什么是数据仓库？如何构建数据仓库？

数据仓库是一个面向主题的（Subject Oriented）、集成的（Integrate）、相对稳定的（Non-Volatile）、反映历史变化（Time Variant）的数据集合，它用于支持企业或组织的决策分析处理。

数仓构建：

1). 前期业务调研 需求调研 数据调研 技术选型

2). 提炼业务模型，总线矩阵，划分主题域；

3). 定制规范 命名规范、开发规范、流程规范

4). 数仓架构分层：一般分为

操作数据层（ODS）、公共维度模型层（CDM）和应用数据层（ADS），其中公共维度模型层包括明细数据层（DWD和汇总数据层（DWS）

公共维度模型层（CDM）：存放明细事实数据、维表数据及公共指标汇总数据，其中明细事实数据、维表数据一般根据ODS层数据加工生成；公共指标汇总数据一般根据维表数据和明细事实数据加工生成。

CDM层又细分为DWD层和DWS层，分别是明细数据层和汇总数据层，采用维度模型方法作为理论基础，更多地采用一些维度退化手法，将维度退化至事实表中，减少事实表和维表的关联，提高明细数据表的易用性；同时在汇总数据层，加强指标的维度退化，采取更多的宽表化手段构建公共指标数据层，提升公共指标的复用性，减少重复加工。

应用数据层（ADS）：存放数据产品个性化的统计指标数据，根据CDM层与ODS层加工生成。

5）.选择合适的数据模型，不同的行业涉选取的模型近不相同，合适的模型，更利于在数据存储，计算，开发，安全，以及数据查询的效率，更能体现数仓的价值。

如何建设数据中台？可简单说下对中台理解与思路

建设数据中台的步骤如下:

- 数据清洗: 清洗数据以确保数据质量。
- 数据整合: 整合来自不同来源的数据。
- 数据存储: 将整合好的数据存储在数据仓库或数据湖中。
- 数据治理: 设置数据治理流程，确保数据的安全性和可追溯性。
- 数据展示: 使用BI工具将数据展示给业务人员和决策者。

数据中台是一个数据管理平台，支持企业各部门使用数据进行决策和运营。它提供了数据清洗、整合、存储、治理和展示等功能，帮助企业提高数据质量和效率。

数据仓库、数据中台、数据湖的理解

数据仓库是一种专门用于存储和管理大量历史数据的系统。数据仓库通常使用数据模型来存储数据，使得数据可以被查询和分析。

数据中台是一种用于统一管理企业数据的平台。数据中台通常包含数据仓库、数据湖等功能，并且还提供数据治理、数据集成等功能。

数据湖是一种用于存储和管理大量原始数据的系统。数据湖通常使用分布式存储系统来存储数据，并且能够支持海量数据的存储和处理。

传统数仓的程度（建模工具、ETL工具、BI报表工具、调度系统）

传统数据仓库 (Data Warehouse) 通常由以下几部分组成：

建模工具：用于创建数据模型和关系图，常用工具有 ERwin、PowerDesigner、Visio等

ETL工具：用于数据抽取、转换和加载到数据仓库，常用工具有 Informatica、DataStage、kettle等

BI报表工具：用于数据展示和分析，常用工具有 superset、cboard、redash、帆软BI/QuickBI/PowerBI等

调度系统:用于管理和调度ETL任务和报表任务的执行，常用工具有 airflow、azkaban、ooize、xxl-job、dolphinscheduler、Zeus、hera、TASKCTL/自研平台等

这些工具通常是集成在一起使用的。

数仓最重要的是什么?

数据的准确性，好多数仓因为数据不准确被终止。

数据的真正价值在于数据驱动决策，通过数据指导运营,在一个不准确的数据驱动下，结果可想而知。

如何保证数据的准确性?

数据准确性的保证通常需要采用多种措施。其中一些常见的方法包括:

- 数据质量控制：在数据采集和输入过程中，对数据进行检查和校验，确保数据的准确性。
- 数据备份和容错：对数据进行备份，并在数据出现错误时能够进行恢复。
- 数据审核和监督：对数据进行审核和监督，确保数据的准确性。
- 可追溯性: 记录并追溯数据的生成和使用历史。
- 数据隔离: 通过使用数据隔离技术，将不同数据集隔离开来，以防止数据污染。
- 使用第三方源数据并进行验证

这些方法并不是简单地采用一种方法就能够解决问题，而是应该根据具体情况进行综合考虑。

用户画像（静态、动态标签，统计、规则、预测标签，衰退系数、标签权重）

用户画像是指对用户的描述，通常包括静态标签（如性别，年龄，教育水平等）和动态标签（如社交媒体活跃度，在线购物习惯等）。这些标签可以通过统计和规则来构建，也可以通过预测模型来预测。衰退系数是指随着时间的推移，用户标签的相对重要性会发生变化的程度。标签权重是指每个标签在描述用户画像中的相对重要性。

为什么要分层的思考？

空间换时间。通过建设多层次的数据模型供用户使用，避免用户直接使用操作型数据，可以更高效的访问数据。

把复杂问题简单化。讲一个复杂的任务分解成多个步骤来完成，每一层只处理单一的步骤，比较简单和容易理解。

而且便于维护数据的准确性，当数据出现问题之后，可以不用修复所有的数据，只需要从有问题的步骤开始修复。

便于处理业务的变化。随着业务的变化，只需要调整底层的数据，对应用层对业务的调整零感知.

数据分层的好处

1）清晰数据结构：每一个数据分层都有它的作用域和职责，在使用表的时候能更方便地定位和理解；

2）减少重复开发：规范数据分层，开发一些通用的中间层数据，能够减少极大的重复计算；

3）统一数据口径：通过数据分层，提供统一的数据出口，统一对外输出的数据口径；

4）复杂问题简单化：将一个复杂的任务分解成多个步骤来完成，每一层解决特定的问题。