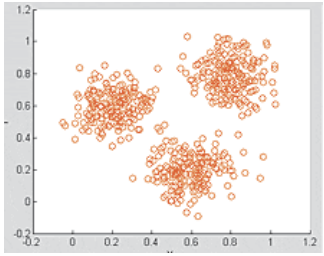


Mining algorithms supported

Algorithm Type	Algorithm	Supported PMML Versions
Classification	Decision Tree	2.0 - 3.0
	Logistic Regression	2.0 - 3.2
	Naïve Bayes	2.0 - 3.2
Regression	Linear Regression	2.0 - 3.0
	Polynomial Regression	2.0 - 3.0
	Transform Regression	2.0 - 3.0
Clustering	Demographic Clustering	2.0 - 3.0
	Kohonen Clustering	2.0 - 3.0
Associations	Association Rules	2.0 - 3.2

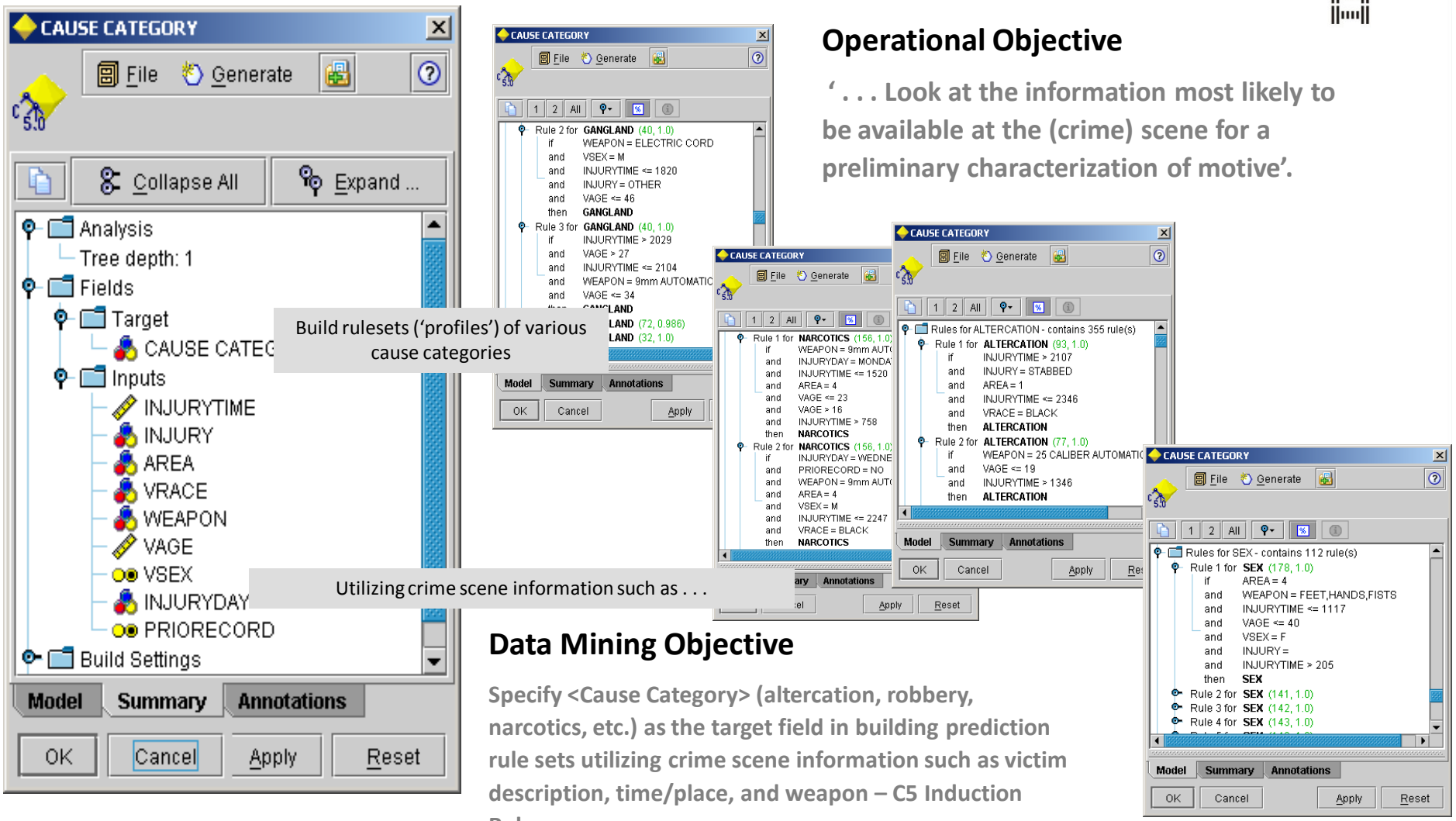






Data Mining & Crime - Rule Induction Model with SPSS

Characterization of 'Motive'

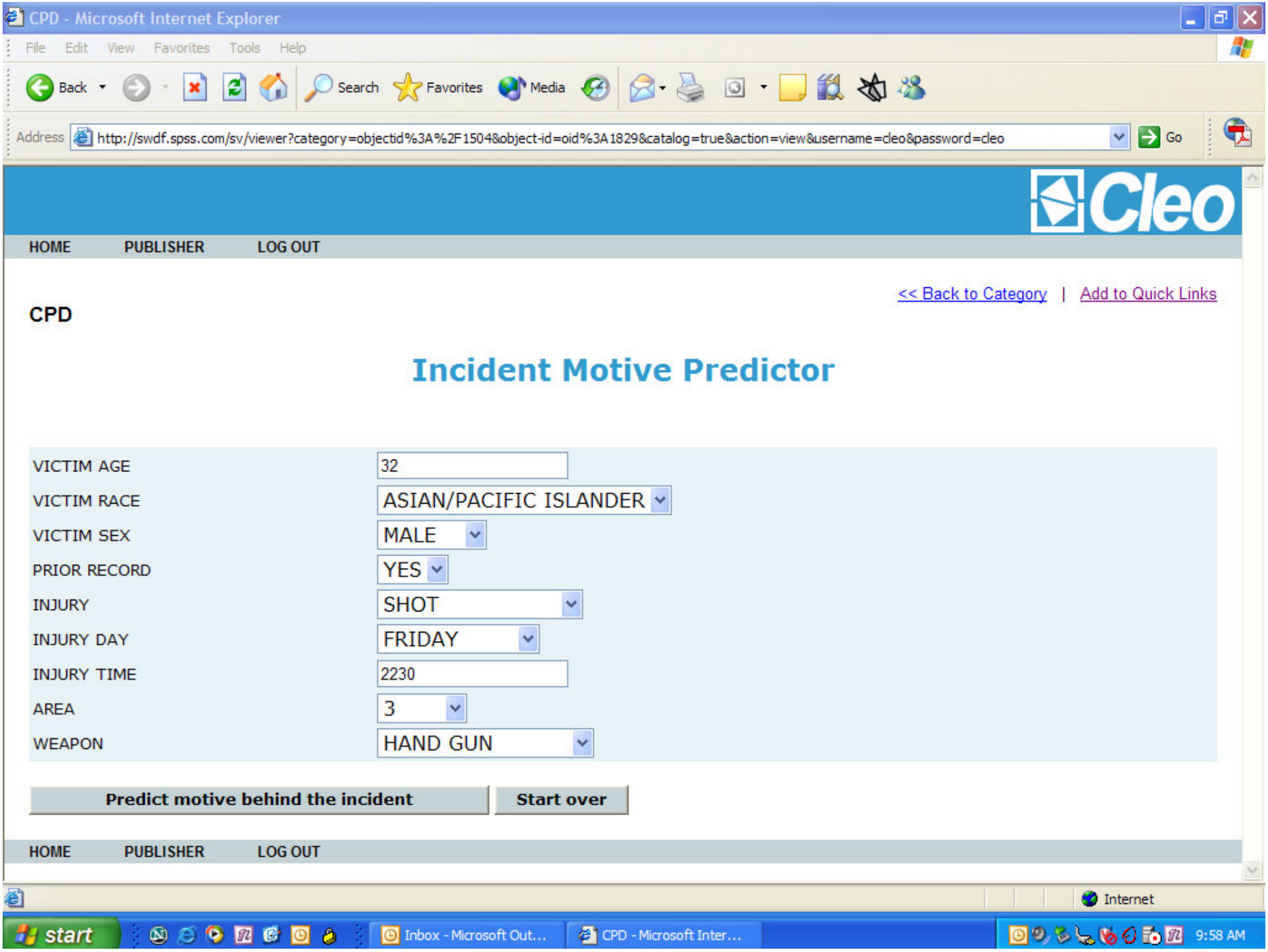


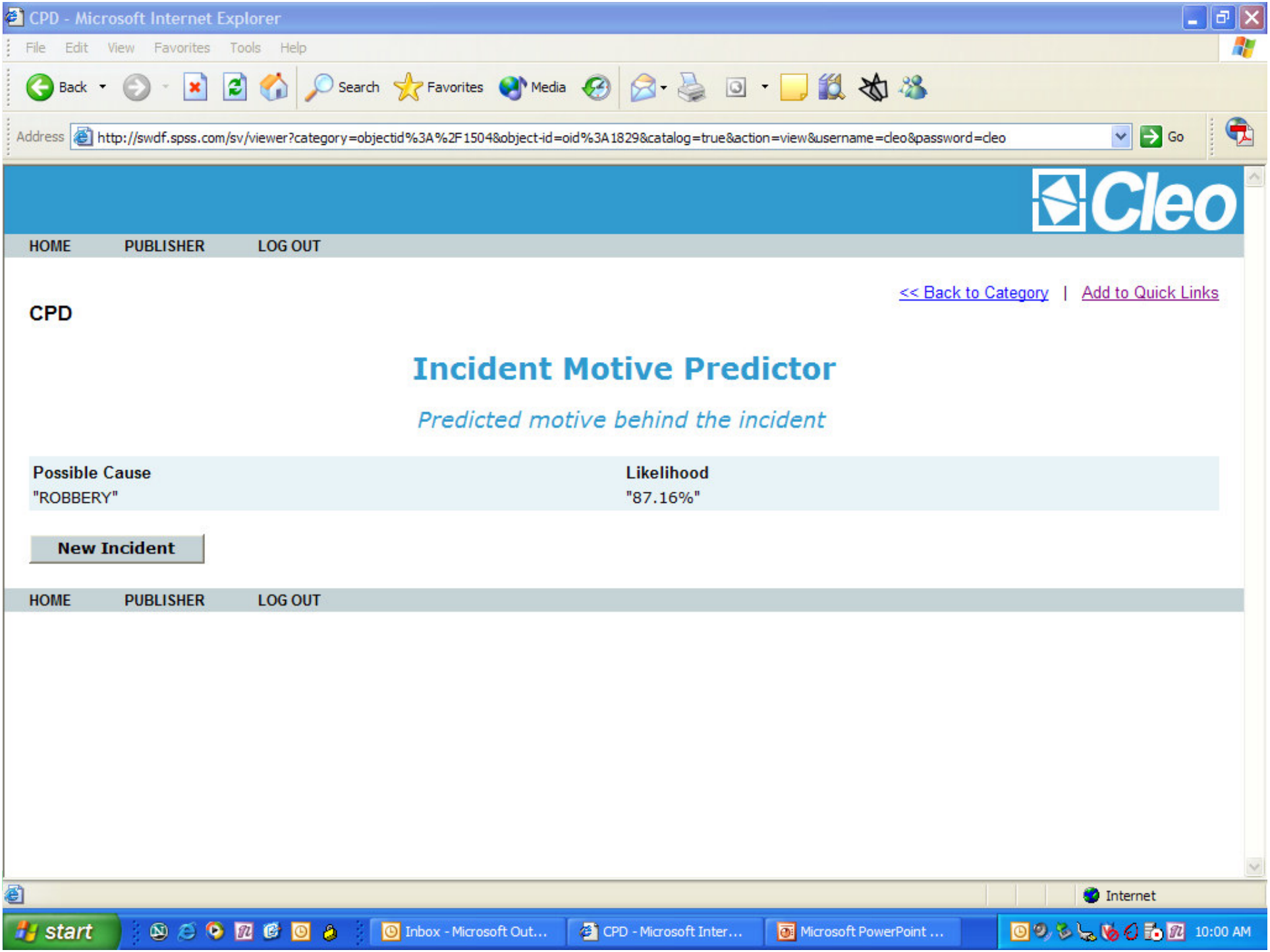
Operational Objective

'. . . Look at the information most likely to be available at the (crime) scene for a preliminary characterization of motive'.

Data Mining Objective

Specify <Cause Category> (altercation, robbery, narcotics, etc.) as the target field in building prediction rule sets utilizing crime scene information such as victim description, time/place, and weapon – C5 Induction Rule


CPD - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://swdf.spss.com/sv/viewer?category=objectid%3A%2F1504&object-id=oid%3A1829&catalog=true&action=view&username=cleo&password=cleo
Go
Cleo
HOME PUBLISHER LOG OUT
<< Back to Category | Add to Quick Links
CPD
Incident Motive Predictor
VICTIM AGE 32
VICTIM RACE ASIAN/PACIFIC ISLANDER
VICTIM SEX MALE
PRIOR RECORD YES
INJURY SHOT
INJURY DAY FRIDAY
INJURY TIME 2230
AREA 3
WEAPON HAND GUN
Predict motive behind the incident
Start over
HOME PUBLISHER LOG OUT
Internet
start
Inbox - Microsoft Out...
CPD - Microsoft Inter...
9:58 AM


CPD - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://swdf.spss.com/sv/viewer?category=objectid%3A%2F1504&object-id=oid%3A1829&catalog=true&action=view&username=cleo&password=cleo
Go
Cleo
HOME PUBLISHER LOG OUT
<< Back to Category | Add to Quick Links
CPD
Incident Motive Predictor
Predicted motive behind the incident
Possible Cause
"ROBBERY"
Likelihood
"87.16%"
New Incident
HOME PUBLISHER LOG OUT
Internet
start
Inbox - Microsoft Out...
CPD - Microsoft Inter...
Microsoft PowerPoint ...
10:00 AM

Accelerating Situational Analysis
Scoring in Streams

Data to be Scored

/home/koranda/homicide/homicide data - unsolved cases.txt - gedit

File Edit View Search Tools Documents Help

New Open Save Print Undo Redo Cut Copy Paste Find Replace

homicide data - unsolved cases.txt

```
victim age|victim race|victim gender|victim gang|victim prior record|injury|injury date|injury day of week|injury time|case cleared|beat|
district|area|location|weapon|weapon recovered|narcotics|intoxicant
25|"White-Hispanic"|"male"|"Black Stones"|"Yes"|"Shot"||"Monday"|1740|"Yes"|"1311"|"13"|"4"|"2400"|"38 Caliber Revolver"|"No"|"No"|"No"
19|"Black"|"male"|"Unknown"|"Yes"|"Shot"||"Monday"|25|"Yes"|"512"|"5"|"2"|"2100"|"Unknown Caliber"|"No"|"Yes"|"Yes"
44|"Black"|"female"|"Unknown"|"No"|"Stabbed"||"Friday"|200|"Yes"|"2531"|"25"|"5"|"1101"|"Carving Knife"|"No"|"No"|"Yes"
18|"Black"|"male"|"Red Devils"|"Yes"|"Shot"||"Tuesday"|1622|"Yes"|"1111"|"11"|"4"|"2400"|"7.62 MM"|"No"|"No"|"No"
18|"Black"|"male"|"Unknown"|"Yes"|"Shot"||"Saturday"|2045|"Yes"|"221"|"2"|"1"|"1312"|"9 MM Automatic"|"Yes"|"No"|"No"
57|"White"|"male"|"Unknown"|"No"|"Shot"||"Monday"|2330|"Yes"|"1614"|"16"|"5"|"1307"|"357 Magnum Revolver"|"No"|"No"|"No"
21|"White-Hispanic"|"male"|"Outlaws"|"Yes"|"Shot"||"Sunday"|350|"Yes"|"434"|"4"|"2"|"2100"|"9 MM Automatic"|"No"|"No"|"No"
23|"White-Hispanic"|"male"|"None"|"No"|"Shot"||"Wednesday"|2207|"Yes"|"912"|"9"|"1"|"2200"|"9 MM Automatic"|"No"|"Yes"|"Yes"
55|"Black"|"male"|"Unknown"|"Yes"|"Stabbed"||"Tuesday"|1930|"Yes"|"934"|"9"|"1"|"1101"|"Kitchen Knife"|"No"|"No"|"No"
31|"Black"|"male"|"Gangster Disciples"|"Yes"|"Shot"||"Thursday"|2230|"Yes"|"221"|"2"|"1"|"2100"|"38 Caliber Automatic"|"No"|"No"|"No"
30|"Black"|"male"|"Four Corner Hustlers"|"Yes"|"Shot"||"Tuesday"|2335|"Yes"|"1121"|"11"|"4"|"2100"|"40 Caliber Automatic"|"No"|"No"|"No"
22|"White-Hispanic"|"male"|"Unknown"|"Yes"|"Shot"||"Wednesday"|327|"Yes"|"1223"|"12"|"4"|"2200"|"38 Caliber Revolver"|"Yes"|"No"|"No"
26|"Black"|"male"|"Unknown"|"Yes"|"Stabbed"||"Thursday"|135|"Yes"|"1532"|"15"|"5"|"2100"|"Carving Knife"|"No"|"No"|"No"
20|"Black"|"male"|"Unknown"|"Yes"|"Shot"||"Sunday"|2243|"Yes"|"824"|"8"|"1"|"2100"|"25 Caliber Automatic"|"No"|"No"|"No"
48|"Black"|"female"|"Unknown"|"No"|"Other"||"Sunday"|630|"Yes"|"2433"|"24"|"3"|"1101"|"Feet Hands Fists"|"Yes"|"No"|"No"
55|"Black"|"male"|"Unknown"|"No"|"Stabbed"||"Saturday"|1725|"Yes"|"224"|"2"|"1"|"2100"|"Kitchen Knife"|"Yes"|"No"|"No"
29|"Black"|"male"|"Joe Boys"|"Yes"|"Shot"||"Tuesday"|50|"Yes"|"931"|"9"|"1"|"2400"|"9 MM Automatic"|"No"|"No"|"No"
37|"White"|"male"|"Unknown"|"Yes"|"Shot"||"Friday"|2350|"Yes"|"1931"|"19"|"3"|"2100"|"Unknown Caliber"|"No"|"No"|"Unknown"
22|"Black"|"male"|"Unknown"|"Yes"|"Shot"||"Saturday"|2207|"Yes"|"1531"|"15"|"5"|"2400"|"9 MM Automatic"|"No"|"No"|"Yes"
38|"White"|"male"|"Unknown"|"Yes"|"Assault"||"Sunday"|522|"Yes"|"1624"|"16"|"5"|"1101"|"Unknown"|"No"|"No"|"Unknown"
42|"Black"|"male"|"Gangster Disciples"|"Yes"|"Shot"||"Saturday"|2051|"Yes"|"1012"|"10"|"4"|"3002"|"Unknown"|"No"|"No"|"Unknown"
24|"Black"|"male"|"TAP Boys"|"Yes"|"Shot"||"Friday"|1805|"Yes"|"1011"|"10"|"4"|"2100"|"Unknown"|"No"|"Yes"|"Unknown"
26|"Black"|"male"|"TAP Boys"|"No"|"Shot"||"Tuesday"|101|"Yes"|"1123"|"11"|"4"|"2100"|"32 Caliber Revolver"|"Yes"|"No"|"Unknown"
21|"Black"|"male"|"Unknown"|"No"|"Shot"||"Tuesday"|2209|"Yes"|"412"|"4"|"2"|"2400"|"Unknown"|"No"|"No"|"Unknown"
45|"Black"|"male"|"Unknown"|"Yes"|"Assault"||"Friday"|1955|"Yes"|"2234"|"22"|"2"|"1107"|"Baseball Bat"|"Yes"|"No"|"Unknown"
62|"Black"|"female"|"Unknown"|"No"|"Stabbed"||"Saturday"|1615|"Yes"|"2124"|"21"|"1"|"1101"|"Unknown"|"No"|"No"|"Unknown"
29|"Black"|"male"|"Sydney Ducks"|"Yes"|"Shot"||"Wednesday"|1545|"Yes"|"1122"|"11"|"4"|"2100"|"40 Caliber Automatic"|"No"|"No"|"Unknown"
20|"White-Hispanic"|"male"|"La Raza"|"No"|"Shot"||"Sunday"|46|"Yes"|"933"|"9"|"1"|"2100"|"357 Magnum Revolver"|"Yes"|"No"|"Unknown"
28|"Black"|"male"|"Gangster Disciples"|"Yes"|"Shot"||"Wednesday"|2320|"Yes"|"231"|"2"|"1"|"2100"|"Unknown"|"No"|"Yes"|"Yes"
16|"White-Hispanic"|"male"|"Unknown"|"Yes"|"Shot"||"Monday"|1447|"Yes"|"2431"|"24"|"3"|"2100"|"9 MM Automatic"|"Yes"|"No"|"Unknown"
```

Ln 1, Col 1 INS

PMML Model that predicts Cause Category

C:\Distillery\Commercialization\SPSS_POC\homicides\homicides\C5 model - cause category.xml - Microsoft Internet Explorer

File Edit View Favorites Tools Help

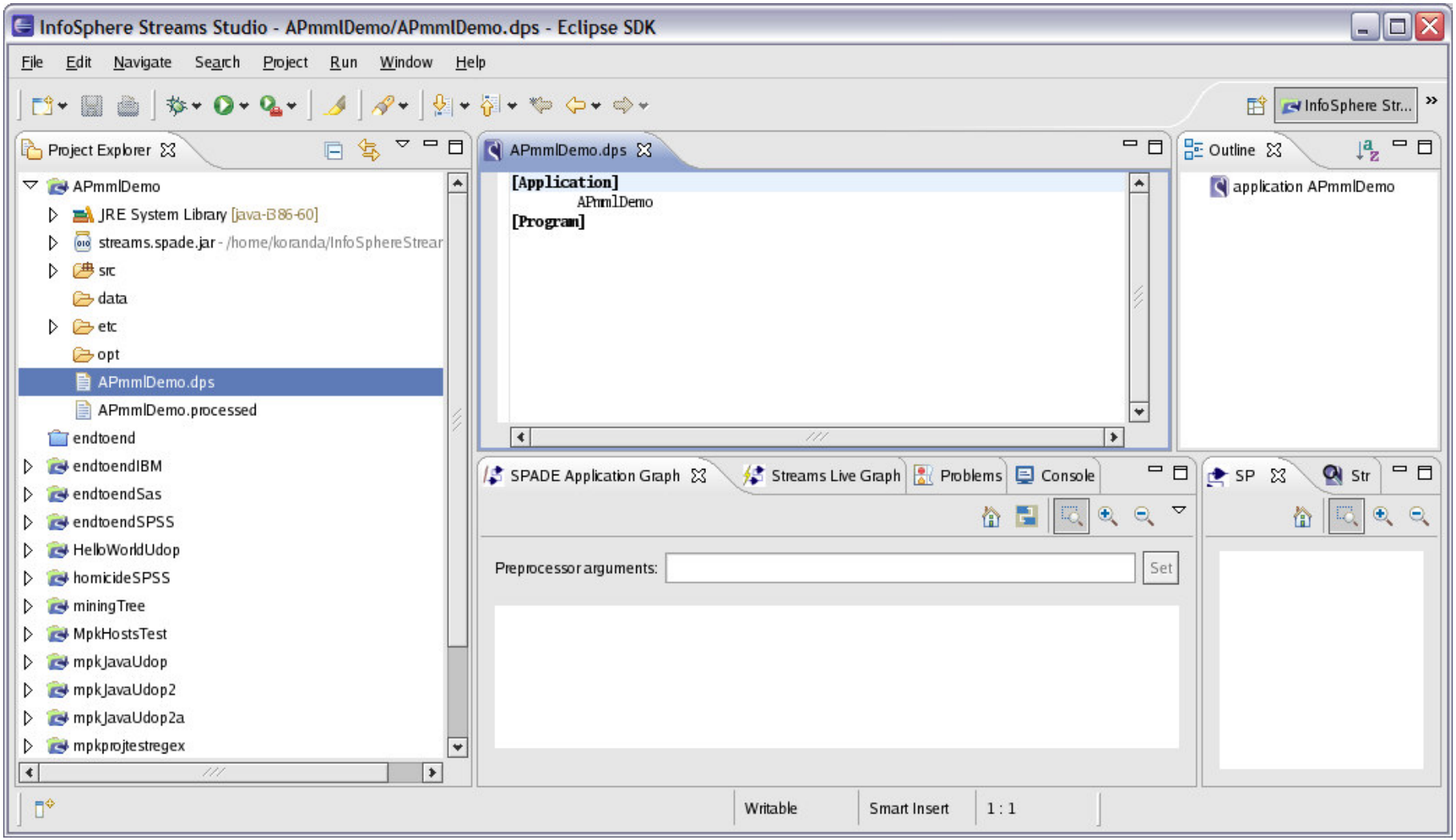
Back Search Favorites

Address C:\Distillery\Commercialization\SPSS_POC\homicides\homicides\C5 model - cause category.xml Go

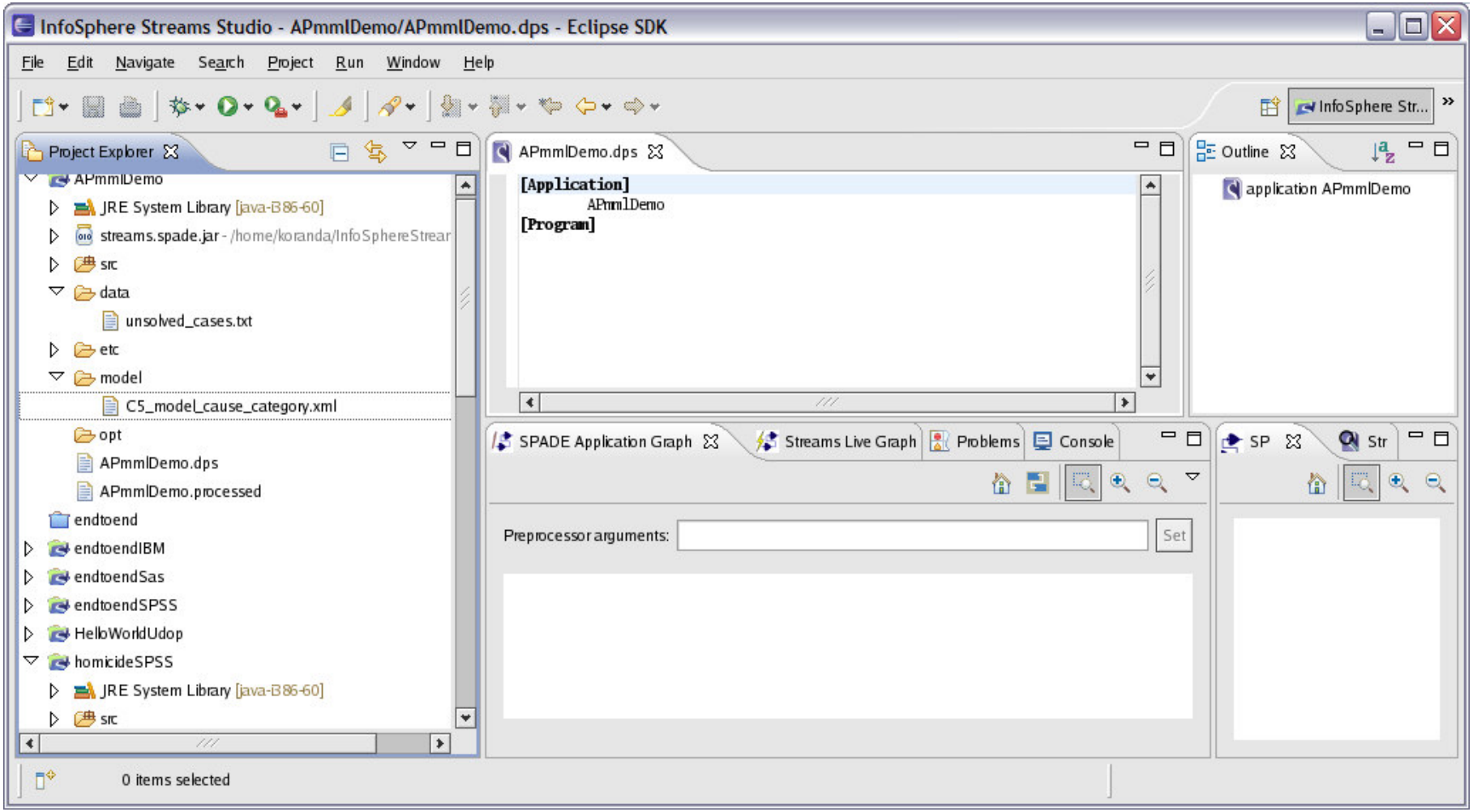
To help protect your security, Internet Explorer has restricted this file from showing active content that could access your computer. Click here for options...

```
<?xml version="1.0" encoding="UTF-8" ?>
- <PMML xmlns="http://www.dmg.org/PMML-3_2" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" version="3.2">
 - <Header copyright="Copyright (c) Integral Solutions Ltd., 1994 - 2009. All rights reserved.">
    <Application name="PASW Modeler" version="13" />
   </Header>
 - <DataDictionary numberOfFields="17">
  - <DataField dataType="string" name="narcotics" optype="categorical">
     <Extension extender="spss" name="storageType" value="string" />
     <Value property="valid" value="No" />
     <Value property="valid" value="Unknown" />
     <Value property="valid" value="Yes" />
    </DataField>
  - <DataField dataType="string" name="injury" optype="categorical">
     <Extension extender="spss" name="storageType" value="string" />
     <Value property="valid" value="Assault" />
     <Value property="valid" value="Other" />
     <Value property="valid" value="Shot" />
     <Value property="valid" value="Stabbed" />
     <Value property="valid" value="Strangulation" />
     <Value property="valid" value="Unknown" />
    </DataField>
  - <DataField dataType="integer" name="area" optype="continuous">
     <Extension extender="spss" name="storageType" value="numeric" />
    </DataField>
  - <DataField dataType="integer" name="injury time" optype="continuous">
     <Extension extender="spss" name="storageType" value="numeric" />
    </DataField>
  - <DataField dataType="integer" name="location" optype="continuous">
     <Extension extender="spss" name="storageType" value="numeric" />
    </DataField>
  - <DataField dataType="integer" name="victim age" optype="continuous">
     <Extension extender="spss" name="storageType" value="numeric" />
    </DataField>
  - <DataField dataType="integer" name="beat" optype="continuous">
     <Extension extender="spss" name="storageType" value="numeric" />
    </DataField>
```

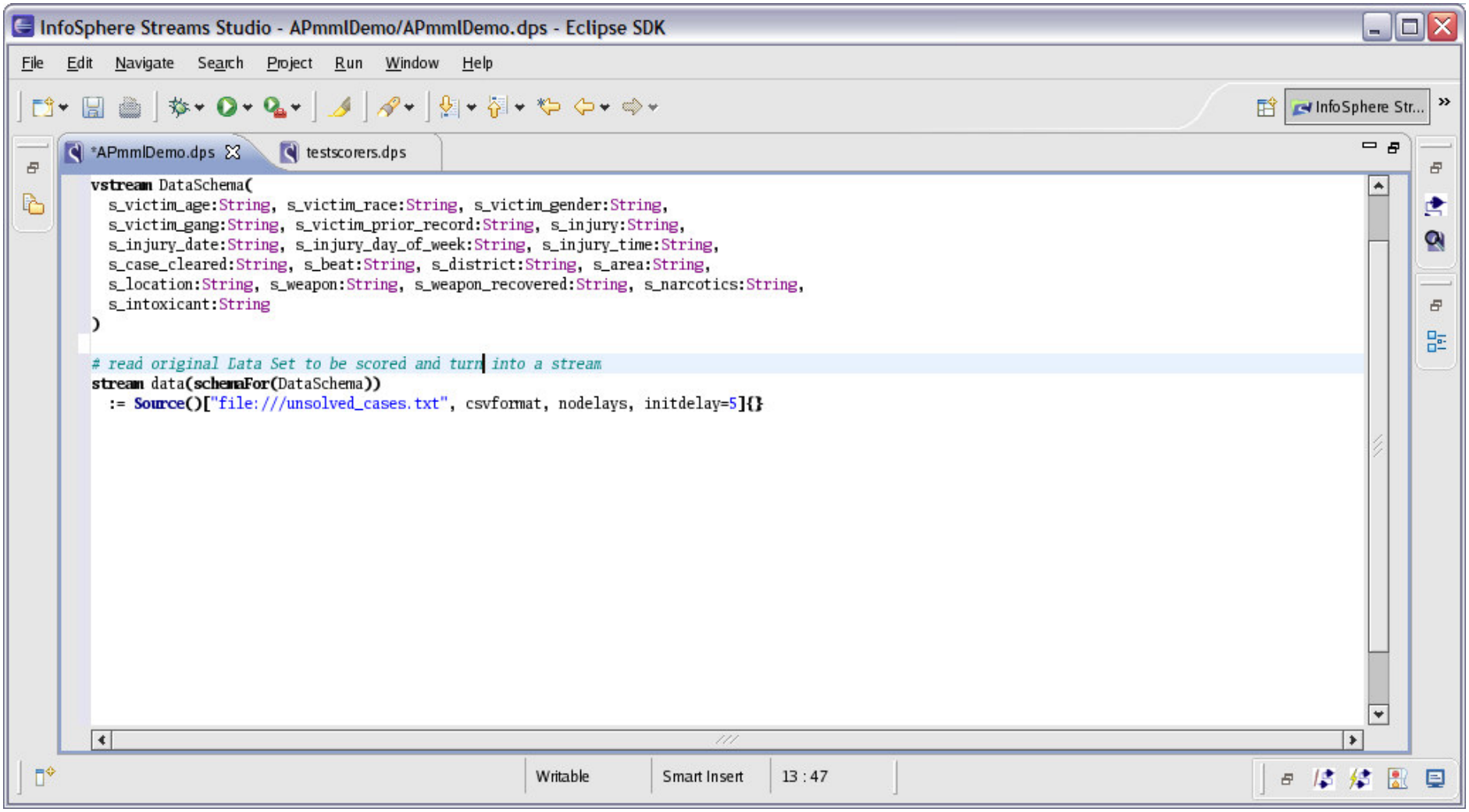
Scoring walkthrough
Create a new blank Streams project



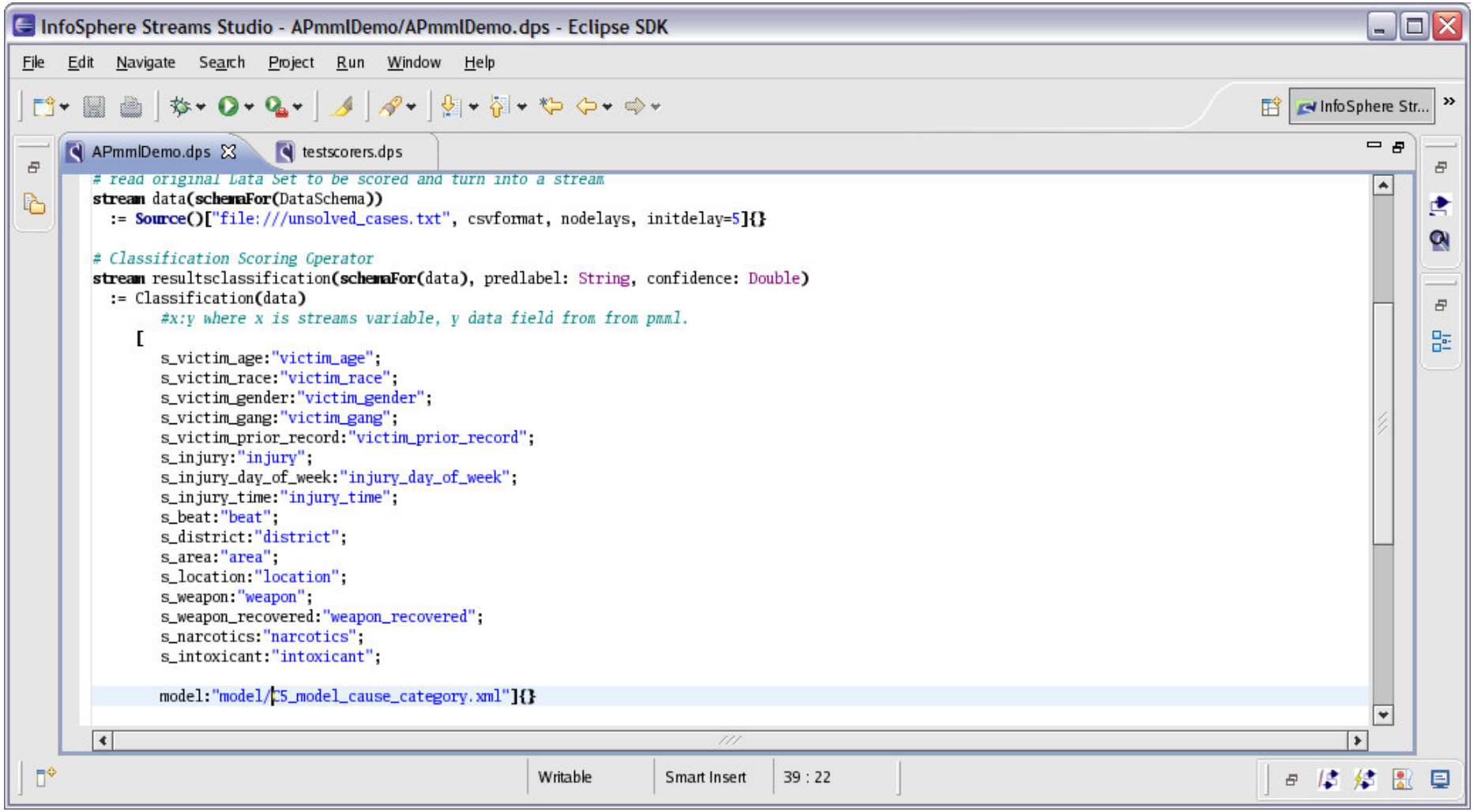
Copy model and data file into the project



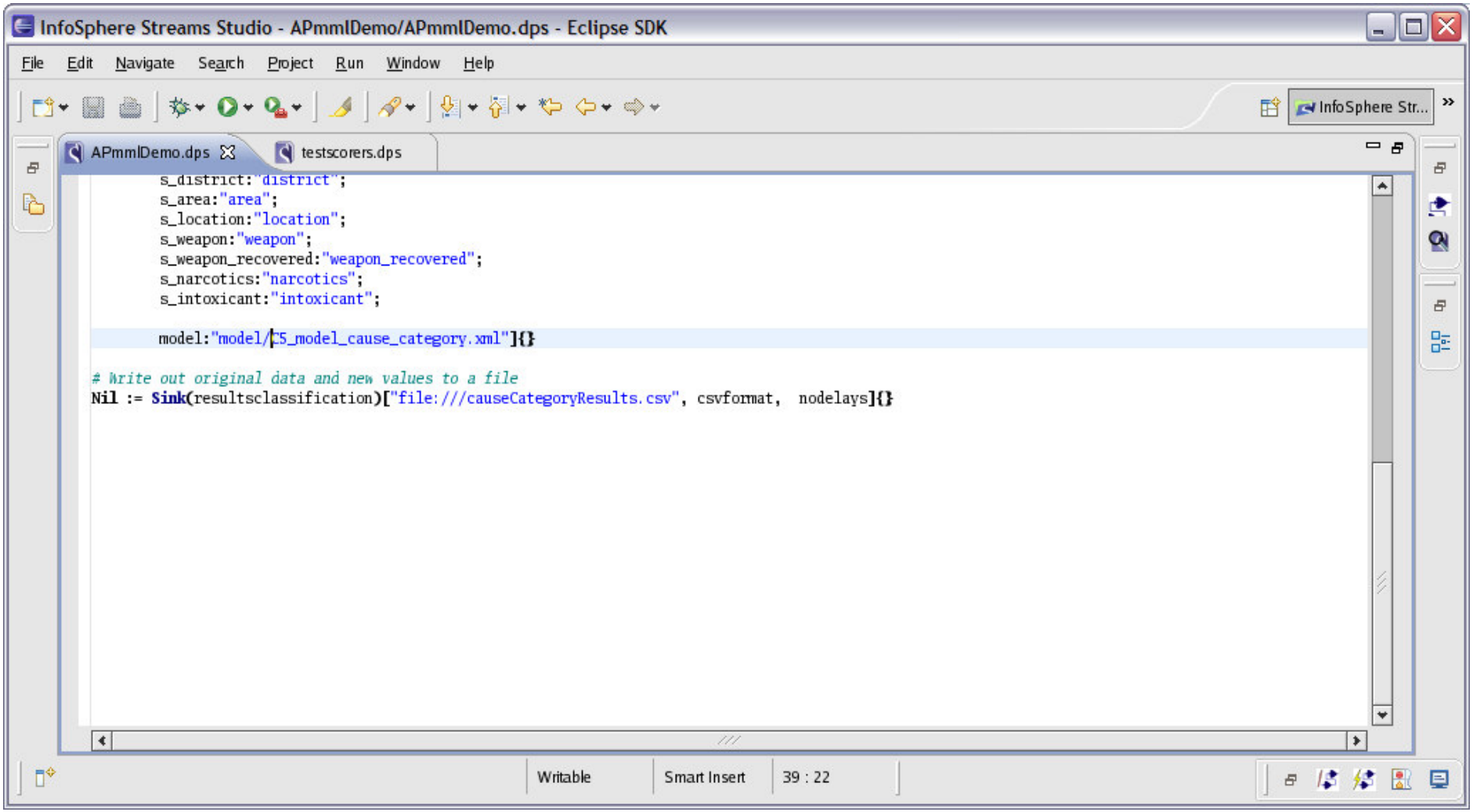
Create the source operator that reads from a file and produces a stream named “data”



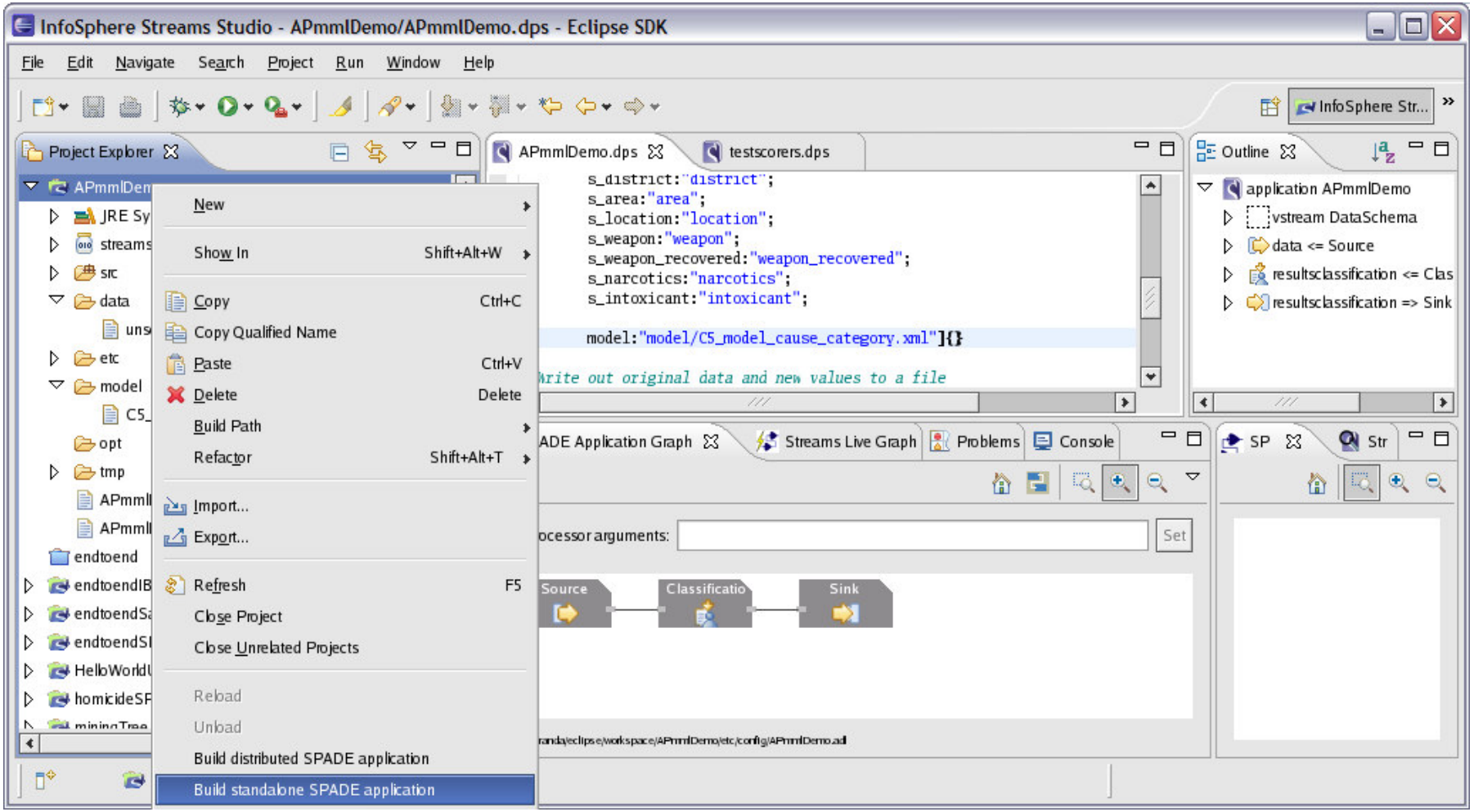
Use Classification operator to process each tuple to produce new output stream called “resultclassification”



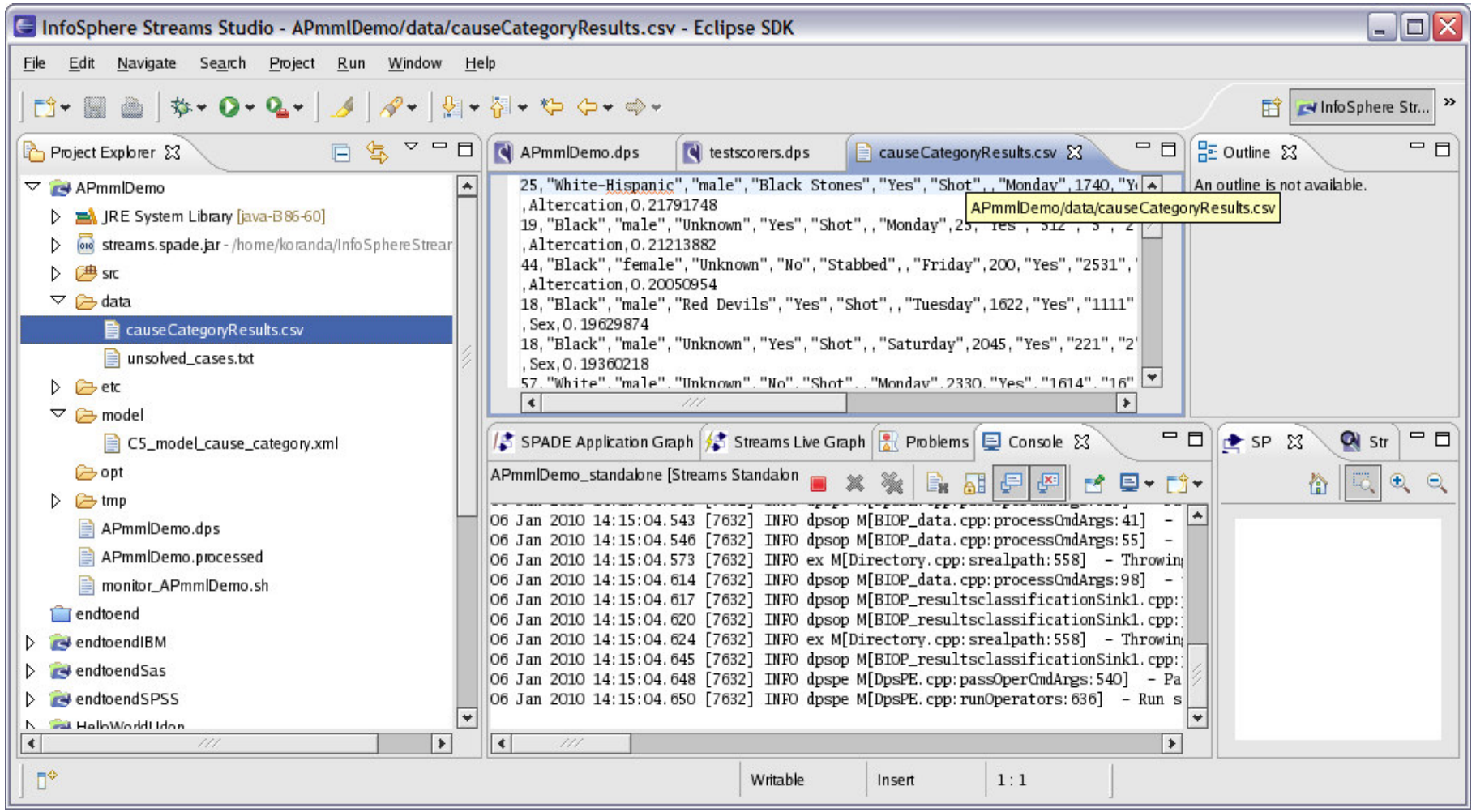
Create the sink operator that writes the resultclassification stream to a file



Compile the program



Run the program and review the results file -- populated with the processed data



System S Research Analytic Assets

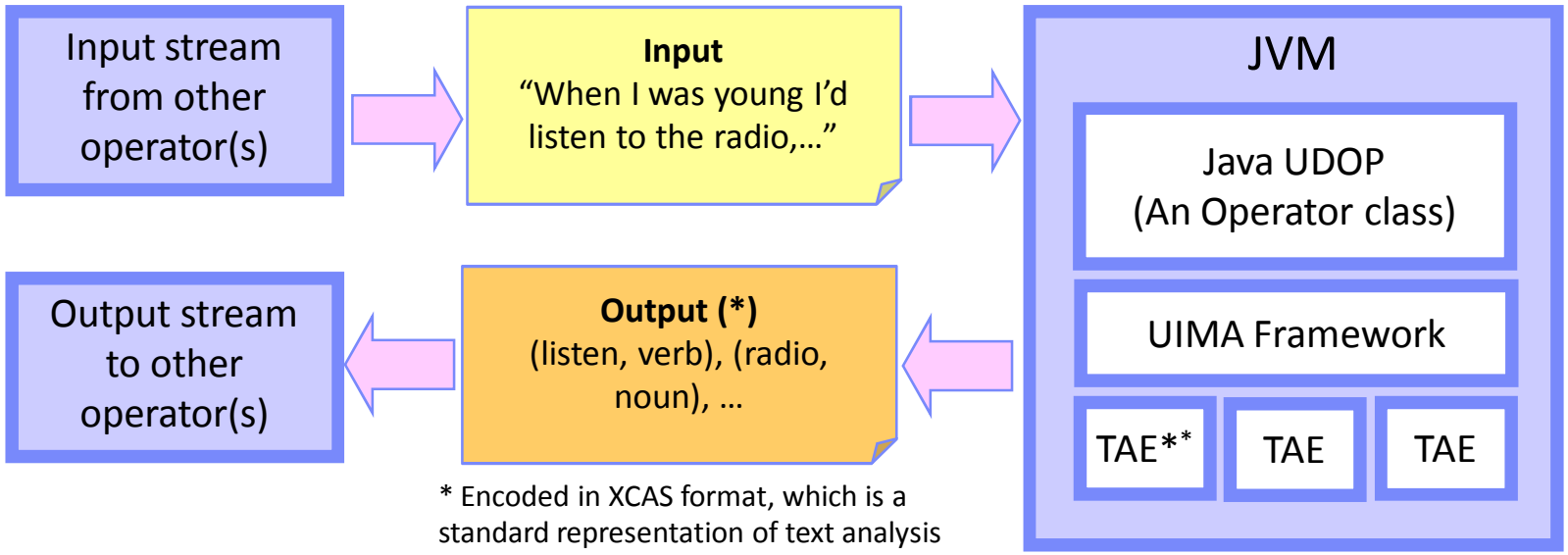
- Automated Analytics Composer
- Workload Generator
- Streams Spatial Extender
- Streams Graph assets
- Time Series Analysis assets
- Telco xDR Hub assets
- Artemis (healthcare) assets
- Market Surveillance Solution
- Intelligent Transportation Application
- Rules on Streams

Research Assets are available outside of IBM only via Technology Evaluation or Services agreements

26

Text analytics used with Streams

- GATE adapter: General Architecture for Text Engineering
- (IBM) UIMA adapter: Unstructured Information Management Architecture
- IBM GNM adapter: Global Name Management
- IBM System T adapter: Fast text analytics
- IBM TALES adapter: Multi-lingual, multi-modal foreign news media monitoring system

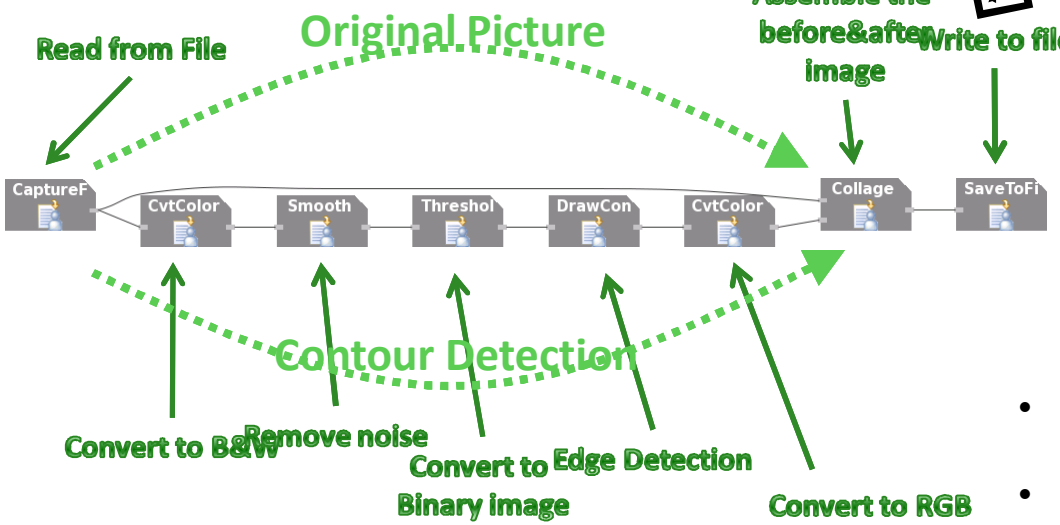


* Encoded in XCAS format, which is a standard representation of text analysis output in UIMA

**TAE:Text Analysis Engine

Video analytics

Example – Contour Detection




```
[Application]
contours

[Program]

vstream IplImage(channels: Integer,
                 depth: Integer,
                 origin: Integer,
                 width: Integer,
                 height: Integer,
                 data: ByteList)

stream vid(schemaFor(IplImage))
  := CaptureFromFile( ) [file: "$ENV{HOME}/demo3.m4v"; repeat:1] { }
    -> partition["P1"]

stream bw_vid(schemaFor(vid))
  := CvtColor(vid) [ ] { data := ~CV_BGR2GRAY() }
    -> partition["P1"]

stream smooth_bw_vid(schemaFor(vid))
  := Smooth(bw_vid) [iteration: 4] { }
    -> partition["P1"]

stream th_vid(schemaFor(vid))
  := Threshold(smooth_bw_vid) [max:255; threshold:125] { data := ~CV_THRESH_TOZERO( ) }
    -> partition["P1"]

stream cntr(schemaFor(vid))
  := DrawContours(th_vid) [ ] { }
    -> partition["P1"]

stream cntr_rgb(schemaFor(vid))
  := CvtColor(cntr) [channels:3] { data := ~CV_GRAY2BGR( ) }
    -> partition["P1"]

stream src_n_cntr(schemaFor(vid))
 := Collage(vid; cntr_rgb) [ ] { }
    -> partition["P1"]

Nil
  := SaveToFile(src_n_cntr) [filename: "$ENV{PWD}/demo3-out.mp4"; rate:15; fourcc:"fmp4"] { }
   -> partition["P1"]
```

- **Use B&W+threshold pictures to compute derivatives of pixels**
- **Used as a first step for other more sophisticated processing**
- **Very low overhead from Streams – pass 200-300 fps per core, once analysis is done, processing overhead is high but can get 30fps on 8 core**

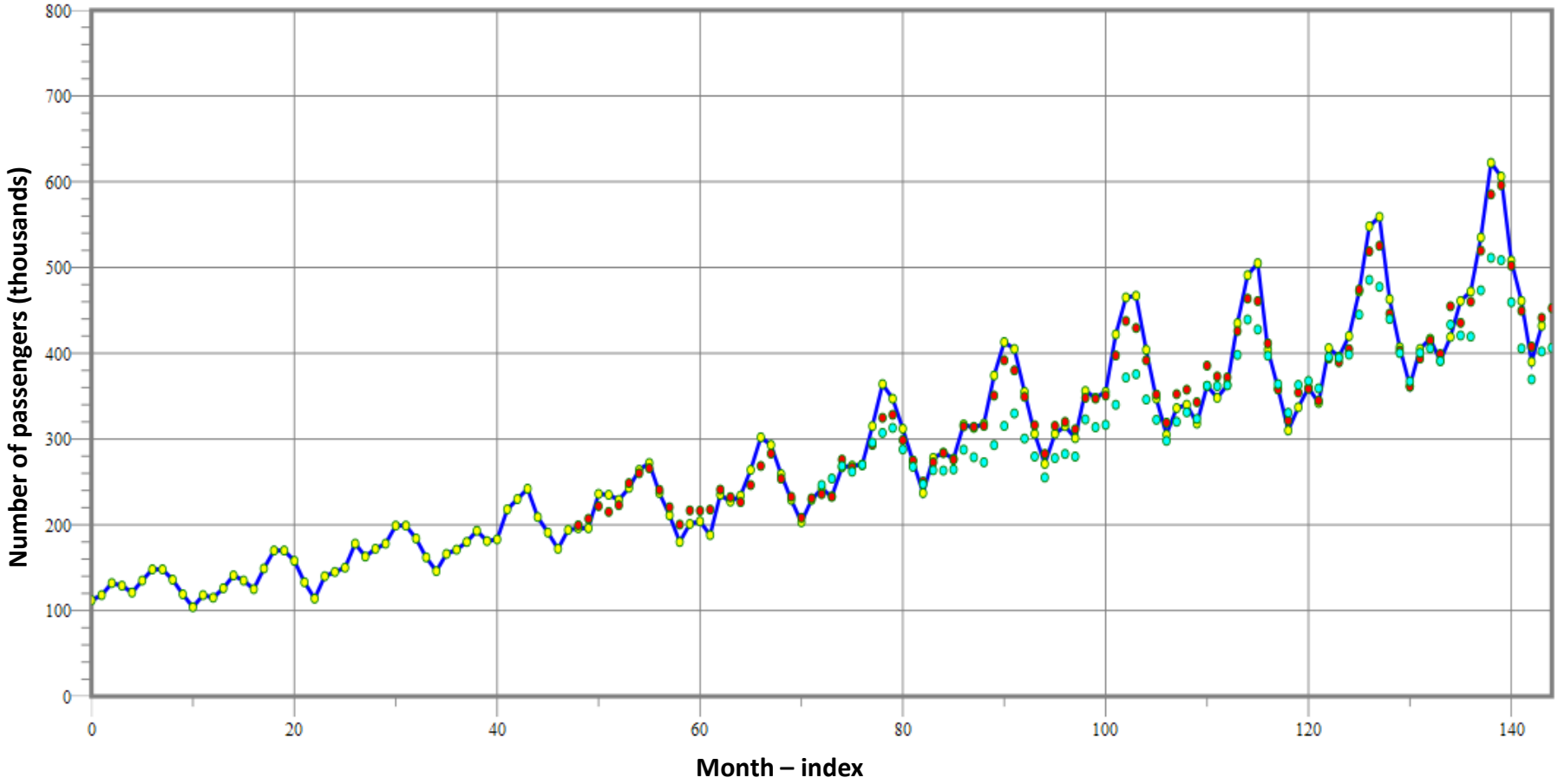
- **Why? Scale, correlate with other types of Streams**

Geospatial Analytics

```
double distance( Point p0, Point p1)
{
   double lat1 = p0.y * M_PI / 180.0 ;
   double lat2 = p1.y * M_PI / 180.0 ;
   double lon1 = p0.x * M_PI / 180.0 ;
   double lon2 = p1.x * M_PI / 180.0 ;

   double sdlat = ::sin((lat2 - lat1)/2.0L);
   double sdlon = ::sin((lon2 - lon1)/2.0L);
   double a = ::sqrt( sdlat * sdlat + ::cos(lat1) * ::cos(lat2) * sdlon * sdlon );
   double d = 2.0L * ::asin(std::min<double>(1.0L,a)) * radius ;
   return d ;
}


 static double distance(const float* p0, const float* p1, const float* p)
 {
        float dv[2] = { (p1[0] - p0[0]), p1[1] - p0[1] };
        if (dv[0] == 0 && dv[1] == 0) return distance(p0, p);

        double r = ( (p[0] - p0[0]) * dv[0] + (p[1] - p0[1]) * dv[1] )
              / (dv[0] * dv[0] + dv[1] * dv[1]);
        if (r < 0) {
           return distance(p0, p);
        } else if (r > 1) {
           return distance(p1, p);
        }
        float md[2] = { ((1 - r) * p0[0] + r * p1[0]), ((1 - r) * p0[1] + r * p1[1]) } ;
        return distance(md, p);
}
```

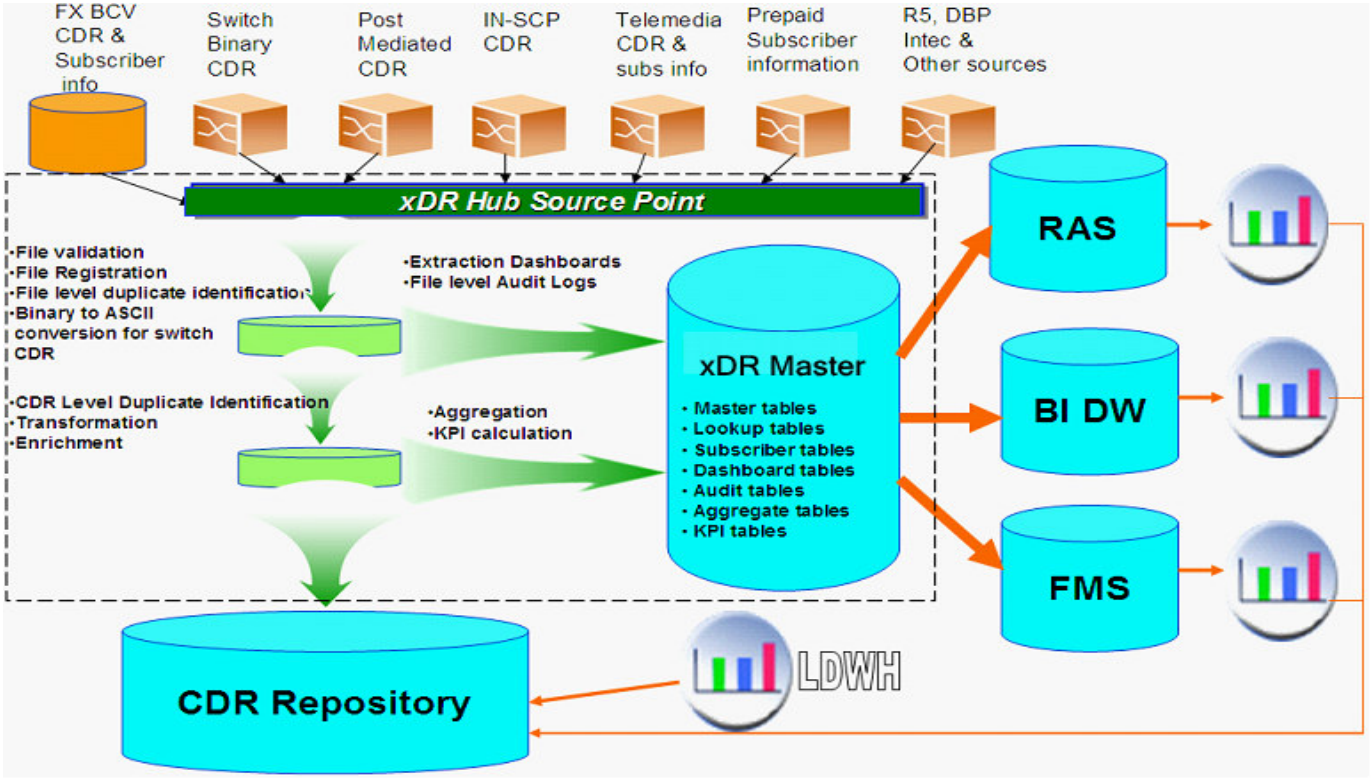
Forecasting Airline passenger data using Streams and time series analytics

monthly data of airline bookings from January 1949 to December 1960
-real-time data in yellow
***- next month prediction** in red*
***- 24 months (or 2 years) prediction** in light blue*



Streaming analytics in practice
Telecommunications CDR Processing

Requirements: Scaling and Price/Performance



8x increase in throughput compared to existing system

Latency reduced from 12 hours to 1 second!

IBM InfoSphere Streams directions

Tools
Streams Studio enhancements
Video/audio analytics
Text/unstructured analytics
Composite data type support
Drag/Drop composition

Runtime
High Availability enhancements
Security enhancements
Installation enhancements
SUSE and AIX support

Adapters
GPFS/HPFS/Hadoop (BigInsights)
Cognos Now
WebSphere MQ
RSS feeds
DB2 Parallel writer
Native XML support

All statements regarding IBM's plans, directions, and intent are subject to change or withdrawal without notice. Any reliance on these statements are at the relying party's sole risk and will not create any liability or obligation for IBM.

Disclaimer

THE INFORMATION CONTAINED IN THIS PRESENTATION IS PROVIDED FOR INFORMATIONAL PURPOSES ONLY. WHILE EFFORTS WERE MADE TO VERIFY THE COMPLETENESS AND ACCURACY OF THE INFORMATION CONTAINED IN THIS PRESENTATION, IT IS PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED. IN ADDITION, THIS INFORMATION IS BASED ON IBM'S CURRENT PRODUCT PLANS AND STRATEGY, WHICH ARE SUBJECT TO CHANGE BY IBM WITHOUT NOTICE. IBM SHALL NOT BE RESPONSIBLE FOR ANY DAMAGES ARISING OUT OF THE USE OF, OR OTHERWISE RELATED TO, THIS PRESENTATION OR ANY OTHER DOCUMENTATION. NOTHING CONTAINED IN THIS PRESENTATION IS INTENDED TO, NOR SHALL HAVE THE EFFECT OF, CREATING ANY WARRANTIES OR REPRESENTATIONS FROM IBM (OR ITS SUPPLIERS OR LICENSORS), OR ALTERING THE TERMS AND CONDITIONS OF ANY AGREEMENT OR LICENSE GOVERNING THE USE OF IBM PRODUCTS AND/OR SOFTWARE.

Thank You

InformationOnDemandIndia2011

The Premier Conference for Information Management

Manage. Analyze. Govern.

February 2, 2011

Hyatt Regency I Mumbai, India