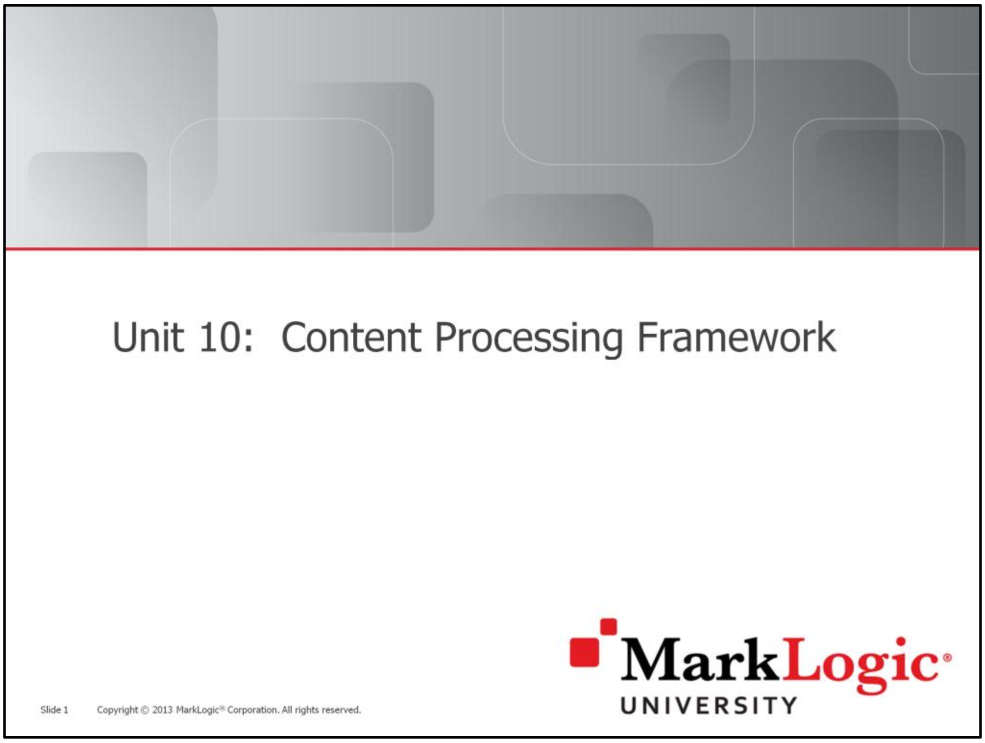
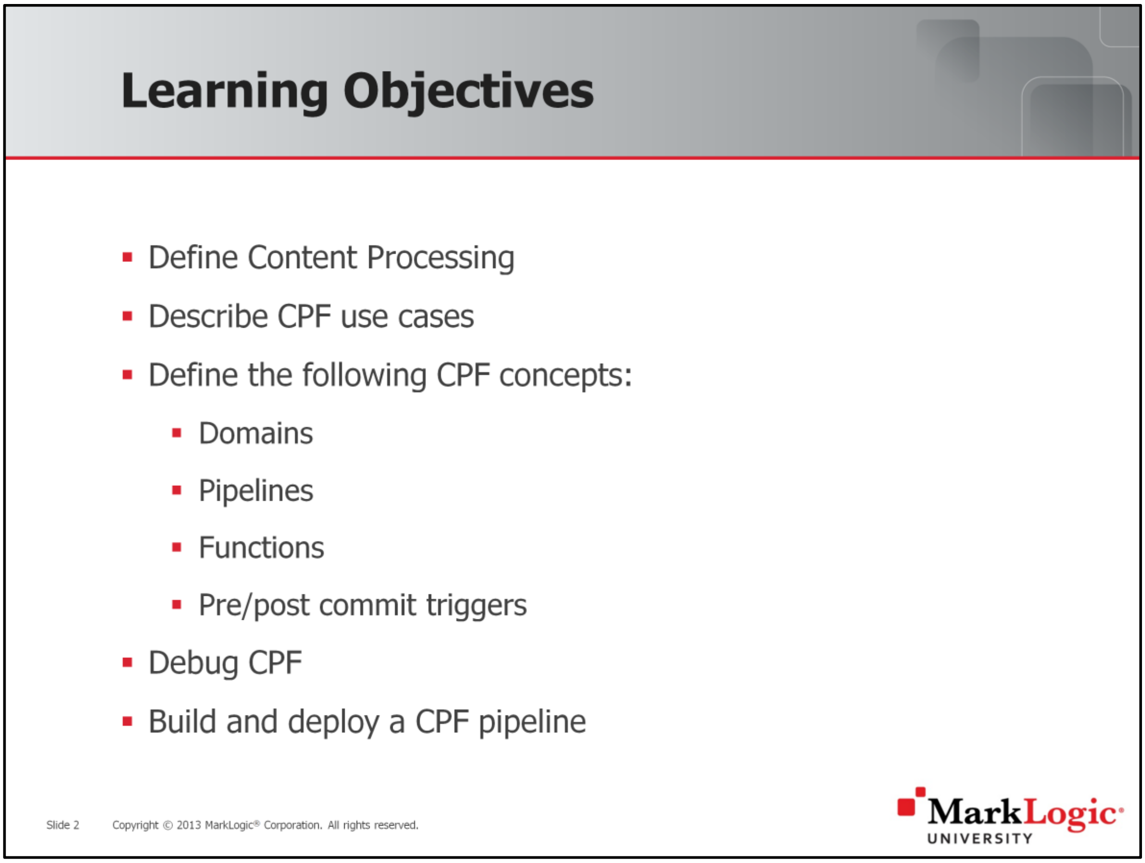


Unit 10: Content Processing Framework

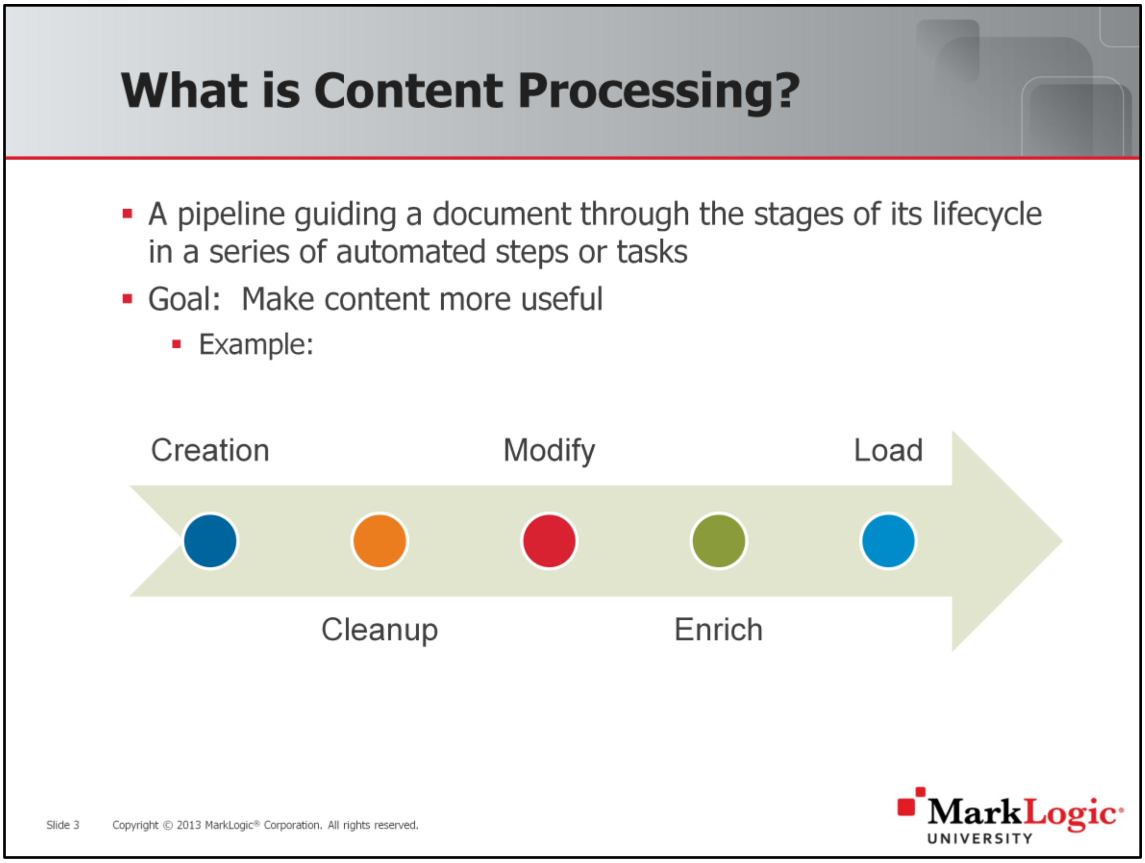
Slide 1 Copyright © 2013 MarkLogic® Corporation. All rights reserved.



Learning Objectives

- Define Content Processing
- Describe CPF use cases
- Define the following CPF concepts:
  - Domains
  - Pipelines
  - Functions
  - Pre/post commit triggers
- Debug CPF
- Build and deploy a CPF pipeline

Slide 2 Copyright © 2013 MarkLogic® Corporation. All rights reserved.



What is Content Processing?

- A pipeline guiding a document through the stages of its lifecycle in a series of automated steps or tasks
- Goal: Make content more useful
  - Example:

Creation

Cleanup

Modify

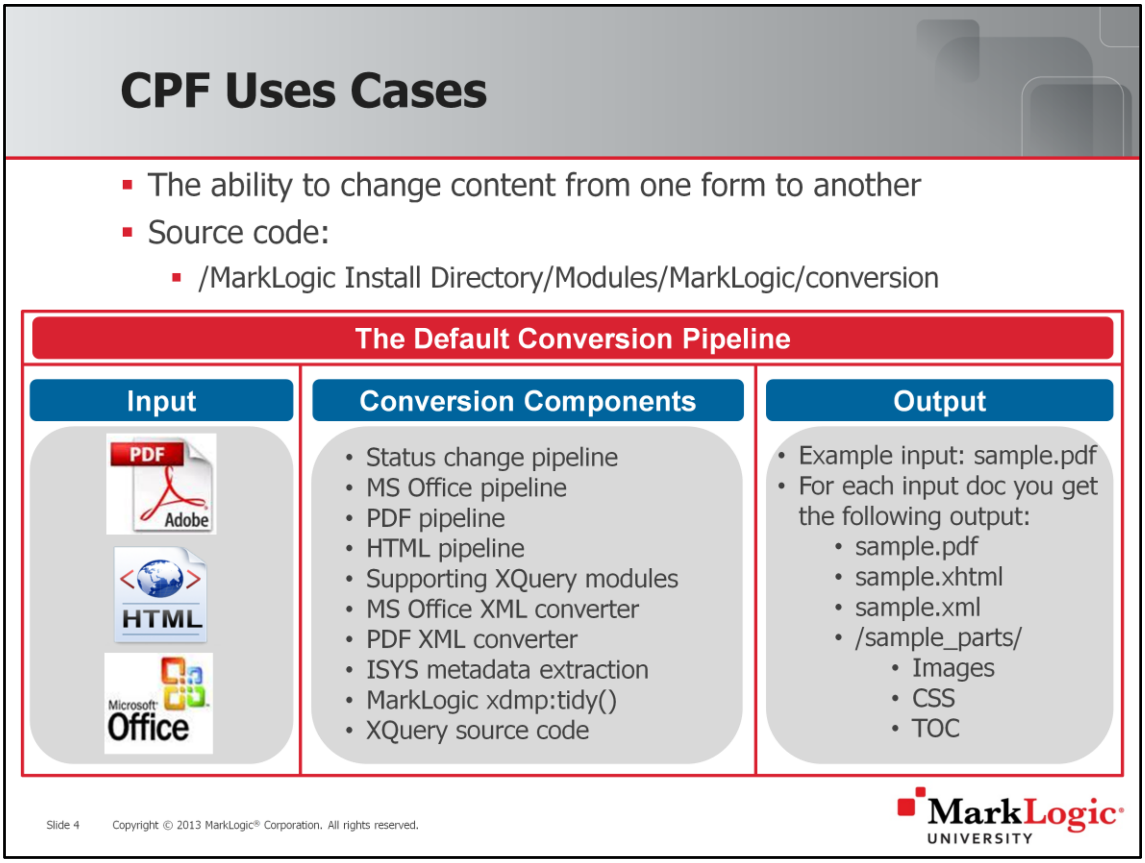
Enrich

Load

Slide 3 Copyright © 2013 MarkLogic® Corporation. All rights reserved.

Content can go through many stages before it is ready to use in an application. Sometimes, these stages might include making the content well-formed XML, some stages might transform one XML structure to another, and some stages might add (or take away) value to the content by examining the contents and combining it with other content or information. The process of content going from one stage to another is called *content processing*.

Every document has a lifecycle. The lifecycle of different documents can vary significantly. A lifecycle typically begins when a document is created, and then continues through various stages of content processing; content processing enables a document to move through the phases of its lifecycle.



CPF Uses Cases

- The ability to change content from one form to another
- Source code:
  - /MarkLogic Install Directory/Modules/MarkLogic/conversion

The Default Conversion Pipeline

Input

PDF
Adobe
HTML
Microsoft Office

Conversion Components

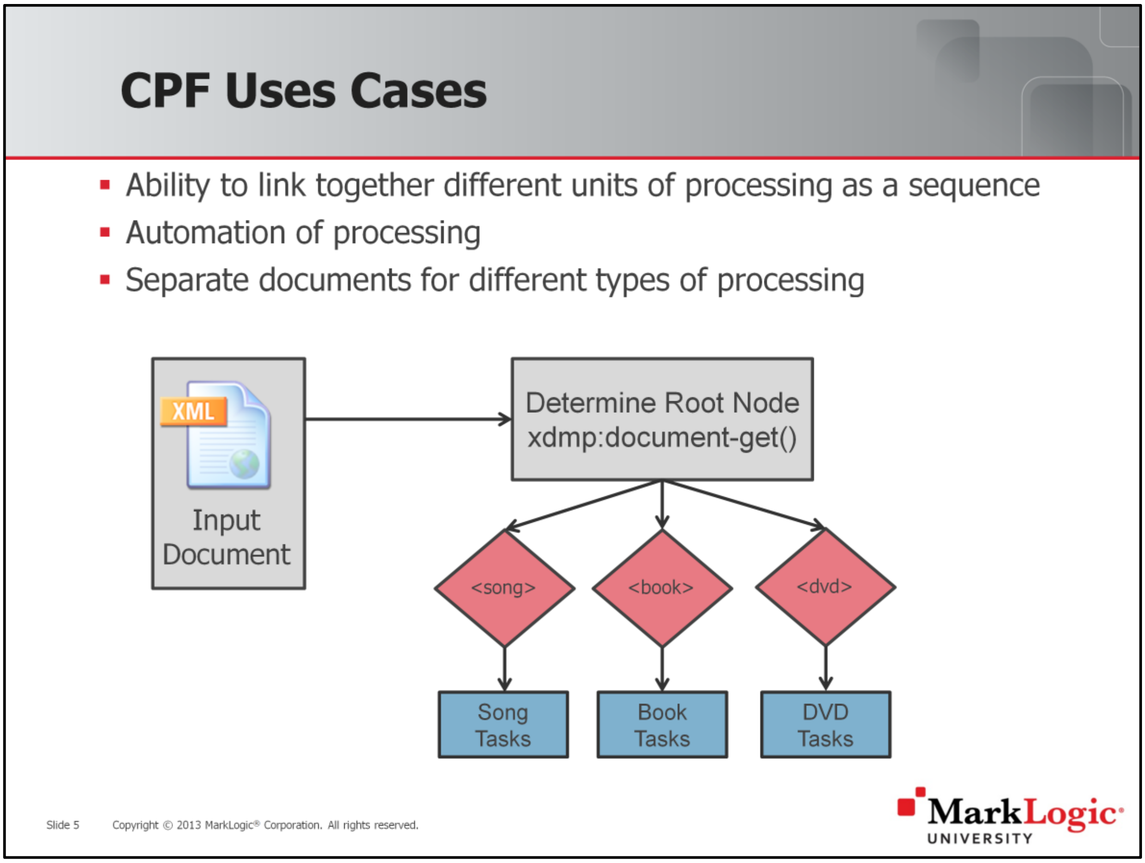
- Status change pipeline
- MS Office pipeline
- PDF pipeline
- HTML pipeline
- Supporting XQuery modules
- MS Office XML converter
- PDF XML converter
- ISYS metadata extraction
- MarkLogic xdmp:tidy()
- XQuery source code

Output

- Example input: sample.pdf
- For each input doc you get the following output:
  - sample.pdf
  - sample.xhtml
  - sample.xml
  - /sample_parts/
    - Images
    - CSS
    - TOC

Slide 4

MarkLogic UNIVERSITY

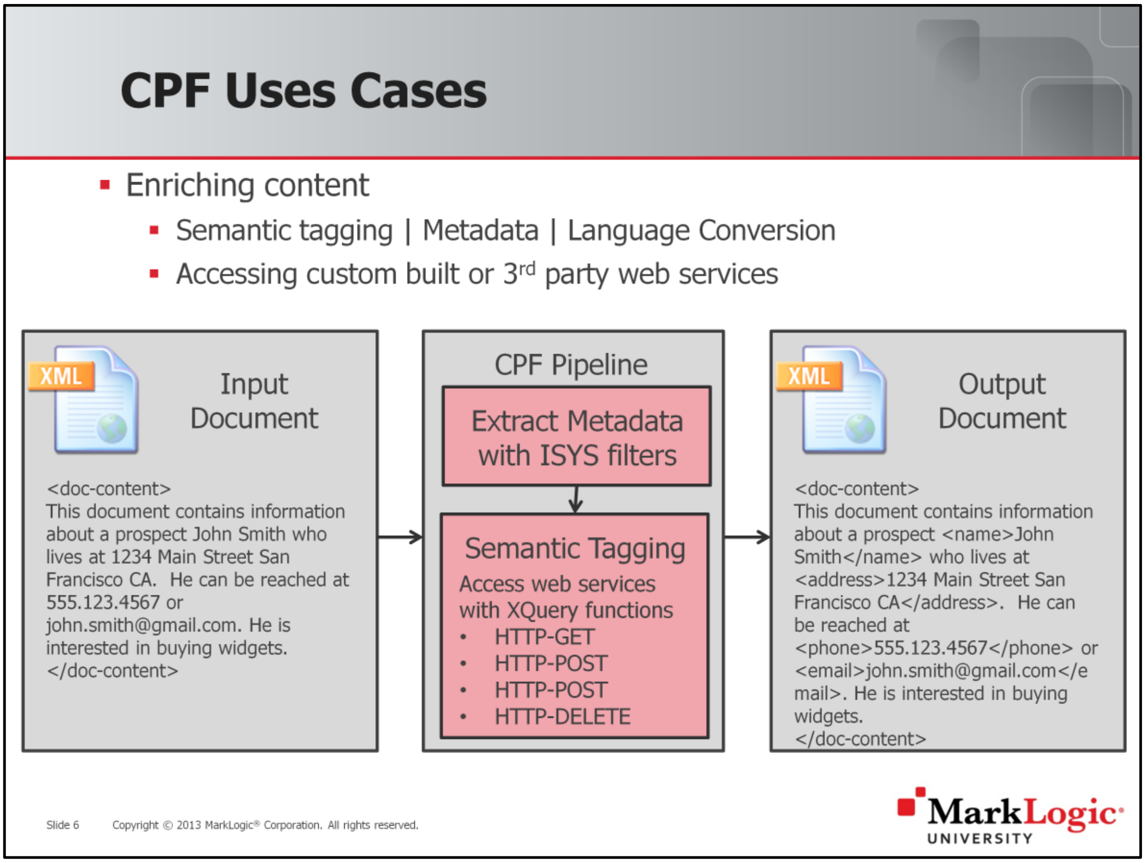


CPF Uses Cases

- Ability to link together different units of processing as a sequence
- Automation of processing
- Separate documents for different types of processing

XML
Input Document
Determine Root Node
xdmp:document-get()
<song>
<book>
<dvd>
Song Tasks
Book Tasks
DVD Tasks

Slide 5 Copyright © 2013 MarkLogic® Corporation. All rights reserved.



CPF Uses Cases

- Enriching content
  - Semantic tagging | Metadata | Language Conversion
  - Accessing custom built or 3rd party web services

<doc-content>
This document contains information about a prospect John Smith who lives at 1234 Main Street San Francisco CA. He can be reached at 555.123.4567 or john.smith@gmail.com. He is interested in buying widgets.
</doc-content>

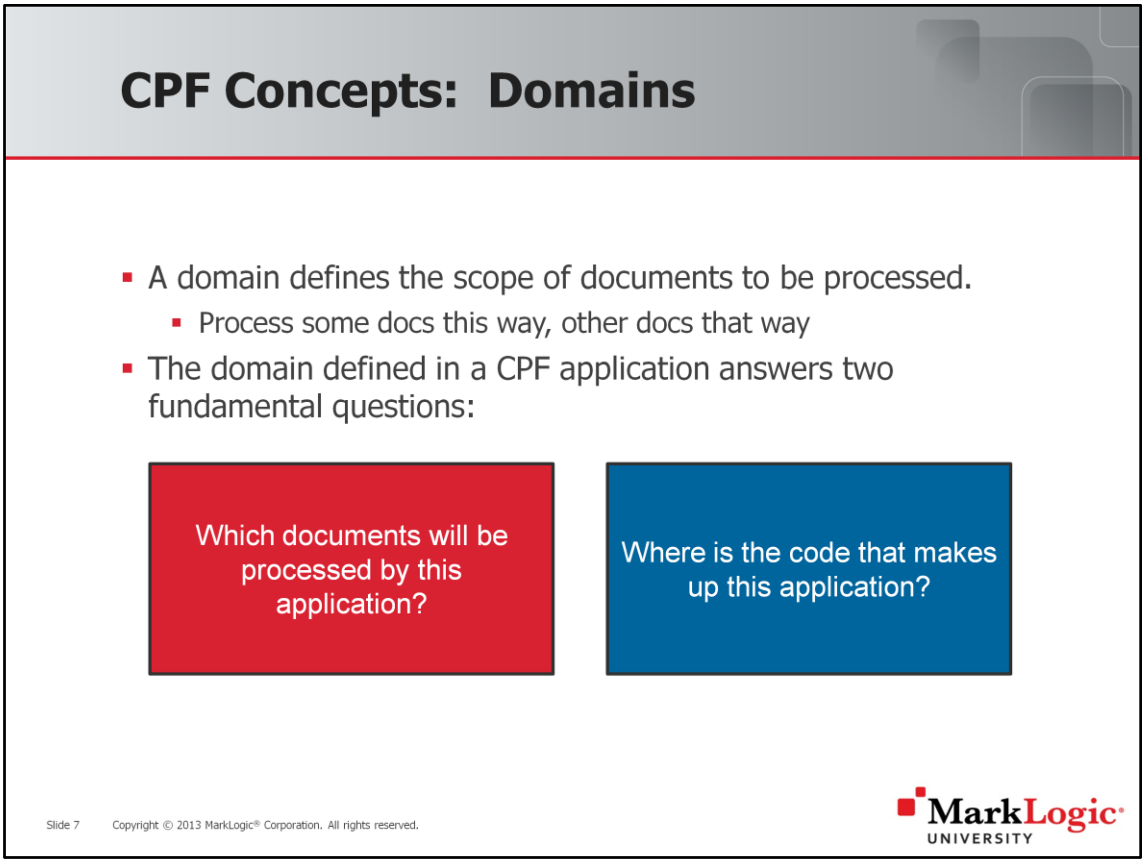
Semantic Tagging

Access web services with XQuery functions

- HTTP-GET
- HTTP-POST
- HTTP-POST
- HTTP-DELETE

<doc-content>
This document contains information about a prospect <name>John Smith</name> who lives at <address>1234 Main Street San Francisco CA</address>. He can be reached at <phone>555.123.4567</phone> or <email>john.smith@gmail.com</email>. He is interested in buying widgets.
</doc-content>

Slide 6



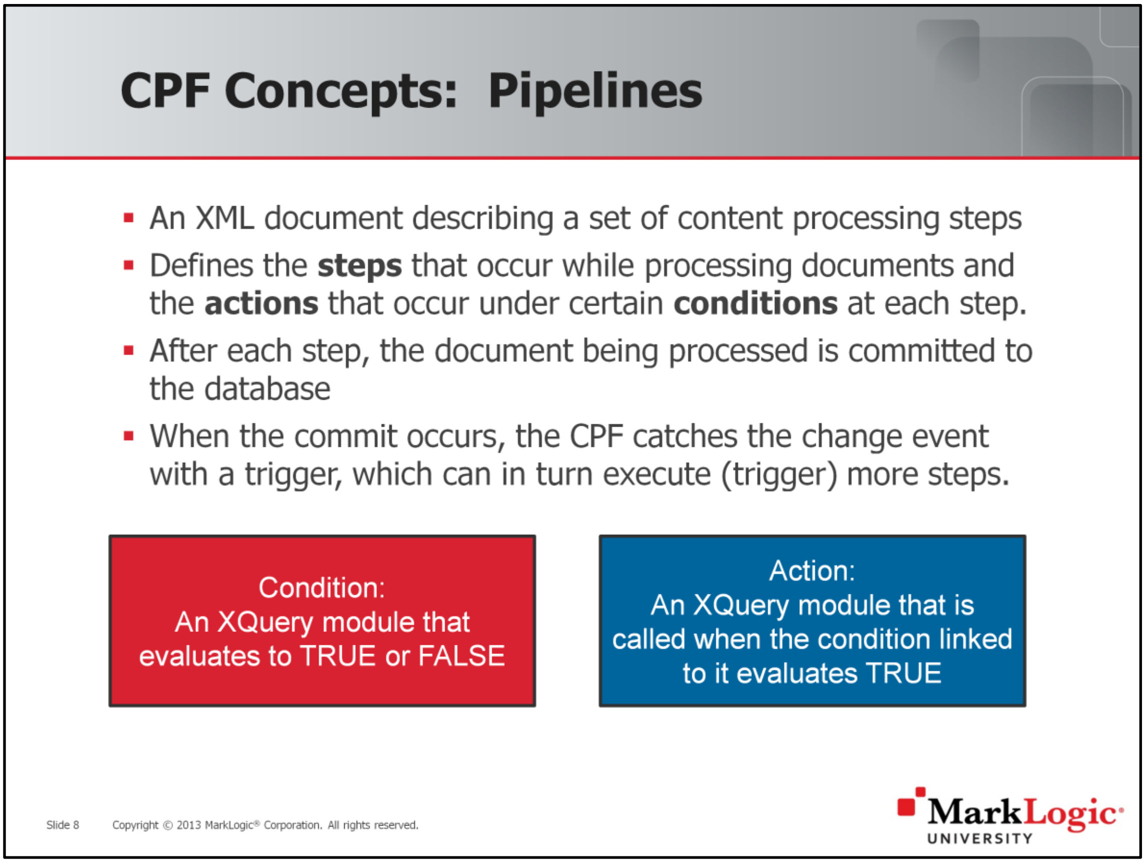
CPF Concepts: Domains

- A domain defines the scope of documents to be processed.
  - Process some docs this way, other docs that way
- The domain defined in a CPF application answers two fundamental questions:

Which documents will be processed by this application?

Where is the code that makes up this application?

Slide 7 Copyright © 2013 MarkLogic® Corporation. All rights reserved.



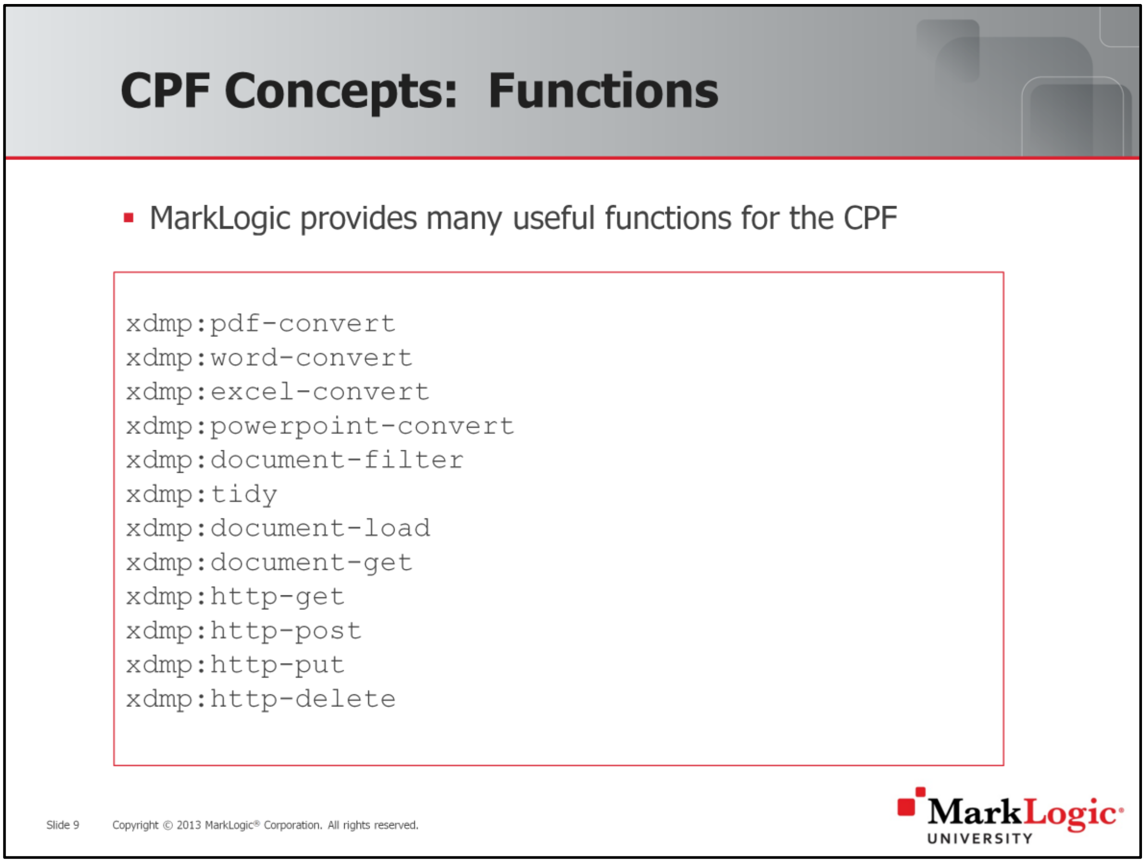
CPF Concepts: Pipelines

- An XML document describing a set of content processing steps
- Defines the **steps** that occur while processing documents and the **actions** that occur under certain **conditions** at each step.
- After each step, the document being processed is committed to the database
- When the commit occurs, the CPF catches the change event with a trigger, which can in turn execute (trigger) more steps.

Condition:
An XQuery module that evaluates to TRUE or FALSE

Action:
An XQuery module that is called when the condition linked to it evaluates TRUE

Slide 8

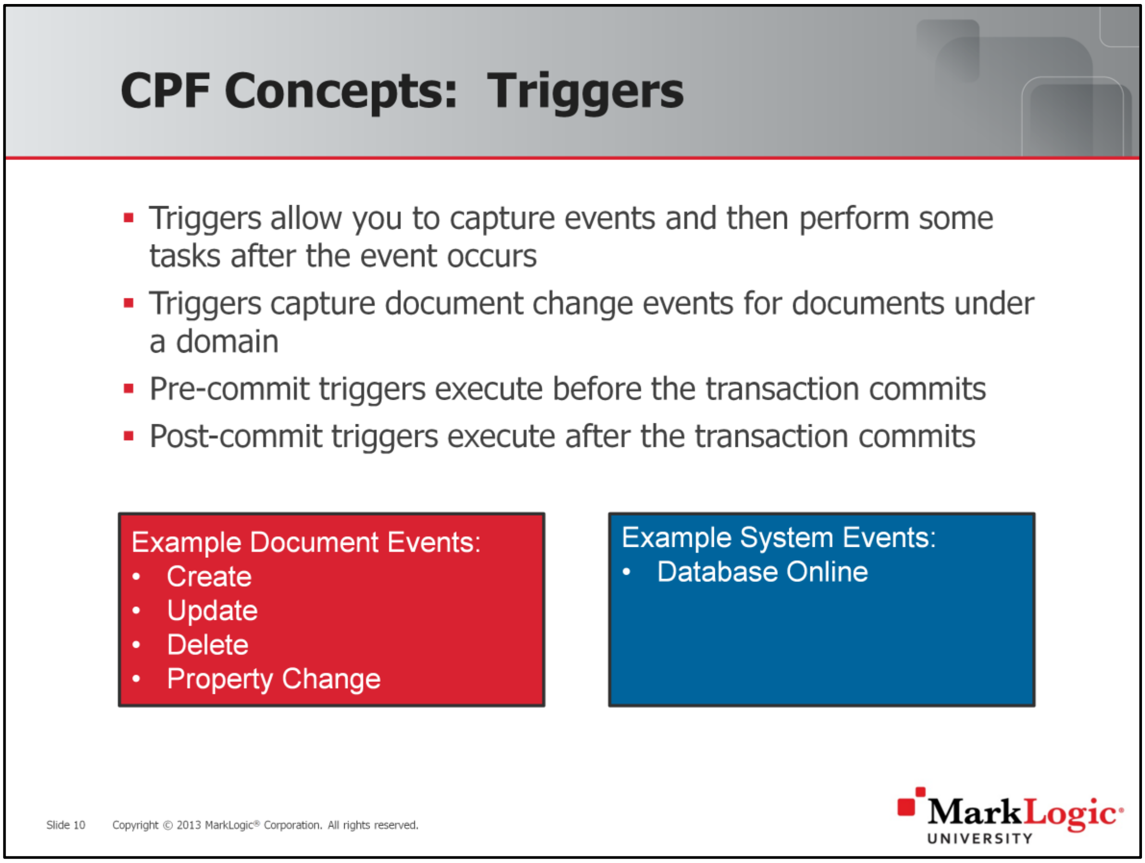


CPF Concepts: Functions

- MarkLogic provides many useful functions for the CPF

```
xdmp:pdf-convert
xdmp:word-convert
xdmp:excel-convert
xdmp:powerpoint-convert
xdmp:document-filter
xdmp:tidy
xdmp:document-load
xdmp:document-get
xdmp:http-get
xdmp:http-post
xdmp:http-put
xdmp:http-delete
```

Slide 9



CPF Concepts: Triggers

- Triggers allow you to capture events and then perform some tasks after the event occurs
- Triggers capture document change events for documents under a domain
- Pre-commit triggers execute before the transaction commits
- Post-commit triggers execute after the transaction commits

Example Document Events:

- Create
- Update
- Delete
- Property Change

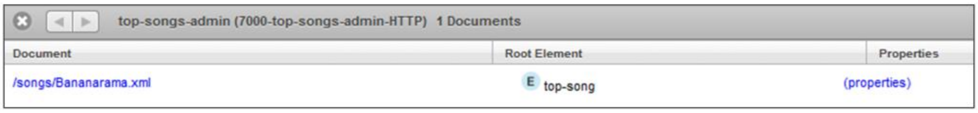
Example System Events:

- Database Online

Slide 10

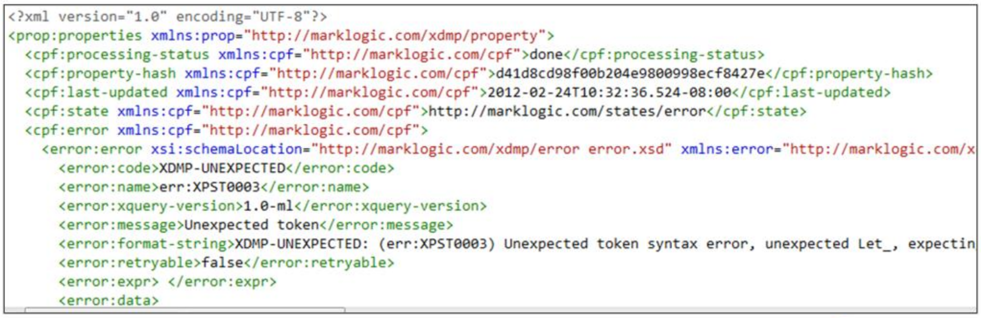
Debugging CPF

- Limit the scope of your testing
  - A single document is a good unit for initial testing
- Check the properties fragment for each document URI



top-songs-admin (7000-top-songs-admin-HTTP) 1 Documents

Document	Root Element	Properties
/songs/Bananarama.xml	E top-song	(properties)



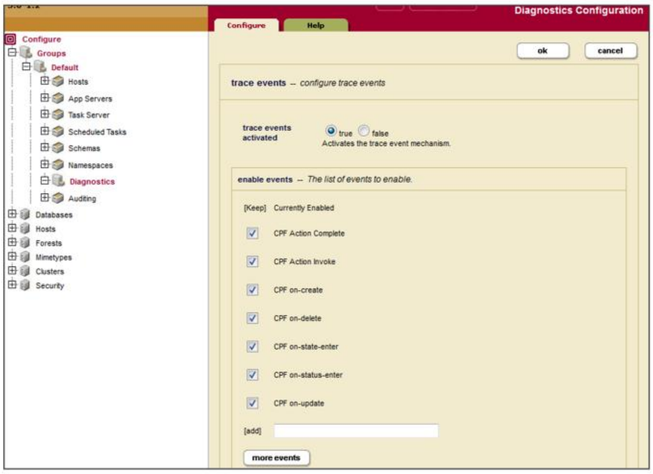
```
<?xml version="1.0" encoding="UTF-8"?>
<prop:properties xmlns:prop="http://marklogic.com/xdmp/property">
  <cpf:processing-status xmlns:cpf="http://marklogic.com/cpf">done</cpf:processing-status>
  <cpf:property-hash xmlns:cpf="http://marklogic.com/cpf">d41d8cd98f00b204e9800998ecf8427e</cpf:property-hash>
  <cpf:last-updated xmlns:cpf="http://marklogic.com/cpf">2012-02-24T10:32:36.524-08:00</cpf:last-updated>
  <cpf:state xmlns:cpf="http://marklogic.com/cpf">http://marklogic.com/states/error</cpf:state>
  <cpf:error xmlns:cpf="http://marklogic.com/cpf">
    <error:error xsi:schemaLocation="http://marklogic.com/xdmp/error error.xsd" xmlns:error="http://marklogic.com/x
      <error:code>XDMP-UNEXPECTED</error:code>
      <error:name>err:XPST0003</error:name>
      <error:xquery-version>1.0-ml</error:xquery-version>
      <error:message>Unexpected token</error:message>
      <error:format-string>XDMP-UNEXPECTED: (err:XPST0003) Unexpected token syntax error, unexpected Let_, expectin
      <error:retryable>false</error:retryable>
      <error:expr> </error:expr>
      <error:data>
```

Slide 11 Copyright © 2013 MarkLogic® Corporation. All rights reserved.

MarkLogic UNIVERSITY

Debugging CPF

- Enable trace events
  - Configure→Groups→Group Name→Diagnostics
  - Turn off when finished debugging!!!



trace events -- *configure trace events*

trace events activated ● true ○ false
Activates the trace event mechanism.

enable events -- *The list of events to enable.*

[Keep] Currently Enabled

☑ CPF Action Complete

☑ CPF Action Invoke

☑ CPF on-create

☑ CPF on-delete

☑ CPF on-state-enter

☑ CPF on-status-enter

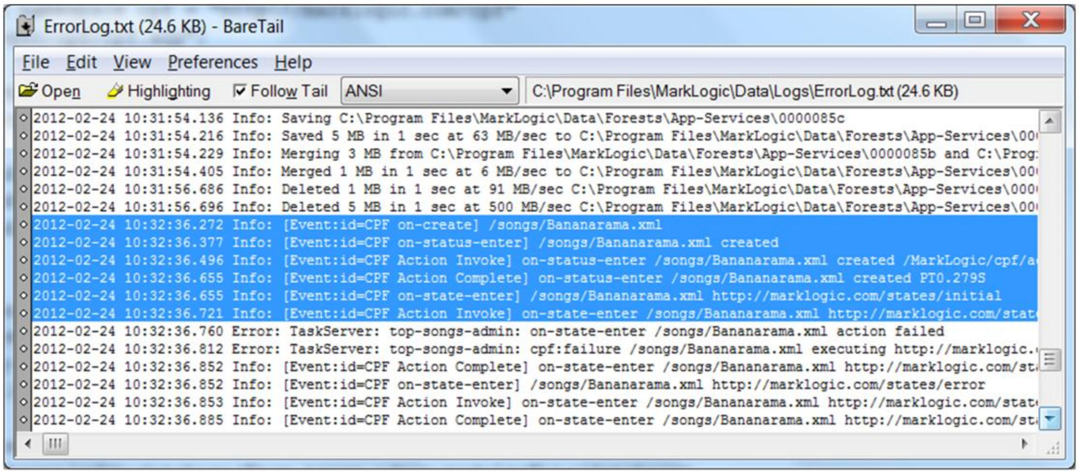
☑ CPF on-update

[add]

Slide 12

Debugging CPF

- Error Log
  - Setup a tail on errorlog.txt to view it live



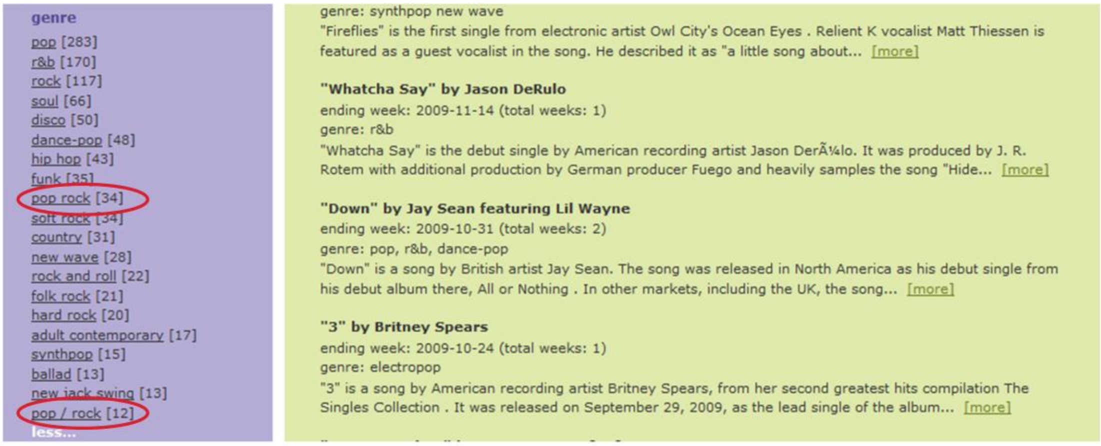
C:\Program Files\MarkLogic\Data\Logs\ErrorLog.txt (24.6 KB)

```
2012-02-24 10:31:54.136 Info: Saving C:\Program Files\MarkLogic\Data\Forests\App-Services\0000085c
2012-02-24 10:31:54.216 Info: Saved 5 MB in 1 sec at 63 MB/sec to C:\Program Files\MarkLogic\Data\Forests\App-Services\00
2012-02-24 10:31:54.229 Info: Merging 3 MB from C:\Program Files\MarkLogic\Data\Forests\App-Services\0000085b and C:\Prog
2012-02-24 10:31:54.405 Info: Merged 1 MB in 1 sec at 6 MB/sec to C:\Program Files\MarkLogic\Data\Forests\App-Services\00
2012-02-24 10:31:56.686 Info: Deleted 1 MB in 1 sec at 91 MB/sec C:\Program Files\MarkLogic\Data\Forests\App-Services\000
2012-02-24 10:31:56.696 Info: Deleted 5 MB in 1 sec at 500 MB/sec C:\Program Files\MarkLogic\Data\Forests\App-Services\00
2012-02-24 10:32:36.272 Info: [Event:id=CPF on-create] /songs/Bananarama.xml
2012-02-24 10:32:36.377 Info: [Event:id=CPF on-status-enter] /songs/Bananarama.xml created
2012-02-24 10:32:36.496 Info: [Event:id=CPF Action Invoke] on-status-enter /songs/Bananarama.xml created /MarkLogic/cpf/a
2012-02-24 10:32:36.655 Info: [Event:id=CPF Action Complete] on-status-enter /songs/Bananarama.xml created PT0.279S
2012-02-24 10:32:36.655 Info: [Event:id=CPF on-state-enter] /songs/Bananarama.xml http://marklogic.com/states/initial
2012-02-24 10:32:36.721 Info: [Event:id=CPF Action Invoke] on-state-enter /songs/Bananarama.xml http://marklogic.com/stat
2012-02-24 10:32:36.760 Error: TaskServer: top-songs-admin: on-state-enter /songs/Bananarama.xml action failed
2012-02-24 10:32:36.812 Error: TaskServer: top-songs-admin: cpf:failure /songs/Bananarama.xml executing http://marklogic.
2012-02-24 10:32:36.852 Info: [Event:id=CPF Action Complete] on-state-enter /songs/Bananarama.xml http://marklogic.com/st
2012-02-24 10:32:36.852 Info: [Event:id=CPF on-state-enter] /songs/Bananarama.xml http://marklogic.com/states/error
2012-02-24 10:32:36.853 Info: [Event:id=CPF Action Invoke] on-state-enter /songs/Bananarama.xml http://marklogic.com/stat
2012-02-24 10:32:36.885 Info: [Event:id=CPF Action Complete] on-state-enter /songs/Bananarama.xml http://marklogic.com/st
```

Slide 13 Copyright © 2013 MarkLogic® Corporation. All rights reserved.

Applying CPF Concepts

- <genre> data in our top-songs application
- The collation in our range index does not handle all variations
- <genre> could be normalized upon ingestion using CPF



genre
pop [283]
r&b [170]
rock [117]
soul [66]
disco [50]
dance-pop [48]
hip hop [43]
funk [35]
pop rock [34]
soft rock [34]
country [31]
new wave [28]
rock and roll [22]
folk rock [21]
hard rock [20]
adult contemporary [17]
synthpop [15]
ballad [13]
new jack swing [13]
pop / rock [12]
less...

genre: synthpop new wave
"Fireflies" is the first single from electronic artist Owl City's Ocean Eyes . Relient K vocalist Matt Thiessen is featured as a guest vocalist in the song. He described it as "a little song about... [more]

"Whatcha Say" by Jason DeRulo
ending week: 2009-11-14 (total weeks: 1)
genre: r&b
"Whatcha Say" is the debut single by American recording artist Jason DerÃ¼lo. It was produced by J. R. Rotem with additional production by German producer Fuego and heavily samples the song "Hide... [more]

"Down" by Jay Sean featuring Lil Wayne
ending week: 2009-10-31 (total weeks: 2)
genre: pop, r&b, dance-pop
"Down" is a song by British artist Jay Sean. The song was released in North America as his debut single from his debut album there, All or Nothing . In other markets, including the UK, the song... [more]

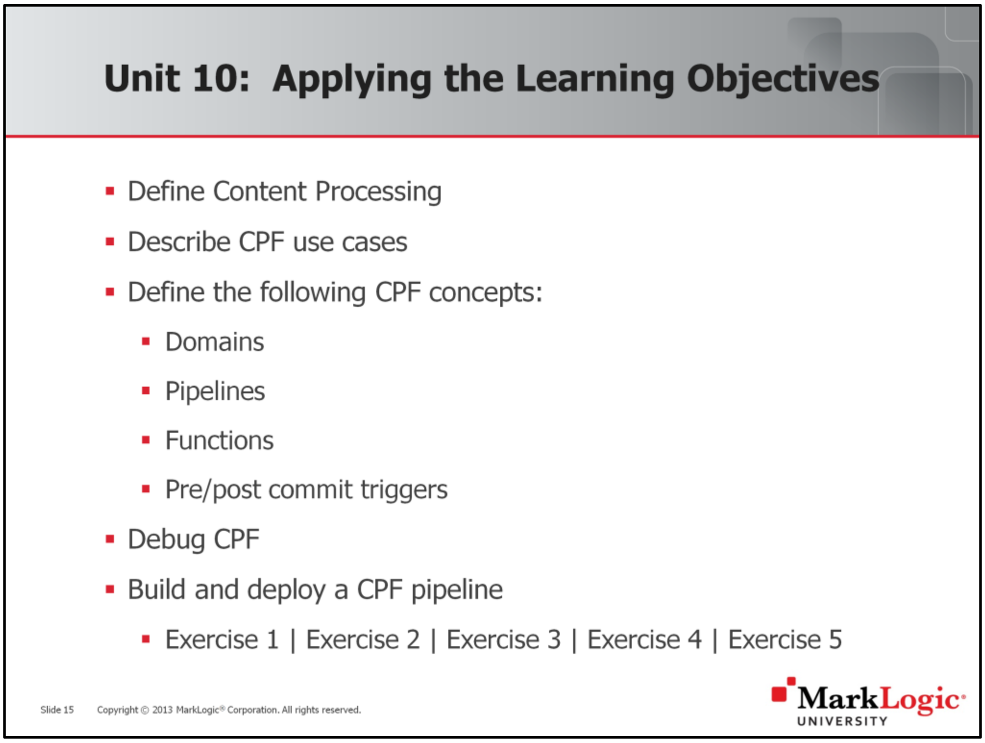
"3" by Britney Spears
ending week: 2009-10-24 (total weeks: 1)
genre: electropop
"3" is a song by American recording artist Britney Spears, from her second greatest hits compilation The Singles Collection . It was released on September 29, 2009, as the lead single of the album... [more]

Slide 14

Our genre facet would be more useful if there was better consistency amongst our genre data in our documents.

For example, look at the facet and notice an entry for “pop rock” and also “pop / rock”.

Another example, what is “synthpop”? And should it differ from “dance-pop”? Or should both simply be “pop”? We can build a CPF application to define those rules and normalize the <genre> data as the documents are loaded.



Unit 10: Applying the Learning Objectives

- Define Content Processing
- Describe CPF use cases
- Define the following CPF concepts:
  - Domains
  - Pipelines
  - Functions
  - Pre/post commit triggers
- Debug CPF
- Build and deploy a CPF pipeline
  - Exercise 1 | Exercise 2 | Exercise 3 | Exercise 4 | Exercise 5

Slide 15