Grid Dynamics

SQA DAYS#28

Автоматизация контроля качества данных

Лянгузов Алексей
QA Architect

April 2021

Grid Dynamics

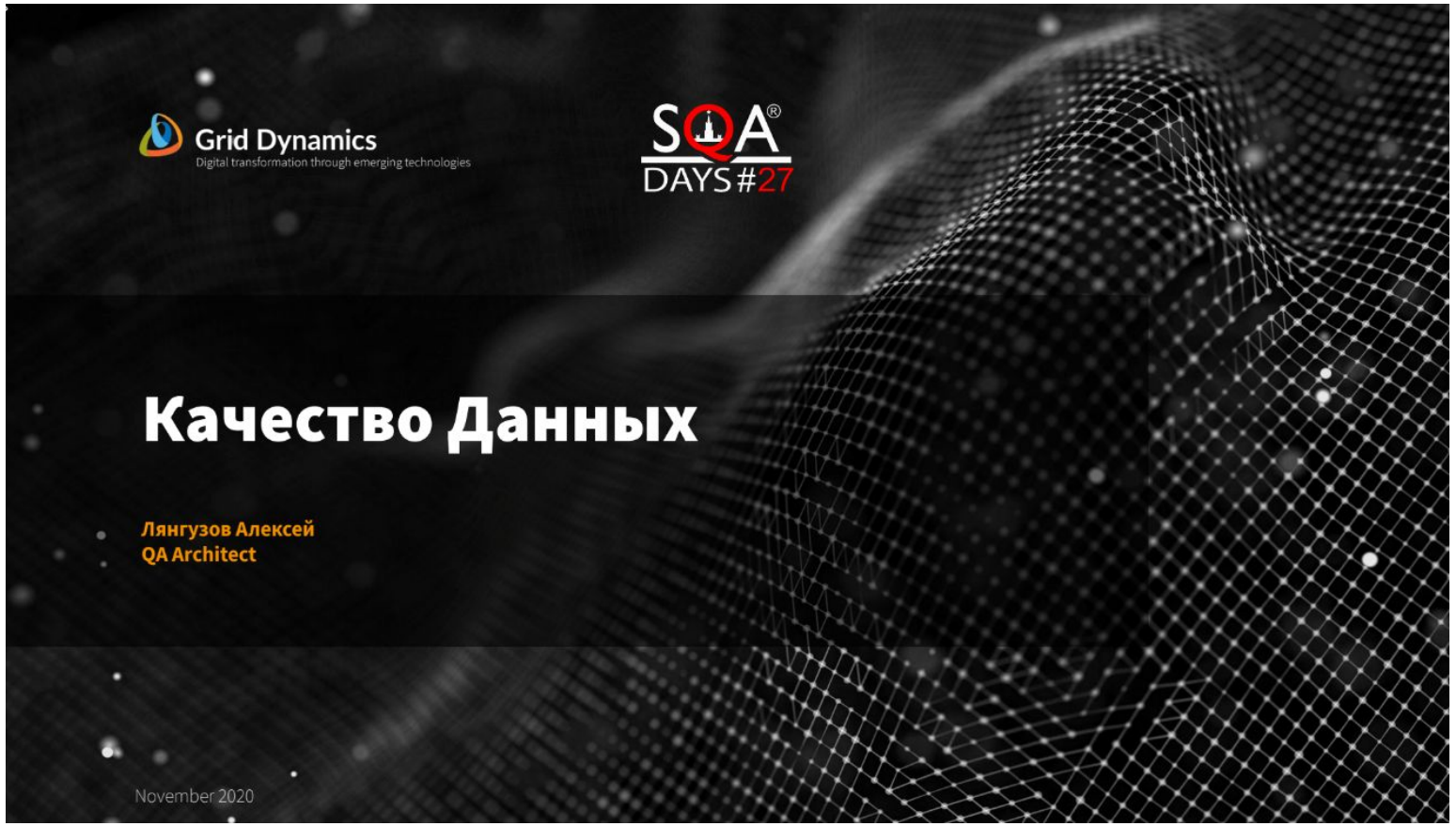
SQA DAYS #28

Автоматизация мониторинга качества данных

Лянгузов Алексей
QA Architect

April 2021

SQA Days 27


Grid Dynamics
Digital transformation through emerging technologies
SQA DAYS#27
Качество Данных
Лянгузов Алексей
QA Architect
November 2020

План

- Сложность задачи контроля качества данных
- Архитектура решения регулярного мониторинга данных
- Плюсы решения
- Примеры

В результате доклада будет понятно, *сложно ли* организовать автоматизированный подход мониторинга качества данных и что для этого нужно.

Сложность задачи контроля качества данных

От чего зависит сложность задачи проверки данных?

Сложность задачи контроля качества данных

От чего зависит сложность задачи проверки данных?

1. От сложности самих данных

Сложность задачи контроля качества данных

От чего зависит сложность задачи проверки данных?

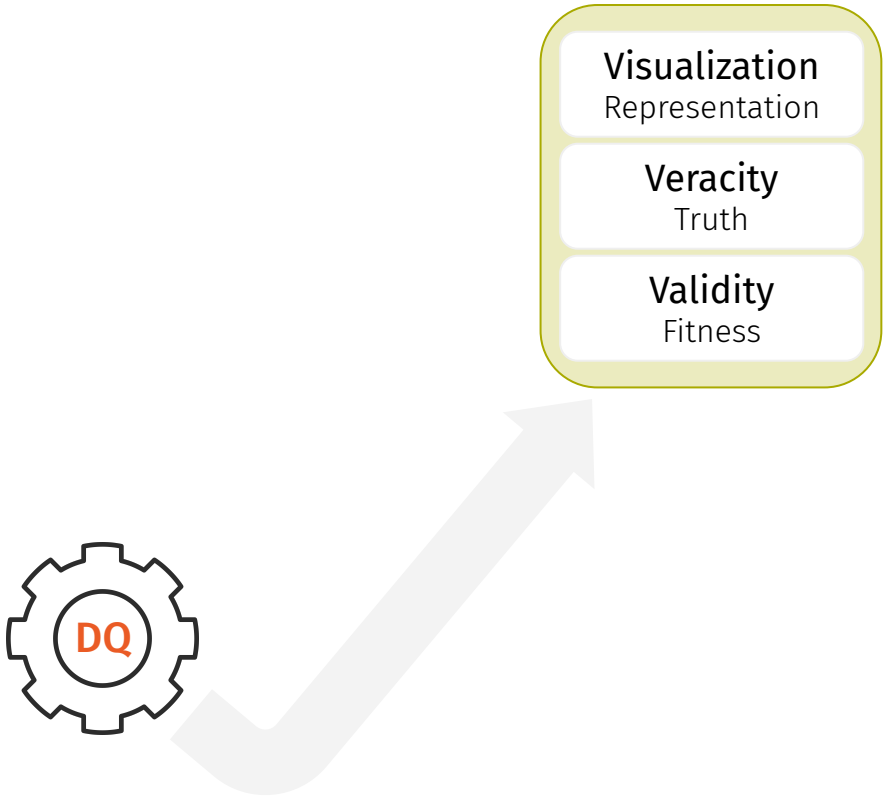
1. От сложности самих данных
2. От глубины и точности проверок

Сложность задачи контроля качества данных

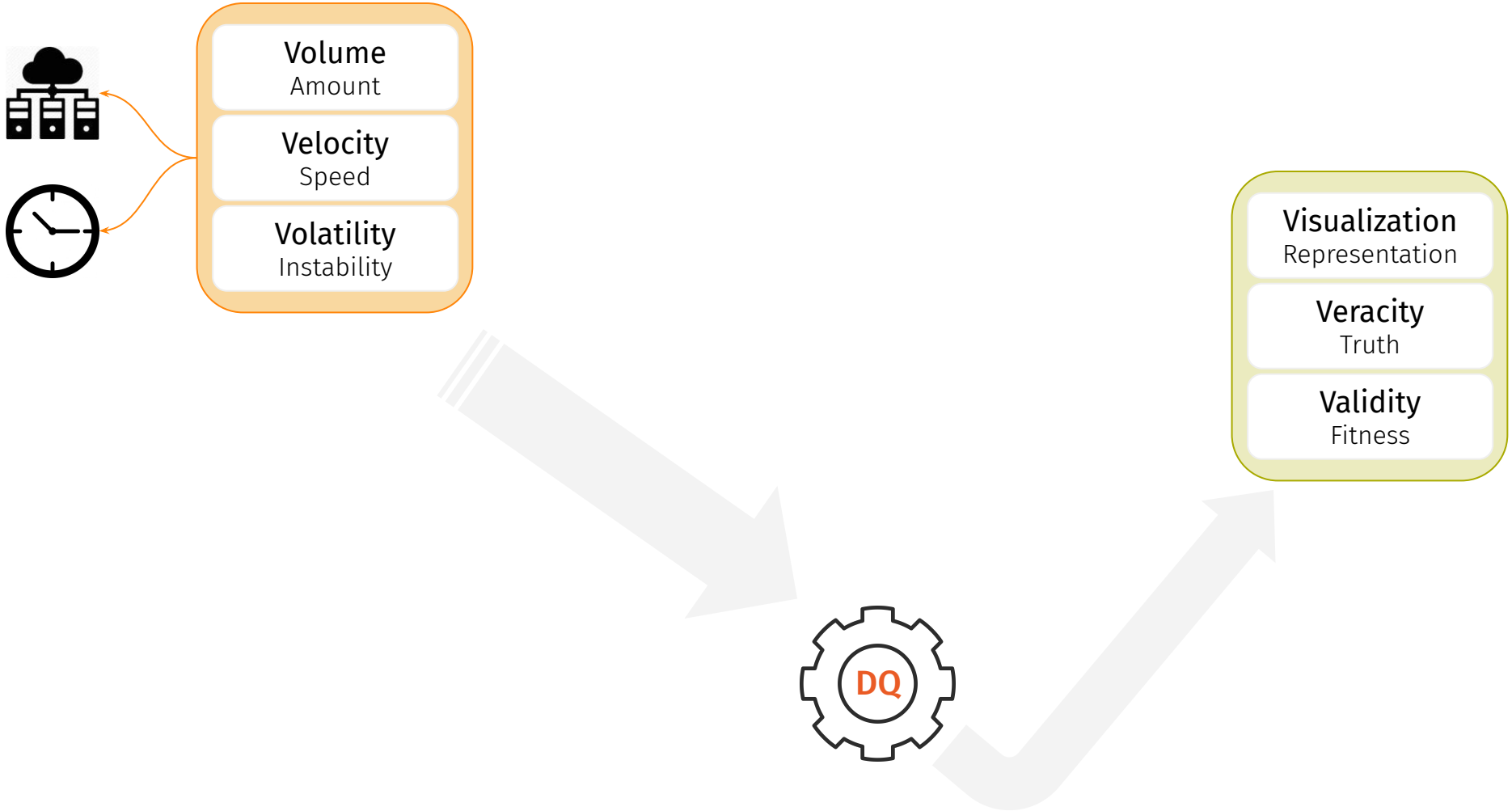
От чего зависит сложность задачи проверки данных?

1. От сложности самих данных
2. От глубины и точности проверок
3. От количества и скорости проверок

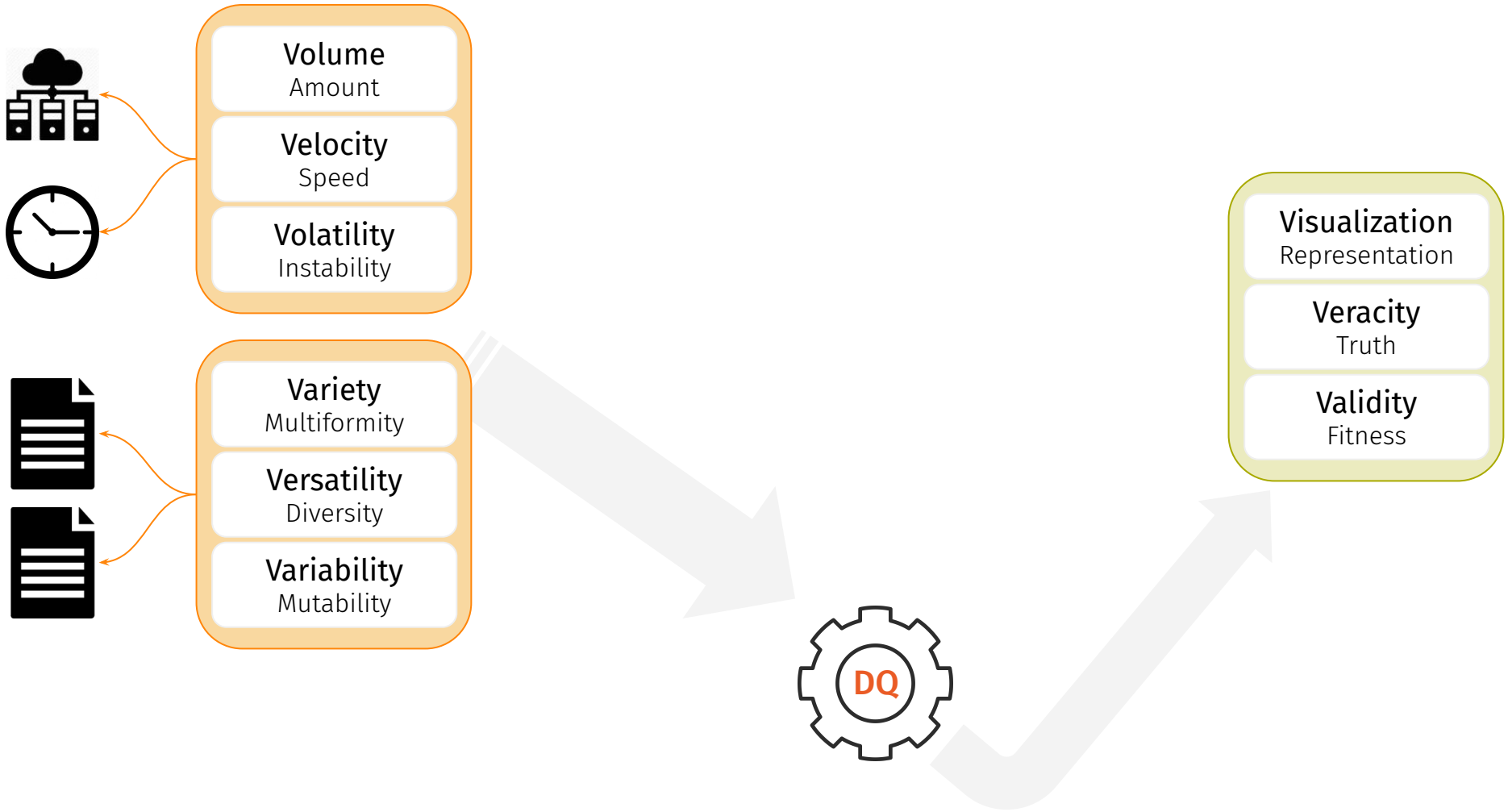
Факторы данных, усложняющие контроль качества



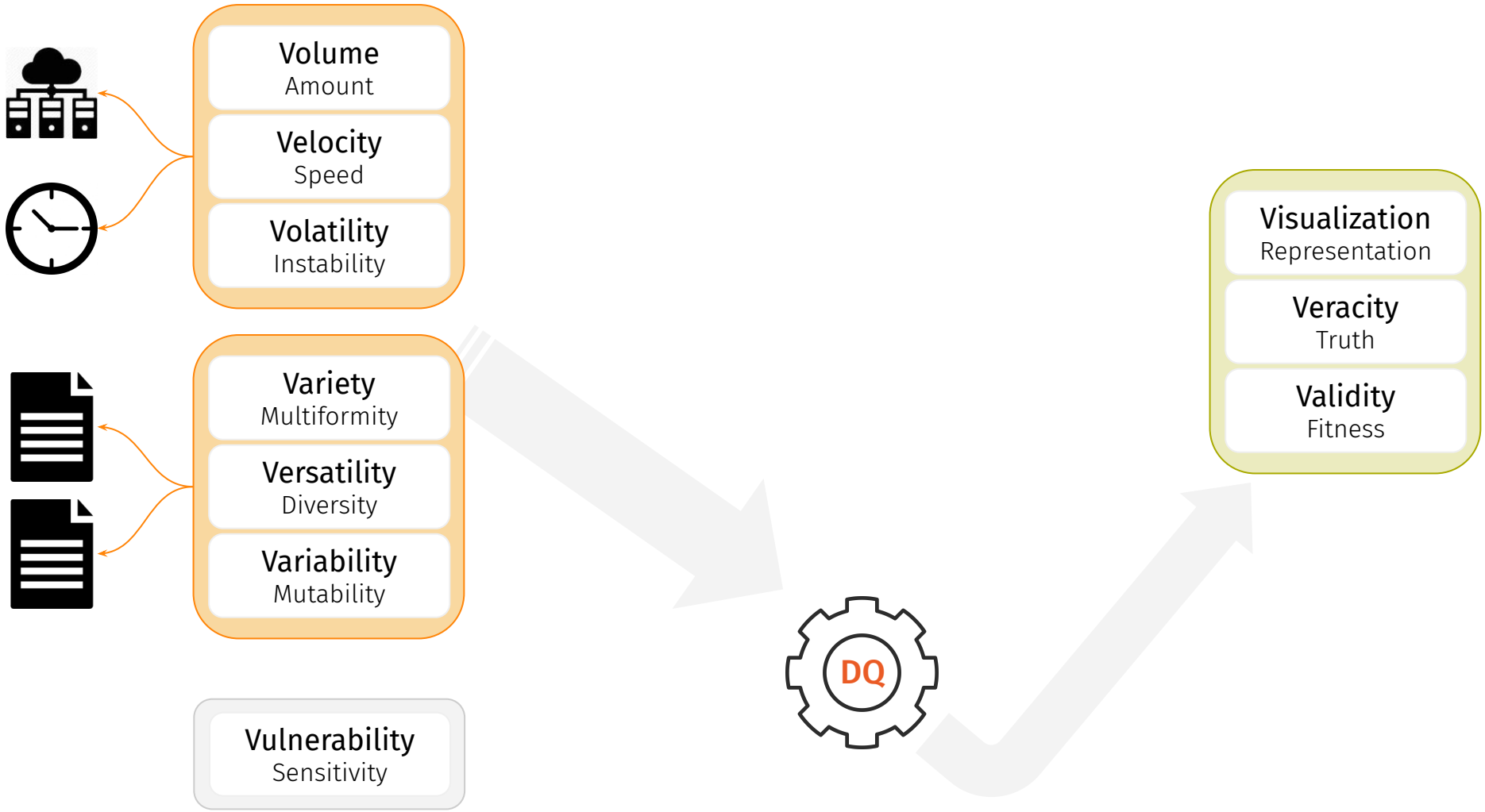
Факторы данных, усложняющие контроль качества



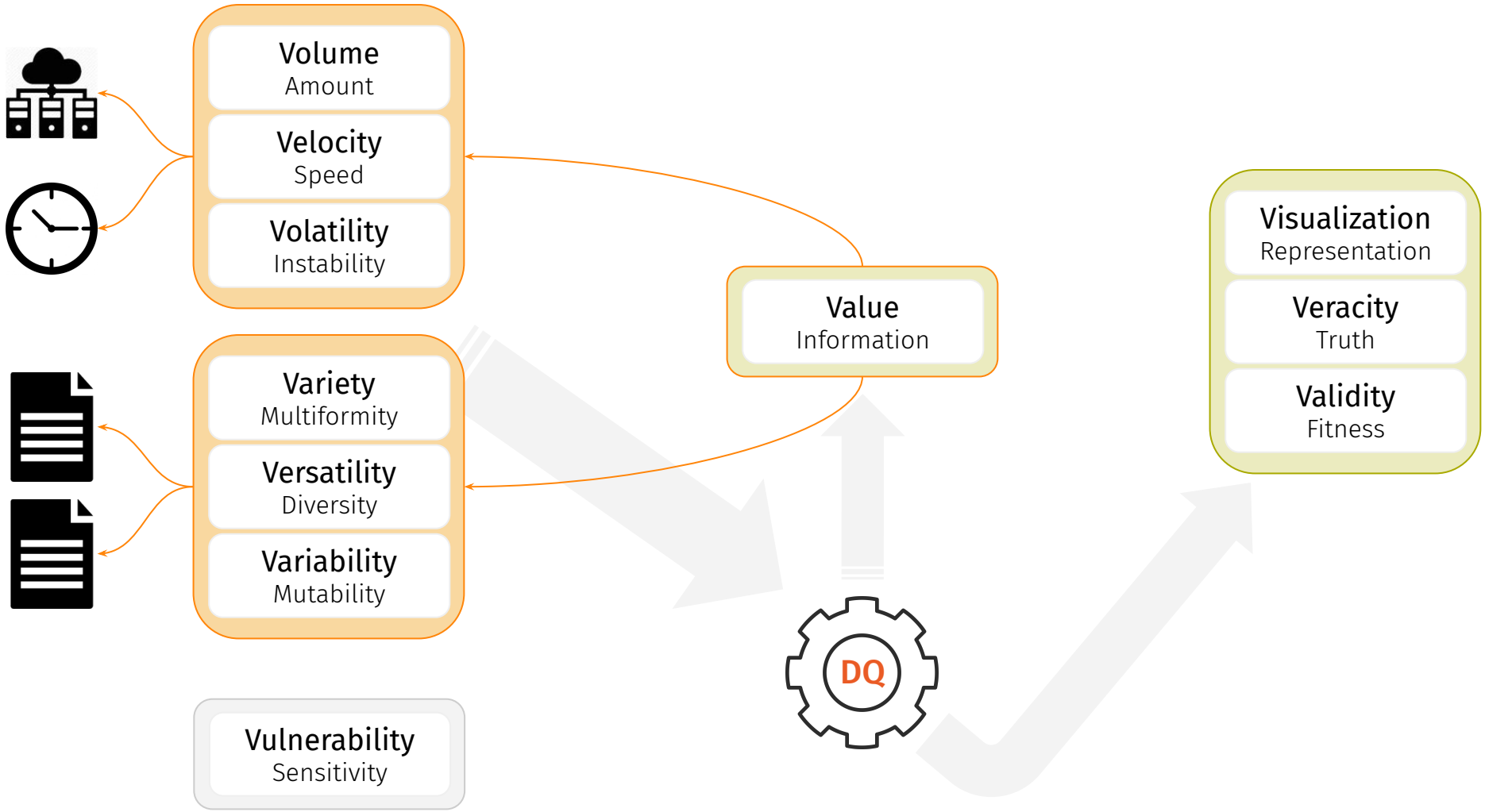
Факторы данных, усложняющие контроль качества



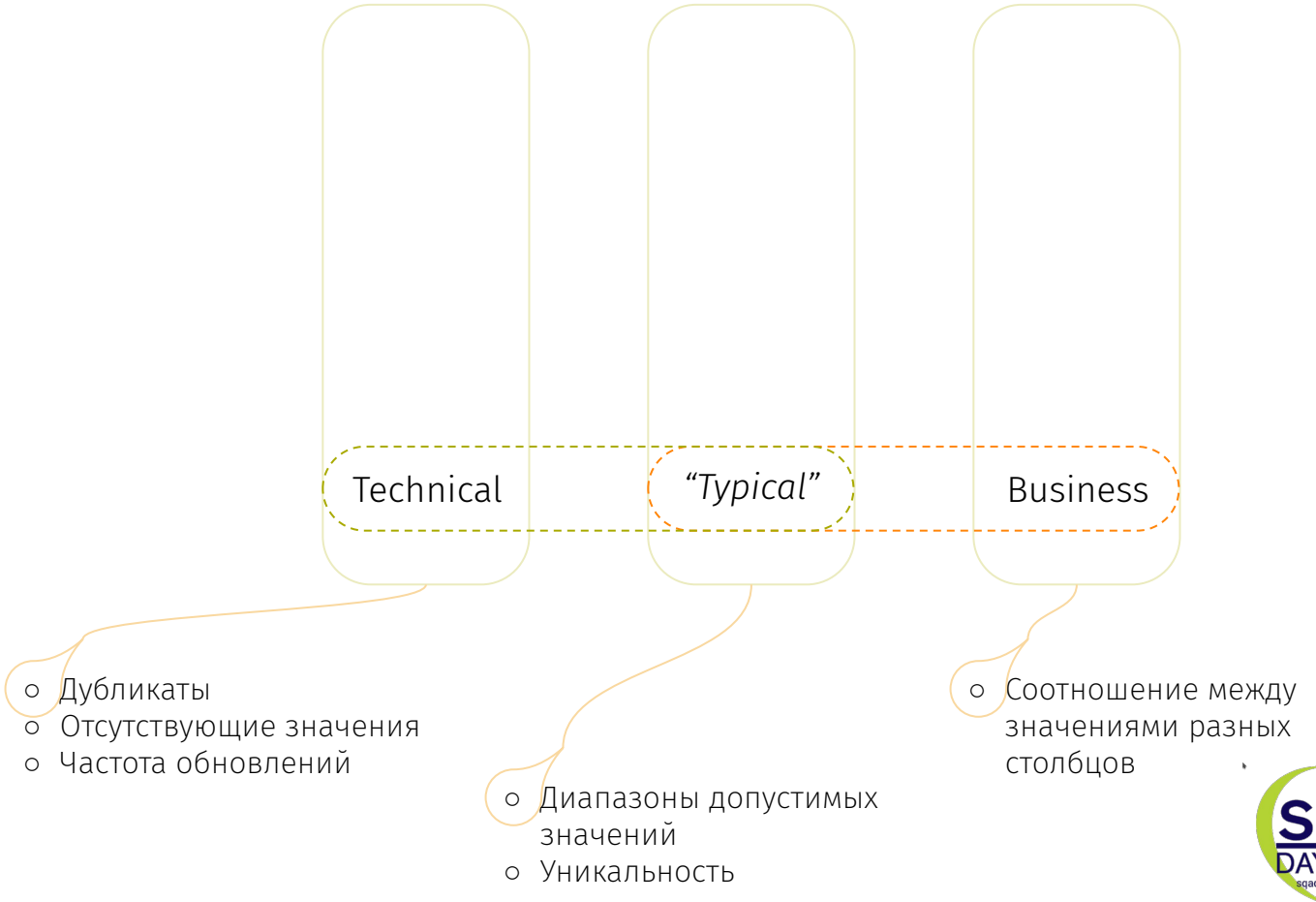
Факторы данных, усложняющие контроль качества



Факторы данных, усложняющие контроль качества

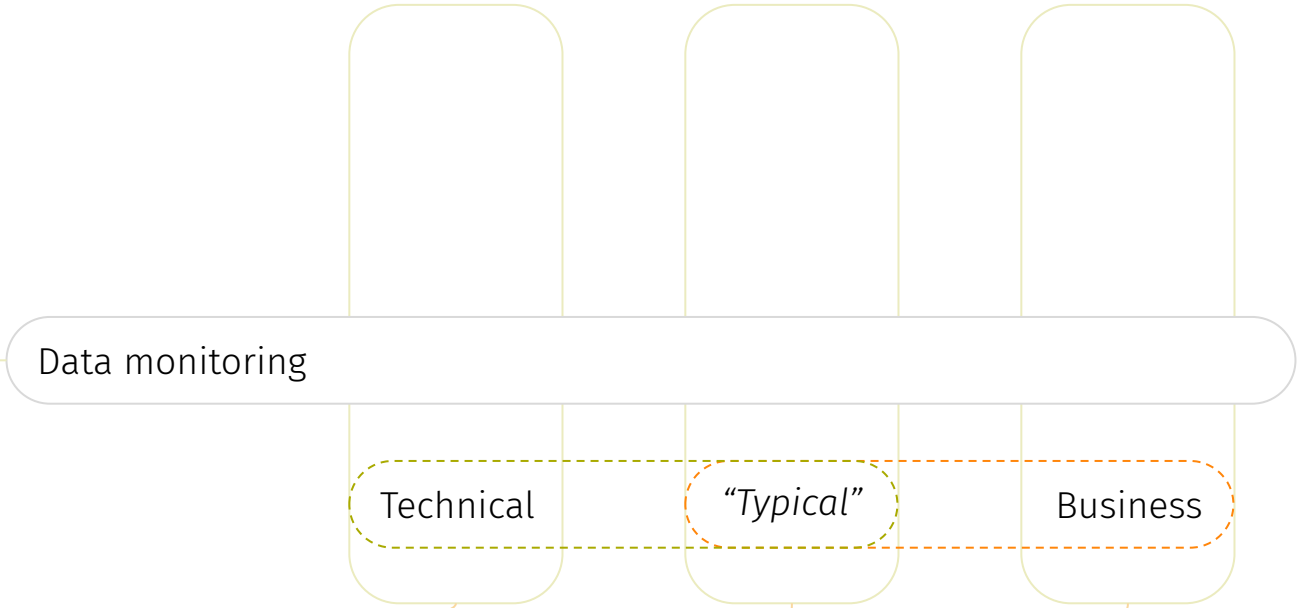


Глубина проверок данных



Глубина проверок данных

- *Цель* мониторинга данных - вовремя определить, что есть проблемы и нотифицировать о них.
- *Точность* мониторинга:
  - Data Source
  - Table
  - Column



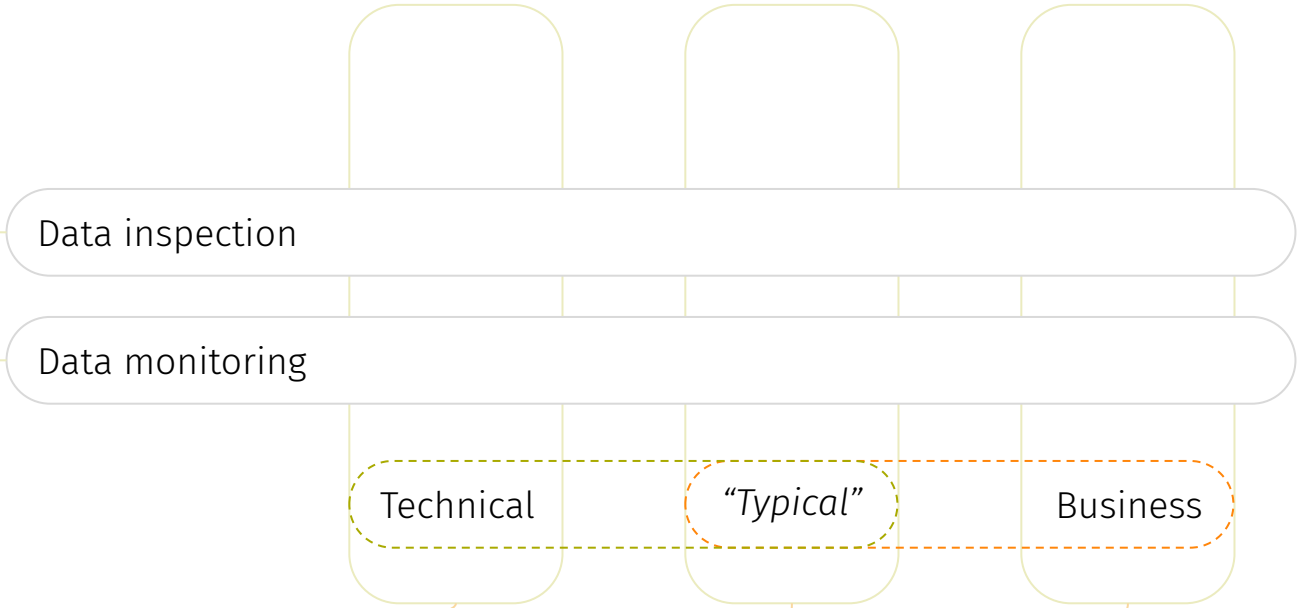
- Дубликаты
- Отсутствующие значения
- Частота обновлений

- Диапазоны допустимых значений
- Уникальность

- Соотношение между значениями разных столбцов

Глубина проверок данных

- *Цель* инспектирования данных - определить к какой категории “хорошести” их можно отнести и объяснить почему. Категории:
  - Duplicated data
  - Incorrect data
  - Damaged data
  - Correct data
- *Точность* инспекций: строка



- *Цель* мониторинга данных - вовремя определить, что есть проблемы и нотифицировать о них.
- *Точность* мониторинга:
  - Data Source
  - Table
  - Column

- Дубликаты
- Отсутствующие значения
- Частота обновлений

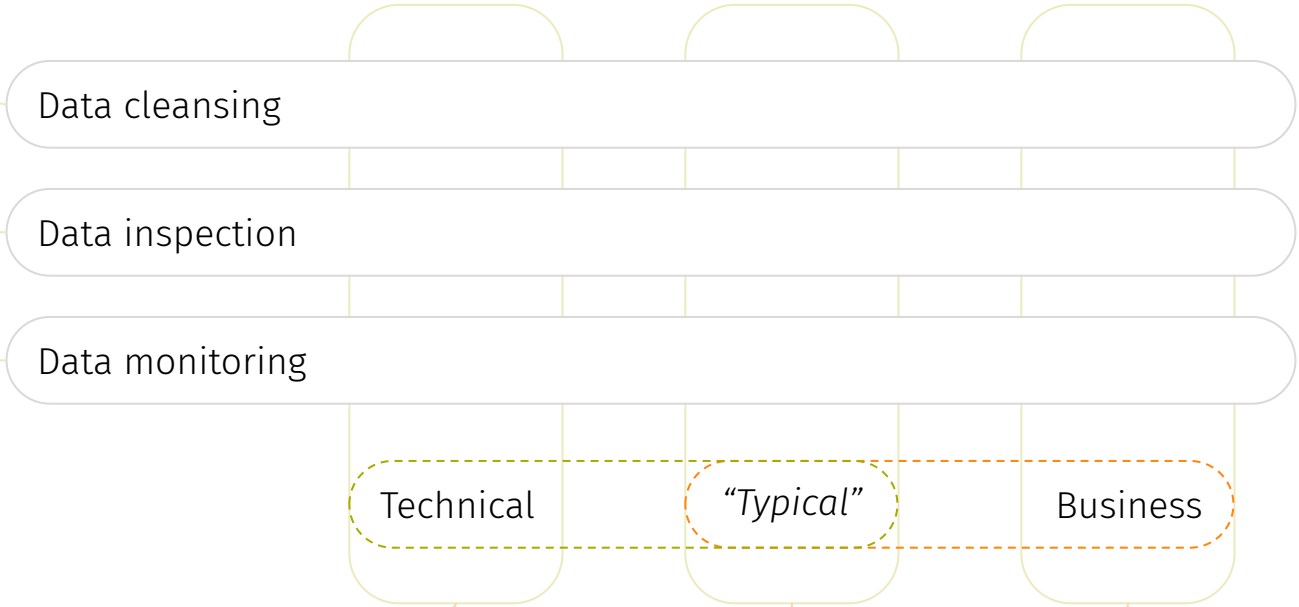
- Диапазоны допустимых значений
- Уникальность

- Соотношение между значениями разных столбцов

Глубина проверок данных

- *Цель* очистки данных - дать рекомендации каким образом исправить данные или сделать это автоматически.
- *Точность* очистки строка, ячейка

- *Цель* инспектирования данных - определить к какой категории “хорошести” их можно отнести и объяснить почему. Категории:
  - Duplicated data
  - Incorrect data
  - Damaged data
  - Correct data
- *Точность* инспекций: строка



- *Цель* мониторинга данных - вовремя определить, что есть проблемы и нотифицировать о них.
- *Точность* мониторинга:
  - Data Source
  - Table
  - Column

- Дубликаты
- Отсутствующие значения
- Частота обновлений

- Диапазоны допустимых значений
- Уникальность

- Соотношение между значениями разных столбцов

Решения в области мониторинга данных

Регулярный автоматизированный подход к мониторингу данных позволяет:

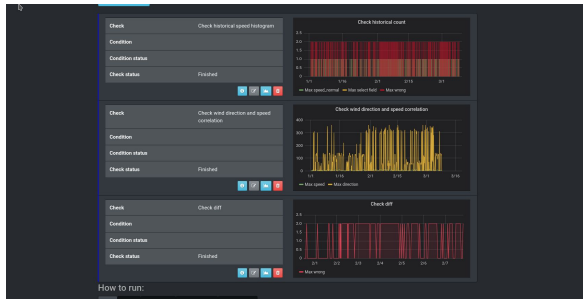
- Вовремя выявлять различного рода ошибки и аномалии в данных или их поведении.
- Собирать метрики о данных.
- Визуализировать данные.
- Отсеивать битые данные на уровне датасетов, таблиц или полей.
- Составлять отчеты об изменении данных со временем.
- Ответить на вопрос “Какой статус данных в настоящий момент?”

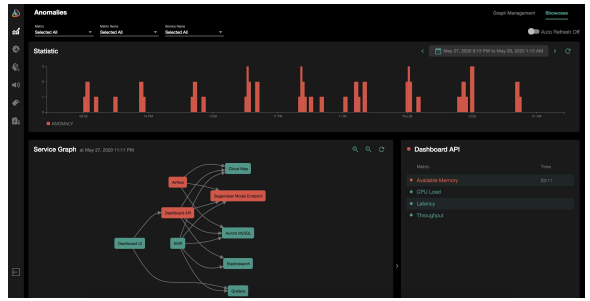
Rule-based monitoring

- Profiling
- Alerting
- Quality dimensions

Real-time anomaly detection

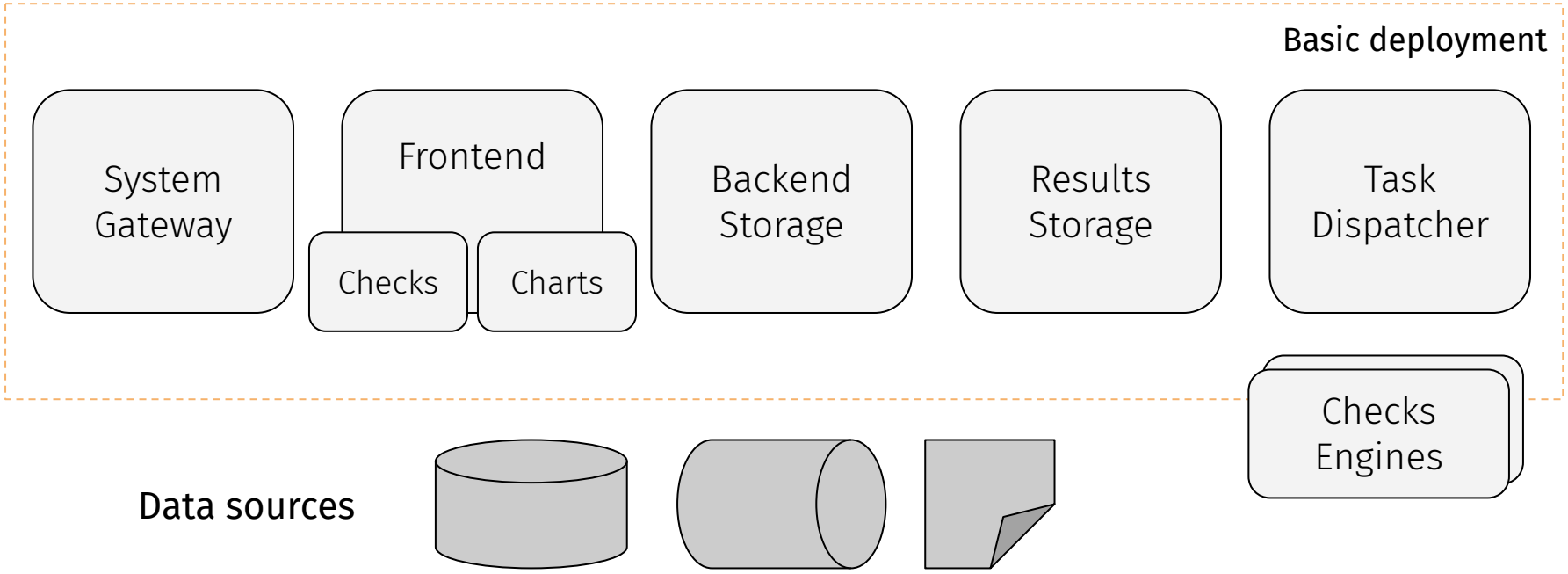
- Real-time analysis
- AI / ML based
- Intelligent alerting



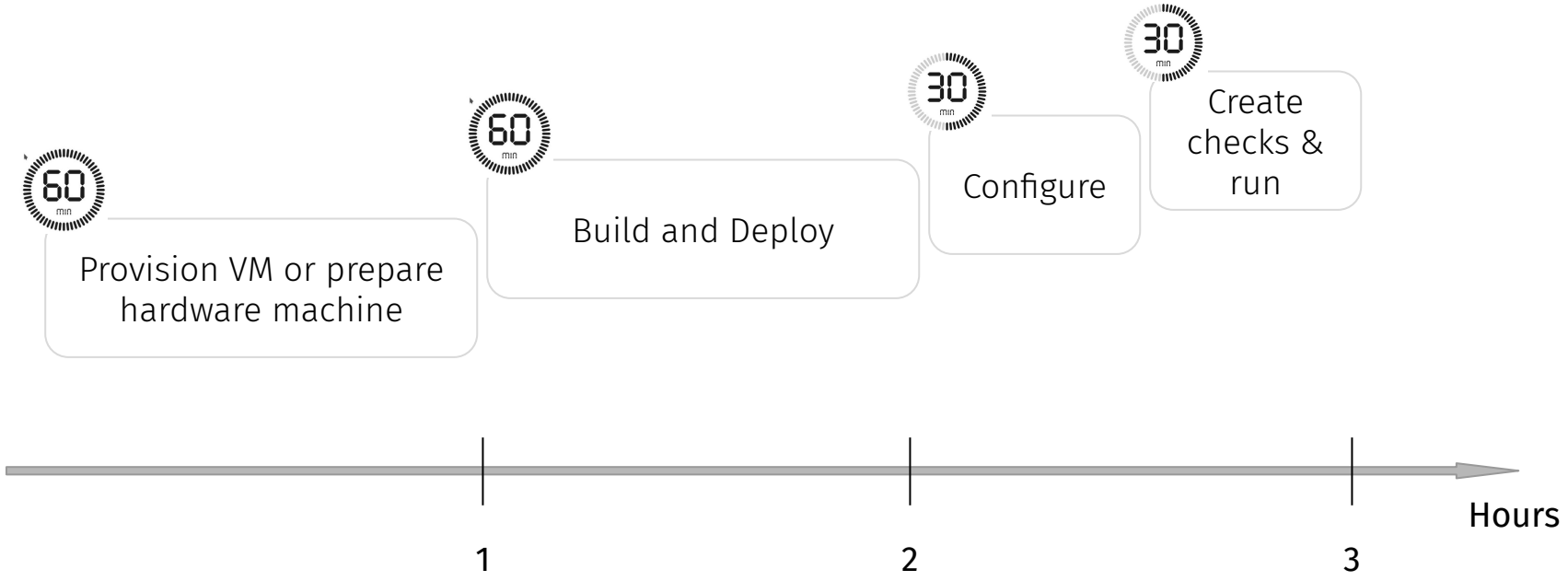


[Статья в блоге]

Архитектура Rule-based решения



Инсталляция Workshop версии



Дальнейшая интеграция требует таких шагов как добавление авторизации, интеграция с планировщиком задач (например Airflow), настройка Retention policy, создание пользовательских коннекторов к данным.
И конечно создание самих проверок.

Плюсы решения

Инсталляция и конфигурация

- Решение базируется на Docker-е
- Cloud-agnostic или on-premise
- t2.xlarge in AWS или продуктивный ноутбук
- 30’ для развертывания базовой версии
- Поддержка любой JDBC-compliant базы данных
- Поддержка файловых форматов CSV and parquet
- Поддержка Kafka-streams
- API для создания пользовательских коннекторов

Репортинг

- Grafana используется для визуализации данных
- Grafana используется для alerting-а
- Grafana используется для оповещений
- Grafana графики интегрированы в UI
- Grafana панели создаются автоматически
- Поддержка репортов по разным “срезам” системы
- Логи последнего запуска доступны в UI

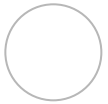
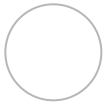
Интеграция с системой обработки данных

- Поддержка любого scheduler-а (e.g. Airflow)
- Шаблоны кода для создания задач (e.g. Lambda)
- Запуск задач через REST API
- Поддержка Sidecar-подхода к мониторингу
- Возможность останавливать основной пайлайн обработки данных
- Возможность использовать Spark cluster
- Возможность использовать Elasticsearch
- Возможность работать с метриками данных без доступа к данным

Создание проверок

- Минимальное требуемое знание - SQL
- Создание параметризируемых шаблонов проверок
- Поддержка 3rd-party тулов проверки данных
- Только r/o доступ к данным
- Возможность запуска из интерфейса

Примеры

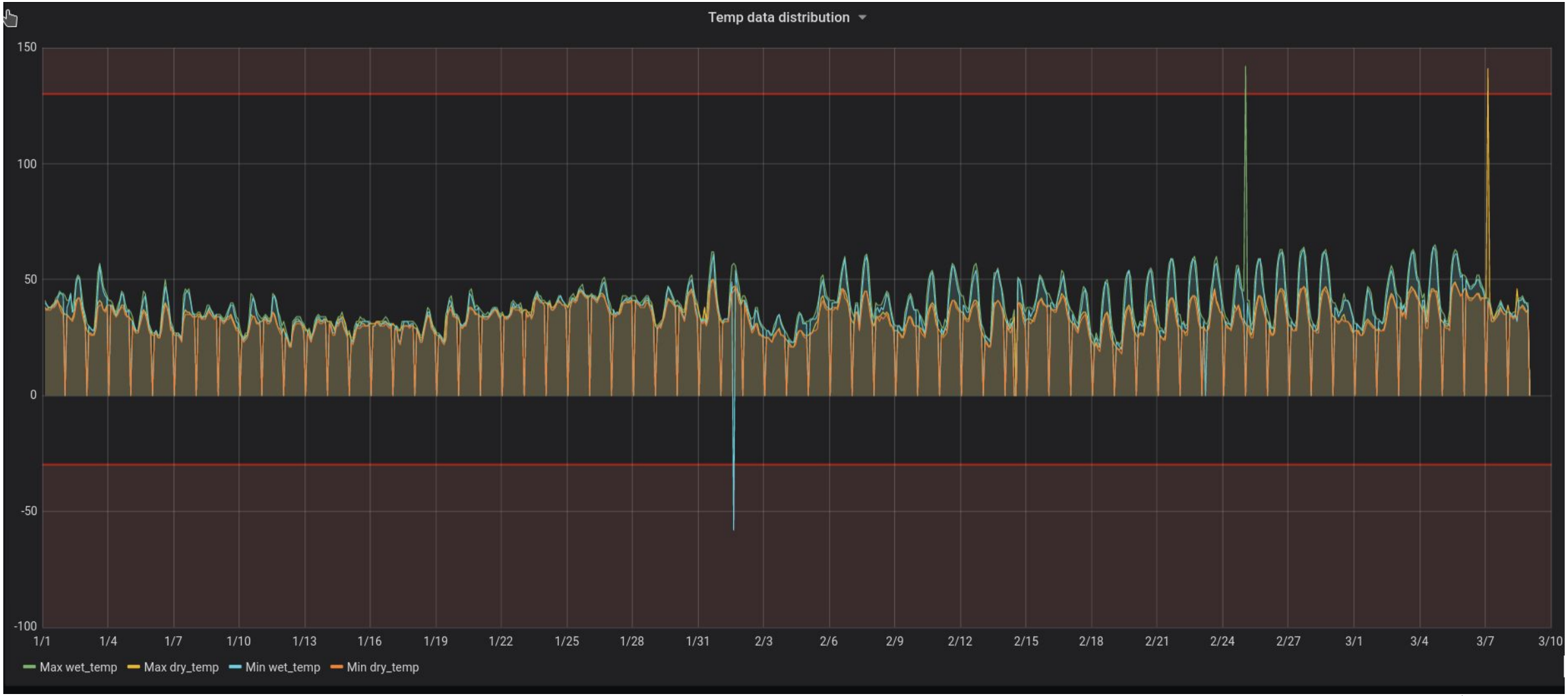
Типы данных \ Техника	Сравнение с образцом	Применение правил оценки	Сравнение со статистикой	Поиск аномалий
Исторические данные				
Пакетная обработка данных				
Потоковая обработка данных				
Миграция данных				

Примеры. Изучаем температуру

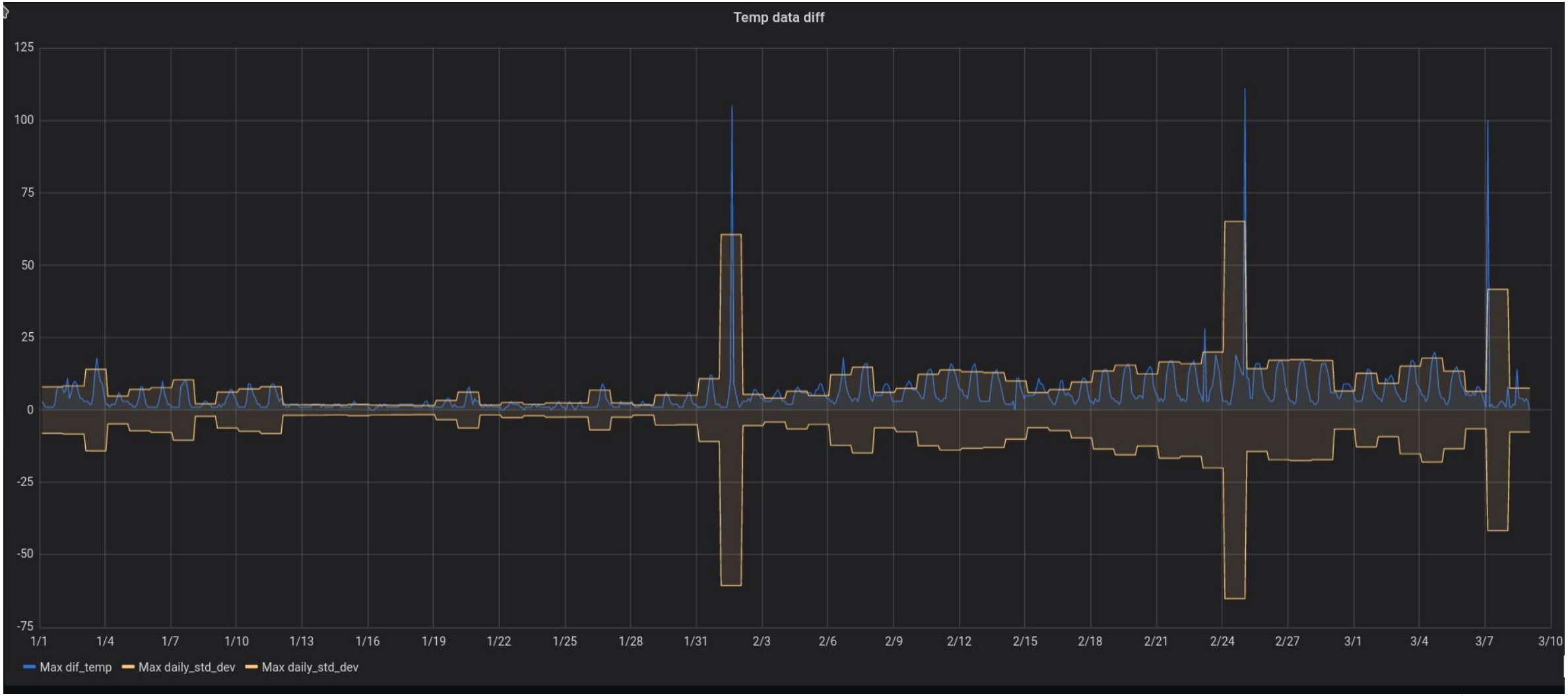
A	B	C
date	dry_temp	wet_temp
2020-01-02T10:56:00	48	39
2020-01-02T11:56:00	50	41
2020-01-02T12:56:00	52	42
2020-01-02T13:56:00	51	42
2020-01-02T14:56:00	51	42
2020-01-02T15:17:00	50	42
2020-01-02T15:27:00	51	42
2020-01-02T15:56:00	50	42
2020-01-02T16:56:00	45	39
2020-01-02T17:56:00	41	37
2020-01-02T18:56:00	38	35
2020-01-02T19:56:00	37	33
2020-01-02T20:56:00	35	31
2020-01-02T21:56:00	33	30
2020-01-02T22:56:00	31	28
2020-01-02T23:56:00	30	27
2020-01-02T23:59:00		
2020-01-03T00:56:00	30	27
2020-01-03T01:56:00	30	27
2020-01-03T02:56:00	29	27
2020-01-03T03:56:00	29	26
2020-01-03T04:56:00	28	26
2020-01-03T05:56:00	28	26

Local Climatological Data (LCD)

Примеры. Изучаем температуру



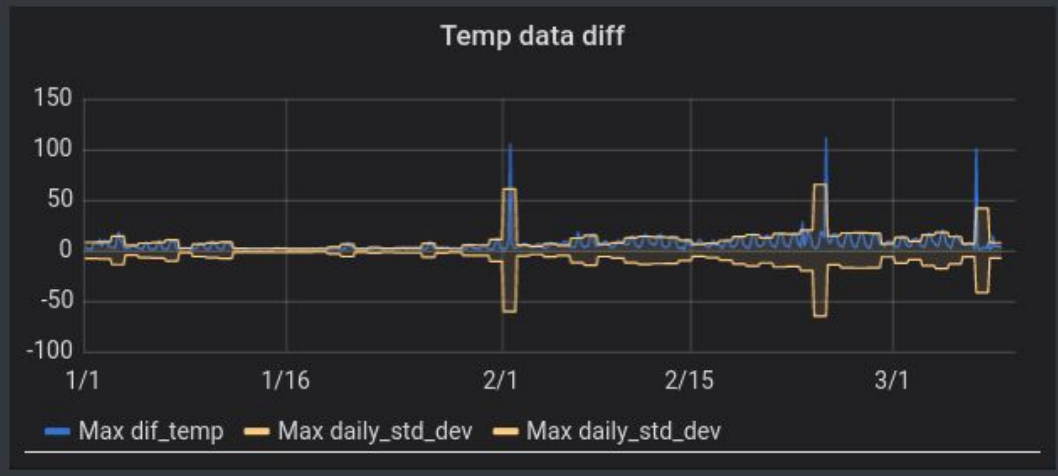
Примеры. Изучаем температуру



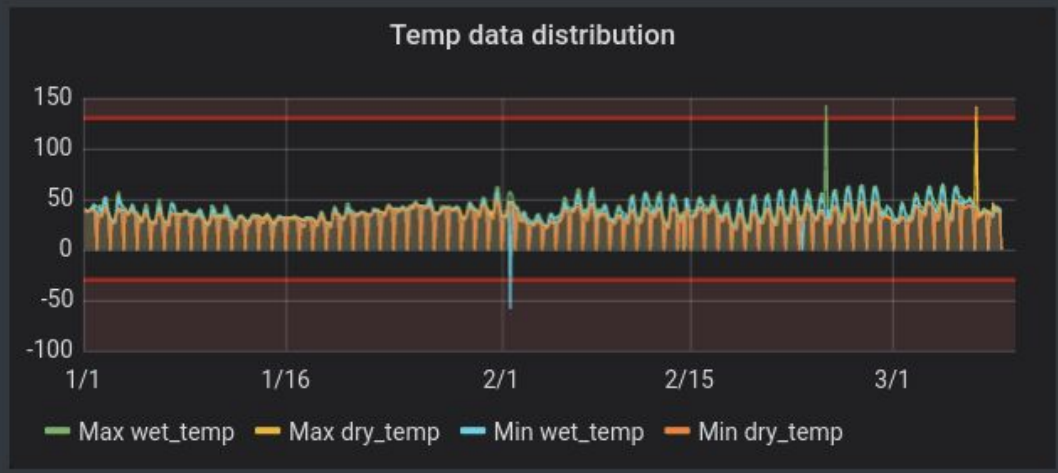
Deja Vu

Примеры. Пакетная обработка данных

Check	Temp data diff
Condition	
Condition status	
Check status	Finished



Check	Temp data distribution
Condition	
Condition status	
Check status	Finished

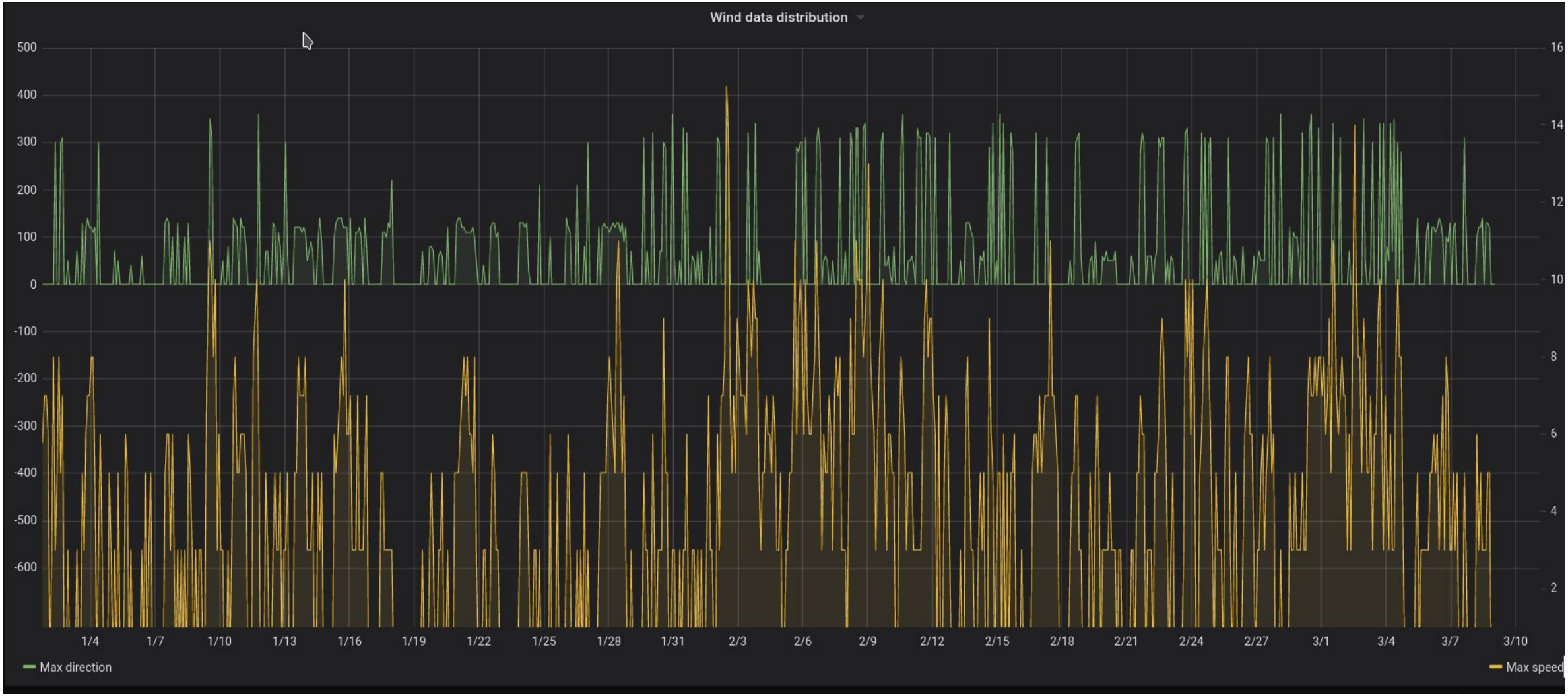


Примеры. Изучаем ветер

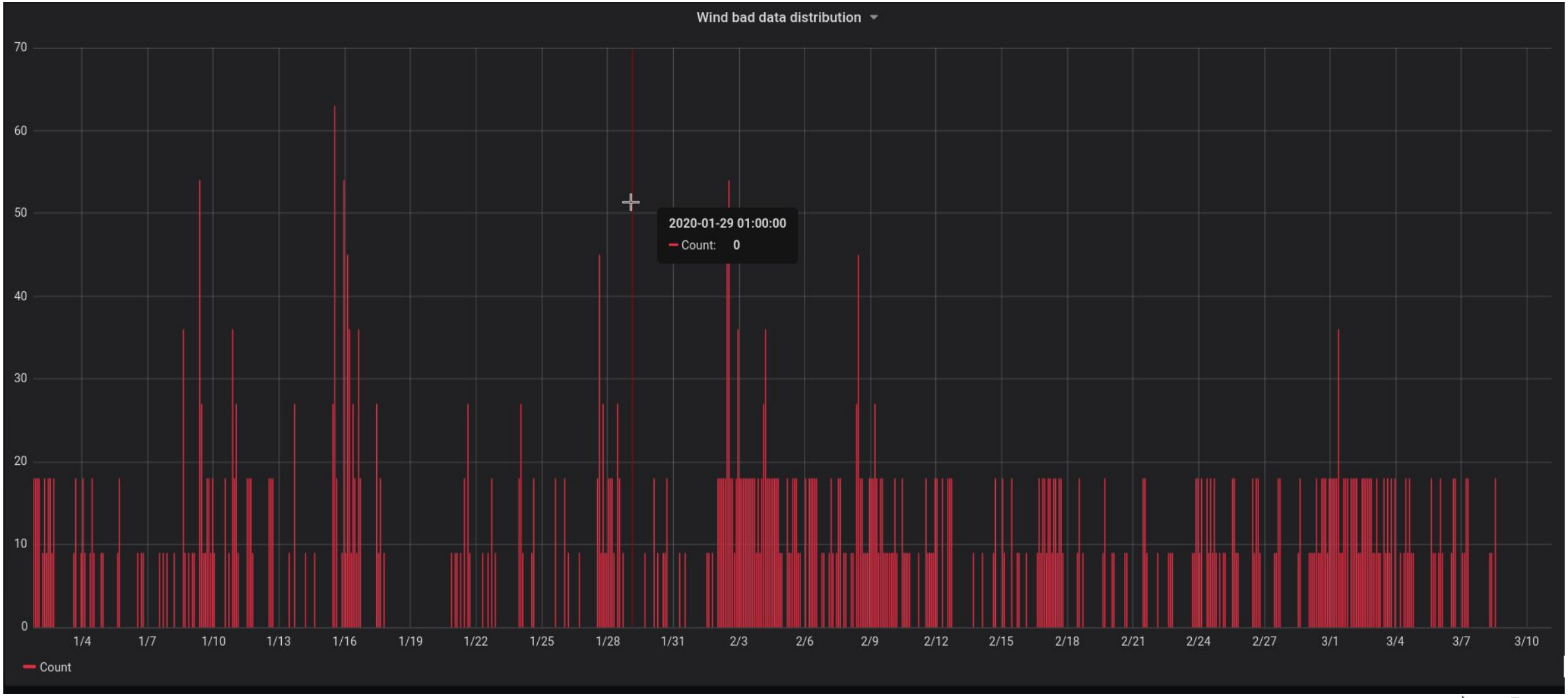
A	B	C	D
date	direction	gust_speed	speed
2020-01-13T16:26:00	100		3
2020-01-13T16:35:00	80		7
2020-01-13T16:45:00	90		7
2020-01-13T16:53:00	100		5
2020-01-13T16:56:00	100		6
2020-01-13T17:11:00	90		7
2020-01-13T17:18:00	90		6
2020-01-13T17:34:00	90		6
2020-01-13T17:42:00	100		6
2020-01-13T17:56:00	110		3
2020-01-13T18:08:00	110		3
2020-01-13T18:19:00	100		7
2020-01-13T18:56:00	0		0
2020-01-13T19:14:00	120		3
2020-01-13T19:28:00	120		3
2020-01-13T19:56:00	110		3
2020-01-13T20:08:00	110		5
2020-01-13T20:27:00	90		7
2020-01-13T20:56:00	90		8
2020-01-13T21:04:00	100		5

Local Climatological Data (LCD)

Примеры. Изучаем ветер



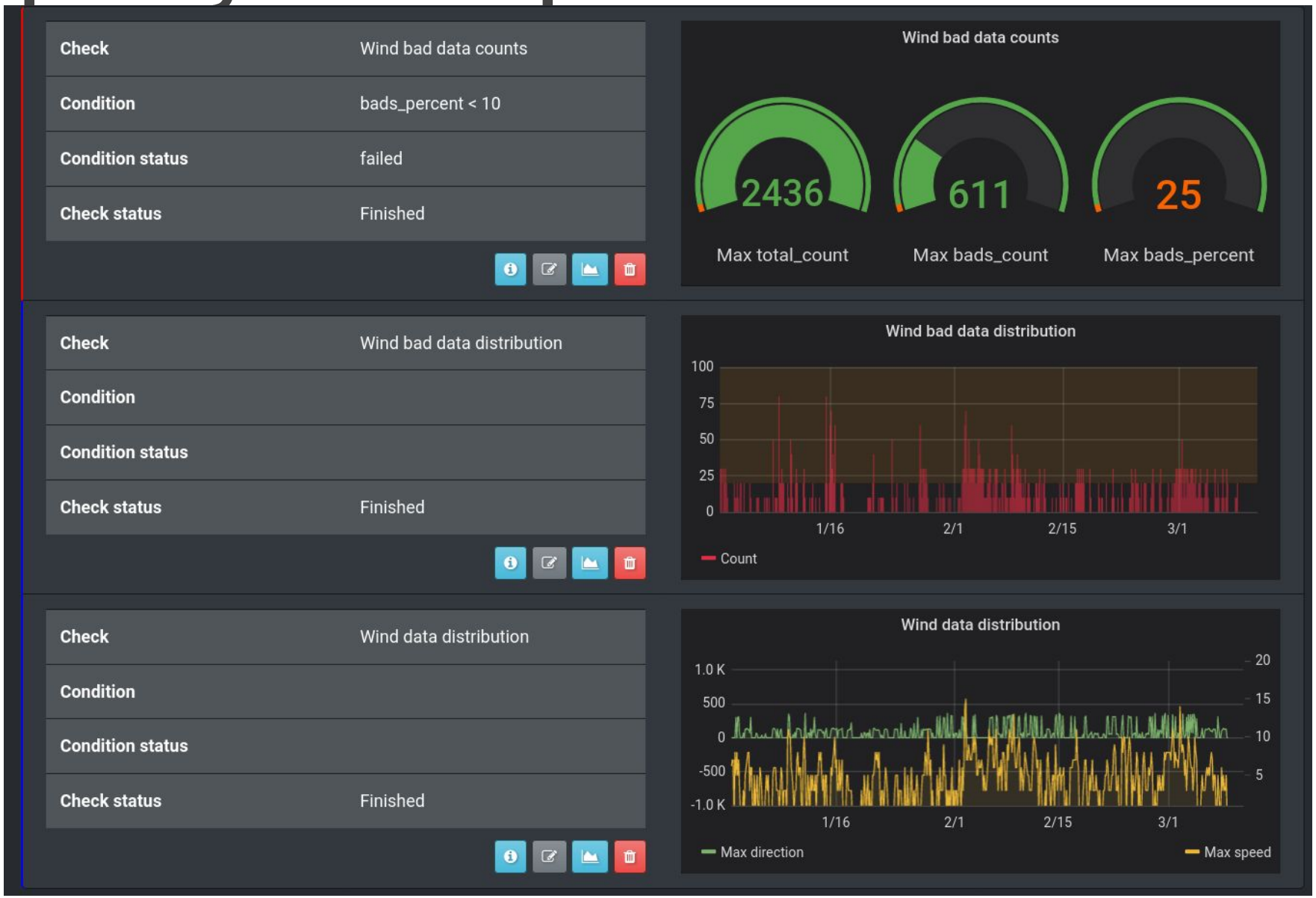
Примеры. Изучаем ветер



Примеры. Изучаем ветер

Примеры. Изучаем ветер

Примеры. Изучаем ветер



Примеры. Сравнение с SoR

JSON

A	B	C	D
date	direction	gust_speed	speed
2020-01-13T16:45:00	90		7
2020-01-13T16:53:00	100		5
2020-01-13T16:56:00	100		6
2020-01-13T17:11:00	90		7

STRING

Database

INTEGER

```
CREATE TABLE public.demo_lcd_wind1 (
    "date" timestamp NULL,
    direction int4 NULL,
    gust_speed int4 NULL,
    speed int4 NULL
);
```

Примеры. Сравнение с SoR

JSON

A	B	C	D
date	direction	gust_speed	speed
2020-01-13T16:45:00	90		7
2020-01-13T16:53:00	100		5
2020-01-13T16:56:00	100		6
2020-01-13T17:11:00	90		7

STRING

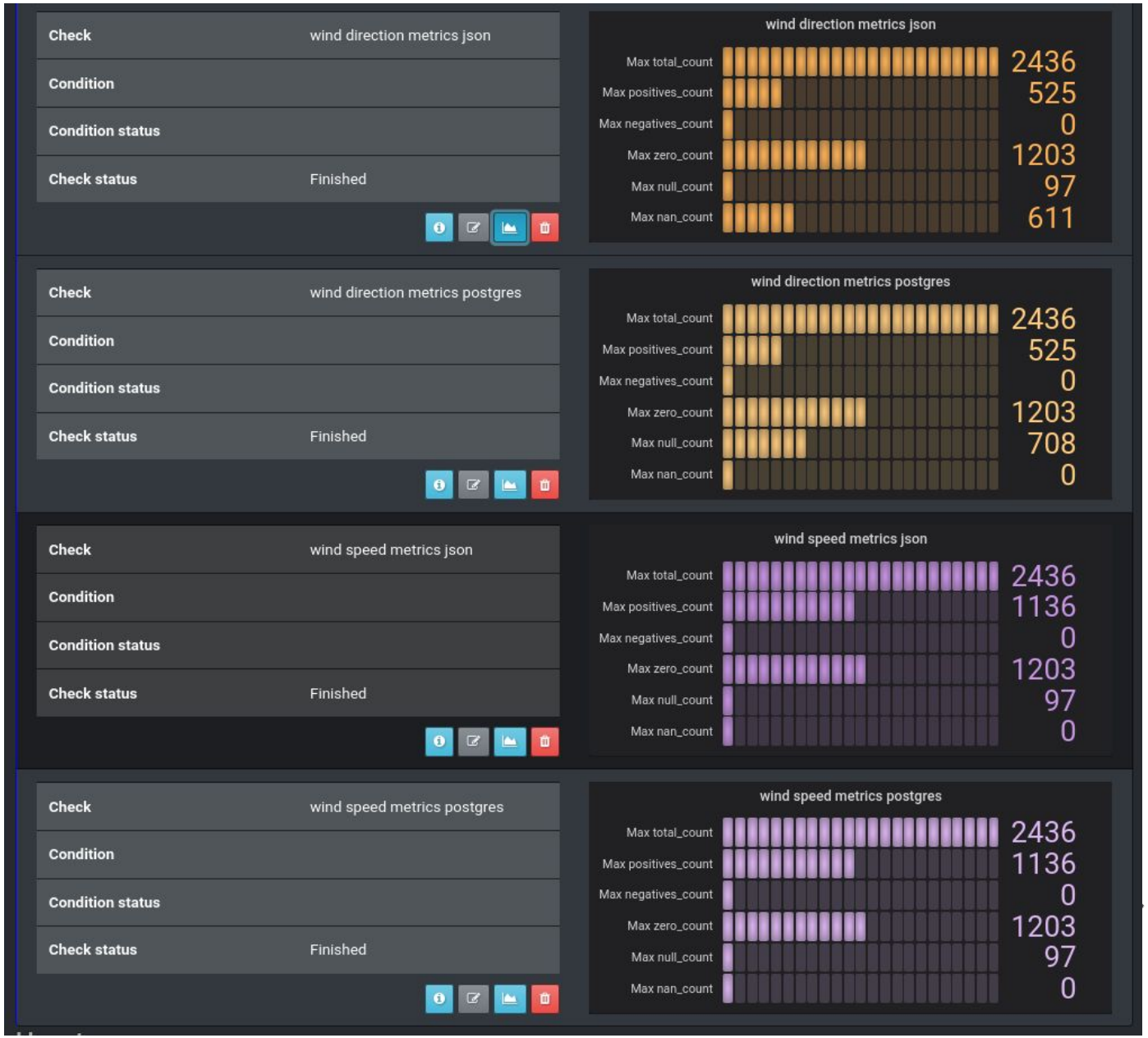
Database

INTEGER

```
CREATE TABLE public.demo_lcd_wind1 (
    "date" timestamp NULL,
    direction int4 NULL,
    gust_speed int4 NULL,
    speed int4 NULL
);
```

Total count:	count(*)
Positives count:	count(*) when > 0
Negatives count:	count(*) when < 0
Zeros count:	count(*) when = 0
Nulls count:	count(*) when is NULL
NaNs count:	count(*) when isnan()

Примеры. Сравнение с SoR



Примеры. Сравнение с SoR

A	B	C	D
date	direction	gust_speed	speed
2020-01-13T14:22:00	110		6
2020-01-13T14:31:00	110		5
2020-01-13T14:40:00	VRB		5
2020-01-13T14:54:00	110		3
2020-01-13T14:56:00	110		5
2020-01-13T15:13:00	VRB		6
2020-01-13T15:19:00	90		7
2020-01-13T15:28:00	90		5
2020-01-13T15:38:00	VRB		3
2020-01-13T15:48:00	100		7
2020-01-13T15:54:00			
2020-01-13T15:56:00			
2020-01-13T16:03:00	90		7
2020-01-13T16:12:00	100		5
2020-01-13T16:17:00	100		3
2020-01-13T16:24:00	100		5
2020-01-13T16:26:00	100		3
2020-01-13T16:35:00	80		7
2020-01-13T16:45:00	90		7
2020-01-13T16:53:00	100		5
2020-01-13T16:56:00	100		6
2020-01-13T17:11:00	90		7
2020-01-13T17:18:00	90		6

Выводы

- Зачастую, данные проходят проверки, но делается это не систематически, редко или вручную. Это неэффективно.
- Изучать данные необходимо с разных сторон.
- Делать это необходимо регулярно и с помощью автоматизированных подходов.
- Организовать мониторинг данных задача, которая по началу может быть несложной, но будет усложняться в зависимости от
  - Сложности данных
  - Глубины проверок
  - Количества проверок на едицу времени

Выводы

- Зачастую, данные проходят проверки, но делается это не систематически, редко или вручную. Это неэффективно.
- Изучать данные необходимо с разных сторон.
- Делать это необходимо регулярно и с помощью автоматизированных подходов.
- Организовать мониторинг данных задача, которая по началу может быть несложной, но будет усложняться в зависимости от
  - Сложности данных
  - Глубины проверок
  - Количества проверок на едицу времени

Не знаете как начать?

Давайте партнериться!

Grid Dynamics

Alexey Lyanguzov - Head of QA Practice
alyanguzov@griddynamics.com

Thank you!

www.griddynamics.com