北京師範大學
BEIJING NORMAL UNIVERSITY

数据价值链路实践 第一讲

导论&数据采集与存储

主讲人：蓝昕

2021.07.31 ~ 08.28

每周六 晚7:00

大纲

1. 导论
2. 数据采集
3. 数据存储

北京師範大學
BEIJING NORMAL UNIVERSITY

导论

系列课程概要

自我介绍

> 16届北师大统计本科，现在某互联网大厂担任数据工程师
>
> 热衷于学习和分享数据行业相关知识

擅长领域

- **数据开发**
- **数据仓库**
- **数据治理**

课程目标

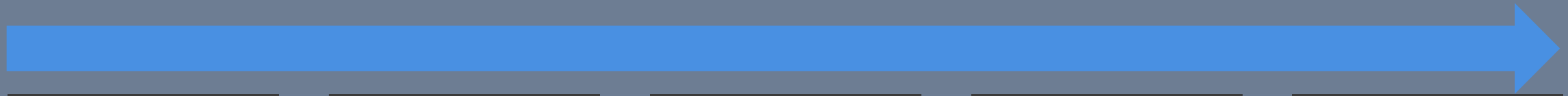
> 通过系列课程，了解并实践业界的数据价值产出流程和关联技术，方便后续有的放矢地深入学习。
>
> 熟悉这一套流程，一个人也可以做数据研究。

涉及全链路的所有岗位

- **爬虫工程**
- **数据开发**
- **数据科学**
- **算法工程**

有问题随时弹幕提问

价值生产链路



采集

收集各种有价值的数据

- 业务日志
- 埋点
- 爬虫
- 数据供应商
- IOT传感器
- 问卷调查

集成

将散落在各处的数据集中存储

- 数据库
- 大数据集群
- 同步程序

清洗&重组

将集中存放的数据重新整理和分类，持续更新，方便下游使用

- ETL流程
- 数据仓库
- 任务调度

分析

挖掘数据中的信息

- 统计对比
- 产业趋势
- 拟合预测
- 模型标记

应用

依据数据给出优化决策

- 投放策略
- 定价策略
- 优惠策略
- 产品策略

系列课程内容

1 采集存储

- 数据来源
- 爬虫技术
- 存储技术

2 清洗加工

- 任务调度
- 数据集成
- SQL
- pandas
- Spark

3 整合重组

- 中台概念
- 管道设计
- 数据仓库

4 挖掘价值

- 数据价值
- 算法框架
- 可视化思想

5 结论报告

- 观点输出
- PPT
- 文章

数据采集

开辟更多的数据源

数据来源

1 业务日志

记录业务活动的信息。如订单明细，销售状况，用户分类等。

常用于分析业务经营状况，制定运营策略。

2 埋点数据

记录用户在使用手机app或者网页时的操作行为。

常用于产品交互的设计优化，通常会结合A/B Test进行

3 爬虫数据

从公开的网络中获取数据

来自公开的网络，因此常用于分析市场状况。其它方面的应用也非常广泛。

作为学生，最低成本的数据来源

4 合作购买

向数据供应商购买/交换数据，形式多种多样。

按需购买，用金钱弥补技术、时间、团队方面的不足

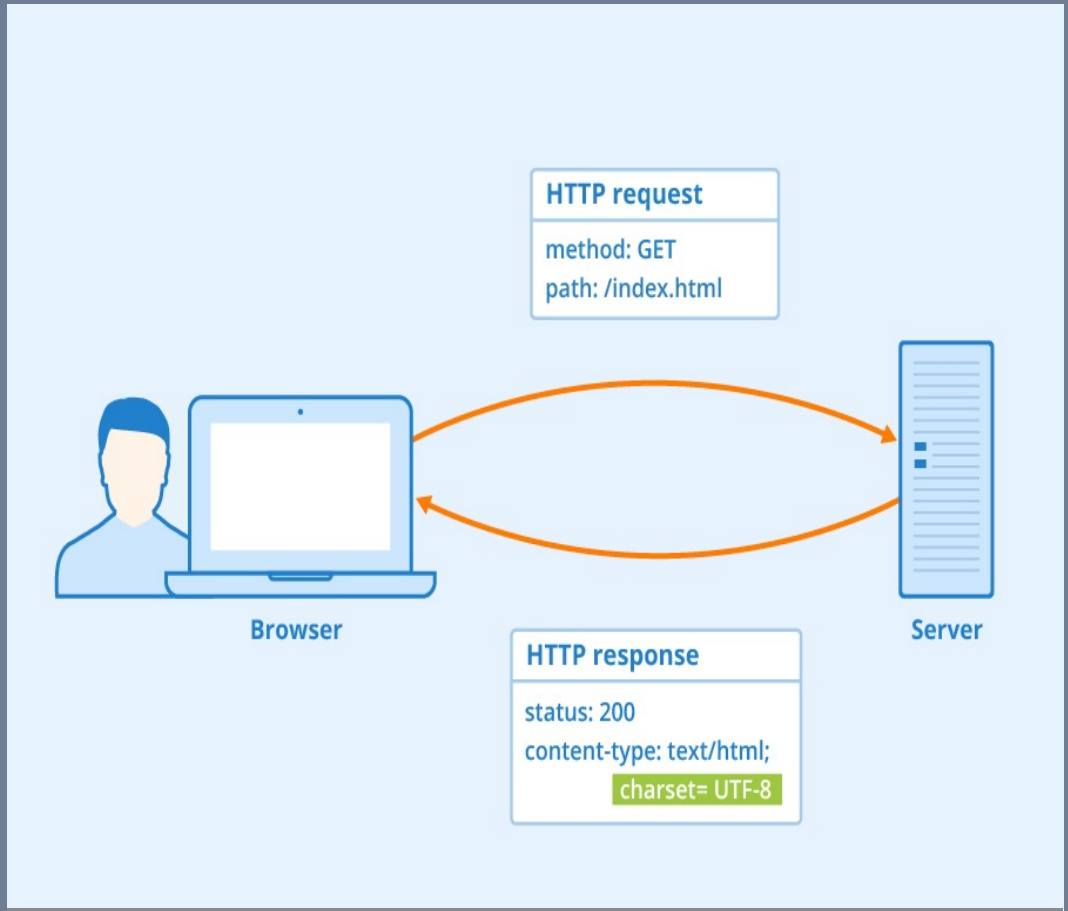
爬虫技术价值

爬虫技术，不仅用于数据采集

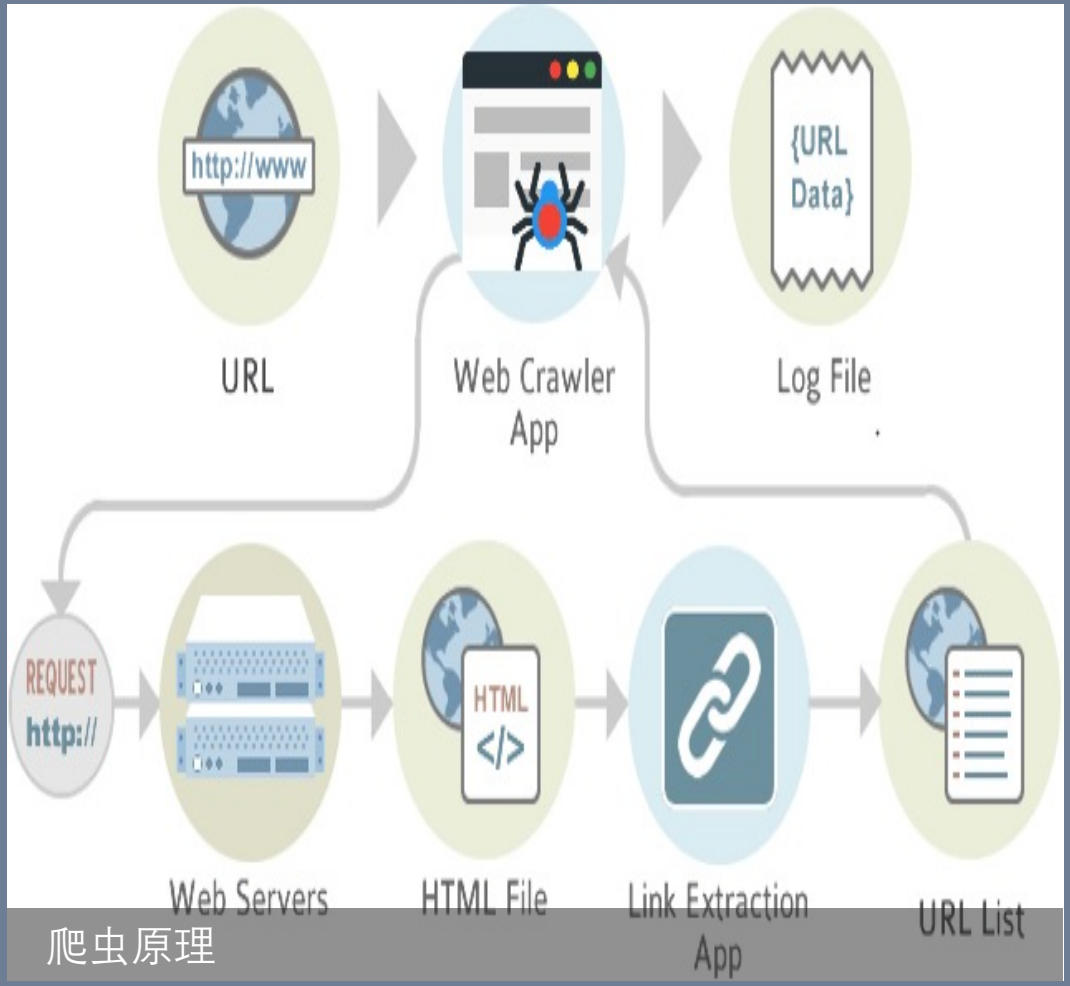
本质是用程序模拟人的网络行为，掌握这一技术，可以大大减轻手工负担。

- 持续收集信息
- 批量下载文件
- 批量提交信息
- 接入API

爬虫原理



Web访问



爬虫原理

编程模型

发送request

- 找到入口
- 突破验证
- 防止被封
- 并发提速

解析数据

- HTML
- XPATH
- CSS选择器
- 正则
- BeautifulSoup
- JSON

持续轮转

- 翻页
- 相关链接

怎么学习

原理

从爬虫的基本原理开始，写出完整可用的爬虫是第一步，了解本质，后续学习理解更容易

教程

阅读书籍或完整教程，深入了解技术细节和原理

《 Python3网络爬虫开发实战》 崔庆才（静觅）

实践

寻找数据入口，应对反爬、并发等问题，解析不同的数据格式。通过实践积累处理需求的经验

框架

学习爬虫框架，工程化，提升开发效率

数据存储

存放和使用数据

存储形式

Ice Cream Preferences

Field	Dislike	Neutral	Like
Pistachio	9	13	4
Vanilla	13	6	7
Strawberry	10	10	6

表格 - 人类阅读

便于阅读、筛选、计算

- CSV，Excel
- DataFrame
- 数据库表


```
{
 "scores": [
   {
     "Away_Score": 2,
     "Away_Team": "Newcastle",
     "Home_Score": 2,
     "Home_Team": "Arsenal"
   },
   {
     "Away_Score": 2,
     "Away_Team": "Napoli",
     "Home_Score": 4,
     "Home_Team": "Liverpool"
   }
 ]
}
```


特殊文本格式 - 机器交流

便于编程使用

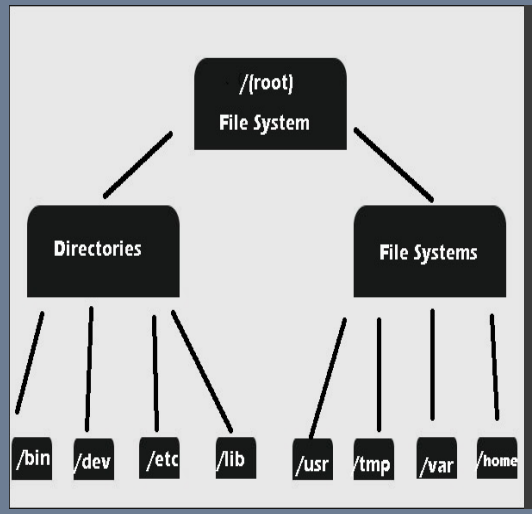
- JSON
- XML
- HTML
- YAML

特殊二进制 - 大规模存储

便于数据库管理

- Parquet
- Orcfile
- 特定数据库文件

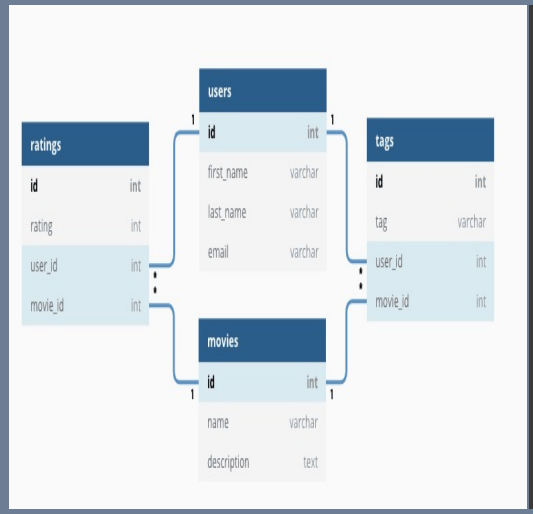
存储系统



文件系统

坚实的底层系统

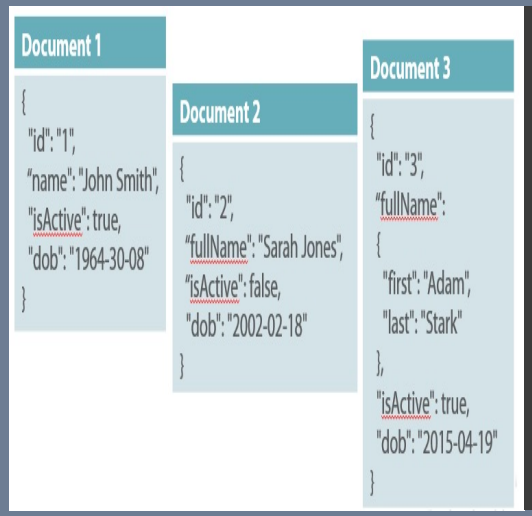
- 系统自带
- 操作符合直觉



关系型数据库

行业事实标准

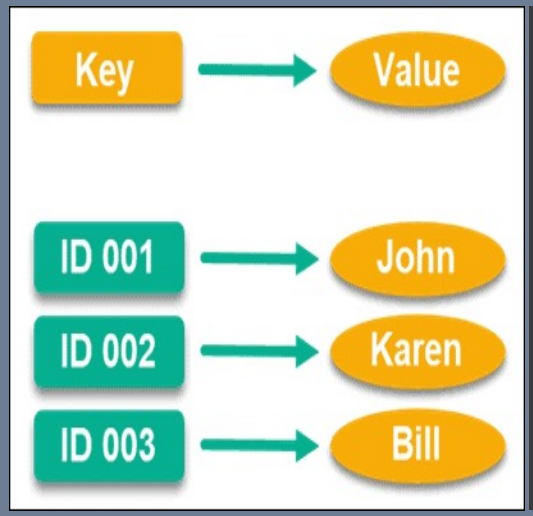
- 关联数据
- 阅读方便
- 查询方法-SQL



文档数据库

关系数据库的补充

- 每条数据一个文档
- 字段增减方便
- 独特的查询语言



Key-Value 数据库

常用于高速读写场景

- 最快的读写速度

分布式与大数据

分布式

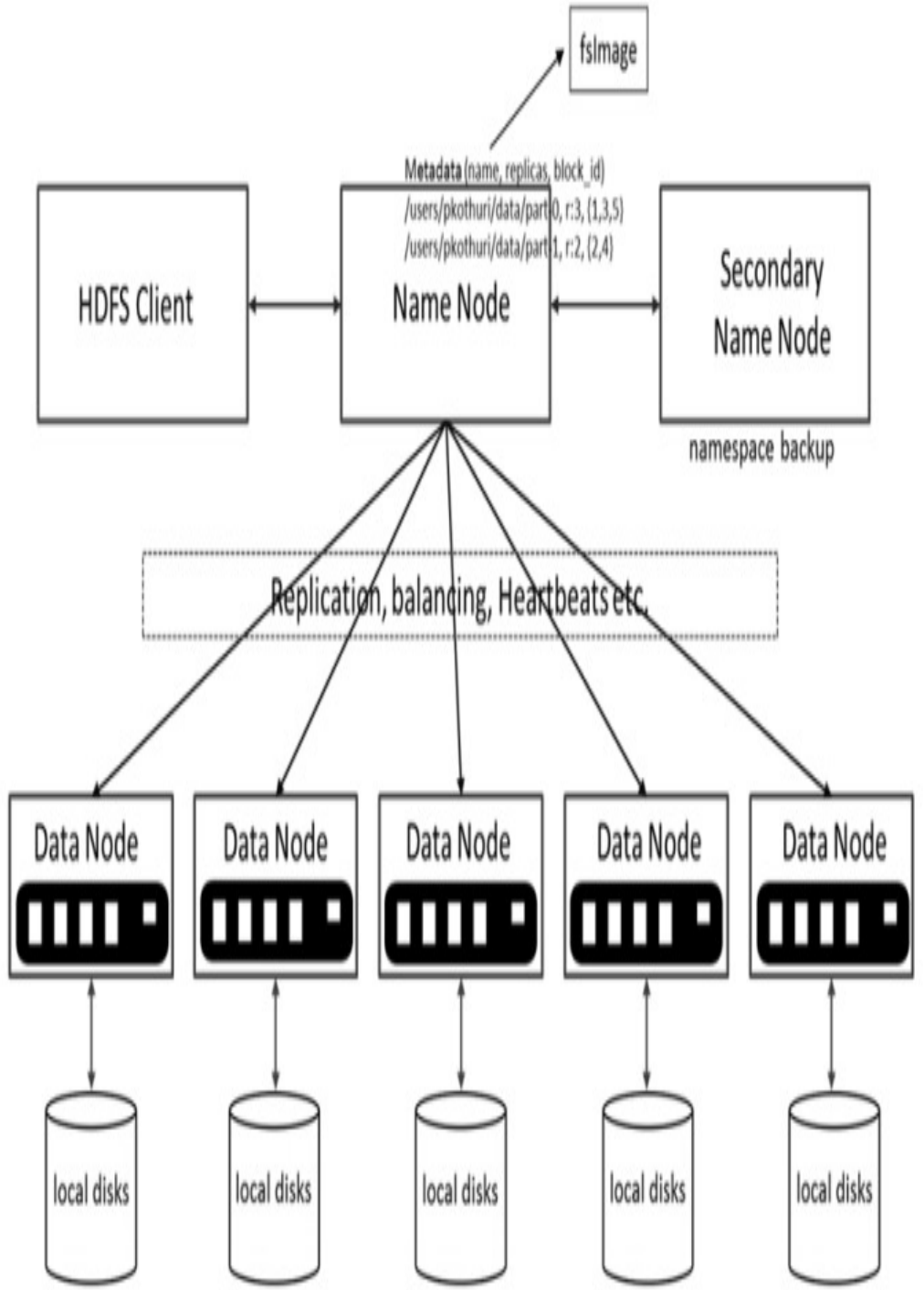
单台计算机性能有上限，通过分布式技术无限扩展机器集群，极大提升了存储与计算的上限

- HDFS：分布式文件系统，解决海量数据存储
- Map-Reduce：分布式计算方法，解决海量数据计算
- Hive : 基于HDFS存储的SQL查询引擎

大数据

分布式技术的出现，演变出了新的解决方案，创造了大数据业务场景。

4V特点：大容量（Volume），多样性（Variety），高速度（Velocity），真实性（Veracity）



北京師範大學
BEIJING NORMAL UNIVERSITY

第二讲 数据处理 2021.08.07

现场弹幕答疑

复杂问题投递邮箱 xingxing_fisher@163.com