MapReduce Project with Inverted Index

赵敏 老师

扫描二维码关注微信/微博
获取最新IT面试情报及权威解答

微信: ninechapter
知乎专栏: http://zhuanlan.zhihu.com/jiuzhang
微博: http://www.weibo.com/ninechapter
官网: www.jiuzhang.com

Course Outline

九章算法

- What is Vagrant & How to use
- What is Docker & How to use
- Design Lucene Search Engine
- Implement search engine with Mapreduce
- Run mapreduce on Hadoop

Why do we learn docker and vagrant?

九章算法

Put it on your resume!!!

Jan Garaj – DevOps | Docker | Kubernetes | AWS | Zabbix | Monitoring

My GitHub codes | www.linkedin.com/in/jangaraj | jan.garaj@gmail.com | +44 79 234 69004
www.jangaraj.com | Infographic about me | My Google codes | +421 949 113 911

Expertise

- **Docker** – 2 years of commercial experience
- **Docker** orchestration – **Kubernetes** and **AWS ECS**
- **900+ GitHub stars** / 200+ forks of my public projects
- **Devops** continuous integration and deployment
- Docker/host/service/app **monitoring**
- **Automation** with Puppet and Ansible
- **Docker** stress and performance testing
- **Docker** image development and optimization

Own Docker monitoring ecosystem

- **Zabbix XXL** – www.github.com/monitoringartist/zabbix-xxl
  Dockerized Zabbix – server, web, proxy, java gateway, snmpd with additional extensions for easy deployment
- **Zabbix agent XXL** – www.github.com/monitoringartist/zabbix-agent-xxl
  Dockerized Zabbix agent with support for:
  - **Zabbix Docker Monitoring** – www.github.com/monitoringartist/zabbix-docker-monitoring
  Zabbix C module and template for low level Docker container monitoring, C module provides high performance with low footprint, Low level discovery concept is used for automatic Docker container detection. Prepared also for Mesos with Chronos and Marathon framework
  - **Zabbix Systemd Monitoring** – www.github.com/monitoringartist/zabbix-systemd-monitoring
  Systemd monitoring of running systemd units
- **Kubernetes Zabbix/Grafana cluster** – www.github.com/monitoringartist/kubernetes-zabbix
  Zabbix 3.0 monitoring with Grafana visualization infrastructure in Kubernetes cluster
- **Grafana XXL Docker image** – www.github.com/monitoringartist/grafana-xxl
  Docker image of Grafana with all available plugins: Zabbix, DalmatinerDB, Ambari, Atsd, Bosun, Druid, ...
- **Zabbix templates image** – www.hub.docker.com/r/monitoringartist/zabbix-templates/

Docker

- A container
- Package your application into a standardized unit
- Share hostOS

Docker

九章算法

Demo

Vagrant

It works on my computer!

Vagrant

场景：

If you have one project, using java7, and your colleague try to run your project, but he’s using java8, what will happen?

Vagrant

Solution:

Change the java version in environment every time he try to run the project.

Vagrant

场景2:

What if you have hundreds of parameters in environment that need to be changed?

Vagrant

九章算法

- Vagrant is computer software that creates and configures virtual development environments.
- It can be seen as a higher-level wrapper around virtualization software such as VirtualBox, VMware and container.

Vagrant

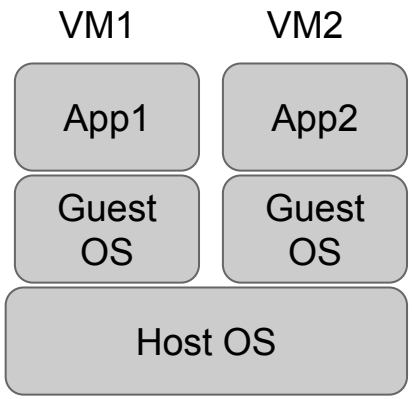
九章算法

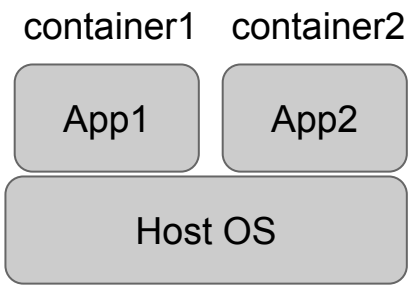
Demo

Docker vs Vagrant

九章算法

Container Vs VM





Q&A

- Why should I install hadoop on docker instead of locally

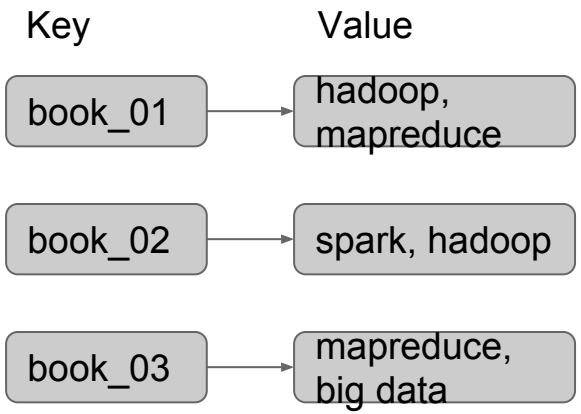
Design Lucene Search Engine

Let’s design a search engine!

Design Lucene Search Engine

Suppose you want to search book talking about data in a library.

- map<book_id, keyword_list>



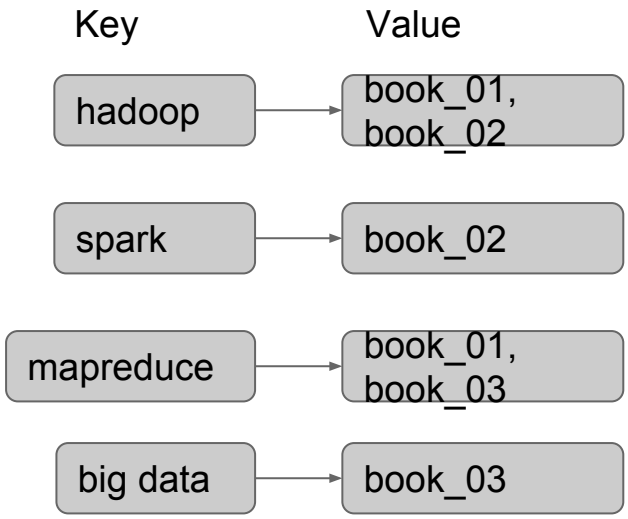
Forward Index

Design Lucene Search Engine

Too slow!

Design Lucene Search Engine

- map<keyword, book_id_list>



Inverted Index

Design Lucene Search Engine

- Fetch the documents
- Map key word to document_id

Design Lucene Search Engine

Sounds easy?

Mapper

doc1.txt:

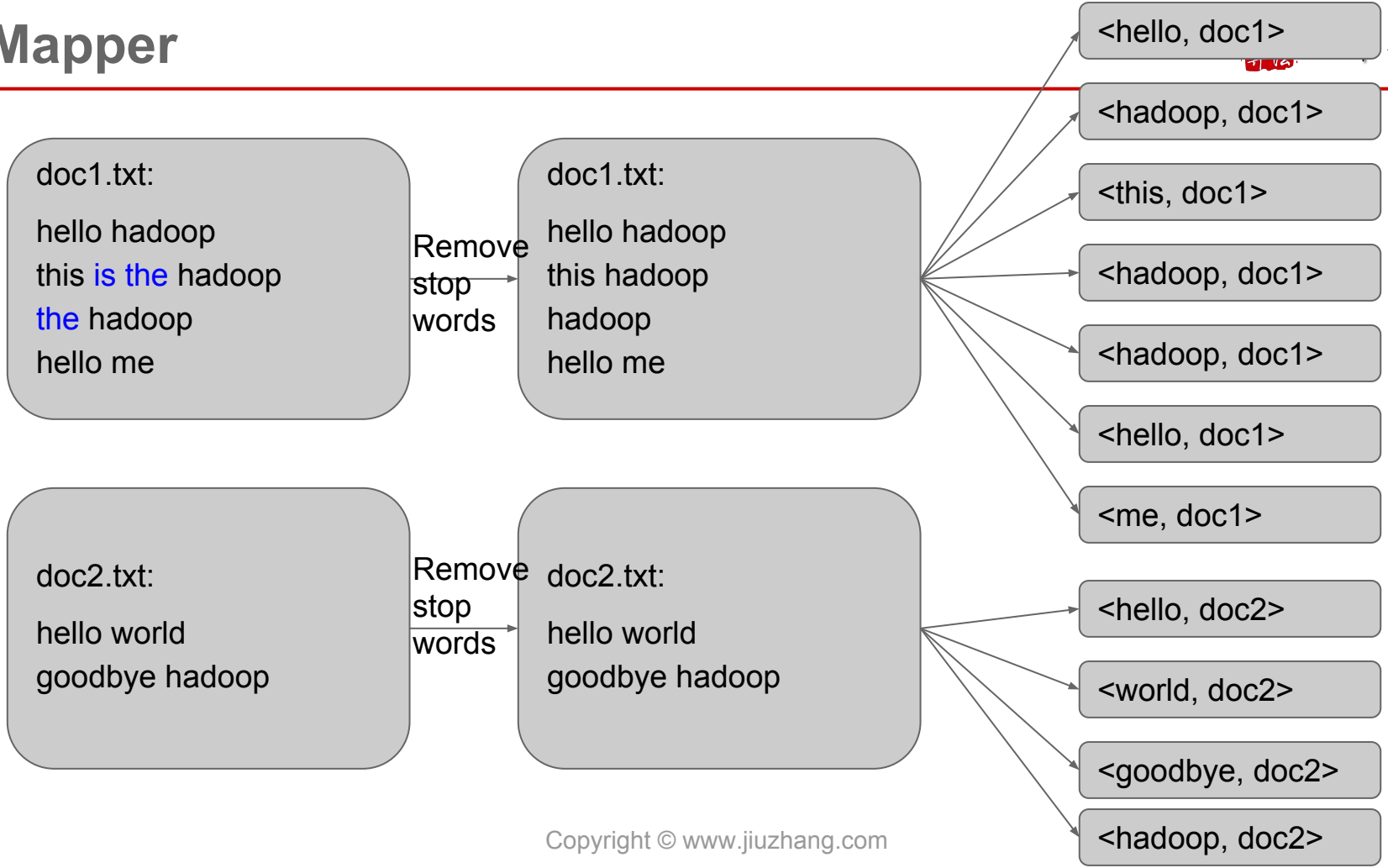
hello hadoop
this is the hadoop
the hadoop
hello me

is → 200

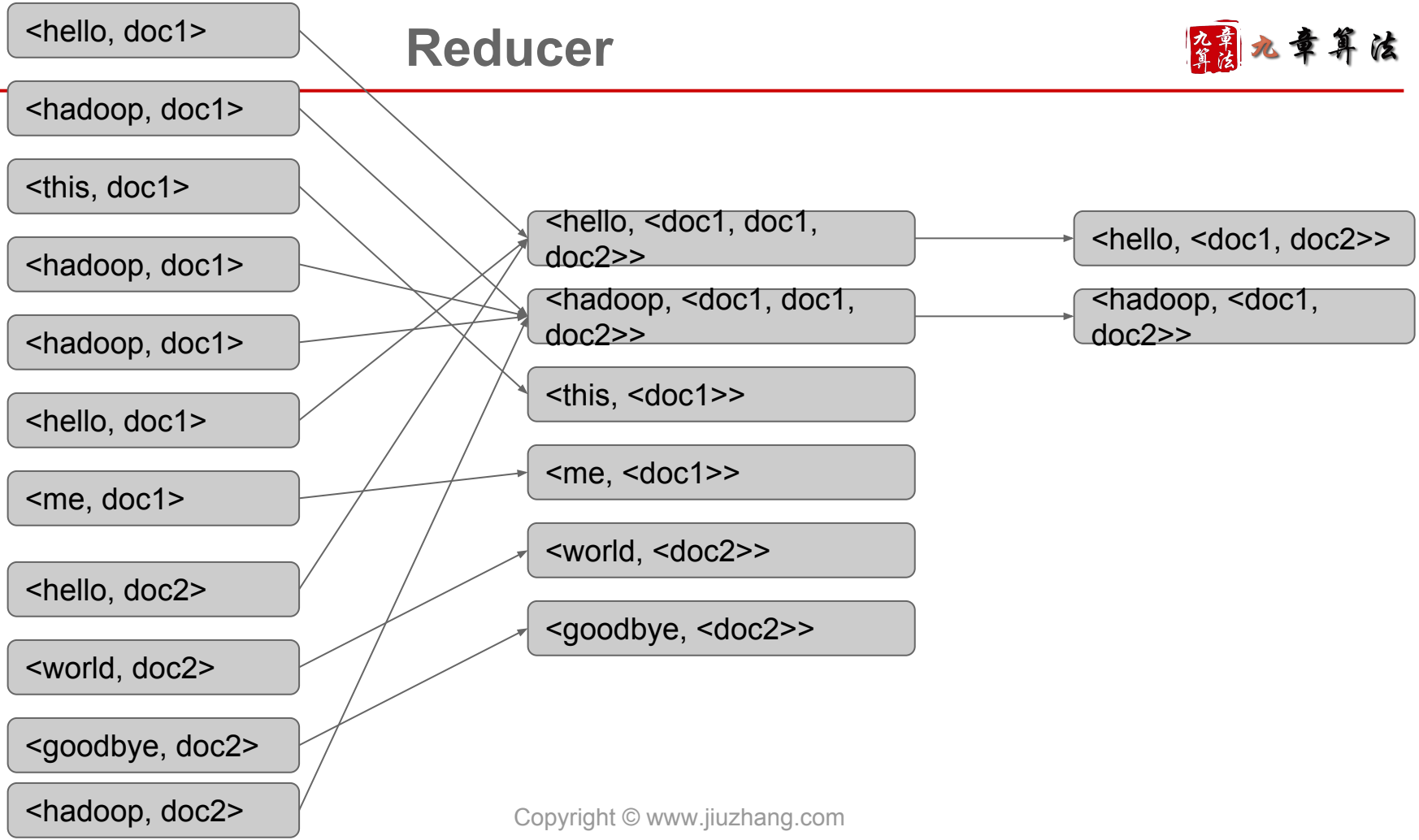
the → 400

Remove stop words

Mapper



Reducer



Implement search engine with Mapreduce

九章算法

Demo

How does the job run

九章算法

- Create directory to store compiled Java classes.
- Generate the jar file.
- Create input directory in HDFS.
- Upload input files into input directory.
- Create output directory in HDFS.
- Run the jar file.
- Check the output results.

What we have learned

九章算法

- What is Vagrant?
- What is Docker?
- The threory behind search engine
- What is inverted index?
- How to implement search engine?