WHAT HAVE WE DONE SO FAR...?

Set up a workflow

Set up a coordinator for a workflow

Specified time based and data based triggers

Now you're ready to manage **data pipelines**

BUNDLES

BUNDLES

We have workflows to define actions
and coordinators to manage workflows
and specify dependencies

Why do we need bundles?

BUNDLES

Consider a very large e-commerce site

Each order on this site is recorded with:

order id

product id

customer id

cost

delivery zip code

and so on...

order id
customer id
product id
cost
delivery zip code

BUNDLES

We might have a job which cleans this data

Another job which calculates daily revenue

Another job which produces per user insights

Another job which calculates produce sales

Another job which sees geographic distribution of sales

order id
customer id
product id
cost
delivery zip code

BUNDLES

Another job which calculates daily revenue

We might have a job which cleans this data

Another job which sees geographic distribution of sales

Another job which calculates produce sales

Another job which produces per user insights

Each job will have a workflow of it's own and a coordinator to run it at required frequencies

order id
customer id
product id
cost
delivery zip code

BUNDLES

Another job which
calculates daily revenue

We might have a job which
cleans this data

Another job which sees
geographic distribution of sales

Another job which
calculates produce sales

Another job which
produces per user insights

All these jobs depend on
the same source data

order id
customer id
product id
cost
delivery zip code

BUNDLES

Another job which calculates daily revenue

We might have a job which cleans this data

Another job which sees geographic distribution of sales

Another job which calculates produce sales

Another job which produces per user insights

All these jobs depend on the same source data

They all probably run at different times and frequencies

order id
customer id
product id
cost
delivery zip code

BUNDLES

Another job which calculates daily revenue

We might have a job which cleans this data

Another job which sees geographic distribution of sales

Another job which calculates produce sales

Another job which produces per user insights

All these jobs depend on the **same source data**

They all probably run at **different** times and frequencies

If for some reason the source data is not available, **all of them** might need to be stopped together

order id
customer id
product id
cost
delivery zip code

BUNDLES

Another job which calculates daily revenue

We might have a job which cleans this data

Another job which sees geographic distribution of sales

Another job which calculates produce sales

Another job which produces per user insights

All these jobs depend on the **same source data**

They all probably run at **different** times and frequencies

And **restarted**, once the data is available once again

order id
customer id
product id
cost
delivery zip code

BUNDLES

We might have a job which cleans this data

Another job which calculates daily revenue

Another job which sees geographic distribution of sales

Another job which calculates produce sales

Another job which produces per user insights

All these jobs depend on the **same source data**

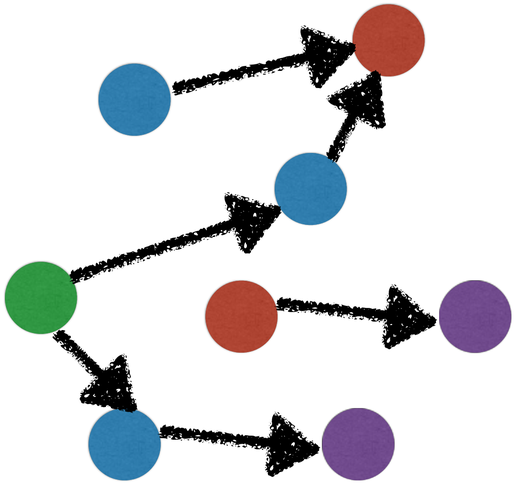
They all probably run at **different** times and frequencies

They should be treated as ONE UNIT

BUNDLES

They should be treated as
ONE UNIT

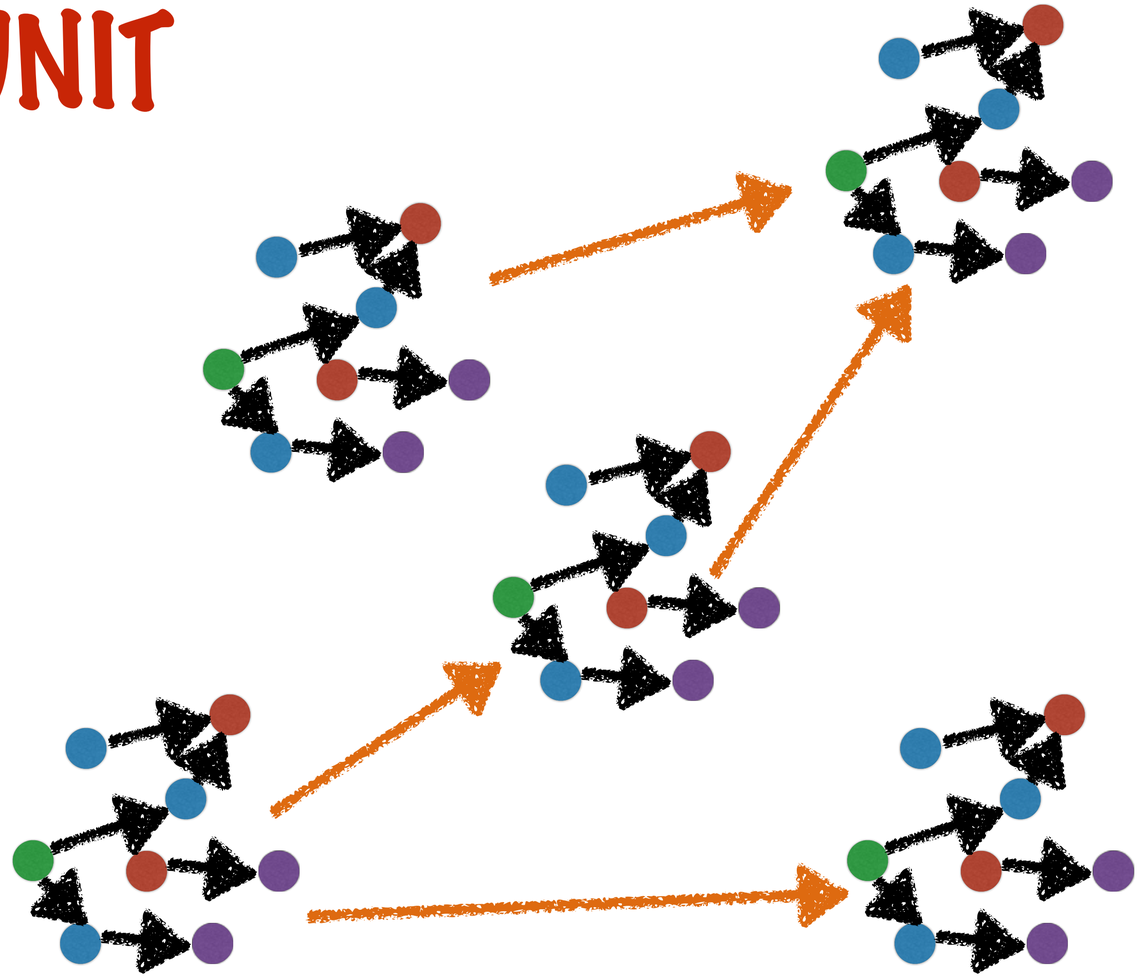
Each workflow can comprise of a number of actions



BUNDLES

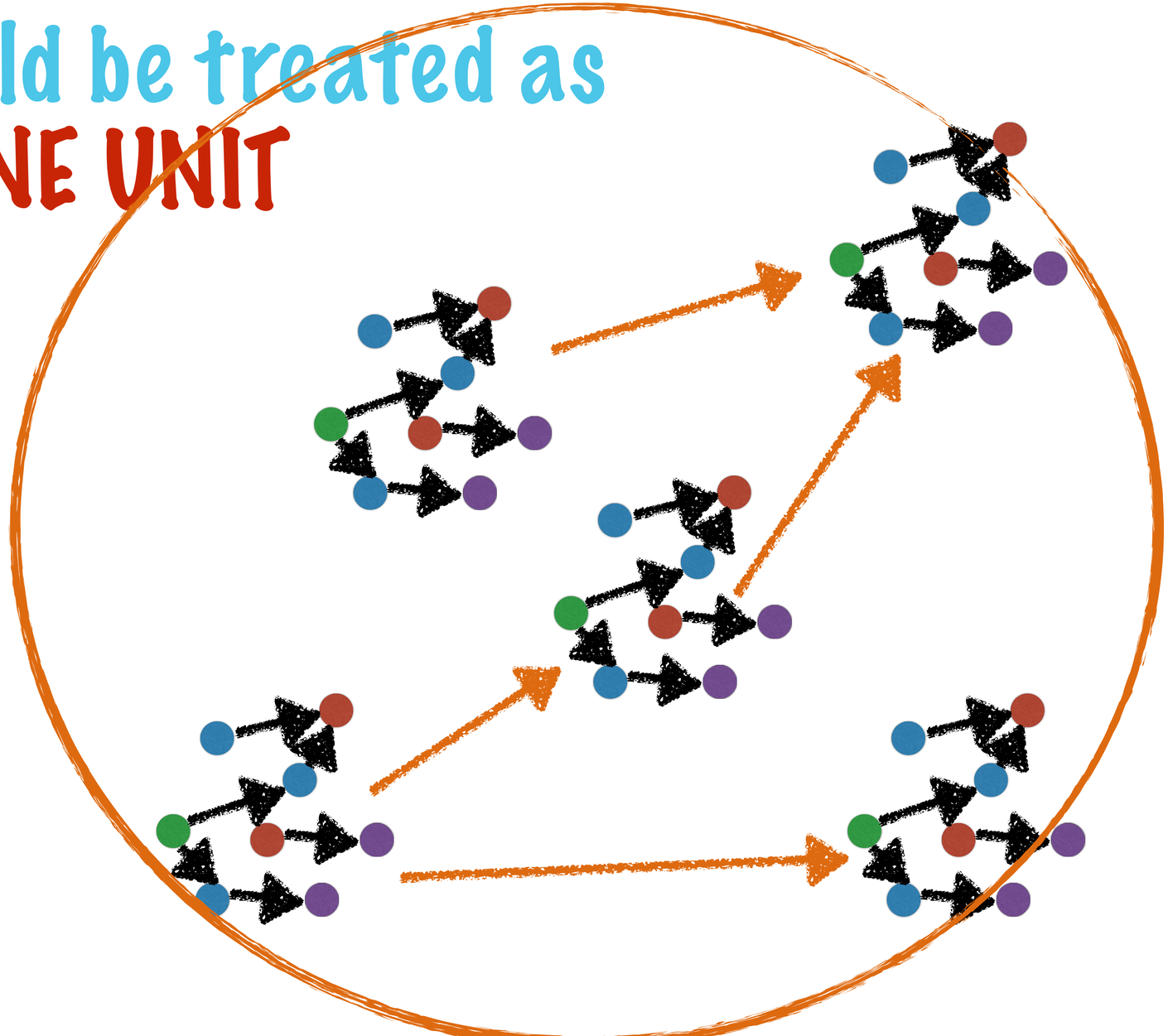
They should be treated as
ONE UNIT

And different workflows can depend on other workflow's data



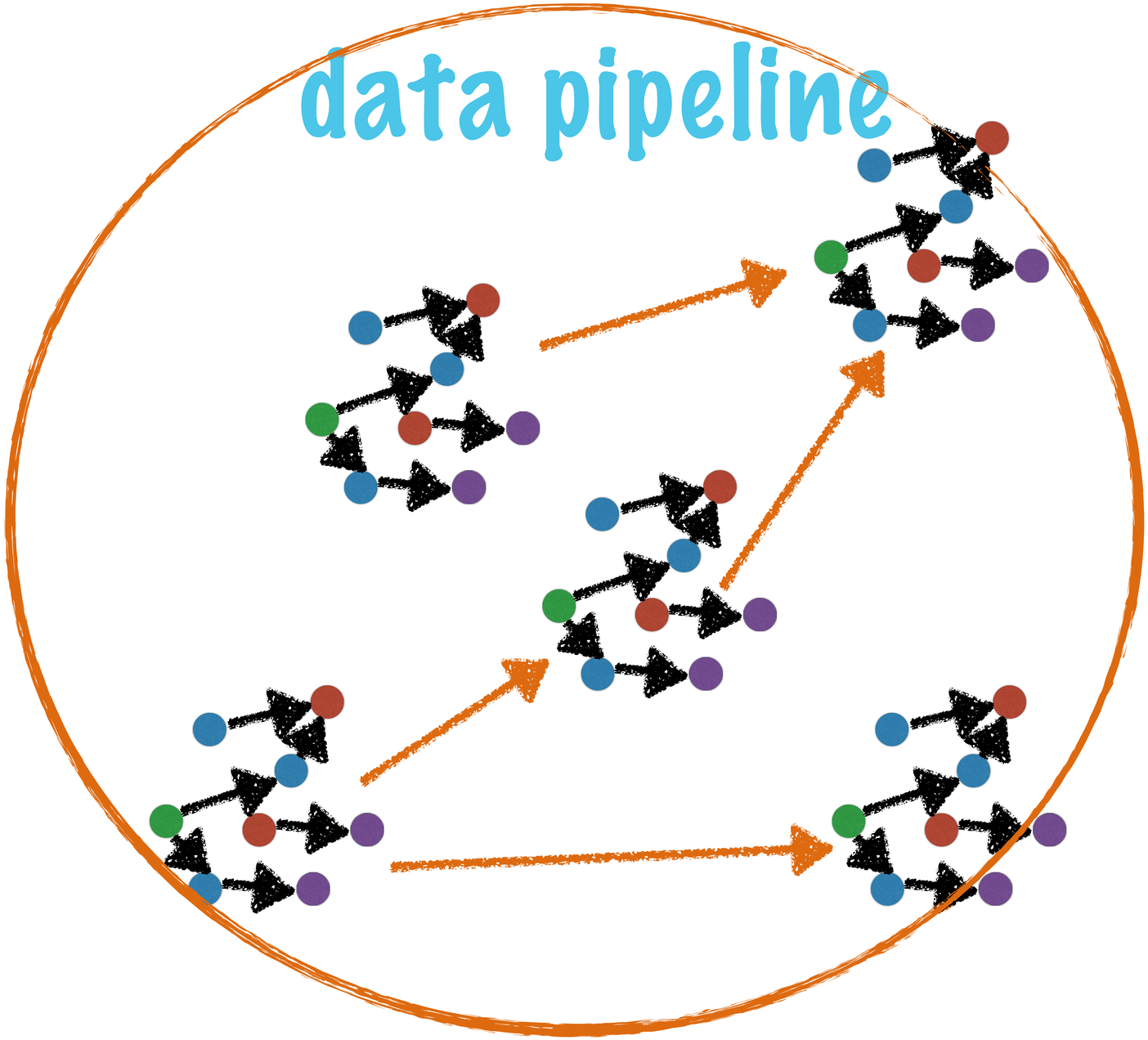
BUNDLES

They should be treated as
ONE UNIT



And they
together form a
data pipeline

BUNDLES



This is where we need a Bundle!

BUNDLES

Bundles are **abstractions** of a data pipeline, which make it **easy** and **convenient** to manage a huge number of workflows

BUNDLES

Insert video 5

BUNDLES

job.properties

```
nameNode=hdfs://localhost:9000
jobTracker=localhost:8032
queueName=default
oozieRoot=oozie

userName=${user.name}
start=2016-01-01T01:00Z
end=2016-01-01T03:00Z
oozie.bundle.application.path=${nameNode}/user/${user.name}/${oozieRoot}/bundle
```

BUNDLES

```
nameNode=hdfs://localhost:9000
jobTracker=localhost:8032
queueName=default
oozieRoot=oozie

userName=${user.name}
start=2016-01-01T01:00Z
end=2016-01-01T03:00Z
oozie.bundle.application.path=${nameNode}/user/${user.name}/${oozieRoot}/bundle
```

Nothing new in these specifications

BUNDLES

Other useful variables, especially the start time and end time

```
nameNode=hdfs://localhost:9000
jobTracker=localhost:8032
queueName=default
oozieRoot=oozie

userName=${user.name}
start=2016-01-01T01:00Z
end=2016-01-01T03:00Z
oozie.bundle.application.path=${nameNode}/user/${user.name}/${oozieRoot}/bundle
```

These can be passed to the coordinators of individual workflows if needed

BUNDLES

Specify the path to the bundle,
this will automatically look for
the bundle.xml file at this location

```
nameNode=hdfs://localhost:8020
jobTracker=localhost:8032
queueName=default
oozieRoot=oozie

userName=${user.name}
start=2016-01-01T01:00Z
end=2016-01-01T03:00Z

oozie.bundle.application.path=${nameNode}/user/${user.name}/${oozieRoot}/bundle
```

BUNDLES

bundle.xml

```
<bundle-app name='bundle-app' xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance' xmlns='uri:oozie:bundle:0.2'>
<parameters>
    <property>
        <name>start</name>
        <value>${start}</value>
    </property>
    <property>
        <name>end</name>
        <value>${end}</value>
    </property>
</parameters>
<controls>
    <kick-off-time>2016-07-01T00:00Z</kick-off-time>
</controls>
<coordinator name='coord-1'>
    <app-path>${nameNode}/user/${userName}/${oozieRoot}/aggregator/coordinator.xml</app-path>
</coordinator>
<coordinator name='coord-2'>
    <app-path>${nameNode}/user/${userName}/${oozieRoot}/aggregator/coordinator.xml</app-path>
    <configuration>
        <property>
            <name>start</name>
            <value>${start}</value>
        </property>
        <property>
            <name>end</name>
            <value>${end}</value>
        </property>
    </configuration>
</coordinator>
</bundle-app>
```

BUNDLES

```
<bundle-app name='bundle-app' xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance' xmlns='uri:oozie:bundle:0.2'>
<parameters>
    <property>
        <name>start</name>
        <value>${start}</value>
    </property>
    <property>
        <name>end</name>
        <value>${end}</value>
    </property>
</parameters>
<controls>
    <kick-off-time>2016-07-01T00:00Z</kick-off-time>
</controls>
<coordinator name='coord-1'>
    <app-path>${nameNode}/user/${userName}/${oozieRoot}/aggregator/coordinator.xml</app-path>
</coordinator>
<coordinator name='coord-2'>
    <app-path>${nameNode}/user/${userName}/${oozieRoot}/aggregator/coordinator.xml</app-path>
    <configuration>
        <property>
            <name>start</name>
            <value>${start}</value>
        </property>
        <property>
            <name>end</name>
            <value>${end}</value>
        </property>
    </configuration>
</coordinator>
</bundle-app>
```

The bundle specifies parameters, controls and coordinators

BUNDLES

The app-path node points to the coordinator.xml files

```
<bundle-app name='bundle-app' xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance' xmlns='uri:oozie:bundle:0.2'>

<coordinator name='coord-1'>
    <app-path>${nameNode}/user/${userName}/${oozieRoot}/aggregator/coordinator.xml</app-path>
</coordinator>
<coordinator name='coord-2'>
    <app-path>${nameNode}/user/${userName}/${oozieRoot}/aggregator/coordinator.xml</app-path>
    <configuration>
        <property>
            <name>start</name>
            <value>${start}</value>
        </property>
        <property>
            <name>end</name>
            <value>${end}</value>
        </property>
    </configuration>
</coordinator>
</bundle-app>
```

These are coordinators which belong to this bundle

BUNDLES

Additional configuration can optionally be specified for each coordinator

```
<bundle-app name='bundle-app' xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance' xmlns='uri:oozie:bundle:0.2'>


<coordinator name='coord-1'>
    <app-path>${nameNode}/user/${userName}/${oozieRoot}/aggregator/coordinator.xml</app-path>
</coordinator>
<coordinator name='coord-2'>
    <app-path>${nameNode}/user/${userName}/${oozieRoot}/aggregator/coordinator.xml</app-path>
    <configuration>
        <property>
            <name>start</name>
            <value>${start}</value>
        </property>
        <property>
            <name>end</name>
            <value>${end}</value>
        </property>
    </configuration>
</coordinator>
</bundle-app>
```

BUNDLES

Any number of coordinators can be specified

```
<bundle-app name='bundle-app' xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance' xmlns='uri:oozie:bundle:0.2'>

<coordinator name='coord-1'>
    <app-path>${nameNode}/user/${userName}/${oozieRoot}/aggregator/coordinator.xml</app-path>
</coordinator>
<coordinator name='coord-2'>
    <app-path>${nameNode}/user/${userName}/${oozieRoot}/aggregator/coordinator.xml</app-path>
    <configuration>
        <property>
            <name>start</name>
            <value>${start}</value>
        </property>
        <property>
            <name>end</name>
            <value>${end}</value>
        </property>
    </configuration>
</coordinator>
</bundle-app>
```

They will all be managed by this bundle

BUNDLES

```
<bundle-app name='bundle-app' xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance' xmlns='uri:oozie:bundle:0.2'>
<parameters>
    <property>
        <name>start</name>
        <value>${start}</value>
    </property>
    <property>
        <name>end</name>
        <value>${end}</value>
    </property>
</parameters>
<controls>
    <kick-off-time>2016-07-01T00:00Z</kick-off-time>
</controls>
<coordinator name='coord-1'>
    <app-path>${nameNode}/user/${user.name}/${oozieRoot}/aggregator/coordinator.xml</app-path>
</coordinator>
<coordinator name='coord-2'>
    <app-path>${nameNode}/user/${user.name}/${oozieRoot}/aggregator/coordinator.xml</app-path>
    <configuration>
        <property>
            <name>start</name>
            <value>${start}</value>
        </property>
        <property>
            <name>end</name>
            <value>${end}</value>
        </property>
    </configuration>
</coordinator>
</bundle-app>
```

The bundle can optionally specify bundle level parameters and default values here

BUNDLES

```
<bundle-app name='bundle-app' xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance' xmlns='uri:oozie:bundle:0.2'>
<parameters>
    <property>
        <name>start</name>
        <value>${start}</value>
    </property>
    <property>
        <name>end</name>
        <value>${end}</value>
    </property>
</parameters>
<controls>
    <kick-off-time>2016-07-01T00:00Z</kick-off-time>
</controls>
<coordinator name='coord-1'>
    <app-path>${nameNode}/user/${userName}/${oozieRoot}/aggregator/coordinator.xml</app-path>
</coordinator>
<coordinator name='coord-2'>
    <app-path>${nameNode}/user/${userName}/${oozieRoot}/aggregator/coordinator.xml</app-path>
    <configuration>
        <property>
            <name>start</name>
            <value>${start}</value>
        </property>
        <property>
            <name>end</name>
            <value>${end}</value>
        </property>
    </configuration>
</coordinator>
</bundle-app>
```

BUNDLES

```
<bundle-app name='bundle-app' xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance' xmlns='uri:oozie:bundle:0.2'>

<controls>
    <kick-off-time>2016-07-01T00:00Z</kick-off-time>
</controls>
</bundle-app>
```

This is interesting and requires a quick discussion of what happens when you run a Bundle on Oozie

BUNDLES

```
oozie job -oozie http://localhost:11000/oozie -
config /Users/jananiravi/Desktop/iMovieLibrary/Oozie/
Bundles/bundle/job.properties -run
```

This is the command used to run a bundle

BUNDLES

```
oozie job -oozie http://localhost:11000/oozie -
config /Users/jananiravi/Desktop/iMovieLibrary/Oozie/
Bundles/bundle/job.properties -run
```

Running a bundle passes in all the coordinator actions to Oozie

Each coordinator then starts off at the specified time in the coordinator.xml file

BUNDLES

```
oozie job -oozie http://localhost:11000/oozie -
config /Users/jananiravi/Desktop/iMovieLibrary/Oozie/
Bundles/bundle/job.properties -run
```

The coordinator actions are passed to Oozie right away

Using -run ignores the kick-off time property

BUNDLES

```
oozie job -oozie http://localhost:11000/oozie -
config /Users/jananiravi/Desktop/iMovieLibrary/Oozie/
Bundles/bundle/job.properties -submit
```

You can also submit an application to Oozie using the -submit option

BUNDLES

```
oozie job -oozie http://localhost:11000/oozie -
config /Users/jananiravi/Desktop/iMovieLibrary/Oozie/
Bundles/bundle/job.properties -submit
```

If no kick-off time is specified then the `-submit` and `-run` options on the command line both do the same thing

BUNDLES

```
oozie job -oozie http://localhost:11000/oozie -
config /Users/jananiravi/Desktop/iMovieLibrary/Oozie/
Bundles/bundle/job.properties -submit
```

no kick-off time

The coordinator actions are sent to Oozie and started at their specified times

Exactly like the -run operation

BUNDLES

```
oozie job -oozie http://localhost:11000/oozie -
config /Users/jananiravi/Desktop/iMovieLibrary/Oozie/
Bundles/bundle/job.properties -submit
```

If a kick-off time is specified then the coordination actions are not passed to Oozie till the kick-off time is reached

BUNDLES

```
oozie job -oozie http://localhost:11000/oozie -
config /Users/jananiravi/Desktop/iMovieLibrary/Oozie/
Bundles/bundle/job.properties -submit
```

Coordinator actions are given to Oozie only at the kick-off time

BUNDLES

```
oozie job -oozie http://localhost:11000/oozie -
config /Users/jananiravi/Desktop/iMovieLibrary/Oozie/
Bundles/bundle/job.properties -submit
```

After the kick-off time is reached the coordinators run based on their specified time and frequency

BUNDLES

```
oozie job -oozie http://localhost:11000/oozie -
config /Users/jananiravi/Desktop/iMovieLibrary/Oozie/
Bundles/bundle/job.properties -submit
```

It is like submitting a job using the -run flag at the kick-off time

BUNDLES

```
<bundle-app name='bundle-app' xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance' xmlns='uri:oozie:bundle:0.2'>
  <parameters>
    <property>
      <name>start</name>
      <value>11:11Z</value>
    </property>
    <property>
      <name>end</name>
      <value>12:12Z</value>
    </property>
  </parameters>

  <controls>
      <kick-off-time>2016-07-01T00:00Z</kick-off-time>
  </controls>
  <coordinator name='coord-1'>
      <app-path>${nameNode}/user/${userName}/${oozieRoot}/aggregator/coordinator.xml</app-path>
  </coordinator>
  <coordinator name='coord-2'>
      <app-path>${nameNode}/user/${userName}/${oozieRoot}/aggregator/coordinator.xml</app-path>
      <configuration>
          <property>
              <name>start</name>
              <value>${start}</value>
          </property>
          <property>
              <name>end</name>
              <value>${end}</value>
          </property>
      </configuration>
  </coordinator>
</bundle-app>
```

BUNDLES

Let's see this as a demo:

```
<bundle-app name='bundle-app' xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance' xmlns='uri:oozie:bundle:0.2'>
  <parameters>
    <property>
      <name>start</name>
      <value>2016-07-01T00:00Z</value>
    </property>
    <property>
      <name>end</name>
      <value>2016-07-03T00:00Z</value>
    </property>
  </parameters>

  <controls>
      <kick-off-time>2016-07-01T00:00Z</kick-off-time>
  </controls>
  <coordinator name='coord-1'>
      <app-path>${nameNode}/user/${userName}/${oozieRoot}/aggregator/coordinator.xml</app-path>
  </coordinator>
  <coordinator name='coord-2'>
      <app-path>${nameNode}/user/${userName}/${oozieRoot}/aggregator/coordinator.xml</app-path>
      <configuration>
          <property>
              <name>start</name>
              <value>${start}</value>
          </property>
          <property>
              <name>end</name>
              <value>${end}</value>
          </property>
      </configuration>
  </coordinator>
</bundle-app>
```

Specify a kick-off time sometime in the future

BUNDLES

Insert video 6

BUNDLES

```
<bundle-app name='bundle-app' xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance' xmlns='uri:oozie:bundle:0.2'>
<parameters>
    <property>
        <name>start</name>
        <value>${start}</value>
    </property>
    <property>
        <name>end</name>
        <value>${end}</value>
    </property>
</parameters>
<controls>
    <kick-off-time>2016-07-01T00:00Z</kick-off-time>
</controls>
<coordinator name='coord-1'>
    <app-path>${nameNode}/user/${userName}/${oozieRoot}/aggregator/coordinator.xml</app-path>
</coordinator>
<coordinator name='coord-2'>
    <app-path>${nameNode}/user/${userName}/${oozieRoot}/aggregator/coordinator.xml</app-path>
    <configuration>
        <property>
            <name>start</name>
            <value>${start}</value>
        </property>
        <property>
            <name>end</name>
            <value>${end}</value>
        </property>
    </configuration>
</coordinator>
</bundle-app>
```

BUNDLES

Bundles are a fairly advanced use case in Oozie

But as your data processing gets more complex, they can be invaluable!