企业数据集市建设要点分析

高并发、低延迟场景下，企业如何建设敏捷高效的数据集市

精选

目 录

一、数据集市的发展态势 3

二、数据集市与数据仓库的对比 5

三、数据集市的项目建设方案及技术点 6

四、数据集市建设所需支持及面临的风险 9

五、数据集市项目验收 11

随着大数据技术的发展，数据量急剧增长，数据的分析挖掘工作也变得越来越重要。在这背后，数据仓库的建设尤为重要，通过数仓，可以为数据分析人员提供全面的数据和计算能力。

数据仓库包括原子级别的数据和轻度汇总的数据，是面向主题的、集成的、不可更新的(稳定性)、随时间不断变化（不同时间）的数据集合，用以支持经营管理中的决策制定过程。但是数据仓库在高并发、快速响应的场景下具有很大的局限性，无法满足海量投资者数据查询服务需求。

数据集市，也叫数据市场，是企业级数据仓库的一个子集，是为满足特定的部门或者用户的需求，只面向某个特定的主题，按照多维的方式进行存储，包括定义维度、需要计算的指标、维度的层次等，生成面向决策分析需求的数据立方体，提升数据查询的效率。

本文围绕着数据集市的需求、架构、方案等方面进行了问题的讨论，对于大家了解和应用数据集市非常具有参考价值。

一、数据集市的发展态势

Q1：当前，金融行业中在什么场景下需要建设数据集市？是否该启动数据集市建设？数据集市的建设是由哪些因素触发的？

A1-1：

数据集市(Data Mart)，也叫数据市场，是企业级数据仓库的一个子集，是为满足特定的部门或者用户的需求，只面向某个特定的主题，数据集市存储为特定用户预先计算好的数据，从而满足用户对性能的需求。数据集市可以在一定程度上缓解访问数据仓库的瓶颈。

数仓解决海量数据统计分析，低用户并发，大量计算。

数据集市低延时和高并发查询的情况下具备足够的支撑能力，可以7×24对外提供数据服务，且不影响原有数据仓库统计分析应用的正常运行。大量生产数据的预处理在数仓进行，数据集市接收数仓预处理后的数据。

数据集市是对数仓的补充，灵活、快速响应业务，支持用户部门自行定制数据进行统计分析，

支持高并发、性价比较高。

A1-2 ：

当用户对大量数据的实时性要求非常迫切，而数据仓库无法满足时，就需要有针对性的对这些

数据进行处理，提升存储和计算的效率，这就需要建立数据集市，通过对数据的预处理，提升

数据查询处理效率。

Q2 ：企业建设数据集市的动力来自哪里？ 建设数据集市对金融行业客户有何现实意义？

A2 ：

数据集市是数据仓库的一个重要补充，满足客户对数据的即时性要求；传统数仓往往比较重存

储，且为了保持数据完整性，而严格按照范式要求保存数据，从而导致计算量很大，时效性较

低。建立数据集市，有助于进一步发挥数据价值，因其面向主题，可以对数据做各种处理，从

而在设计上提升数据存储和查询的效率。

Q3 ：目前金融行业中，规划或者已经建设了数据集市项目的企业多吗？数据集市在金融行业中

发展的态势如何？

A3-1 ：

数据集市适合行业内拥有海量数据的公司，需对特定领域服务，通过互联网，对公众、监管机

构提供数据服务，好处是快速灵活、降低原有数据仓库成本。

A3-2 ：

金融行业积累了非常多的数据，数据质量也很高，随着金融业与技术的紧密结合，面向某一领

域的数据需求变得越来越强烈，比如行情数据，很多应用都需要，这时如果形成行情的数据集

市，提供数据服务，不仅避免重复建设，而且有利于数据的统一管理。

Q4 ：数据集市项目的直接用户和间接用户是哪些方面？数据集市会为用户带来哪些好处？

A4-1 ：

直接用户是各业务部门、各业务系统，数仓不直接为各业务系统直接提供数据服务，通过数据集市提供；间接用户是各业务系统的用户，如通过互联网办理业务的投资者用户。直接用户是可以定制化数据服务，方便、快捷，间接用户是快速响应、更好用户体验、高性能和可用性。

A4-2 ：

直接用户是业务数据分析人员、业务系统负责人员，对数据有强烈的需求和敏锐的嗅觉，能够将数据价值抽象出来；间接用户就是一线业务人员，可以直接获取该数据价值，对用户及时反馈。数据集市提升了数据交互的时间，使数据的使用变得流畅。

二、数据集市与数据仓库的对比

Q1 ：数据集市的出现，会使数据仓库消亡吗？

A1-1 ：

不会的，两者在概念上并不完全平级，就像工厂和门店的关系一样，两者应该是相辅相成，充分发挥数据的价值的。

Q2 ：数据集市、数据仓库在方法论上有什么差异，为什么选择不同的技术和平台？相比而言，数据集市和数据仓库各有什么优点和缺点？

A2-1 ：

本质上没有区别，如豪华汽车和中端汽车，目的是解决不同场景的问题。

数据集市 (Data Mart) ，也叫数据市场，是企业级数据仓库的一个子集，是为满足特定的部门或者用户的需求，只面向某个特定的主题，数据集市存储为特定用户预先计算好的数据，从而满足用户对性能的需求。数据集市可以在一定程度上缓解访问数据仓库的瓶颈。

数仓解决海量数据统计分析，低用户并发，大量计算。

数据集市低延时和高并发查询的情况下具备足够的支撑能力，可以 7 × 24 对外提供数据服务，且不影响原有数据仓库统计分析应用的正常运行。

A2-2 ：

数据仓库用于存放着非常多的各类数据，而数据集市是把某些方面的数据包装对外服务的。所以数仓的数据全，粒度细，没有很好的加工，而数据集市可能会改变原有数据，所以粒度是较粗的。数据仓库的优点是数据全面，但是分析时效性慢，而数据集市更关注于时效性和应用场景，不关心数据本身结构。

三、数据集市的项目建设方案及技术点

Q1 ：在软 / 硬件、开发、运行、维护和管理等各个方面面临哪些困难？如何克服？

A1-1 ：

可能主要在三方面：

1、技术管理：专门的开发运维团队，和数仓（大仓）间的职责分工，集市技术人员与业务人员间的职责分工明确

2、技术架构： 集市数据架构设计， 从数仓 （大仓） 数据加载， 数据集市运维监控， 集市高性能、高可用机制，业务人员灵活定制数据服务与集市日常数据服务、集市数据加载间的资源分配，新技术掌控力度等

3、信息安全：开发、运维分离，生产操作各项安全审计，数据库本身权限最小、数据隔离机制等

A1-2 ：

类比大数据平台，数据 ETL 是个非常重要且比较有难度的技术点，往往会有人提出数据导的不正常或丢失、数据未正常到达等情况，这个没有较好的解决方法，只能做好监控，尽量保证各

个环节正常流转。在运维方面，集群往往面临着查询和存储压力，如何在遇到问题时有效解决或者预警，这也是非常有挑战的，可以通过对平台底层的了解，不断对集群进行优化和监控。在开发方面，需要分析人员对业务需求非常了解，才能设计出合理的数据模型，提升数据服务的效率。

Q2 ：从数据的采集、数据的存储、数据的处理、数据的访问 / 供应等几个方面，那种方案合适，以及如何进行设计？

A2-1 ：

一般分为数据采集、数据传输、数据处理、数据存储、平台服务等方面，具体每个部分技术方案需要看企业的需求，关键点包括：数据治理、数据加载、数据存储等方面。

A2-2 ：

整体架构包括采集、存储、处理和访问等层次，就像建立大数据平台一样，这都是需要重点考虑的技术点， 至于哪种方案合适， 首先还是要了解各个层次对应的技术组件和成熟的解决方案，对各个技术点重点测试，掌握优缺点后才能合理组合各个组件。每个公司都会找到适合自己的整体方案，毕竟，并没有一种方案能完全适应各个公司的数据情况。

Q3 ：数据集市完成的目标和使命是什么？从这个意义上讲，应该为数据集市的成功标准设计什么样的 KPI ？

A3-1 ：

主要分为业务需求和技术需求两部分。业务需求主要是技术服务部门或系统的业务场景，建模即业务建模，需要对业务关系、实体充分了解，目前各厂商针对各个行业，如证券都有成熟业务模型。技术需求主要包括可维护性、性能、安全、监控运维等。

A3-2 ：

需求模型的确定需要详细了解业务需求，同时具备数据建模思维，将需求转化为数学模型，从

而通过数据集市计算引擎实现。 数据集市是否有效的衡量标准有扩展性、 性能、 稳定性、 安全、易用性等方面。

Q4 ：金融行业在数据集市项目中，是否涉及到主机、存储、数据库、系统管理等的选型？如何选型？

A4-1 ：

在数据集市项目中，重点关注数据集市的系统技术方案，好的数据存储和计算引擎是数据集市项目成功与否的关键，存储格式、计算性能、扩展性、稳定性、安全性、数据 ETL 等方面都是要考察的技术点，就像大家选择 oracle 做 OLTP 系统一样，选择一个好的底层引擎，才有可能将数据集市做大做强。

A4-2 ：

一般数据集市和企业数据仓库或大数据平台一起进行选型，主要考虑点包括分布式系统，性价比低，国产厂商，金融行业成功案例，满足各项需求指标等。

Q5 ：金融行业在规划数据集市项目时，有哪些技术路线可选？如何选？还有，有哪些数据集市的供应商可选？如何考察？

A5 ：

最主要在于分布式数据库选型，如传统关系型如 oracle 、db2 ， mpp 架构 GP、TD、Gbase ，新兴分布式架构，如阿里云、华为、星环大数据平台等。

Q6 ：金融行业在实施数据集市项目时，成本组成有哪些？

A6-1 ：

主要四方面，软件平台厂商支持、团队建设、应用系统开发、硬件维护。

A6-2 ：

主要包括硬件服务器、软件平台、平台管理人员、数据分析团队（建模、数据服务化、系统开发）。

Q7 ：数据集市中有大量数据时，数据之间的关系应该怎么管理呢？场景问题：面对大量的数据集市中的数据，有可能针对同一个属性从不同的表里得到是存在不一致的情况的。这种情况怎么处理？

A7-1:

这需要两方面：

1、充分理解业务场景，进行元数据管理，定义各表中各字段的业务含义

2、加强数据治理，在数据 ETL 时对业务系统数据进行按照元数据规则进行预处理。

A7-2:

这个也是数据分析时遇到的共性问题，数据仓库同样由此问题，通常的做法就是建立元数据标准，在数据 ETL 的时候进行标准化处理，避免数据不一致的情况。

Q8 ：实施数据集市的项目时，对于项目参与人员，应该如何规划培训学习进行技能储备？

A8-1 ：

数据集市的项目跟建立大数据平台一样，是个大工程，都面临着数据 ETL、存储、计算、数据管理等技术方面的挑战，而比大数据平台更复杂的是数据集市是个面向业务人员的 OLAP 型系统，如何满足业务各种各样的建模需求也是建立数据集市面临的一个大问题。所以，项目人员既要懂底层技术，又要懂上层业务，只有这样才能发挥数据集市的优势。在技术方面，可以类比大数据平台，可以储备数据 ETL、分布式存储、分布式计算、管理和运维的经验；业务上就要深入了解业务需求，转化为数据模型。

四、数据集市建设所需支持及面临的风险

Q1 ：数据集市的建设方应该是那个部门？哪些部门需要提供支持？以及什么样的支持？

A1-1 ：

建设方为技术部门的数据仓库组或技术平台组，所服务业务部门，技术部门各业务系统组，采购、财务部门，按照各部门分工支持配合。

A1-2 ：

业务部门、技术部门（主要为数据分析团队、数据仓库团队）应当作为核心部门对此支持，主要为实施、分析和使用；其余部门应当作为辅助。

Q2 ：比如在技术上、组织授权上、人员力量上、流程上、管理上、资金投入上等各个方面，会有哪些风险，如何规避？

A2-1 ：

1、面临着管理需求同实际情况不能严格匹配。

2、项目推动过程需要各业务部门有效的统一组织和规划安排。

3、 IT 人员对技术的掌握和对项目控制很关键。

A2-2 ：

风险无处不在，应注意识别并预防，主要以下三方面

（1）管理制度、组织架构

专门的开发运维团队，和数仓（大仓）间的职责分工，集市技术人员与业务人员间的职责分工

（2）技术

集市数据架构设计，从数仓（大仓）数据加载，数据集市运维监控，集市高性能、高可用机制，业务人员灵活定制数据服务与集市日常数据服务、集市数据加载间的资源分配，新技术掌控力度

（3）信息安全

开发、运维分离，生产操作各项安全审计，数据库本身权限最小、数据隔离机制

A2-3 ：

不光是在数据集市项目上，几乎所有大型项目上，都存在技术、管理、组织、资金等方面的风险，如何规避这些风险，主要还是要从政策上支持、技术上把关、实施上严谨、分析人员业务精通、与业务人员时刻沟通避免方向错误等方面入手，严防风险的发生。

Q3 ：数据集市项目的直接用户和间接用户是哪些方面？数据集市会为用户带来哪些好处？

A3-1 ：

直接用户是各业务部门、各业务系统，数仓不直接为各业务系统直接提供数据服务，通过数据集市提供；间接用户是各业务系统的用户，如通过互联网办理业务的投资者用户。直接用户是可以定制化数据服务，方便、快捷，间接用户是快速响应、更好用户体验、高性能和可用性。

A3-2 ：

直接用户是业务数据分析人员、业务系统负责人员，对数据有强烈的需求和敏锐的嗅觉，能够将数据价值抽象出来；间接用户就是一线业务人员，可以直接获取该数据价值，对用户及时反馈。数据集市提升了数据交互的时间，使数据的使用变得流畅。

五、数据集市项目验收

Q1 ：如何根据数据集市项目设定的建设目标和标准，进行数据集市功能的测试和验收？

A1 ：

数据集市的建模是终端用户驱动的，终端用户必须参与数据集市的建模过程，因为他们显然是要使用该数据集市的人。所以建设目标要紧贴业务需求，解决用户最关注的的数据获取问题，数据集市的建立就是要建立一套面向业务的数据服务；建设标准一般根据自己公司的特点定制，

没有行业唯一的标准。测试和验收工作跟一般项目差不多，只要能够满足业务需求，用户可以很方便且无差错地通过数据集市获取数据，架构设计扩展性高，性能满足要求即可。