大数据处理通用框架流程

大数据定义与特点

大数据的定义：无法用现有的软件工具提取、存储、搜索、共享、分析和处理的海量的、复杂的数据集合。

大数据的特点：

1. Volume：数据量大，包括采集、存储和计算的量都非常大。
2. Variety：种类和来源多样化。
3. Value：数据价值密度相对较低贵。
4. Velocity：数据增长速度快，处理速度也快，时效性要求高。
5. Veracity：数据的准确性和可信赖度，即数据的质量。

大数据的发展阶段

- **萌芽阶段**：20世纪90年代到21世纪，数据库技术与数据挖掘理论成熟，也称数据挖掘阶段。
- **突破阶段**：2003---2006年，非结构化的数据大量出现，传统的数据库处理难以应对，也称非结构化数据阶段。
- **成熟阶段**：2006---2009年，谷歌公开发表两篇论文《谷歌文件系统》和《基于集群的简单数据处理:MapReduce》，其核心的技术包括分布式文件系统GFS，分布式计算系统框架MapReduce，分布式数据库BigTable，这期间大数据研究的焦点是性能，云计算，大规模的数据集并行运算算法，以及开源分布式架构（Hadoop）
- **应用阶段**：2009年至今，大数据基础技术成熟之后，学术界及及企业界纷纷开始转向应用研究，2013年大数据技术开始向商业、科技、医疗、政府、教育、经济、交通、物流及社会的各个领域渗透，因此2013年也被称为大数据元年。

数据来源获取

1. **数据采集**：通过使用公司摄像头按照算法需求的数据要求，进行数据采集获取。如疲劳驾驶、活体检测、手势识别项目。

2. **网络爬取**：一方面是从网络下载开源数据集，另一方面通过网络爬取的方式，获取所需要的数据集。在网络爬虫方面通常使用Python进行网络爬取，常用的库有urllib、requests、scrapy、selenium和phantomjs。

采集的数据一定要经过数据处理与清洗，否则数据直接交付给算法和模型，只能是trash in，trash out。

1. **删除无用数据**：获取的数据含有一定不符合要求的数据，需要进行删除清洗。以确保算法得到的数据是高质量准确的数据。

2. **标记数据**：通过标记工具对数据进行人工标记处理，二维手势识别和3D手指关节点识别，都需要标记工具进行标记。

3. ***构造特征***：*通过其他方式处理数据，获取到算法需要的特征。*

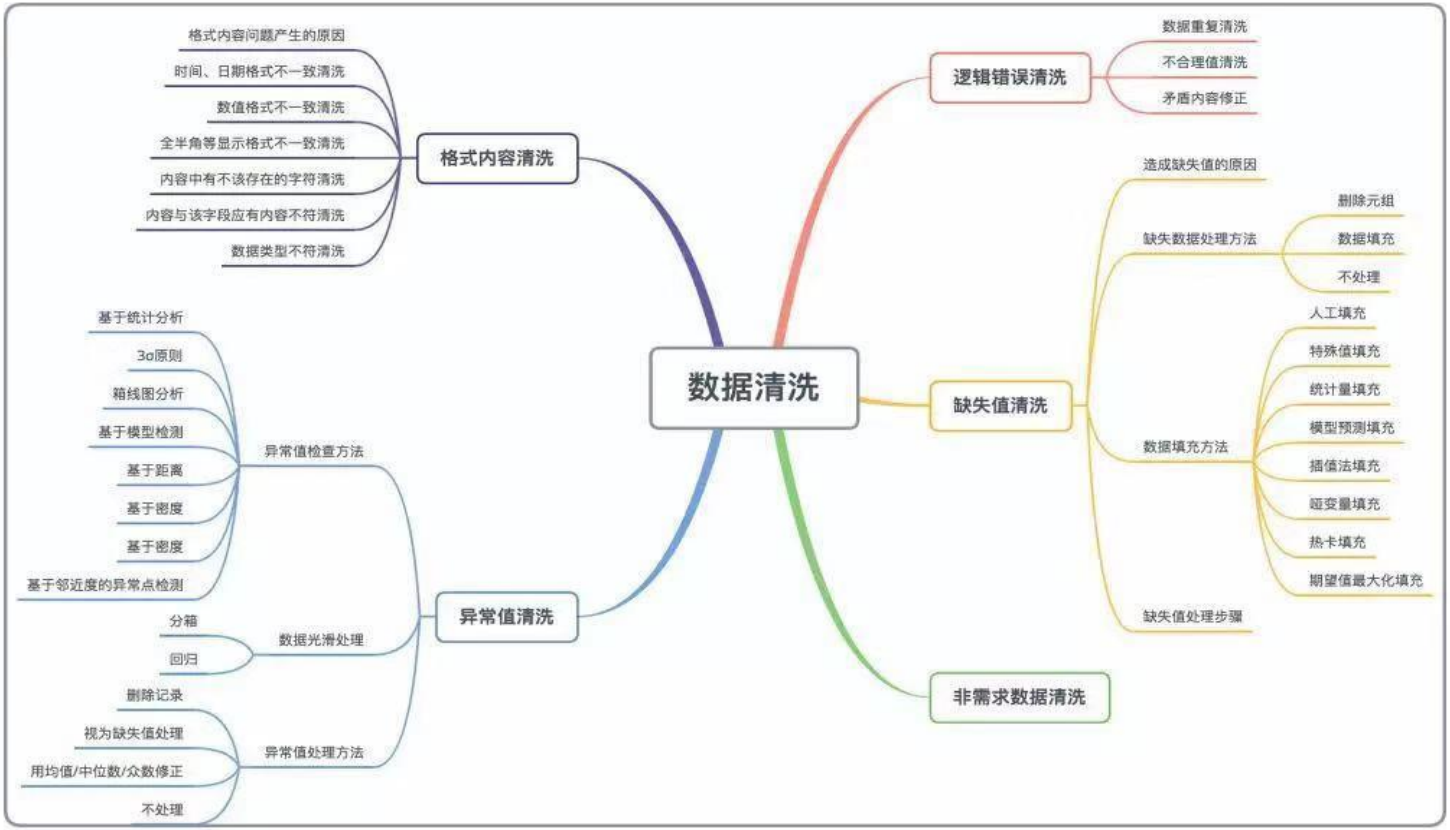
4. ***缺失处理***：*对于缺失数据进行增添处理。*

5. ***数据不均衡***：*下采样和上采样处理样本数据。*

6. ***异常值删除***：*删除不符合数据分布的异常点。*

Python数据处理与清洗过程中应用的库有os、shutil、numpy和pandas。

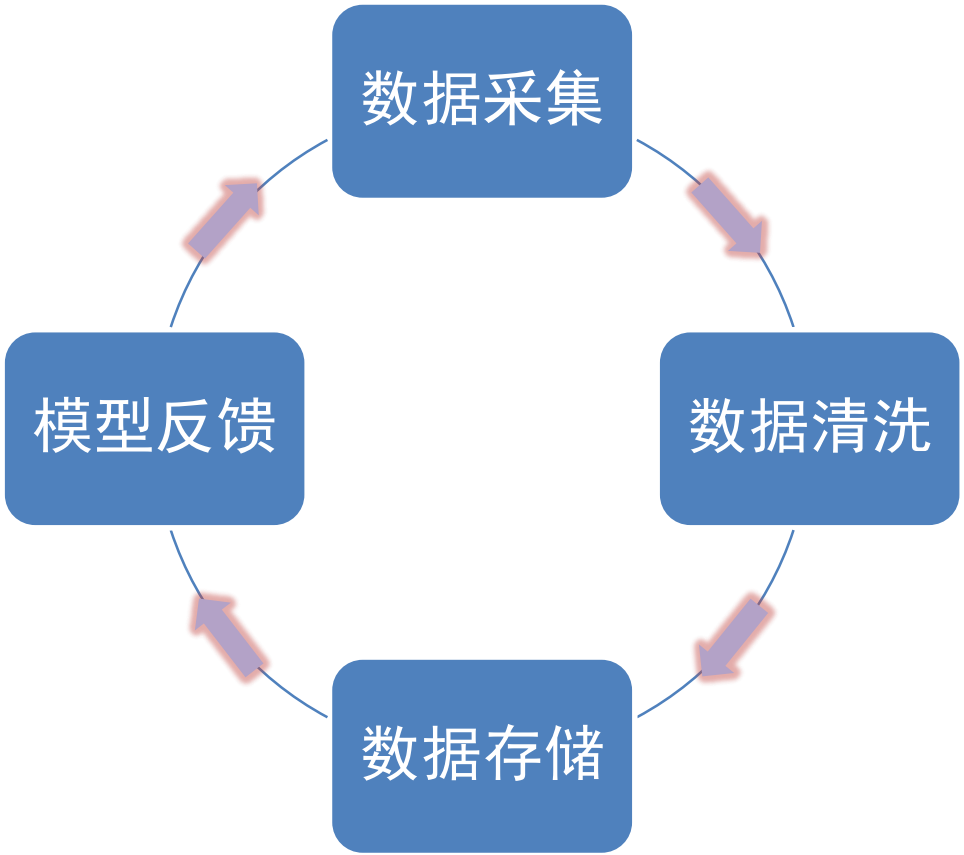
数据处理清洗


数据清洗
格式内容清洗
格式内容问题产生的原因
时间、日期格式不一致清洗
数值格式不一致清洗
全半角等显示格式不一致清洗
内容中有不该存在的字符清洗
内容与该字段应有内容不符清洗
数据类型不符清洗
逻辑错误清洗
数据重复清洗
不合理值清洗
矛盾内容修正
缺失值清洗
造成缺失值的原因
缺失数据处理方法
删除元组
数据填充
不处理
数据填充方法
人工填充
特殊值填充
统计量填充
模型预测填充
插值法填充
哑变量填充
热卡填充
期望值最大化填充
缺失值处理步骤
异常值清洗
异常值检查方法
基于统计分析
3σ原则
箱线图分析
基于模型检测
基于距离
基于密度
基于密度
基于邻近度的异常点检测
数据光滑处理
分箱
回归
异常值处理方法
删除记录
视为缺失值处理
用均值/中位数/众数修正
不处理
非需求数据清洗

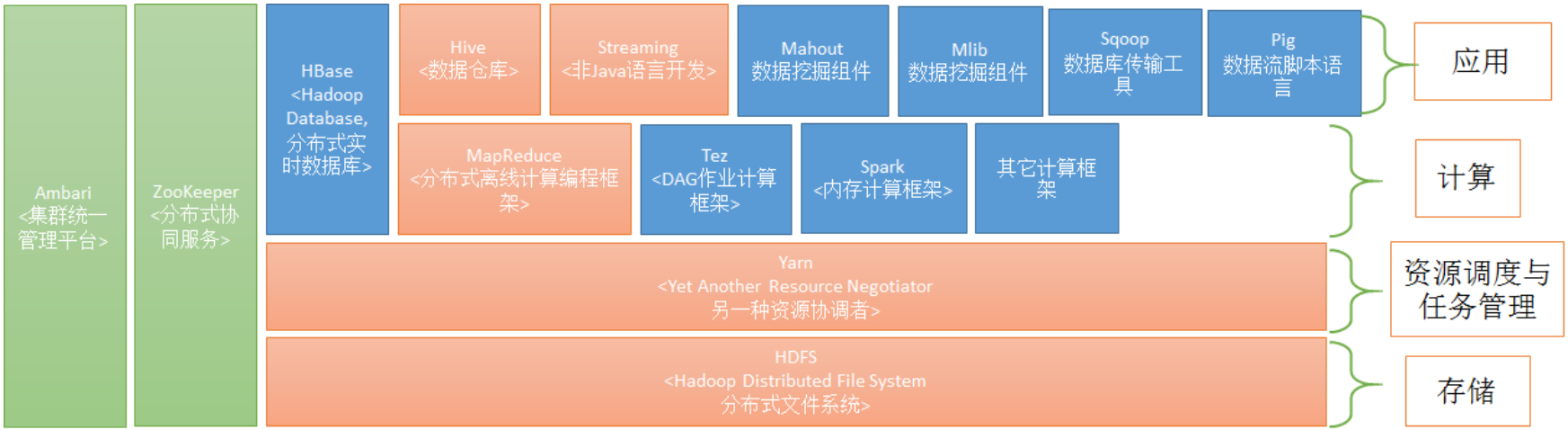
21世纪数据就是石油，谁拥有了数据，谁就拥有了未来。

1. FTP传输存储：将处理好的数据上传到FTP进行传输存储。
2. 网盘传输存储：将处理好的数据上传到网盘进行传输存储。
3. 本地硬盘存储：将处理好的数据存储到本地硬盘。
4. *数据库存储：结构化数据库MySQL、Oracle、SQL Sever，非结构化数据库MongoDB、Redis。*

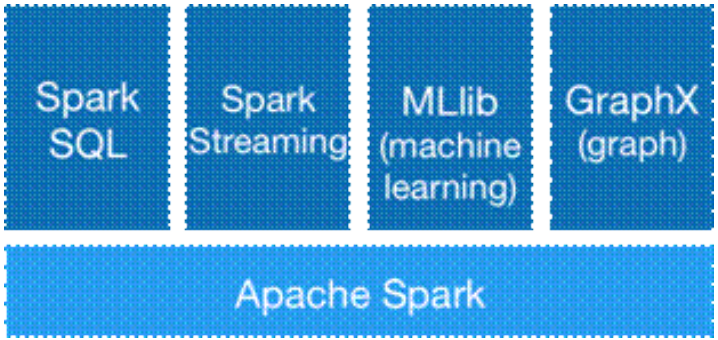
数据流程



大数据框架---Hadoop生态圈



Hadoop是一种分布式大数据处理框架，主要功能为HDFS和MapReduce，另外Hadoop生态圈还具有Hbase数据库和Hive数据仓库存储查询数据，Mahout提供机器学习接口，Yarn提供资源调度与管理。Hadoop的出现对大数据处理提供了强有力的解决方案。



基于Hadoop的离线计算框架，响应处理速度慢、MapReduce函数单一、基于磁盘读写无法进行迭代等问题，这些问题统统被Spark大数据计算框架解决。Spark基于内存计算、提供比MapReduce更丰富的函数、Spark Streaming提供流计算可达到秒级响应速度甚至毫秒级响应，Mlib提供机器学习算法接口，GraphX提供图计算，Spark SQL提供SQL查询使用。可以说，Spark可以完全替代Hadoop中的MapReduce。

大数据的未来

大数据+算法+算力 ⟶ 人工智能

现阶段还处于对数据处理投入大量的人工，才能产生相应的智能。

随着数据财富的日益积累，社会将会以数据驱动未来。

Thanks for your listening