## 大数据容器化可行性分析

### 技术说明

#### elasticsearch

##### 部署

官网提供了Helm部署的方式，要求Helm版本在2.8以上，kubernetes版本在1.8以上，提供了elasticsearch的配置项说明及使用，参照下面的链接

<https://github.com/elastic/helm-charts/blob/7.2.0/elasticsearch/README.md>

<https://github.com/bitnami/charts/tree/master/bitnami/elasticsearch>

参考资料：

[https://sematext.com/blog/kubernetes-elasticsearch/#toc-how-to-deploy-elasticsearch-on-kubernetes-6](https://sematext.com/blog/kubernetes-elasticsearch/" \l "toc-how-to-deploy-elasticsearch-on-kubernetes-6)

亦可以通过自定义资源模板的方式部署：

<https://rancher.com/blog/2018/2018-11-22-deploying-elasticsearch/>

以及operator的方式

<https://github.com/zalando-incubator/es-operator>

<https://akomljen.com/kubernetes-elasticsearch-operator/>

##### 监控

使用 Metricbeat Prometheus 模块收集指标

[Metricbeat](https://www.elastic.co/cn/beats/metricbeat) 是 Elastic 的轻量型指标采集器。 Metricbeat 随附的 [Prometheus 模块](https://www.elastic.co/guide/en/beats/metricbeat/7.0/metricbeat-module-prometheus.html)可以采用三种方式 收集指标：

通过端口 9090 连接到 Prometheus 服务器， 并使用 Prometheus 联合 API 拉取已收集的指标 （以获得 Prometheus 收集的指标）

使用 /metrics 端点 通过端口 9090 连接到 Prometheus 服务器（Prometheus 自我监测）

连接到各个 Prometheus 导出器并解析 公开格式

##### 集群安全

elasticsearch版本7之后免费提供了一些安全功能，包括集群安全

<https://www.elastic.co/cn/blog/getting-started-with-elasticsearch-security>

<https://www.elastic.co/guide/en/elasticsearch/reference/7.x/configuring-security.html>

#### Zookeeper

目前在k8s官网中已经提供了对zk的安装部署文档支持，地址如下所示:

<https://kubernetes.io/docs/tutorials/stateful-application/zookeeper/>

博客资料示例

<https://blog.csdn.net/weixin_34032792/article/details/91949785>

而安装部署使用的是headless的方式部署，相对于一般的工具部署在配置yaml上复杂一点，但总体方案可行.

#### Kafka

目前暂时没有在官网找到kafka的k8s支持文档，但是在github中找到第三方的部署资料。

<https://github.com/Yolean/kubernetes-kafka>

但是关于该方案的部署还有待考虑主要是配置参数的设置和具体情况了解等

业界部署kafka方案

<https://lxkaka.wang/2019/10/25/k8s-kafka/>

helm部署方式

<https://github.com/helm/charts/tree/master/incubator/kafka>

#### GreenPlum

###### 安装部署

Greenplum官网提供了Greenplum for kubernetes的helm版本的包，通过这个和参考部署手册，基本可以解决部署方面的问题.

以下内容是详细的参考地址.

[https://netw](https://network.pivotal.io/products/greenplum-for-kubernetes)[ork.pivotal.io/products/greenplum-for-kubernetes](https://network.pivotal.io/products/greenplum-for-kubernetes)

官网下载Greenplum for kubernetes的包

<https://greenplum-kubernetes.docs.pivotal.io/1-12/gp4k1120ReleaseNotes.html>

这是目前所支持的GP版本的参考资料.

<https://greenplum-kubernetes.docs.pivotal.io/1-11/installing.html>

这是安装gp for k8s的operator地址.

<https://greenplum-kubernetes.docs.pivotal.io/1-11/deploy-operator.html>

这是gp for k8s的安装集群的参考地址

<https://greenplum-kubernetes.docs.pivotal.io/1-11/operator-reference.html>

这是安装的时候使用的yaml的参数的参考地址.

备注:目前官网给出的最高的Greenplum for kubernetes的包是1.11版本的，对应GP版本是5.24，而docker hub上有第三方打包好的对应GP版本为6.1的，但是目前建议使用官网提供的版本.

#### flink

##### 部署

目前官网有提供相应的支持，具体链接如下所示：

<https://ci.apache.org/projects/flink/flink-docs-release-1.10/zh/ops/deployment/kubernetes.html>

#### MongoDB

目前官网有提供相应的支持，具体链接如下所示:

<http://k8smongodb.net/>

官网给出的关于在k8s上部署mongoDB的链接.

<https://www.jianshu.com/p/dbd7fc2d687b>

业界部署方案

<https://github.com/helm/charts/tree/master/stable/mongodb>

helm部署方案.

#### Hadoop

Hadoop是大数据核心组件之一，目前该组件的部署方案已经较为成熟，因此相对较为容易.

<https://blog.csdn.net/z294155673/article/details/78754128>

业界部署方案参考.

<https://github.com/helm/charts/tree/master/stable/hadoop>

Helm官网在github上关于部署hadoop的参考资料.

### 技术难点

#### Kafka

目前kafka没有明确地官方支持文档等信息，但业界已经有较多部署方案，因此部署目前难度可控

#### GreenPlum

目前官方已经提供了helm版本的包，但是可供选择的GP版本最高为5.24，最新的6.x版本暂时未提供，该版本官网并未提供相应的包，最低版本为5.10， 因此如果后续需要升级版本，建议可以暂时搁置该版本的方案，先熟悉6.x版本新特性，等到后续需要的时候再升级.或者考虑使用第三方docker hub上打包好的版本的包，或者考虑自行打包，但是打包难度较大，并且需要熟悉新特性等。

### 性能指标分析

需待真正部署完成后进一步测试。

### 可靠性分析

在Kubernetes的解决方案中，一个服务可以自我扩展、自我诊断，并且容易升级，在收到服务扩容的请求后，Kubernetes会触发调度流程，最终在选定的目标节点上启动相应数量的服务实例副本，这些服务实例副本在启动成功后会自动加入负载均衡器中并生效，整个过程无须额外的人工操作。另外，Kubernetes会定时巡查每个服务的所有实例的可用性，确保服务实例的数量始终保持为预期的数量，当它发现某个实例不可用时，会自动重启该实例或者在其他节点上重新调度、运行一个新实例，这样，一个复杂的过程无须人工干预即可全部自动完成。

### 安全性分析

安全分析主要考虑两部分内容，一个是容器云的安全认证方面这部分内容，目前有待继续完善第二部分就是关于kerberos认证方面的支持程度。

<http://docs.kubernetes.org.cn/148.html>

这是k8s中文社区关于k8s的安全授权认证方面的资料.

<https://github.com/apache-spark-on-k8s/kubernetes-HDFS/blob/master/charts/README.md>

这是github上关于k8s上部署kberberos hdfs认证方面的资料

<https://zhuanlan.zhihu.com/p/73485888>

知乎上的相关内容

### 风险分析

目前风险主要考虑集中在几个方面:

1.开发人员对k8s技术方面的不熟悉，水平参差不齐，设计方案和改造与安全不符。

2.目前大部分大数据组件在官方都有支持，但是对于部分组件版本可能会滞后或者支持度不足，如Greenplum目前生产使用的版本在官网因为版本太旧不支持了.因此需要考虑升级新版本带来的风险。

4.大数据平台监控告警模块的内容，因为有部分功能和容器云自身的模块重合因此这部分的内容需要进行取舍。

5.关于多租户授权这块，目前需要对容器云的相关技术进行深入了解。

6.扩容，大数据平台中的扩容概念与k8s中的扩容并不是相同的概念，前者是指在一个新的机器节点加入到集群之后，资源加入到yarn中进行管理，而后者主要是指副本集的动态扩展缩减。