翼支付数仓建设与数据治理实践之路

黄洛 翼支付 高级数据仓库开发

翼支付介绍

天翼电子商务有限公司是中国电信集团有限公司的成员企业，是国资委双百改革和发改委第四批混改“双试点”企业，也是“双试点”企业中唯一的金融科技公司。公司以翼支付APP为载体，提供支付方案、会员权益、民生服务、分期借贷、保险理财、消费电商等服务内容，依托区块链、云计算、大数据、人工智能等技术，致力于推动包括生活服务、金融服务的数字化升级

秉持“响应监管、服务民生、资源共享、合作多赢”的理念，聚焦“开放、安全、便捷”的核心产品力，翼支付坚持通过服务投入与产品升级，构建贴合需求的管理与业务体系，以交流融合的业务实践，推动产业各方实现数字化转型。

讲师简介

黄洛
高级数仓开发

9年数据仓库开发及数据治理经验，曾就职于众安保险。2019年加入翼支付，现就职于大数据与人工智能研究院，负责金融版块业务数仓建设及治理。有丰富的数仓建设、数据治理实践及大数据平台应用经验。

翼支付 DataFun.

CONTENTS

数据治理背景

代码冗余、任务时效不稳定

祖传代码严重，任务链路长，烟囱式开发严重，任务时效得不到保障。

元数据信息严重缺失

缺少建表责任人、字段中文备注、分区字段随意等，导致库表清理及新人上手难度很大。

数据安全风险高

敏感数据未加密，数据下载入口多或无下载记录等，数据安全风险高。

数据口径差异明显

在一些整合数据口径下，由于各自整合口径来源不同，后续指标口径不一致

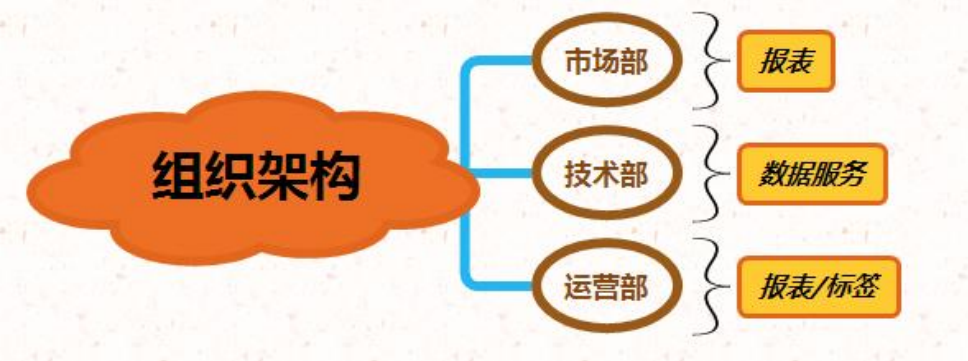
数据治理建设内容

组织协同	成立数据治理委员会(**牵头各组织协同推进治理进程，为治理分歧的最终决策组织**)、技术架构委员会(**公司系统信息架构审核，基础数据规范推行落地，提升原始数据质量**)、治理实施小组(**治理的落地组织，由业务、研发、大数据组成，统一考核，统一调度**)，在数据治理委员会的统筹下，紧密协作，形成统一、顺畅、敏捷的组织协同链路
平台建设	依托数据开发平台、BI平台、元数据管理平台、数据资产平台，构建统一的数字化和数据平台架构
数据应用治理	通过提升数据易用性、缩短计算和查询时效、提升数据质量、降低计算存储成本，构建敏捷的商业分析和数据洞察能力
数据规范	通过规范业务生产系统数据保证源数据的质量，构建数仓规范、主数据&元数据管理、数据分类分级保证数仓数据治理的质量等，形成完成全面的数据治理标准
数据安全	从数据存储、数据传输、数据使用三个方向进行数据安全链路改造，让企业数据符合国家对于数据安全的合规要求

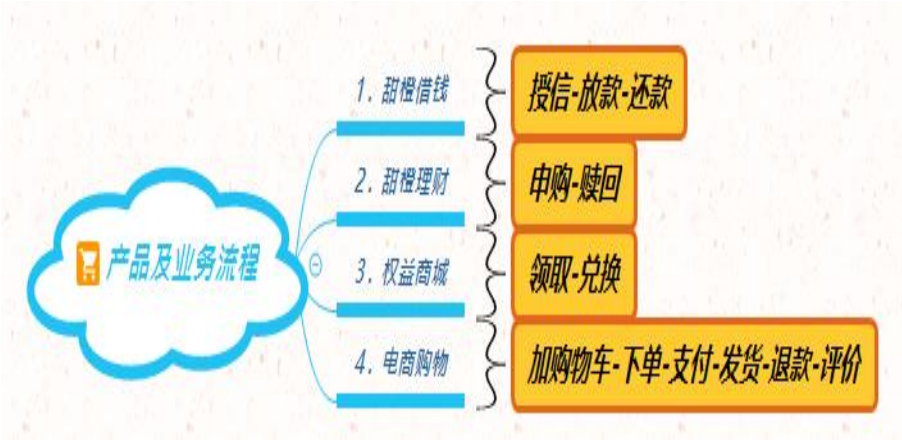
企业级数仓建设-调研阶段

01

02



03



04

企业级数仓建设-平台护航

平台是骨架，数据是血液！

数据开发平台： hive->spark计算引擎升级，双环境调度开发保障、数据质量监控保障、数据运行监控保障、任务运维等

即席查询平台：提供数据探查入口、管理数据下载审批流程等

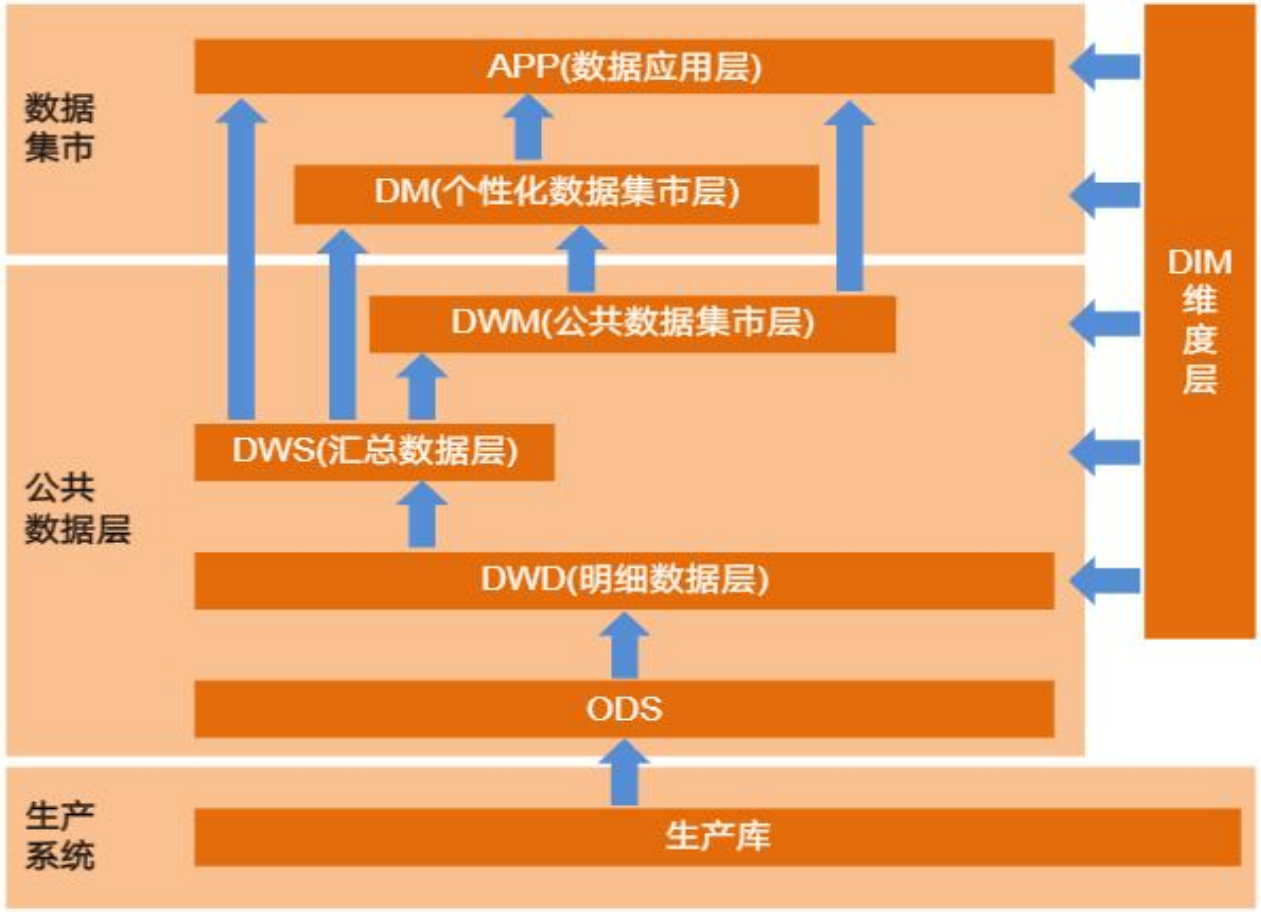
自研报表平台：自研可视化平台，推动国产化进程

元数据平台：覆盖建表管理、数据地图管理、血缘分析管理、表生命周期管理、冷热数据自动化管理、安全分类分级自动化管理

指标管理平台：管理原子指标、派生指标、衍生指标的相关元数据信息

企业级数仓建设-数仓分层

■ 数据分层架构示意图



数仓分层	简述
APP	应用场景驱动，定制化拼装
DM	各业务线个性化指标及业务分析主题数据加工

数仓分层	简述
DWM	业务分析主题驱动，业务分析主题宽表
DWS	明细宽表及汇总模型层
DIM	一致性维度层
DWD	明细事实层，业务过程建模驱动
ODS	贴源层，源业务系统的数据快照

企业级数仓建设-维度建模

维度建模四步曲

选择业务过程：以业务为基础，选择需要建模的业务过程，如授信、放款、还款、催收等均为不同的业务过程

声明粒度：粒度是维度建模中非常重要的要素之一，在同一事实表中，必须有相同的粒度，不能混用多种粒度(如grouping sets)， 如授信表中的授信流水号，订单表中订单号等

确认维度：维度在业务分析中占据了核心地位，一个好的数仓模型，通常会在不改变粒度的情况话退化一些常用的维度信息

确认事实：事实一般是数值类可累加的、不可重复的，在维度建模中一定要注意维度退化造成的数据重复问题，后续模型建设好也需要配置数据质量监控

企业级数仓建设-维度建模

命名规范

表命名规范：包含层级、数据域、产品线、业务过程、刷新周期、增全量标识等。如：dwd_{数据域缩写}_{产品缩写}_[业务过程缩写]_[自定义命名缩写]_{刷新周期标识}{单分区增量全量标识}

字段命名规范： 建设数仓共有的词根命名规范，原子指标命名如crdt_cnt(代表次数)，crdt_num(代表人数)；派生指标命名不可以在已命名的原子指标中间穿插修饰词，如： succe_crdt_num ,而不是crdt_succe_num

代码规范：代码中必须有任务名称(任务中文名+任务英文名)，功能描述、创建人、创建时间、修改记录等，用于后续数据异常追踪

企业级数仓建设-维度建模

资产沉淀

元数据建表：采取规范建表模式，除需要填写基本的库表信息及表的业务描述外，重要的是需要填写表的生产周期、分区的保留策略、数据层级、数据域、业务条线等信息来完善数据目录，还需要填写字段的分类分级、重要数据等保障数据安全资产沉淀。

数据开发任务上线：按照规范创建好表、准备好代码，先在数据开发平台测试通过，经系统审核后才可以发布到生产环境。

指标配置：完成任务上线后，需要在指标管理平台维护原子指标、派生指标、衍生指标的业务口径、技术口径及沉淀指标目标的相关元数据信息。

企业级数仓建设-维度建模

任务保障

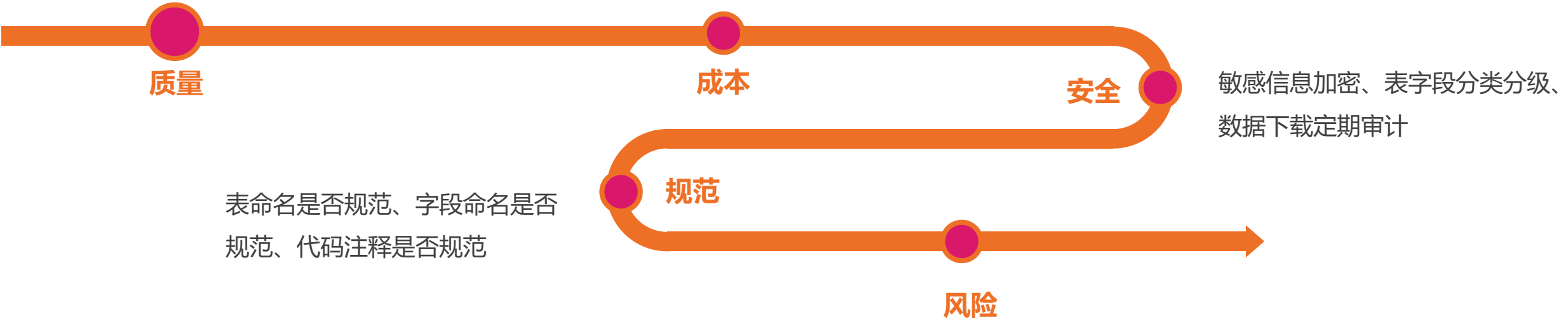
任务资源保障：在需求承接的时候，需要沟通好需求的保障等级，若是需要保障的，根据数仓规范中的定义等级将任务调整到响应的资源队列保障凌晨任务有足够的资源队列。

任务质量监控：对需要质量监控的任务配置 主键唯一性、枚举值是否有空、非空检查、长度检查、字段数值范围检查、数据波动性等

任务调度监控：任务失败预警、任务运行相对最近7天运行平均时长过长预警、指定时间未完成预警等。

企业级数仓建设-数据监控

数据治理是一个长期的过程，不是一次性的！



数据治理成效

成本&时效

- 一站式治理，2023年治理多个部门，平均资源降低**86.18%，计算成本降低近千万元/年**，平均时效提升**4.72h**。

元数据

- 保障完整的元数据信息、血缘关系、表生命周期、冷热数据标识，次均治理人力节省**3PD**。

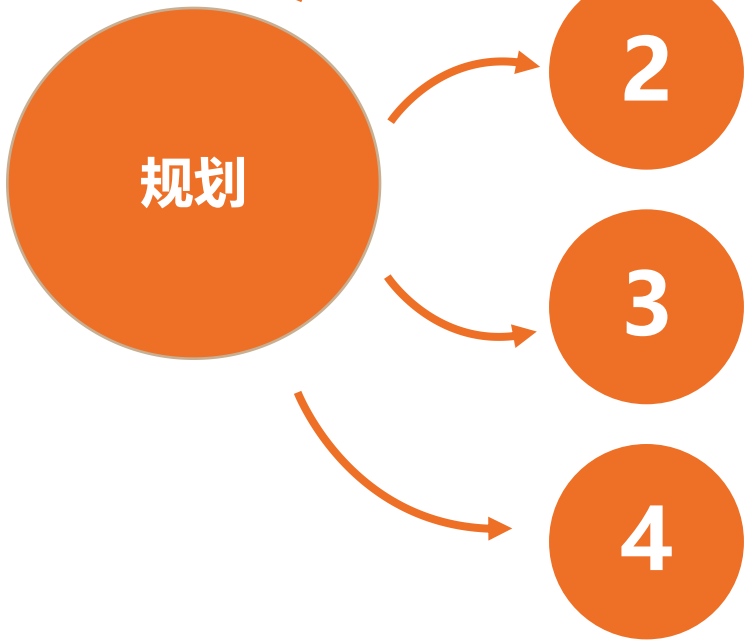
数据安全

- 敏感数据完成分类分级，且L3及以上数据**100%**加密，数据下载需要审批且下载记录留存。

指标管理

- **从0-1建设**，通过原子指标、统计粒度、业务限定、统计周期四要素来定义管理指标。

未来规划



数仓驾驶舱

数仓自己的管理驾驶舱，一览当前数仓模型中存在的不规范操作及异常监控情况等。

资产管理

数据资产总视图，总览计算、存储、小文件、安全等问题，提供一站式治理方案

指标管理

通过指标管理四要素配置预生产代码，直接在生产使用，减少指标的重复建设

数据赋能

数据的最终要素还是提供生产力，努力去探索更多的数据赋能场景

感谢您的观看

翼支付 DataFun.