Configurable Analytic Flows at Scale: A New Challenge for the BPM Community

Richard Hull, IBM T.J. Watson Research Center

8 September 2014

Presented at the 3rd Intl. Workshop on Data- and Artifact-Centric BPM (DAB), as part of the 12th Intl. Conf. on Business Process Management (BPM), held in Eindhoven, The Netherlands
Big Data Analytics

“Big Data”
- Volume
- Variation
- Velocity
- Veracity

“Analytics”
- Descriptive
- Predictive
- Prescriptive

Big Data Analytics is bringing value across all industries
- Logistics
- Retail
- Healthcare
- Energy
- Education
- Born-on-the-web companies
- ...

Volume
- Data at Rest
  Terabytes to exabytes of existing data to process

Velocity
- Data in Motion
  Streaming data, milliseconds to seconds to respond

Variety
- Data in Many Forms
  Structured, unstructured, text, multimedia

Veracity*
- Data in Doubt
  Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Prescriptive

Predictive

Descriptive

- Typically for human consumption
- Useful for embedding into BP’s
- Raises challenge of continual improvement
“Big Data Analytics”: A major force in early 21\textsuperscript{st} century

- McKinsey
  - Big Data will become the basis for competition
  - Big Data will underpin new waves of productivity growth
  - 140,000 to 190,000 more deep analytical talent positions in US
  - 1.5 Million more data-savvy managers needed in US
- Key sectors include healthcare, retail, manufacturing, also education

From McKinsey 2011:
Big Data: The next frontier of innovation, competition and productivity
Example: Supply Chain Management

Big Data Analytics increasingly relevant to Business Operations

How can I reduce my CO₂ emissions?

What if?

- Sustainability Metrics
  By site, operation, SKU, product group, geo, etc.

How can we reduce our costs?

What if?

- Inventory Metrics
  By site, SKU, product group, geo, etc.

E.g., where to place warehouses to:
- Maximize speed of delivery
- Minimize on-hand inventory
- Minimize green-house gases

E.g., an IBM analytics project saved pharma McKesson over $1B through Supply Chain costs

How can I increase inventory turns?

What if?

- Operations Managers

How can we improve our P & L?

What if?

- CFO Office

P&L - Revenue Analysis
By vendor, customer, site, SKU, product group, geo, etc.

Cost-to-serve
By vendor, customer, site, SKU, product group, process, geo etc.
Example: Social Media & Text Analytics

- RetailerXX wants to sell to the “Millennials” - ages 16-25
- Who are the Millennials, anyway, and how do they shop??

- IBM analyzed over 3 BILLION tweets
- Created 7 “clusters” of Millennial shoppers

For example...

**Fashion on a Dime Persona**

- Loves going to the Mall, whether it is to purchase at a department store or at Forever 21
- Young Millennial who has a positive sentiment towards RetailerXX’s but is not brand loyal
- Prefers discounts and is highly incentivized by personalized offers
- Shares everything with their friends, from their latest purchase to their dream vacation
- Follows latest fashion news and gossip, dreams of going to Fashion Week, and feels like they belong when their “look” matches their favorite stars

**Illustrative Info Extraction / Text Analytics** increasingly present in Big Data applications
"Actionable Customer Satisfaction" for B2B sales

**Analytics infer customer sat, key drivers, mediating actions**

---

**Data Cleaning and Preprocessing**

**Data Transformation**

**Intermediate Insights**

**Analytics Execution**

**Company NPS info and rec'd actions (refreshed daily)**

---

**Customer experience data** (e.g., Purchasing Provisioning, Tickets, ...)

---

**Representative Insight & Mitigating Actions**

- **Service Delay Threshold:**
  
  Customers averaging >xx day ticket response show lower customer satisfaction

- **Potential Actions:**
  
  - Auto escalate tickets after 4 days
  - Pre-communicate to customer when repair estimate >xx days

---

**Predicting drivers through Mechanistic Models**

- **Intrinsic distribution**
- **Exponential non-linearity**
- **Probabilistic spiking**

- **Based on recently**

---

**Cust Sat Explained from Drivers to Recommended Actions**

**Short-term Actions**

- More client contact
- Offer discounts and/or free services

**Long-term Actions**

- Design training program

---

**Repeated execution of analytics flow embedded into on-going Business Process**

---

**Company NPS info and rec'd actions (refreshed daily)**

---

**NPS Dashboard**

---

**B2B Seller NPS Dashboard**

---

**Customer Satisfaction Drivers**

- Market spectrum
- Design training program
- More client contact
- Offer discounts and/or free services

---

**Company NPS info and rec'd actions (refreshed daily)**

---

**Cust Sat & Driver Scores**

---

**Service Delivery Metrics**

---

**Personnel & Execution**

---

**Contextual factors**

---

**Market spectrum**

---

**Unsatisfied Clients**

---

**Satisfied Clients**
LARIAT adds timely listening to traditional approaches to B2B Lead-to-Revenue management
LARIAT output: Data about Companies (detail from Smarter Process sales team view)

<table>
<thead>
<tr>
<th>Company</th>
<th>DUNS number</th>
<th>Priority level</th>
<th>Industry</th>
<th>Employee count</th>
<th>Revenue</th>
<th>Relevant indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medline Industries, Inc.</td>
<td>25460908</td>
<td>95 Q:50</td>
<td>Surgical appliances and supplies, nsk</td>
<td>7230</td>
<td>1019000000</td>
<td>? ✓ Smart Process: Mergers and Acquisitions (Medline Industries, Inc., Professional Hospital Supply, Inc.) (1)</td>
</tr>
<tr>
<td>Zynga Inc.</td>
<td>15495485</td>
<td>55 Q:50</td>
<td>Prepackaged software</td>
<td>3058</td>
<td>1281267000</td>
<td>? x Smart Process: Healthcare Processing Challenged (Medline Industries, Inc.) (1)</td>
</tr>
<tr>
<td>Nationwide Corporation</td>
<td>7902026</td>
<td>70 Q:50</td>
<td>Life insurance, nsk</td>
<td>34417</td>
<td>12084628674</td>
<td>? x Smart Process: Loan Processing Challenged (Nationwide Corporation) (1)</td>
</tr>
</tbody>
</table>

IBM Copyright © 2014
A new kind of BPM

- **Traditional BPM**
  - Focus on managing business transactions
  - Process-centric

- **Case Management (and Artifact-centric)**
  - Focus on managing highly variable business processes
  - Data-centric

- **Analytics Process Mgmt “APM”**
  - Focus on managing the creation, deployment, use, & maintenance of analytics processes
  - Focus on Prescriptive analytics
  - Results guiding BP decisions

Increase in knowledge-worker driven BP’s

Analytics will support next wave of business productivity improvements
Research community has not been thinking about repeating analytics flows used by BPs

- CACM Survey of Business Intelligence [Chaudhuri, Dayal, Narasayya 2011]
  - The “product” of analytics is for human consumption, not BPs

- CACM Technical Challenges in Big Data [Jagadish et al 2014]
  - Again, the “product” of analytics is for human consumption, not BPs

- [Troung and Dustar 2012]: “Research on how to manage analysis algorithms and how to provide an open platform for third parties to develop, search and share algorithms is quite open.”
Agenda

- Drill-down on representative Analytics Processes
- What makes APM different/hard?
- The ProkoFieV framework
  - Functionality
  - Variation
  - Provenance
- Relevant techniques/tools
- Some foundational research questions
"Actionable Customer Satisfaction" for B2B sales

**Analytics infer customer sat, key drivers, mediating actions**

Customer experience data (e.g., Purchasing Provisioning, Tickets, ...)

- Service Delay Threshold: Customers averaging >xx day ticket response show lower customer satisfaction
- Potential Actions:
  - Auto escalate tickets after 4 days
  - Pre-communicate to customer when repair estimate >xx days

Representative Insight & Mitigating Actions

Predicting drivers through Mechanistic Models

Repeated execution of analytics flow embedded into on-going Business Process
Actionable Customer Satisfaction: Production Flow and “Feeder Analytics”

Production Analytics Flow (run daily)

- Ad-hoc, exploratory analytics
- Found “5-Day Tipping Point”
- Need to validate/ & possibly refine monthly

Combining driver scores using predictive model

Statistical insights incorporated into Production Flow

Trouble Tickets Driver

B2B Seller NPS Dashboard
Actionable Customer Sat Analytics Flow (abstracted)

**Run monthly and as needed**

- Derive Driver 1 score & actions
- Derive Driver 2 score & actions
- Derive Driver 3 score & actions

**Run daily**

Combine drivers to infer Customer Sat and prioritize actions

These sub-flows produce statistical models (policies, algorithms) for inferring driver scores.

The driver models are embedded into the top-level customer sat flow.

Training data may be used to create statistical model for the top-level flow.

Daily output includes customer sat, driver scores, prioritized actions.

Training data
LARIAT adds timely listening to traditional approaches to B2B Lead-to-Revenue management

Traditional Approaches

- Dun & Bradstreet
- Txn History
- Contact History
- Marketing Content
- SEC filings
- Propensity to Buy
- Business Ontology
- Business Value Maps

Foundations

LARIAT addition

- News, blogs, analysts, SEC filings

Intelligent Listening

- Business Goals
- Capabilities Needed
- Business Opportunities
- Indicators

Salesperson/Client Rep

Prioritized listing of leads with recent events and rationale

Analytics Flow is running continuously
LARIAT Functional Components and Processing Flow overview

For daily use in BPs for lead identification & nurturing

Presentation Layer

Business Information Repository Layer

Information Processing Layer

Input Documents → Structured Enterprise Data

- Crawler & Syntactic Filtering
- Semantic Filtering
- Aggregation & Entity Resolution
- Validation Pruning
- Scoring

Sales Team

Power Business User UI
Spreadsheets
RSS Feeds
Aggregated Statistics

Propensity to Buy
Stakeholders around an Analytics Flow Solution (example)

P/L Owner
- Manages ROI
- Very concerned about metrics
- Wants explanations

Visualization

Sales Mgmt

Sellers

Measure & refine

Manual validation team

IT Support

Source acquisition

Source selection

ETL Specialist

Statistical Analytics Team

Maintain/tune Propensity-to-buy

Propensity to Buy

Structured Enterprise Data

Information Processing Layer

Input Documents

Crawler & Syntactic Filtering

Semantic Filtering

Aggregation & Entity Resolution

Validation Pruning

Scoring

Text Analytics Specialist

Create text extractors based on business patterns

Ontologies

Business causality relationships

Subject Matter Experts

Identify business patterns

Identify

Stakeholders

Stakeholders around an Analytics Flow Solution (example)
Top-level BPs for a repeating analytics flows*

- Each of these top-level BPs is knowledge-worker intensive
- Case Management/Biz Artifacts is natural approach to support these
- This will enable strong measurement & governance of the effectiveness of both the analytics flows and the personnel that are working on/with them

*Related to, and expanding on, CRISP-DM
The core entity type: Configurable Analytics Flow

- Full flow might execute, or a subgraph
- Multiple points of configurability
  - Mainly based on changing data or logic
- A vehicle for retaining provenance of computed data
  - Prospective: flow design
  - Retrospective: info about instance
- Provides anchor for measurements and identifying attributions

Notes:
- Flows are Directed Acyclic Graphs (DAGs)
- Evolution/Variation can be accomplished with simple manipulations, e.g., add node, delete node, etc.
Broader Perspective: A 3-dimensional view of this space

- **Functionality**
- **Variation**
- **Provenance**

Think of this as one analytics flow schema

Think of these as executed instances of that schema

**Functionality**
- BP embedding, measurement
- Reports, Viz
- Production flows
- Ad hoc exploration
- ETL, Data Fusion
- Raw Data, Streams

**Provenance**

**Variation**
### Relevant techniques/tools

<table>
<thead>
<tr>
<th>Technique/Tool</th>
<th>Description</th>
</tr>
</thead>
</table>
| **ETL** (Extract-Transform-Load) | Broad array of techniques for gathering/curating data for use in analytics/data mining  
No higher-level tools to help workers manage/record/govern their ETL work |
| **CRISP-DM** (CRoss-Industry Standard Process for Data Mining) | Framework for Data Mining (including refinements)  
Primarily a methodology; comprehensive mgmt platforms not available  
Focus on finding one-off insights |
| **Case Management** | Good fit: The top-level BPs for APM are very knowledge-worker driven  
We should identify some template schemas |
| **Scientific Workflow** | The analytics flows themselves are quite similar to scientific workflows  
However, analytics flows emphasize *measurement, attribution, refinement* |
| **BPM Adapatability** | Frameworks/tools to manage variation of BPs, at instance level and schema level  
Analytics flows are DAGs (simpler); but provenance and queries against collections of flows are important |
| **IT Governance** | Standardized practices for ensuring that IT processes are effectively serving business objectives  
Analytics flows are a blend of biz and IT |
Some key challenges (overview)

- A precise model of Configurable Analytics Flows
  - **Capabilities**: Provenance, Support for Measurement, Variation/Evolution
  - **Abstraction** over the heterogeneity of underlying components/tools

- From exploratory flow(s) to a reified flow
  - The challenge of being *light-weight*

- Extract-Transform-Load (ETL)
  - Some tools are fairly mature but the work is *still very time-consuming*

- Enabling rich collaboration in Analytics Flow eco-system
  - Vision for *factorization of logical components* to enable broad-scale, cloud-hosted crowd sourcing across all areas of APM

Additional challenges

- Case Mgmt templates for the top-level BPs for APM
- How should Case Mgmt be extended to work better on Analytics Flows?
Good models go beyond description - they support action

- Selecting the right model for the job matters

Example: “Game of 15”
Winner: First one to reach exactly 15 with any 3 chips

First model - A is [ ] and B is [ ] - what is B’s move?
Second model - [ ] [ ] - B’s move is 6!
Configurable Analytics Flows as a key abstraction layer

Flow from LARIAT

- Article stream
- Syntactic filtering
  - Syntactic query
- Semantic info extraction
- AQL Extractors
- Entity Resolution policy/algorithm
- Match article to D&B info
  - Dunn & Bradstreet Company Info
- Join
  - Propensity to-buy
- Manual validation
- Validation Policy
- Worker
- Prioritization
- Prioritization Rules

Flow from Actionable Customer Sat. (abstracted)

- Derive Driver 1
  - score & actions
- Derive Driver 2
  - score & actions
- Derive Driver 3
  - score & actions

- Combine drivers to infer Customer Sat and prioritize actions

Is this the useful model? Is there a more useful one?
Configurable Analytics Flows: Requirements and Approaches

**Requirements**

- Intuitive, conceptually transparent
- Numerous ways to work with the flows
  - Ad hoc exploration
  - Re-use, including re-use of sub-flows
  - Rich query ability over large sets of flows, including visualize answers
- Enables easy comparison between flows based on measurements
  - Crucial for achieving ultimate biz goal of the analytics
- Provenance of flow outputs is intuitive, conceptually transparent
  - Important for measurement, compliance, governance

**Starting points from Scientific Workflow**

- Considerable work on provenance, executability, optimization, tools

**Additional research needed**

- Adapt query/visualization to better support measurement
- Develop a theory of sub-flows, sub-flow composition, queries on sub-flows
- Find simple/intuitive ways to describe flows, to enable “executive level” explanations for flow outputs and differences between flows
VisTrails: Flows and Flow Provenance Tree
From ad hoc flows to reified flow

- **Context:**
  - Data scientists often explore a variety of perspectives and analytics models before identifying insights that can bring deep value
    - Heterogeneous data/tools may be used
  - Several flows might be used for testing/measurement
  - Finally, some flow(s) will be reified and put into production use
    - Perhaps in a tool different from the ad hoc exploration tool(s)

- **Challenge:** Data Scientists typically can’t keep track of their flows
  - Capture of flows
  - Access to flows (and sub-flows): Queries over flow collections
  - Mapping from highly flexible ad hoc tools to production tool

- **Starting points from Scientific Workflow**
  - E.g., Kepler, Taverna, SWIFT, VisTrails use flow models, with query support
  - Approaches to “capture” of flows
    - Use operating system logs (e.g., PASS)
    - Logically centralized workflow tool - record all (e.g., Kepler, VisTrails) or delegate prov capture to components (e.g., Provenance-Aware SOA project/standard)
  - VisTrails designed to support ad hoc, exploratory flow creation
    - Focus on outputs used by humans, not embedded into BPs
    - *Representation of sets of flows, and query access, needs strengthening*
    - *Can we create something even more light-weight, unobtrusive (cf REST, JSON)*
The ETL Challenge

- [NY Times 8/17/2014] -- 50% to 80% of Analytics work is “data wrangling” or “data munging” or “data janitor work”
  - Timothy Weaver, CIO of Del Monte Foods: data wrangling big data’s “iceberg” issue

- State of the art in ETL (e.g., [Chaudhuri, Dayal, Narasayya 2011]):
  - Gather data and place into a warehouse
  - Variety of tools are now mature
    - Consistency mgmt, e.g. “…, California, Canada”
    - String manipulations, entity resolution, e.g., “California” -> “CA”
    - Extracting structure from strings, e.g., parse “Coby MP3 512MB MP-C756 - Blue.”
    - Instance-level key/foreign-key idenification
    - Data load and refresh (e.g., by triggers, by log scraping)

- The data-centric BPM community can provide help!
  - We know: process, data mgmt, variation, knowledge work, collaboration
  - Starting point may be to apply ideas from Configurable Analytics Flows to ETL
    - Extend warehouse focus to include process-centric data capture
    - Enable capture of ad hoc ETL explorations
    - Simplify reification of “good” ETL flows
    - Enable better re-use through use of ontologies, semantic web
Vision for Factoring Analytics Flow (Illustration)

- An environment where multiple parties can contribute to different portions of the LARIAT flow?
- Data-centricity & basic analytics flow provide backbone
  - Cf. variation in traditional BPM
- Multi-tenancy:
  - Different end-users given access to subsets of flow & output
- Compensation based on Attribution
  - Challenge: how to determine attribution

---

[Callery et al, SCC, 2014]
Conclusions / Call to Action

- Analytics Process Management (APM) is the next big research challenge in BPM

- Data-centric BPM community is best positioned group to attack this

- Case Management is well-suited for the top-level BPs of APM

- Configurable Analytics Flows are a good abstraction layer for modeling the fundamental activity of APM

- While Scientific WF provides a starting point, there are many challenges in adapting to the BP context
  - Stemming from repeating flows, heterogeneity of stakeholders, measurement & feedback loop, explanation to executives, ...

- [Troung and Dustar 2012]: “Research on how to manage analysis algorithms and how to provide an open platform for third parties to develop, search and share algorithms is quite open.”
Acknowledgements

This thinking is based on projects/collaborations with several people, including

- LARIAT: Matt Callery, Terry Heath, Danny Oppenheim, Noi Sukaviriya, Roman Vaculin
- Actionable Customer Satisfaction: Krishna Ratakondra, Jeff Robinson, Anshul Sheopuri, Dashun Wang
- BOLO: Elham Kabheri, Yang (Daniel) Li, Matt Reiman, Roman Vaculin
Backup slides
CRISP-DM: Standardized method for performing iterative Data Mining

- Identify business challenges & questions
- Understand the available data
- Prepare data
  - Cleansing
  - Transformation
  - Integration
- Create analytical model(s)
  - Myriad of alternatives to fit broad variety of applications
- Evaluate & refine models
- Deploy
- Iterate
  - 1 or 2 month cycle
  - Each iteration builds value, infrastructure and experience
IT Governance (COBIT)

The 5 focus areas in COBIT

- COBIT assumes a fairly rigid separation between IT and Biz roles
- With Analytics Flows, some roles lie at interface of IT and Biz, e.g., Data Scientist, UI Designer/implementer
- Approaches to manage and measure these roles requires an extension of COBIT
Querying sets of flows in Scientific WF Systems

- **REDUX**: SQL against underlying relational store
- **VisTrails**: domain-specific query language
- **MyGrid**: SPARQL against RDF store

---

**Figure 5.** Provenance query implemented by three different systems. REDUX uses SQL, VisTrails uses a language specialized for querying workflows and their provenance, and myGrid uses SPARQL.

[Freire et al, CISE, 2008]