容器化大数据应用部署实践

肖德时 @数人云

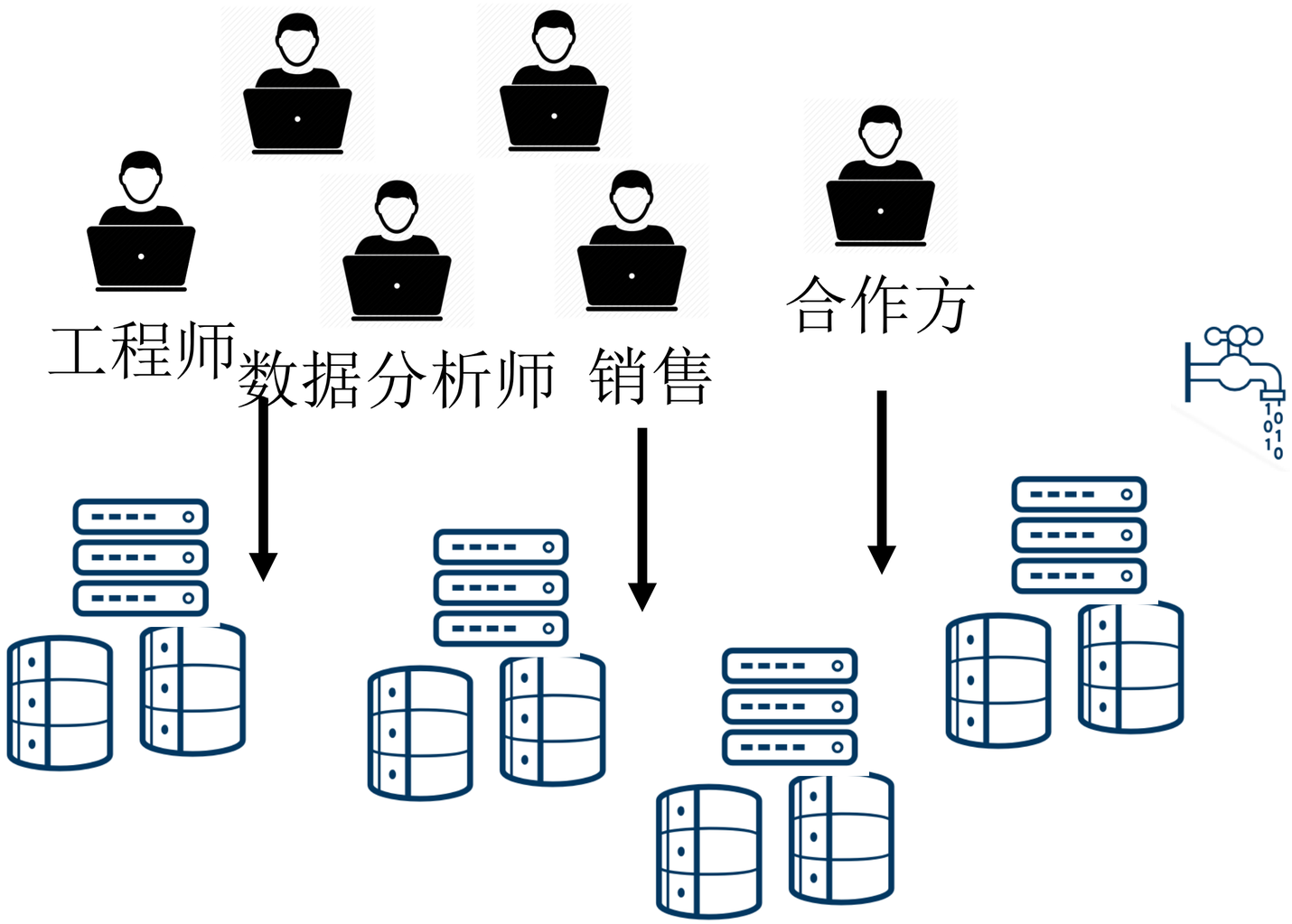
2016年 3月

数人云 shurenyun.com | 下一代 DCOS

议题

- 大数据应用部署的现状
- 容器化变革尝试
- 容器化应用部署实践 - Hadoop Cluster
- 容器化大数据的好处
- 容器化大数据的未来

大数据应用部署的现状



资源利用率低于30%

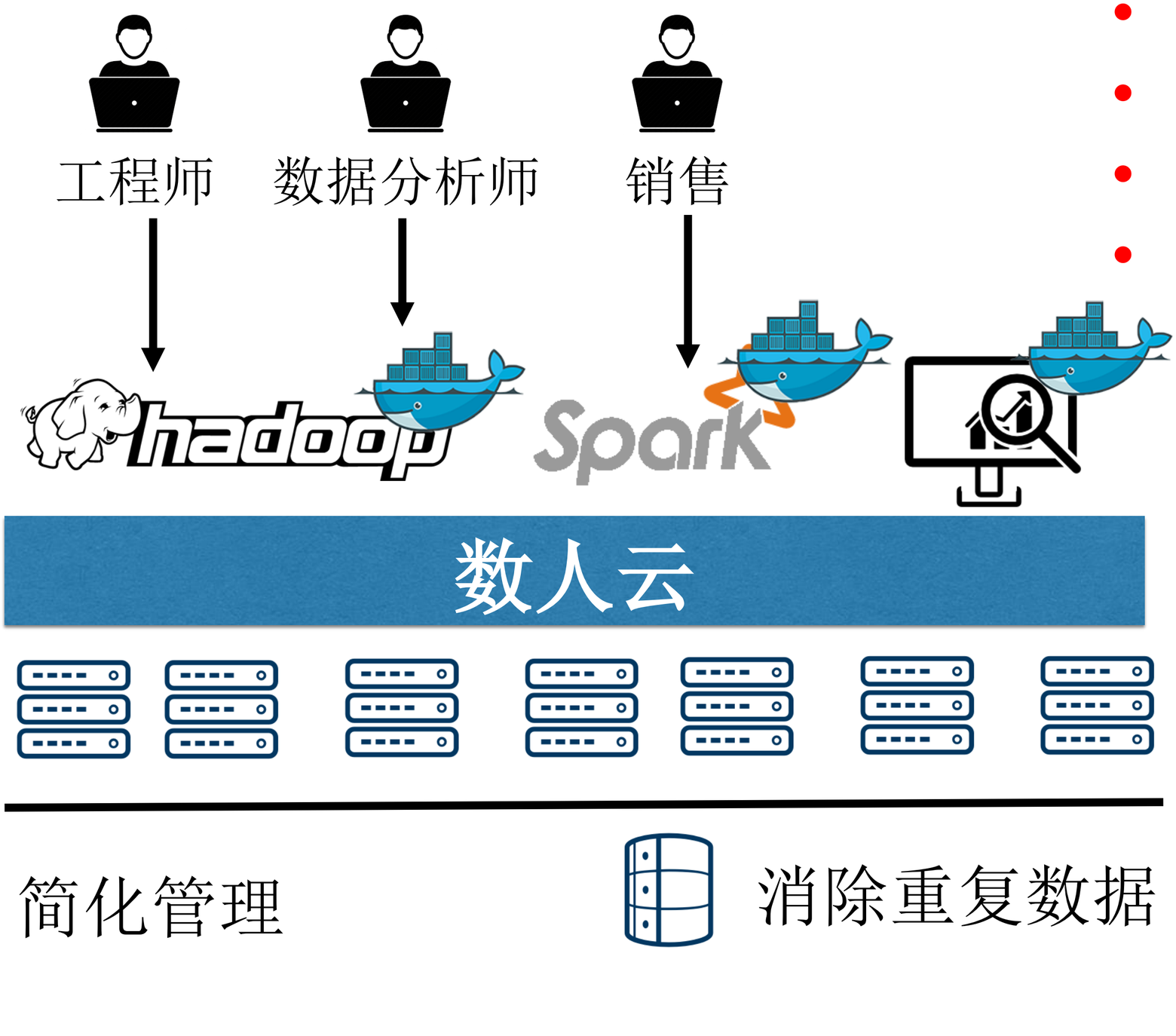
管理复杂，重复数据过多

大数据应用部署的现状

工具越来越多，安装部署都不一样。重

- Apache Kylin(麒麟) - OLAP 引擎
- Apache storm
- Spark Streaming
- Apache Hive & Hadoop
- Apache Flink: Scalable Batch and Stream Data Processing
- Presto: Distributed SQL Query Engine for Big Data

容器化变革尝试



特点：

- 统一平台
- 资源混合部署
- 基于容器分发
- 资源利用率大于90%

简化管理

消除重复数据

容器化应用部署实践 - Hadoop Cluster

- 使用 Docker 容器避免底层依赖

PUBLIC | AUTOMATED BUILD

sequenceiq/hadoop-docker ☆

Last pushed: 4 months ago

Repo Info | Tags | Dockerfile | Build Details

Short Description

An easy way to try Hadoop

Full Description

#Apache Hadoop 2.7.0 Docker image

Note: this is the master branch - for a particular Hadoop version always check the related branch

A few weeks ago we released an Apache Hadoop 2.3 Docker image - this quickly become the most popular Hadoop image in the Docker registry.

Following the success of our previous Hadoop Docker images, the feedback and feature requests we received aligned with the Hadoop release cycle, so we have released an Apache Hadoop 2.7.0 Docker image - same as the previous version, it's available as a trusted and automated build on the official Docker registry.

FYI: All the former Hadoop releases (2.3, 2.4.0, 2.4.1, 2.5.0, 2.5.1, 2.5.2, 2.6.0) are available in the GitHub branches or our Docker Registry - check the tags.

Docker Pull Command

```
docker pull sequenceiq/hadoop-docker
```

Owner

SEQUENCEIQ sequenceiq

Source Repository

sequenceiq/hadoop-docker

容器化应用部署实践 - Hadoop Cluster

- 通过 Pachyderm 减少对 MapReduce 的依赖

Pachyderm File System (pfs)

Pfs is a copy-on-write distributed file system built to deploy on containerized infrastructure.

Version controlled data

Pfs is a commit-based distributed file system that offers complete version control for your data. Pfs lets you take space-efficient snapshots of your entire cluster so you can track how your data has changed over time and instantly revert back to any previous state. This makes pfs ideally suited for large data sets that change over time — such as production database dumps or log files.

Isolated development environments.

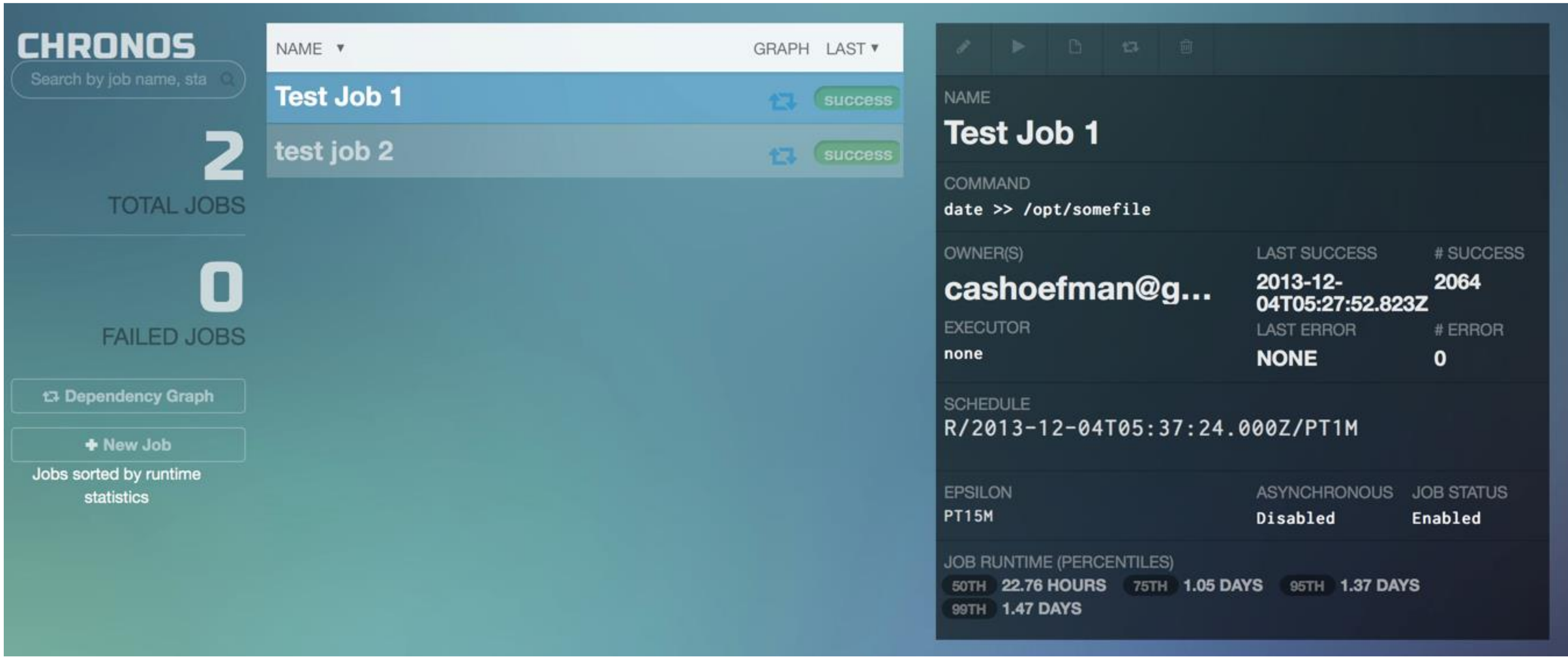
Pfs supports branching and merging, just like your VCS tools for code, but on your entire data set! Just give each developer or data scientist a personal branch and it'll feel like they've got the cluster all to themselves. They can manipulate data and develop analysis in complete isolation and then merge it into the main line when completed.

Easy dashboarding

In Pachyderm, all data is accessible via HTTP so you can serve dashboards directly out of the file system. You can even have the results of an analytics pipeline be a dashboard that automatically updates every time new data is committed.

容器化应用部署实践 - Hadoop Cluster

- 使用 Chronos 来执行调度分析任务



容器化应用部署实践 - Hadoop Cluster

- 使用 Ferry 在本地搭建大数据应用开发环境

Big Data Development Environment using Docker

Ferry helps you create big data clusters on your local machine. Define your big data stack using YAML and share your application with *Dockerfiles*. Ferry supports Hadoop, Cassandra, Spark, GlusterFS, and Open MPI.

Here's an example Hadoop cluster:

```
backend:
   - storage:
        personality: "hadoop"
        instances: 2
        layers:
           - "hive"
connectors:
   - personality: "hadoop-client"
```

容器化应用部署实践 - Hadoop Cluster

·使用轻量级 PaaS 快速搭建生产级别大数据环境

一键部署 「快速搭建企业生产环境」

可一键部署 Docker 容器化应用。（支持企业原有容器化应用，及 Docker Hub 等第三方平台的 Docker 镜像）；
可一键部署 Spark、Hadoop、 Cassandra、Jenkins、Kafka 等多种常用分布式应用，从而快速搭建企业生产环境，开发流行的微服务和大数据应用。

容器化大数据的好处

- 快速安装大数据组件
- 开发/QA/生产使用同一套镜像和流程
- 一个节点到多个节点，步骤都是一样的启动服务

容器化大数据的未来

数人云 shurenyun.com

后台业务支撑平台软件

企业级 AppStore
云操作系统

快速
发布

自动
扩容

自动
容错

安全
监控

IDC或者私有云

基础运维

openstack

vmware

UCLOUD 专业基础云计算服务商

阿里云 aliyun.com

腾讯云

HUAWEI

数人云 shurenyun.com | 下一代 DCOS

关注数人云