Day11 让编排变的简单-数据湖工厂

大数据处理

大数据技术，是指从各种各样类型的数据中，快速获得有价值信息的能力。

大数据处理的过程就像做饭一样。

买菜

洗菜

配菜

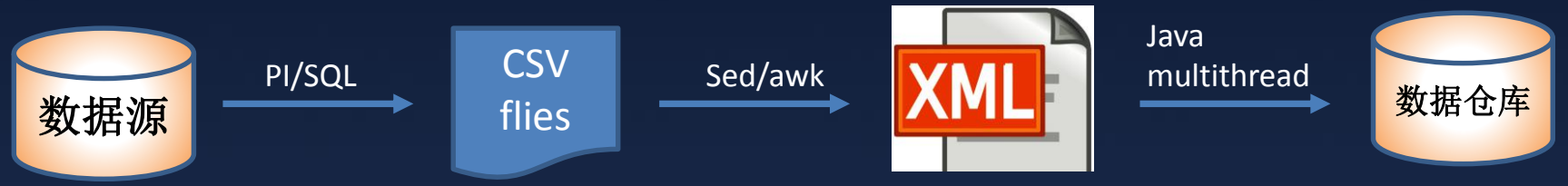
炒菜

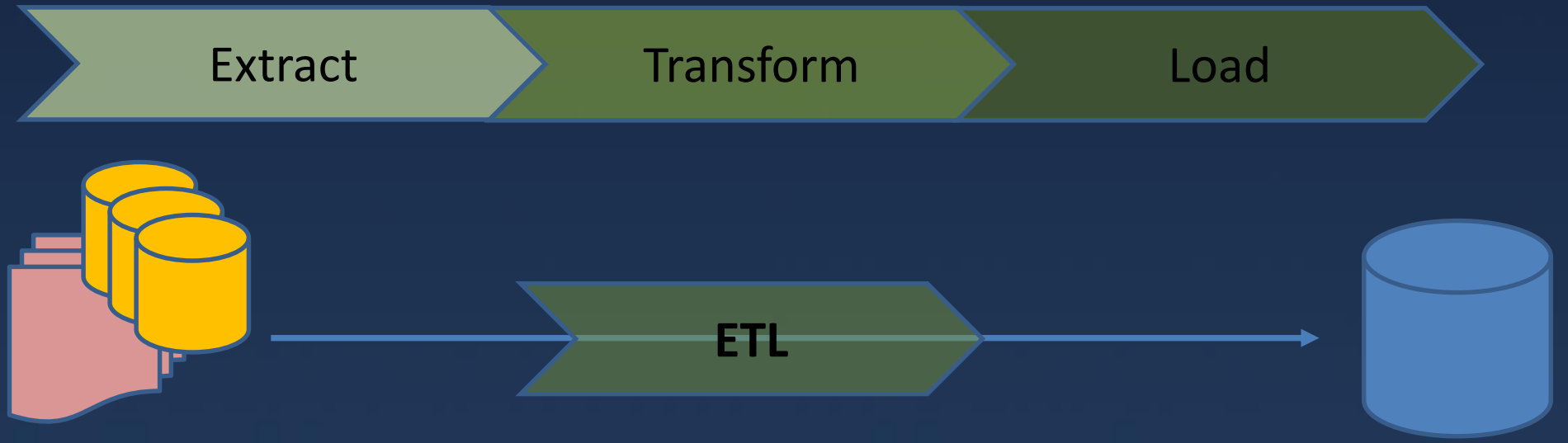
数据集成 → 数据清洗 → 数据建模 → 数据加工

HUAWEI

大数据处理工具——ETL

ETL，是英文 Extract-Transform-Load 的缩写，用来描述将数据从来源端经过萃取（extract）、转置（transform）、加载（load）至目的端的过程。





HUAWEI

使用ETL工具会遇到的一些问题……

- 数据没有按时处理完成怎么办？
- 数据处理失败了怎么办？
- 数据处理过程中其中一个步骤失败怎么办？
- 如何对数据处理的过程进行运维和监控？
- 如何控制数据处理工作流的调度？

HUAWEI

DLF解决这些问题

大数据处理一般较为复杂，数据建模、数据集成、脚本开发、作业调度、运维监控需要使用多种工具和服务组合，使用门槛较高。

大数据门槛高	➢ 涉及多个大数据服务，要串起来一个业务技术要求高 ➢ 每个服务存储格式，访问接口各不相同
业务开发难	➢ 打通本地开发工具与服务配置复杂，安全性差 ➢ 每个服务开发工具各不相同
运维专业性强	➢ 工作流缺乏可视化编排与调度 ➢ 业务执行过程各自监控，运维难度高

开发难

运维难

HUAWEI

DLF服务介绍

DLF是EI大数据核心服务MRS、DWS、DLI、CloudTable、MLS的配套服务，基本上各种EI大数据的业务场景都可以使用，它是简化用户使用华为云大数据服务的工具。

产品优势

一站式IDE平台：**一站式**建设云上数仓，无需切换多个工具

数据湖开发：管理**多种**大数据服务，可实现**跨服务**作业编排和调度

简单易用：在线SQL/Shell脚本编辑调试；预设**10多种**任务类型，拖拽式工作流编排

调度稳定高效：丰富的调度配置策略，**百万级别**的作业调度能力

HUAWEI

DLF服务介绍

DLF为serverless多租户架构，即开即用，而且目前是免费的。DLF使用的流程是按照用户使用大数据业务的流程来划分自身的功能。

DLF目前支持6种数据连接：

- 数据湖探索DLI
- 数据仓库DWS
- MapReduce服务Hive
- MapReduce服务SparkSQL
- 表格存储服务CloudTable
- 云数据库RDS

数据集成：

- CDM
- DIS
- CloudStream

脚本开发功能：

- 支持立即执行
- 支持格式化SQL
- 支持导出执行结果
- 支持参数设置

工作流开发功能：

- 支持多种节点
- 支持拖拽式开发
- 支持立即执行做调试
- 支持配置运行参数
- 支持添加便签

作业调度与监控功能：

- 支持启用、暂停或停止调度
- 支持查看调度执行历史
- 支持查看执行日志
- 支持可视化调度依赖关系
- 支持对单个原子重试、暂停、跳过、强制成功等操作

DLF工作流编排界面

节点库

工作流操作按钮

工作流画布

工作流调度配置

工作流运行日志

HUAWEI

典型场景1：BI报表开发和自动化调度

典型的报表工作流：

1、使用CDM服务例行将数据集成到数仓

2、使用多个SQL脚本进行报表数据生成

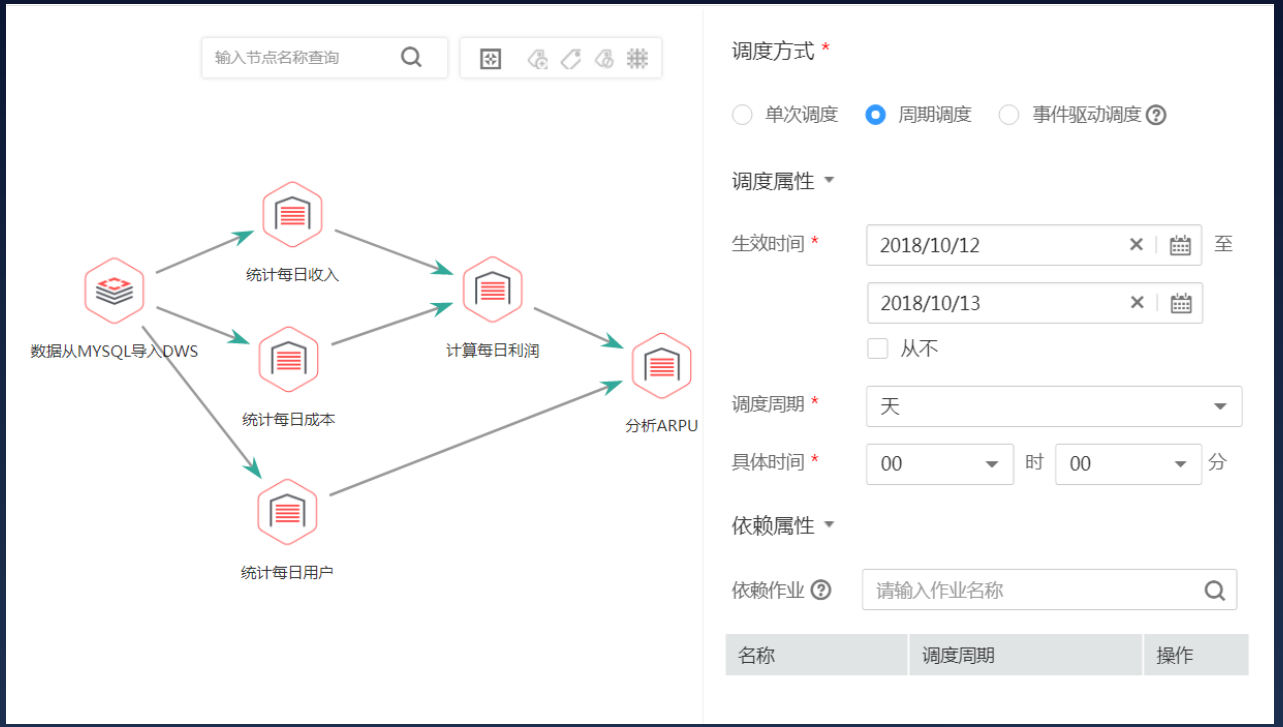
3、数据生成后再用可视化BI软件展现报表结果

DLF在BI报表生成场景下的价值：

1、支持报表脚本在线开发和调试

2、支持工作流编排，从数据集成到多个BI报表脚本

3、支持管理员方便进行运维，比如单个节点重试，暂停，跳过，强制成功等



典型场景2：机器学习流程自动化（如销售预测）

典型的机器学习流程：

1、将数据导入到大数据存储

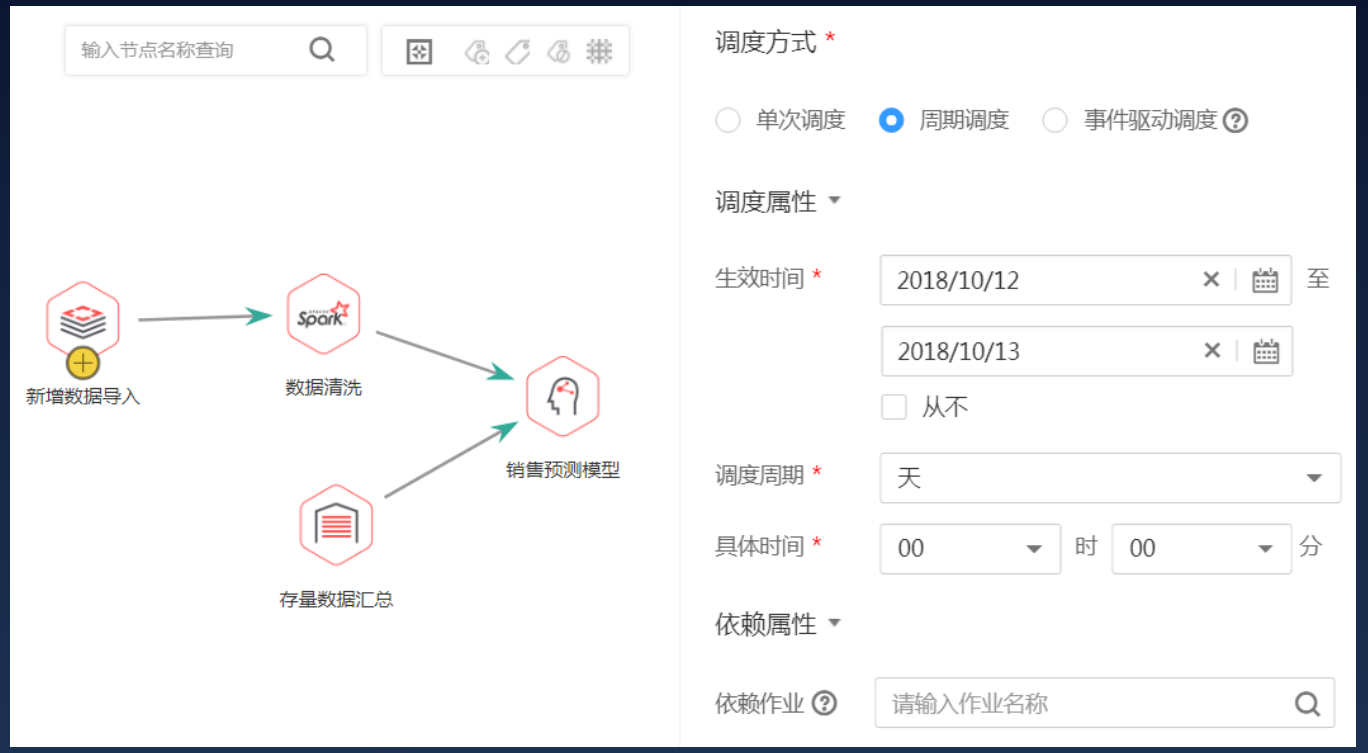
2、使用数据处理服务进行数据清洗，整理

3、使用机器学习任务进行数据挖掘

DLF在机器学习场景下的作用：

1、支持跨服务的混合流程编排

2、机器学习流程端到端可视化



HUAWEI

典型场景3：电商商品推荐

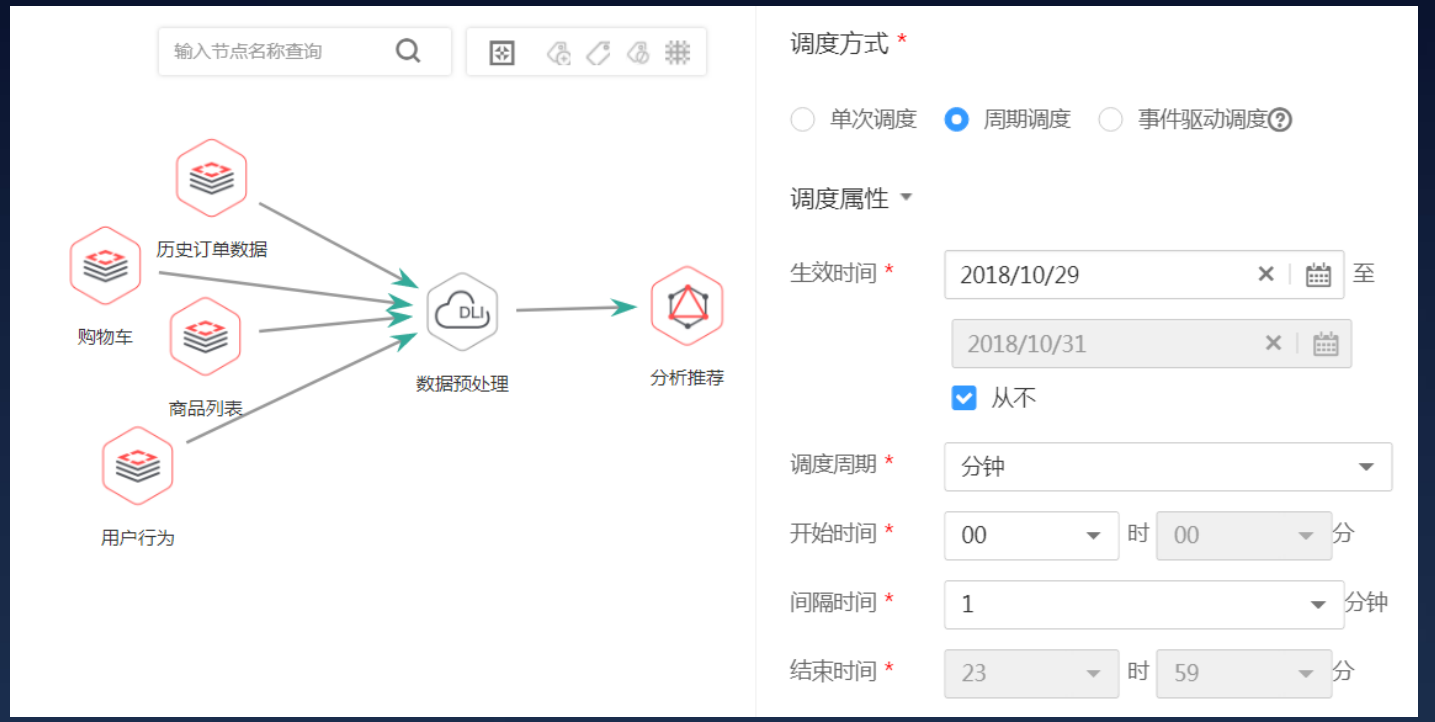
典型的推荐系统流程：

1. 将电商商品推荐的信息统一导入OBS
2. DLI SQL对OBS上的数据进行预处理
3. 使用分析推荐系统对OBS上预处理过的数据进行分析，输出推荐列表

DLF在推荐系统场景下的作用：

1、支持多数据源定时导入

2、调用推荐系统API定时完成推荐



数据湖工厂帮助文档详情查看：
https://support.huaweicloud.com/dlf/index.html

HUAWEI

Thank You.

华为云数据湖工厂服务DLF