翼支付 | DataFun.

翼支付数据治理实践之元数据管理

王平　翼支付产品总监

鲍旭　翼支付资深大数据研发工程师

目录
CONTENT

翼支付 | DataFun.

翼支付 | DataFun.

01 元数据的定位

数据治理与元数据治理

数据治理与元数据治理

“清洁数据成就卓越运营，智慧数据驱动有效增长。”元数据是数据治理的基础设施，在数据治理过程中有着核心作用

企业数据治理面临的核心问题

问题1：数据质量和时效不高

问题2：核心数据识别困难，数据一致性差

问题3：数据治理前清后乱，难以维持

问题4：数据安全风险居高不下

问题5：数据开发烟囱化严重，基础数据面临多次重复建设

…

元数据在数据治理中的核心作用

成本
效率
质量
安全

通过资产元数据识别数据表价值，通过数据血缘识别任务链路，推进核心/低价值任务治理

通过主数据治理及数据质量提升，提升数据一致性及数据质量

通过数据分类分级及数据安全治理，降低生产及大数据侧数据存储、传输和使用安全风险

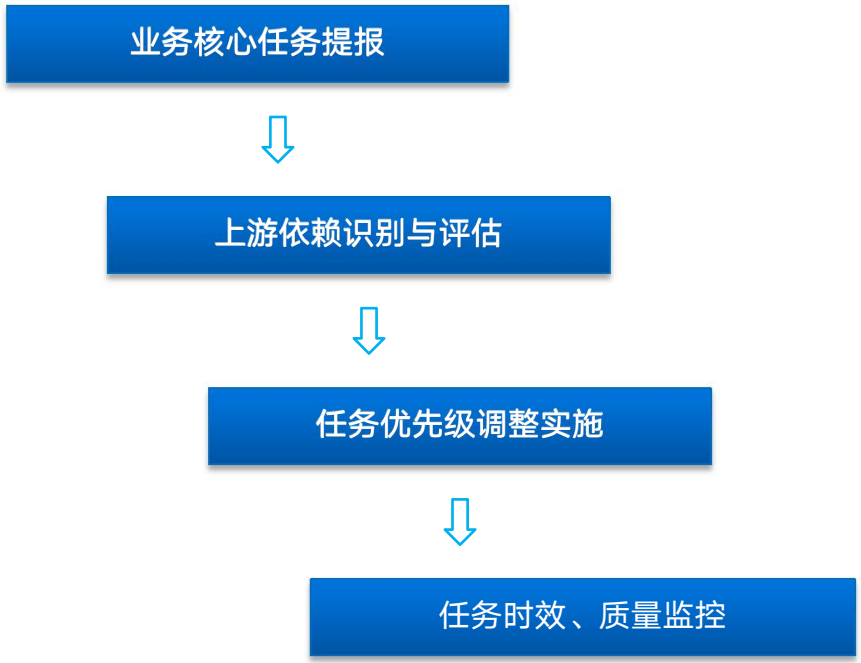
翼支付 | DataFun.

02 元数据治理体系

核心数据保障、主数据治理、数据规范体系建立、产品架构

核心数据保障

问题1：数据质量和时效不高



资源分配方案：

项目空间：支持多租户及资源分配，可控制每个空间的队列资源、优先级及任务最早启动时间

队列划分：核心>重要>一般（核心任务由数仓统一管控）

资源策略：5点前，所有资源优先核心队列任务，5点后按优先级及依赖进行资源分配

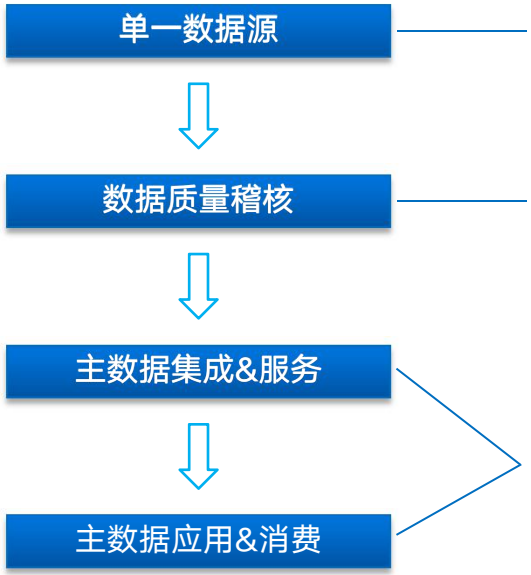
任务优先级变更控制规范：

事业群申请->数仓评估范围->大数据领导审批->数仓实施->运维监控保障

主数据治理

问题2：核心数据识别困难，数据一致性差

通过主数据治理实现同源多用+数据质量提升，逐步建立起主数据的权威。



主数据定义

主数据是具有高价值、跨流程、跨系统重复使用的数据，需要从基础数据中逐步识别和完善。需重点确保主数据的唯一性和权威性。

质量管控与提升

主数据的质量问题直接影响数据应用的稳定性及主数据本身的推广，建立严格的数据稽核机制及质量管控策略是主数据治理成败的关键。

数据\流程\IT协同

主数据的应用遵循规范的企业流程（新增数据属于主数据范畴，不允许新建；存量进行识别\排期整改），统一进行主数据集成并在元数据中标记出主数据表，再根据业务需求进行服务构建和数据消费。

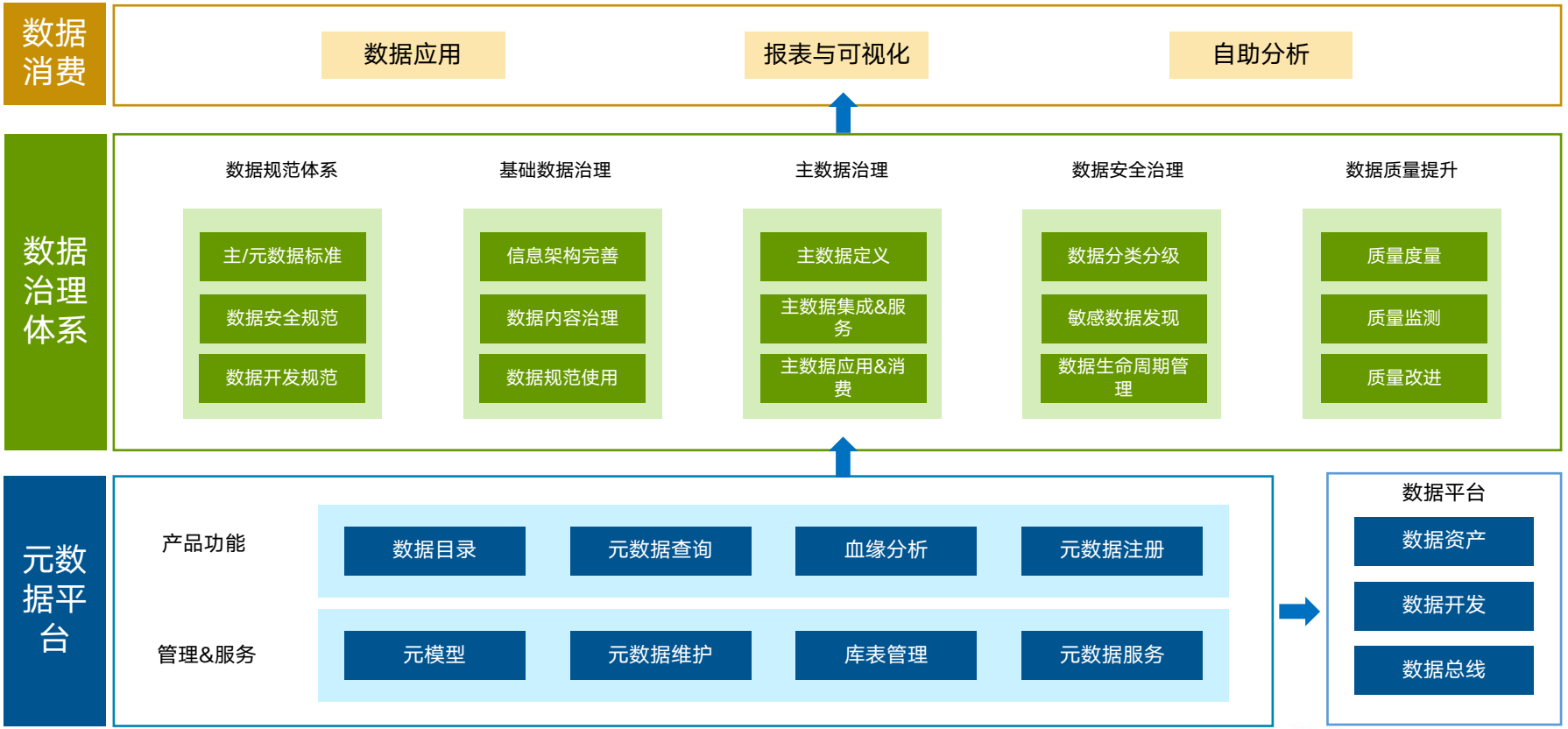
数据规范体系建立

问题3：数据治理前清后乱，难以维持； 问题4：数据安全风险居高不下

坚持生产源头治理并行：（杜绝前清后乱）

1. 数据安全治理（存储、传输、使用）
2. 生产元数据治理（库表字段命名规范统一）
3. 主数据识别与应用

产品架构



翼支付 | DataFun.

03 元数据平台技术介绍

系统架构、元数据采集、全链路血缘

设计理念

Goods: Organizing Google's Datasets

Alon Halevy[2*], Flip Korn[1], Natalya F. Noy[1], Christopher Olston[1], Neoklis Polyzotis[1], Sudip Roy[1], Steven Euijong Whang[1]

[1]Google Research [2]Recruit Institute of Technology

alon@recruit.ai, {flip, noy, olston, npolyzotis, sudipr, swhang}@google.com

ABSTRACT

Enterprises increasingly rely on structured datasets to run their businesses. These datasets take a variety of forms, such as structured files, databases, spreadsheets, or even services that provide access to the data. The datasets often reside in different storage systems, may vary in their formats, may change every day. In this paper, we present Goods, a project to rethink how we organize structured datasets at scale, in a setting where teams use diverse and often idiosyncratic ways to produce the datasets and where there is no centralized system for storing and querying them. Goods extracts metadata ranging from salient information about each dataset (owners, timestamps, schema) to relationships among datasets, such as similarity and provenance. It then exposes this metadata through services that allow engineers to find datasets within the company, to monitor datasets, to annotate them in order to enable others to use their datasets, and to analyze relationships between them. We discuss the technical challenges that we had to overcome in order to crawl and infer the metadata for billions of datasets, to maintain the consistency of our metadata catalog at scale, and to expose the metadata to users. We believe that many of the lessons that we learned are applicable to building large-scale enterprise-level data-management systems in general.

exist for managing datasets. We argue that developing principled and flexible approaches to dataset management has become imperative, lest companies run the risk of internal siloing of datasets, which, in turn, results in significant losses in productivity and opportunities, duplication of work, and mishandling of data.

Enterprise Data Management (EDM) is one common way to organize datasets in an enterprise setting. However, in the case of EDM, stakeholders in the company must embrace this approach, using an EDM system to publish, retrieve, and integrate their datasets. An alternative approach is to enable complete freedom within the enterprise to access and generate datasets and to solve the problem of finding the right data in a post-hoc manner. This approach is similar in spirit to the concept of *data lakes* [4, 22], where the lake comprises and continuously accumulates all the datasets generated within the enterprise. The goal is then to provide methods to "fish" the right datasets out of the lake on the as-needed basis.

In this paper, we describe Google Dataset Search (Goods), such a post-hoc system that we built in order to organize the datasets that are generated and used within Google. Specifically, Goods collects and aggregates metadata about datasets after the datasets were created, accessed, or updated by various pipelines, without interfering with dataset owners or users. Put differently, teams and engineers continue to generate and access datasets using the tools

[1]

Apache Atlas

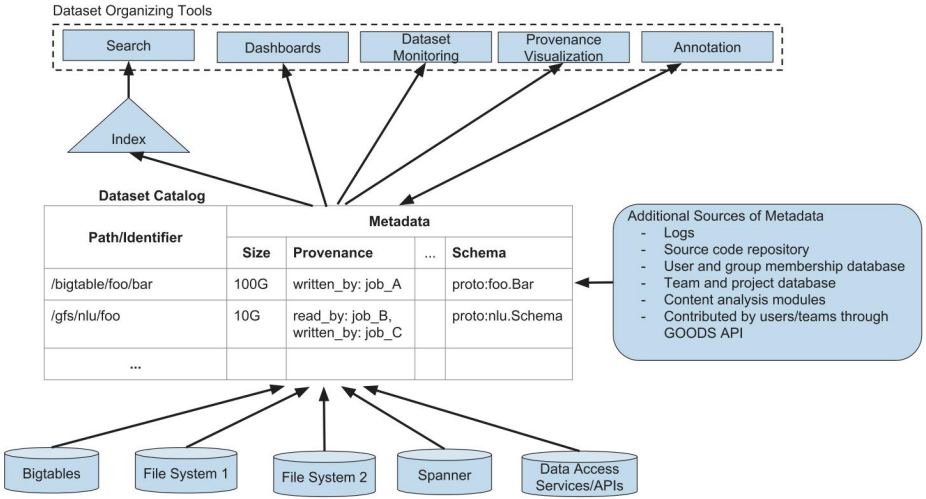
Amundsen

DataHub

翼支付 | DataFun.

设计理念

[1]



[1]

Metadata Groups	Metadata
Basic	size, format, aliases, last modified time, access control lists
Content-based	schema, number of records, data fingerprint, key field, frequent tokens, similar datasets
Provenance	reading jobs, writing jobs, downstream datasets, upstream datasets
User-supplied	description, annotations
Team and Project	project description, owner team name
Temporal	change history

Table 2: Metadata in the GOODS catalog.

设计理念

1. a post-hoc system
2. use bigtable as the storage medium: “blind writes”
3. a large number of diverse batch-processing jobs
4. a scoring function in dataset search

架构设计

接收层

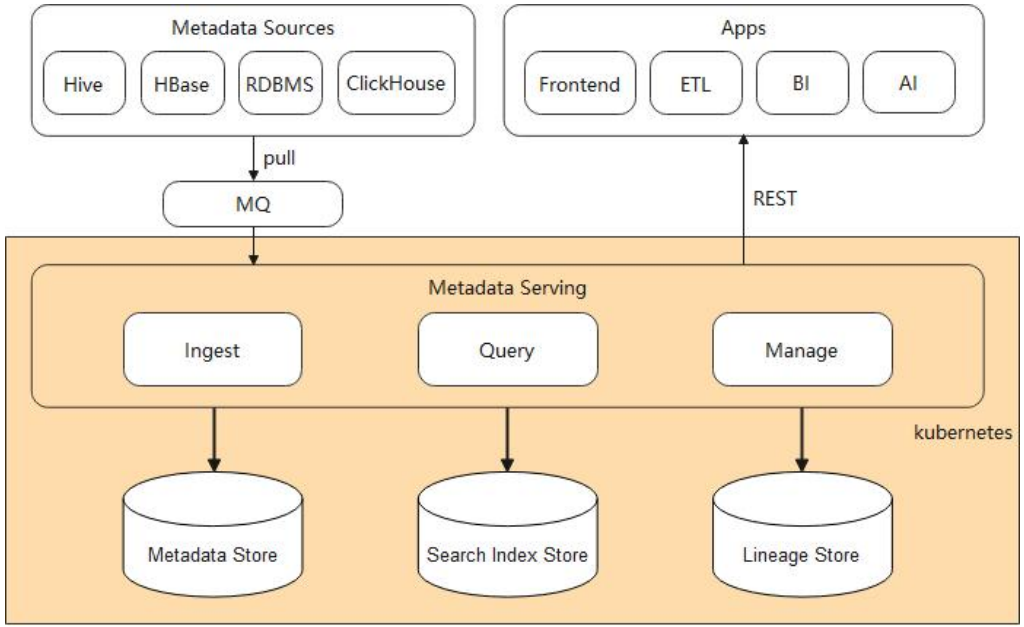
适配不同的数据源，接收元数据信息

服务层

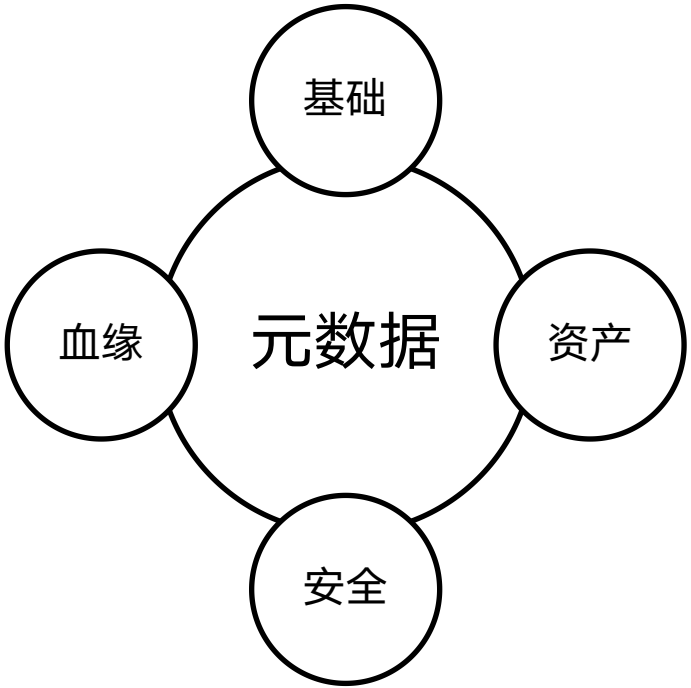
提供元数据查询、管理和分析服务

存储层

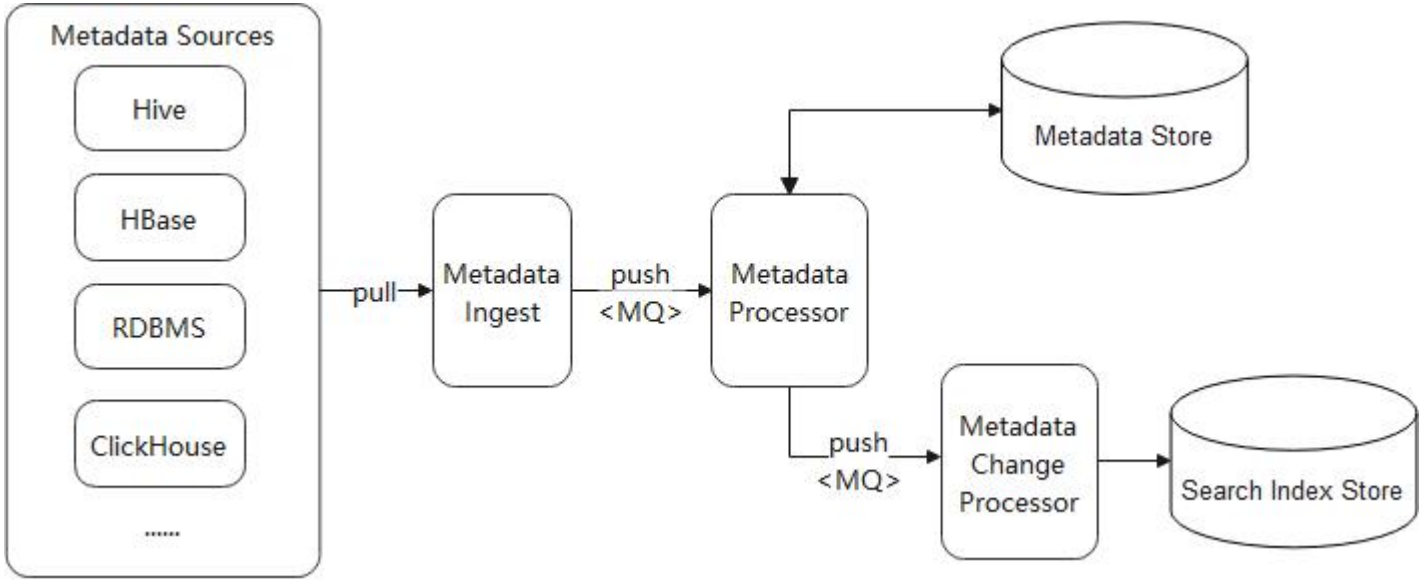
利用不同的存储系统存储元数据信息、查询索引和血缘信息



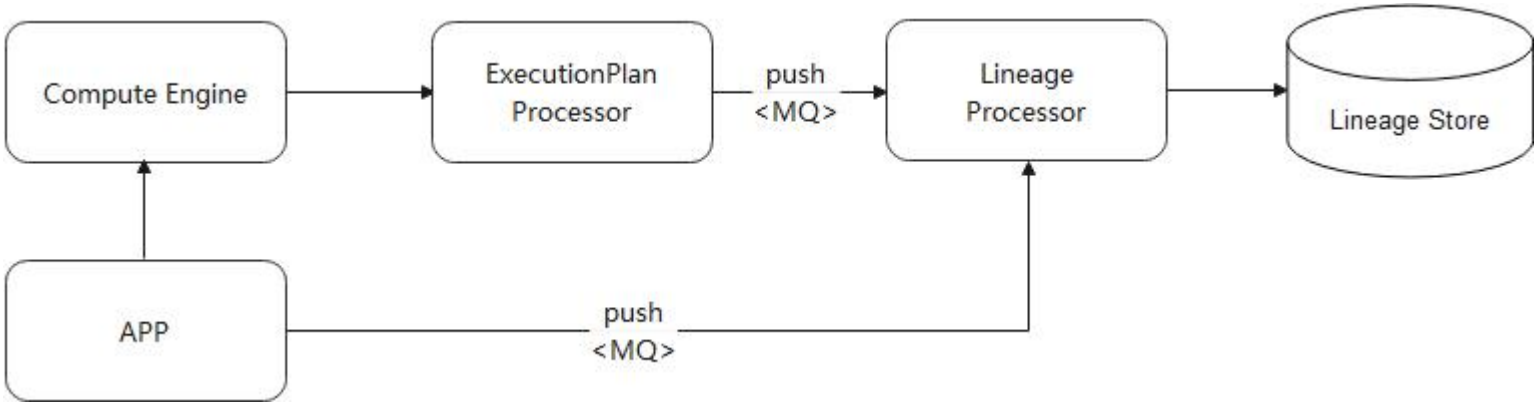
元数据模型



元数据采集



全链路血缘



翼支付 | DataFun.

04 未来展望

异构数据、容灾、智能推荐

未来展望

引用

[1] Alon Halevy, Flip Korn, Natalya F. Noy, Christopher Olston, Neoklis Polyzotis, Sudip Roy, & Steven Euijong Whang (2016). Goods: Organizing Google's Datasets international conference on management of data. https://readpaper.com/paper/2438792749

非常感谢您的观看

翼支付 | DataFun.