接入方案

背景：

数据平台定位为对外提供灵活数据分析服务，支持业务人员通过数据的指导来进行流程改进、业务创新、产品优化等工作。数据平台是企业实现数据驱动，数智化转型的基础平台。

使用数据平台的人员通常为项目PO，业务人员，数据分析师等，他们迫切希望能关注业务发展现状，从数据中去识别问题或者发掘洞见。基于这样的目的，数据分析常常是探索性的，业务分析人员希望能尽量广泛的分析现有业务的方方面面，这就带来了分析需求的千变万化。目前平台需要支持自助式探索性数据分析，定期输出指标数据，对接BI系统进行报表展示等。

接入方案：

当前的数据平台采用T+1的时效性进行数据接入，暂无实时数据接入需求。

一般而言，数据接入分为两类：

1. 全量数据：仅需要第一次数据接入时，可能有大量数据交换
2. 增量数据：每天定时读取前一天的数据，数据量一般较少

基于我们对各个业务系统的理解，数据平台连接业务数据的方案总共可以有4种（按照优先选择顺序排列）：

ID	方案说明	建议指数	优点	风险点
方案一 **直联数据库**	1. 业务方提供数据库（推荐从库）的只读账号供数据平台获取数据 2. 获取数据：协商一个业务系统负载较低的时间段进行	√√√ √√	1. 直联业务系统数据库，只需要只读权限； 2. 提高跨团队沟通效率； 3. 业务方无需开发工作，节约成本； 4. 没有中间数据转换过程，不容易引起问题 5. 可以快速适应数据需求变更 6. 业界的通用实践，高效而稳定	1. 对业务系统负载有一定的影响（对策：使用业务系统低负载的时间运行） 2. 业务系统数据结构变更可能带来数据平台数据接入任务失败（对策：数据平台通过监测机制实现对应，业务系统无操作）
方案二 数据库Dump	1. 业务方不提供数据库的读取账号，而是利用数据库导出数据的工具（如mysqldump）定期将数据导出存放于某一个共享目录 2. 获取数据： a. 全量：业务方做一次全量数据导出 b. 增量：业务方实现定时任务，每天晚上定期导出前一天的数据	√√	1. 没有中间数据转换过程，不容易引起问题 2. 可以快速适应数据需求变更业务 3. 系统需要少量的开发维护工作	1. 需要业务系统做一定的开发和运维成本及开发的时间成本 2. 跨团队协作成本高（需要沟通确定共享目录地址等） 3. 增加额外的出错可能性（比如某一天业务系统导出数据任务失败），需要做更多的容错性处理 4. 需要数据平台搭建与业务方相同的数据库来先做一次数据的导入，然后再从数据库中接入到数据平台，带来额外的成本 5. 业务系统数据结构变更可能带来数据平台数据接入任务失败（对策：数据平台通过监测机制实现对应，业务系统无操作）
方案三 开发通用接口	1. 业务方不提供数据库的读取账号，但是提供一个通用的接口导出数据，由数据平台提供配置（需要导出的库、表），业务方实现接口定期将数据导出为一个中间格式（如csv）存放于某一个共享目录 2. 获取数据： a. 全量：业务方根据数据平台的配置做一次全量数据导出 b. 增量：业务方根据数据平台的配置每天晚上定期导出前一天的数据	√	1. 可以较快速适应数据需求变更业务 2. 接口相对稳定	1. 需要业务系统做较多的开发和运维，及较多的时间成本 2. 跨团队协作成本高（需要沟通确定共享目录地址等） 3. 业务系统数据结构变更可能带来数据平台数据接入任务失败（对策：可以通过一定的监测机制实现，由数据平台来负责处理，业务系统无需关心） 4. 存在中间数据转换过程，增加更多的额外出错可能性（比如某一天业务系统导出数据任务失败，比如csv格式中的null空字符串带来问题等），需要做更多的容错性处理
方案四 开发定制接口	1. 业务方不提供数据库的读取账号，根据数据平台的数据需求来开发特定接口 2. 获取数据： a. 全量：业务方根据需求做一次全量数据导出 b. 增量：业务方根据需求每天晚上定期导出前一天的数据	不推荐	定制接口相对稳定，不会由于业务数据结构变化而带来数据平台的额外工作	1. 需要业务系统做大量的开发和运维，及大量的时间成本 2. 跨团队协作成本很高； 3. 新增数据接入，接口需要重新开发，无法应对数据分析需求的多变性 4. 存在中间数据转换过程，增加更多的额外出错可能性（比如某一天业务系统导出数据任务失败，比如csv格式中的null空字符串带来问题等），需要做更多的容错性处理 5. 每次需求变更，每个接口都需要重新对接数据需求，重新开发上线

数据平台本身是对外提供灵活数据分析服务的，业务需求会经常变，如果变了，就要重新核对需求；因此方案四不建议使用。

方案二和方案三本身会增加业务系统和数据平台的工作量，因此也不推荐。

建议使用方案一

PS：如果业务系统无法支持方案一，需要确认原因以及是否有解决方案，综合考量各个因素进行最终方案的选择。

一些业务系统关心的问题及应对：

" dump文件中会有数据库的用户/密码信息，数据安全性上是否可行？
一般的系统设计，在安全性上会禁止数据库中的以明文的形式存储密码。一个参考的方式是加盐hash，这可以较大程度上提升安全性。这样一来，即便有数据库信息也是无法反向计算用户密码的。另外，数据平台这边有完善的数据访问权限控制及脱敏规则支持，这也给数据安全提供了保障。如果还有安全性的担心，我们可以考虑在导出时，将这些数据设置为空值。

" 需要dump的数据量是否能够在我们服务器的空闲时间内完成？
这个需要SCRM这边进行评估。一般而言，第一次全量数据导出，会花费的时间较长，不过可以考虑分库分表导出（如每次只导出一部分）。每日的增量数据，由于数据量小，可以很快完成，一般不会成为问题。

" 另外一旦连携中出现我们dump出的文件你们导入出现问题，该如何调查原因？比如需要我们再准备环境测试导入协助调查吗？
这个问题数据平台团队会主要负责处理。我们这边一般的做法是会搭建一个与业务系统同样版本的数据库，然后进行数据导入，再将数据从这个中间数据库导入到数据平台，所以这个环节出问题的可能性不大。