北京師範大學
BEIJING NORMAL UNIVERSITY

数据价值链路实践 第二讲

数据处理

主讲人：蓝昕

2021.07.31 ~ 08.28

每周六 晚7:00

大纲

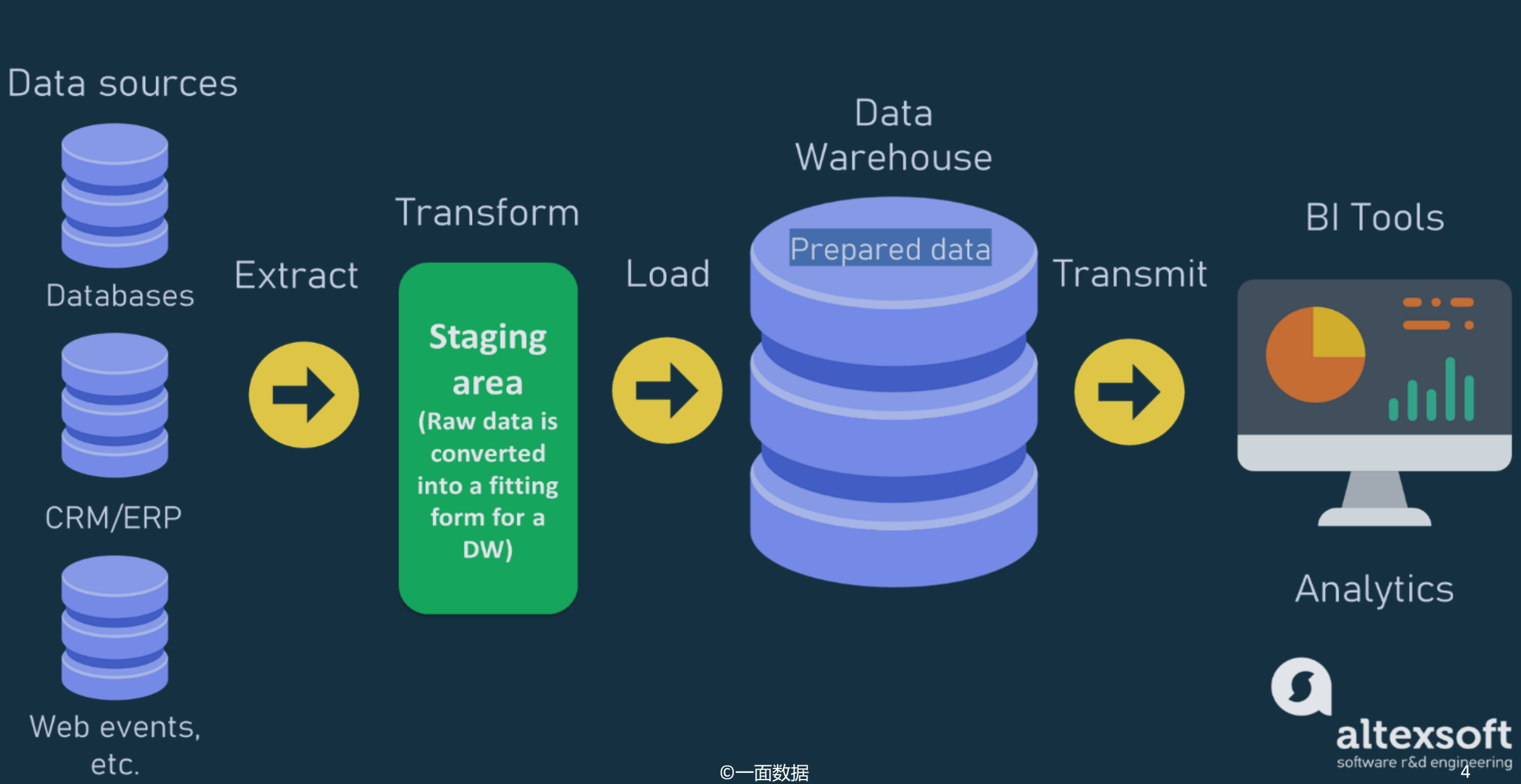
1. 业务场景抽象
2. DAG与任务调度
3. 数据管道

北京師範大學
BEIJING NORMAL UNIVERSITY

业务场景抽象

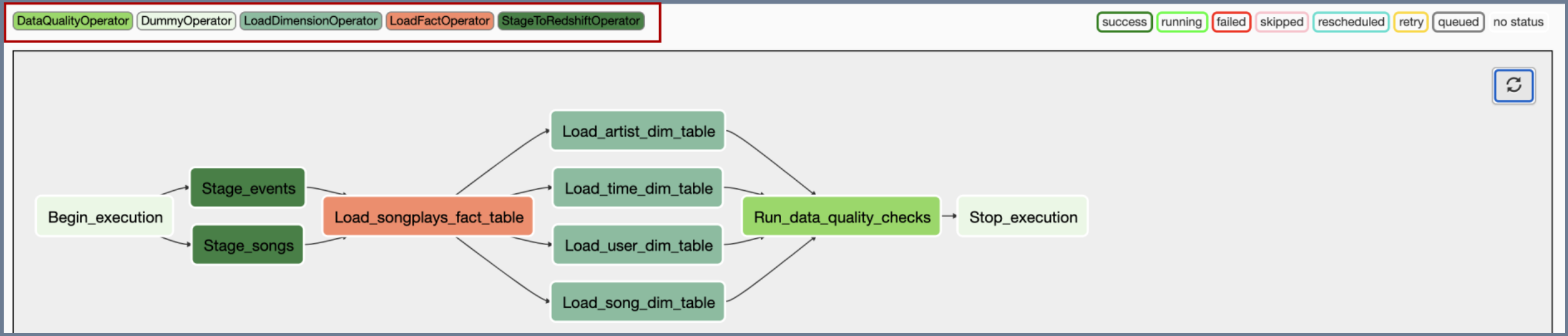
理解业务，串联知识

数据流转链路


ETL: EXTRACT, TRANSFORM, LOAD
Data sources
Databases
CRM/ERP
Web events, etc.
Extract
Transform
Staging area
(Raw data is converted into a fitting form for a DW)
Load
Data Warehouse
Prepared data
Transmit
BI Tools
Analytics
altexsoft
software r&d engineering

核心概念

Operators:



组成

想要数据持续流转，需要有引擎和管道共同配合

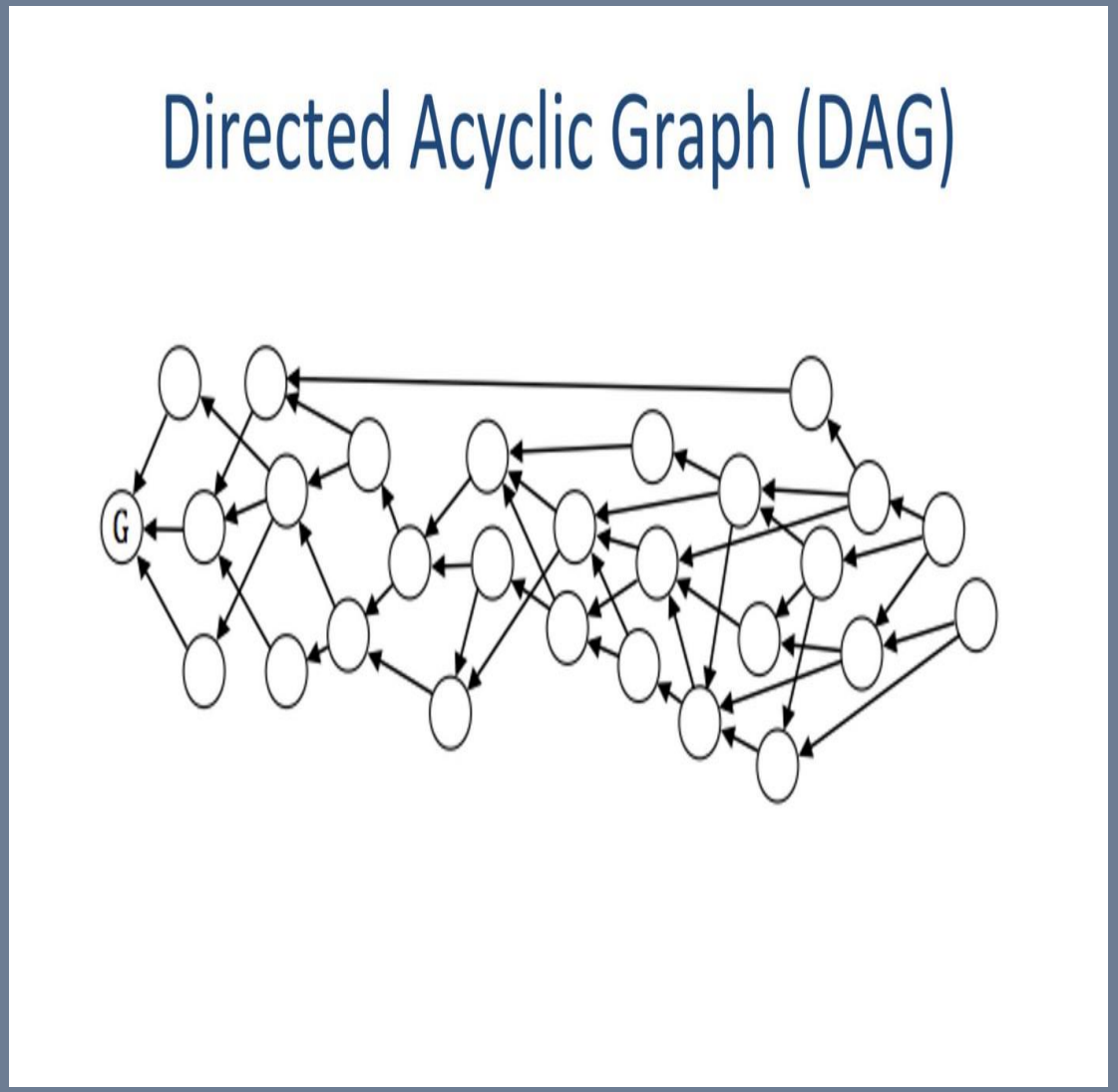
- Job – 数据管道：构成原子
- DAG – 任务管理：串联Job
- Scheduler - 定时器：定时启动

北京師範大學
BEIJING NORMAL UNIVERSITY

DAG与任务调度

数据流动的引擎

DAG



数据处理流程是一个DAG，有方向，而且不可形成自循环的闭环

- 处理后的数据不能写入源数据表
- 下游数据是经过层层处理得到的
- 一个中间结果可供若干下游数据使用

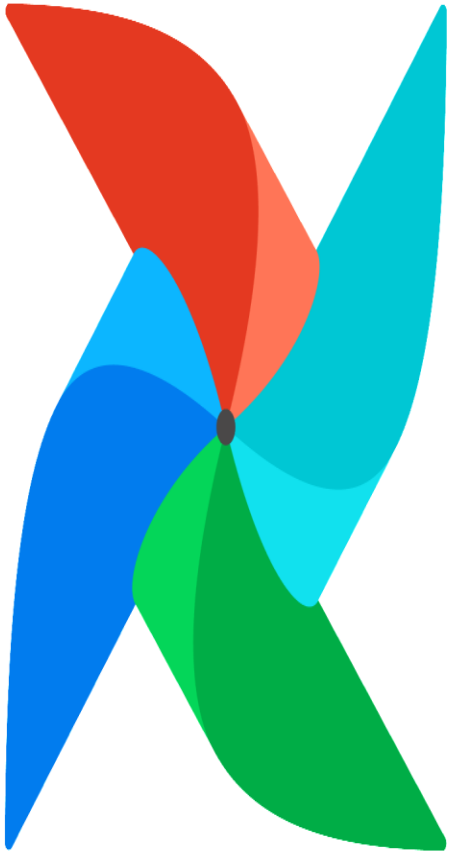
任务调度系统

需求

- 管理任务依赖，检测闭环
- 定时启动
- 监控任务状态，出错报警
- 其它高级需求如SLA等

流行的开源调度系统

- Airflow，Python，Airbnb开发（推荐）
- DolphinScheduler，Java，易观开发
- Azkaban，Java， Linkedin开发



数据管道

数据流动的纽带

数据开发基本原则

整体

以表、列为整体进行处理

链路

有清晰的数据流转链路，能画出DAG

场景

先理解数据加工的需求，再学习具体技术

加工场景

Map

计算前后不改变数据粒度

- 四则运算，取对数等
- 文本加工
- 窗口计算

聚合

将数据分成若干组，分别对每组计算整体结果。常用于对明细数据聚合计算，得到的聚合指标用于指定维度的对比分析，如月份，商品品类等。

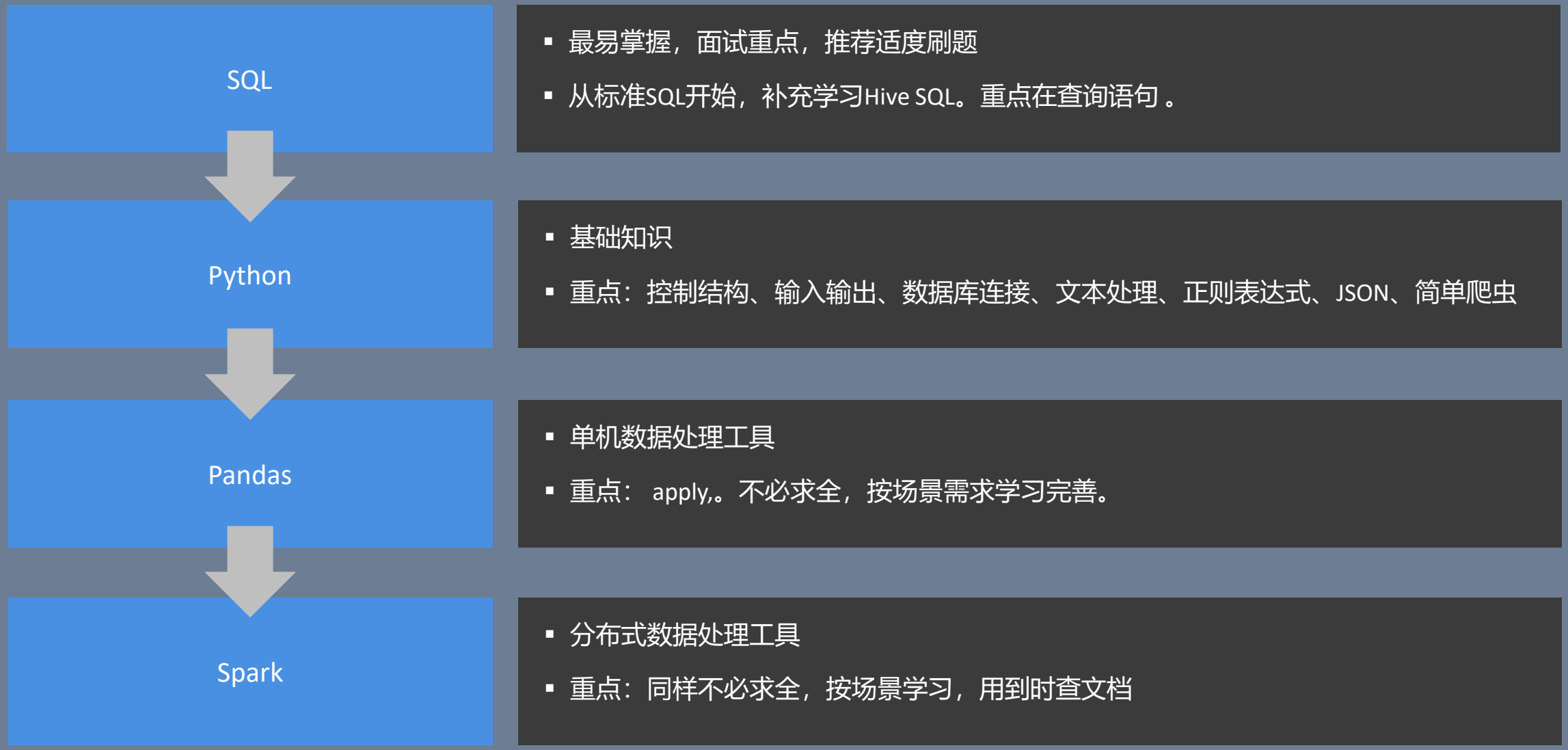
- 计数，求和，求均值等
- 每组的前N项
- 全组数据融合

下钻

从现有数据裂变，获得更详细的信息

- 展开一个复杂类型的数据列，如List，Map等，将其中元素展平
- 关联一份更详细的数据，获得详情信息。如一份月报数据，关联日报数据，即可得到日报信息+月报信息的完整展示

数据处理技术


SQL
最易掌握，面试重点，推荐适度刷题
从标准SQL开始，补充学习Hive SQL。重点在查询语句 。
Python
基础知识
重点：控制结构、输入输出、数据库连接、文本处理、正则表达式、JSON、简单爬虫
Pandas
单机数据处理工具
重点： apply,。不必求全，按场景需求学习完善。
Spark
分布式数据处理工具
重点：同样不必求全，按场景学习，用到时查文档

北京師範大學
BEIJING NORMAL UNIVERSITY

第三讲 数据整合 2021.08.14

现场弹幕答疑

复杂问题投递邮箱 xingxing_fisher@163.com