

Modern Storages and Data Warehousing Week 10 - Business Intelligence

Попов Илья, i.popov@hse.ru

1 - Homework Q&A

2 - Homework #4

Ресар прошлых занятий

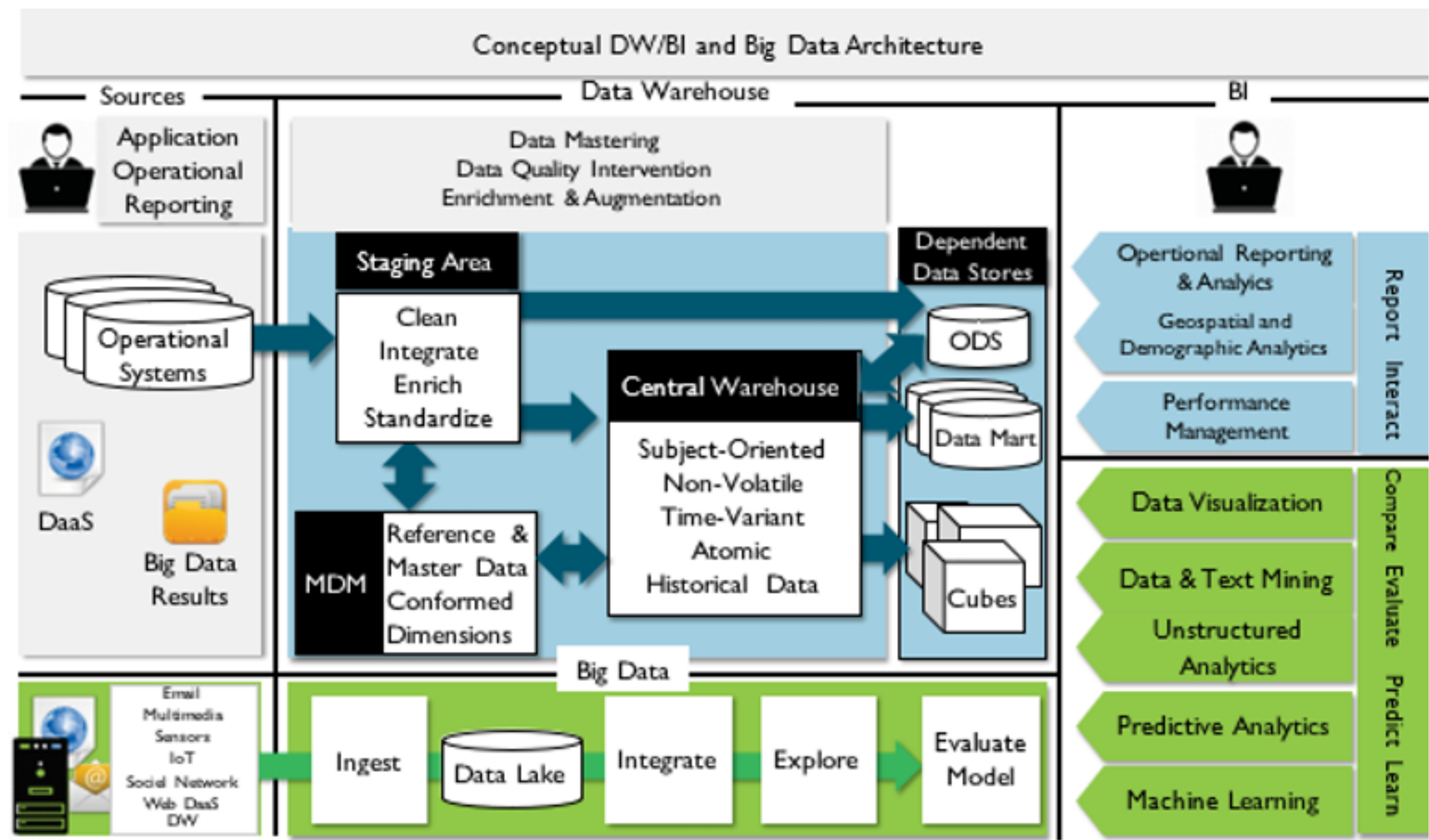


Figure 5: Data Warehouse Concept

3 - BI tools

Мотивация

- › Считать цифры - это, конечно, хорошо, но их еще и хочется видеть в удобном виде
- › Гуманитарии усиленно не принимают таблицы и формулы - им нужны визуализации, фильтры и кнопочки
- › Кроме того, что на отчетность хотят смотреть гуманитарии, мы еще хотим в удобном виде смотреть на мониторинги (а лучше - повесить команде телевизор и смотреть доту на метрики)
- › Для всего этого нам нужны BI-инструменты

3.1 - Мониторинги

Для чего они нужны

- › Когда у нас есть ODS-слой, и мы хотим следить за изменением оперативных метрик (клики, продажи, так далее)
- › Когда у нас есть ODS-слой, и мы хотим оперативно отлавливать события (500-ки; послыки, которые не успели в свой SLA)
- › Когда у нас есть Timeseries DB (например, Prometheus или Solomon), в который мы пушим сообщения, и мы хотим их читать / собирать из них метрики

Золотой стек

В 99.9999% компаний вы увидите именно такой стек для мониторинга



Prometheus



Grafana

Золотой стек



Prometheus



Grafana

Кроме Яндекса, где у вас будет:



Solomon



Monitoring

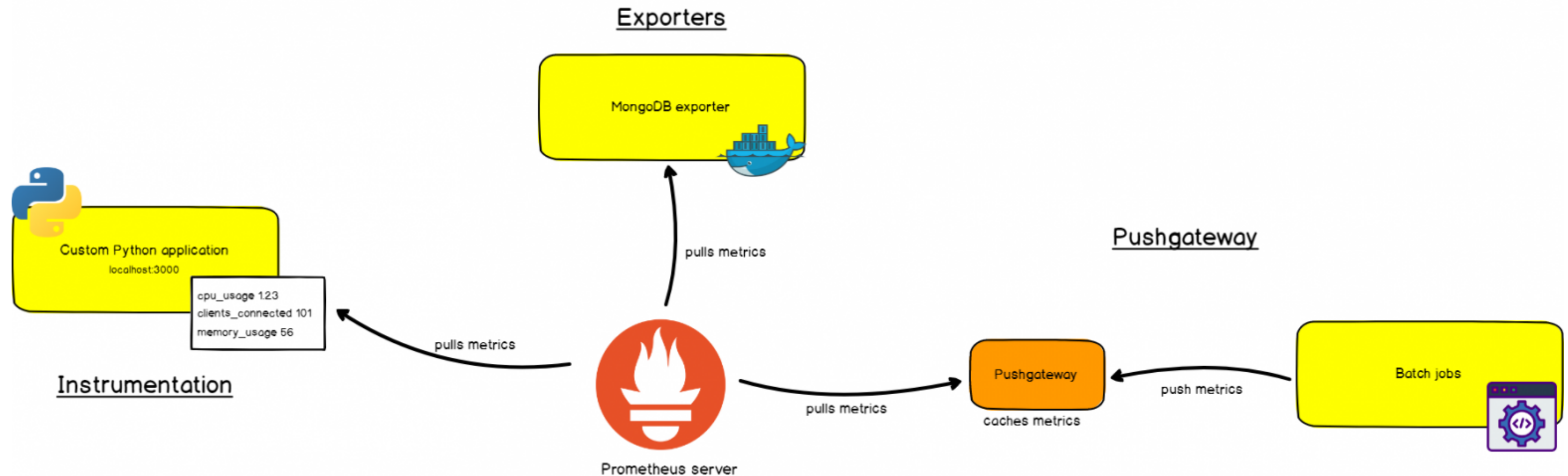
Prometheus



- › Создали внутри Soundcloud в 2012 году
- › Что умеет Prometheus:
 - Timeseries DB для хранения метрик
 - Умеет по REST API ходить к вашей инфре и получать метрики
 - Имеет Pushgateway - чтобы приложения могли сами ходить и отправлять свои метрики
 - Умеет кидать Webhook на алерт по заданным правилам

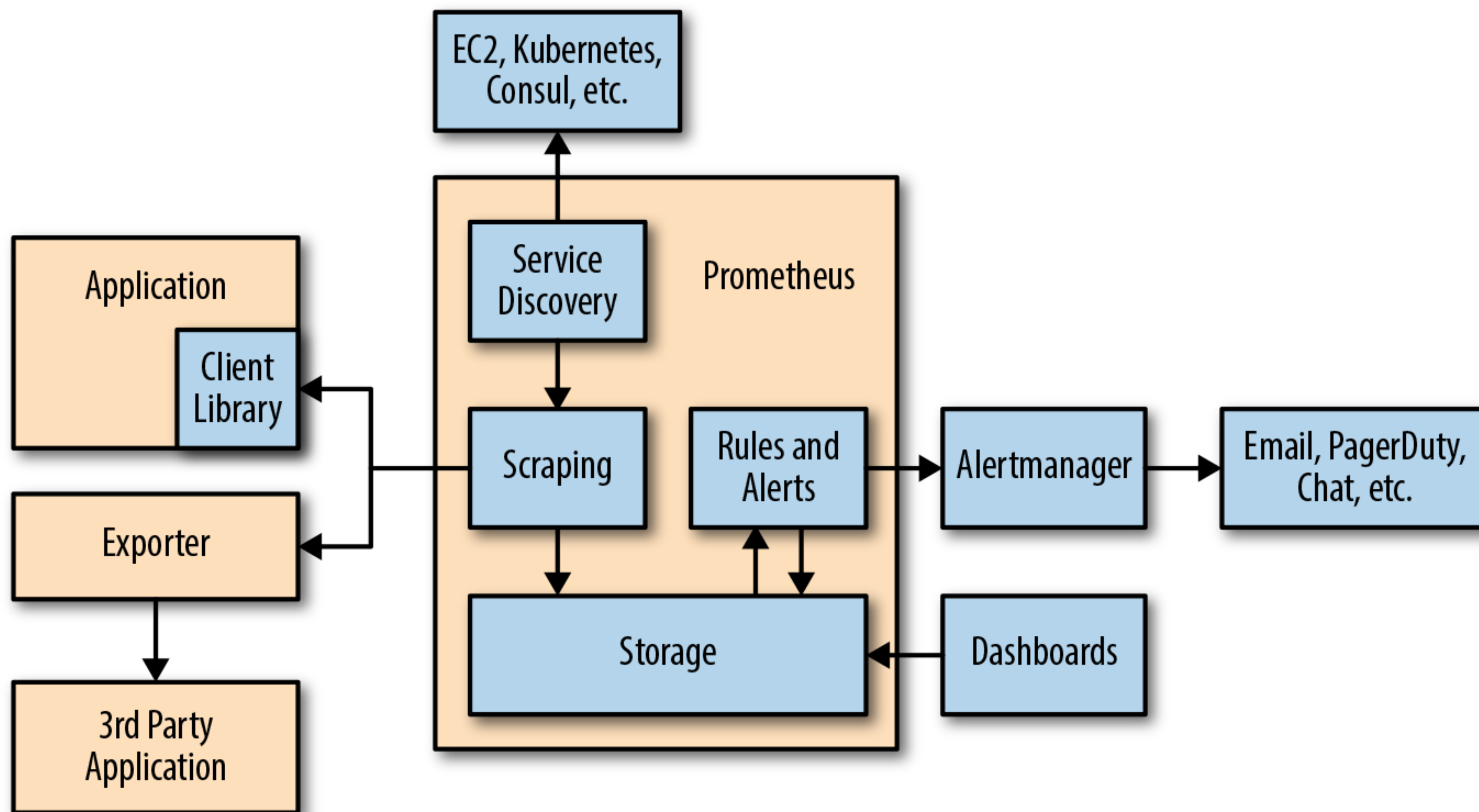
Prometheus - polling / pulling

Ways to gather metrics in Prometheus



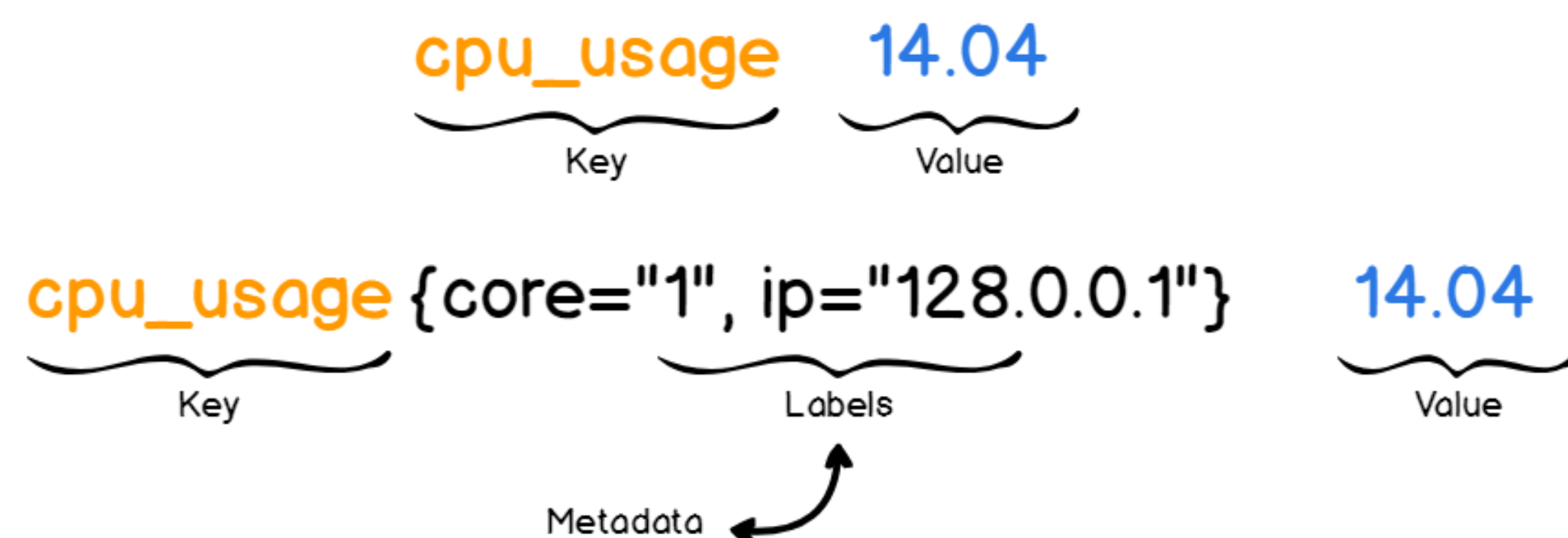
Подробнее: <https://habr.com/ru/companies/slurm/articles/455290/>

Prometheus - архитектура



Prometheus - метрики

1 Prometheus Data Model



2 Data Model Filtering



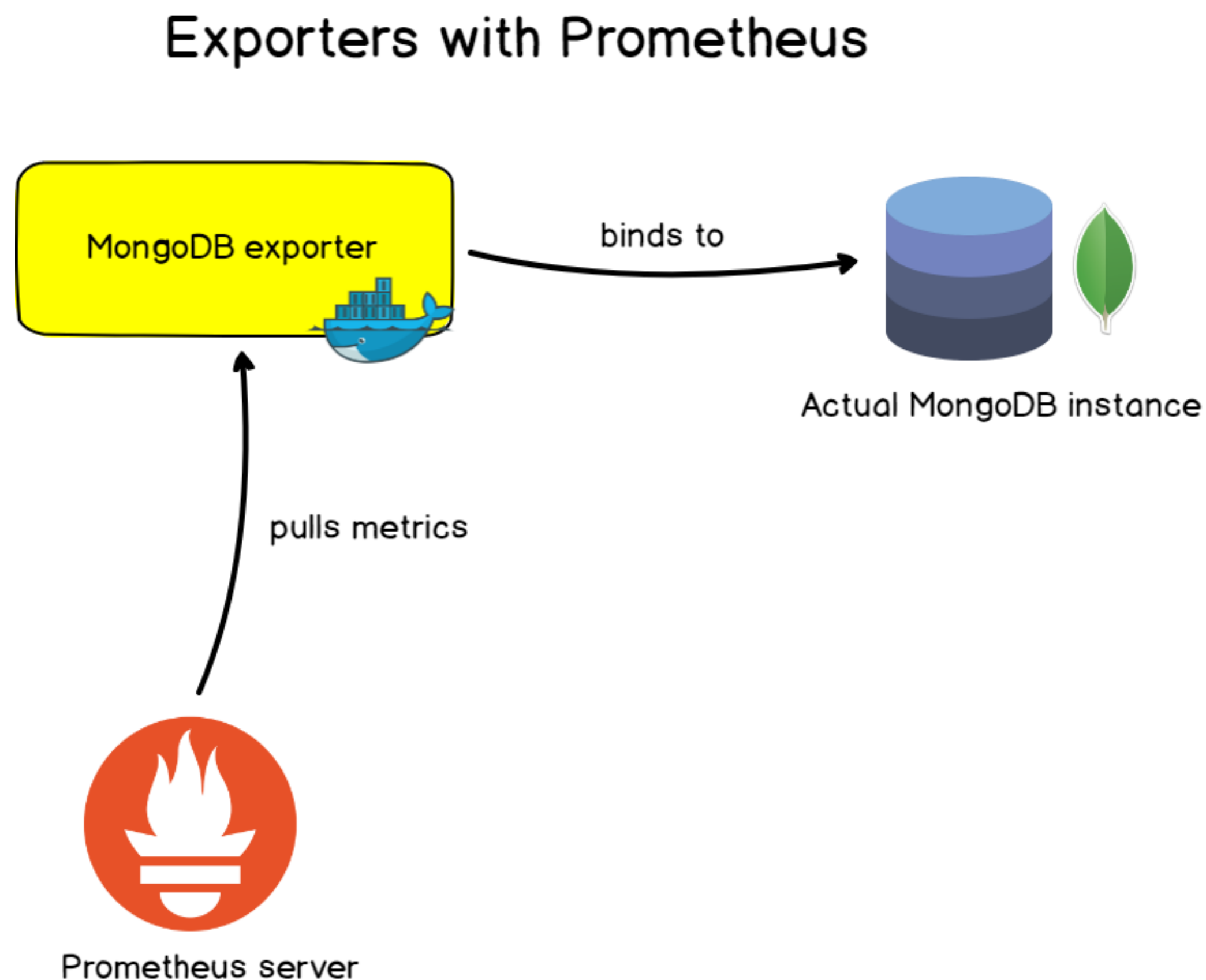
› Prometheus работает с парами «ключ-значение». Ключ описывает, что мы измеряем, а значение хранит фактическую величину в виде числа.

› Ярлыки дают больше сведений о метриках, добавляя дополнительные поля.

› Виды метрик:

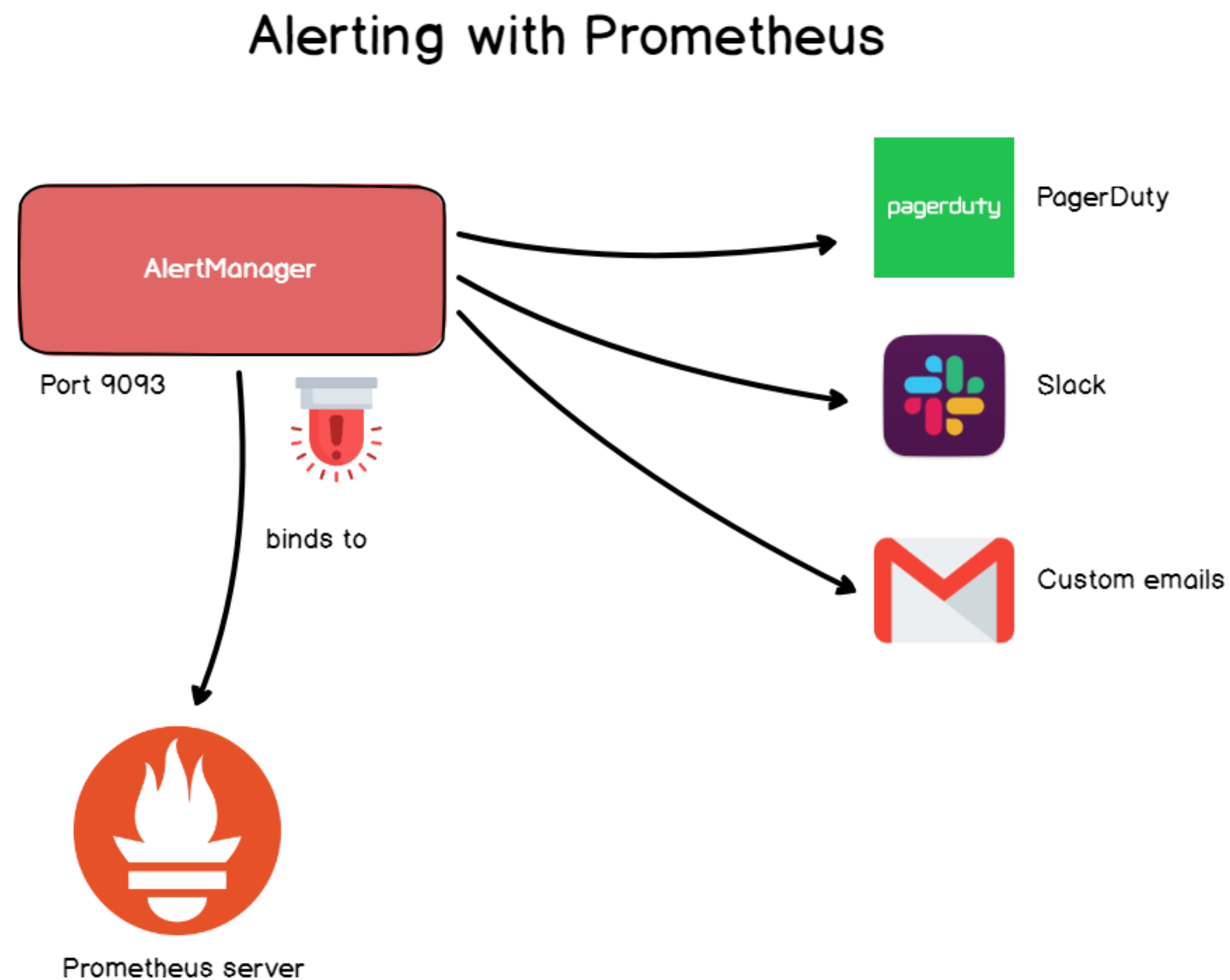
- Counter
- Gauge
- Histogram
- Summary

Prometheus - экспортеры



- › Для известных приложений, серверов и баз данных Prometheus предлагает экспортеры, с помощью которых можно мониторить целевые объекты.
- › Эти экспортеры обычно представлены в виде образов Docker и легко настраиваются.
- › Они предоставляют готовый набор метрик и часто готовые панели мониторинга, с которыми можно настроить мониторинг за считанные минуты.
- › Примеры:
 - **Экспортеры баз данных:** MongoDB, PostgreSQL, MySQL и другие
 - **Экспортеры HTTP:** HAProxy, Apache, NGINX и другие
 - **Экспортеры Unix**

Prometheus - алерты



- › Менеджер оповещений — это отдельный инструмент, который присоединяется к Prometheus и запускает кастомные оповещатели.
- › Оповещения фиксируются в конфиг-файле, если условие срабатывает - инициируется оповещение.
- › В качестве получателя можно указать электронный адрес, вебхук Slack, PagerDuty и кастомные HTTP-объекты.


Grafana



- › Создали внутри Orbitz (американский Aviasales) в 2014 году
- › Первый UI был написан на основе Kibana (еще одна opensource BI)
- › Что умеет Grafana:
 - Ходить в Timeseries DB (InfluxDB, Prometheus)
 - Ходить с помощью SQL по JDBC
 - Ходить в другие хранилища данных (например, в Graylog / Elasticsearch)
 - Умеет строить борды
 - Умеет кидать Webhook на алерт по заданным правилам

Демо

3.2 - Отчетность



Главная задача отчета -
быстро бросить взгляд,
понять, все хорошо или
плохо, и идти дальше

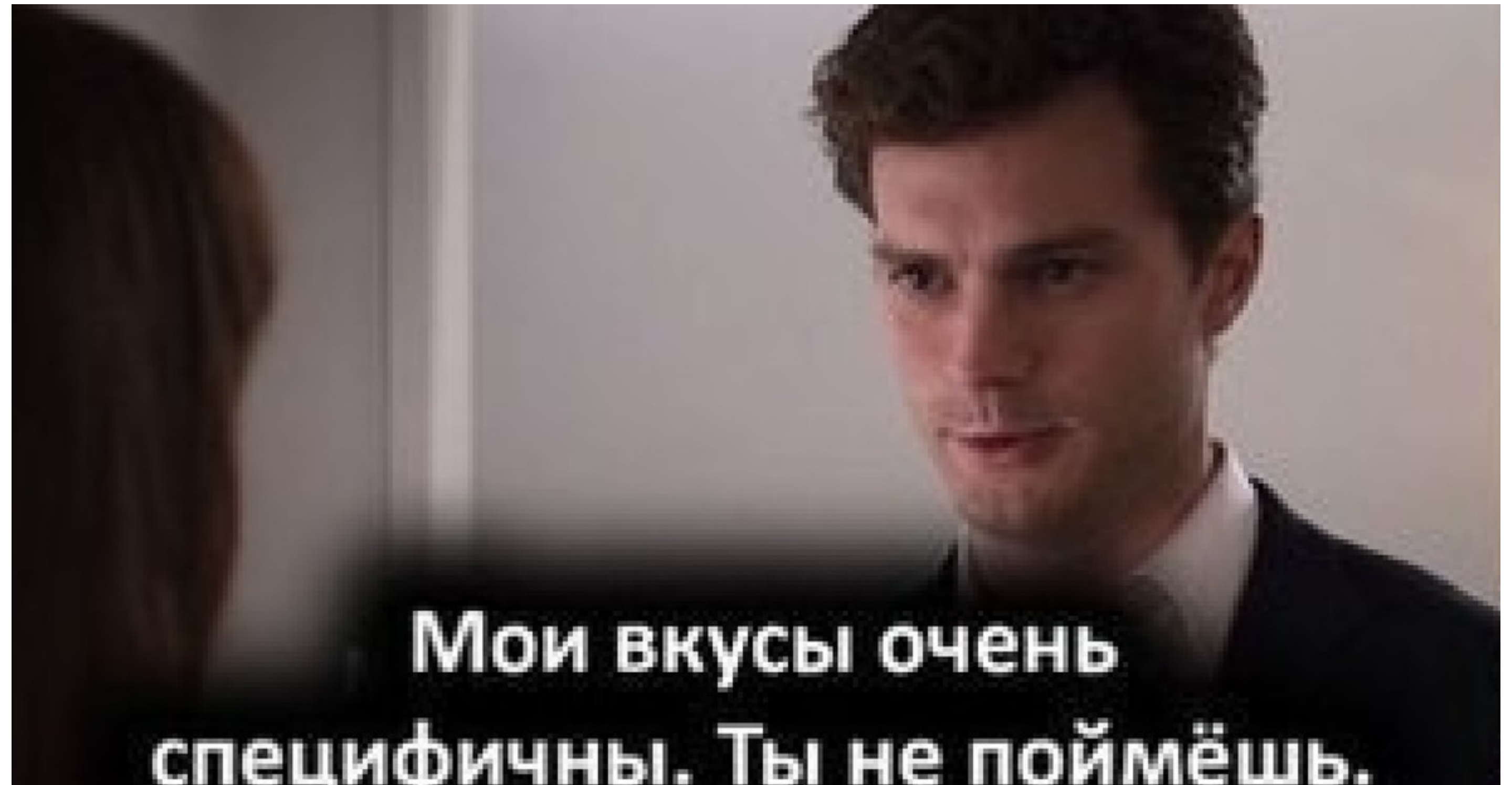
Андрей Павлович

Примитивные методы отчетности

- › Отправка в мессенджер изображения / текста / pdf / excel по расписанию (например, с помощью Apache Airflow)
- › Публикация HTML-файла с отчетом на статическом сервере
- › Написание чат-бота, который будет выполнять запрос к БД по требованию пользователя и отправлять его в чат

Примитивные методы отчетности

› Excel с макросом на VBA, в котором есть обращение к БД



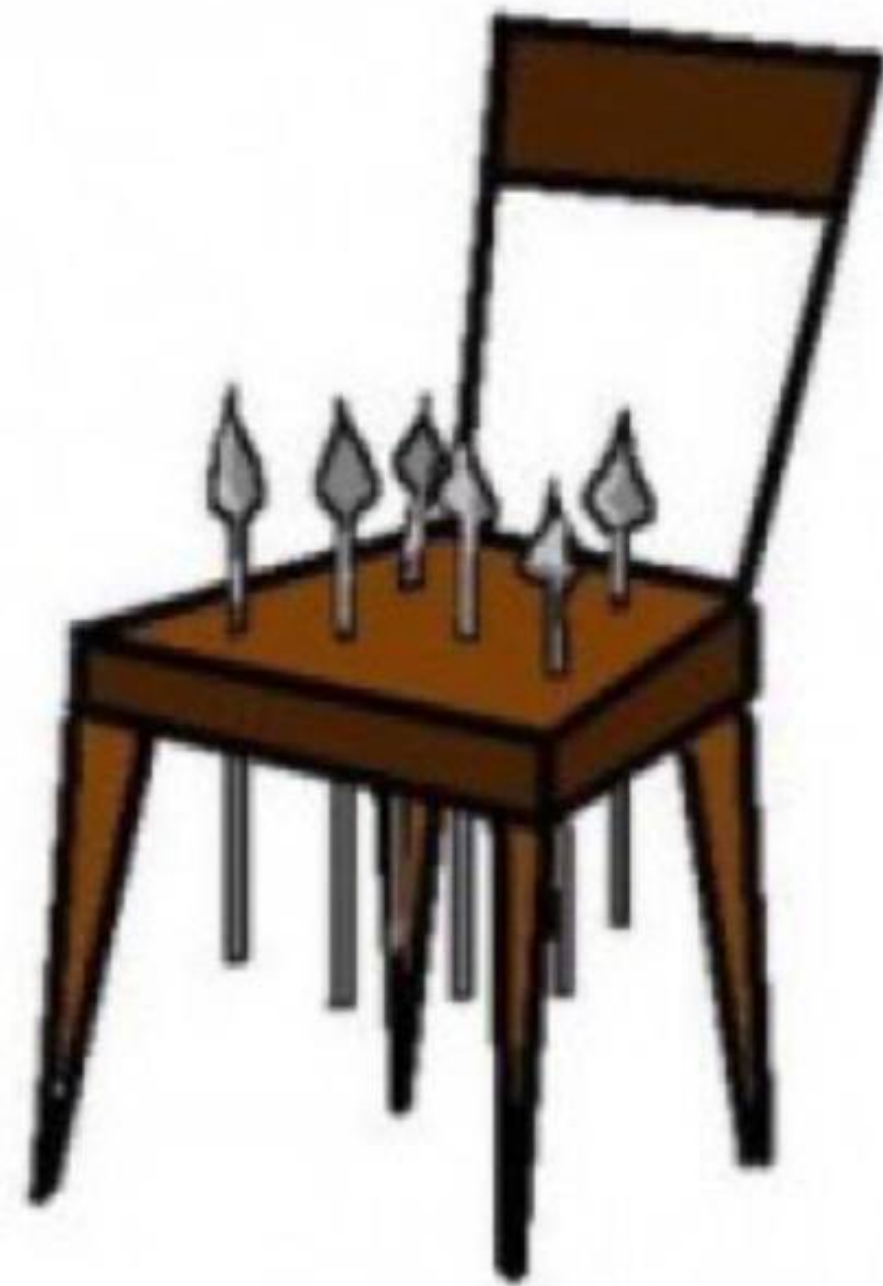
Продвинутые BI- инструменты

Почему Grafana - не абсолютное решение

- › Заточена под работу с realtime и timeseries данными
- › Не умеет в кэширование (из коробки)
- › Набора визуализаций не хватает для использования как полноценного BI
- › Высокий порог входа (нужно знать SQL и понимать хотя бы основы того, как оно работает под капотом)
- › Риск задудосить базу под капотом

Что делать?

У нас снова есть два стула:



Написать своё

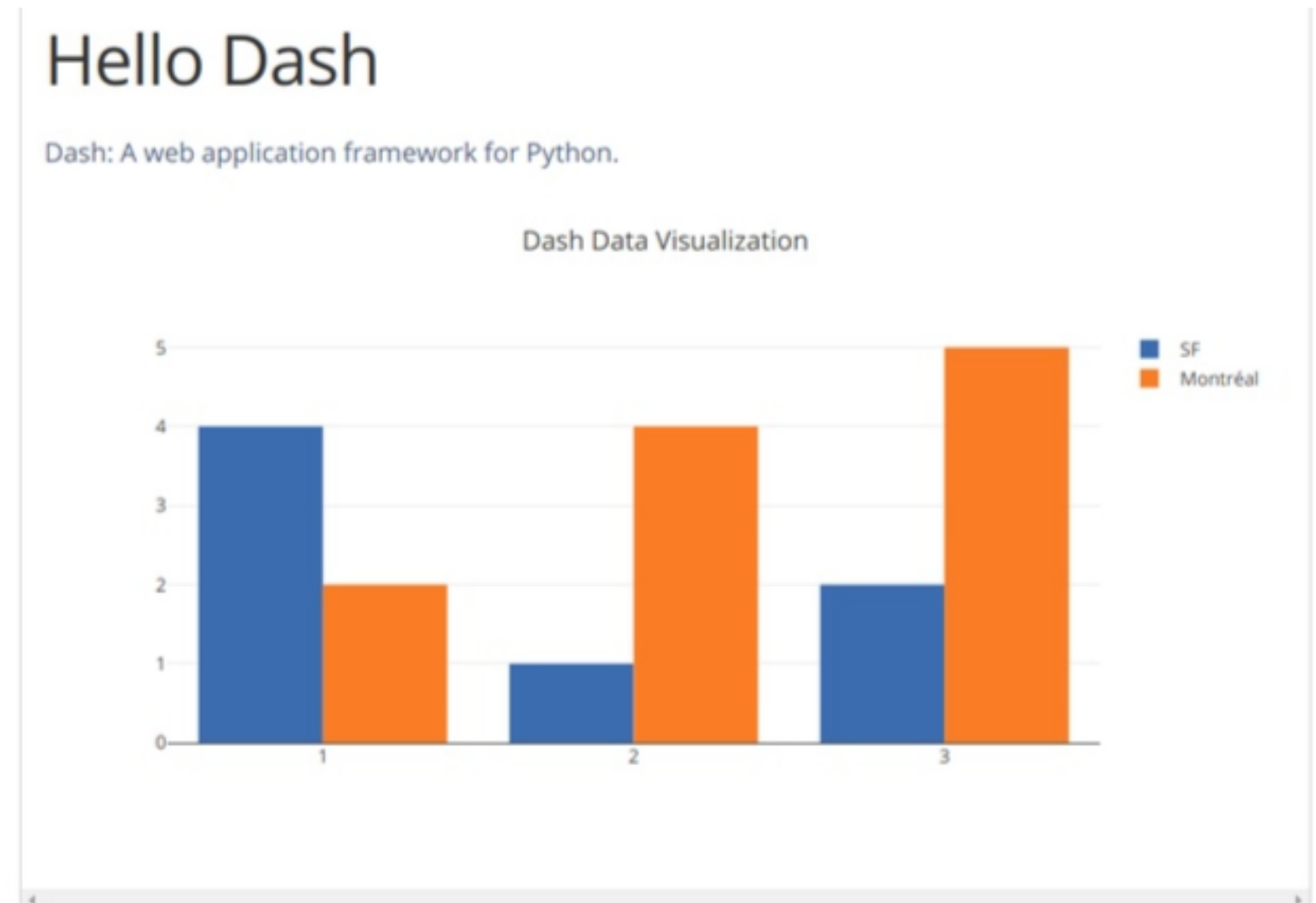


Использовать готовое

Написать своё

Пример:

› Написать что-то свое на plotly dash



Подробнее: <https://habr.com/ru/articles/431754/>

Написать своё

Пример:

› Написать что-то свое на plotly dash

› Использовать streamlit



Подробнее: <https://streamlit.io>

Написать своё

Пример:

- › Написать что-то свое на plotly dash
- › Использовать streamlit
- › Написать что-то свое на flask



Flask

Написать своё

Плюсы:

- › Бесконечные сценарии кастомизации
- › Независимость от возможностей классических BI-фреймворков - если что-то можно сделать какой-то сложной логикой руками, то это всегда можно сделать и на python

Минусы:

- › Доступ только из корпоративной сети, иначе - начинаются пляски с ИБ
- › Аналитики превращаются в фуллстеков - вместо найма обычных стажеров надо искать квалифицированных людей
- › Это дело надо где-то хостить
- › Надо пробрасывать до хоста возможность подключения к БД, что не всегда возможно из-за требований ИБ

А что у нас на другом стуле?

А там - еще два стула:

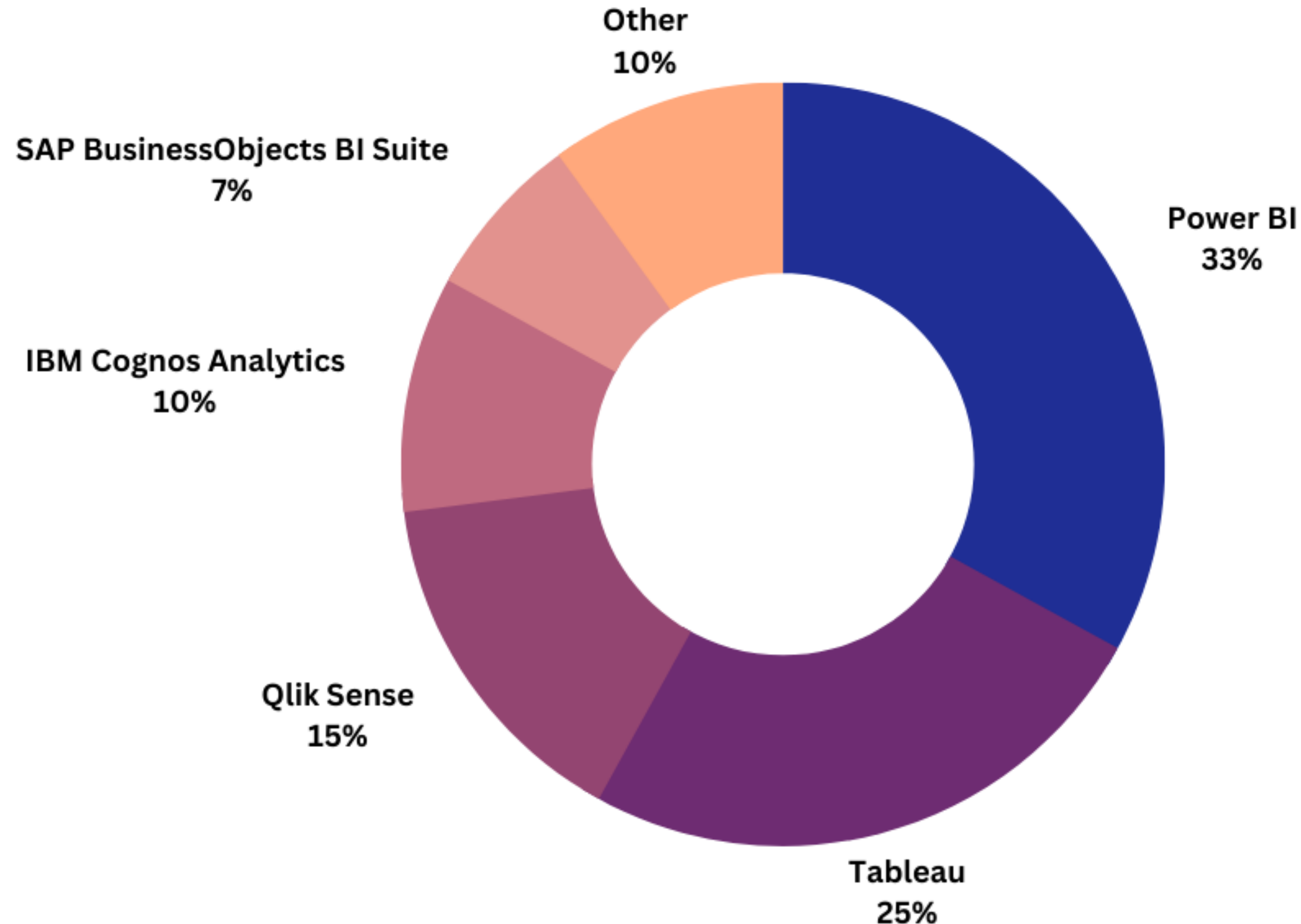


Лицензируемый
софт



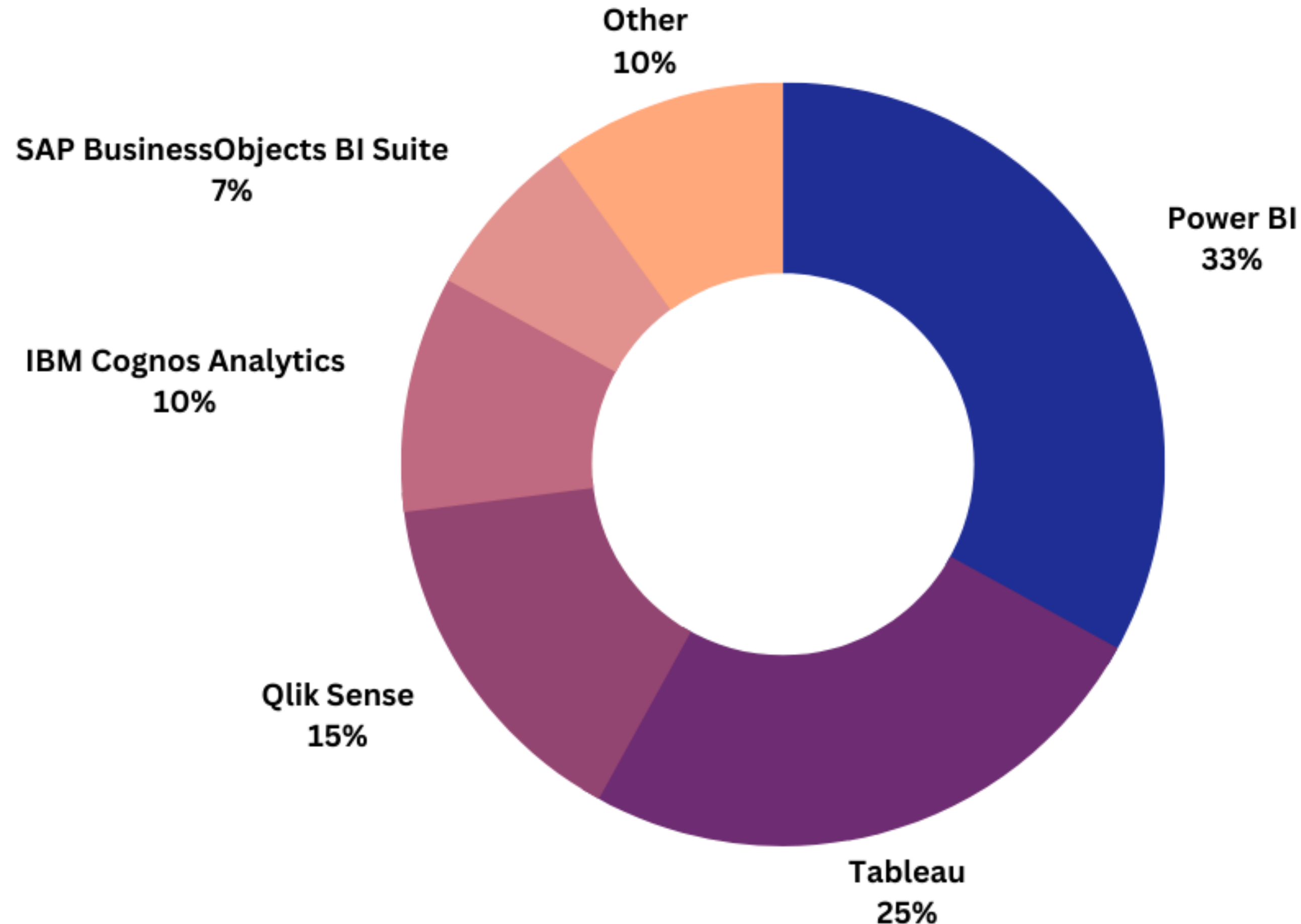
Opensource

Промышленные BI-решения



- › На 2023 безоговорочные лидеры рынка - Power BI и Tableau
- › Почти все лидеры (кроме Tableau) - тесно интегрированные экосистемные решения

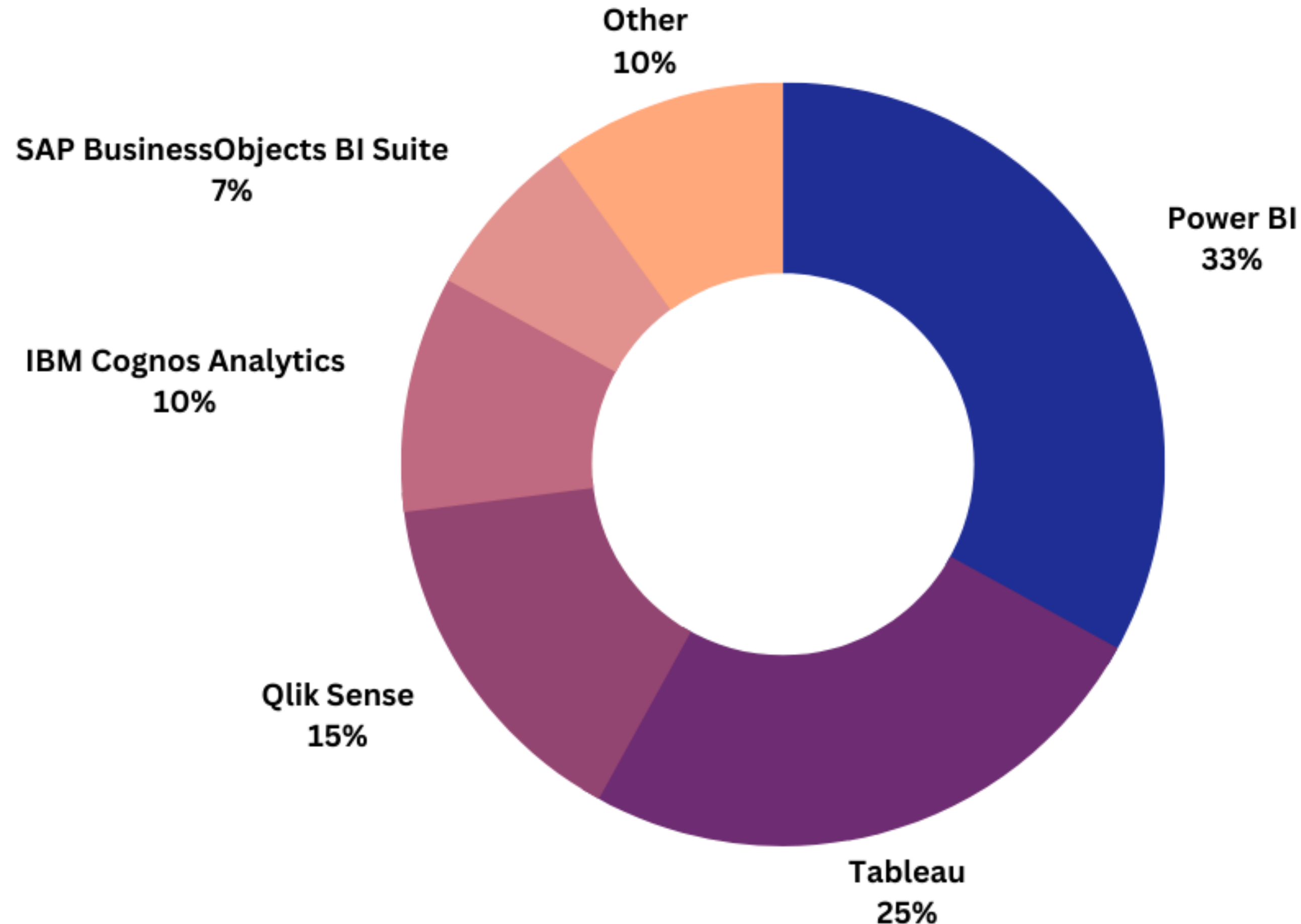
Промышленные BI-решения



Плюсы:

- › Стандарт индустрии: гуманитарии знают, как этим пользоваться, а BI'щики - как с этим работать
- › Минимум затрат на администрирование
- › Кэширование (!)

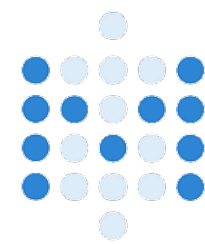
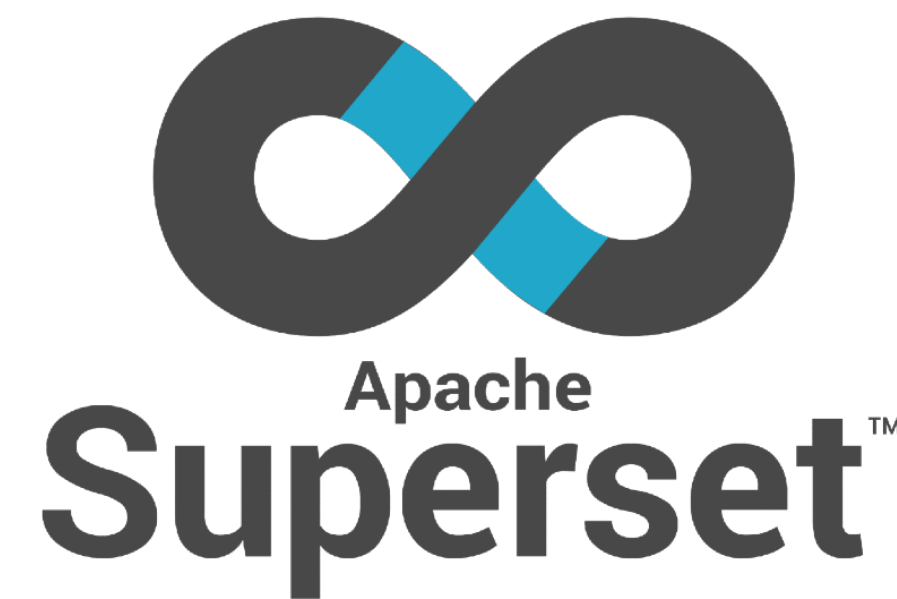
Промышленные BI-решения



Минусы:

- › Очень дорого
PowerBI Pro - 10\$ / m
PowerBI Premium - 20\$ / m
- Tableau Viewer - 15\$ / m,
Explorer - 42\$ / m, Creator -
75\$ / m
- › Санкции
- › Если нет inhouse server -
проблемы с ИБ

А что в opensource



Metabase




А в opensource сейчас в основном встречаются 4 игрока:


- › Datalens (Yandex)
- › Superset (Apache)
- › Metabase
- › Redash

А что в opensource


Google open source bi tools

 Logz.io
https://logz.io › blog › busin... · [Перевести эту страницу](#)


18 Free and Open Source Business Intelligence Tools
18 Free and **Open Source Business Intelligence Tools** · 1. BIRT · 2. ClicData · 3. The ELK Stack · 4. Helical Insight · 5. Jedox · 6. JasperReports Server · 7.

 Holistics
https://www.holistics.io › blog · [Перевести эту страницу](#)


12 Best Open Source BI Tools Data Teams Recommend
23 апр. 2022 г. — II. 12 Free & **Open-source BI Tools** Data Teams Love · 12. FineReport · 11. Abixen · 10. ART - A Reporting Tool · 09. Tableau Public · 08.

 Monte Carlo Data
https://www.montecarlodata.com › ... · [Перевести эту страницу](#)


Top Open Source BI Tools In 2023 (Quick Reference Guide)
10 февр. 2023 г. — **Open Source Business Intelligence Tools** · Eclipse BIRT · Apache Superset · Seal Report & ETL · Jaspersoft Business Intelligence Suite Community ...

 Netguru
https://www.netguru.com › о... · [Перевести эту страницу](#)

Top 5 Commercial And Open-Source BI Tools
17 окт. 2023 г. — The best options for your **business intelligence** projects include Tableau, Microsoft Power **BI**, Qlik, and Looker, while more niche **tools** include ...

 MEDevel.com
https://medevel.com › bi-das... · [Перевести эту страницу](#)

22 Open-source Business Intelligence (BI) Dashboards
2 июл. 2023 г. — Some popular **open-source BI** dashboards include Apache Superset, Metabase, and Redash. These **tools** offer a range of features, including data ...

 Dataconomy
https://dataconomy.com › ор... · [Перевести эту страницу](#)

Top 15 Open Source Business Intelligence Software
10 мая 2023 г. — SpagoBI is a comprehensive **open-source business intelligence** suite that comprises various **tools** for reporting, charting, and data-mining. The ...

› BI-тулзов в Opensource - как грязи

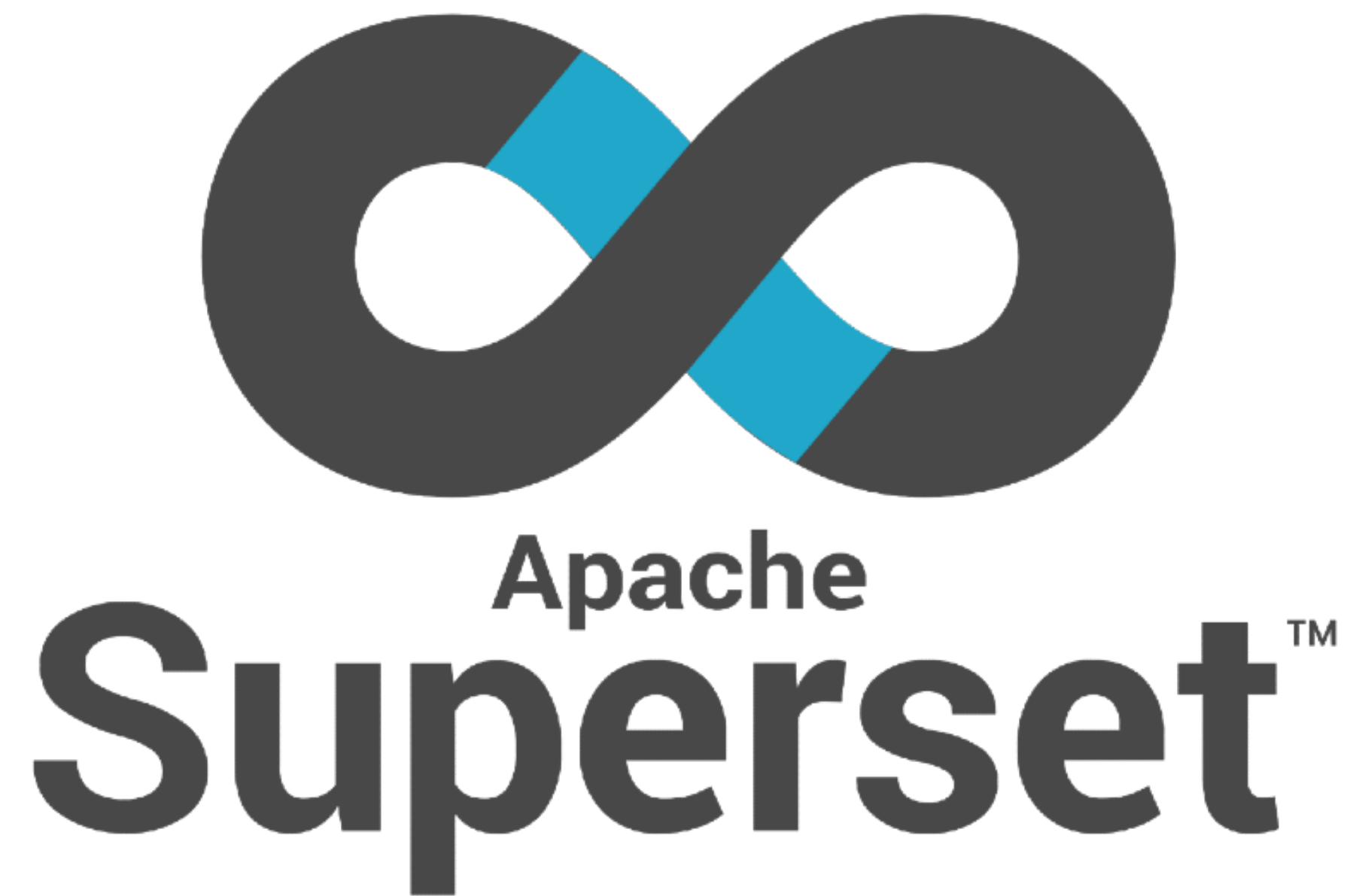
› Вопрос - что из этого действительно работает, чему можно доверять и на чем хочется строить свое решение

Apache Superset



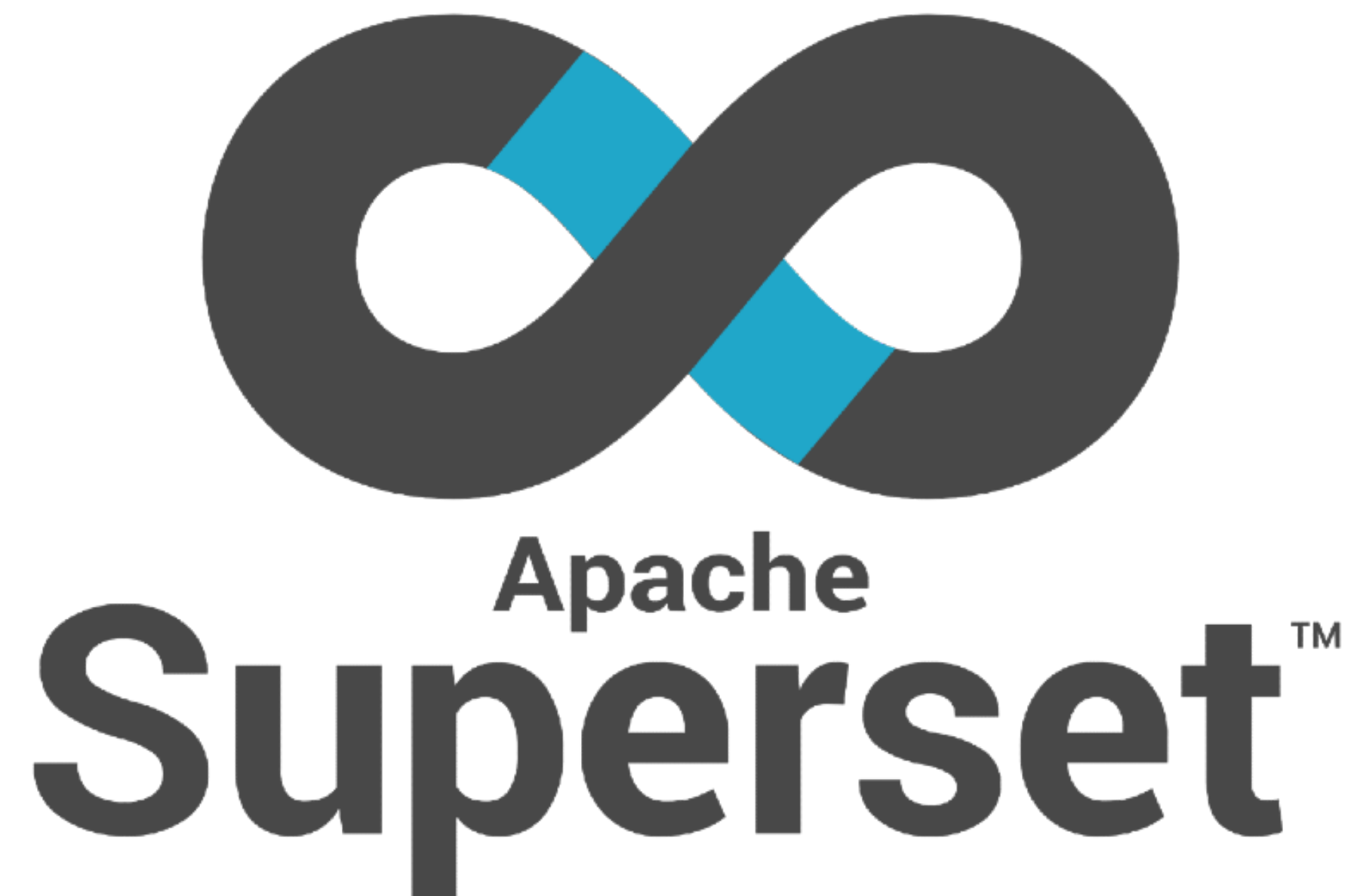
- › Создали в Airbnb (у истоков - тот же Maxime Beauchemin, что стоял и у истоков Airflow)
- › 2017 - релиз в составе Apache Incubator
- › 2021 - core-проект Apache Software Foundation
- › Используется в Airbnb, Dropbox, Lyft, Netflix и X (Twitter) как основной BI-инструмент
- › В России точно знаю про OZON и Avito

Apache Superset



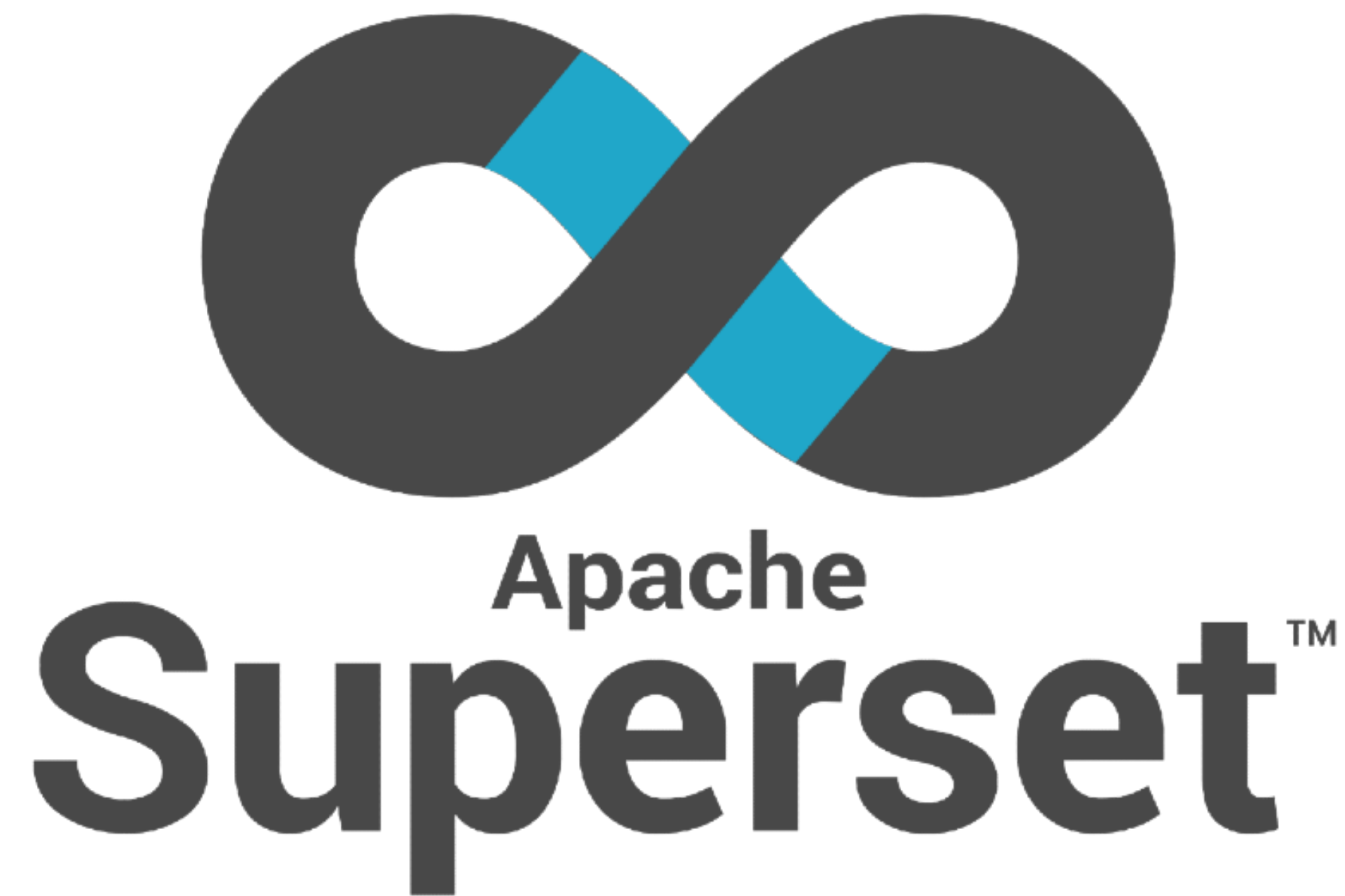
- › Бэкенд - Python
- › Фронтэнд - JS
- › Писать модули и интеграции в него способен любой первокурсник ПТУ

Apache Superset



- › Бэкенд - Python
- › Фронтэнд - JS
- › Подключения - все, у чего есть API, или что может съесть SQLAlchemy
- › Огромная библиотека пользовательских методов визуализации
- › Из коробки поддерживает нужные корпоративным клиентам интеграции (например, SSO)

Apache Superset



Минусы:

- › Высокий порог входа (нужно знать SQL и Jinja)
- › Нет кэширования - бэкенд нещадно спамит базу запросами от каждого чарта
- › Внутренняя архитектура, не сильно рассчитанная на highload

Datalens

- Дашь домашку списать?

- Дам, но списывай не точь-
в-точь, чтобы не спалили

- Ок



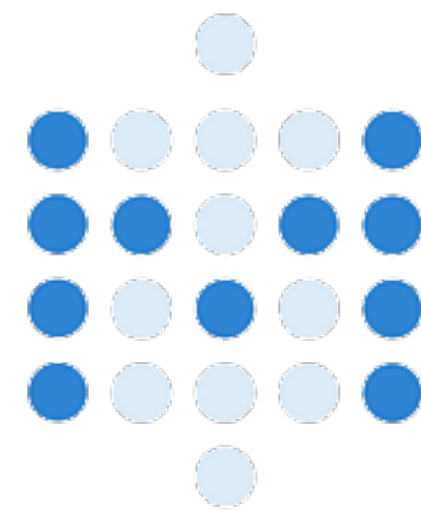
› Русский BI (!)

› Бэкенд - Python

› Фронтэнд - JS

› Подозрительно похож на Apache Superset

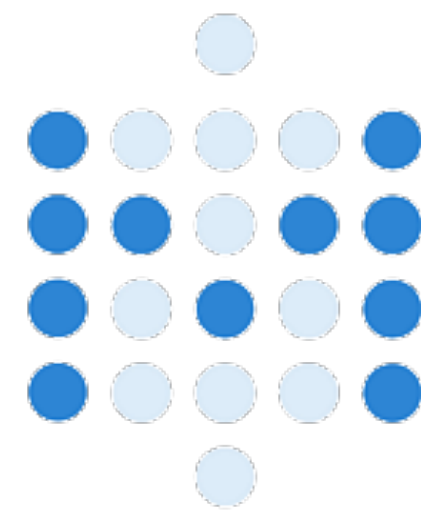
Metabase



Metabase

- › Сразу писался как Open Source проект командой под руководством Sameer Al-Sakran в Exra Ventures
- › В 2015 публикуются первые релизы на Github
- › Крупные пользователи: N26, Revolut
- › Позиционируется как BI для людей, и имеет практически нулевой порог входа

Metabase

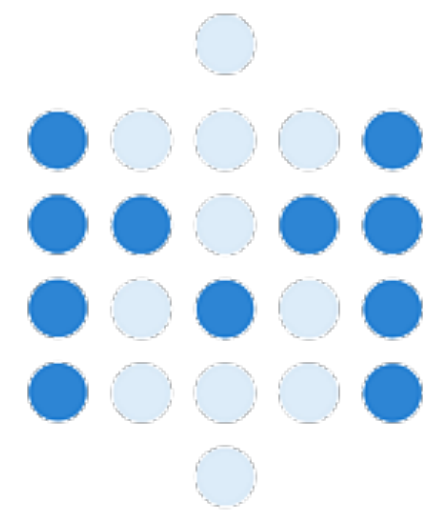


Metabase

Плюсы:

- › Нулевой порог входа
- › Легко разворачивается inhouse
- › Абсолютно не требователен к обслуживанию или администрированию, поэтому часто используется в маленьких (даже не IT) компаниях, стартапах и других местах, где надо быстро и дешево внедрить аналитику

Metabase



Metabase

Минусы:

- › Все, что легко для пользователя - сложно для базы данных
- › Встроенный генератор запросов работает очень топорно
- › Чтобы делать сценарии сложнее, чем дэшборд с фильтрами на колонки - нужно уже изучать SQL / Jinja / JS