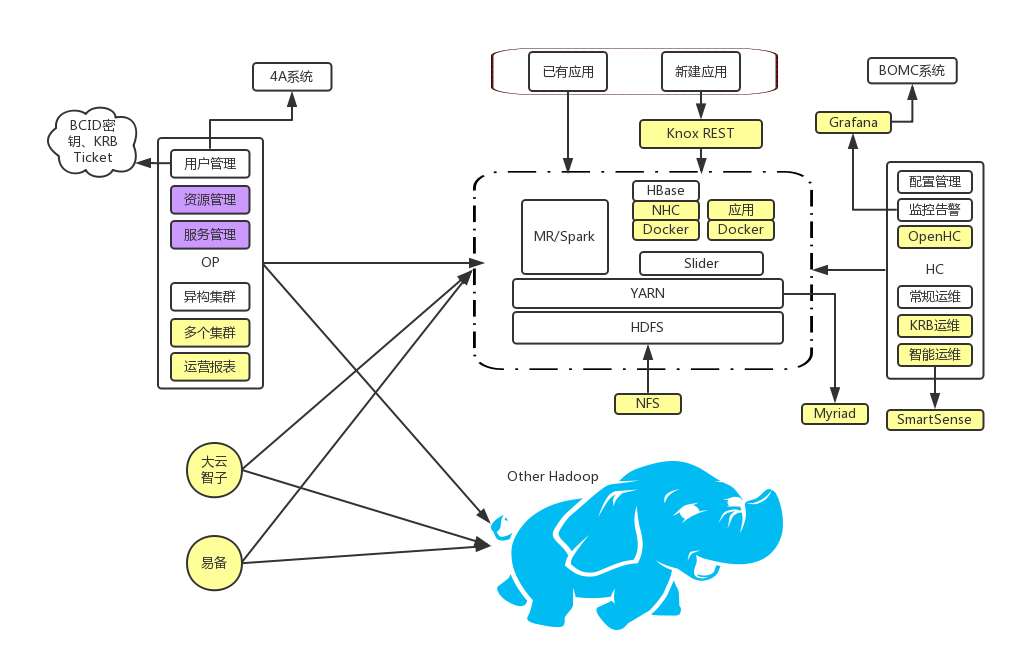
Hcontrol on yarn

Background

当前通过Slider启动的应用如HBase,Storm等，缺少监控、滚动升级、配置管理等基本功能模块，从而App on Yarn的可用性不高。Hcontrol有完整的应用管理机制，通过在Yarn Containe中运行Hcontrol，Hcontrol来接管App的部署与管理，是一个可行的方案。总体框架如下图所示：



总体思路是在Docker中启动Hcontrol模块，然后Hcontrol来部署应用。Hcontrol无法接管的应用，直接部署的Docker中或者部署的集群中。

集群外的客户端访问新建的应用，通过Knox Rest来实现认证与代理。

Requirements

实现Hcontrol on Yarn，需要解决下面的问题：

1）Slider 能够实现对Hcontrol Component Docker的编排、部署、监控与管理

2）Slider能够在Hcontrol on yarn启动后，一键部署所有应用

3）Slider具有多种调度策略，包括Strict,Label,Host-Constraints等

4）启动的Hcontrol 组件的Docker Container保证一直处于可用状态，要保证NM挂掉以后，可以进行无数据丢失的迁移

5）基本原则是最大程度的有状态的应用中启动进程与数据分离，做到App Container run anywhere

KeyPoint

1）Slider 编排Hcontrol Cluster，根据需要启动所需要的Hcontrol Server与Agent，包括个数，启动主机及网络，涉及到调度，Container网络配置

2）在Hcontrol中一键式部署所需要的应用，可以通过Ambari Shell根据Blueprint创建应用，在Slider中选择启动的App，并生成Blueprint

3）实现Hcontrol Docker Container的迁移，由于Container是有状态的，将该Container迁移至其他节点，解决的难点是实现数据的迁移

4）应用进程相关数据与应用本身数据的分离（这点有待进一步的去研究），任意的应用都包括两个方面：

进程运行所需要的数据

a)包括配置，安装文件等

b)组件之间交互所需要的数据，例如Storm 、Kafka中的zk path中保存的主机及 端口信息等

进程所要处理的数据，例如HBase运行所处理的表数据（包括元数据）

Design

下图是Slider当前的运行架构图：



当前启动应用有两种方式：

1）bin/bash，直接在物理机上启动服务，目前正在使用的Hbase,Storm,Jstorm均已这种方式来启动

2）Docker Engine，目前还没有进行使用，对于Hcontrol on yarn，在Docker中启动Hcontrol Master与Agent进程

为了实现Hcontrol on Yarn，系统架构图如下所示：



系统分成三个层次：

1）Slider Cluster

Slider与YARN交互，启动Slider Cluster，其中包括Slider AppMaster与Slider Agent，Slider AppMaster的作用为启动与管理Slider Agent

2）Docker Daemon与Docker

Slider Agent与Docker Daemon交互，启动Docker Container，其中Container中运行HC-Master和HC-Agent

3）HContorl Cluster

包括HControl Master和Agent，启动和管理应用

4）其他

Hadoop生态系统中的服务，数据一般存放在HDFS上，部分元数据与运行数据放到ZK中

**1.Slider Cluster的编排**

Slider Docker App的创建，定义3个文件，以memcached为例：

1）metainfo.json //定义app的组成部分及image

{

"schemaVersion":"1.0",

"application":{

"name":"MEMCACHED",

"components":[

{

"name":"MEMCACHED",

"type":"docker",

"dockerContainers":[

*{*

*"name":"memcached",*

*"commandPath:":"/usr/bin/docker",*

*"image":"memcached"*

*}*

]

}

]

}

}

2）appConfig.json

app所需要的配置：

{

"schema":"http://example.org/specification/v2.0.0",

"metadata":{},

"global":{},

"components":{

"MEMCACHED":{

"memcached.commandPath":"/usr/bin/docker"

}

}

}

3）resources.json

docker app的组成部分

{

"schema":"http://example.org/specification/v2.0.0",

"metadata":{},

"global":{},

"components":{

"slider-appmaster":{},

"MEMCACHED":{

"yarn.role.priority":"1",

"yarn.component.instances":"1",

"yarn.memory":"512"

}

}

}

**2.Hcontrol on Docker**

1) Ambari in Docker初探

2) Ambari in Docker

Development in Docker

https://cwiki.apache.org/confluence/display/AMBARI/Development+in+Docker

3) Ambari-cluster-on-Docker

http://blog.sequenceiq.com/blog/2014/06/17/ambari-cluster-on-docker/

详情见文档，可参考Ambari-cluster-on-Docker，但是该项目仅是在一个主机上启动Ambari cluster，我们要解决的问题解决Ambari Cluster分布在多个节点的网络问题（可参考宇轩原来做的Hadoop On Docker的网络处理方案）

**3. Hcontrol App与外部HDFS及ZK的交互**

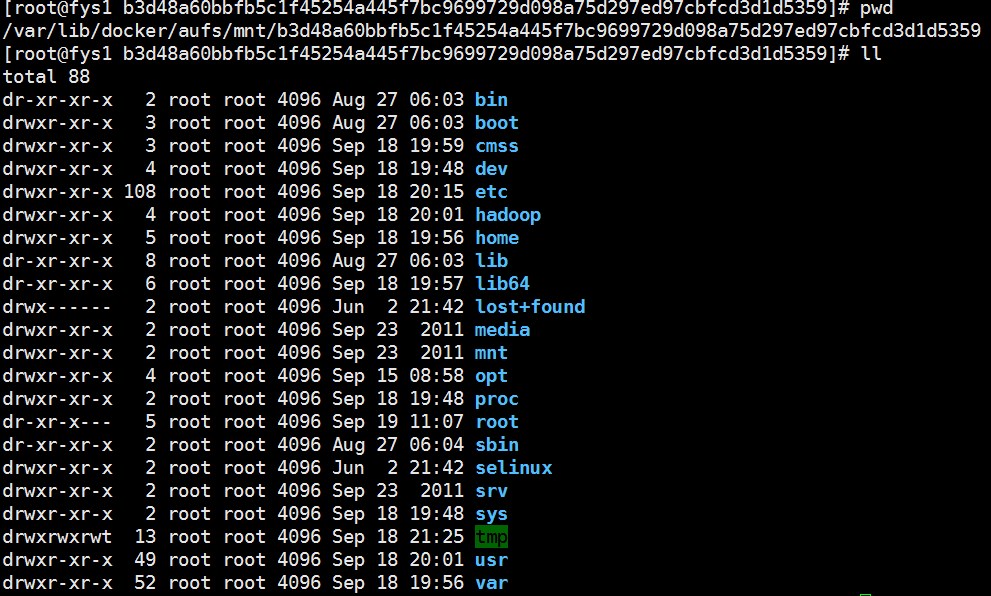
流程图如下所示：



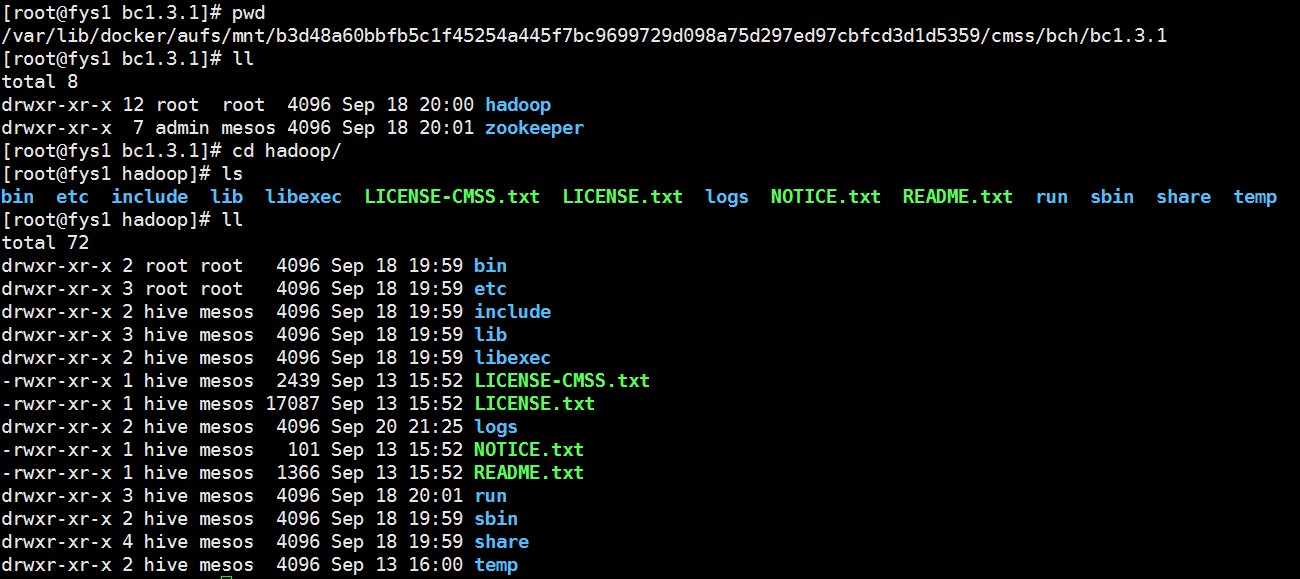
**4.Hcontrol Docker 迁移**

当出现NM或者Docker Daemon宕机的情况，Slider无法重启在该节点启动的Docker container，从而导致无法启动对应的服务，当出现这个情况时，应该进行Docker Container的迁移，涉及到的数据：

1）Docker Container的执行目录（可能是操作系统的镜像），例如：



cmss的hadoop的安装目录，如下：



该目录是存在本地的/var/lib/docker目录

2）Container的Image

在Container的运行过程中，会根据base Image生成一个运行时image，在迁移时这个目录也要迁移

[root@fys1 layers]# ll

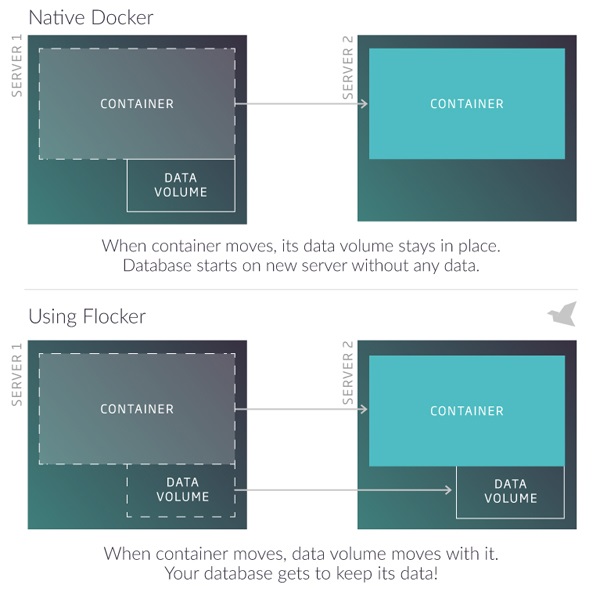
b3d48a60bbfb5c1f45254a445f7bc9699729d098a75d297ed97cbfcd3d1d5359

b3d48a60bbfb5c1f45254a445f7bc9699729d098a75d297ed97cbfcd3d1d5359-init

还要其他的参数需要考虑，手动迁移几乎不可能

解决方案：采用Flocker Plugin，在Slider Cluster运行过程中如何调用，需要进一步调研

Flocker是一个数据卷管理器和多个主机Docker集群管理工具，使用Flocker，可以控制数据使用与无状态应用程序相同的工具：

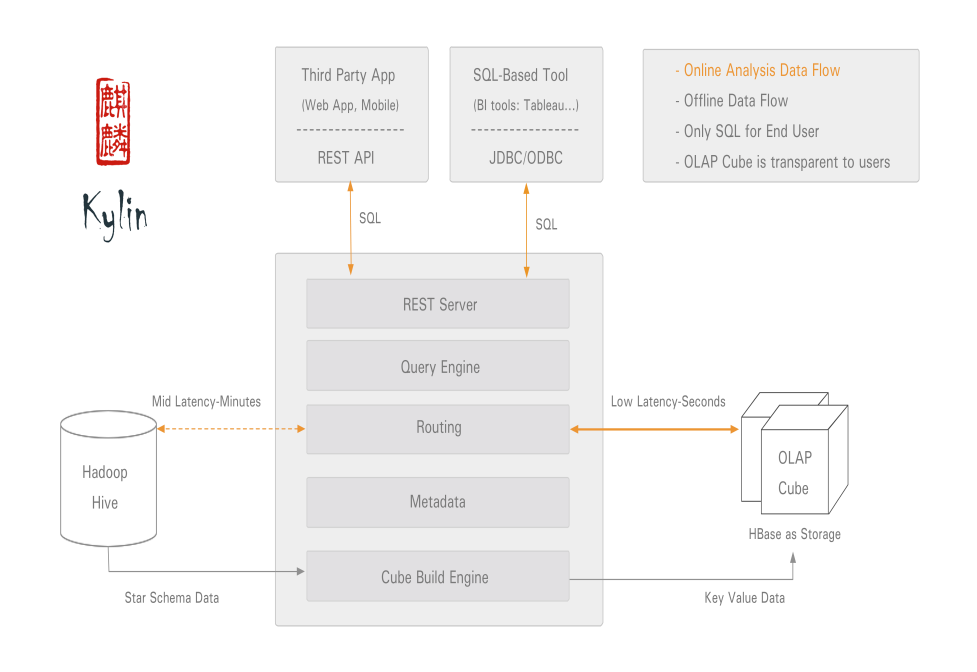


在3中将App的数据写到Docker之外的存储的一个考虑，就是最小化Docker Container迁移的代价。

**5.其他**

Case

以Kylin为例，系统架构图如下：



Kylin完整的运行系统涉及到：

1）Kylin Server

2）HBase

3）Hive

部署框架如下所示：

