Container workflows for data science and machine learning

William Benton • @willb • willb@redhat.com
Container workflows for data science and machine learning

William Benton • @willb • willb@redhat.com
About you
What do data scientists do?
\( f(\_\_\_\ _) = \begin{bmatrix} 0.67 & 0.57 & 0.84 & \cdots & 0.08 & 0.42 & 0.01 \end{bmatrix} \)
codifying problem and metrics → data collection and cleaning
data collection and cleaning → feature engineering → model training and tuning
data collection and cleaning → feature engineering → model training and tuning
feature engineering → model training and tuning → model validation
model validation → model deployment → monitoring and validation
feature engineering
model training and tuning
model validation
model deployment
monitoring and validation

codifying problem and metrics
data collection and cleaning
feature engineering
model training and tuning
model validation
model deployment
monitoring and validation
A conventional hash table (or hash table-backed set structure) consists of a series of buckets. Hash table insert looks like this:

1. First, use the hash value of the key to identify the index of the bucket that should contain it.
2. If the bucket is empty, update the bucket to contain the key and value (with a trivial value in the case of a hashed set).
3. If the bucket is not empty and the key stored in it is not the one you've hashed, handle this hash collision. There are several strategies to handle hash collisions precisely; most involve extra lookups (e.g., having a second hash function or going to the next available bucket) or
A conventional hash table (or hash table-backed set structure) consists of a series of buckets. Hash table insert looks like this:

1. First, use the hash value of the key to identify the index of the bucket that should contain it.
2. If the bucket is empty, update the bucket to contain the key and value (with a trivial value in the case of a hashed set).
3. If the bucket is not empty and the key stored in it is not the one you've hashed, handle this hash collision. There are several strategies to handle hash collisions precisely; most involve extra lookups (e.g., having a second hash function or going to the next available bucket) or
What’s a container?
%pip install numpy
<table>
<thead>
<tr>
<th>executable</th>
<th>/usr/bin/pip</th>
</tr>
</thead>
<tbody>
<tr>
<td>arguments</td>
<td>pip install numpy</td>
</tr>
<tr>
<td>environment</td>
<td>LANG=en_US USER=willb ...</td>
</tr>
<tr>
<td>virtual memory</td>
<td></td>
</tr>
<tr>
<td>file handles</td>
<td></td>
</tr>
<tr>
<td>root filesystem</td>
<td>/</td>
</tr>
<tr>
<td>process table</td>
<td></td>
</tr>
<tr>
<td>network routes</td>
<td></td>
</tr>
<tr>
<td>executable</td>
<td><code>/usr/bin/pip</code></td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>arguments</td>
<td><code>pip install numpy</code></td>
</tr>
<tr>
<td>environment</td>
<td><code>LANG=en_US USER=willb ...</code></td>
</tr>
<tr>
<td>virtual memory</td>
<td></td>
</tr>
<tr>
<td>file handles</td>
<td></td>
</tr>
<tr>
<td>root filesystem</td>
<td><code>/</code></td>
</tr>
<tr>
<td>process table</td>
<td></td>
</tr>
<tr>
<td>network routes</td>
<td></td>
</tr>
</tbody>
</table>
Software Failure. Press left mouse button to continue.
Guru Meditation #00000004.0000AAC0
Software Failure. Press left mouse button to continue.
Guru Meditation #00000008.0000A8C8
<table>
<thead>
<tr>
<th>Executable</th>
<th><code>/usr/bin/pip</code></th>
</tr>
</thead>
<tbody>
<tr>
<td>Arguments</td>
<td><code>pip install numpy</code></td>
</tr>
<tr>
<td>Environment</td>
<td><code>LANG=en_US USER=willb ...</code></td>
</tr>
<tr>
<td>Virtual Memory</td>
<td></td>
</tr>
<tr>
<td>File Handles</td>
<td></td>
</tr>
<tr>
<td>Root Filesystem</td>
<td><code>/var/lib/envs/main</code></td>
</tr>
<tr>
<td>Process Table</td>
<td></td>
</tr>
<tr>
<td>Network Routes</td>
<td></td>
</tr>
<tr>
<td>category</td>
<td>value</td>
</tr>
<tr>
<td>---------------------</td>
<td>--------------------------------------------</td>
</tr>
<tr>
<td>executable</td>
<td>/usr/bin/pip</td>
</tr>
<tr>
<td>arguments</td>
<td>pip install numpy</td>
</tr>
<tr>
<td>environment</td>
<td>LANG=en_US USER=willb ...</td>
</tr>
<tr>
<td>virtual memory</td>
<td></td>
</tr>
<tr>
<td>file handles</td>
<td></td>
</tr>
<tr>
<td>root filesystem</td>
<td>/var/lib/envs/main</td>
</tr>
<tr>
<td>process table</td>
<td></td>
</tr>
<tr>
<td>network routes</td>
<td></td>
</tr>
</tbody>
</table>
Immutable images
Immutable images

configuration and installation recipes

base image
Immutable images

user application code

configuration and installation recipes

base image
Immutable images

- User application code
  - a6af9d1e
  - 6b8cad3e
- Configuration and installation recipes
  - 33721112
  - e8ca4e6f
  - 2bb6ab16
  - a8296f7e
- Base image
  - 979229b9
Stateless microservices
Stateless microservices
Stateless microservices
Stateless microservices
Declarative app configuration
Declarative app configuration

https://devconf19.willb.io
Integration and deployment
Integration and deployment
Integration and deployment

OK!

- base image
- configuration and installation recipes
- application code
Integration and deployment

OK!

base image

configuration and installation recipes

application code

base image
Integration and deployment
What containers offer data scientists
FROM centos:centos7
RUN yum install -y \  
    python python-pip \  
    java java-devel git
ENTRYPOINT /bin/bash
FROM centos:centos7
RUN yum install -y \
    python python-pip \
    java java-devel git
ENTRYPOINT /bin/bash
No friction: mybinder.org
More flexible: source-to-image
More flexible: source-to-image

https://github.com/openshift/source-to-image
willb@echo % oc new-app --name model \
quay.io/willbentong/simple-model-s2i:demo\~https://github.com/willb/example-model-s2i-notebook
willb@echo % oc new-app --name model \
quay.io/willbenton/simple-model-s2i:demo\n~https://github.com/willb/example-model-s2i-notebook
model validation

model deployment

monitoring and validation
data scientists

application developers

data engineers

machine learning engineers

data scientists

- events
- databases
- file, object storage

- transform
- transform
- transform

- models

- train
- federate
- developer UI

web and mobile

archive

reporting

management
radanalytics.io
Open Data Hub
What did we talk about today?