# 数据质量管理

随着业务发展和数据量的增加，大数据应用开发已成为部门应用开发常用的开发方式，由于部门业务特点的关系，spark和hive应用开发在部门内部较为常见。当处理的数据量达到一定量级和系统的复杂度上升时，数据的唯一性、完整性、一致性等等校验就开始受到关注，而通常做法是根据业务特点，额外开发job如报表或者检查任务，这样会比较费时费力。

目前遇到的表大部分在几亿到几十亿的数据量之间，并且报表数量在不断增加，在这种情况下，一个可配置、可视化、可监控的数据质量工具就显得尤为重要了。Griffin 数据质量监控工具正是可以解决前面描述的数据质量问题的开源解决方案。

# Apache Griffin

Griffin起源于eBay中国，并于2016年12月进入Apache孵化器，Apache软件基金会2018年12月12日正式宣布Apache Griffin毕业成为Apache顶级项目。

Griffin是属于模型驱动的方案，基于目标数据集合或者源数据集(基准数据)，用户可以选择不同的数据质量维度来执行目标数据质量的验证。支持两种类型的数据源：batch数据和streaming数据。对于batch数据，我们可以通过数据连接器从Hadoop平台收集数据。对于streaming数据，我们可以连接到诸如Kafka之类的消息系统来做近似实时数据分析。在拿到数据之后，模型引擎将在spark集群中计算数据质量。

## 特性

* 度量：精确度、完整性、及时性、唯一性、有效性、一致性。
* 异常监测：利用预先设定的规则，检测出不符合预期的数据，提供不符合规则数据的下载。
* 异常告警：通过邮件或门户报告数据质量问题。
* 可视化监测：利用控制面板来展现数据质量的状态。
* 实时性：可以实时进行数据质量检测，能够及时发现问题。
* 可扩展性：可用于多个数据系统仓库的数据校验。
* 可伸缩性：工作在大数据量的环境中，目前运行的数据量约1.2PB(eBay环境)。
* 自助服务：Griffin提供了一个简洁易用的用户界面，可以管理数据资产和数据质量规则；同时用户可以通过控制面板查看数据质量结果和自定义显示内容。

### 数据质量指标说明

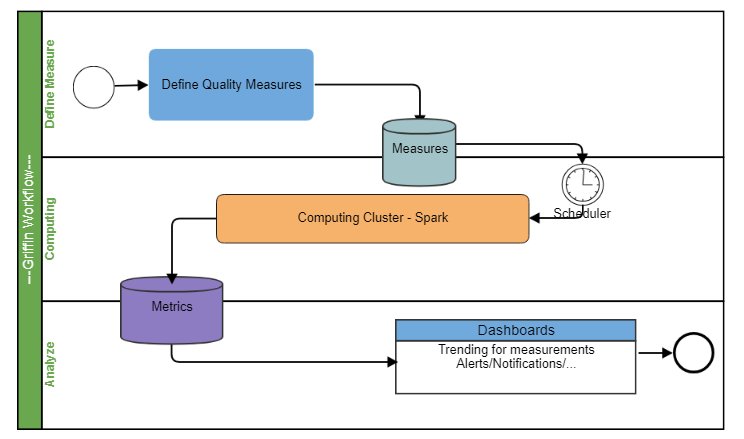
* 精确度：度量数据是否与指定的目标值匹配，如金额的校验，校验成功的记录与总记录数的比值。
* 完整性：度量数据是否缺失，包括记录数缺失、字段缺失，属性缺失。
* 及时性：度量数据达到指定目标的时效性。
* 唯一性：度量数据记录是否重复，属性是否重复；常见度量为hive表主键值是否重复。
* 有效性：度量数据是否符合约定的类型、格式和数据范围等规则。
* 一致性：度量数据是否符合业务逻辑，针对记录间的逻辑的校验，如：pv一定是大于uv的，订单金额加上各种优惠之后的价格一定是大于等于0的。

## 优势

* 可配置、可自定义的数据质量验证。
* 基于spark的数据分析，可以快速计算数据校验结果。
* 历史数据质量趋势可视化。

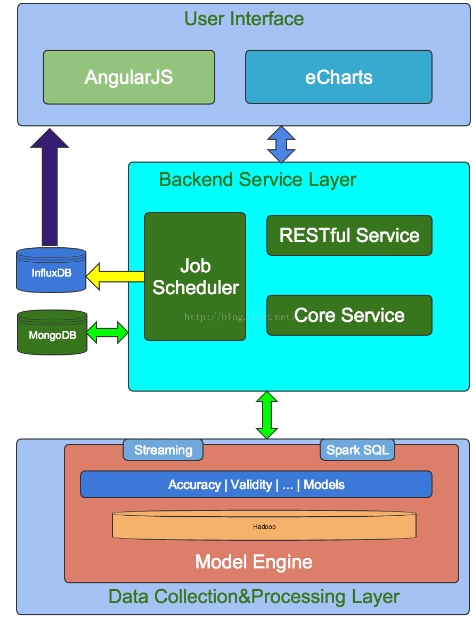
## 工作流程

1. 注册数据，把想要检测数据质量的数据源注册到griffin。
2. 配置度量模型，可以从数据质量维度来定义模型，如：精确度、完整性、及时性、唯一性等。
3. 配置定时任务提交spark集群，定时检查数据。
4. 在门户界面上查看指标，分析数据质量校验结果。

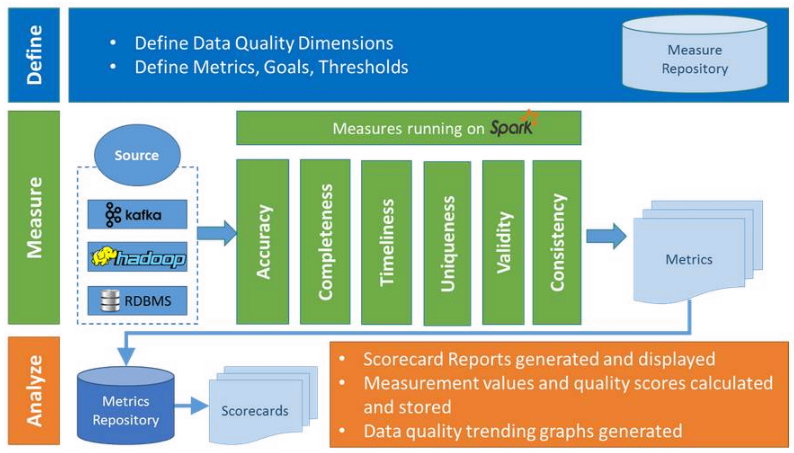


## 系统架构

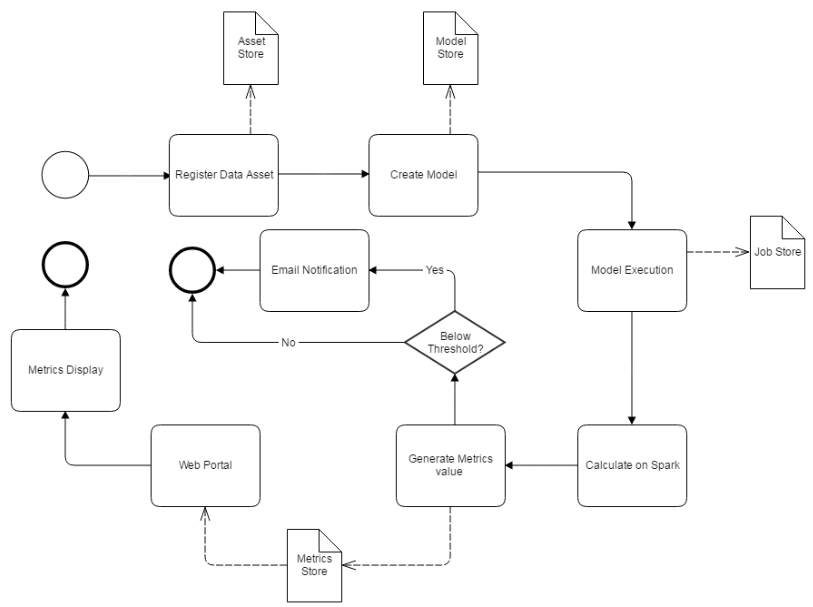
Griffin 系统主要分为：数据收集处理层（Data Collection&Processing Layer）、后端服务层（Backend Service Layer）和用户界面（User Interface），如图：



系统数据处理分层结构图：

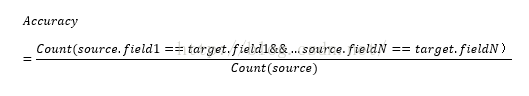


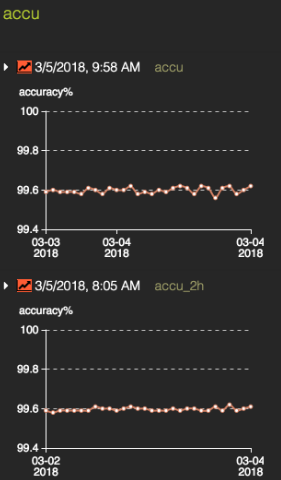
系统处理流程图：



## 数据验证逻辑

### 精确度验证(accurancy),从hive metadata中加载数据源，校验精确度

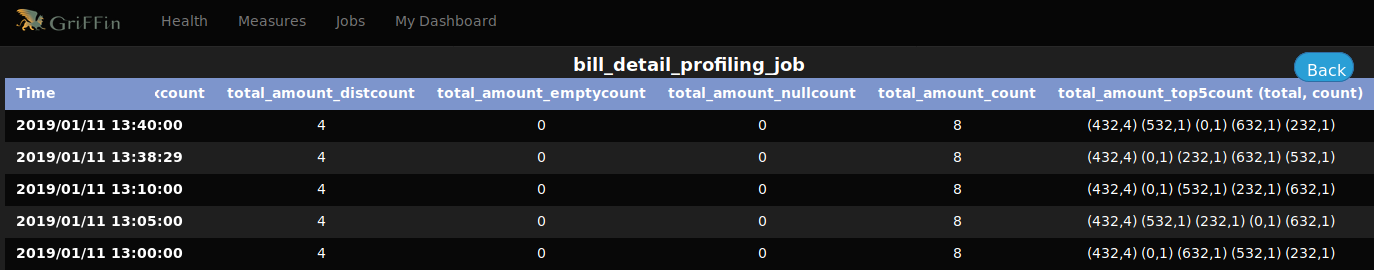
1. 选择source表及列
2. 选择target表及列
3. 选择字段比较规则（大于、小于或者相等）
4. 通过一个公式计算出结果
5. 最后在控制面板查看精确度趋势



### 数据统计分析(profiling)

* 选择需要进行分析的数据源，配置字段等信息。
* 简单的数据统计：用来统计表的特定列里面值为空、唯一或是重复的数量。例如统计字段值空值记录数超过指定一点阈值，则可能存在数据丢失的情况。
* 汇总统计:用来统计最大值、最小值、平均数、中值等。例如统计年龄列的最大值最小值判断是否存在数据异常。
* 高级统计:用正则表达式来对数据的频率和模式进行分析。例如邮箱字段的格式验证，指定规则的数据验证。
* 数据分析机制主要是基于Spark的MLlib提供的列汇总统计功能，它对所有列的类型统计只计算一次。

控制面板分析数据

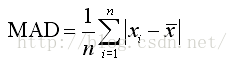


### 异常检测

异常检测的目标是从看似正常的数据中发现异常情况，是一个检测数据质量问题的重要工具。通过使用BollingerBands和MAD算法来实现异常检测功能，可以发现数据集中那些远远不符合预期的数据。

以MAD作为例子，一个数据集的MAD值反映的是每个数据点与均值之间的距离。可以通过以下步骤来得到MAD值：

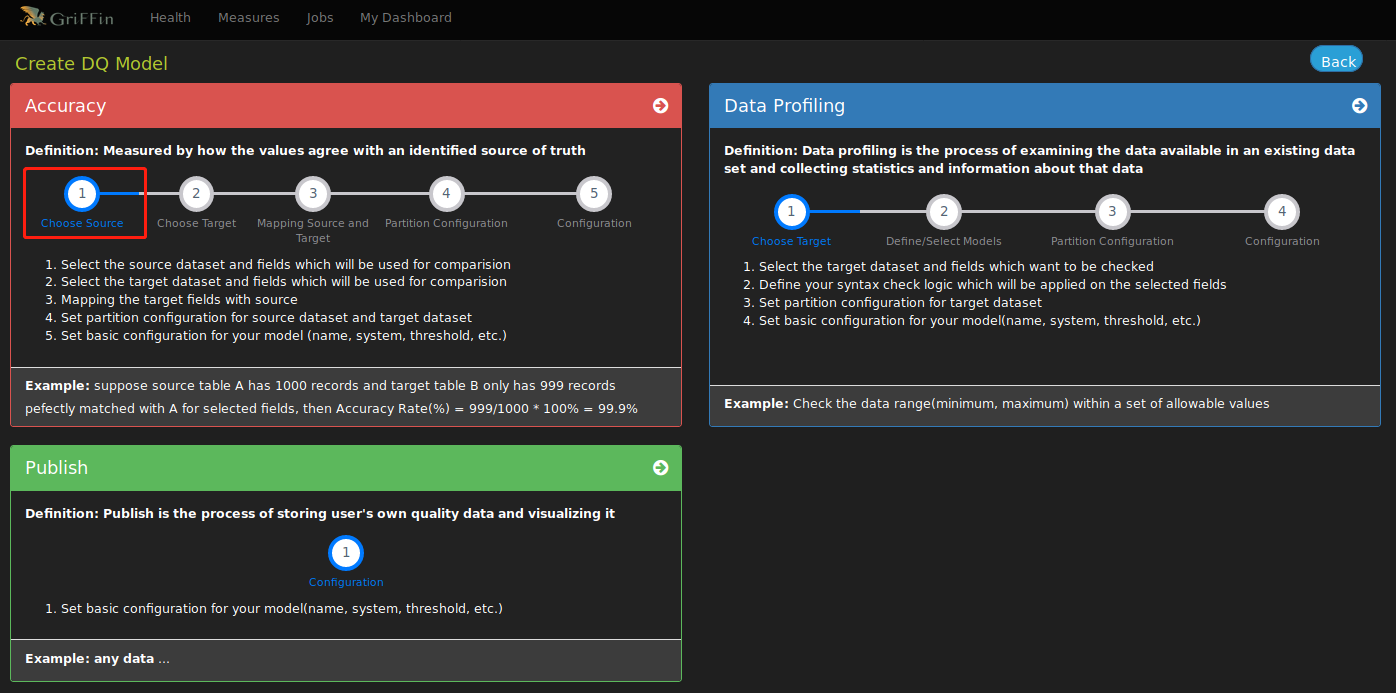
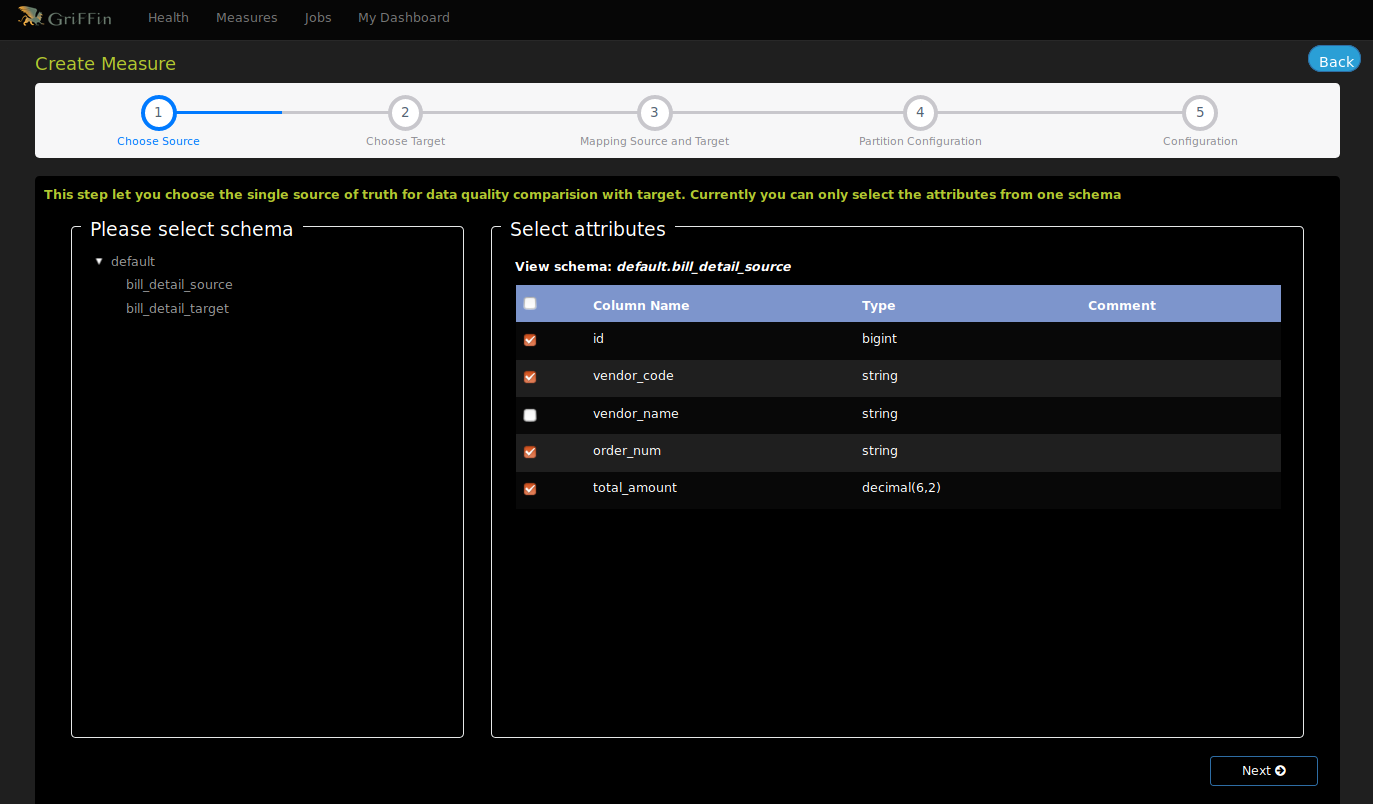
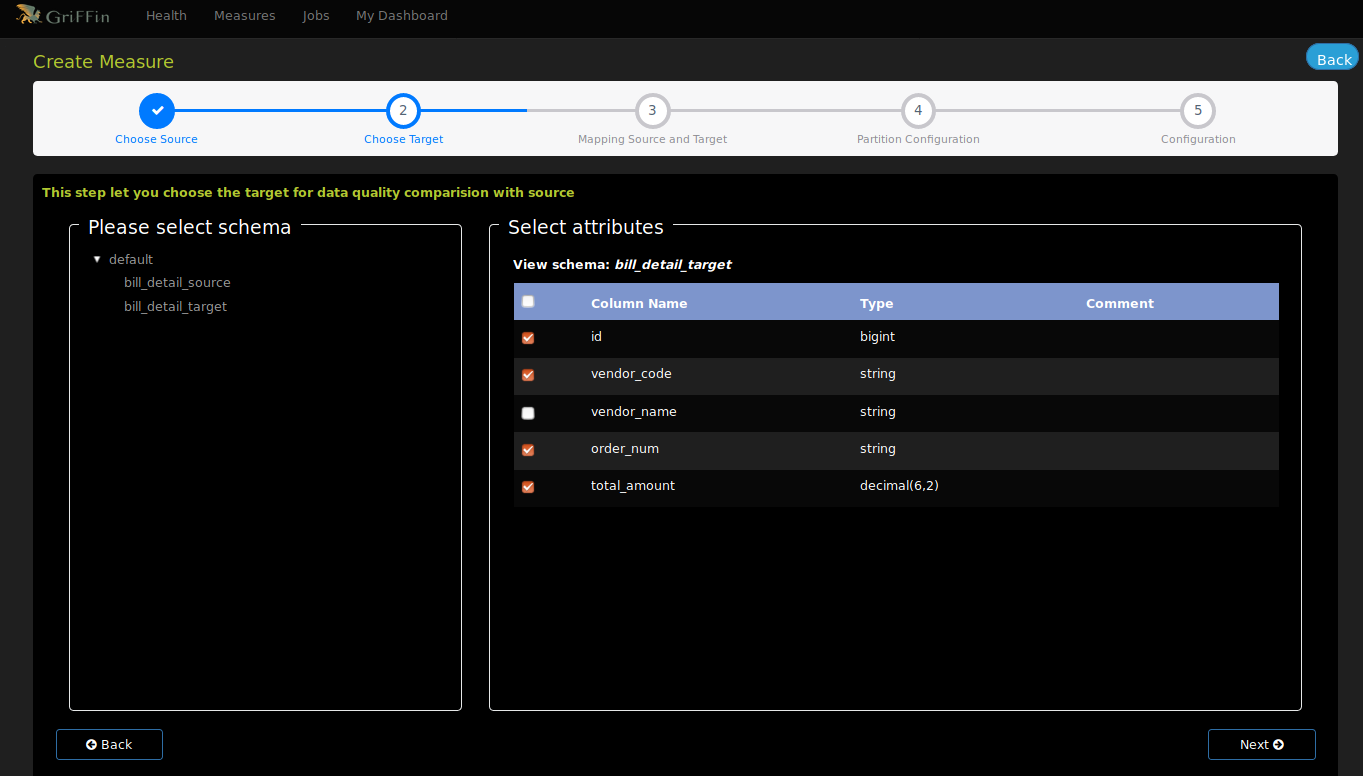
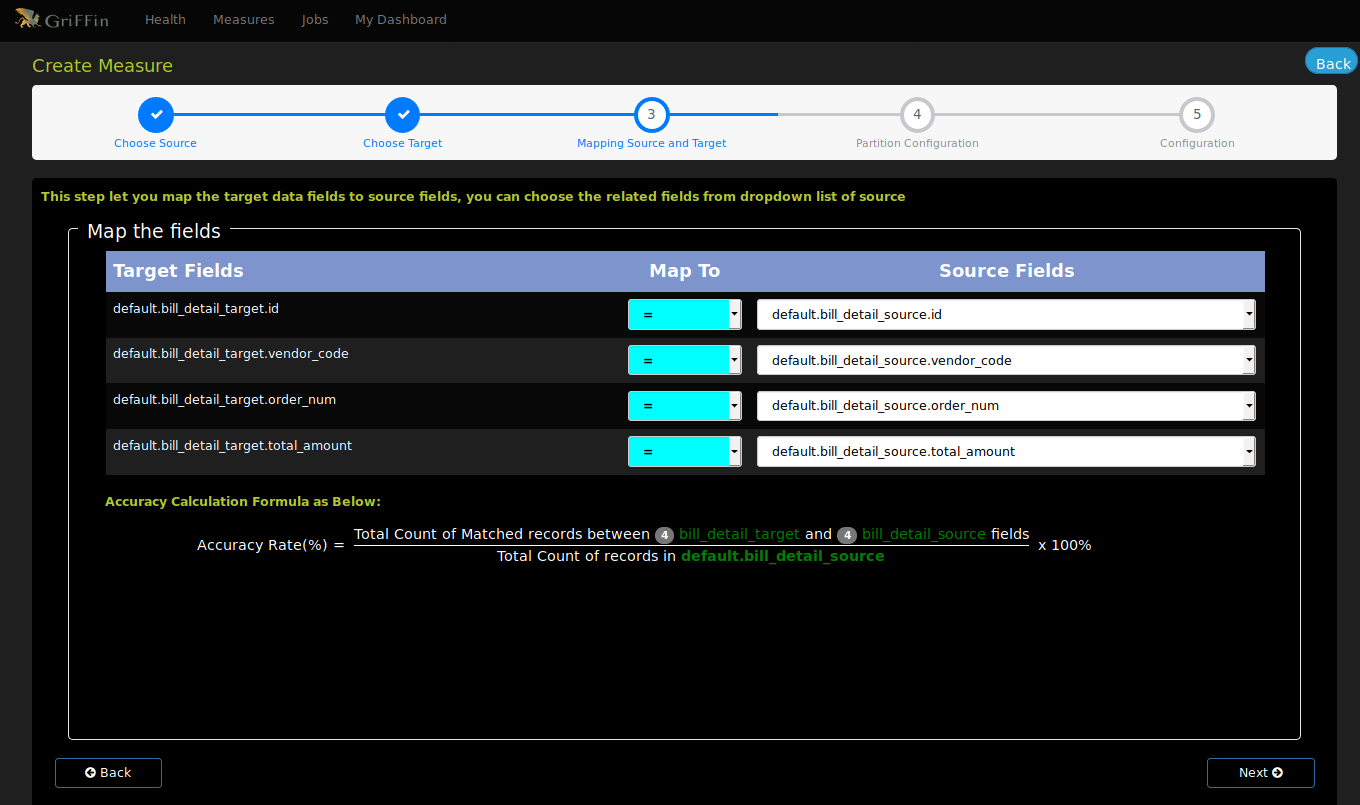
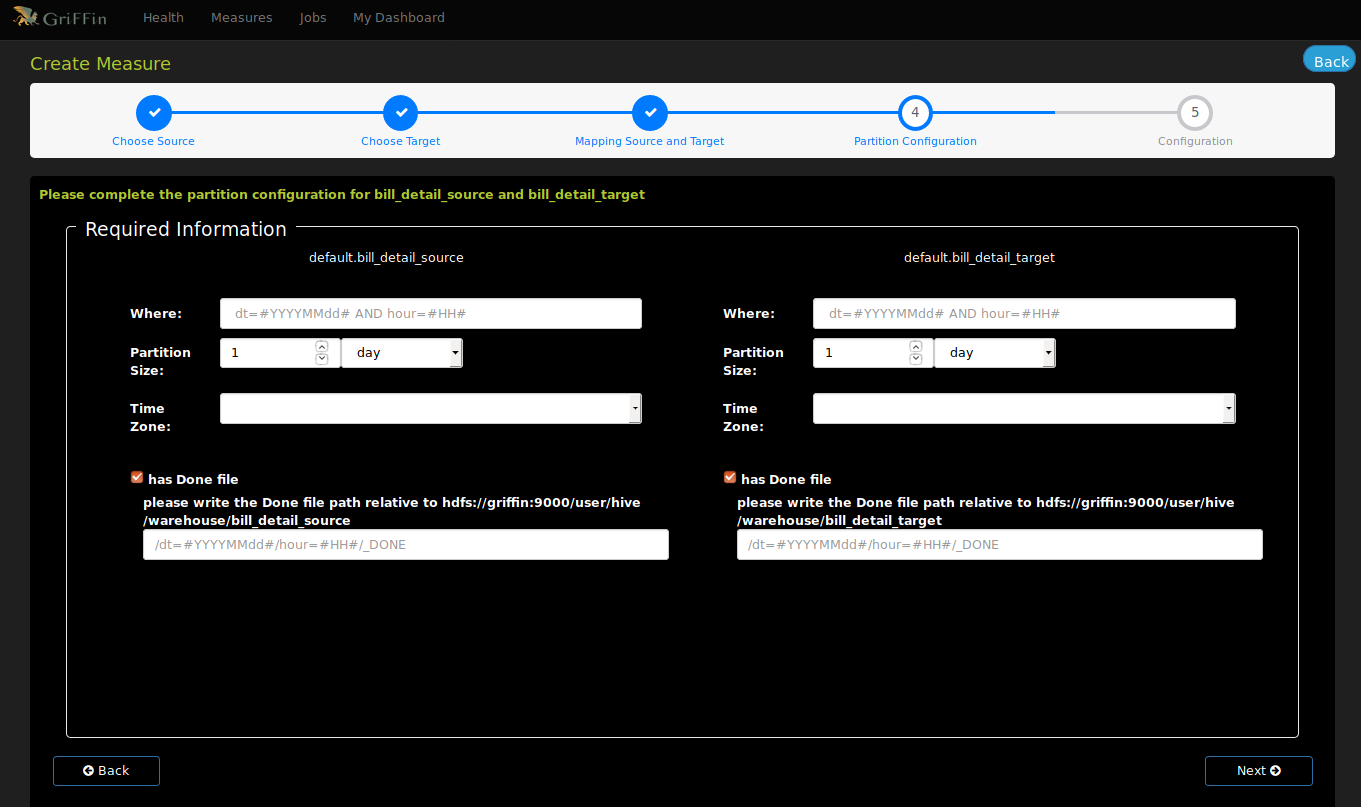
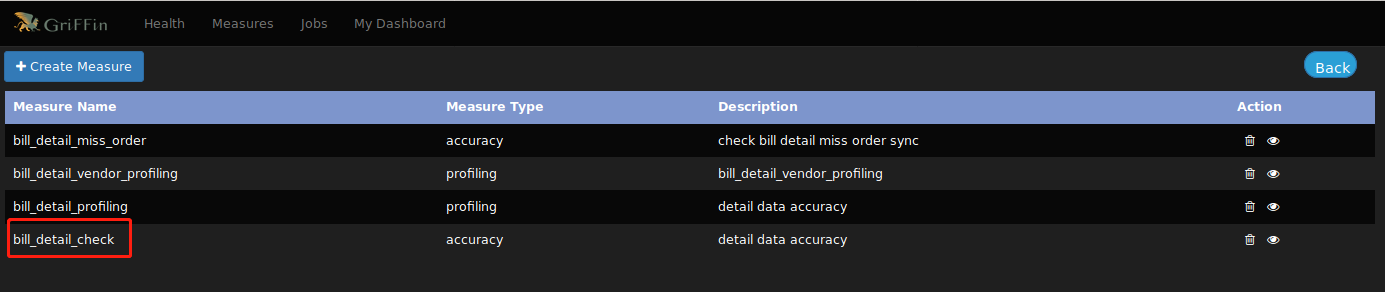
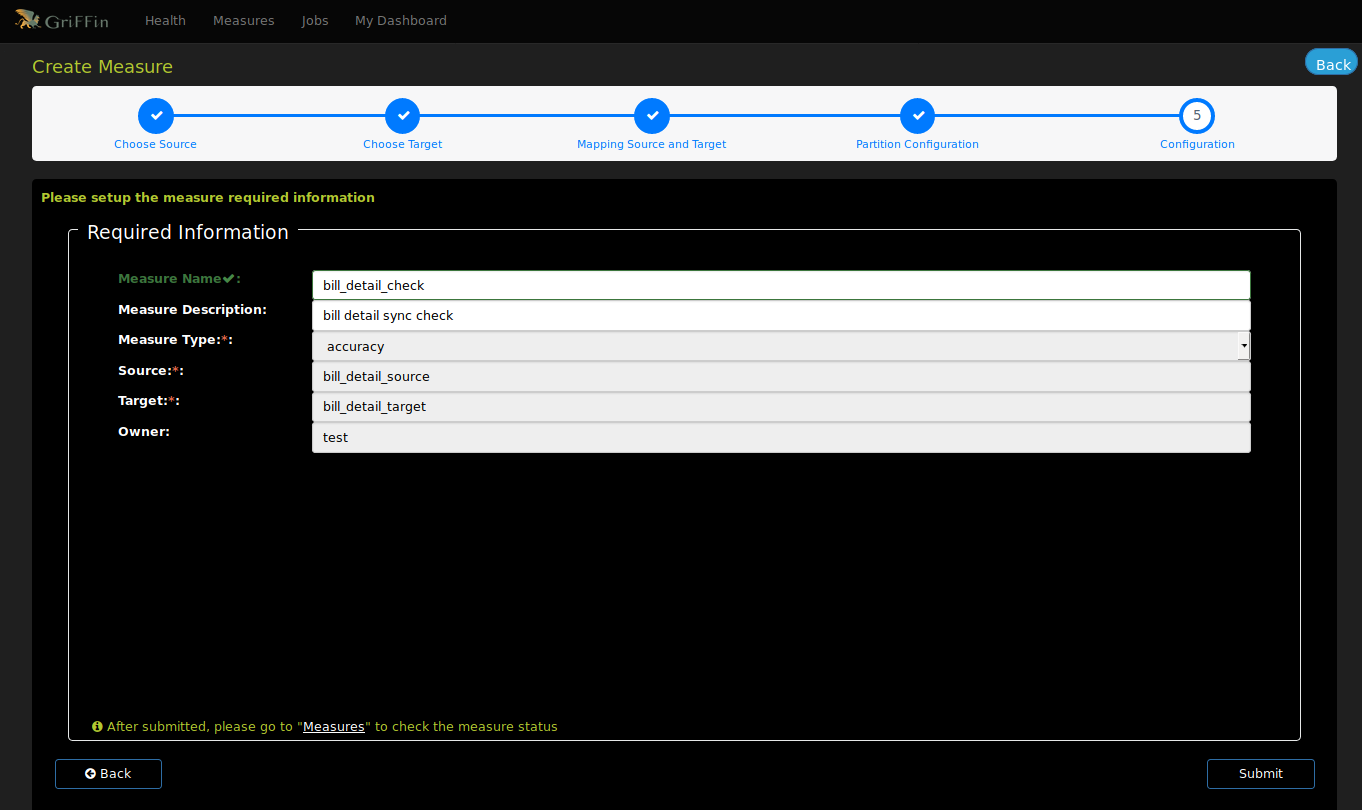
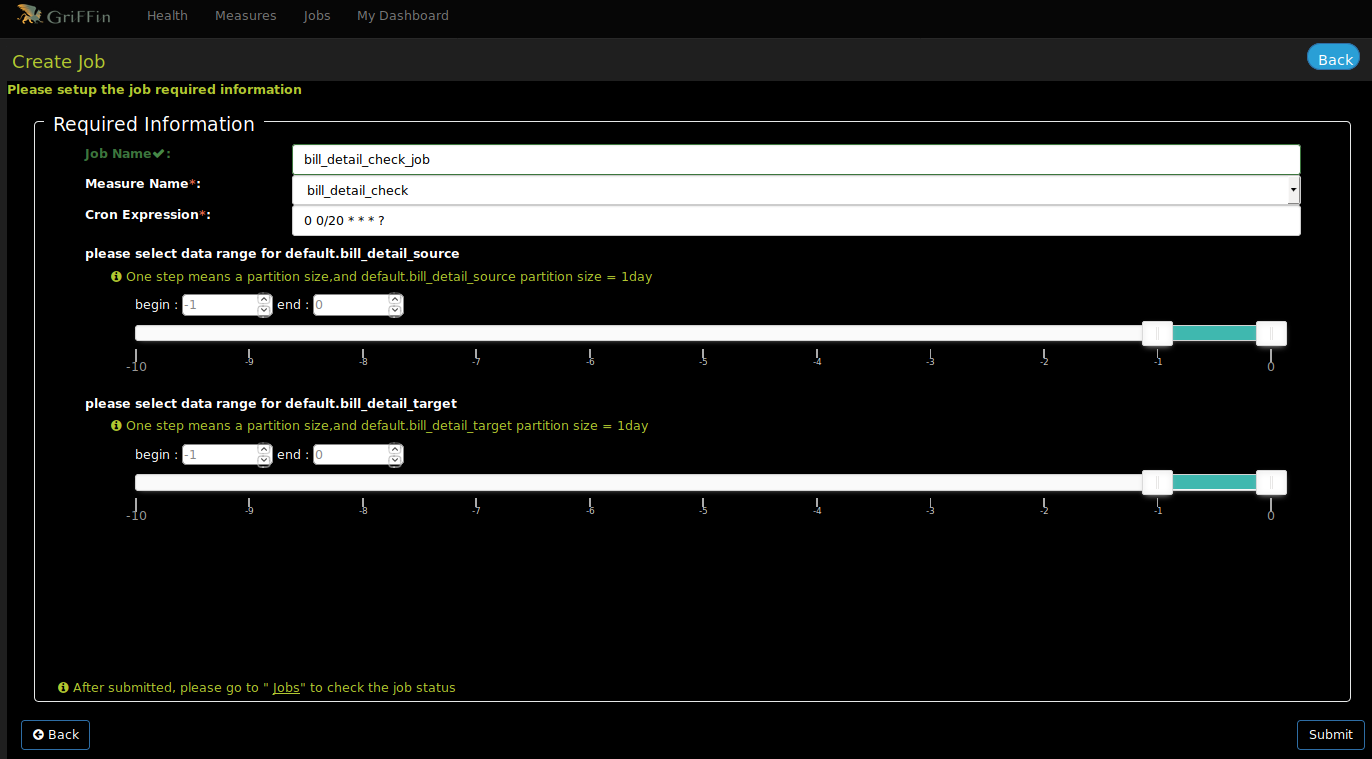
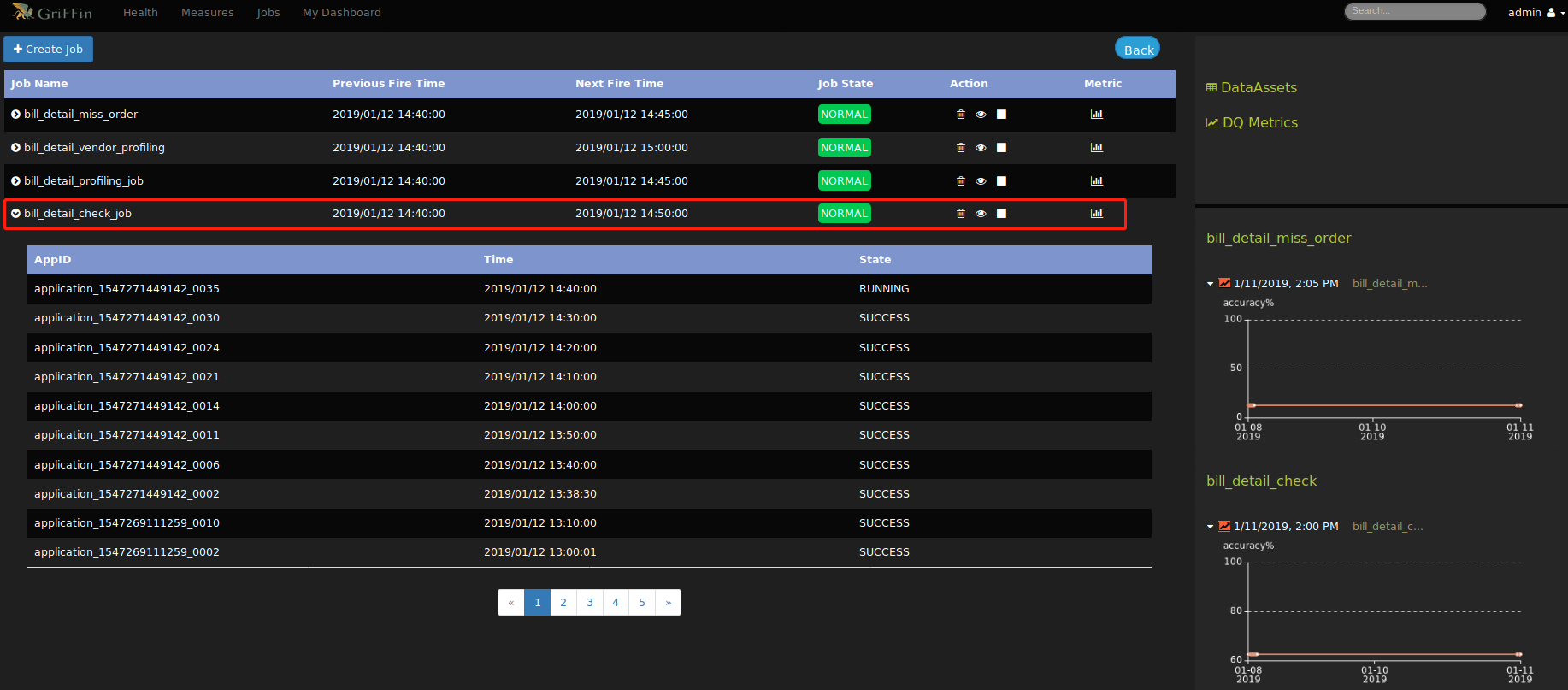
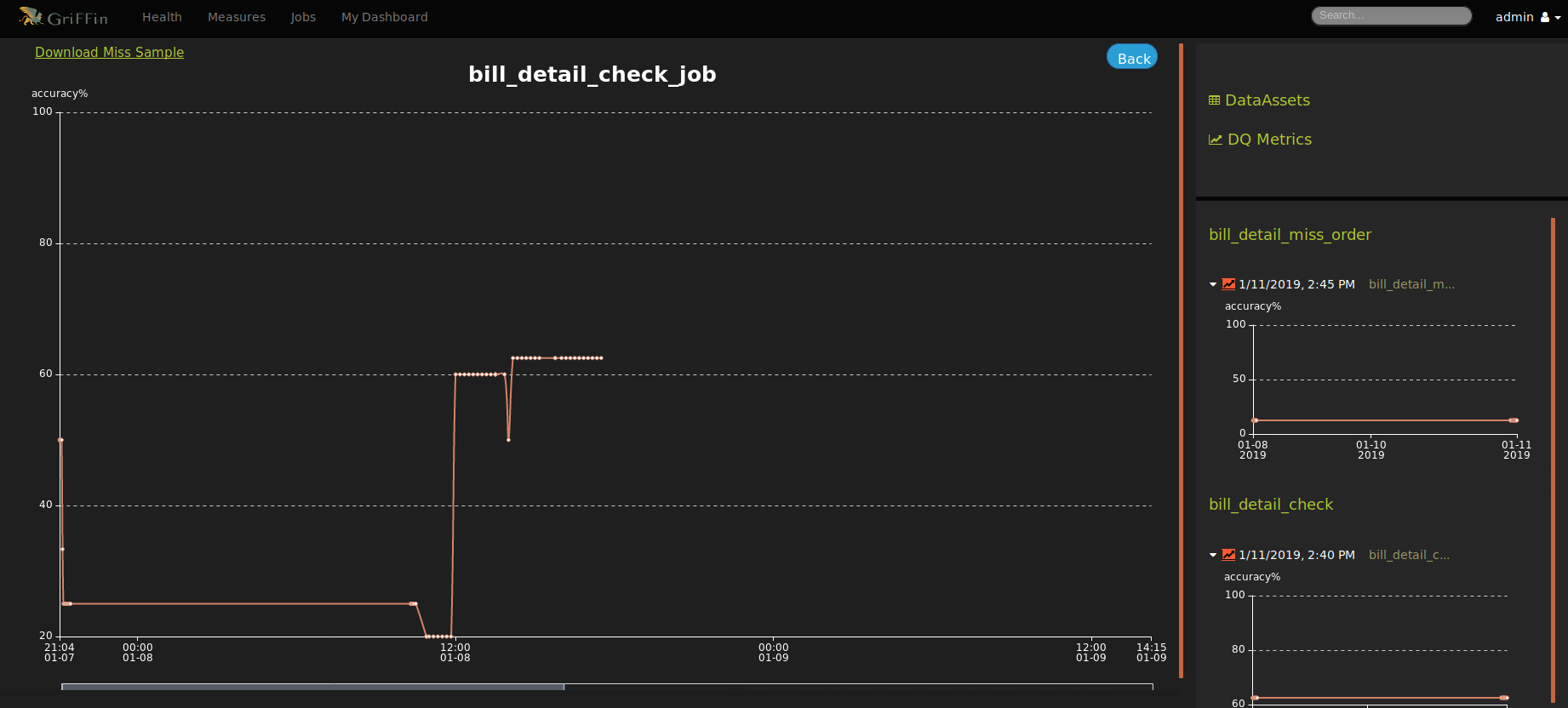
1. 算出平均值
2. 算出每一个数据点与均值的差
3. 对差值取绝对值
4. 算出这些差值取绝对值之后的平均值
5. 公式如下：



通过异常检测可以发现数据值的波动大小是否符合预期，数据的预期值则是在对历史趋势的分析中得来的，用户可以根据检测到的异常来调整算法中必要的参数，让异常检测更贴近需求。

## Demo

以检测供应商账单明细表的同步精确度为例，配置数据检测，如图：

1. 选择数据源
2. 选择账单明细源表字段
3. 选择账单明细目标表字段
4. 设置源表和目标表的校验字段映射关系
5. 选择数据分区、条件和是否输出结果文件。（无分区表可以跳过）
6. 设置验证项目名称和描述，提交后就可以在列表看到度量的信息了
7. 创建了数据模型度量后，需要相应的spark定时任务来执行分析，接下来就是创建spark job和调度信息了。在job菜单下，选择Create Job 
8. 创建job界面中需要选择源表和目标表数据范围，如上图所示是选择t-1到当前的数据分区，即昨天的数据分区。设置定时表达式，提交任务后即可在job列表中查看：
9. 到这里，数据验证度量和分析任务都已配置完成，后面还可根据你的指标设置邮件告警等监控信息，接下来就可以在控制面板上监控你的数据质量了，如图：

## 后台提交监控任务

除了用户在控制面板创建数据验证任务，也可以通过后台生成指标信息，提交spark任务进行数据检测，提供了良好的数据接入和配置的扩展性，api配置数据检测可查看官网快速指引。

实时数据检测目前未有界面配置，可以通过api的方式提交实时数据监控，详细内容可以参考：Streaming Use Cases。