

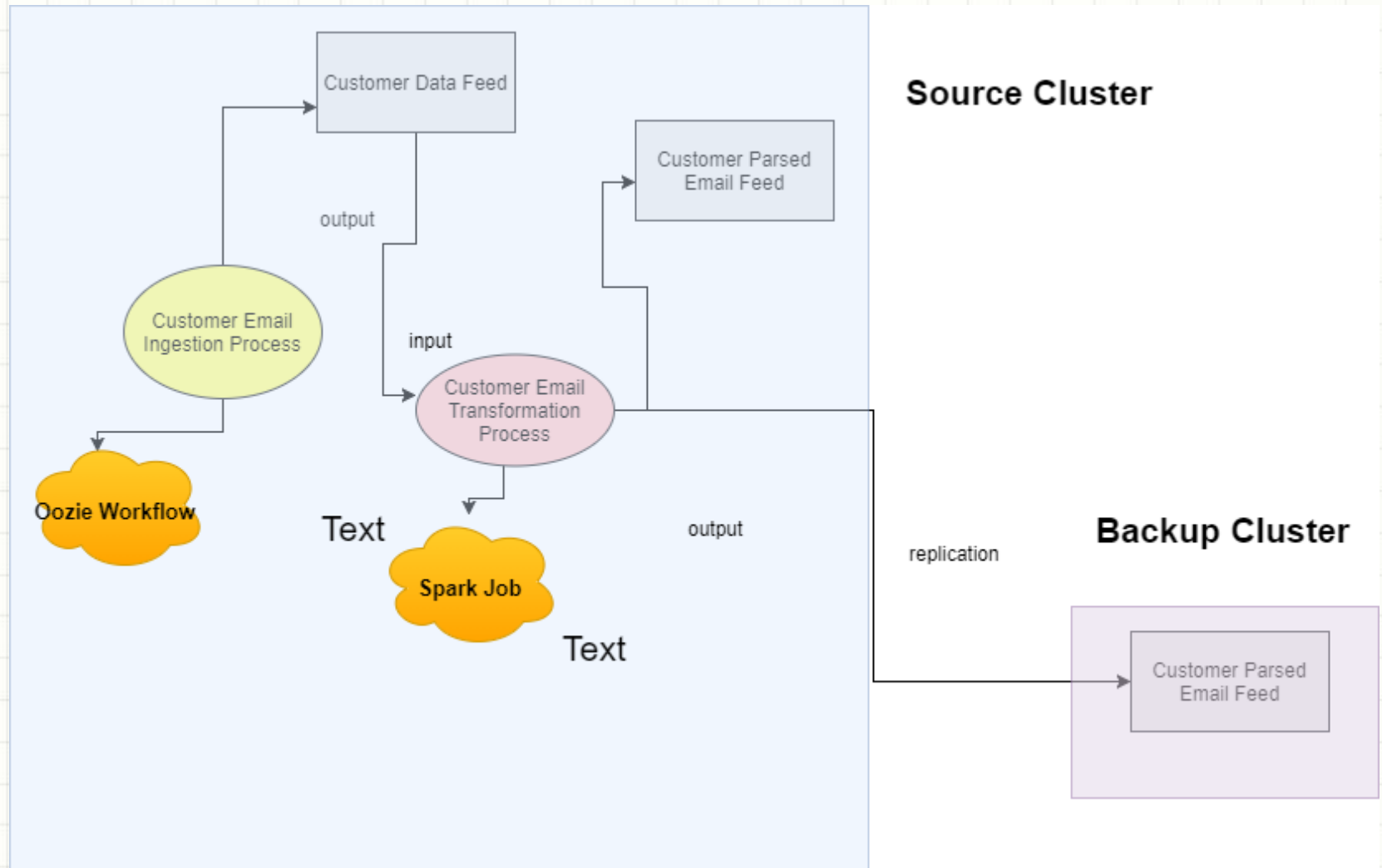


# **APACHE FALCON DATA MANAGEMENT LABS**

Bin Jiang

04/08/2017

# Customer Churning Lab



# Customer Churning Lab

- **Customer Email Format**

```
-----  
hdfs://sandbox-hdp.hortonworks.com:8020/user/root/falcon/customer/input/2017-12-27-21/106590.txt  
-----
```

Message-ID: <20866019.1075863331634.JavaMail.evans@thyme>

Date: Mon, 5 Mar 2001 19:23:00 -0800 (PST)

From: drew.foosum@enron.com

To: darrell.schoolcraft@enron.com

Subject: TW Gas Sales: PRIVILEGED AND CONFIDENTIAL ATTORNEY CLIENT PRIVILEGE

Cc: danny.mccarty@enron.com, steven.harris@enron.com, kevin.hyatt@enron.com

Mime-Version: 1.0

Content-Type: text/plain; charset=us-ascii

Content-Transfer-Encoding: 7bit

Bcc: danny.mccarty@enron.com, steven.harris@enron.com, kevin.hyatt@enron.com

X-From: Drew Fossum

X-To: Darrell Schoolcraft <Darrell Schoolcraft/ET&S/Enron@ENRON>

X-cc: Danny McCarty <Danny McCarty/ET&S/Enron@Enron>, Steven Harris <Steven Harris/ET&S/Enron@ENRON>, Kevin Hyatt <Kevin Hyatt/Enron@EnronXGate>

X-bcc:

X-Folder: \DFOSSUM (Non-Privileged)\Fossum, Drew\Sent Mail

X-Origin: Fossum-D

X-FileName: DFOSSUM (Non-Privileged).pst

In anticipation of potential litigation involving TW's operational activities, please prepare an analysis for me of the reasons for TW's sale of excess natural gas at the California border. I am aware of several of these sales and have been informed that excess pressure at the border is the basic reason for them. I'd like a more specific explanation that includes the following information:

# Customer Churning Lab

- **Change Falcon Default Port**

localhost:8080/#/main/services/FALCON/configs

Ambari Sandbox 0 ops 0 alerts Dashboard Services Hosts Alerts Admin admin

Summary Configs Quick Links Service Actions

Group Default (1) Manage Config Groups Filter...

V2 internal 2 months ago HDP-2.6 V1 admin 2 months ago HDP-2.6

V2 internal authored on Fri, Nov 10, 2017 09:17 Discard Save

Falcon Server

Falcon Server host sandbox-hdp.hortonworks.com

Falcon data directory /hadoop/falcon

Falcon server port 15500

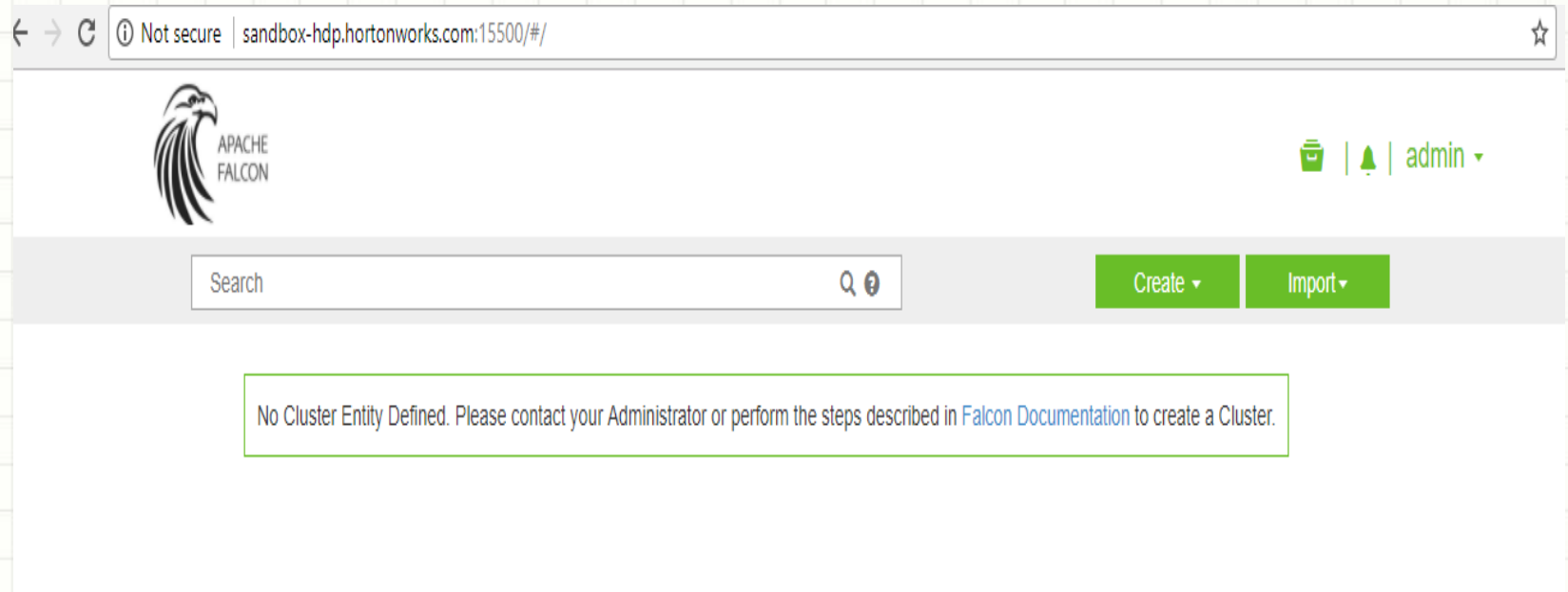
Falcon startup.properties

\*.ConfigSyncService.impl org.apache.falcon.resource.ConfigSyncService

\*.ProcessInstance org.apache.falcon.resource.InstanceManager

# Customer Churning Lab

- **Apache Falcon Console**



# Customer Churning Lab

- Source Cluster



## NEW CLUSTER

Cluster Name\*

SourceCluster

Name available

Data Center or Colo Name\*

SourceDC

Description

Tags

SourceTag

DataIngestion

+ add tag

## INTERFACES

Type

Endpoint

File System Read Endpoint Address

hftp://sandbox-hdp.hortonworks.com:50070

File System Default Address

hdfs://sandbox-hdp.hortonworks.com:8020

Yarn Resource Manager Address

sandbox-hdp.hortonworks.com:8032

Workflow Address

http://sandbox-hdp.hortonworks.com:11000/oozie/

Message Broker Address

tcp://sandbox-hdp.hortonworks.com:61616?daemon=true

☐ Metadata Catalog Registry

thrift://<hostname>:9083

☒ Spark

☐ Yarn Cluster ☐ Yarn Client ☒ Local ☐ Custom

## PROPERTIES

# Customer Churning Lab

- Customer Feed

## NEW FEED

Feed Name\*

customerFeed

Name available

Description

customer data feed

Tags

customerFeedTag

customerFeed

+ add tag

Feed Groups (comma separated)

churnAnalysisDataPipeline

Type\*

HDFS

☐ Enable Replication

Source(s)

Cluster\*

SourceCluster

Statistics Path

/tmp/\${YEAR}-\${MONTH}-\${DAY}-\${HC

Data Path\*

/user/root/falcon/customer/input/

Start\*

12/30/2017

[MORE OPTIONS >](#)

Time\*

11

: 31

AM

End\*

12/31/2099

Time\*

# Customer Churning Lab

- **Customer Email Ingestion Process**

## NEW PROCESS

Process Name\*

customerIngestProcess

Tags

customerIngestionTag

customerIngestion

+ ADD

### Details

Engine\*

Oozie

INPUT(S)

+ ADD

Workflow Name\*

customerIngestWorkflow

Workflow Path\*

/user/root/falcon

Cluster\*

SourceCluster

OUTPUT(S)

Name\*

output

Feed\*

customerFeed

INSTANCE

Instance\*

now(0,0)

- delete

+ ADD

Run Duration\*

Start

12/30/2017

11

34

AM



# Customer Churning Lab

- Customer Email Ingestion Process

## Frequency

Repeat Every\*

minutes ▼

Timezone\*

UTC ▼

## ADVANCED OPTIONS ^

### Retry Policy

Type

Periodic ▼

Delay Up to

minutes ▼

Attempts

### Performance & Ordering

Max Parallel Instances

1 ▼

Order

FIFO ▼

### Properties

jobTracker

sandbox-hdp.hortonworks.com:8032

- delete

nameNode

hdfs://sandbox-hdp.hortonworks.com:8020

- delete

queueName

default

- delete

+ ADD

### Access Control List

Owner\*

root

Group\*

users

### Permissions\*

Read

Write

Execute

# Customer Churning Lab

- Transformation Cluster



## NEW CLUSTER

Cluster Name\*

TransformationCluster

Name available

Data Center or Colo Name\*

TransformationDC

Description

Tags

TransformationTag

DataTransformation

+ add tag

## INTERFACES

Type

Endpoint

File System Read Endpoint Address

hftp://sandbox-hdp.hortonworks.com:50070

File System Default Address

hdfs://sandbox-hdp.hortonworks.com:8020

Yarn Resource Manager Address

sandbox-hdp.hortonworks.com:8032

Workflow Address

http://sandbox-hdp.hortonworks.com:11000/oozie/

Message Broker Address

tcp://sandbox-hdp.hortonworks.com:61616?daemon=true

☐ Metadata Catalog Registry

thrift://<hostname>:9083

☒ Spark

☐ Yarn Cluster

☐ Yarn Client

☒ Local

☐ Custom

## PROPERTIES

# Customer Churning Lab

- Transformation Cluster

Type	Endpoint
File System Read Endpoint Address	<input type="text" value="hftp://sandbox-hdp.hortonworks.com:50070"/>
<div>URI for read operations Eg. hdfs://localhost:50070   webhdfs://localhost:50070   hftp://localhost:50070</div>	<input type="text" value="hdfs://sandbox-hdp.hortonworks.com:8020"/>
Hadoop Resource Manager Address	<input type="text" value="sandbox-hdp.hortonworks.com:8032"/>
Workflow Address	<input type="text" value="http://sandbox-hdp.hortonworks.com:11000/oozie/"/>
Message Broker Address	<input type="text" value="tcp://sandbox-hdp.hortonworks.com:61616?daemon=true"/>
<input type="checkbox"/> Metadata Catalog Registry	<input type="text" value="thrift://&lt;hostname&gt;:9083"/>
<input checked="" type="checkbox"/> Spark	<div><input type="radio"/> Yarn Cluster <input type="radio"/> Yarn Client <input checked="" type="radio"/> Local <input type="radio"/> Custom</div>

PROPERTIES

Property Name	Value	
<input type="text" value="name"/>	<input type="text" value="value"/>	+ add property

LOCATION

Location Name	Path	
Staging*	<input type="text" value="/apps/falcon/TransformationCluster/staging"/>	
Temp*	<input type="text" value="/tmp"/>	
Working*	<input type="text" value="/apps/falcon/TransformationCluster/working"/>	+ add location

ADVANCED OPTIONS ▼

CANCEL

NEXT

# Customer Churning Lab

- **Parsed Customer Email Feed**

## NEW FEED

Feed Name\*

parsedCustomerFeed

Name available

Description

parsed customer emails

Tags

parsedCustomerTag

parsedCustomer

+ add tag

Feed Groups (comma separated)

churnAnalysisDataPipeline

Type\*

HDFS

☒ Enable Replication

### Source(s)

Cluster\*

SourceCluster

Statistics Path

/tmp/\${YEAR}-\${MONTH}-\${DAY}-\${HOUR}

Data Path\*

/user/root/falcon/customer/output

Start\*

12/30/2017

[MORE OPTIONS >](#)

Time\*

11

:

42

AM

End\*

12/31/2099

### Target(s)

Cluster\*

TransformationCluster

Statistics Path

/tmp/\${YEAR}-\${MONTH}-\${DAY}-\${HOUR}

Data Path\*

/user/root/falcon/customer/output

Start\*

12/30/2017

[MORE OPTIONS >](#)

Time\*

11

:

42

AM

End\*

12/31/2099

# Customer Churning Lab

- Customer Email Transformation Process

## NEW PROCESS

Process Name\*

customerTransformationProcess

Name available

Tags

customerTransformationTa

customerTransformation

+ ADD

### Details

Engine\*

Spark

Workflow Path\*

/user/root/falcon

Cluster\*

SourceCluster

Name\*

CustomerTransformationApplication

Application\*

/user/root/falcon/DataManagementOnFal

Main Class

ca.training.bigdata.falcon.churn.Custome

Runs On

Local

Spark Options

--driver-memory 2G --executor-memory 2

Spark Arguments

### INPUT(S)

Name\*

input

Feed\*

customerFeed

#### INSTANCE

Start\*

now(0,0)

End\*

now(0,0)

- delete

+ ADD

### OUTPUT(S)

Name\*

output

Feed\*

parsedCustomerFeed

#### INSTANCE

Instance\*

now(0,0)

- delete

+ ADD

Run Duration\*

Start

12/30/2017

11

: 45

AM

# Customer Churning Lab

- Customer Email Transformation Process

12/29/2017 11 : 45 AM

End

12/31/2099 11 : 59 AM

Frequency

Repeat Every\*

30 minutes ▾

Timezone\*

UTC ▾

ADVANCED OPTIONS ^

Retry Policy

Type

Periodic ▾

Delay Up to

30 minutes ▾

Attempts

3

Performance & Ordering

Max Parallel Instances

1 ▾

Order

FIFO ▾

Properties

oozie.action.sharelib.for.spark spark2 + ADD

Access Control List

Owner\*

root

Group\*







users

Permissions\*

Read Write Execute

# Customer Churning Lab

- Submitted Feeds and Process

<input type="text" value="Search"/>		 		Create ▾	Import ▾
<input type="checkbox"/> Name ↑↓	Tags	Cluster	Type	Status	
▶ Schedule ▶ Resume    Pause ⚙ Edit ↗ Copy 🗑 Delete ⚡ XML					
<input type="checkbox"/>  parsedCustomerFeed	parsedCustomerTag=parsedCustomer	SourceCluster, TransformationCluster		SUBMITTED	
<input type="checkbox"/>  customerFeed	customerFeedTag=customerFeed	SourceCluster		SUBMITTED	
<input type="checkbox"/>  customerTransformationProcess	customerTransformationTag=customerTransformation	SourceCluster		SUBMITTED	
<input type="checkbox"/>  customerIngestProcess	customerIngestionTag=customerIngestion	SourceCluster		SUBMITTED	
					1

# Customer Churning Lab

- **Parsed Customer Email Schema**

```
17/12/31 03:27:31 INFO StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint
```

```
root
```

```
|-- message_id: string (nullable = true)
|-- edate: string (nullable = true)
|-- efrom: string (nullable = true)
|-- eto: string (nullable = true)
|-- subject: string (nullable = true)
|-- cc: string (nullable = true)
|-- mime_type: string (nullable = true)
|-- content_type: string (nullable = true)
|-- content_transfer_encoding: string (nullable = true)
|-- bcc: string (nullable = true)
|-- x_from: string (nullable = true)
|-- x_to: string (nullable = true)
|-- x_cc: string (nullable = true)
|-- x_bcc: string (nullable = true)
|-- x_folder: string (nullable = true)
|-- x_origin: string (nullable = true)
|-- x_filename: string (nullable = true)
```

```
17/12/31 03:27:35 INFO ParquetFileFormat: Using default output committer for Parquet: org.apache.parquet.hadoop.ParquetOutputCommitter
```



# Customer Churning Lab

- **Create External Table**

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:1> CREATE EXTERNAL TABLE IF NOT EXISTS customer_churning ( message_id string, edate string, efrom string, eto string, subject string, cc string, mime_type string, content_type string, content_transfer_encoding string, bcc string, x_from string, x_to string, x_cc string, x_bcc string, x_folder string, x_origin string, x_filename string) STORED AS PARQUET LOCATION '/user/root/falcon/customer/output/2017-12-29-11';
```

No rows affected (1.469 seconds)

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:1> show tables;
```

```
+-----+
|  tab_name  |
+-----+
| business  |
| customer_churning |
+-----+
```