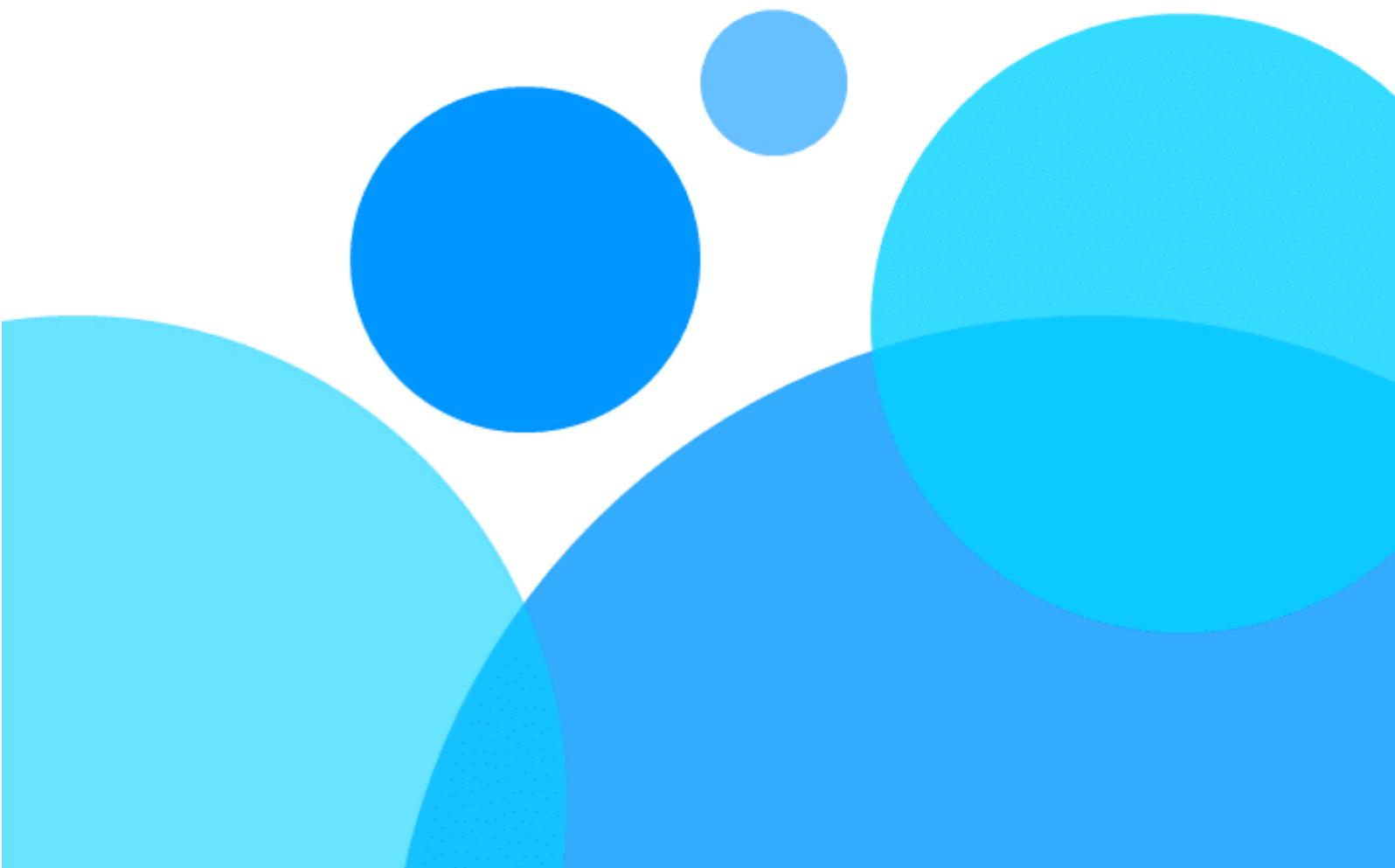




## 自由释放数据潜能

# FineDirect V1.0 应用方案建议

2017年05月27日



## 目录

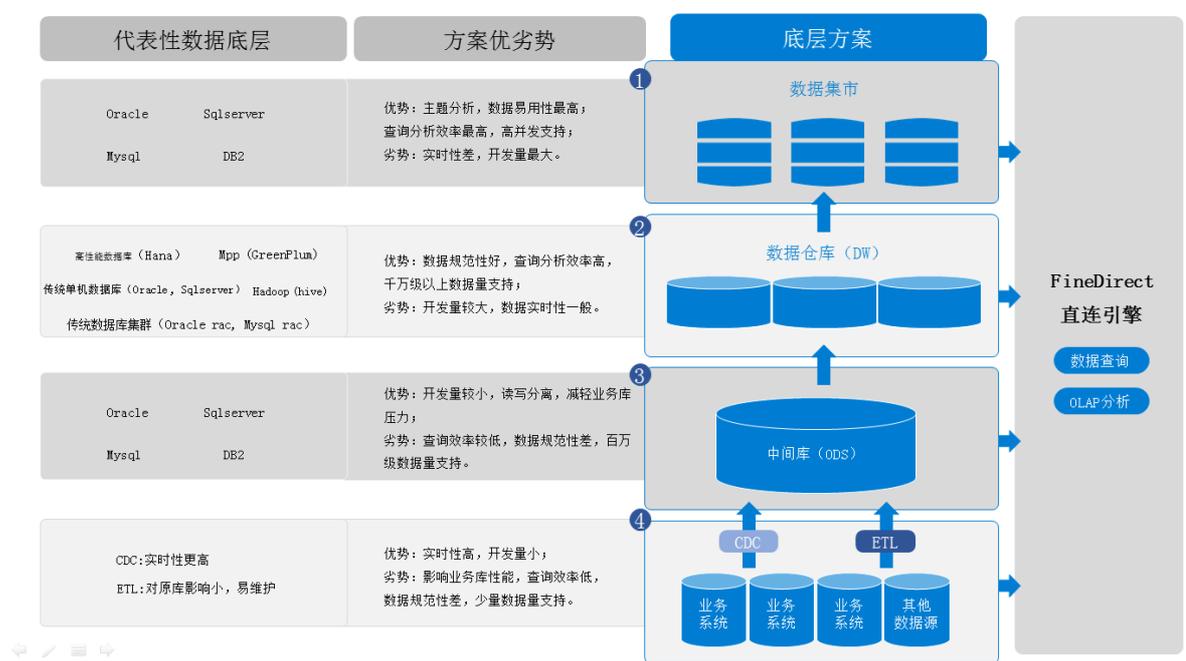
概述 .....	3
方案① 直连业务系统数据库.....	4
实际用户案例: .....	4
方案② 中间库 (ODS) .....	5
1、CDC 同步 .....	5
2、ETL 抽取 .....	5
实际用户案例: .....	6
方案③ 数据仓库 (DW) (推荐) .....	7
1. 传统关系型单机数据仓库: oracle, sqlserver, mysql, db2; .....	9
2. 传统关系型集群数据仓库: .oracle rac, mysql rac; .....	9
3. 高性能数据仓库: hana.....	9
4. hadoop 类数据仓库: hive, kylin; .....	10
5. mpp 类数据仓库: greenplum, vertica, teradata; (推荐) .....	10
实际客户案例: .....	11
方案④ 数据集市 (DM) .....	12
实际客户案例: .....	13

## 一、概述

FineBI 推出的直连引擎主要是为了解决大数据量的 OLAP 分析和分析的实时性。基于用户前端的可视化分析操作，通过 FineDirect 数据引擎将其转化为适应各类数据库方言的查询语言，因此底层数据库的选型与 BI 系统的应用效果有密切的关系，本文基于帆软多年行业经验和 FineDirect 客户实际案例，将常见的底层方案做整理，对各方案利弊做一定分析，希望给 FineDirect 分析引擎用户提供 BI 系统应用建设方面的一些建议。

文档主要整理四类常见方案：

直连业务系统数据库；中间库（ODS）；数据仓库（DW）；数据集市（DM）四种方案，每一种方案又细分数据库选型建议。结合帆软客户实际案例，给企业用户提供建议，供参考（参见图一）。



图一、数据分析平台建设底层方案图

## 二、方案 1：直连业务系统数据库

企业安装部署 FineBI，通过 FineDirect 分析引擎数据模块直接连接业务系统数据源，通过 FineDirect 分析模块，针对业务系统的原始数据源通过易用的拖拽操作进行 OLAP 分析和可视化的展示。

### 总体建议：

#### 优势：

- 1.几乎不需要实施时间，一天内安装部署完成，即可实现针对业务系统数据的可视化分析；
- 2.获取数据的实时性和准确的最高；

#### 劣势：

1. 查询和分析操作针对业务系统，事务性系统对于查询和分析操作响应效率差，分析效果差，仅可以完成较小数据量的查询操作；
2. 查询和分析会影响业务系统的正常运行；
3. 业务系统数据结构设计多数不适合查询分析操作；

### 实际用户案例：

少数情况可以使用，但不推荐使用。

### 三、方案 2： 中间库（ODS）

操作数据存储 ODS（Operational Data Store）是数据仓库体系结构中的一个可选部分，ODS 具备数据仓库的部分特征和 OLTP 系统的部分特征，它是“面向主题的、集成的、当前或接近当前的、不断变化的”数据。

企业将业务系统有查询和分析需要的数据抽取到中间库（ODS 层），安装部署 FineBI，通过 FineDirect 数据模块对接 ODS 层进行查询分析操作。

数据流从业务库到 ODS 层常用的有两种方式：

#### 1、CDC 同步

这种模式的原理是根据业务库的日志进行数据的同步，因此需要开启业务库日志。

优点：实时性高，数据响应快，基本能达秒级。

缺点：对业务库性能有一定影响，容错率较小，并且运维难度较大，对数据仓库的数据库种类有要求。例如通过 oracle 的 OGG 中间件实现实时数据同步。

#### 2、ETL 抽取

这种模式的更新要求业务库数据有时间戳或者其他相关的更新字段，需要在数据库设计时考虑进去。

优点：不影响业务系统使用性能，维护量小容错率高。

缺点：有一定的数据结果延迟，根据数据量大小，进行 ETL 时序设计，

当前数据量下不会超过 N+1 天模式。有一定开发量。

常用作 ODS 层选型数据库：Greenplum 等 mpp 数据库；oracle，sqlserver，mysql 等传统关系型数据库。

### 总体建议：

总体来说采用建立中间库(ODS)来为 FineDirect 提供数据源的方案：

优势：

- 1.中间库一般不需要太多的数据库设计，因此实施和开发周期都比较短；
- 2.中间库的数据延迟低，能做到实时分析。
- 3.读写分离，一定程度提高查询效率，减轻业务库压力。

劣势：

1.对于 OLAP 分析来说，由于没有良好的数据处理机制，数据的易用性和查询效率都不够高。

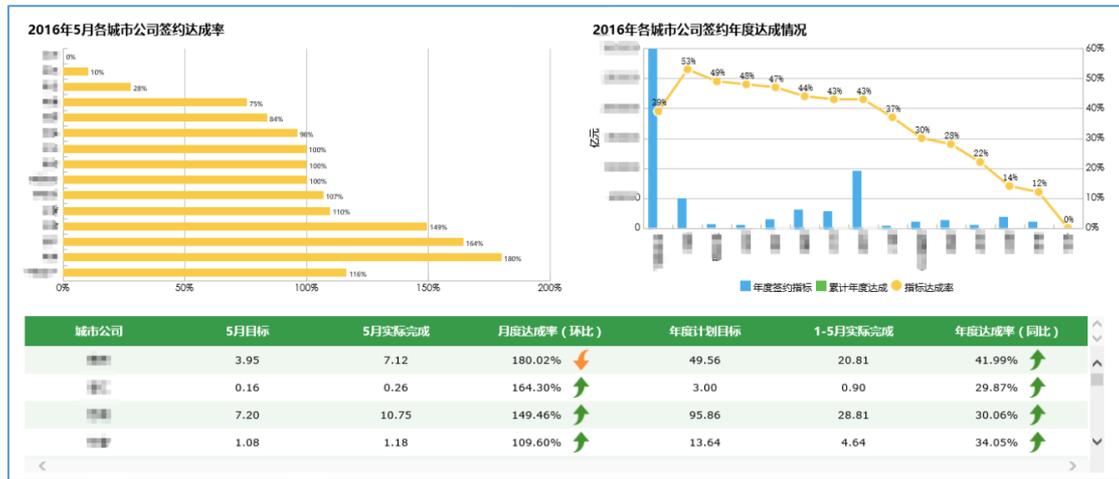
2.通常情况中间库的选型，无法支持较大的数据量，一般场景数据量在百万级别。

### 实际用户案例：

某地产集团：采用 Greenplum 数据库作为 ODS 层，通过 CDC 方式实现实时数据同步，实现亚秒级实时更新。Greenplum 部署 4 个节点，方案可以处理总数据量 1T，增量数据 1G/每天。当前实际数据总量 40 多 G，增量 300M/每天，同场景查询执行效率与 Sqlserver2008 相比提高 80 倍以上。

用户通过 FineDirect 数据引擎连接 Greenplum 中间层，实现实时数据报表查看和 OLAP 数据分析。

截图：



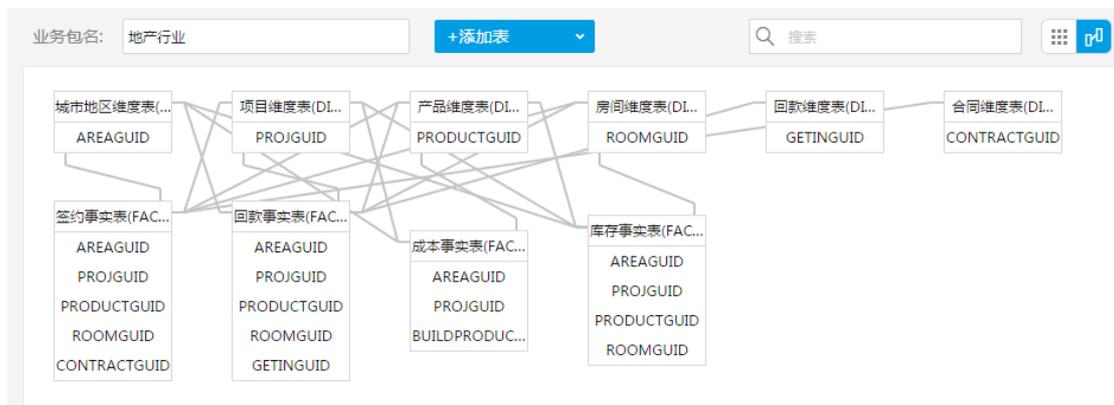
#### 四、方案③ 数据仓库 (DW) (推荐)

数据仓库 (Data Warehouse) 是决策支持系统和联机分析应用数据源的结构化数据环境。数据仓库研究和解决从数据库中获取信息的问题。数据仓库的特征在于面向主题、集成性、稳定性和时变性。

FineBI 的 FineDirect 数据引擎对接用户的数据仓库层是最好的选择，面向主题的特征正好配合 FineBI 的业务包管理方式，能让管理员更敏捷的准备数据，同时让用户更好找到想要进行的分析数据，集成性解决了数据规范性，数据一致性，数据孤岛的问题，更规范完善的数据能让数据分析的结果更有实际意义。



FineBI 按照主题进行数据管理



FineBI 数据模型管理

管理员在配置好业务包之后，分析人员可以根据有权限的数据，进行自助拖拽，进行上钻、下钻、切片、换维、同期环期等 OLAP 分析，更能通过 FineBI 丰富的可视化效果实现分析结果的良好呈现。



当然对于查询的效率和用户的体验来说，当中的数据仓库的建设是商业智能（BI）十分重要的一部分，为了更好的搭配 FineBI 直连分析引擎，下面给出一些常见的数据仓库方案建议。

### 1. 传统关系型单机数据仓库：oracle, sqlserver, mysql, db2;

这类数据仓库是当前大部分企业都具备的，不过针对日益增长的数据量，和对响应时间的要求的苛刻，这部分数仓已经不那么能满足需求。不过由于这类数据库，是以事务库为基础的，因此虽然在进行 olap 分析时，效果平平，但是对于并发的支持效果较好，因此更适合于大量固定报表需求的 BI 系统，百万级别数据量。

### 2. 传统关系型集群数据仓库：.oracle rac, mysql cluster;

同传统单机型数据仓库，集群的应用让 BI 系统的并发支持量更好，适应更大体量的企业用户。

### 3. 高性能数据仓库：hana

以 hana 为代表的高性能内存数据库，对于千万级别数据量的用户的分析需求，通过

FineDirect 对接效果是比较好的，通常的分析查询都可以做到 5s 内响应。

#### 4. hadoop 类数据仓库：hive, kylin;

Hadoop 系列的数据仓库，天生可以处理超大数据量，亿级，十亿级别，数据量不是问题。但是 hive 这类数据仓库，对于 olap 分析支持效果一般，即查询效率不够高；kylin 是专门为 olap 分析而生的数据仓库，建模之后分析的效果非常好，但是建模过程久，而且不够灵活，有一定局限性。

#### 5. mpp 类数据仓库：greenplum, vertica, teradata; (推荐)

这一类 NewSQL 数据仓库，兼具大数据量的处理能力和 OLAP 分析支持的能力，对接 FineDirect 引擎拥有最好的效果，而且不同于 kylin，这类数据库不需要建模，提前生产 cube，因此能充分利用 FineBI 分析功能的灵活性，能真正的解放 IT 人员，实现业务人员的自助分析。值得一提的是 Greenplum 的开源，也让这款性能良好的数据库能更广泛的在企业选型中被采用。而 FineBI 直连分析引擎，也搭配 GP 数据仓库方案他成功解决了一个个企业用户的大数据 BI 分析的难题。这类数据仓库的缺陷是对于超大并发支持的可能不够好，但是对于数据分析这种场景这类数据仓库的并发支持也是足够的。

#### 总体建议：

总体来说采用建立数据仓库(DW)来为 FineDirect 提供数据源的方案：

##### 优势：

- 1.经过设计的数据仓库，数据的规范性，易用性，复用性都较高，极大提升用户的查询效率。
- 2.数据仓库能够处理的数据量较大，根据用户的场景，从百万级到亿级，千亿级都有相应的解决方案。

##### 劣势：

1.DW 模型开发量较大，糟糕的数据仓库设计对之后 BI 应用有很大影响。

2.经过层层 ETL 处理，数据的实时性不能保证，更多的是针对历史数据，静态数据的分析。

### 实际客户案例：

某母婴零售：SQLserver2012，单机数据库数据仓库方案，基于执行计划和索引优化的单机数据仓库也能支持其要求苛刻的 OLAP 分析。部分分钟级别的执行场景优化后可以秒级出结果。

某证券企业：FineDirect 直连引擎对接 oracle rac 双节点集群的数据仓库，处理单表最大 10 亿行，20~30 列，整个 BI 系统用户 600，并发数量 50~100。



证券

金融资产统计分析

期初日期: 2016-11-8

统计组件1

营业部号	营业部	营业部类别	融资融券合计(万)	保证金合计(万)	港股总市值(...)	个股期权市值(...)	托管市值合计(...)
- 10155003	-	A	0	0	0	0	0
- 10155004	-		0	0	0	0	0
- 10155005	-		0	0	0	0	0
- 10155006	-		0	0	0	0	0
- 10155007	-		0	0	0	0	0
- 10155008	-		0	0	0	0	0
- 10155009	-		0	0	0	0	0
- 10155010	-		0	0	0	0	0
- 10155011	-		0	0	0	0	0
- 10155012	-		0	0	0	0	0
- 10155013	-		0	0	0	0	0
- 10155014	-		0	0	0	0	0
- 10155015	-		0	0	0	0	0
- 10155016	-		0	0	0	0	0
- 10155017	-		0	0	0	0	0
- 10155018	-		0	0	0	0	0

某市公安局:FineDirect 直连引擎对接 Greenplum 分布式数据仓库，6 个节点，预计数据总量达到 5T，增量数据 1G/每月。当前实际数据总量 100G，增量 600M/每月，用 Greenplum 数据仓库之后之前 oracle 中 600 秒的 sql，执行时间 3 秒以内。数据时效性:T+1。



## 五、方案④ 数据集市 (DM)

数据集市(Data Mart)，也叫数据市场，数据集市就是满足特定的部门或者用户的需求，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。

从范围上来说，数据是从企业范围的数据库、数据仓库，或者是更加专业的数据仓库中抽取出来的。数据中心的重点就在于它迎合了专业用户群体的特殊需求，在分析、内容、表现，以及易用方面。

常用作 DM 层选型的数据库：oracle, sqlserver, mysql 等传统关系型数据库；也有华为 elk 这样的大数据平台作为数据加速的数据集市层。

通常情况下数据集市是在数据仓库的加的一层，一种应用场景是按照部门主题从企业大的数据仓库切分出一块，用来满足该部门的需求，变相解决查询效率和并发问题。另一种应用场景是，从企业的业务库中，抽取该部门需要的数据，来建立该部门的数据集市。

## 总体建议：

总体来说采用建立数据集市(DM)来为 FineDirect 提供数据源的方案：

### 优势：

- 1.根据业务需求，从数据池整合成结果数据集市。能更方便的管理，并专注于主题进行分析。
- 2.将部门数据进行切分，能极大提升查询效率，并且解决多部门访问而给数据仓库带来的并发压力。
- 3.多数情况下，数据集市会存放计算的结果集，因此对于复杂的计算指标，直接展现效率会很高。减少 SQL 会话和事物的处理时间。

### 劣势：

1. 多数数据集市不会再采用高成本的数据库，因此数据集市的展现需要依赖数据库的性能。
- 2.多层 ETL，导致数据的及时性和实时性更难保证，而且有赖于 ETL 的效率。

## 实际客户案例：

某市公安局：oracle 数据集市，用户采用了的是 Greeplum 数据仓库加 oracle 数据集集市的方案，大量的明细数据分析在 Greenplum 中计算，常用的复杂指标，提前处理好存放在 oracle 数据集市中，这样即满足了用户对查询展现时间的要求，有能让资源合理利用。同时因为固定的指标面向的访问人群量大，因此 oracle 数据集市又解决了大并发的的问题。





