

大数据开发技术

东北林业大学

卢洋

第二章

Hadoop序列化

1. 序列化概念;
2. 自定义bean对象实现序列化接口(Writable);
3. 序列化案例.

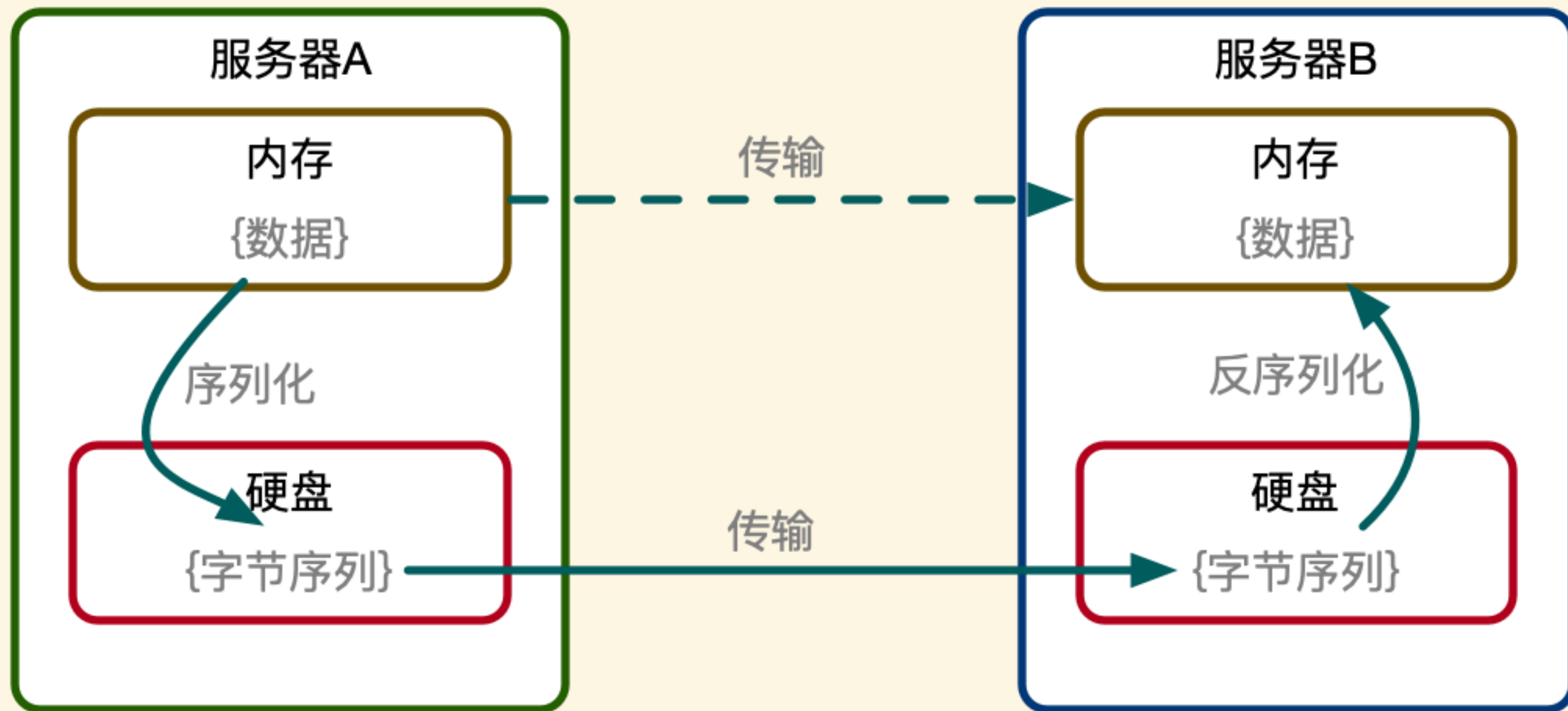
2.1

序列化概念

1 什么是序列化?

- 序列化就是将内存中的对象，转化成字节序列(或其它数据传输协议)以便于存储到磁盘(持久化)和网络传输.
- 反序列化就是将接收到的字节序列(或其它数据传输协议)或者磁盘持久化数据，转换成内存中的对象.

1 什么是序列化?



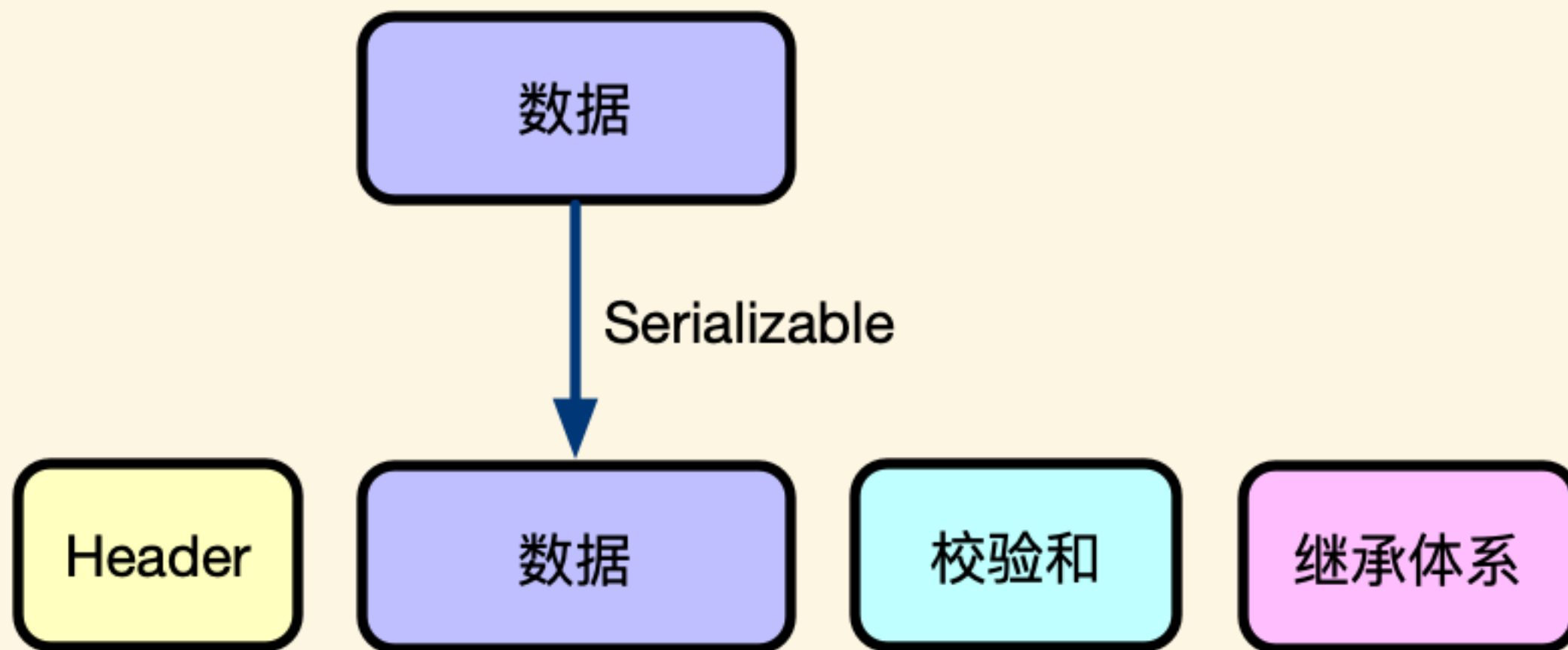
2 为什么要序列化?

- ❶ 一般来说，“活的”对象只生存在内存之中，关机断电就没有了.
- ❷ 而且，“活的”对象只能由本地进程使用，不能被发送到网络上的另外一台计算机.
- ❸ 然而，序列化可以存储“活的”对象，可以将“活的”对象发送到远程计算机.

3 为什么不用Java的序列化?

- Java的序列化是一个重量级序列化框架(Serializable);
- 一个对象被序列化后，会附带很多额外信息(各种校验信息、Header、继承体系等);
- 不便于在网络中高效传输;
- Hadoop自己开发了一套序列化机制(Writable).

3 为什么不用Java的序列化?



3 为什么不用Java的序列化?

👁 Hadoop序列化特点:

- (1) 紧凑: 高效使用存储空间;
- (2) 快速: 读写数据的开销小;
- (3) 可扩展: 通信协议的升级, 可升级;
- (4) 互操作性: 支持多语言的交互.