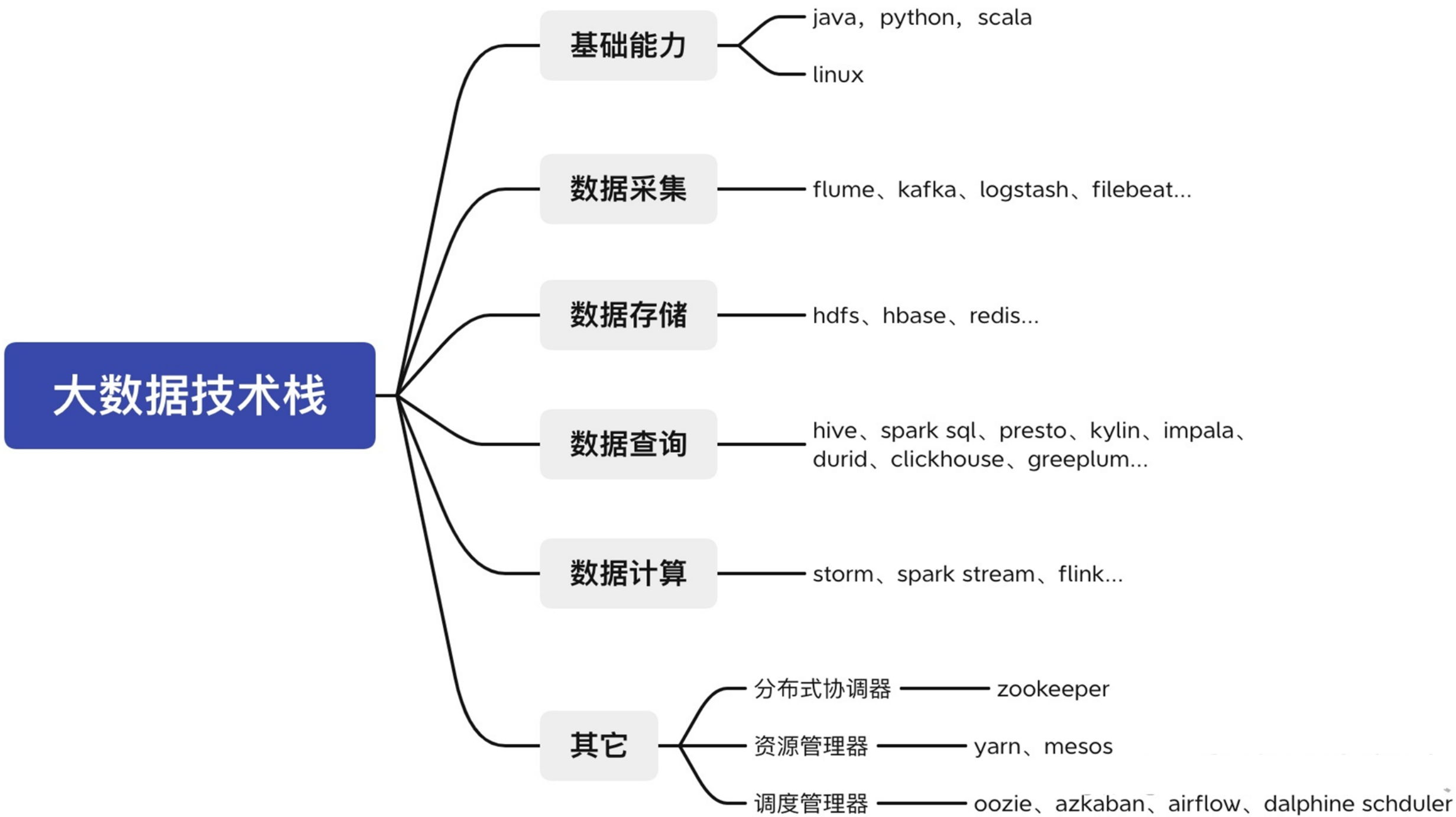
第一章

大数据技术全栈


大数据技术栈
基础能力
java，python，scala
linux
数据采集
flume、kafka、logstash、filebeat...
数据存储
hdfs、hbase、redis...
数据查询
hive、spark sql、presto、kylin、impala、
durid、clickhouse、greeplum...
数据计算
storm、spark stream、flink...
其它
分布式协调器
zookeeper
资源管理器
yarn、mesos
调度管理器
oozie、azkaban、airflow、dalphine schduler

大数据技术知识体系

基础技术

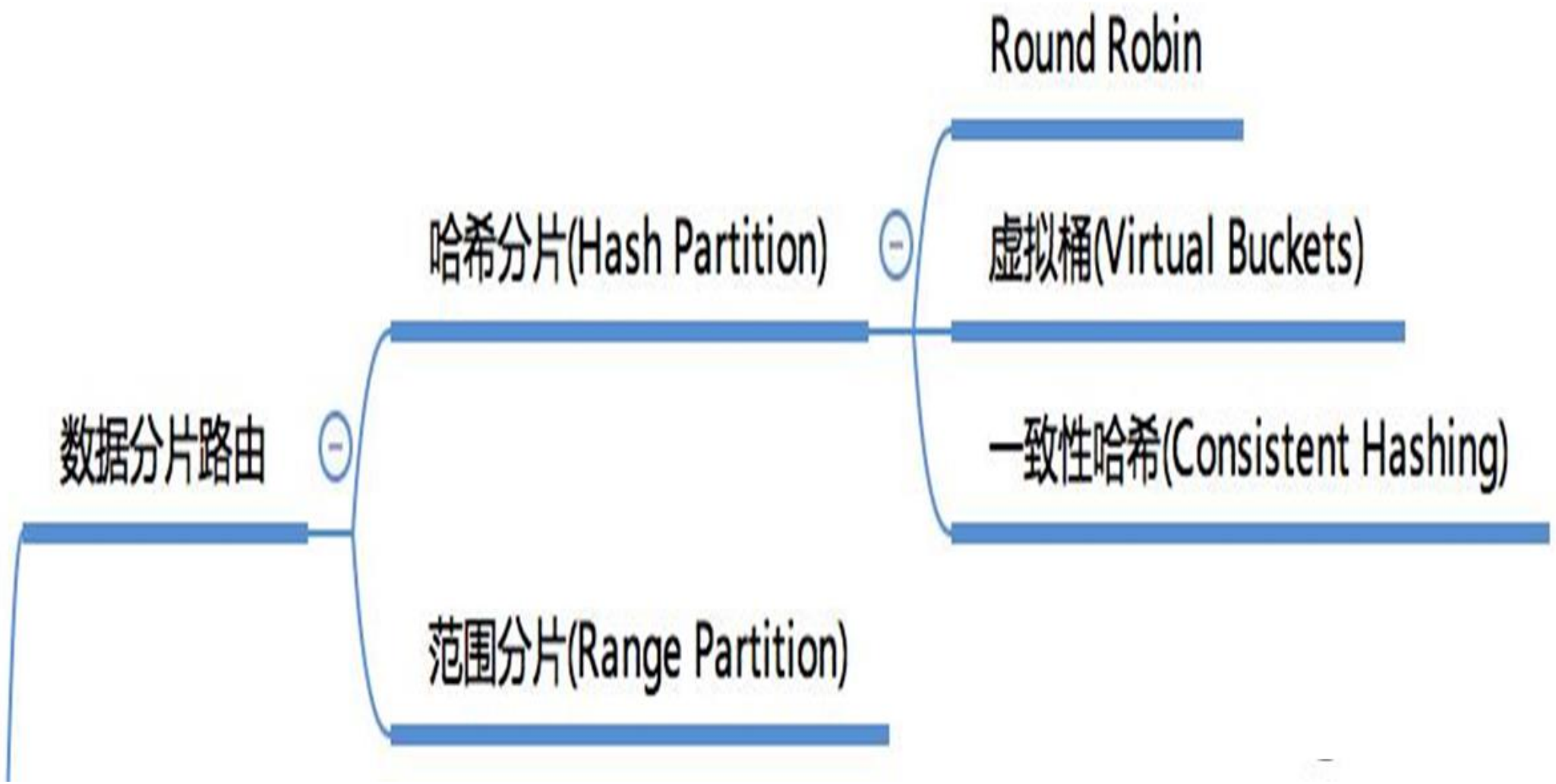
数据应用

数据组织集成

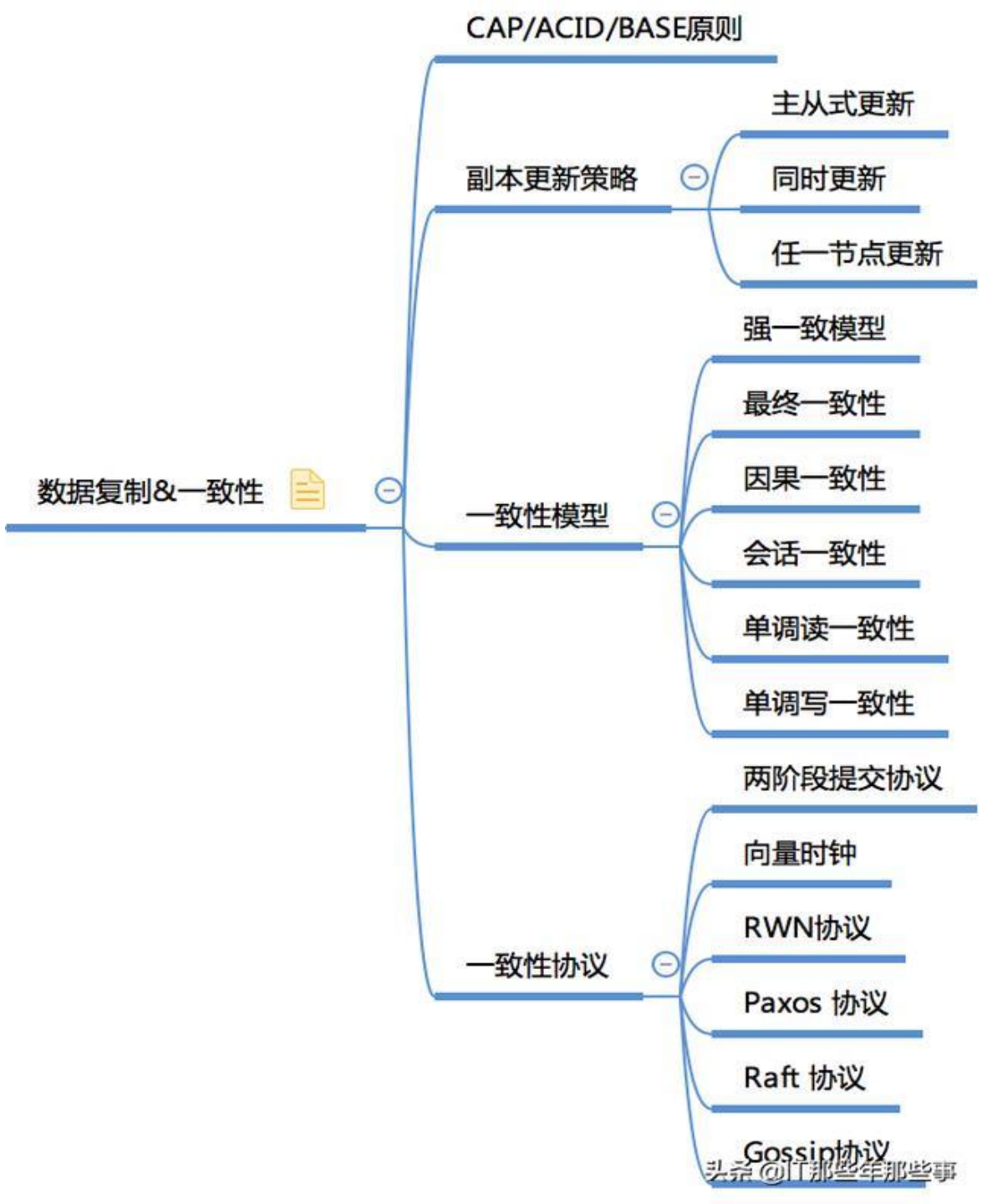
数据传输

数据采集

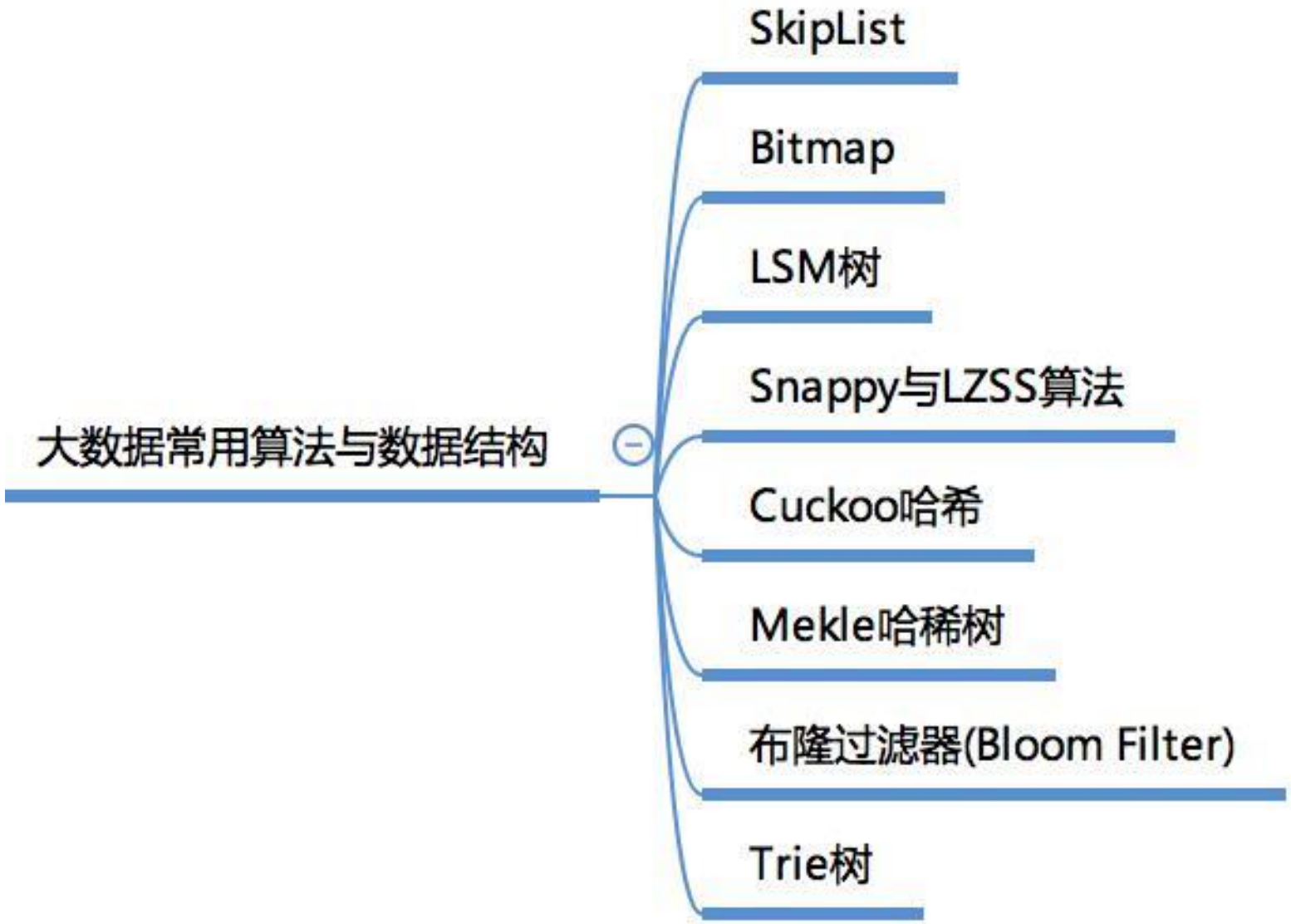
数据治理


数据分片路由
哈希分片(Hash Partition)
Round Robin
虚拟桶(Virtual Buckets)
一致性哈希(Consistent Hashing)
范围分片(Range Partition)

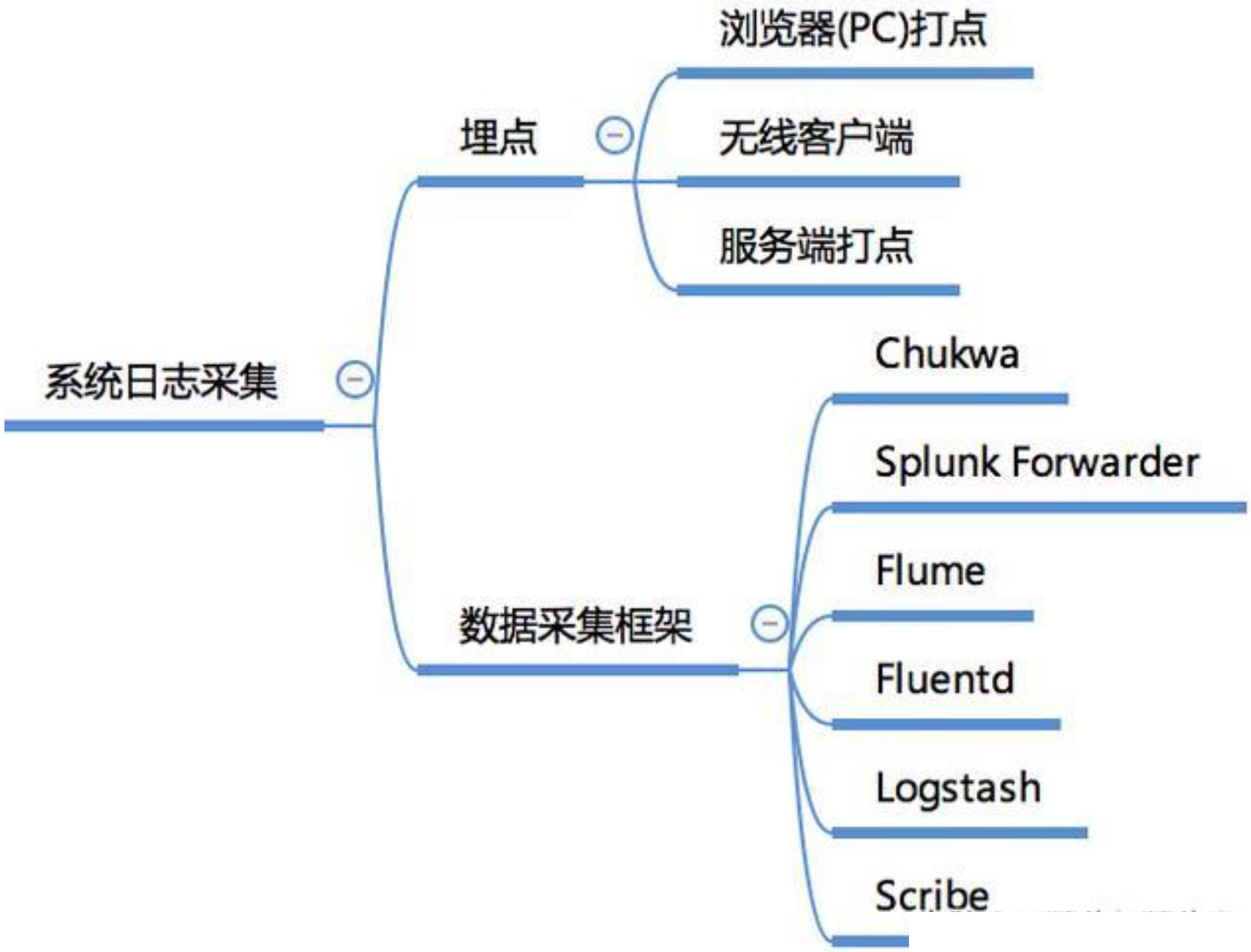
大数据基础技术



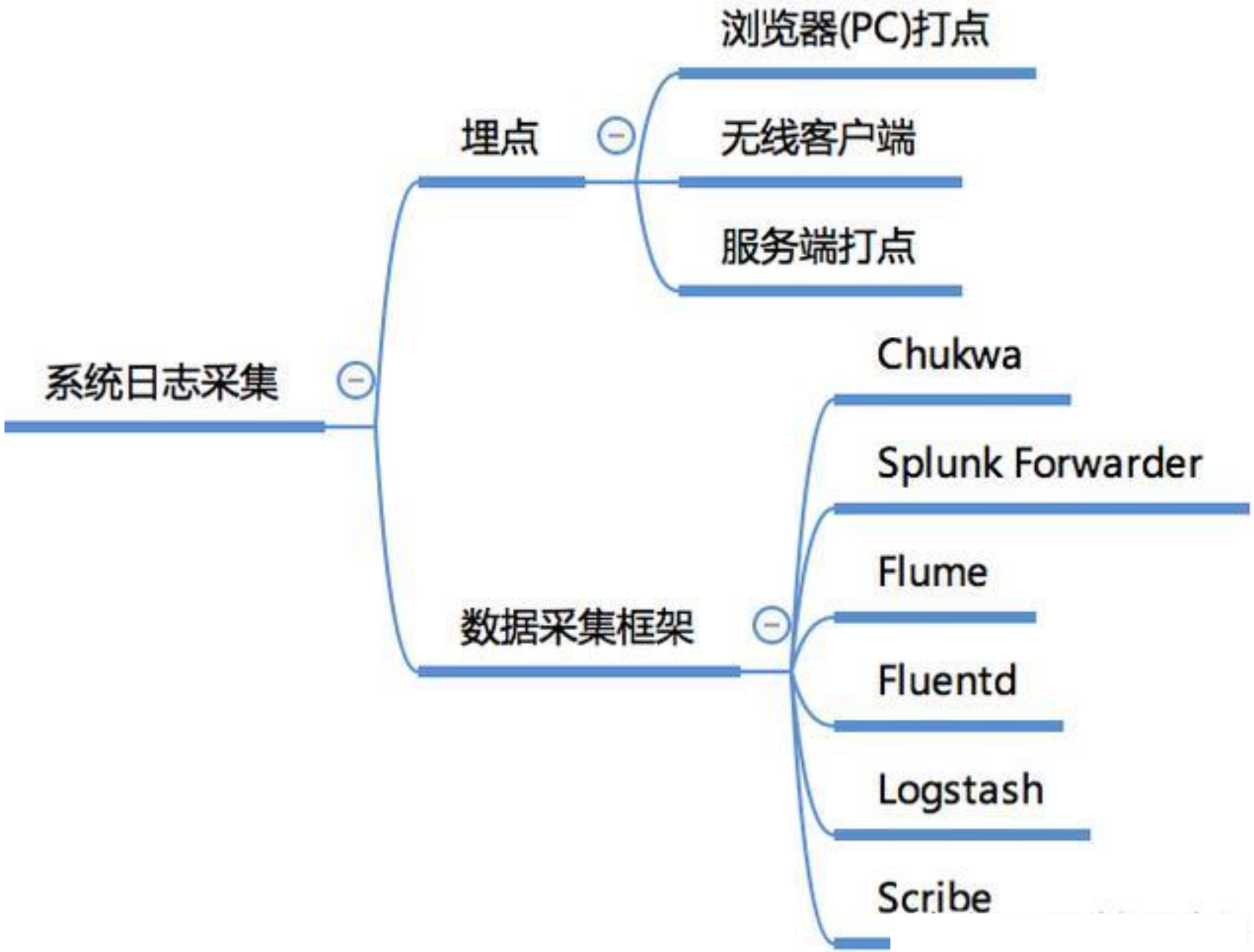
大数据基础技术


大数据常用算法与数据结构
SkipList
Bitmap
LSM树
Snappy与LZSS算法
Cuckoo哈希
Mekle哈稀树
布隆过滤器(Bloom Filter)
Trie树

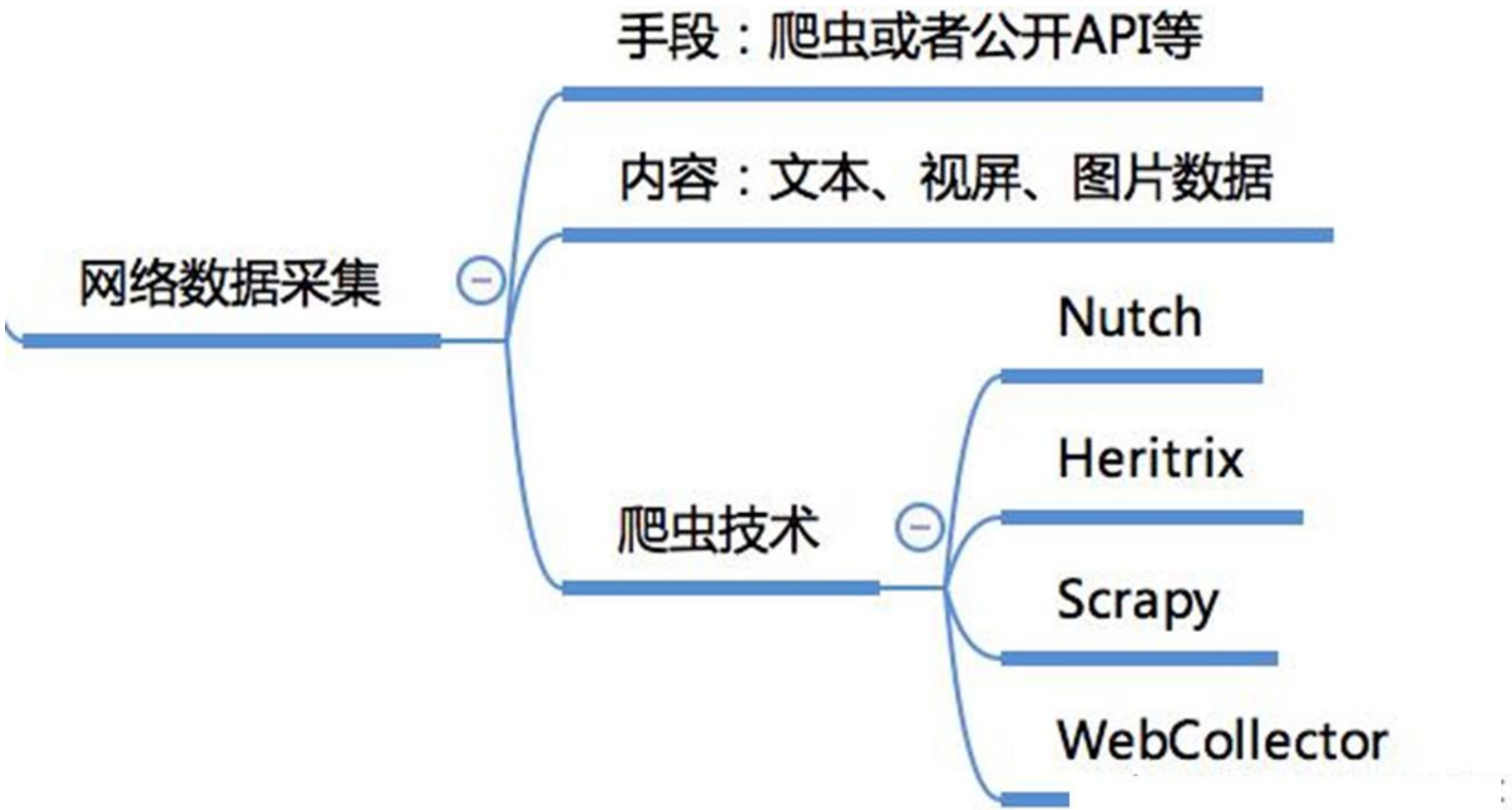
数据采集


系统日志采集
埋点
浏览器(PC)打点
无线客户端
服务端打点
数据采集框架
Chukwa
Splunk Forwarder
Flume
Fluentd
Logstash
Scribe

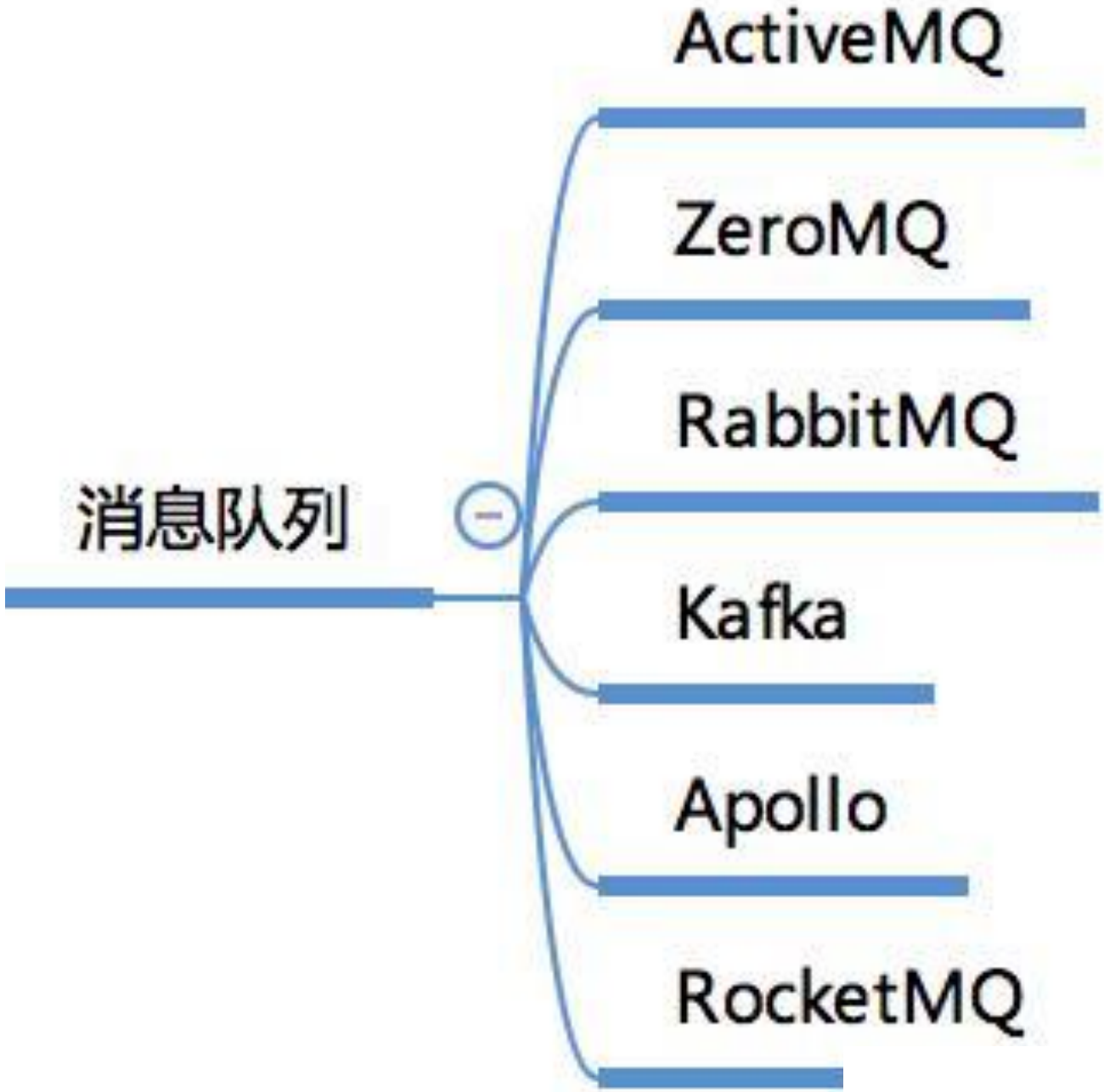
数据采集


系统日志采集
埋点
浏览器(PC)打点
无线客户端
服务端打点
数据采集框架
Chukwa
Splunk Forwarder
Flume
Fluentd
Logstash
Scribe

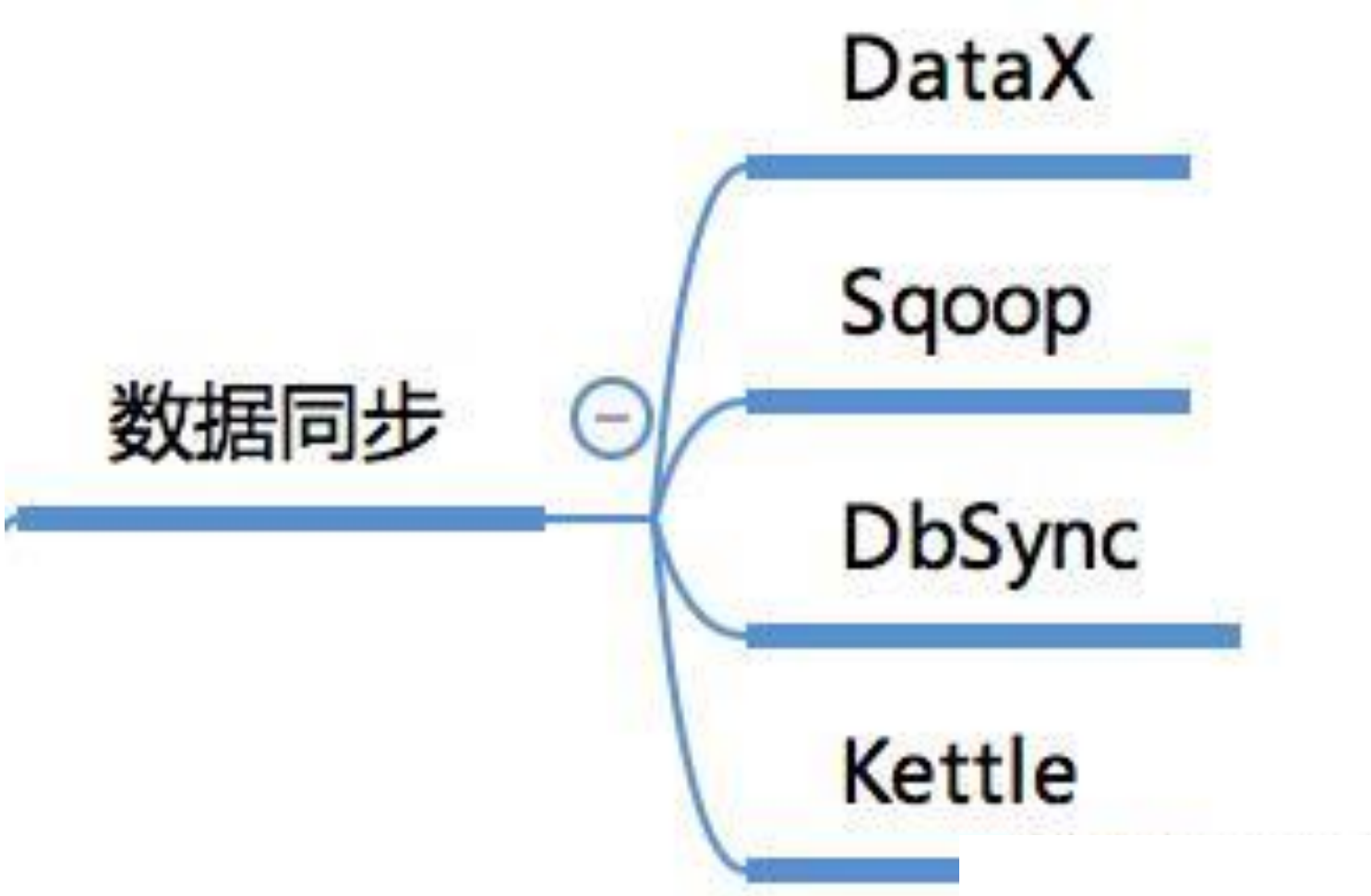
数据采集


网络数据采集
手段：爬虫或者公开API等
内容：文本、视屏、图片数据
爬虫技术
Nutch
Heritrix
Scrapy
WebCollector

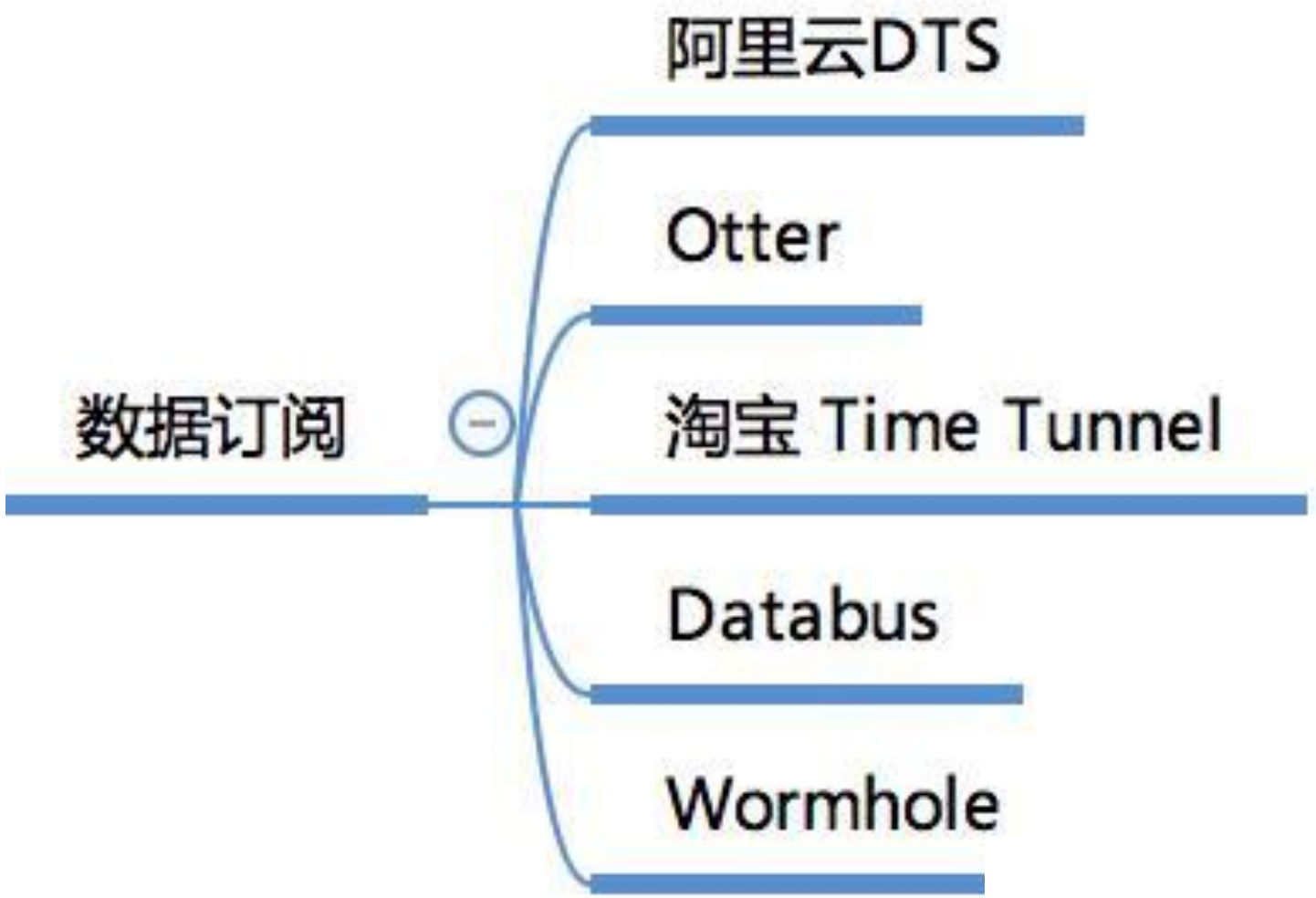
数据传输



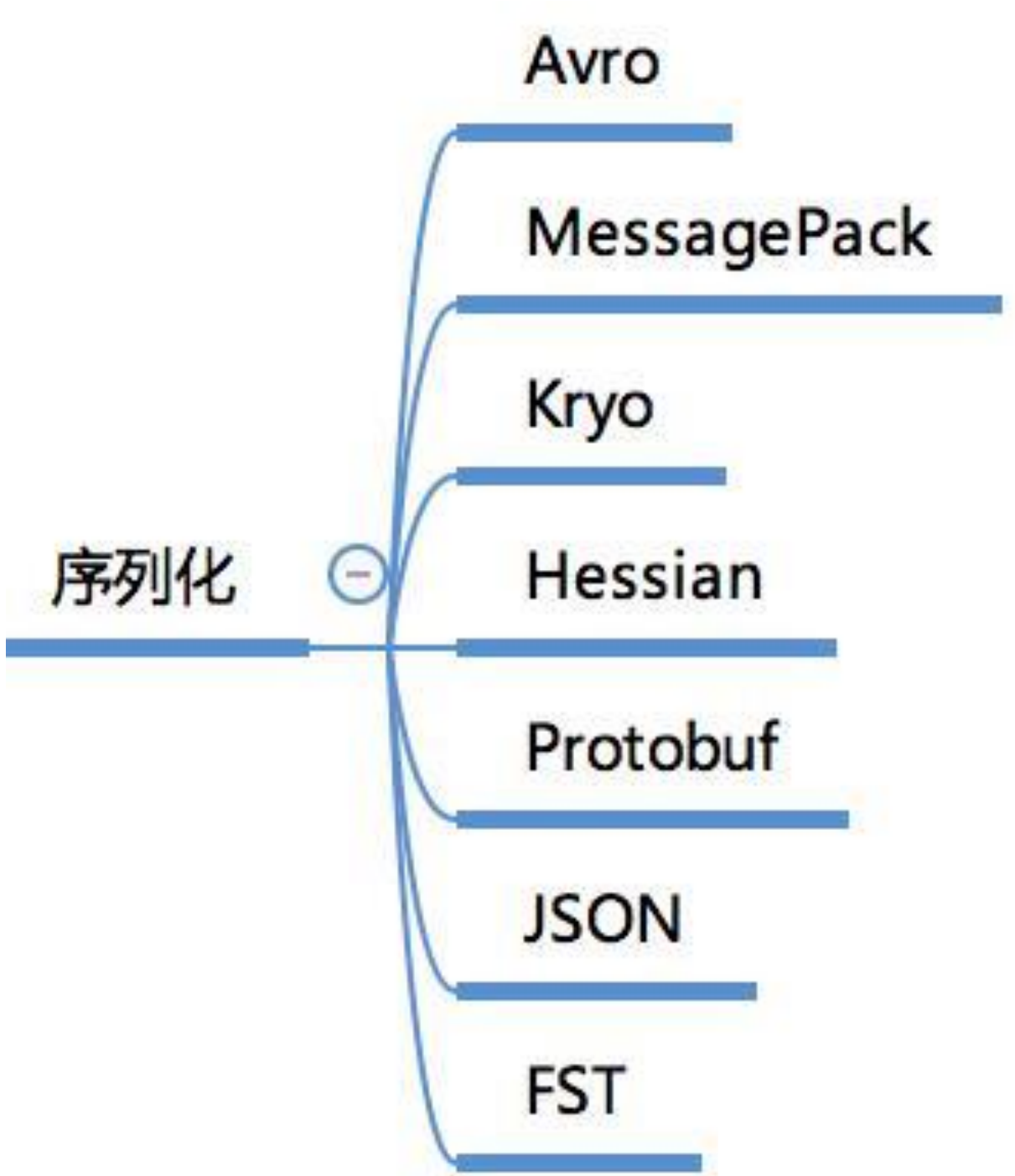
数据传输



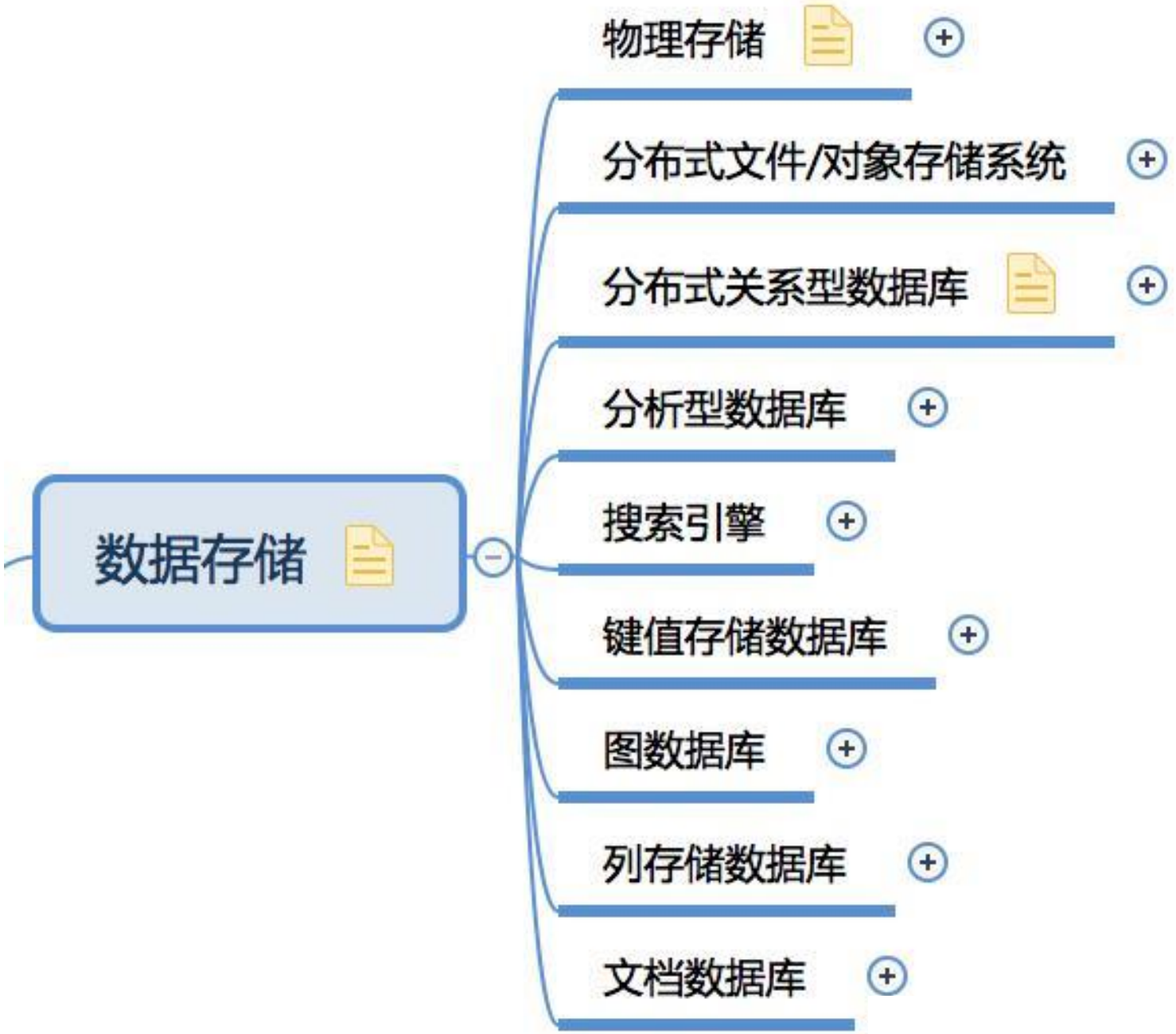
数据传输



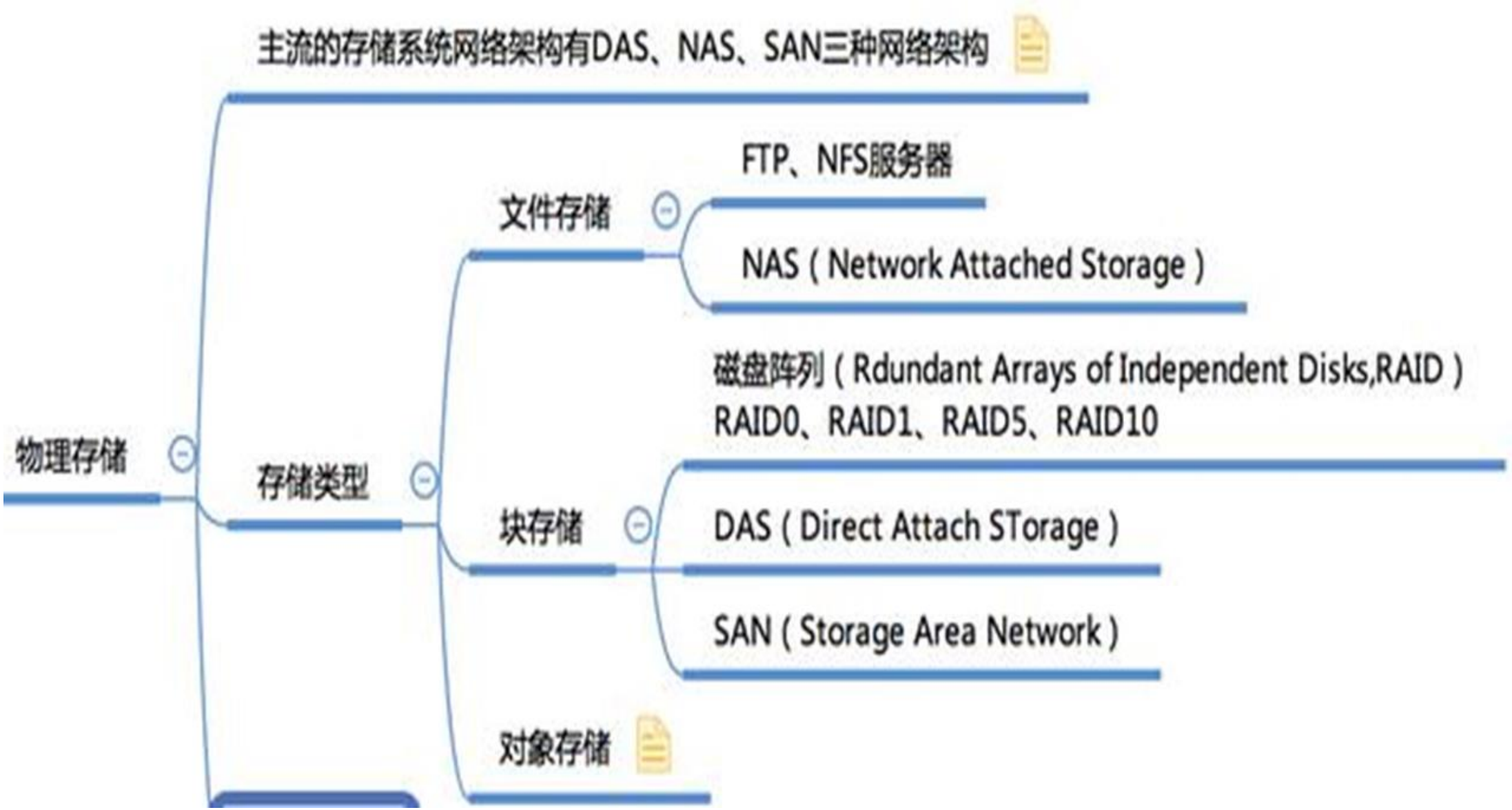
数据传输



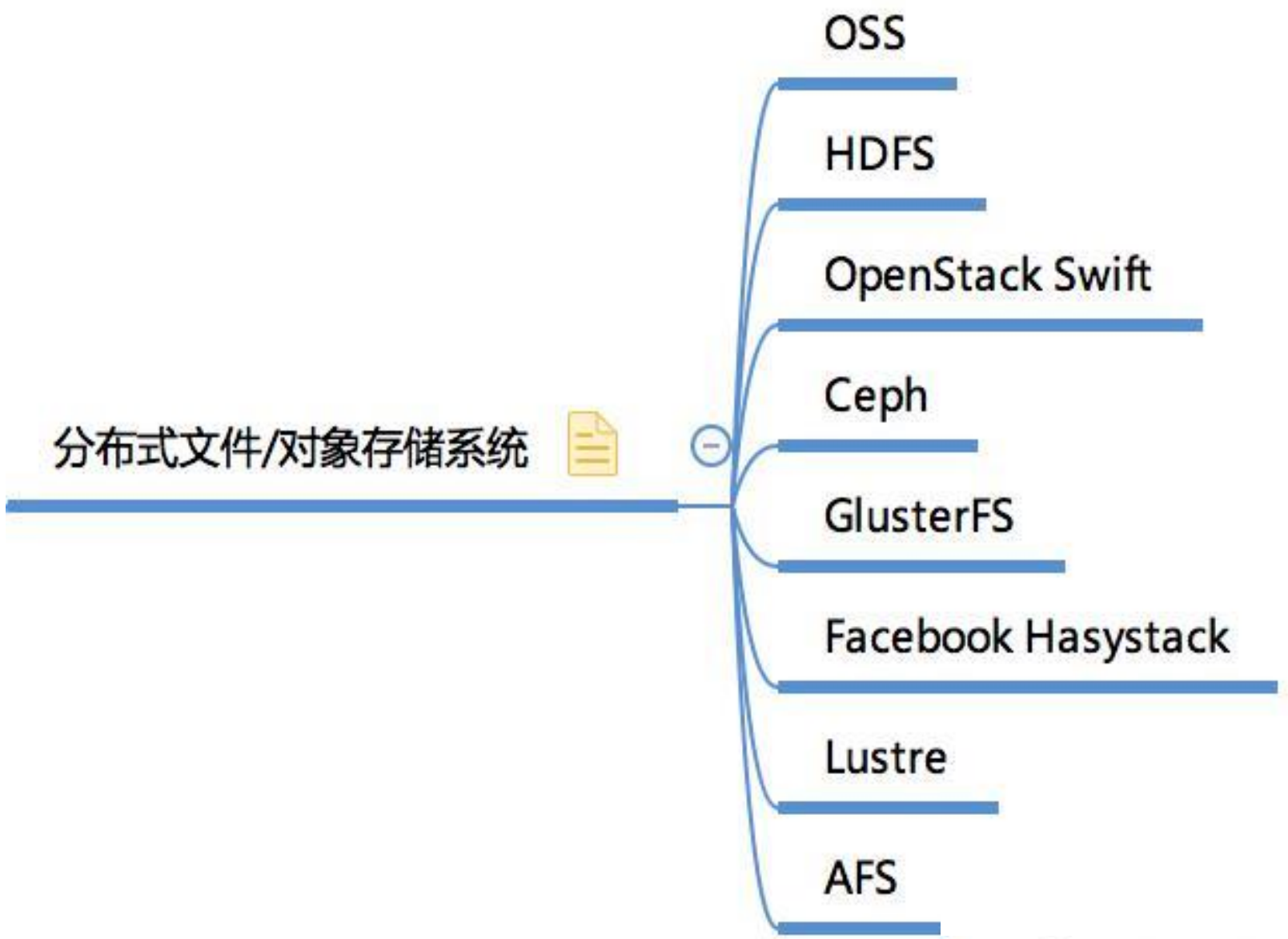
数据组织集成


数据存储
物理存储
分布式文件/对象存储系统
分布式关系型数据库
分析型数据库
搜索引擎
键值存储数据库
图数据库
列存储数据库
文档数据库

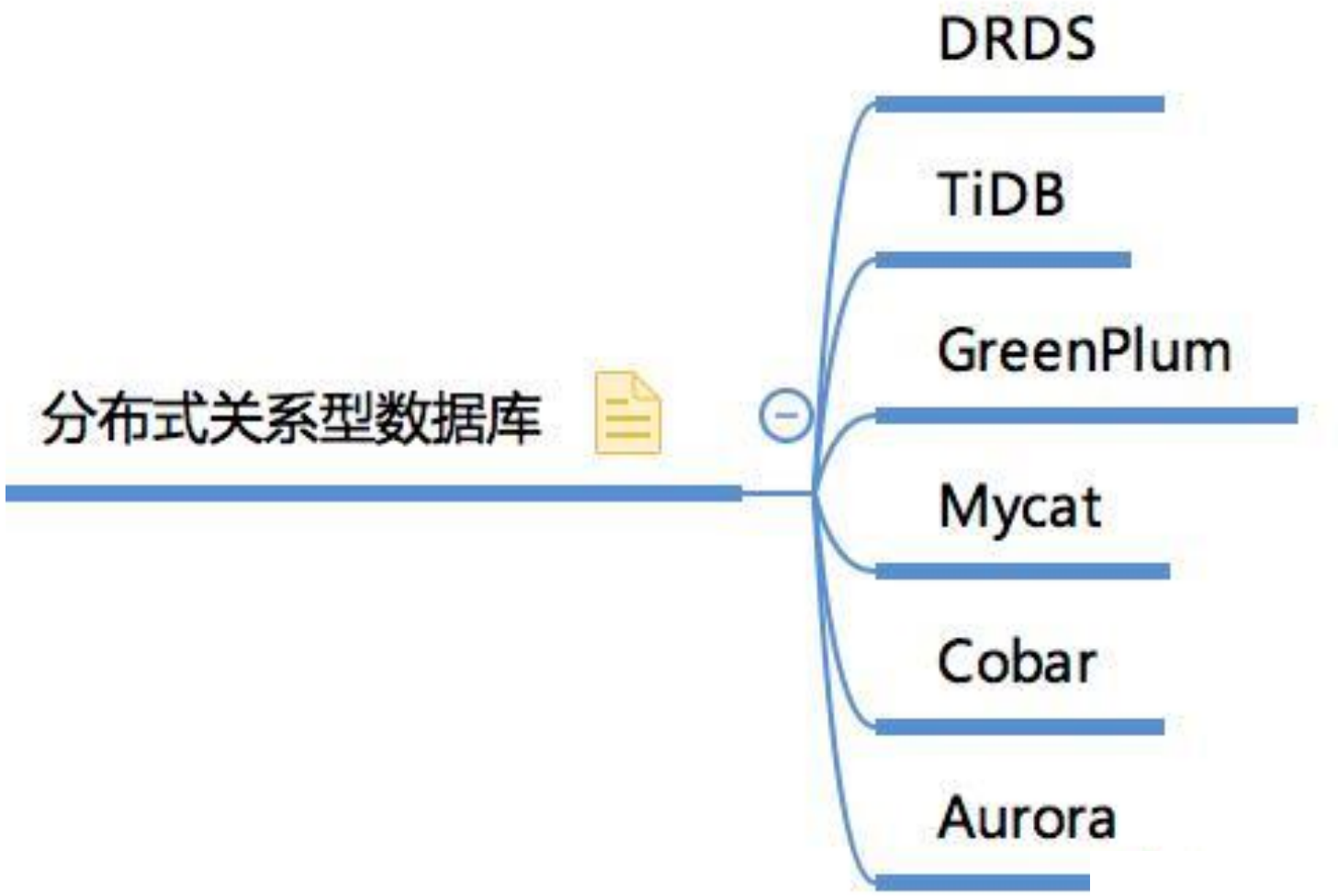
数据存储


主流的存储系统网络架构有DAS、NAS、SAN三种网络架构
物理存储
存储类型
文件存储
FTP、NFS服务器
NAS (Network Attached Storage)
块存储
磁盘阵列 (Rdundant Arrays of Independent Disks,RAID)
RAID0、RAID1、RAID5、RAID10
DAS (Direct Attach STorage)
SAN (Storage Area Network)
对象存储

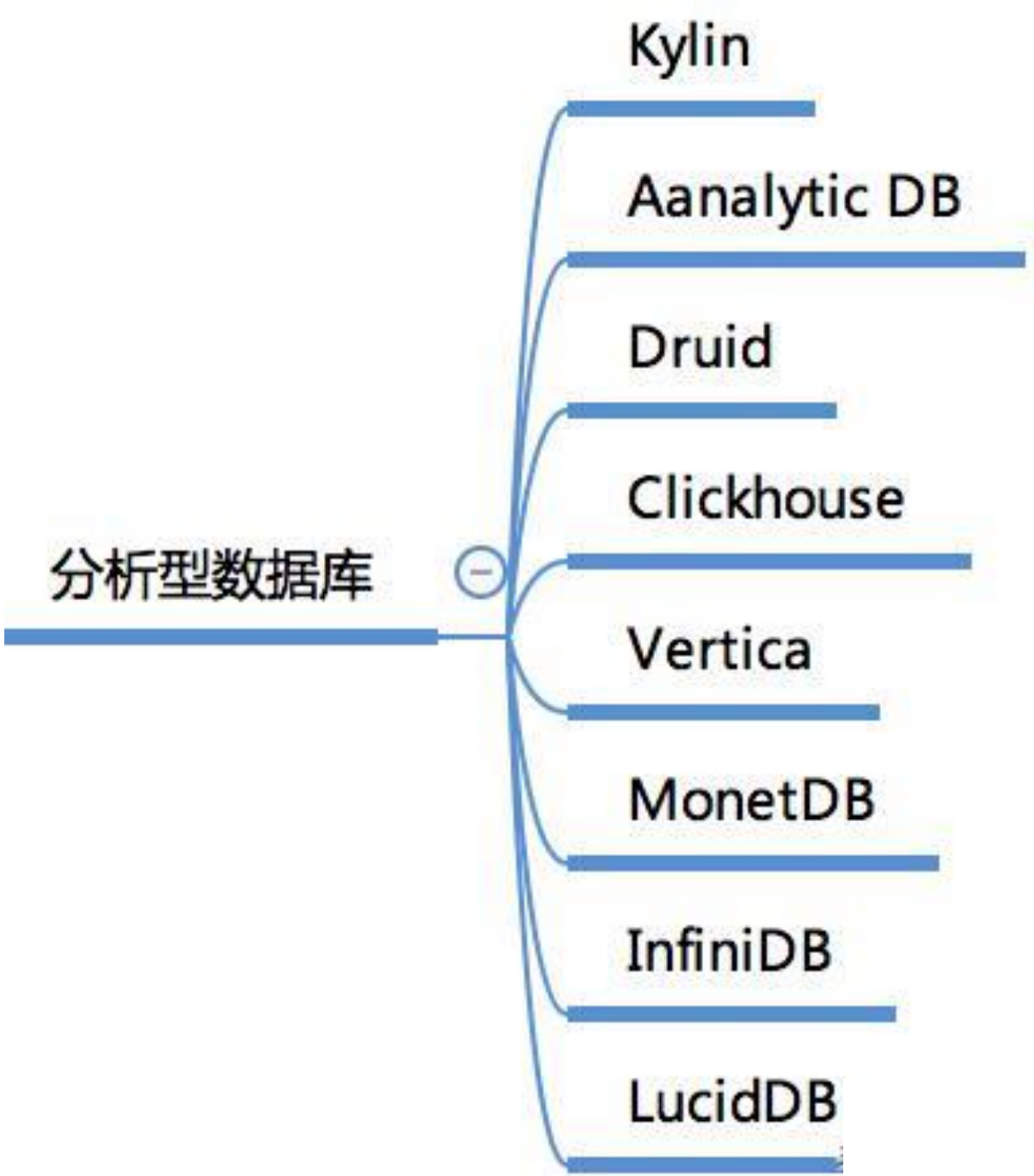
数据存储


分布式文件/对象存储系统
OSS
HDFS
OpenStack Swift
Ceph
GlusterFS
Facebook Hasystack
Lustre
AFS

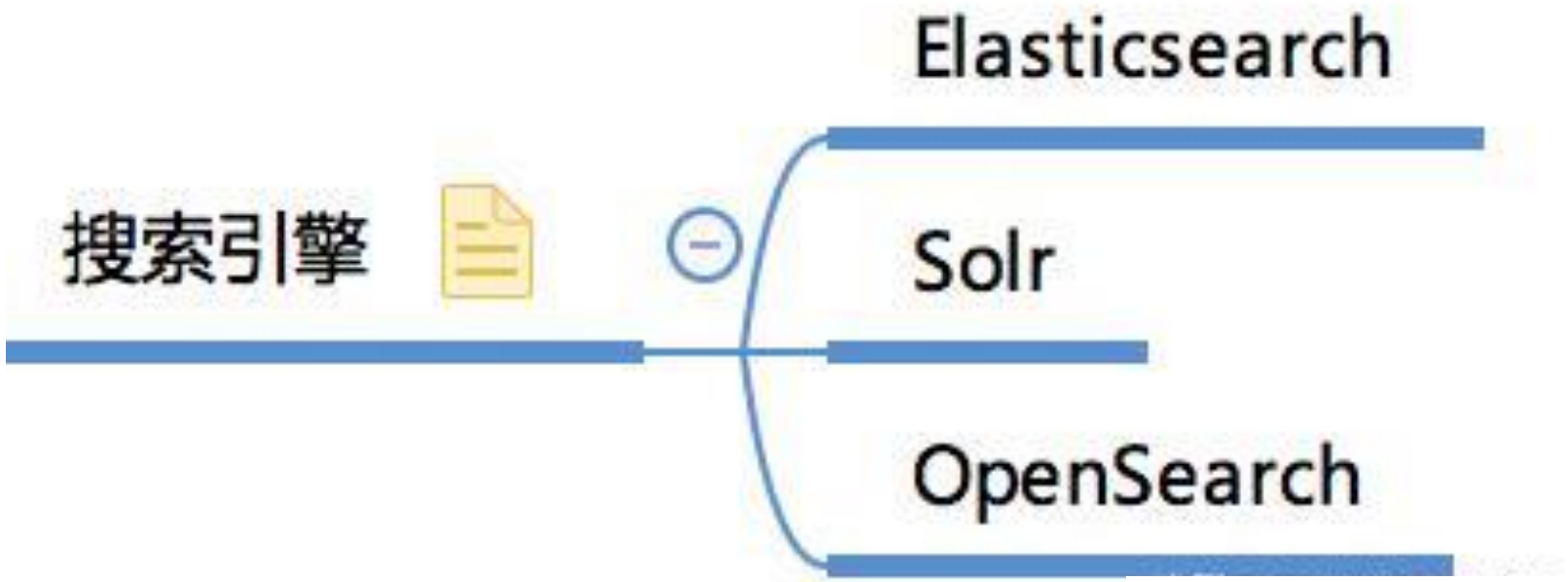
数据存储



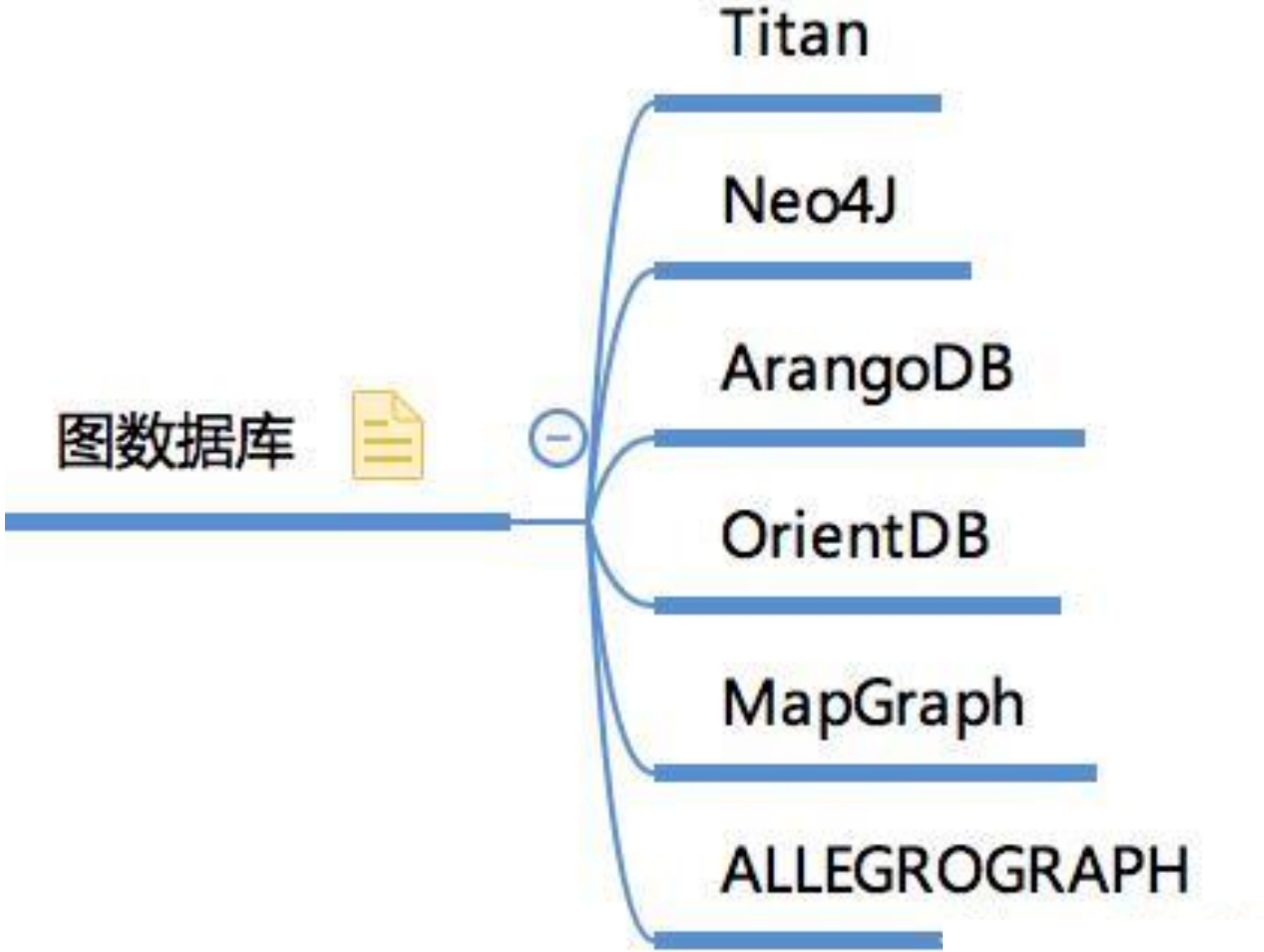
数据存储


分析型数据库
Kylin
Aanalytic DB
Druid
Clickhouse
Vertica
MonetDB
InfiniDB
LucidDB

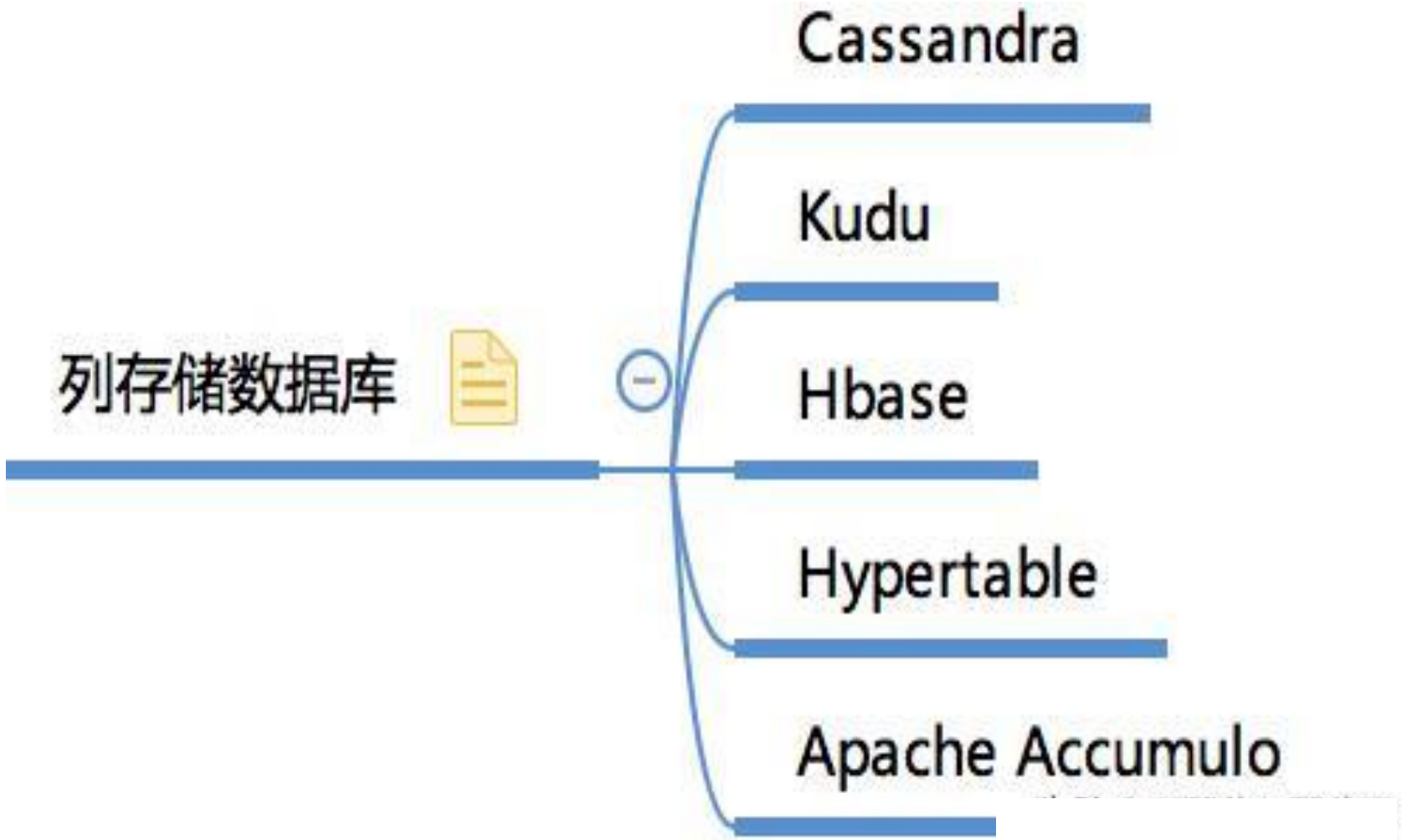
数据存储



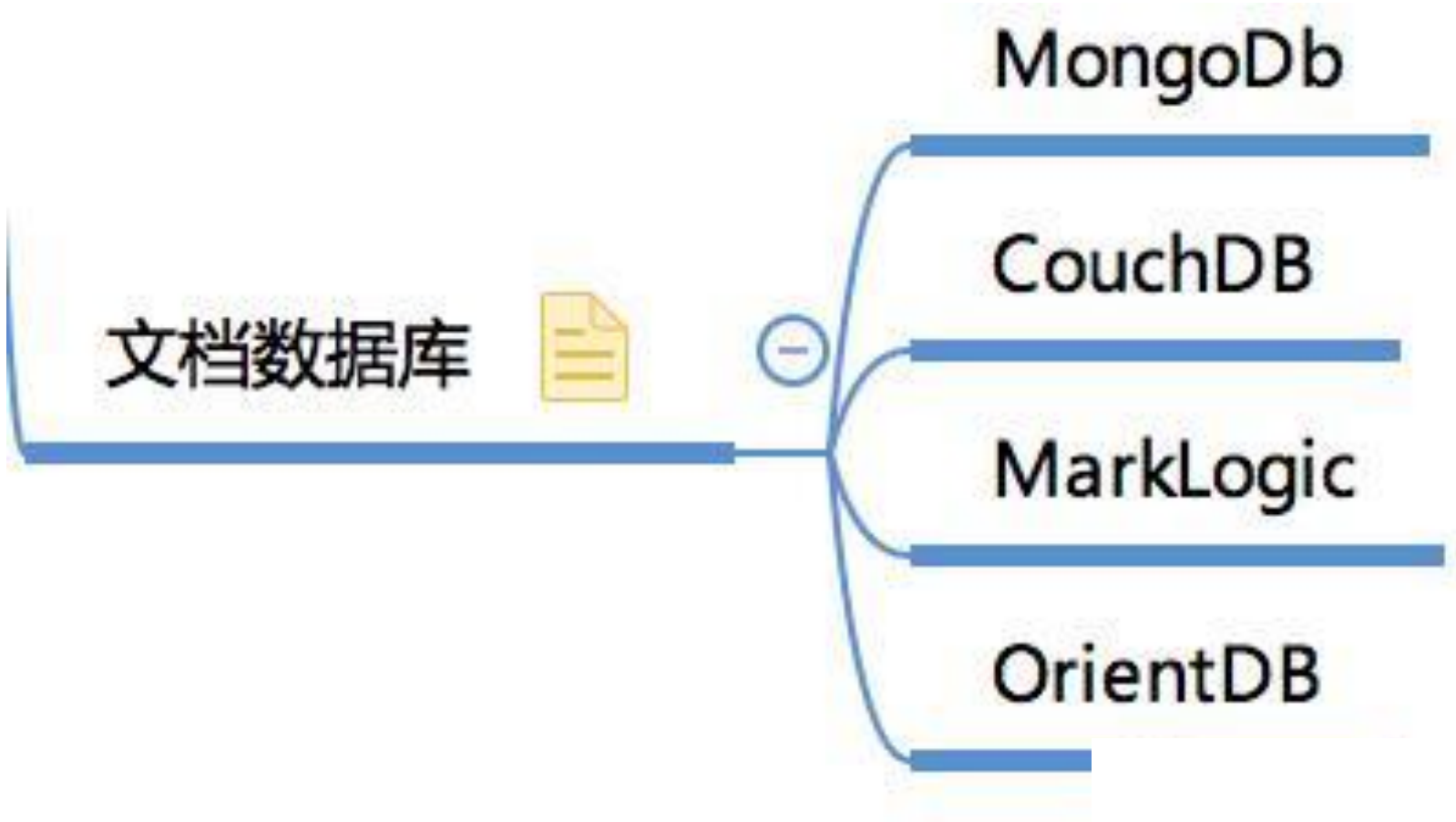
数据存储


图数据库
Titan
Neo4J
ArangoDB
OrientDB
MapGraph
ALLEGROGRAPH

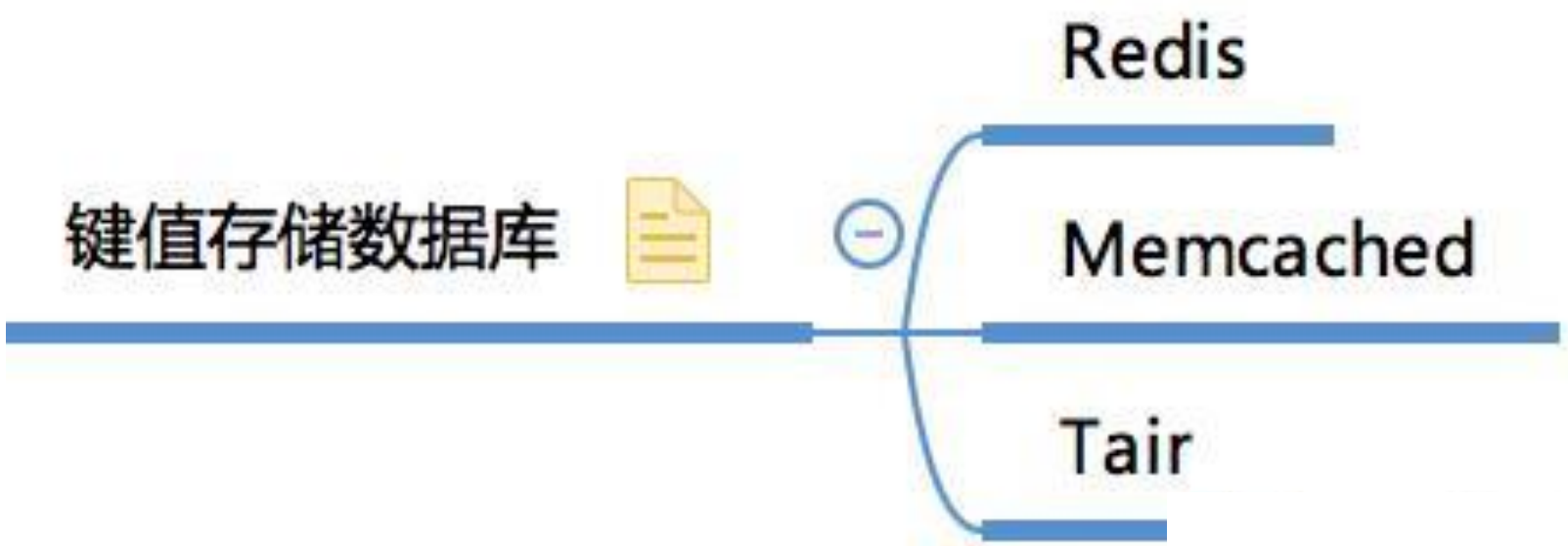
数据存储



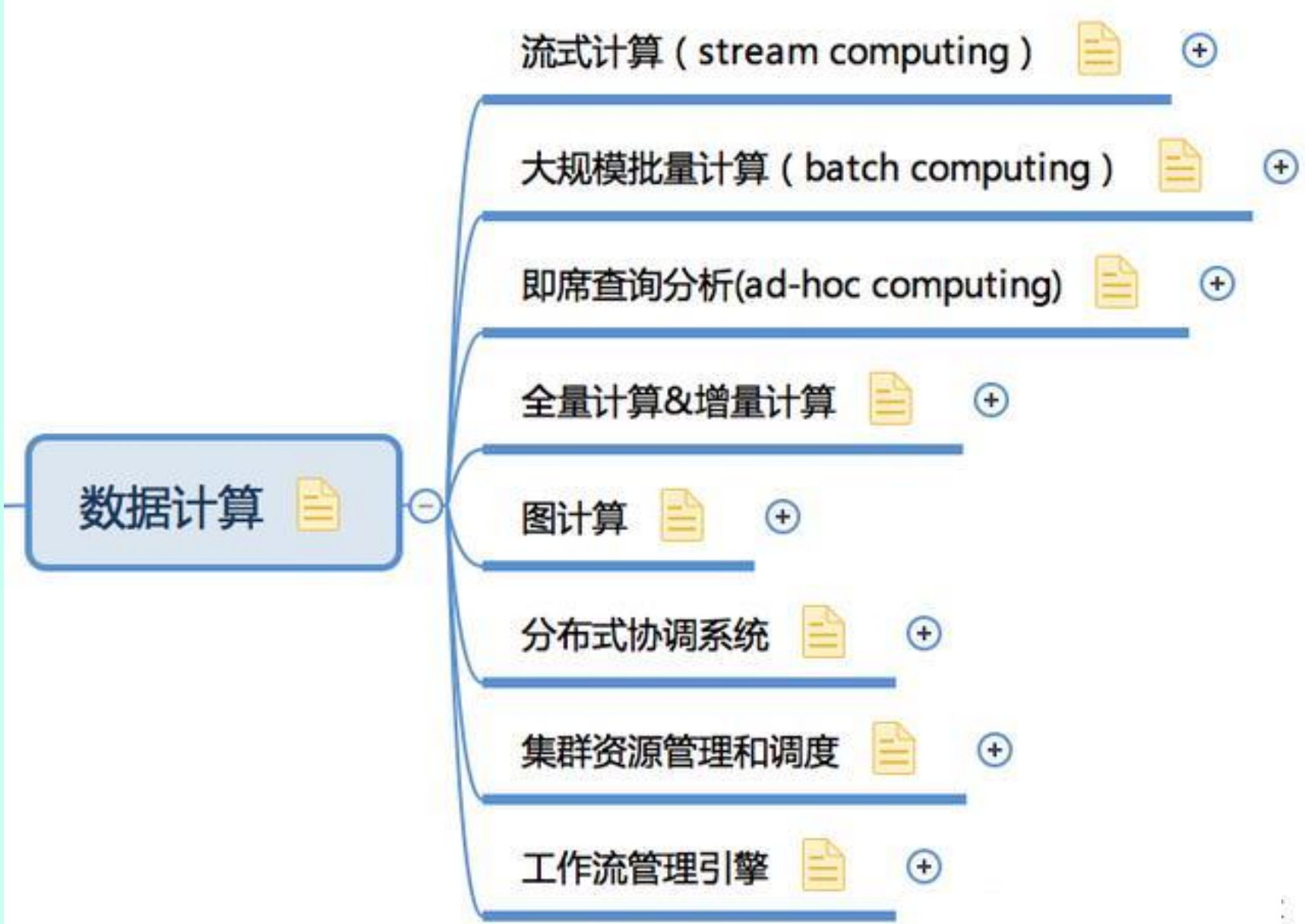
数据存储



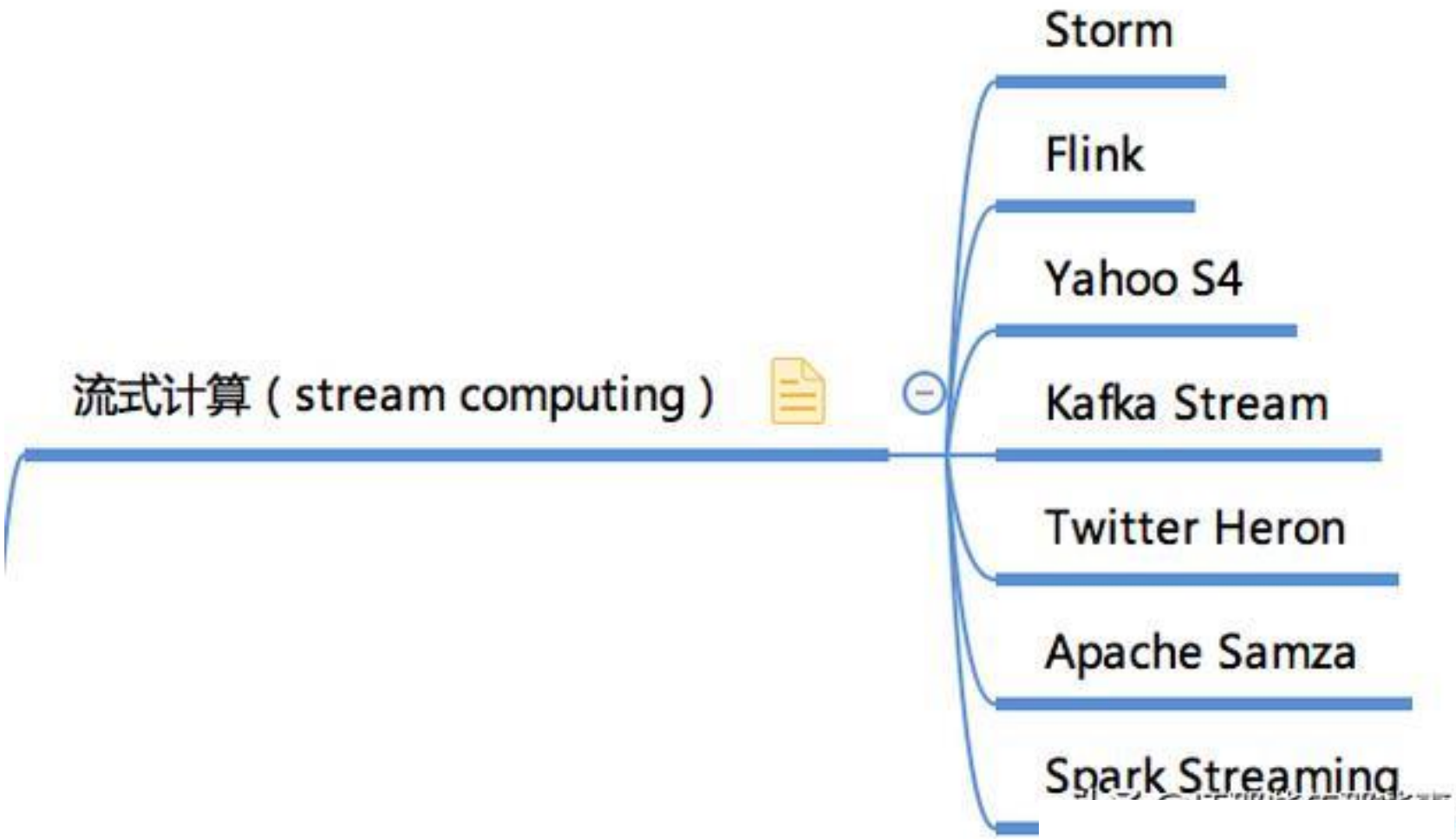
数据存储



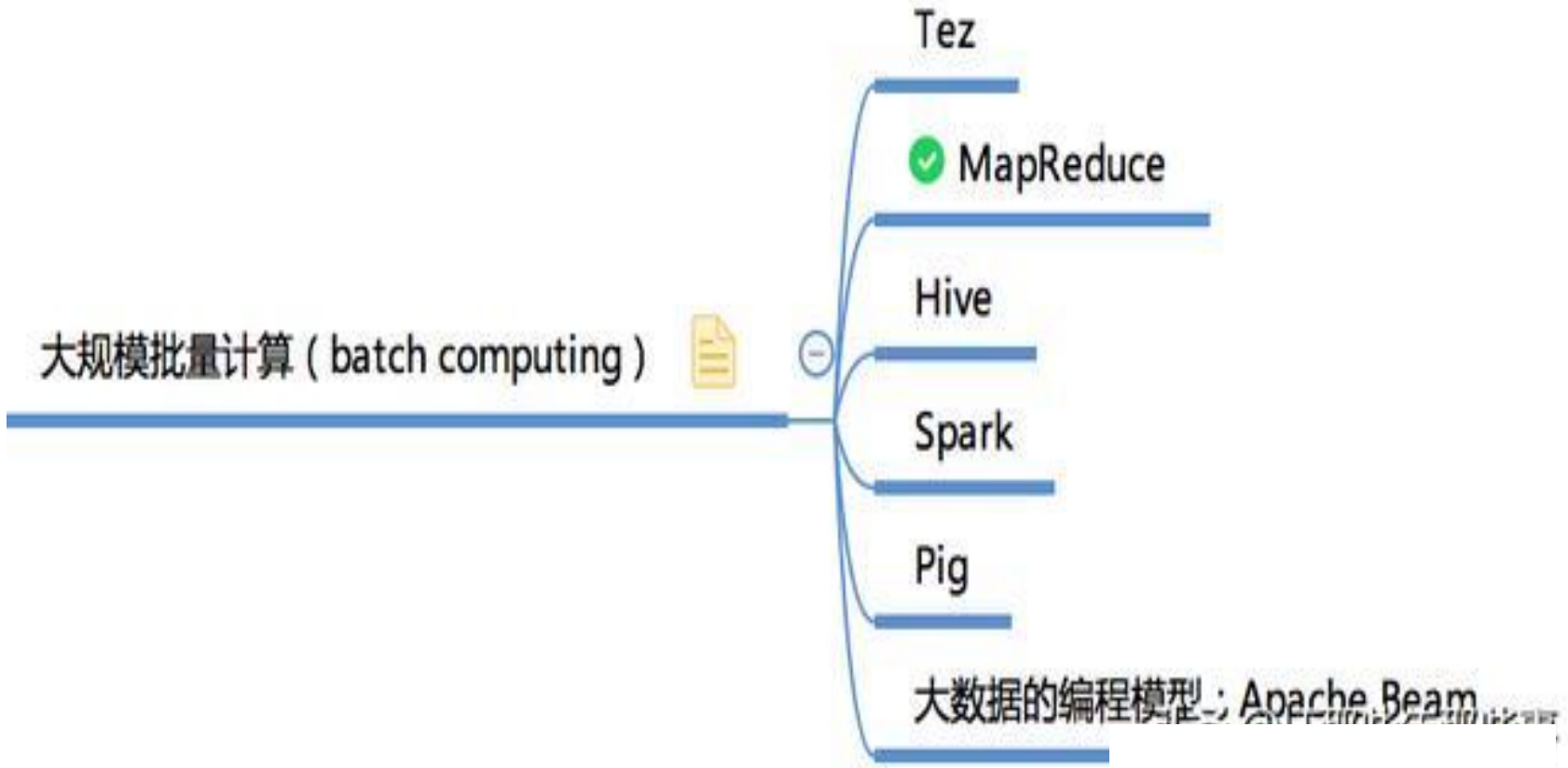
数据组织集成


数据计算
流式计算（stream computing）
大规模批量计算（batch computing）
即席查询分析(ad-hoc computing)
全量计算&增量计算
图计算
分布式协调系统
集群资源管理和调度
工作流管理引擎

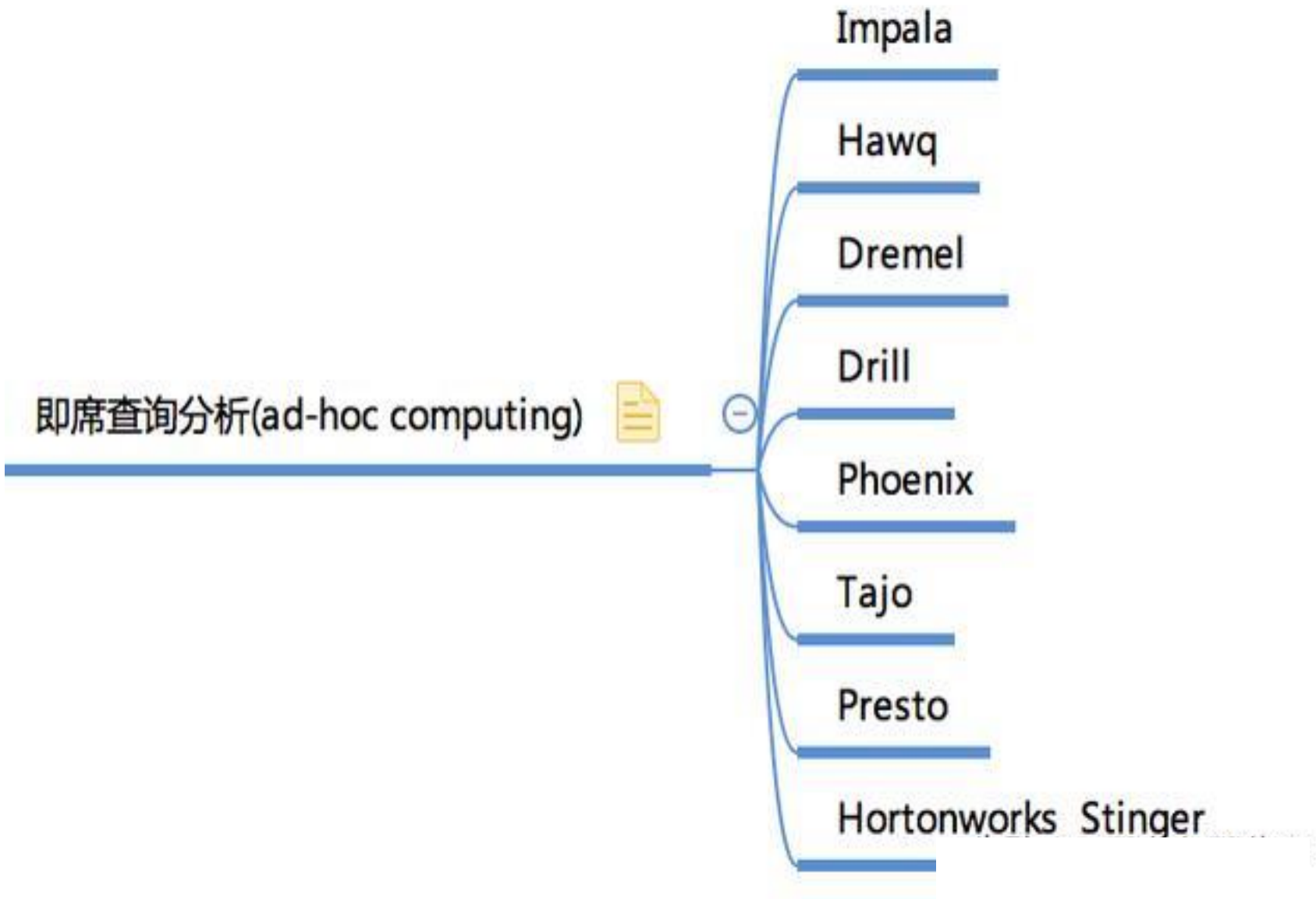
数据计算


流式计算（stream computing）
Storm
Flink
Yahoo S4
Kafka Stream
Twitter Heron
Apache Samza
Spark Streaming

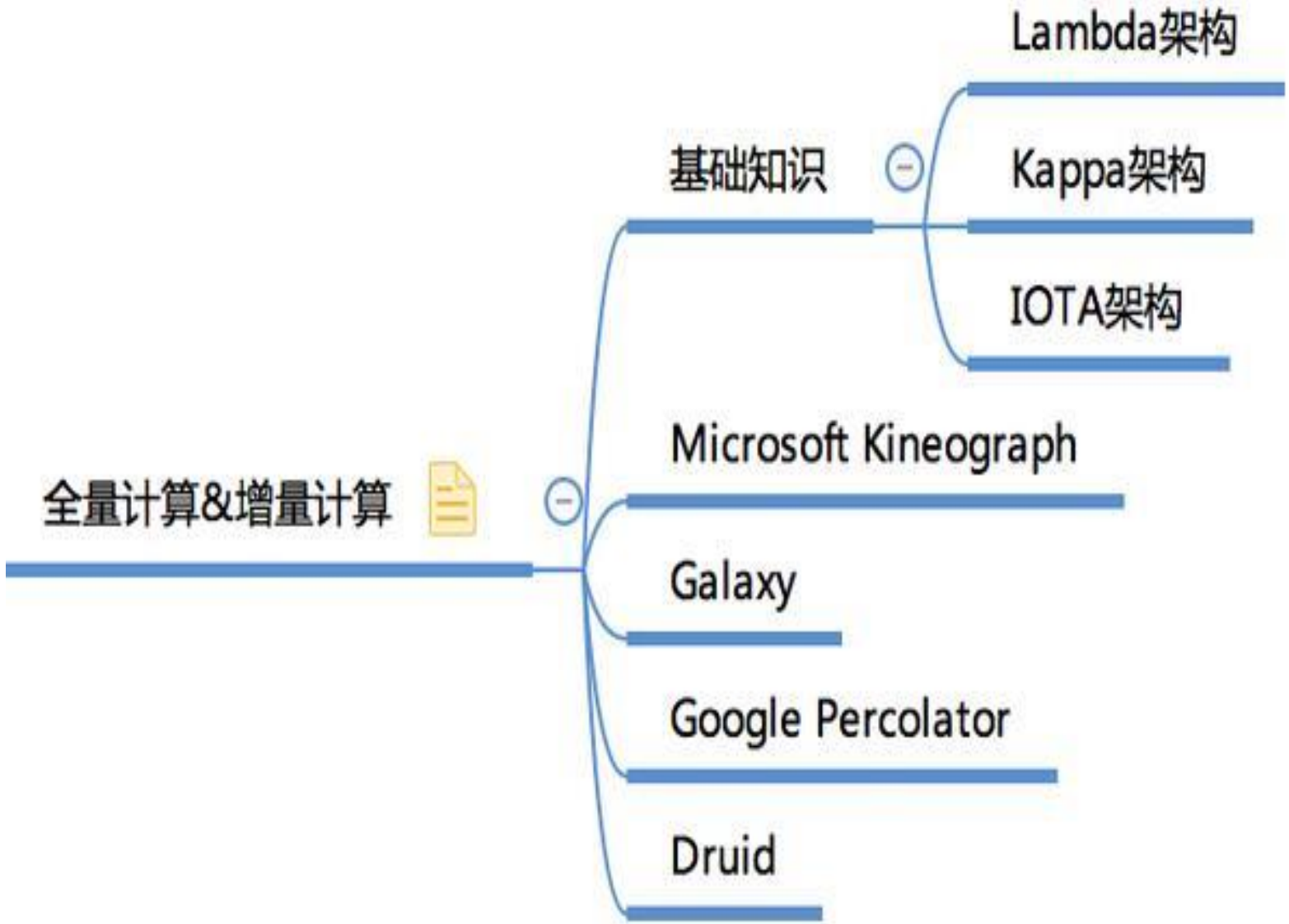
数据计算


大规模批量计算（batch computing）
Tez
MapReduce
Hive
Spark
Pig
大数据的编程模型：Apache Beam

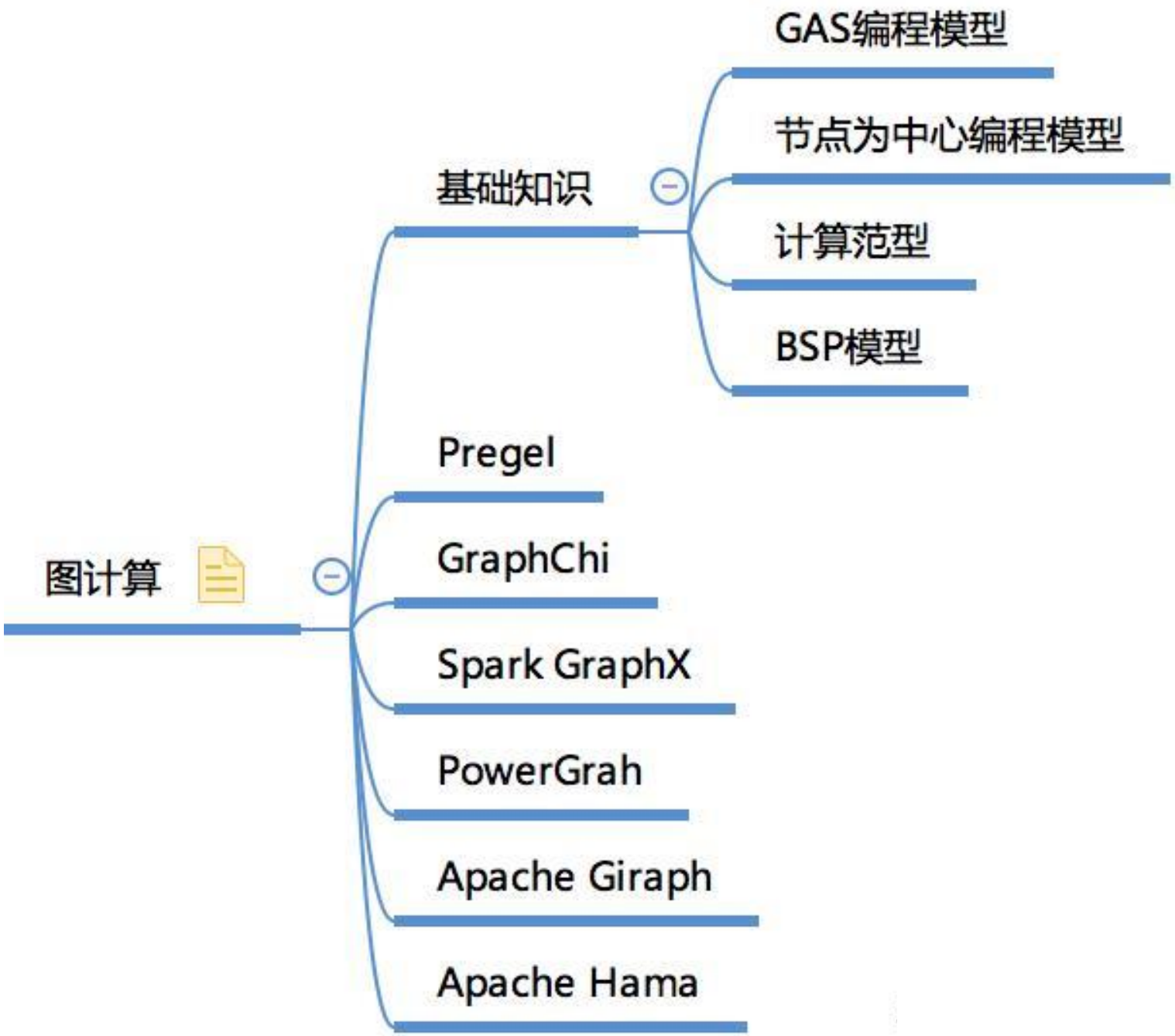
数据计算


即席查询分析(ad-hoc computing)
Impala
Hawq
Dremel
Drill
Phoenix
Tajo
Presto
Hortonworks Stinger

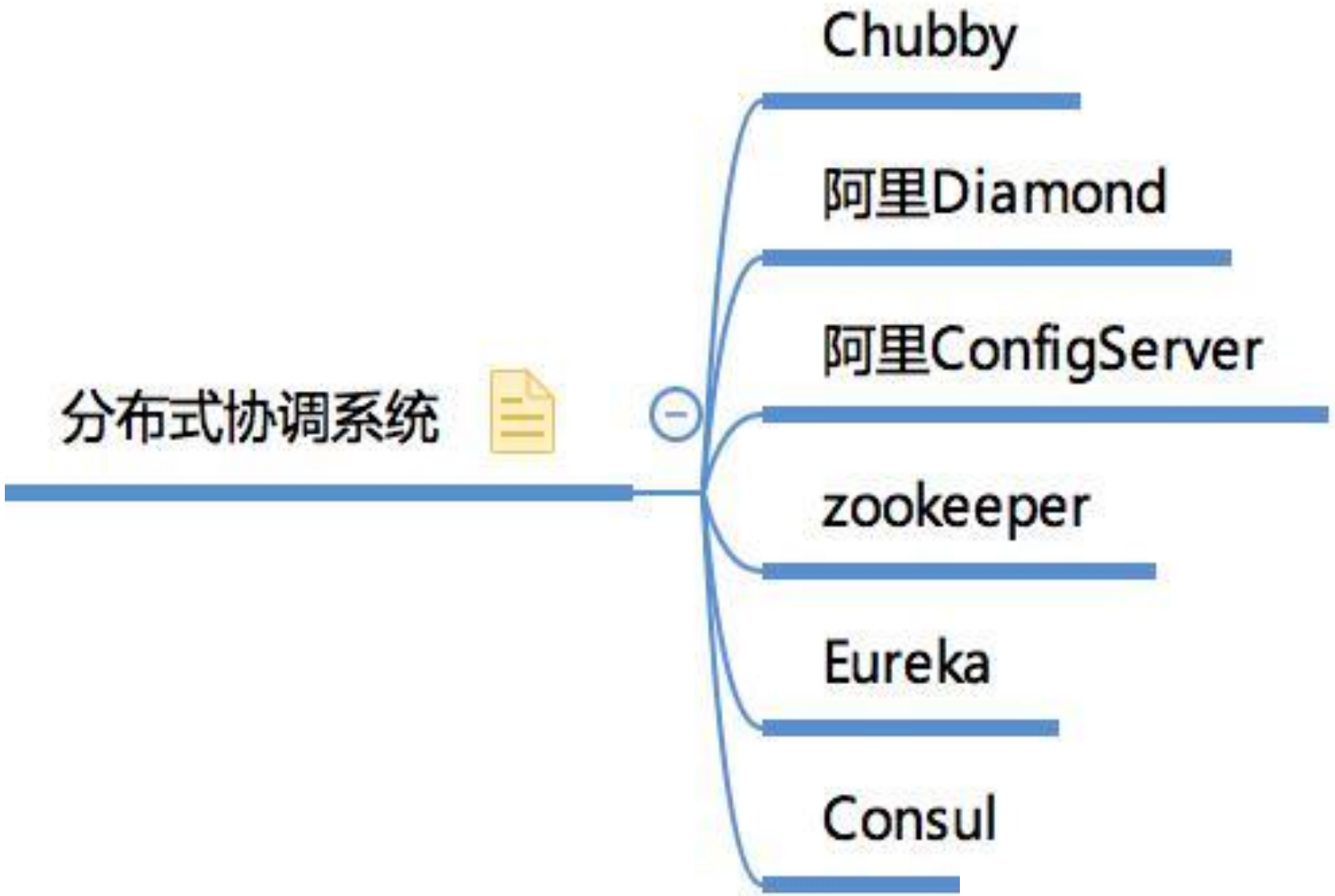
数据计算


全量计算&增量计算
基础知识
Lambda架构
Kappa架构
IOTA架构
Microsoft Kineograph
Galaxy
Google Percolator
Druid

数据计算


图计算
基础知识
GAS编程模型
节点为中心编程模型
计算范型
BSP模型
Pregel
GraphChi
Spark GraphX
PowerGrah
Apache Giraph
Apache Hama

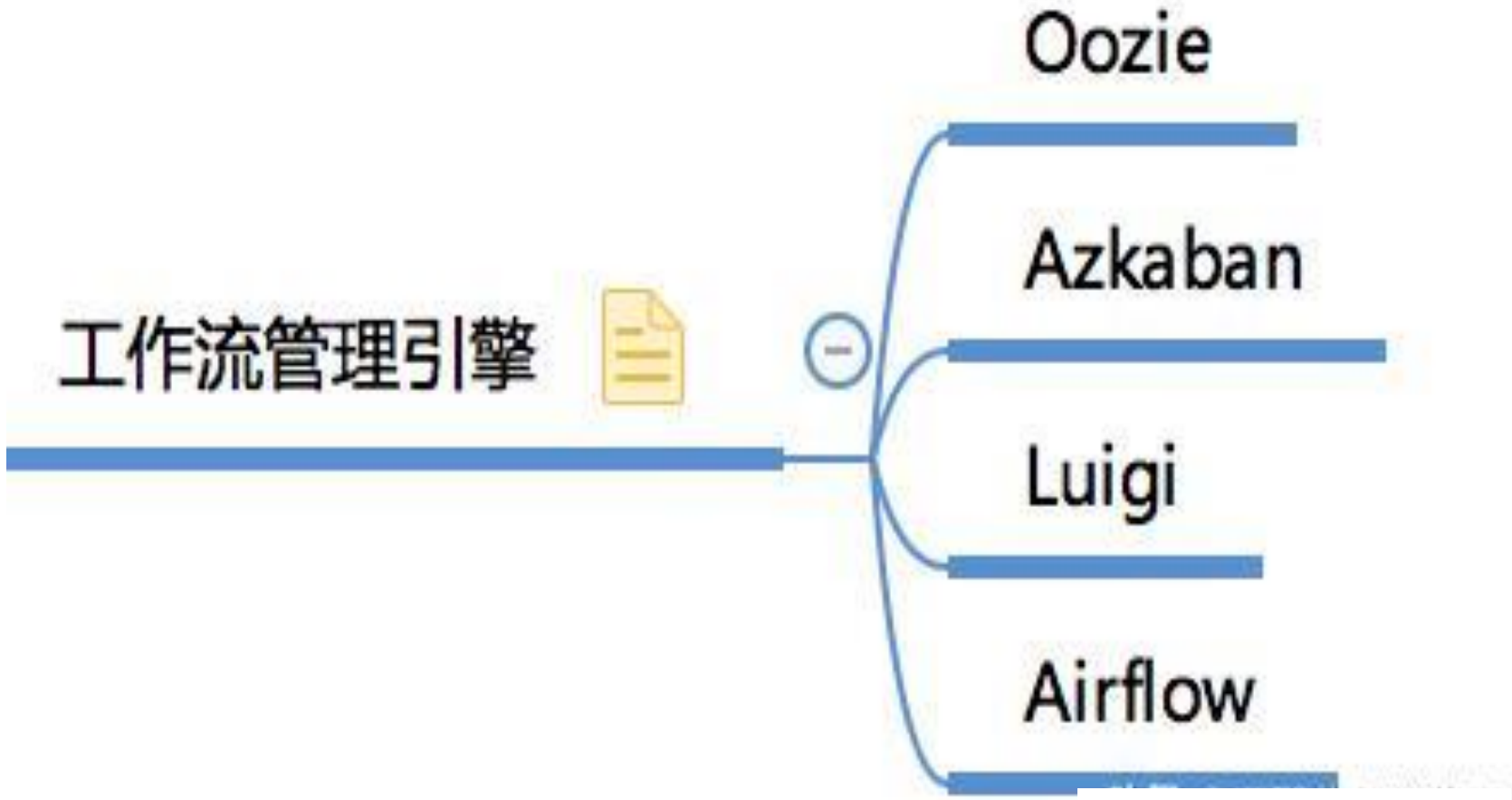
数据组织集成


分布式协调系统
Chubby
阿里Diamond
阿里ConfigServer
zookeeper
Eureka
Consul

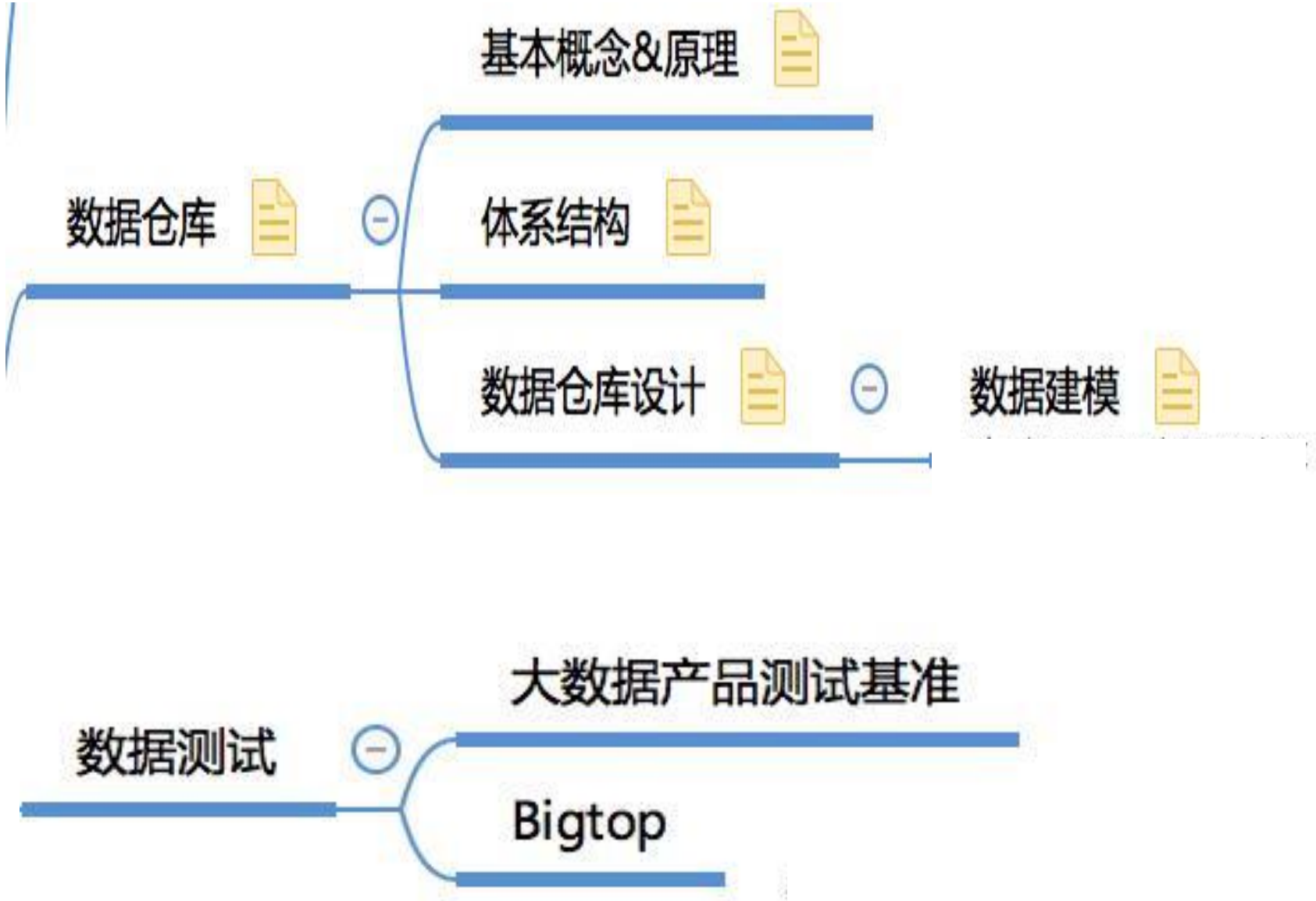
数据组织集成


集群资源管理和调度
管理调度框架
Omega
Brog
Mesos
Corona
Yarn
Torca
管理和监控工具
Ambari
Chukwa

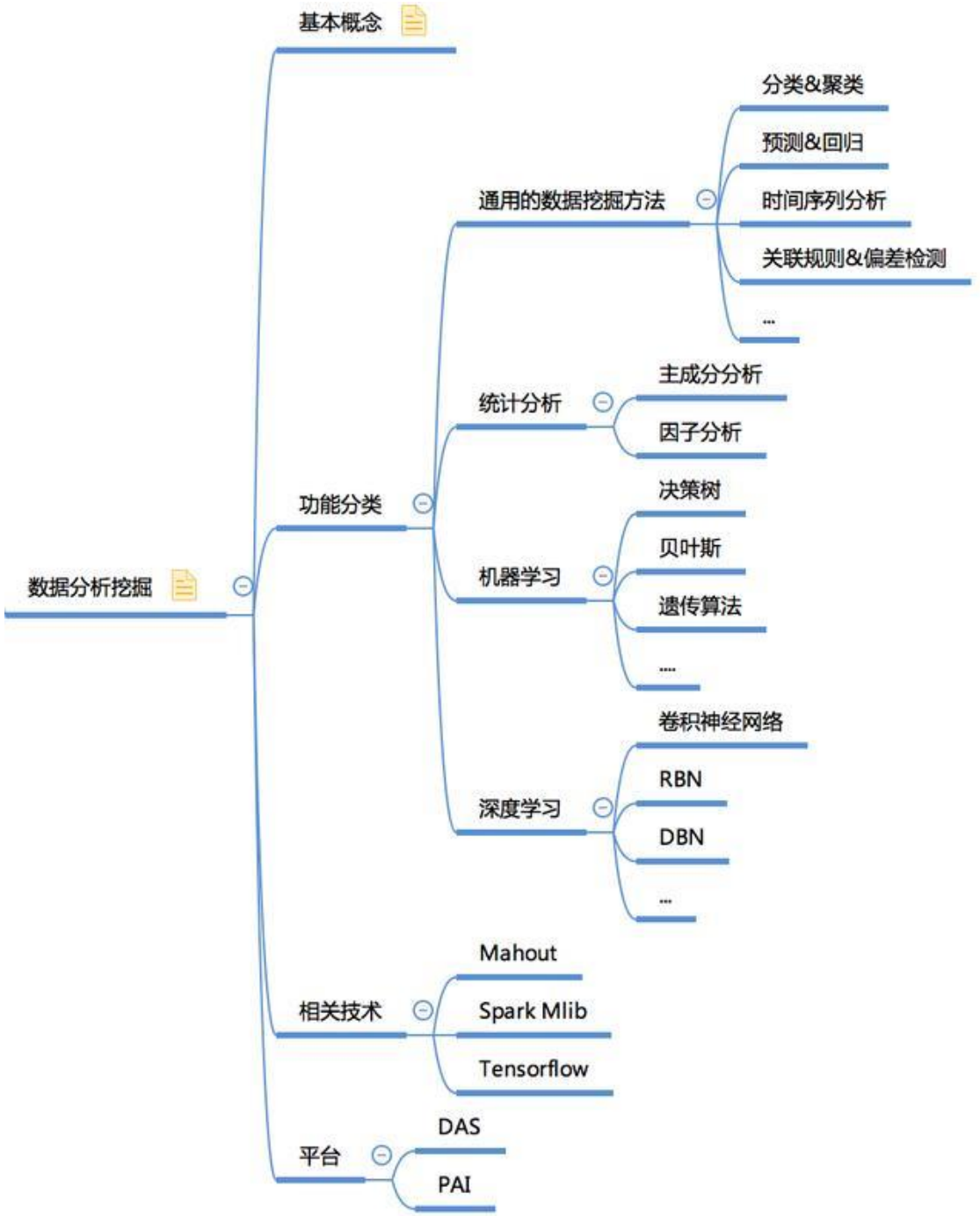
数据组织集成



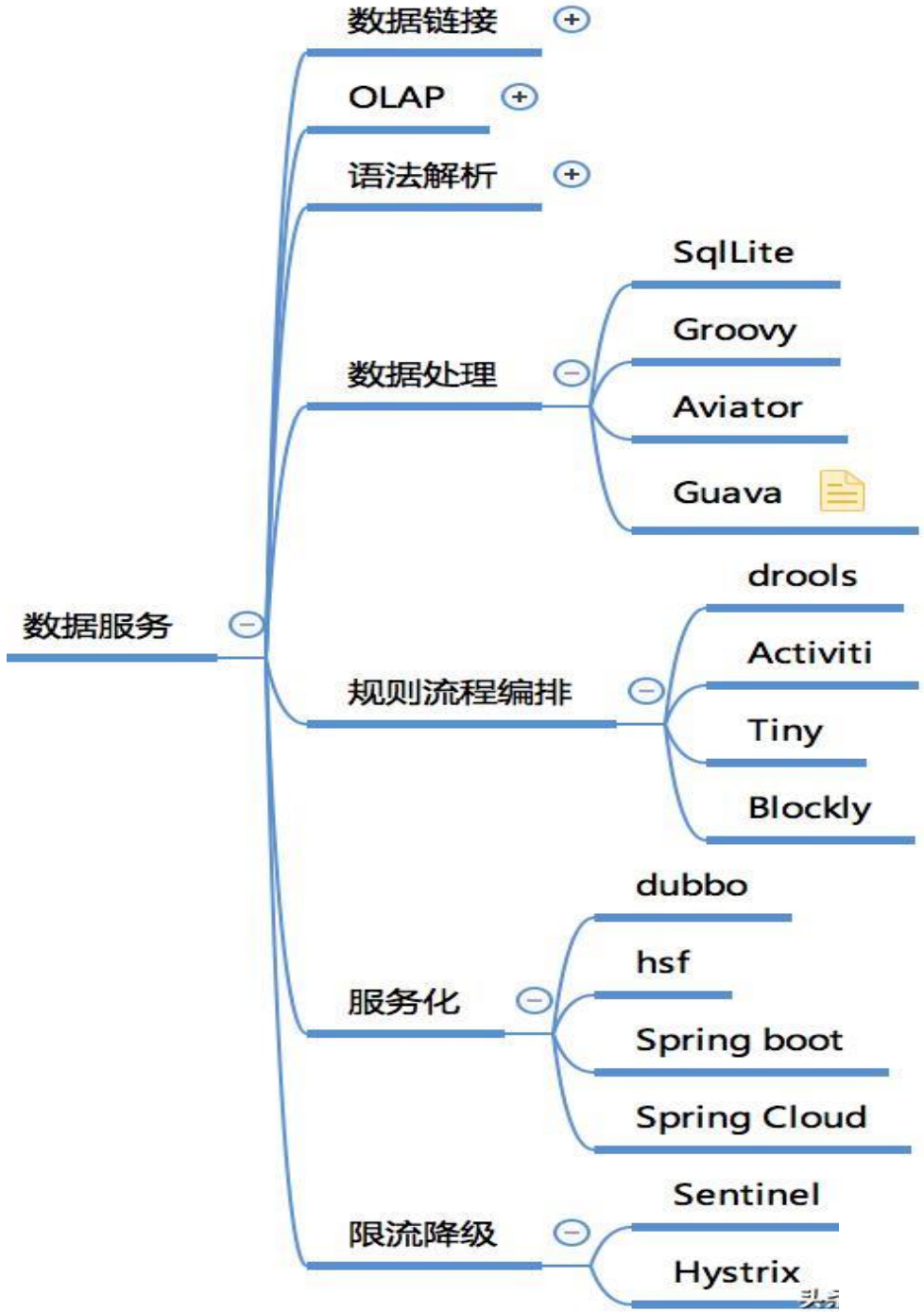
数据组织集成



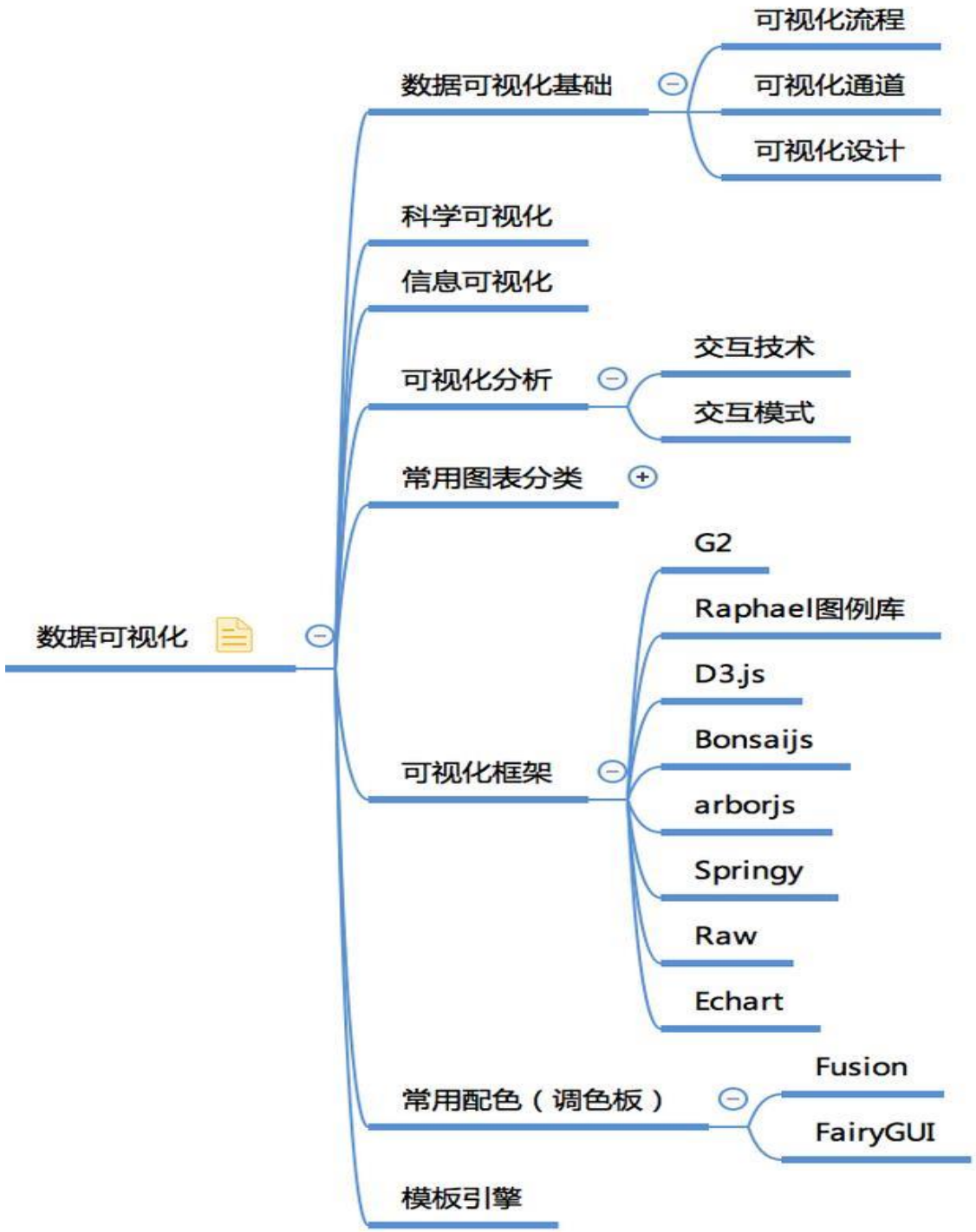
数据应用


数据分析挖掘
基本概念
功能分类
通用的数据挖掘方法
分类&聚类
预测&回归
时间序列分析
关联规则&偏差检测
...
统计分析
主成分分析
因子分析
机器学习
决策树
贝叶斯
遗传算法
....
深度学习
卷积神经网络
RBN
DBN
...
相关技术
Mahout
Spark Mlib
Tensorflow
平台
DAS
PAI

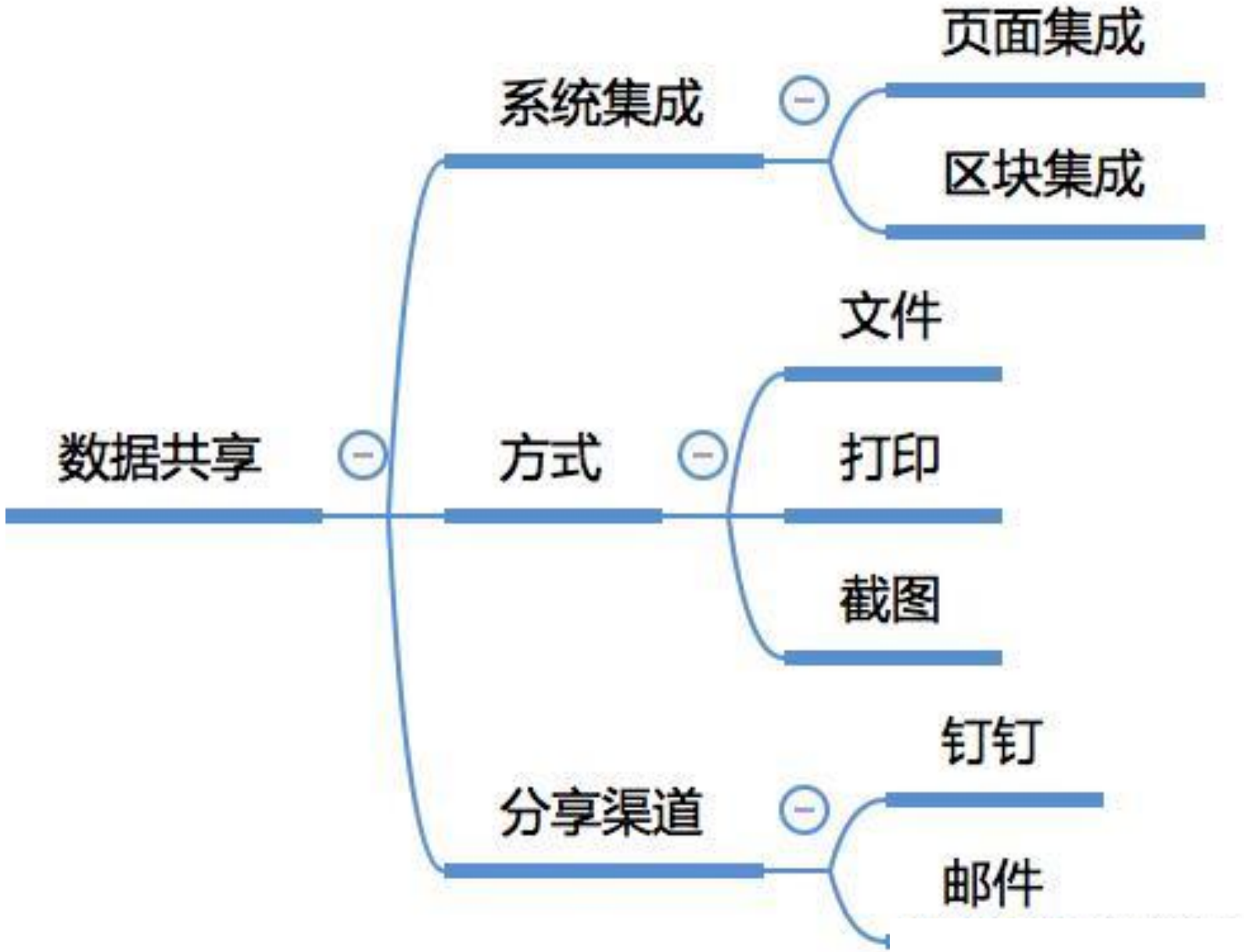
数据应用


数据服务
数据链接
OLAP
语法解析
数据处理
SqlLite
Groovy
Aviator
Guava
规则流程编排
drools
Activiti
Tiny
Blockly
服务化
dubbo
hsf
Spring boot
Spring Cloud
限流降级
Sentinel
Hystrix

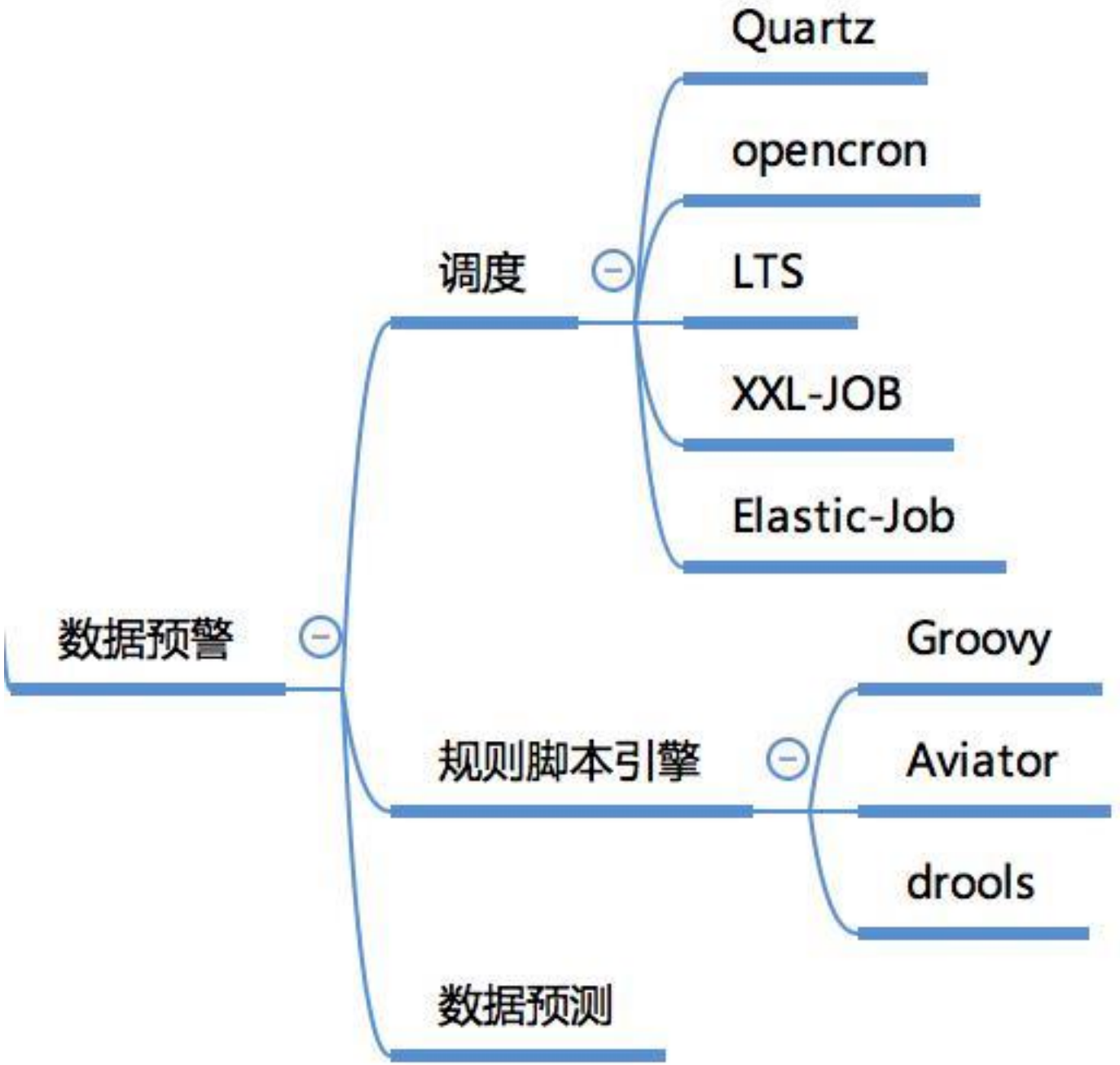
数据应用


数据可视化
数据可视化基础
可视化流程
可视化通道
可视化设计
科学可视化
信息可视化
可视化分析
交互技术
交互模式
常用图表分类
可视化框架
G2
Raphael图例库
D3.js
Bonsaijs
arborjs
Springy
Raw
Echart
常用配色（调色板）
Fusion
FairyGUI
模板引擎

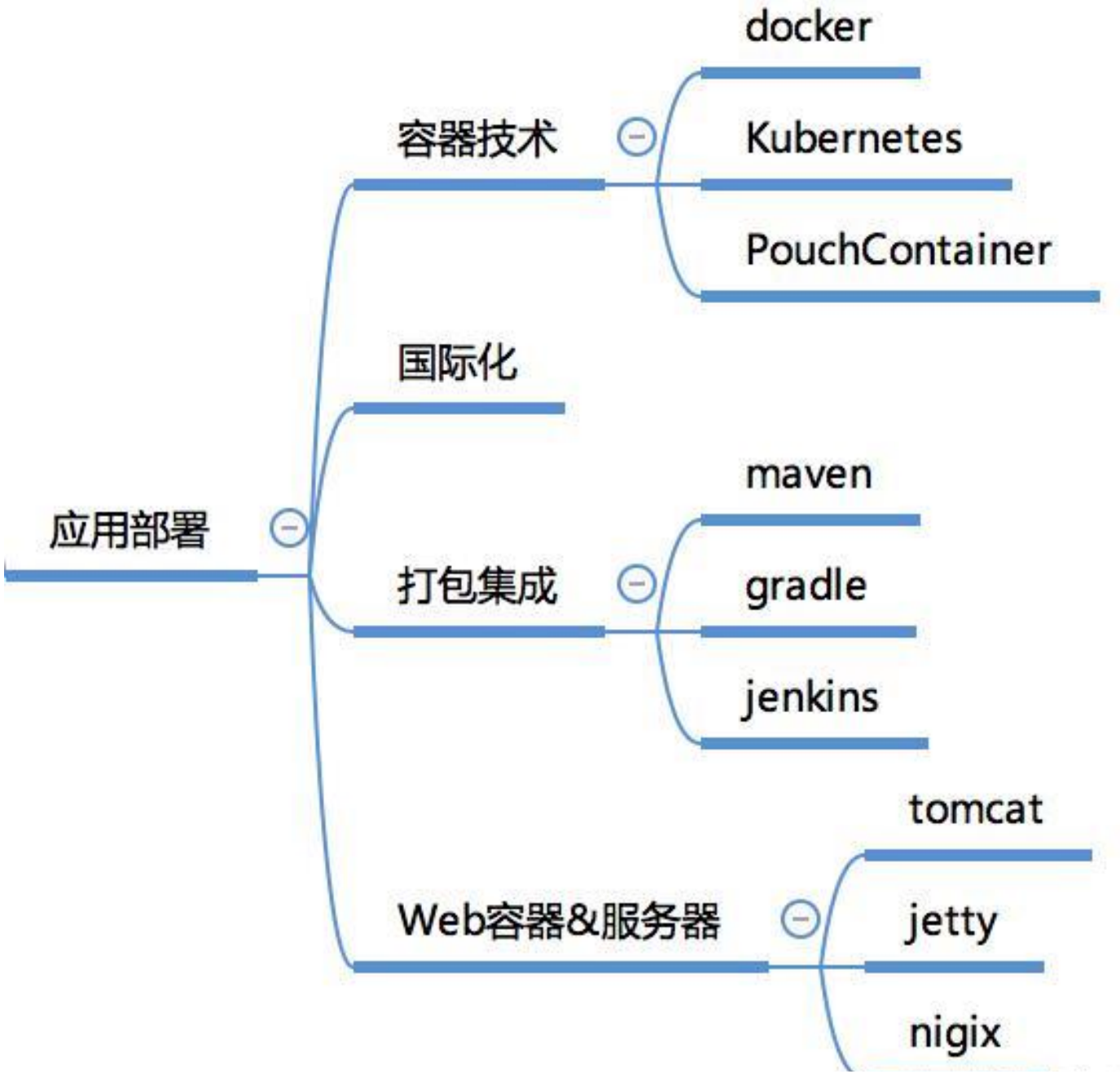
数据应用



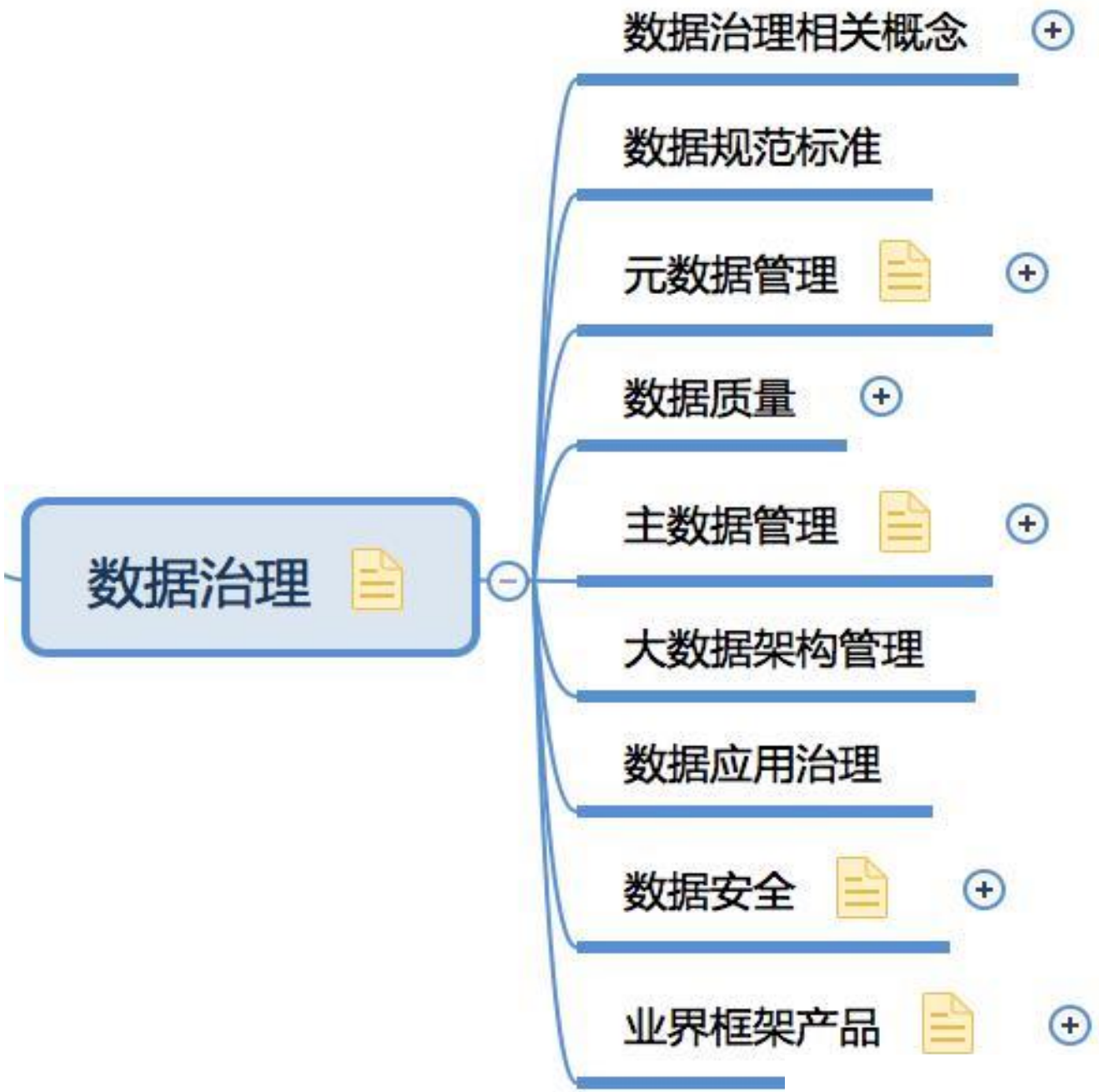
数据应用


数据预警
调度
Quartz
opencron
LTS
XXL-JOB
Elastic-Job
规则脚本引擎
Groovy
Aviator
drools
数据预测

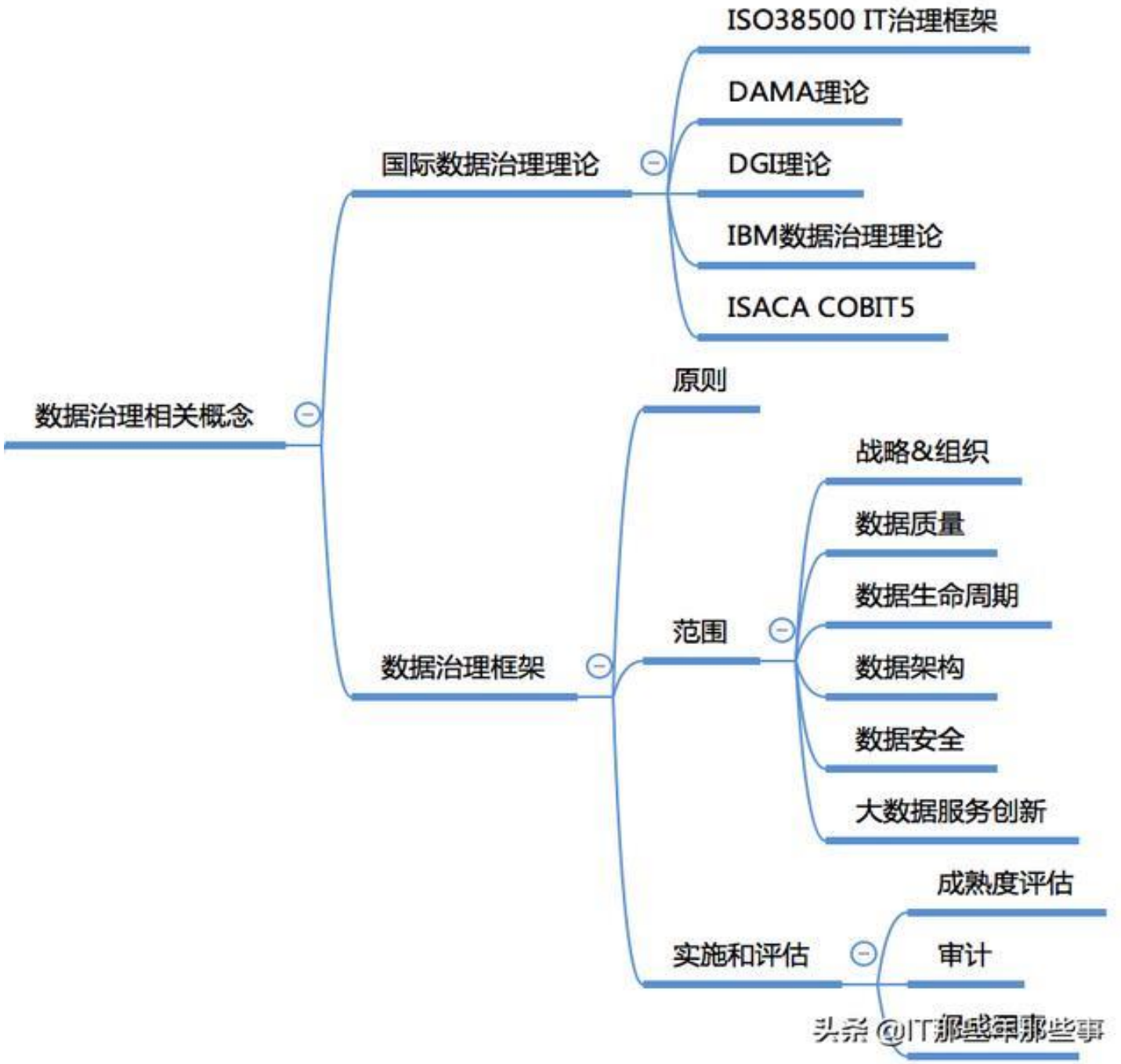
数据应用


应用部署
容器技术
docker
Kubernetes
PouchContainer
国际化
打包集成
maven
gradle
jenkins
Web容器&服务器
tomcat
jetty
nigix

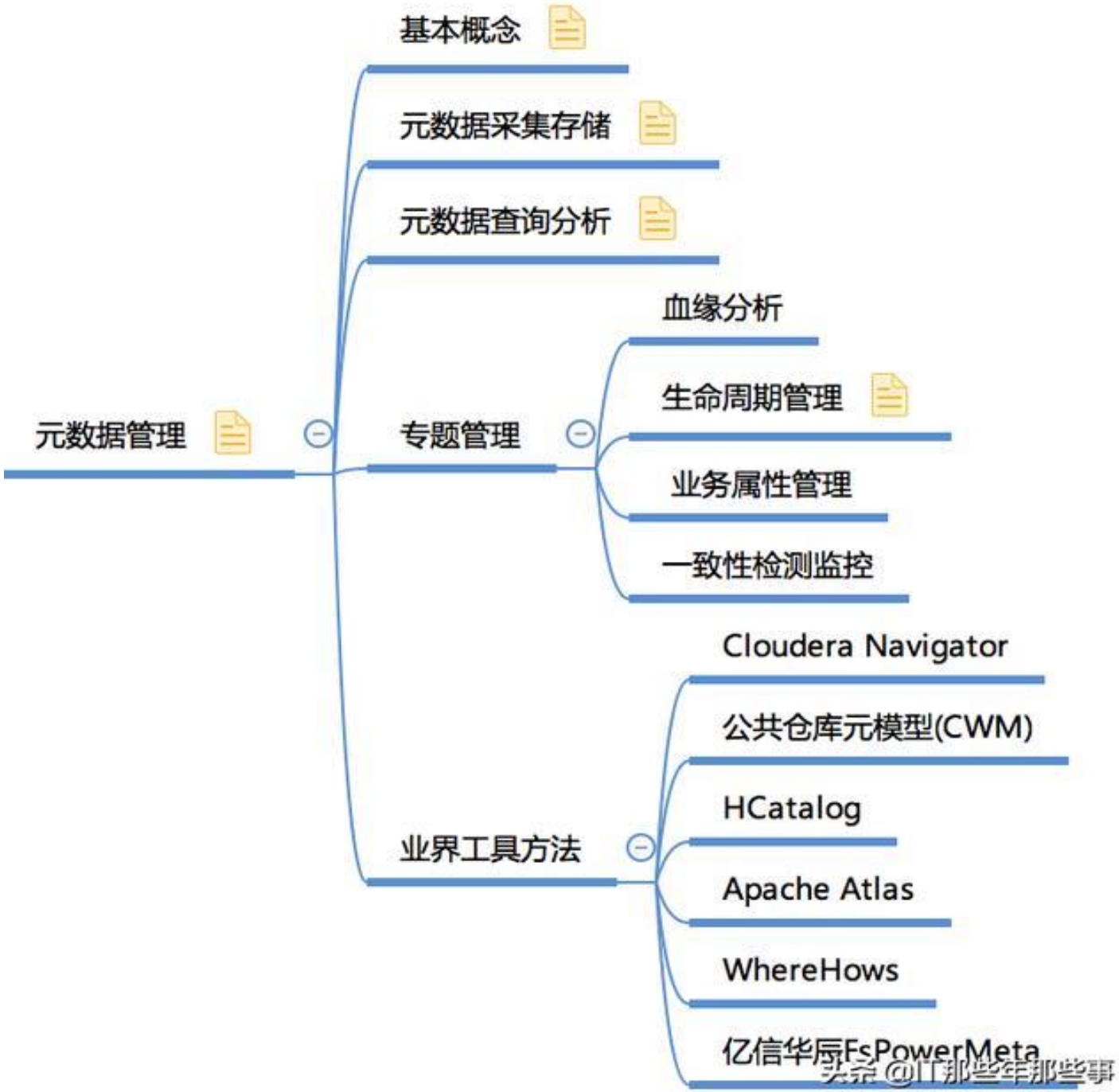
数据治理


数据治理
数据治理相关概念
数据规范标准
元数据管理
数据质量
主数据管理
大数据架构管理
数据应用治理
数据安全
业界框架产品

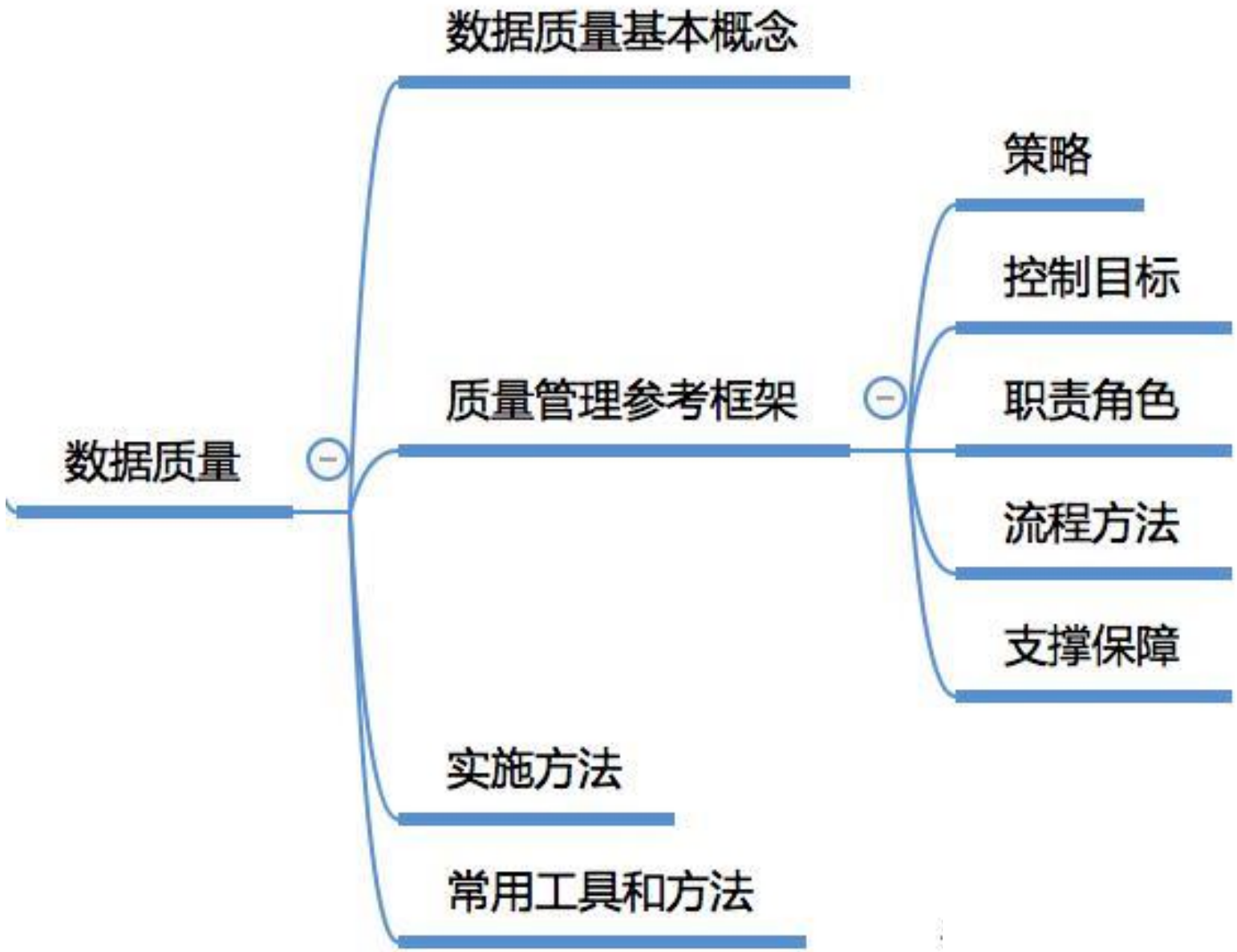
数据治理



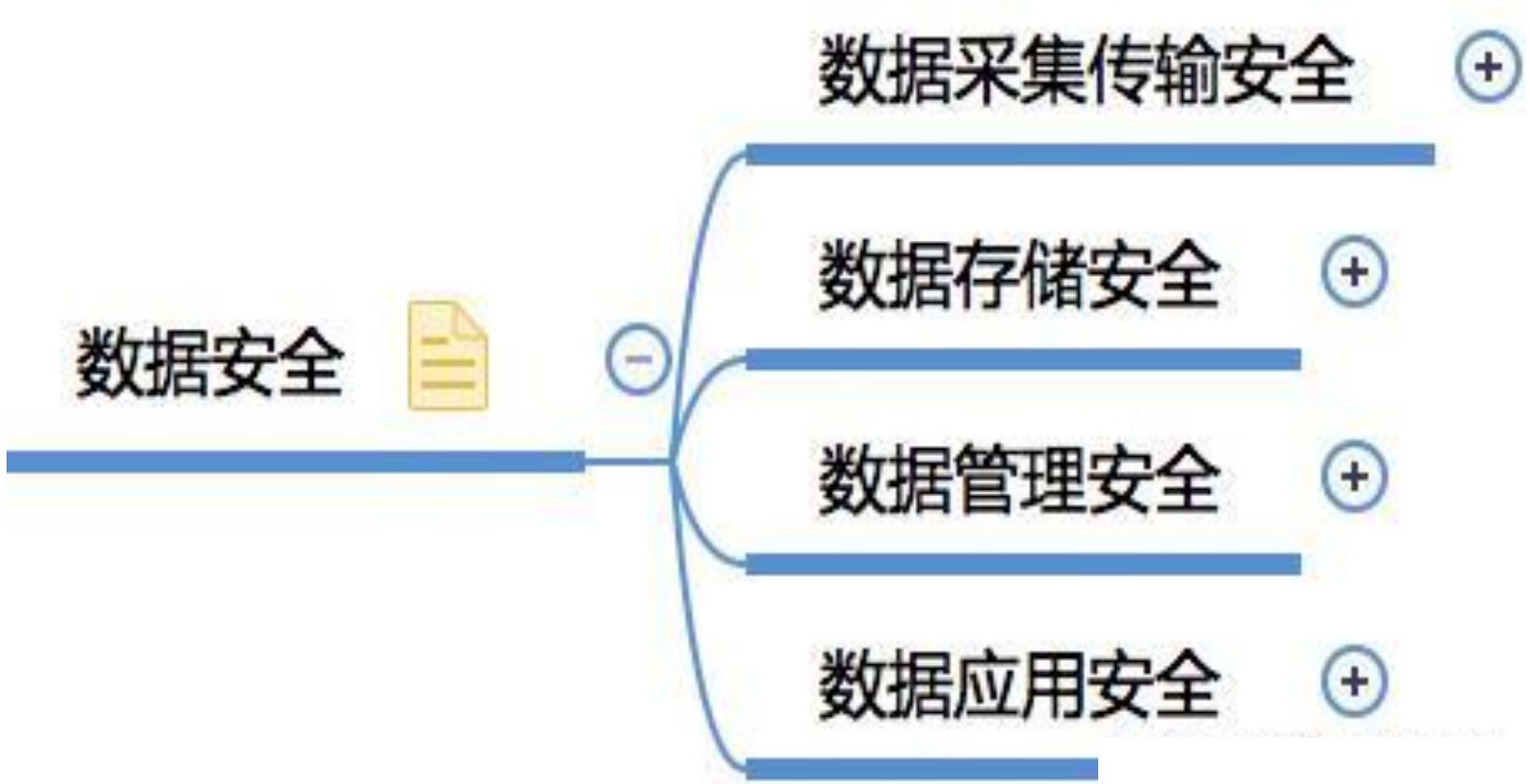
数据治理



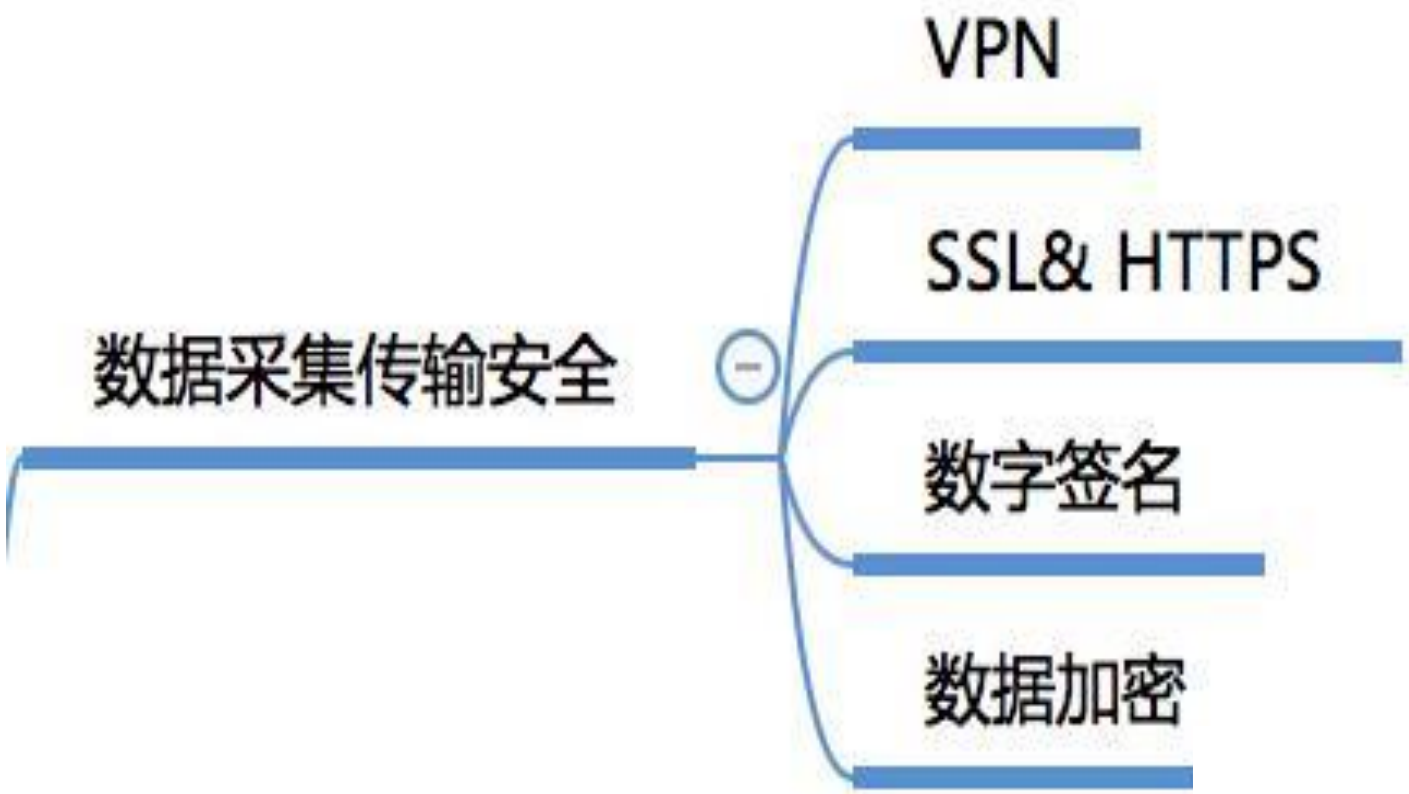
数据治理



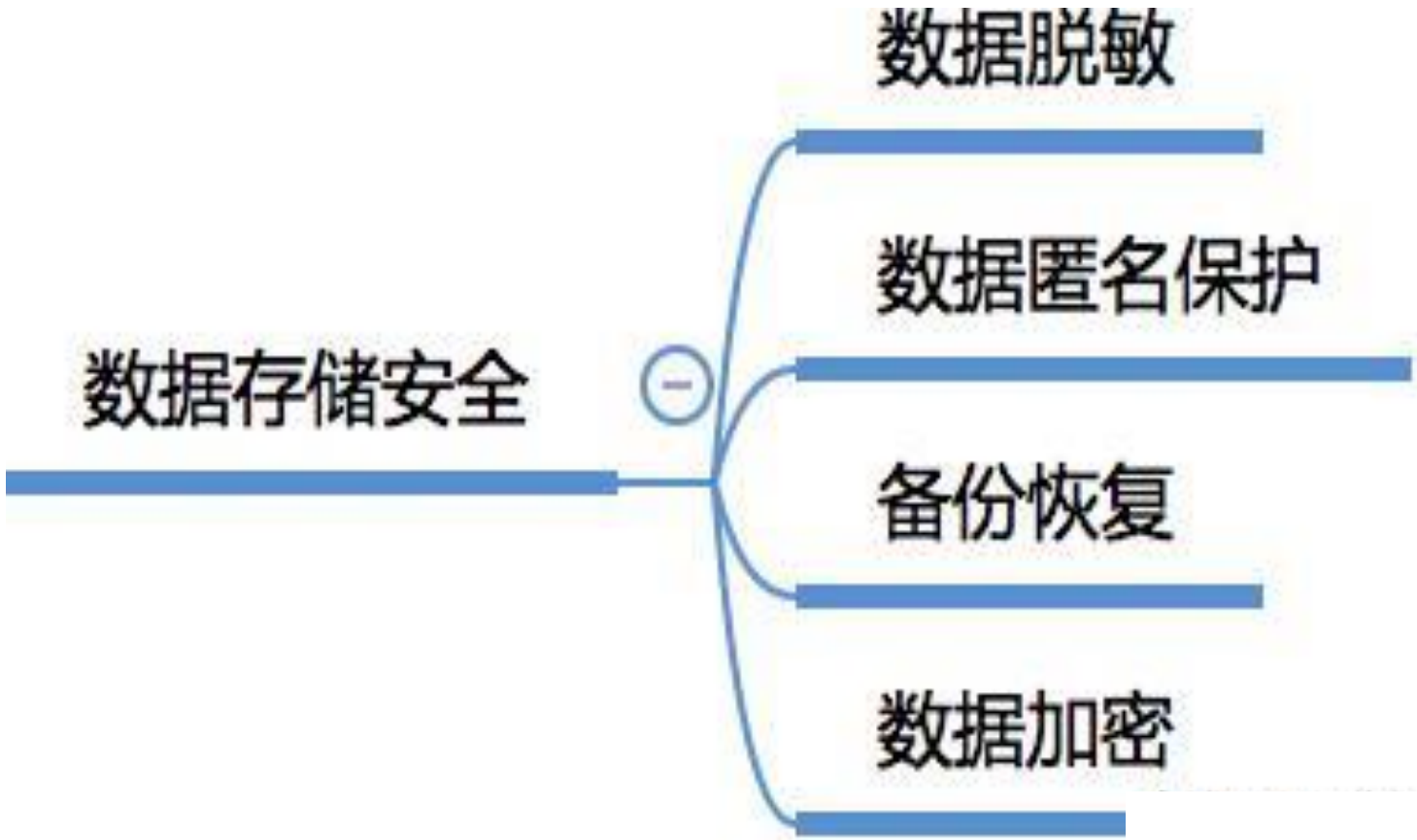
数据治理



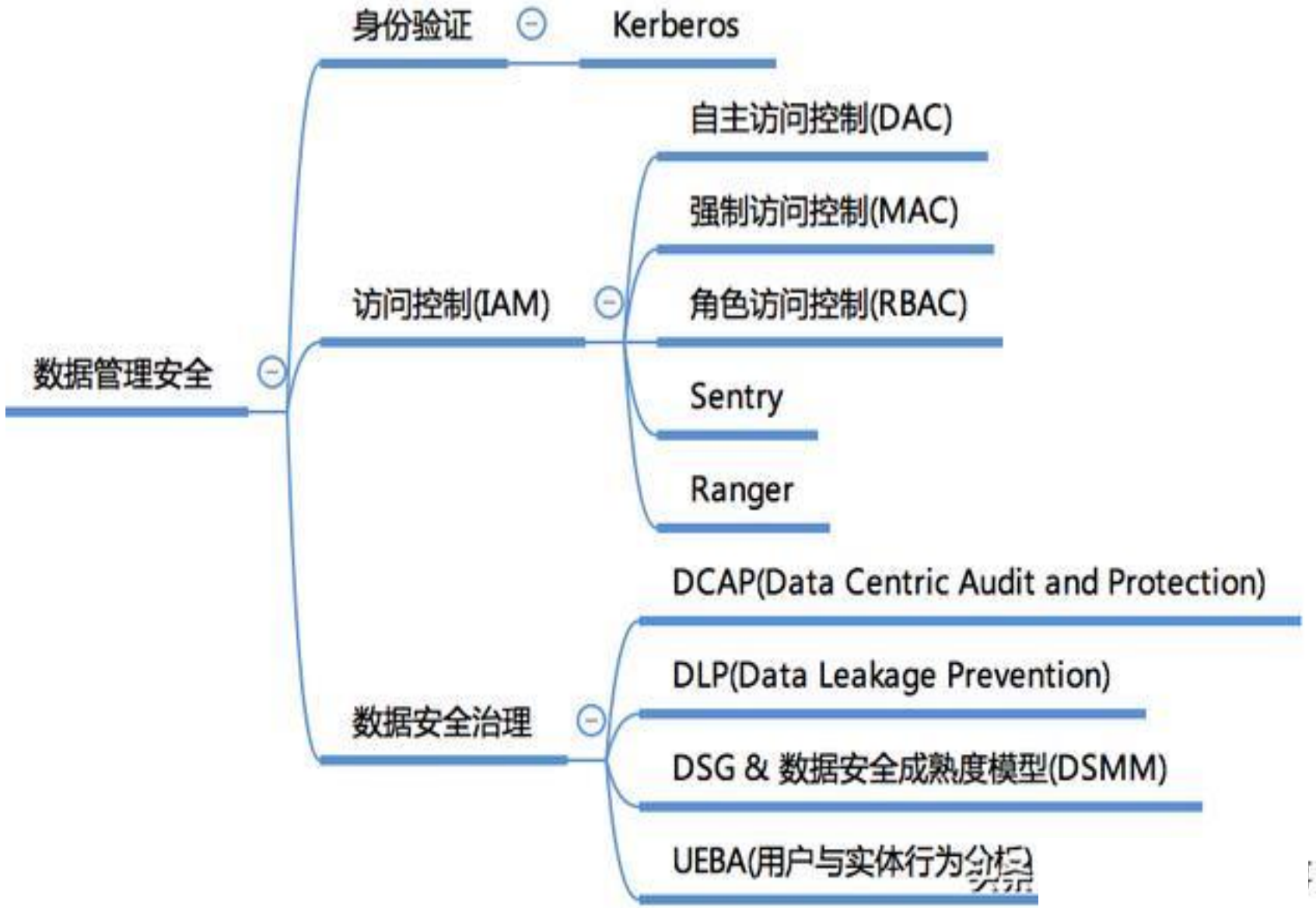
数据安全



数据安全



数据安全


数据管理安全
身份验证
Kerberos
访问控制(IAM)
自主访问控制(DAC)
强制访问控制(MAC)
角色访问控制(RBAC)
Sentry
Ranger
数据安全治理
DCAP(Data Centric Audit and Protection)
DLP(Data Leakage Prevention)
DSG & 数据安全成熟度模型(DSMM)
UEBA(用户与实体行为分析)

数据安全