Data intensive flows
Knowledge objectives

1. Recognize the importance of usability and taking a user-centered approach
2. Remember BPMN elements regarding flow objects, swimlanes, connections, and data artifacts
Understanding Objectives

1. Assign ETL uses to BPMN elements
Application Objectives

1. Given a description of an ETL process, model it using BPMN
User centered design

“It is users and not data that are important.”

- Focus on the users
- Needed activities
  - Specify the context of use
  - Specify the user and business requirements
  - Design the product
  - Evaluate the design
Usable systems

- Effectiveness
  - Does it do the job?

- Efficiency
  - How easily does it do the job?

- Satisfaction
  - How enjoyable is it to do the job?
Consistent representation

- **Complete**
  - The user can get all information

- **Correct**
  - The user cannot derive any other information

---

Table 2.1: Example of database (Note, position relative to Rome).

<table>
<thead>
<tr>
<th>Town</th>
<th>People</th>
<th>#Position</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rome</td>
<td>4,000,000</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Milan</td>
<td>1,800,000</td>
<td>North</td>
<td>600</td>
</tr>
<tr>
<td>Naples</td>
<td>1,500,000</td>
<td>South-East</td>
<td>200</td>
</tr>
<tr>
<td>Pisa</td>
<td>150,000</td>
<td>North-West</td>
<td>350</td>
</tr>
<tr>
<td>Pescara</td>
<td>200,000</td>
<td>East</td>
<td>220</td>
</tr>
</tbody>
</table>
Classification of business processes
Workflow management system
Technological Challenges

- Business Process Management
- Service Composition
- Service Infrastructure and Management

![Diagram showing Business Processes, Business Services, and Utility Services with related technologies like DB, MapReduce, SCM, CRM, MDM, ERP]
Comparison between ETL and BPM

- Benefits of treating ETL as a type of process
  - Provide an abstract view (implementation independent)
  - Monitor and report in terms of the abstract view

- ETL is batch oriented, while BPM is event oriented
  - We can also consider pipelining ETL
    - This is more appropriate for streaming
ETL operations

- Extraction
- Schema modification
  - Projection
  - Field splitters
  - Attribute addition
- Aggregation
- Value derivation
  - Value mapper
  - Lookups
  - String processing
  - Scripting
  - Cryptography
- Dataset alteration
  - Filtering
  - Duplicate removal
  - Sampling
- External calls
  - Check for existence
  - Send e-mail
  - Write to log
- Others
  - Delay row
  - Blocking step
  - Abort
Process Modelling Viewpoints

- When?
- Which?
- What?
- Organization
- Who?
- Data / Service / Product
- Function
- Process

http://fundamentals-of-bpm.org
BPMN idea

A BPMN process model is a graph consisting of four types of elements (among others):

- **Event**
- **Task**
- **Flow**
- **Gateway**

[http://fundamentals-of-bpm.org](http://fundamentals-of-bpm.org)
BPMN main elements

Connections (when)
- Message
- Association
- Flow

Swimlanes (who)
- Pool
- Lane
- Data Store

Flow Objects (what)
- Gateway
- Event
- Activity

Artifacts (which)
- Text Annotation
- Data Object

http://fundamentals-of-bpm.org
Flow elements

- Start Event
- Task
- End Event
- Flow
- AND-Join
- XOR-Decision
- XOR Merge
- AND-Split

http://fundamentals-of-bpm.org/
Gateways

- **Exclusive Decision / Merge**
  - Indicates locations within a business process where the sequence flow can take two or more alternative paths
  - **Only one** of the paths can be taken

- **Parallel Fork / Join**
  - Provide a mechanism to synchronize parallel flow and to create parallel flow
  - Depicted by a diamond shape that *must* contain a marker that is shaped like a plus sign

http://fundamentals-of-bpm.org
Example of gateways

Check stock availability → Reject order

Node:

Confirm order

Send invoice

Ship goods

Node:
Sub-processes

- An activity in a process can “invoke” a separate (sub-)process
- Use this feature to:
  1. Break down large models into smaller ones, making them easier to understand and maintain → process hierarchies
  2. Share common fragments across multiple processes → shared subprocesses
  3. Identify parts of a process that should be:
     - Repeated
     - Executed multiple times in parallel
     - Cancelled
- Good practice is that the top-level process should be simple (no gateways) and should show the main phases of the process
  - This is sometimes called a “value chain”
  - Each phase then becomes a sub-process

http://fundamentals-of-bpm.org
Example of process hierarchies

Level 3

Process Inquiry and Quote → Receive and Validate Order → ...

Level 4

Receive Order → Enter Order → Check Credit → ...

Level 5

Access Credit Record → Credit Available? → Clear Order → Contact customer account rep. → ...

http://fundamentals-of-bpm.org
Multiple instance marker

- \( \equiv \): Sequential repetition of an activity/sub-process
- \( ||| \): Parallel repetition of an activity/sub-process

Useful when the same activity should be executed for multiple entities or data items,

- Examples:
  - Request quotes from multiple suppliers
  - Check the availability for each line item in an order separately
  - Send and gather questionnaires for multiple witnesses in the context of an insurance claim
Example of Multiple instance activity

For each supplier

Obtain Quote → Select Best Quote → Send PO

http://fundamentals-of-bpm.org

November 2013 22 Alberto Abelló & Oscar Romero
Resource elements

Resource classes are captured using:

- **Pools** – independent organizational entities
  - E.g., Customer, Supplier, East-Tallinn Hospital, Tartu Clinic
- **Lanes** – resource classes in the same organizational space and sharing common systems
  - Sales Department, Marketing Department
  - Clerk, Manager, Engineer

Resource class is a set of resources with shared characteristics

- E.g., Clerk, Manager, Insurance Officer

A resource class may be a

- **Role** (skill, competence, qualification)
  - Classification based on what a resource can do or is expected to do
- **Group** (department, team, office, organizational unit)
  - Classification based on the organization’s structure

http://fundamentals-of-bpm.org
Pools and Swimlanes

Data intensive flows

November 2013 Alberto Abelló & Oscar Romero

http://fundamentals-of-bpm.org
Example of Pools
Example of Lanes

Data intensive flows

November 2013 26 Alberto Abelló & Oscar Romero

http://fundamentals-of-bpm.org
Artifacts

- **Data Objects** are a mechanism to show how data is required or produced by activities:
  - Are depicted by a rectangle that has its upper-right corner folded over
  - Represent input and output of a process activity

- **Data stores** are containers of data objects that need to be persisted beyond the duration of a process instance

- **Associations** are used to link artifacts such as data objects and data stores with flow objects (e.g., activities)

http://fundamentals-of-bpm.org/
Data intensive flows

Example of Artifacts

- Check stock availability
- Purchase Order
- Confirm order
- Reject order
- Send invoice
- Ship goods

Set PO to approved
Set PO to rejected

http://fundamentals-of-bpm.org/
# Events

<table>
<thead>
<tr>
<th></th>
<th>Start</th>
<th>Intermediate</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
<td>Event Sub-Process</td>
<td>Boundary</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Interrupting</td>
<td>Interrupting</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Event Sub-Process</td>
<td>Boundary</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-Interrupting</td>
<td>Non-Interrupting</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Catching</td>
<td>Throwing</td>
</tr>
</tbody>
</table>

### None: Untyped events, indicate start point, state changes or final states.

### Message: Receiving and sending messages.

### Timer: Cyclic timer events, points in time, time spans or timeouts.

### Error: Catching or throwing named errors.

### Compensation: Handling or triggering compensation.

### Link: Off-page connectors. Two corresponding link events equal a sequence flow.

### Terminate: Trigging the immediate termination of a process.
Example of Events

Receive PO

Register PO

Check Availability

Send PO Response

Next working day

weekend/holiday

Receive PO

Register PO

Change

Receive PO

Register PO

Change
Data-intensive flows

Data-based vs. event-based decision

- In an XOR-split gateway, one branch is chosen based on expressions evaluated over available data
  - Choice is made immediately when the gateway is reached

- Sometimes, the choice must be delayed until something happens
  - Choice is based on a "race between events"

- BPMN distinguishes between:
  - Exclusive decision gateway (XOR-split)
  - Event-based decision gateway
Example of Event-based Decision

Receive PO Response

- Receive Error Message
- After 24 hours

Process PO Response

Notify Purchasing Officer

http://fundamentals-of-bpm.org
Boundary events

- Sometimes during a sub-process execution, some event may occur that needs some action...
- Such events are placed at the boundaries of the sub-process (boundary events)
- Two flavors:
  - Interrupting boundary events
  - Non-interrupting boundary events
Boundary Events – Example

- Purchase order received
- Check stock availability
  - New customer details received
  - Order cancellation request received
- Product not in stock
  - Order canceled
- Product in stock
  - New customer details registered
  - Handle order cancellation

http://fundamentals-of-bpm.org
Exception handling (error events)

- Exceptions are events that deviate a process from its “normal” course.
- Handling exceptions often involves stopping a sub-process and performing a special activity.
- Achieved using two event nodes:
  - An “end error event” that stops the enclosing subprocess execution.
  - An “intermediate error event” attached to the enclosing subprocess – this is where the process execution will continue after the error.
Example of Error events

Data intensive flows

Handle PO

Next working day

weekend

Check Availability

Send PO Response

PO Changed

default

OK

http://fundamentals-of-bpm.org

November 2013 36 Alberto Abelló & Oscar Romero
Event sub-processes

- An event sub-process are processes attached to a parent process, that are triggered when an event happens.
- Alternative to putting a boundary non-interrupting event around the parent process.
Example of Event sub-processes

- Acquire raw materials
- Manufacture product
- Retrieve product from warehouse
- Confirm order
- Ship and invoice
- Archive order

- Check stock availability
  - Purchase order received
  - product not in stock
  - product in stock

- Order enquiry
  - Order enquiry received
  - Order enquiry handled

- Equire order status

http://fundamentals-of-bpm.org
Choreographies

- Focus on interactions occurring between two or more parties
  - Each interaction involves an exchange of messages (one or more)

- Each activity element contains the information of the participants
  - Light band for the initiator
  - Dark band for the recipient
Example of choreographies

Data intensive flows

November 2013 Alberto Abelló & Oscar Romero

http://fundamentals-of-bpm.org
Good practices

- Hierarchical design
  - a) By using BPMN levels (1&2) notation
    i. Main flow
    ii. Exception handling
  - b) By drilling down activities into subprocesses

- Completeness

- Clarity (unambiguous)

- Shareability between business and IT

- Structural consistency (use standards)
Activity

- **Objective:** Use BPMN to model an ETL process

- **Tasks:**
  1. (15’) Individually draw a proposal of the corresponding ETL part
  2. (15’) Match all three proposals in to one
  3. Hand in the merged proposal

- **Roles for the team-mates during task 2:**
  a) Explains his/her material
  b) Asks for clarification of blur concepts
  c) Mediates and **controls time**
### Summary

<table>
<thead>
<tr>
<th>ETL</th>
<th>BPMN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraction/Load</td>
<td>Data store</td>
</tr>
<tr>
<td>Input/Output</td>
<td>Data objects</td>
</tr>
<tr>
<td>Parallelism</td>
<td>AND-gateway</td>
</tr>
<tr>
<td>Load balance</td>
<td>XOR-gateway</td>
</tr>
<tr>
<td>Complex task</td>
<td>Subprocess</td>
</tr>
<tr>
<td>Pipelining</td>
<td>Multiple instance marker</td>
</tr>
<tr>
<td>Multiple components</td>
<td>Swimlanes</td>
</tr>
<tr>
<td>Multiple resources</td>
<td>Pools</td>
</tr>
<tr>
<td>Exception handling</td>
<td>Error events</td>
</tr>
<tr>
<td>Compensation actions</td>
<td>Compensation events</td>
</tr>
<tr>
<td>Control flow</td>
<td>Even based decisions &amp; Boundary events &amp; Event subprocess</td>
</tr>
</tbody>
</table>
Bibliography

Resources

- http://www.signavio.com
- http://www.bpmb.de/images/BPMN2_0_Poster_EN.pdf
- http://oozie.apache.org
- https://sqoop.apache.org
- http://flume.apache.org