## 课程目标：

* 影响数据质量的因素，数据质量问题类型
* 了解MaxCompute(DataIDE)产品
* 通过DataIDE设计数据质量监控流程
* 独立解决常见的数据质量监控需求

## 1 数据质量基本介绍

### 1.1数据质量相关概念：

数据的质量，数据就是拿来做分析的，正确的数据能得出正确的分析。

数据质量直接影响系统应用的成功。

数据质量管理，是指对数据从计划，获取，存储，共享，维护，应用，消亡生命周期的每个阶段里可能引发的各类数据质量问题。进行识别，度量，监控，预警等一系列管理活动。

数据质量管理，不仅包括了对数据质量的改善，还包括了对组织的改善。针对数据的改善和管理，主要包括数据分析，数据评估，数据清洗，数据监控，错误预警等内容。针对组织的改善和管理，主要包括确立组织数据质量改进目标，评估组织流程，指定组织流程改善计划，指定组织监督审核机制，实施改进，评估改善效果等多个环节。

数据质量管理好处：

1. 能够直接影响业绩
2. 对决策造成影响
3. 随着业务发展，分析系统中的数据流转和处理环节越来越多
4. 保证数据更好的为公司及企业的战略提供正确，有力的支撑

### 1.2数据质量影响因素：

**需求过程引发**

需求设计，开发，测试，上线等过程中引发的数据质量问题。

管理机制和流程不健全

**数据源引发**

上游数据源的不完善和不规范，导致下游系统收到影响。

这个式分析系统数据质量问题的主要来源，主要体现在：

信息不正确

信息不完整

信息不一致

**统计口径引发**

主要指KPI，报表等指标口径中存在的数据质量问题，包括指标准确性，一致性及完整性问题。

此类问题有以下几种情况：

不同源系统或不同业务部门，对相同名称的指标定义和口径不同，造成最终的指标统计结果出现差异。

业务部门再描述或定义指标口径时，在在许多不明确和不完善的地方，造成了指标口径的二义性。

**系统自身**

主要指的是系统再开发建设，日常运营和维护过程中，引发的数据质量问题，比如数据模型质量问题，系统升级数据信息丢失问题，ETL数据清洗不够彻底带来的问题以及数据仓库流程调度等问题。主要原因有：

数据仓库的建设过程中，缺乏规范化，体系化的建设思路，造成系统架构，数据模型，处理流程等不够优化和合理

在系统运营中，犹豫缺乏完善的日常管理和维护流程，在处理，监控数据是，没有可以一句和规范的处理手段，导致操作出现失误或一楼，造成数据质量问题。

### 1.3数据质量问题类型：

错误值：

数据存放错误。

重复值：

存在完全一样的重复记录

数据不一致：

数据记录是否符合规范。数据记录的规范和数据逻辑的一致性。

数据的完整性：

数据的记录和信息是否完整，是否存在缺失。

缺失值：

正常的数据信息记录中，存在信息缺失的情况。

异常值：

数据记录中出现明显得数据偏差或者错误的记录。

### 1.4MaxComput,DataIDE介绍：

MaxComput:

提供针对TB/PB级数据，实时性要求不高的分布式处理能力，应用于数据分析，挖掘，商业智能等领域。

特点：

分布式

安全：沙箱运行

易用：全面支持基于SQL的数据处理。

管理和授权：

DataIDE

大数据开发继承环境。提供可视化开发界面，离线任务调度运维，快速数据继承，多人协同工作。拥有强大的Open API 为数据应用开发者提供良好的在创作生态。

数据开发，数据管理，作业调度。

数据开发（DataIDE）常见场景

数据产生 -> 数据收集与储存 -> 数据分析与处理 -> 数据提取 ->数据展现与分享

中间三个流程为数据开发的流程。

DataIDE开发操作

在控制台里，选择大数据开发套件功能模块，选择对应的项目，进行开发操作。DataIDE主要包括数据集成，数据开发，数据管理，运维中心，项目管理，机器学习平台等功能模块。

此次的数据核查主要设计数据开发，运维中心两大模块。

* 数据开发页面是根据业务需求，设计数据计算流程，并实现为多个相互以来的任务，供调度系统自动执行的主要操作页面。
  + 任务开发，包含周期属性和依赖关系，是数据计算的主要载体，支持多种类型的任务和节点适应不同的场景。
  + 脚本开发时对周期任务的补充，通常用于辅助数据开发过程，主要用于实现非周期的临时数据处理，如临时表的增删改等，因此不包含周期属性和依赖关系。
  + 在大数据开发套件中，当一个任务被成功提交后，底层的调度系统从第二天开始，将会每天按照该人物有的时间属性生成实例，并根据上有依赖的实例运行结果和时间点运行。23：30之后提交成功的任务从第三天开始才会生成实力。
* 运维中心是日常运维的主要工具，可对已提交的工作流及其节点任务进行管理和维护，同时也可针对节点来添加监控报警。主要发呢为运维概览，任务管理，任务运维和监控报警四个模块。

### 1.5电商数据分析平台数据质量核查：

任务背景：

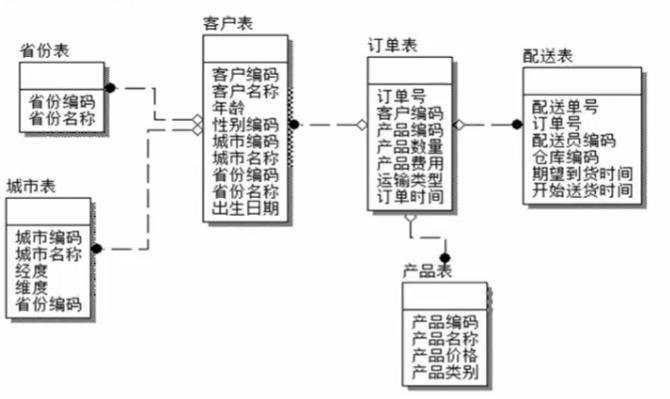
ABC是一家销售公司，其客户可以通过网站下单订购该公司经营范围内的商品，并使用信用卡，银行卡，转账的方式付费。付费成功后，ABC公司会根据客户地址一句就近原则选择自己的货舱，指派何是的快递人员配送商品，

经过几年的经营，该公司积累了一批经营数据，他们依托于阿里云数据计算服务，大数据开发套件搭建了一个企业级的数据平台，讲历史数据以及每天产生的数据都同步到该平台上去。由于种种遗留问题，造成这些数据中存在一些数据质量问题，为了能得到更准确的数据分析结果，数据分析师希望你这个刚加入的助理大数据工程师能合理的使用掌握的MaxCompute的技术，发现数据中的质量问题。

任务涉及业务数据

此次数据质量监控实验，主要涉及如下业务表。

* 订单表：保存客户订购产品所产生的订单信息
* 客户表：保存客户的姓名，地址信息
* 配送表：保存配送员与订单的对应信息
* 产品表：保存产品相关的信息
* 省份表：保存省份信息
* 城市表：保存城市信息



任务目标：

主要针对相关业务数据实现如下几种监控规则，并输出数据质量监控报告。

错误值：订单表某些时间格式存在问题，导致数据库人员将部分时间字段设成了字符串型

重复值：订单系统中部分记录关键信息重复（同样的人在同样的时间下了不同的订单），导致客户投诉。

数据不一致：城市信息名称未标准化，导致在数据分析时，未能吧相同低于的数值汇总在一起。

数据的完整性：配送的订单在订单表中不存在，导致物流人员空跑，效率下降

缺失值：部分客户性别信息确实，影响后续使用

异常值：单月购买次数异常（当月购买次数大于10次）



思考和任务：

