Аналитика Больших данных: технологии

С.Д. Белов, ЛИТ ОИЯИ
belov@jinr.ru

Осенняя Школа по информационным технологиям ОИЯИ 2022
14-19 ноября 2022г., ЛИТ ОИЯИ, Дубна

Современные тенденции в анализе и управлении данными

Gartner Top Data and Analytics Trends for 2022

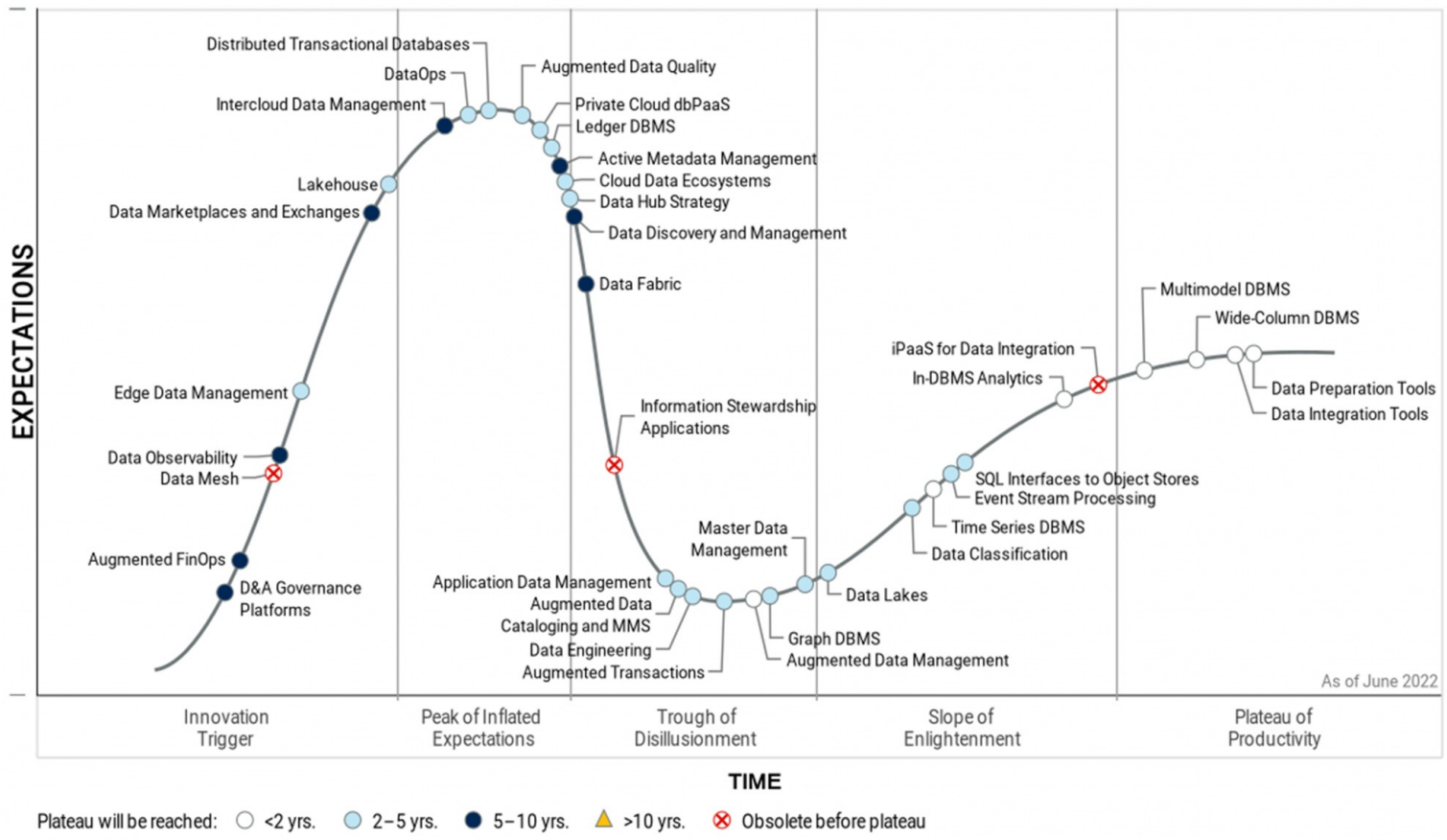
Activate Dynamism and Diversity	Augment People and Decisions	Institutionalize Trust
• Adaptive AI Systems	• Context-Enriched Analysis	• Connected Governance
• Data-Centric AI	• Business-Composed D&A	• AI Risk Management
• Metadata-Driven Data Fabric	• Decision-Centric D&A	• Vendor and Regional Ecosystems
• Always Share Data	• Skills and Literacy Shortfall	• Expansion to the Edge

gartner.com

Gartner

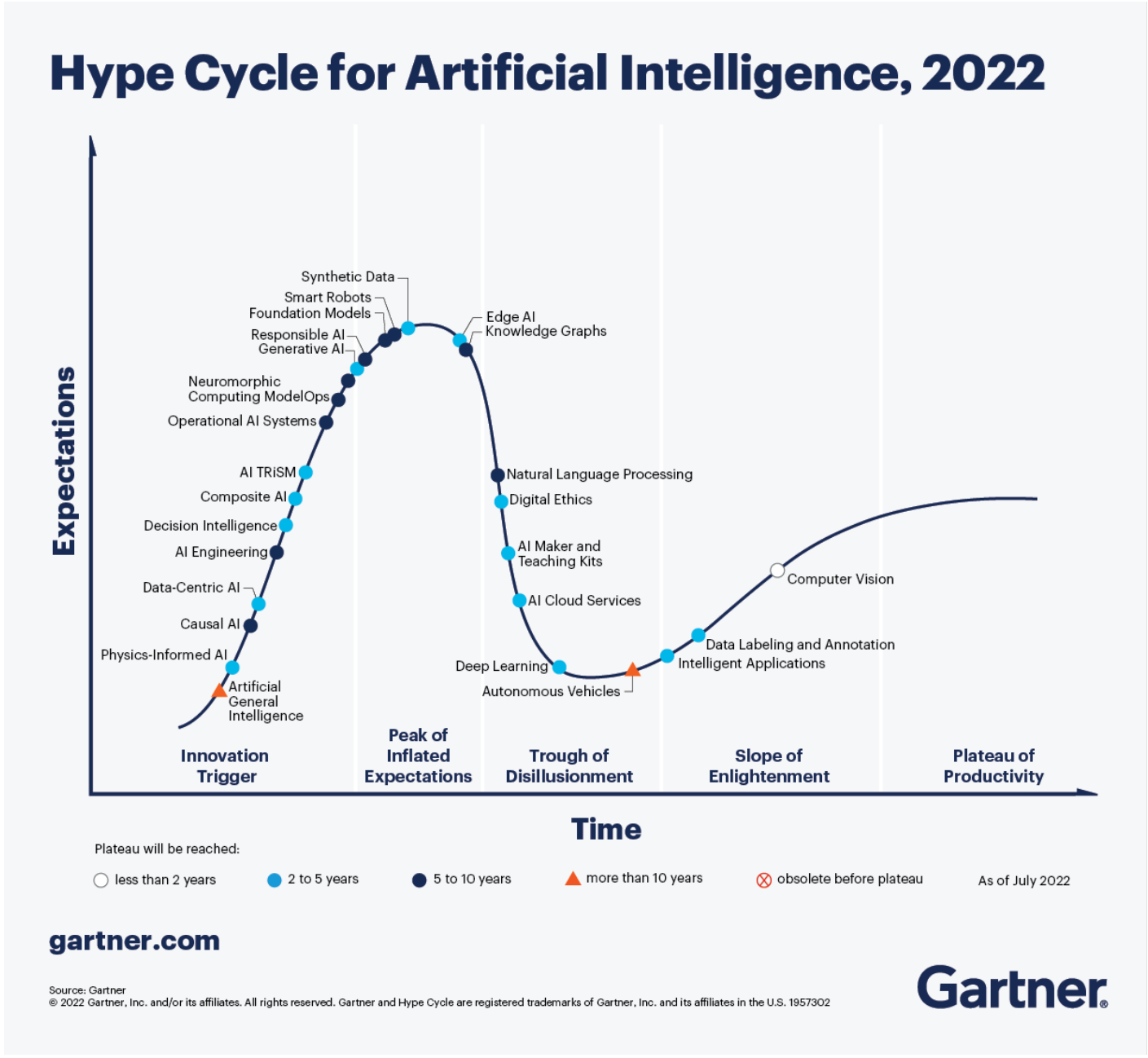
https://www.gartner.com/en/articles/12-data-and-analytics-trends-to-keep-on-your-radar

Hype Cycle for Data Management, 2022



Source: Gartner (June 2022)
https://www.denodo.com/en/document/analyst-report/2022-gartner-hype-cycle-data-management

Современные тенденции развития технологий искусственного интеллекта



https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2022-gartner-hype-cycle

Классификация анализа Больших данных по типам использования

Операционная деятельность

TICKET BOOKING

ONLINE SHOPING

SOCIAL MEDIA

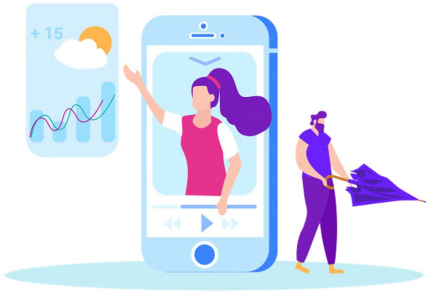
ORGANISATION

Аналитика

STOCK MARKET

SPACE MISSION

WEATHER FORECAST



MEDICAL

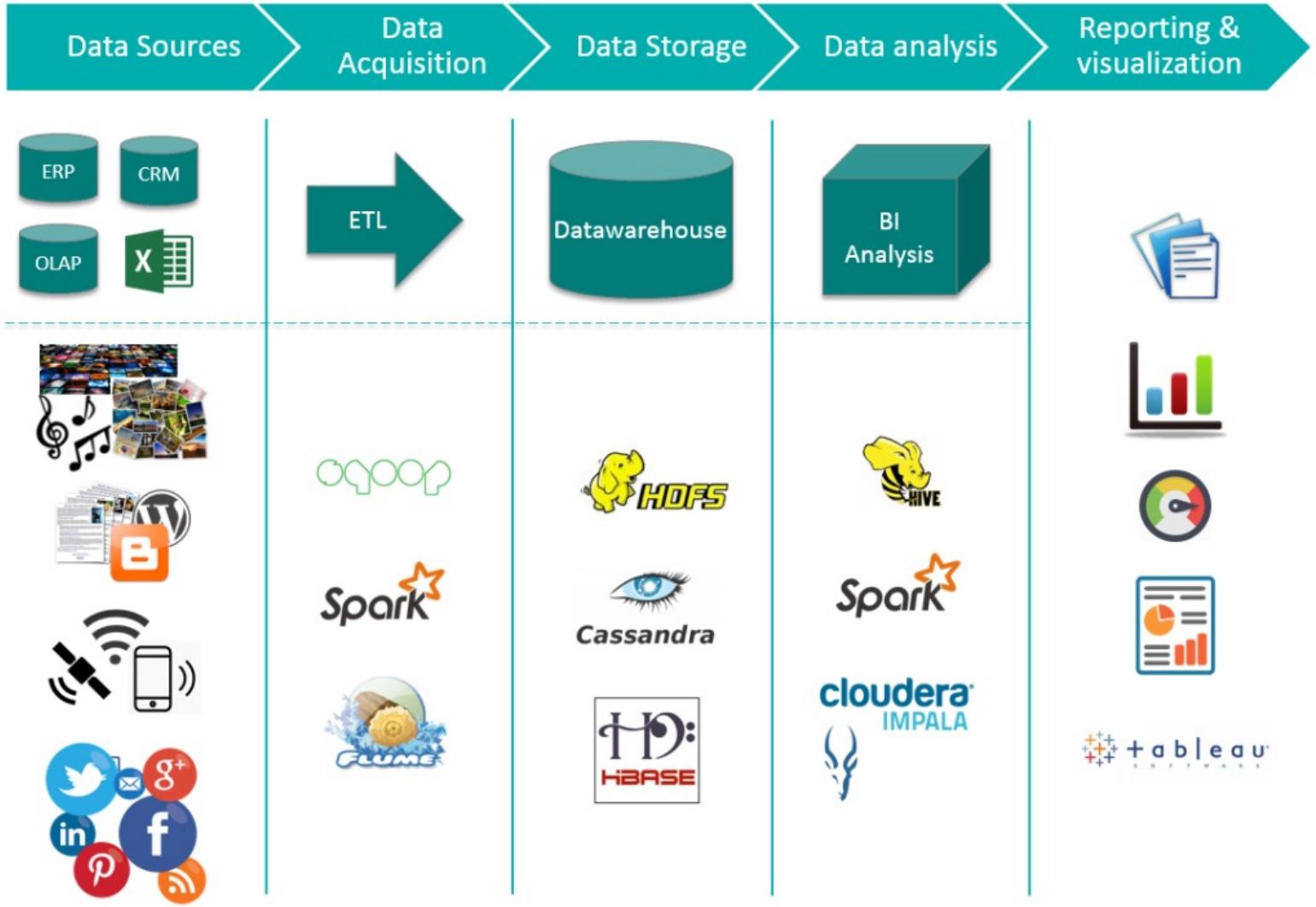
https://www.edureka.co/blog/top-big-data-technologies/

Классификация технологий по функционалу

- **Основные группы технологий Больших данных**
- Хранилища данных (Data Storage)
- Извлечение/добыча данных (Data Mining)
- Аналитика (Data Analytics)
- Визуализация (Data Visualization)

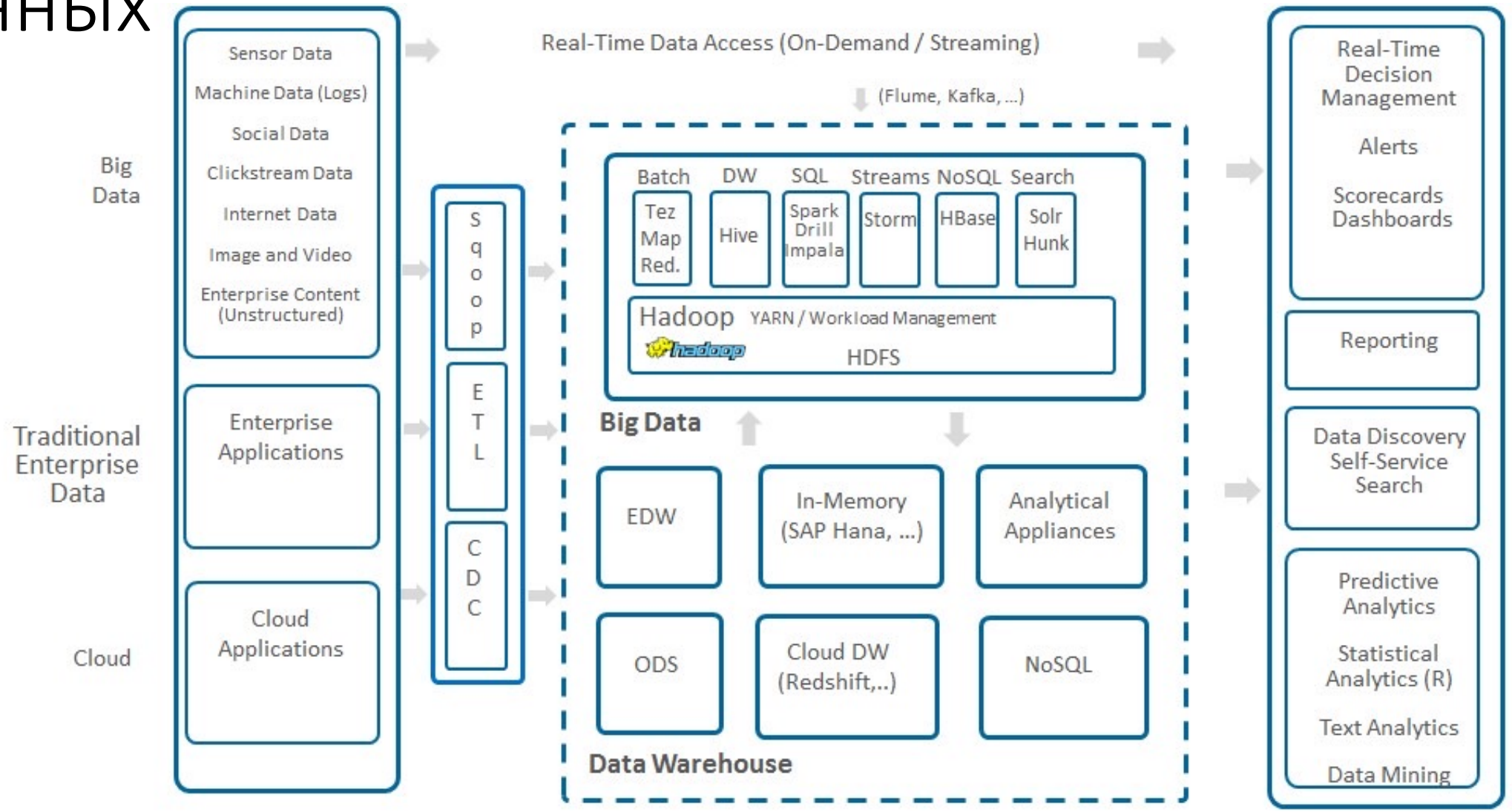
https://www.edureka.co/blog/top-big-data-technologies/

Пример цепочки анализа Больших данных



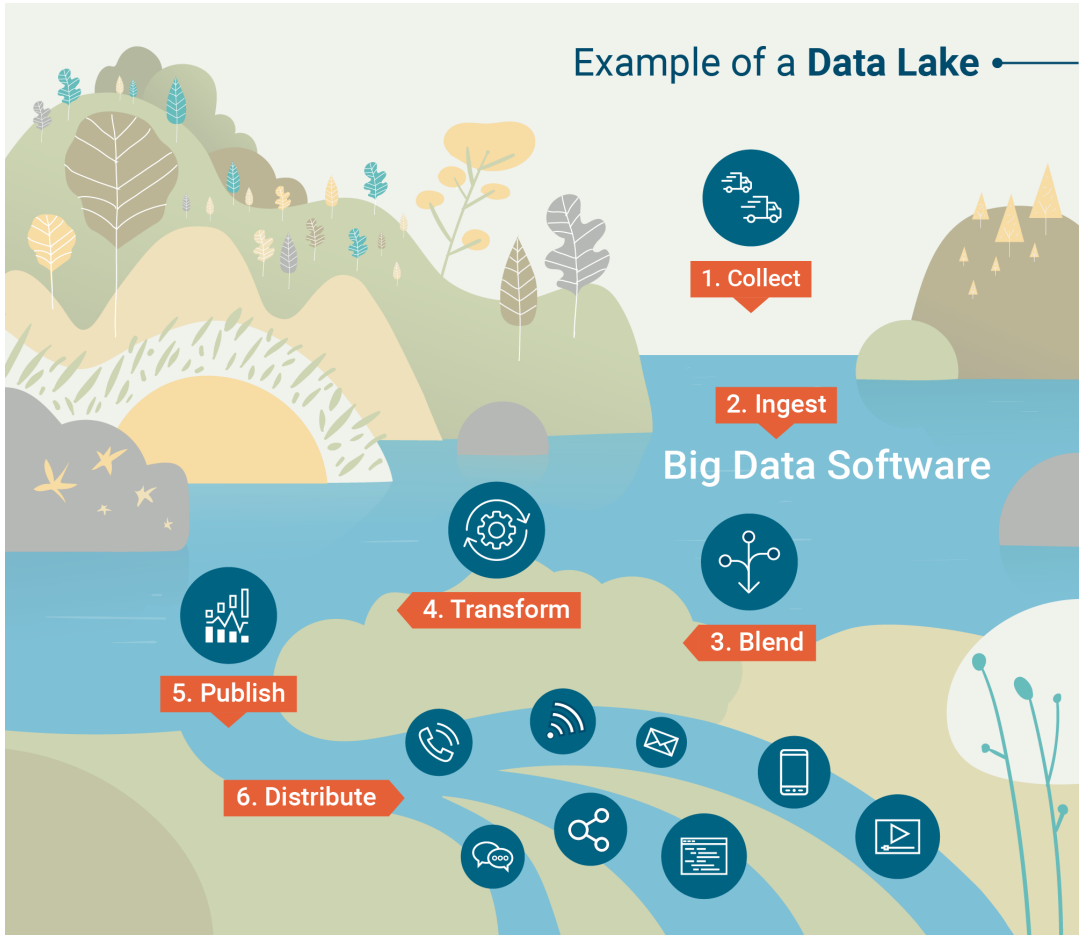
https://www.clearpeaks.com/big-data-ecosystem-spark-and-tableau/

Пример архитектуры обработки Больших данных



https://www.datavirtualizationblog.com/logical-architectures-big-data-analytics/

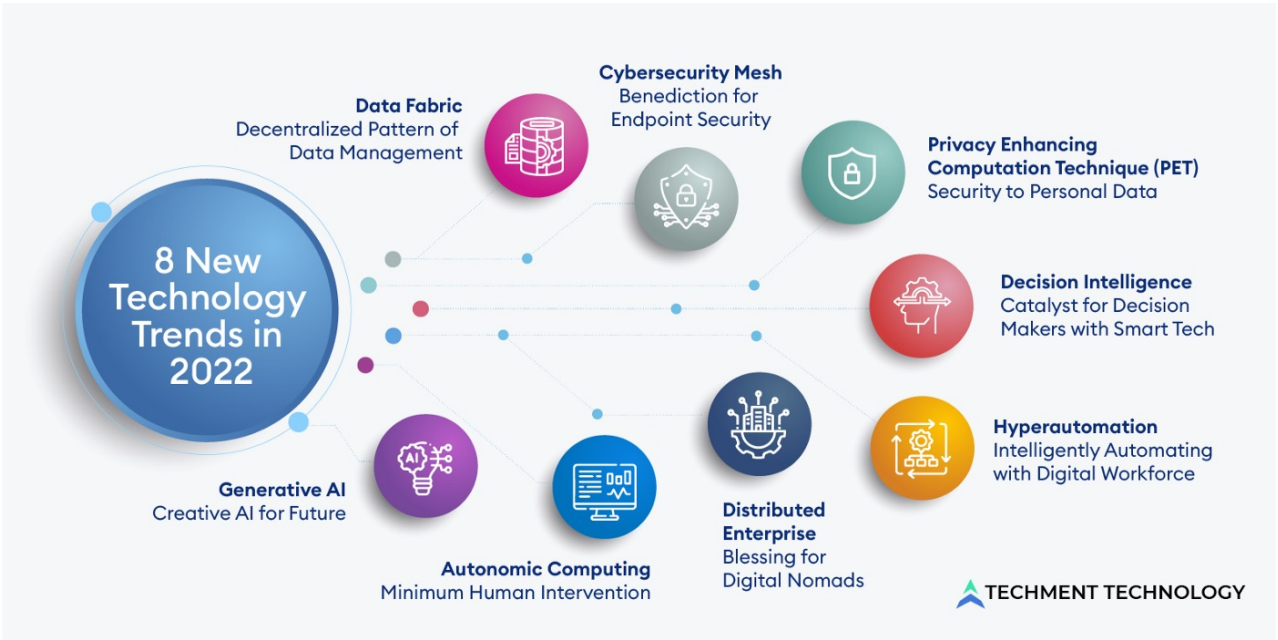
Озёра данных (Data Lakes)



https://www.g2.com/articles/what-is-a-data-lake

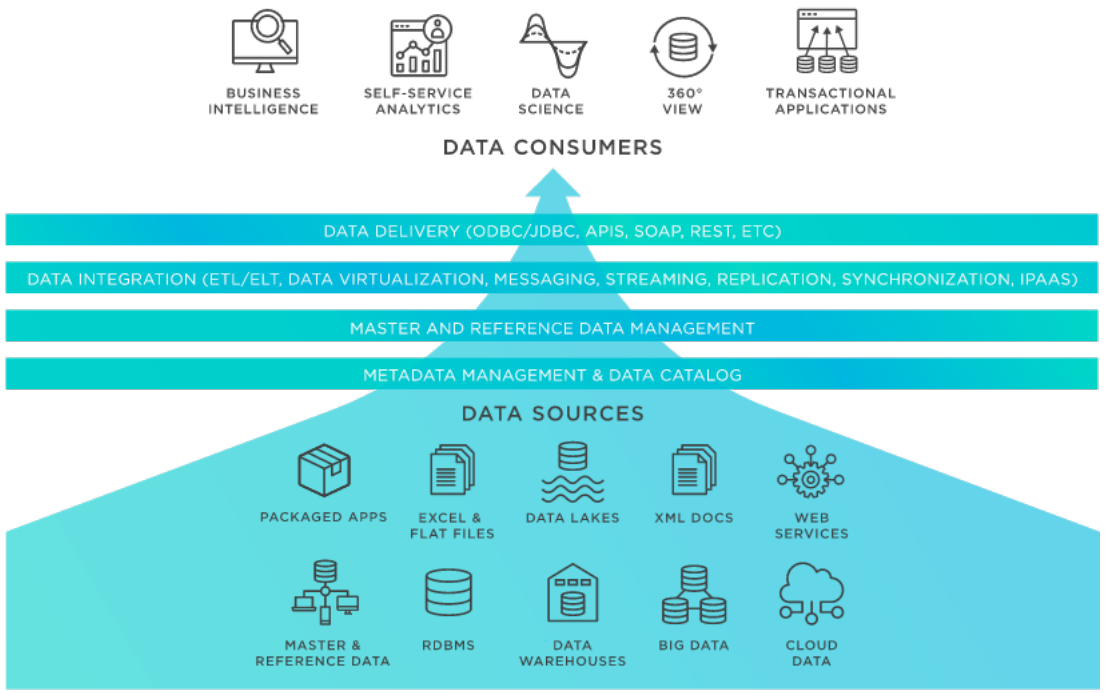
https://vitalflux.com/data-lake-design-principles-best-practices/

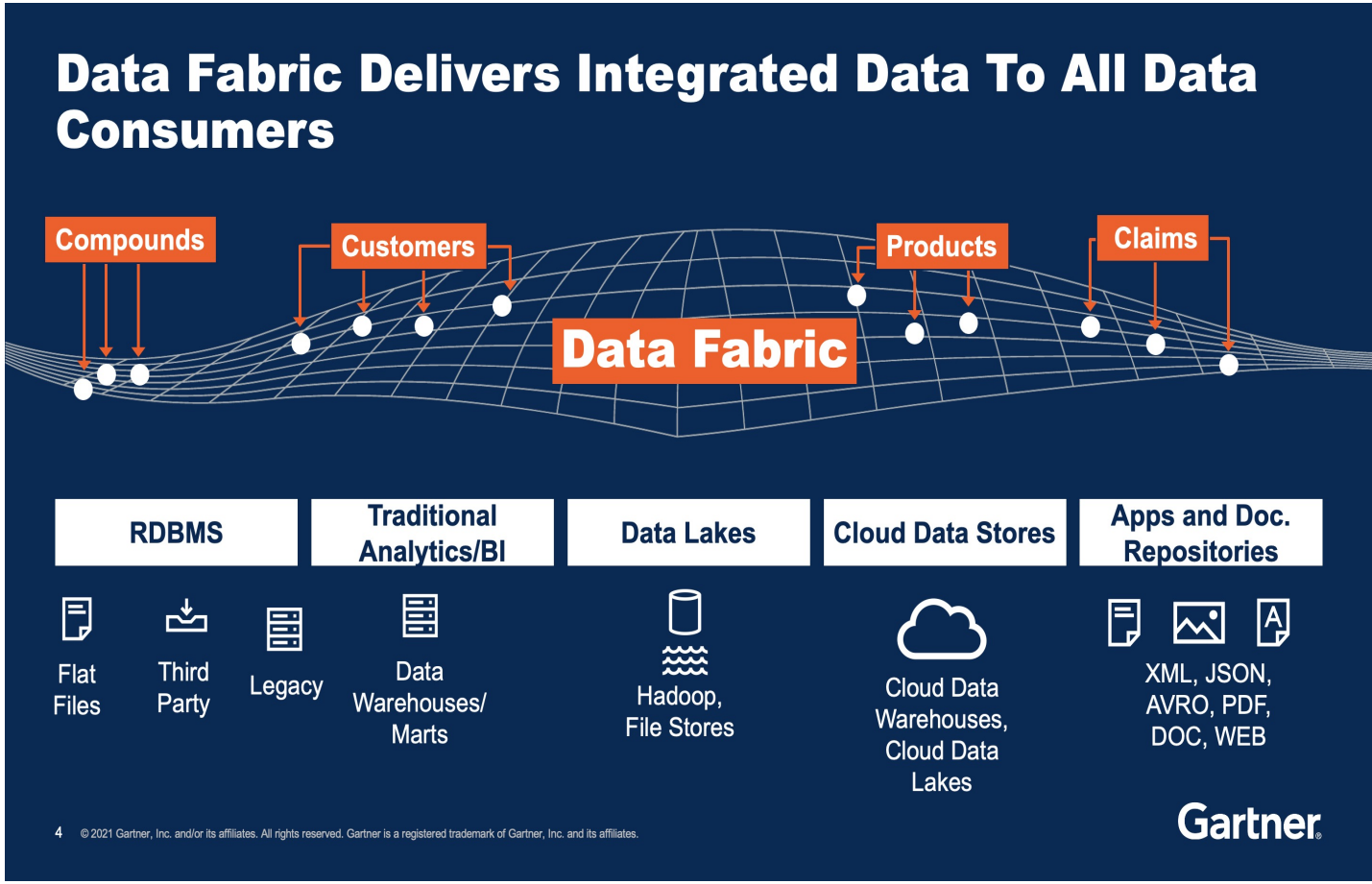
Некоторые технологические тренды



https://techment.medium.com/ctos-must-bet-on-these-future-technology-trends-for-digitalization-and-sustainable-business-growth-a3163dc5e2f3

Data Fabric

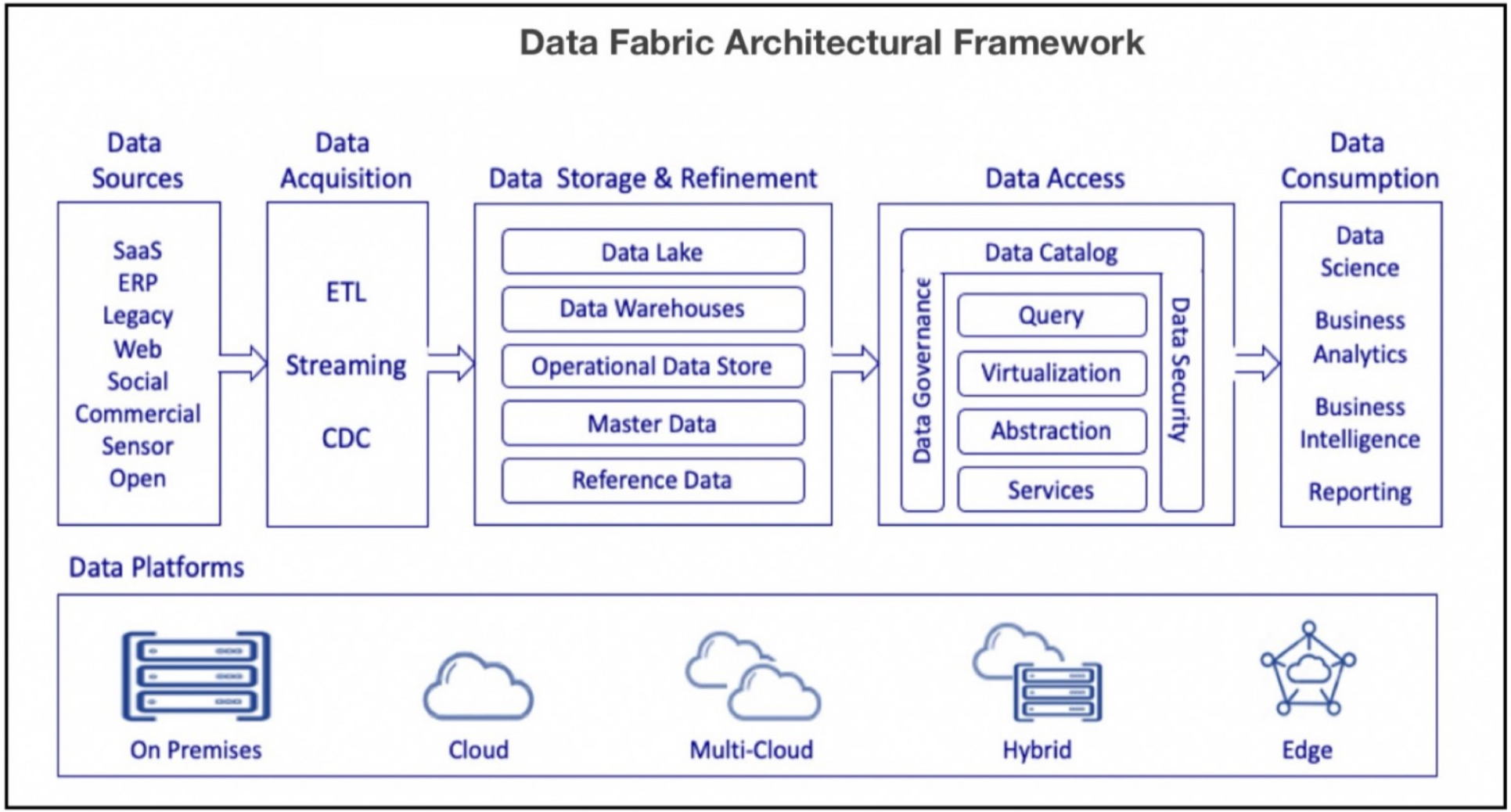




https://www.tibco.com/reference-center/what-is-data-fabric

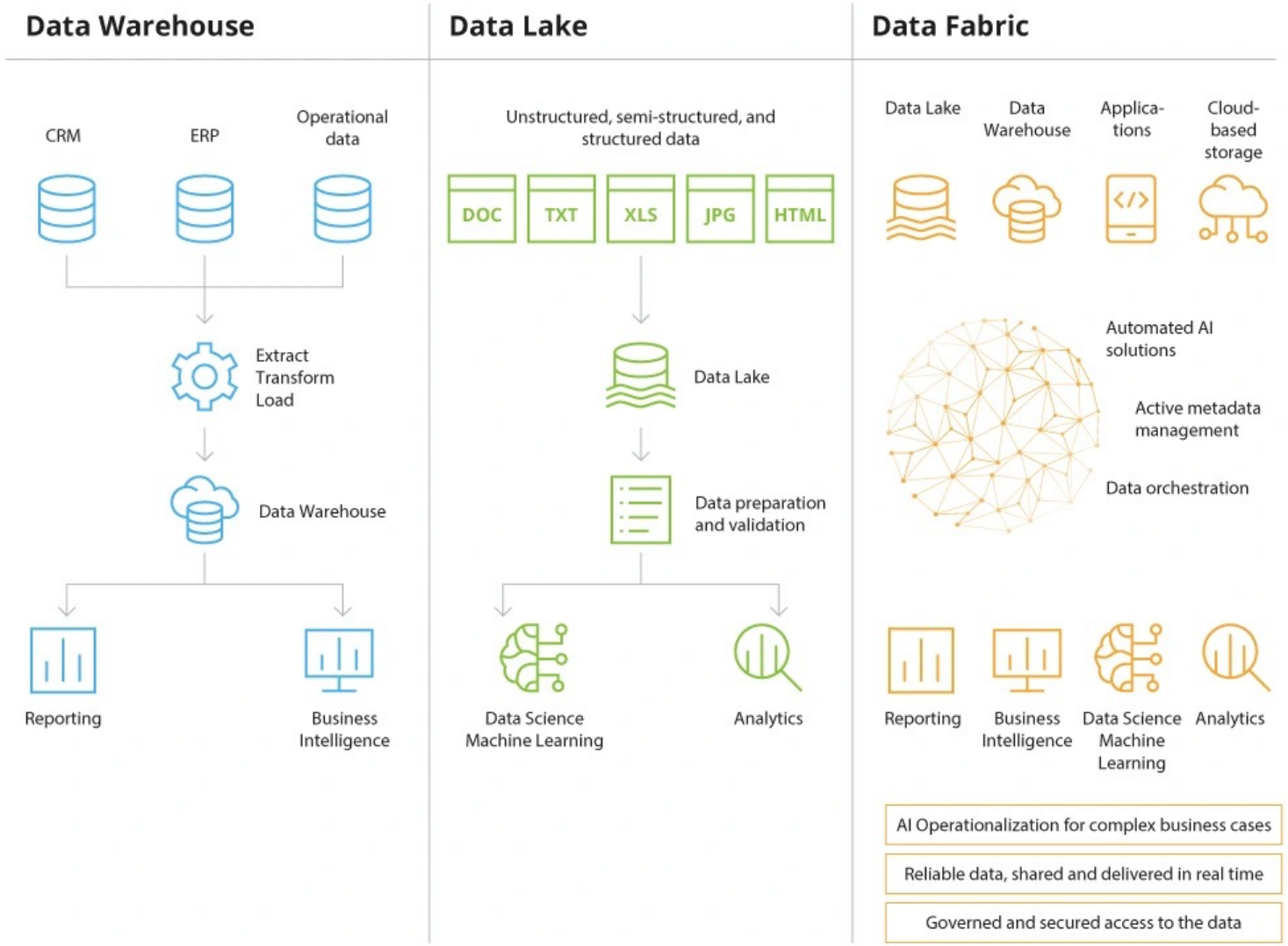
https://www.k2view.com/what-is-data-fabric

Архитектура Data Fabric



https://habr.com/ru/post/647379/

Сравнение Data Warehouse, Data Lake и Data Fabric

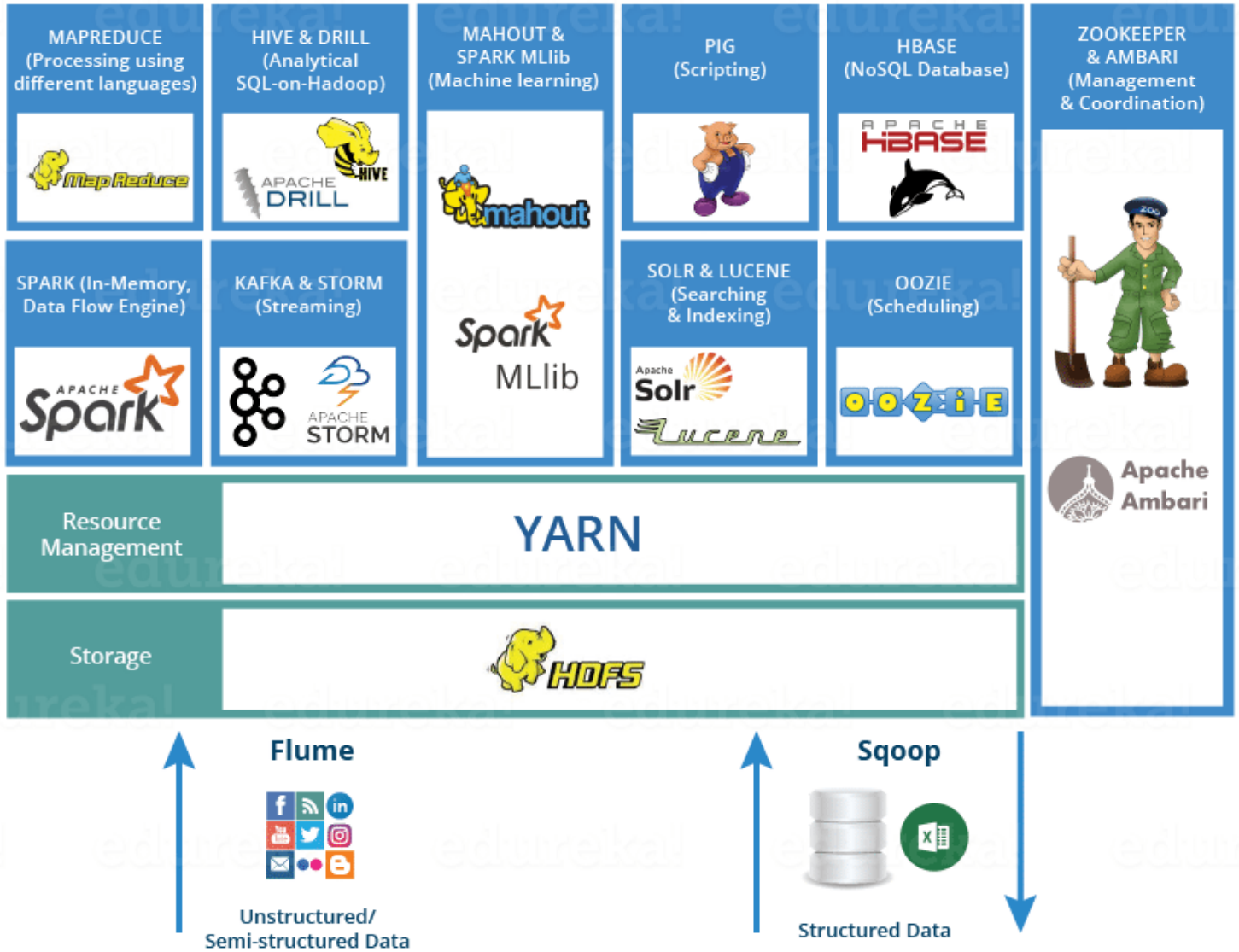


https://www.infopulse.com/blog/dwh-vs-data-lake-vs-data-fabric

Пакеты с открытым исходным кодом

https://informationit27.medium.com/big-data-open-source-tools-150d6a68214a

Экосистема Hadoop



https://www.edureka.co/blog/hadoop-ecosystem

Представление и хранение данных

https://hadoop.apache.org/

- Фреймворк для разработки и выполнения распределенных программ

https://www.mongodb.com /

- Документоориентированная нереляционная (NoSQL) СУБД

https://cassandra.apache.org/

- Распределённая нереляционная СУБД

Представление и хранение данных

https://kylin.apache.org/

- Интеграция различных источников данных, предоставление SQL-интерфейса и аналитика

https://calcite.apache.org/

- Динамическое управление разнородными данными, интеграция данных

https://ignite.apache.org/

- Распределённая СУБД с хранением данных в памяти (для высокопроизводительных вычислений)

Представление и хранение данных

https://neo4j.com/

- Графовая база данных

https://hbase.apache.org/

- Масштабируемая распределённая база данных для работы с большими данными

Извлечение/добыча данных – Data Mining

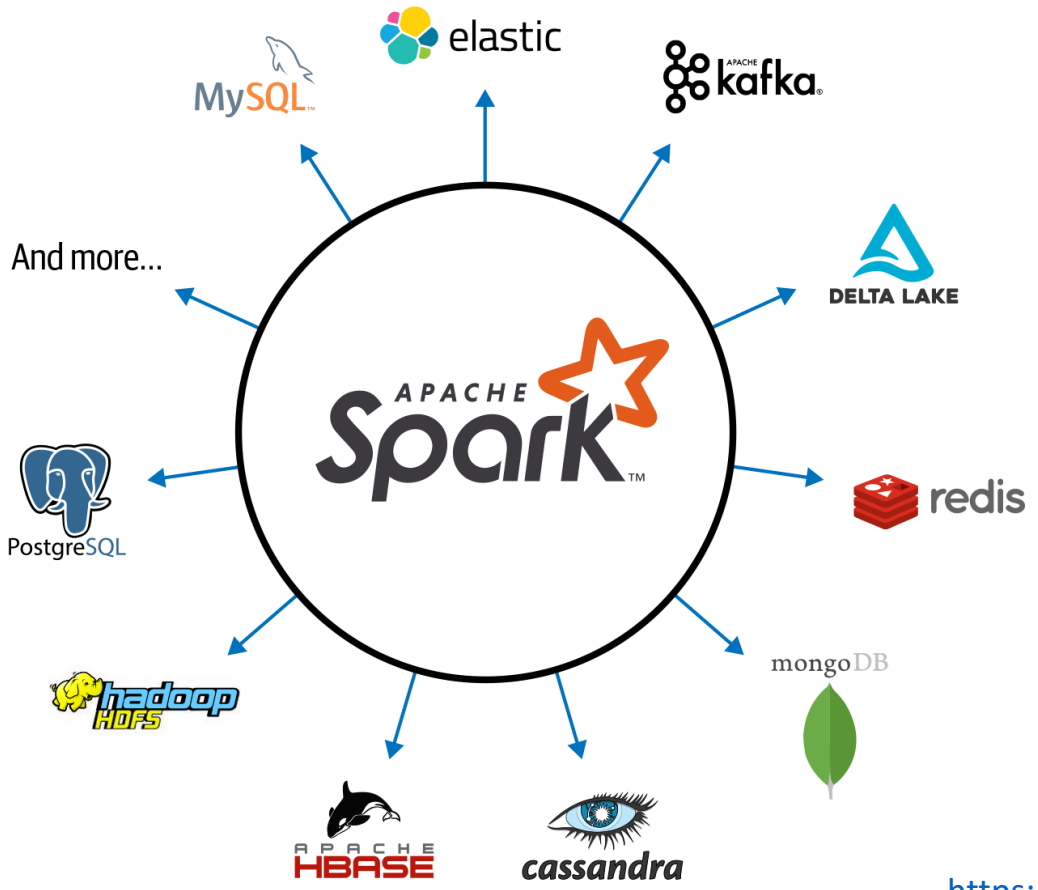
https://prestodb.io/

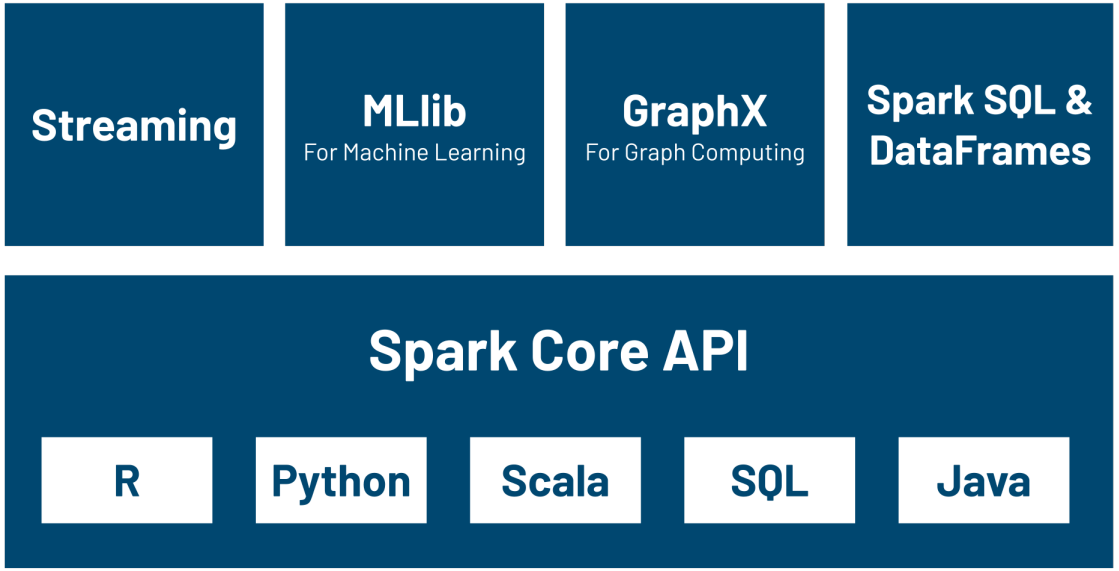
https://github.com/rapidminer/

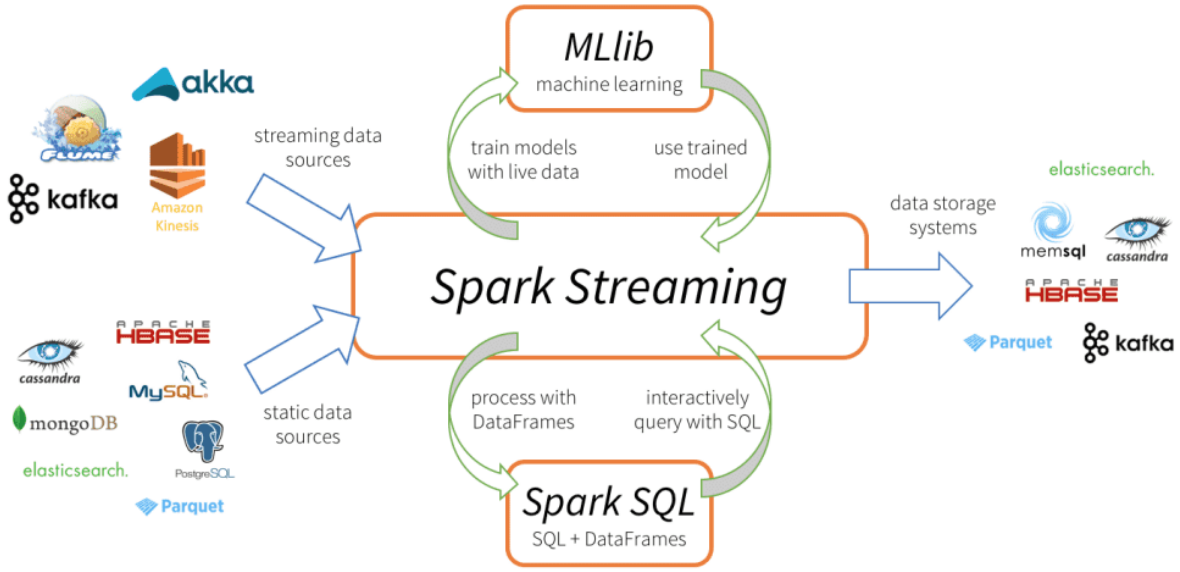
https://www.elastic.co/

- Система SQL-запросов к различным СУБД (в том числе нереляционным)
- Платформа для анализа данных
- Программная поисковая система

Анализ данных – Apache Spark

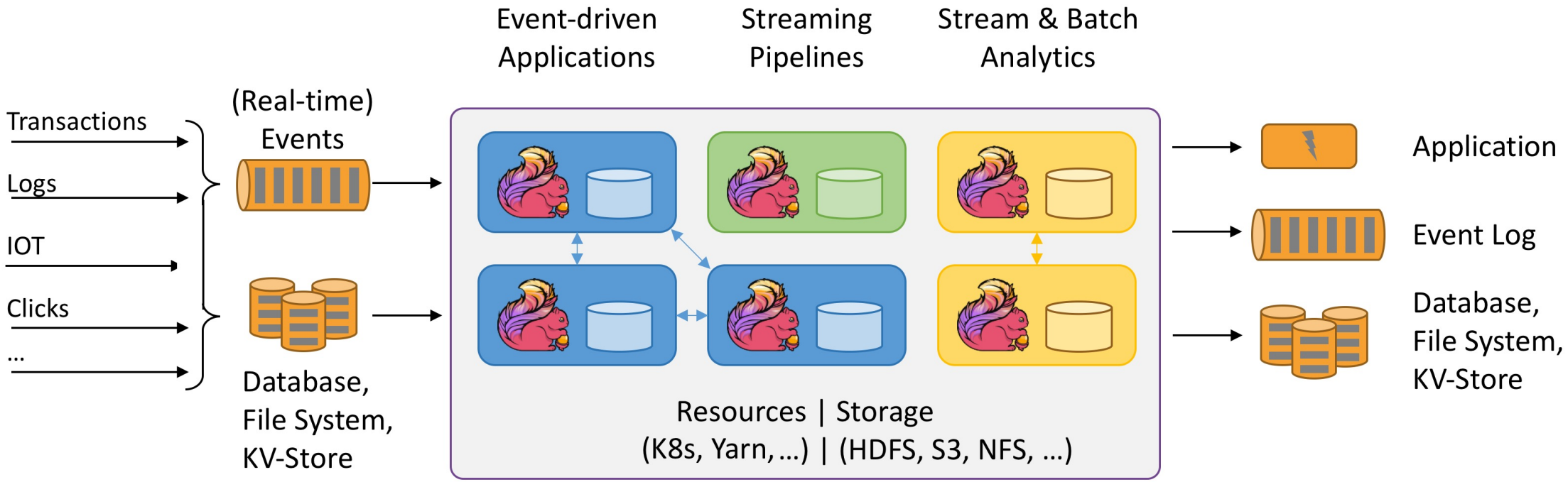






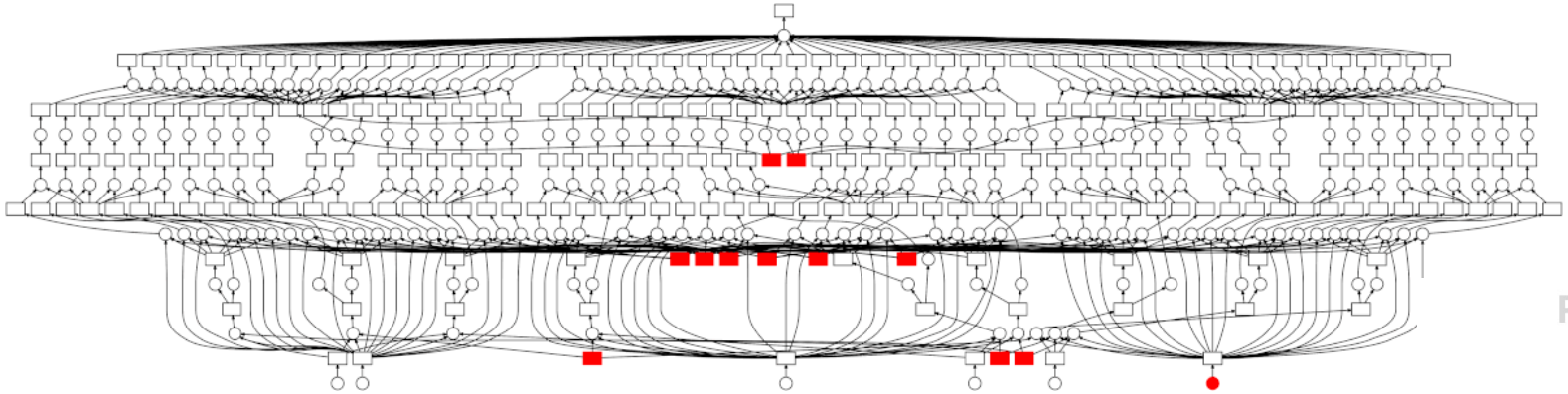
https://spark.apache.org/

Анализ данных – Apache Flink



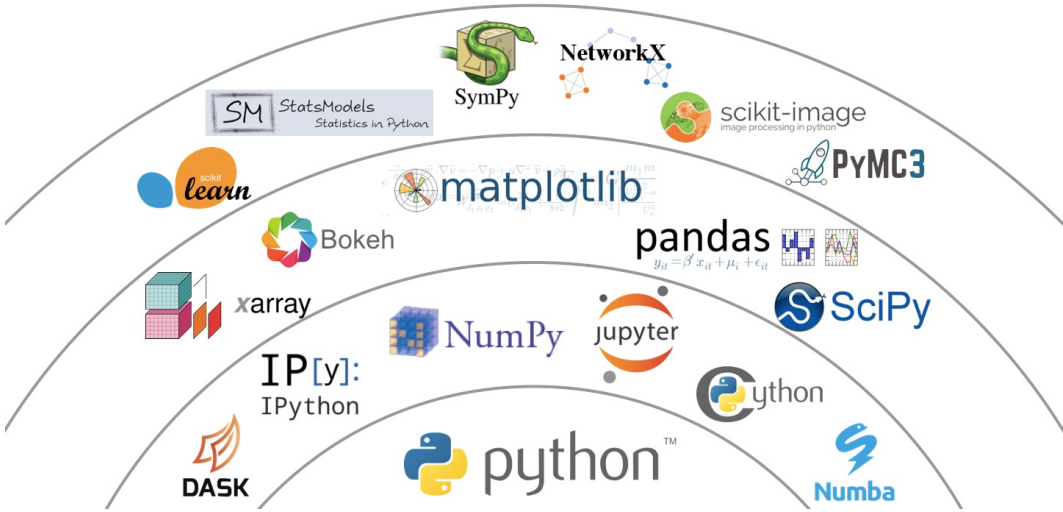
https://flink.apache.org/

Анализ данных - dask



Python's Scientific Stack

Jake Vanderplas PyCon 2017 Keynote

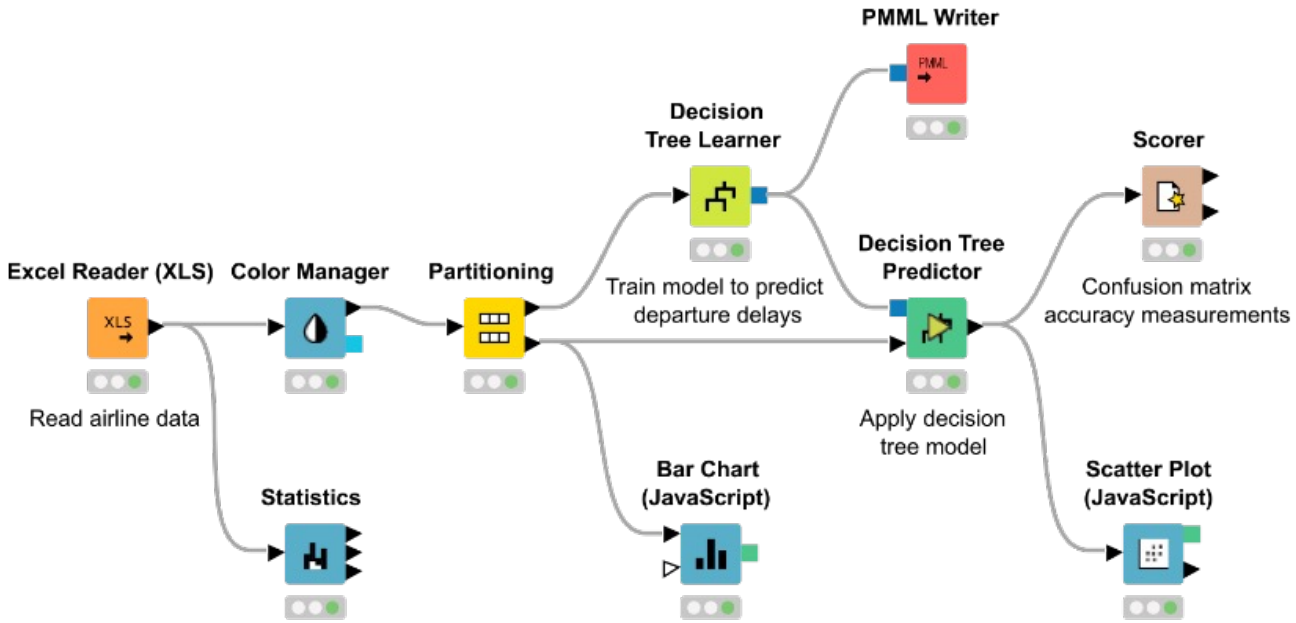


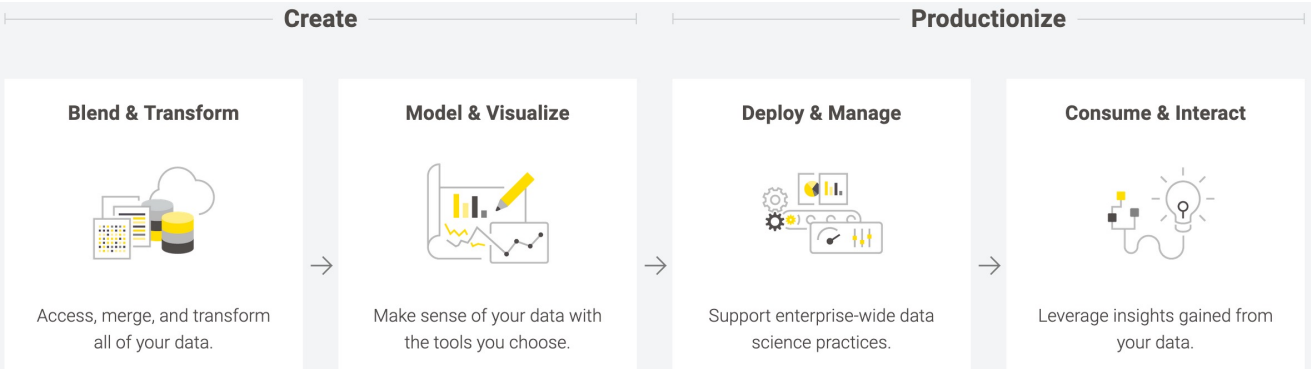
https://www.dask.org/

Масштабируемые параллельные вычисления для Python

Аналитика - KNIME

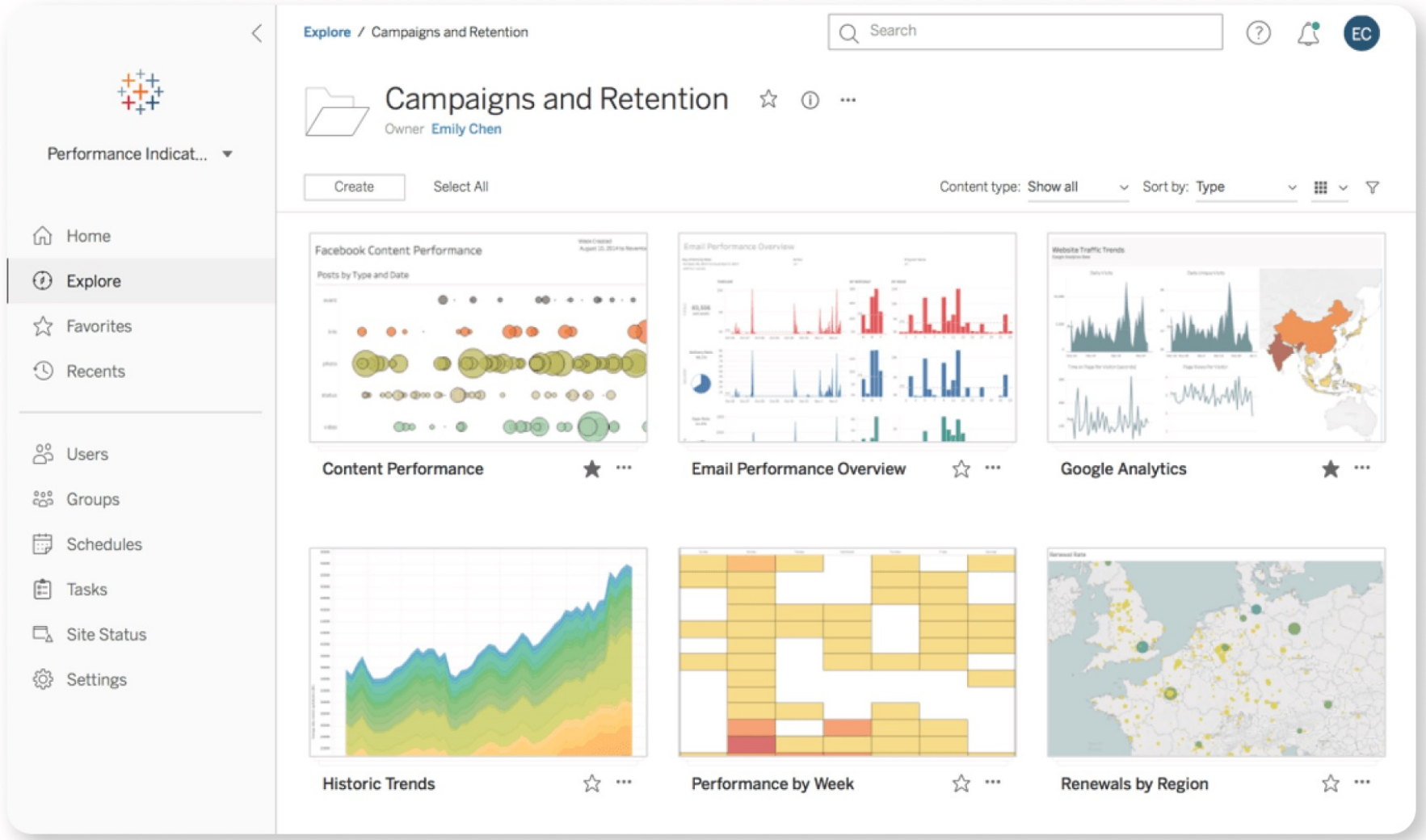
Open for Innovation
KNIME





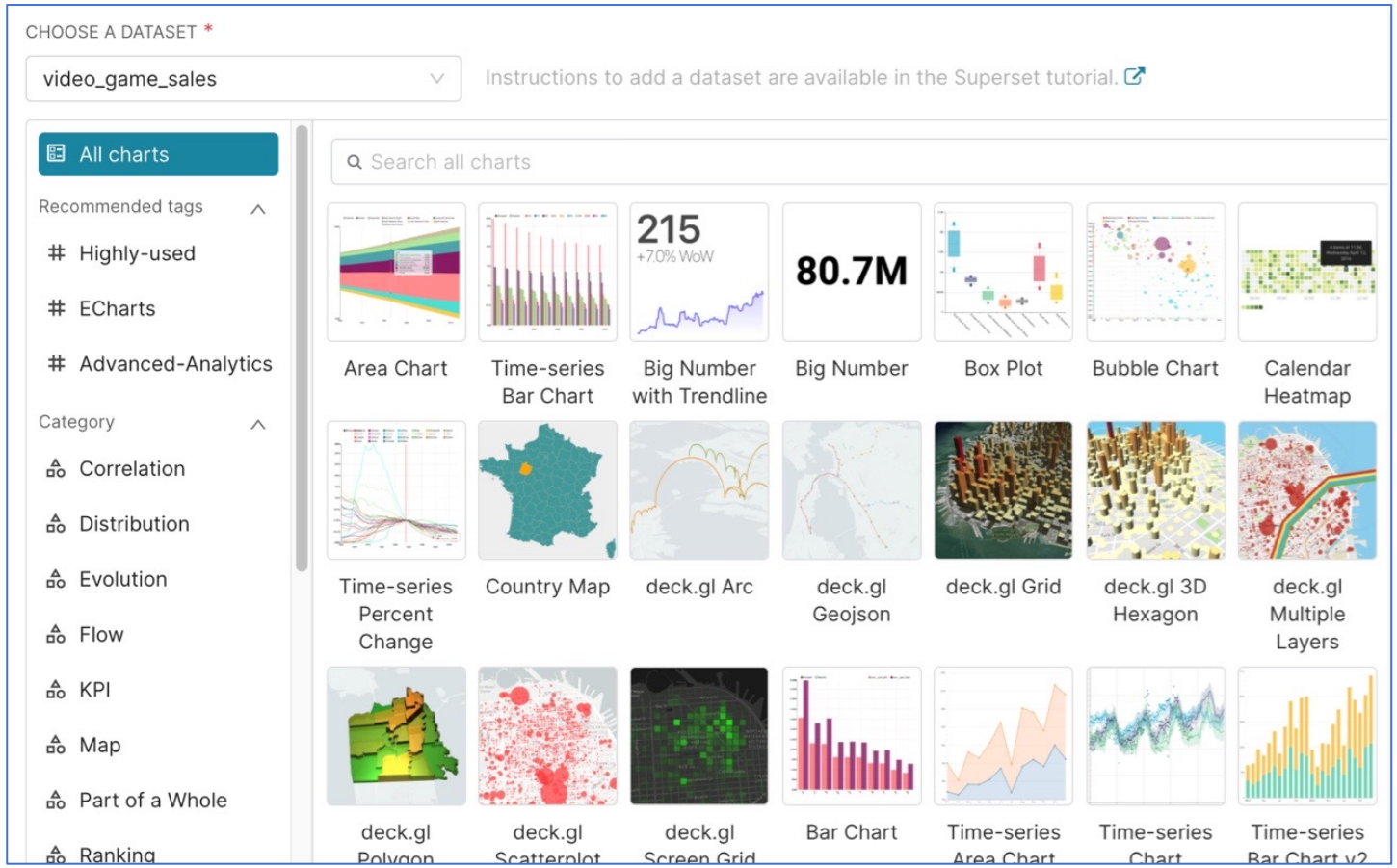
https://www.knime.com/

Анализ, визуализация - Tableau



https://www.tableau.com/en-gb

Визуализация, анализ – Apache Superset



Apache Superset™

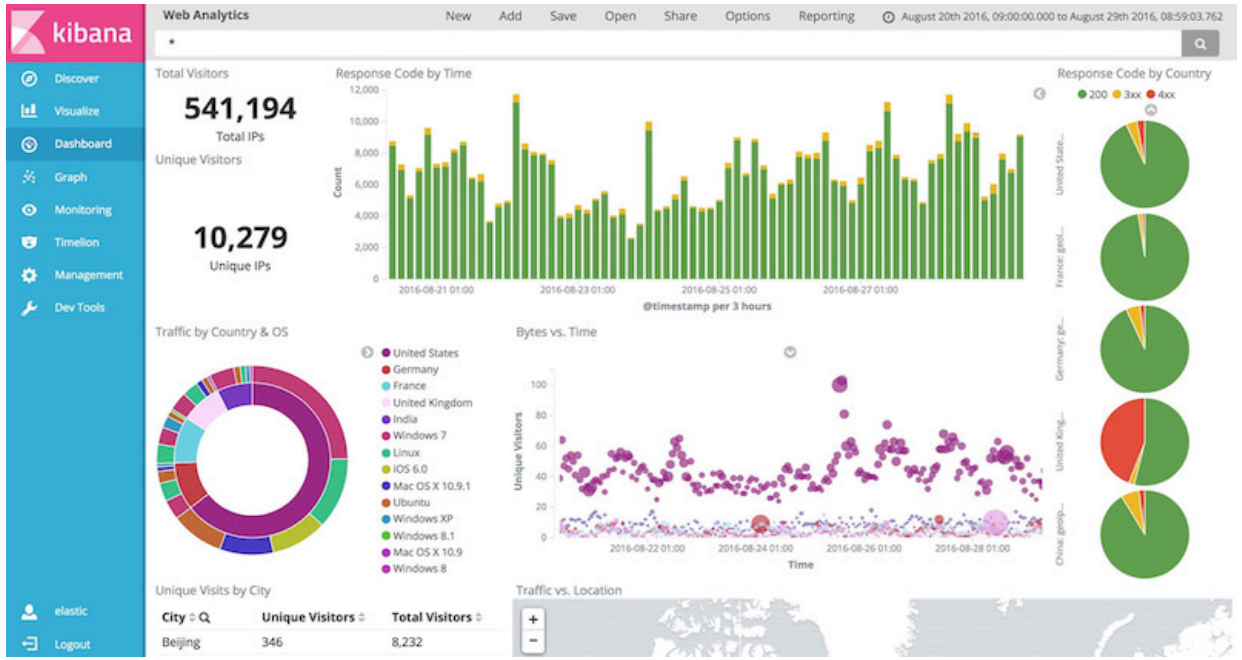
amazon REDSHIFT · druid · Google BigQuery · ClickHouse · dremio · databricks · Exasol · Firebird · Greenplum · IBM DB2 · MySQL · SQL Server · monetdb · ORACLE DATABASE · PostgreSQL · presto · snowflake · SQLite · trino · ROCKSET · VERTICA · teradata. · Timescale · yugabyteDB

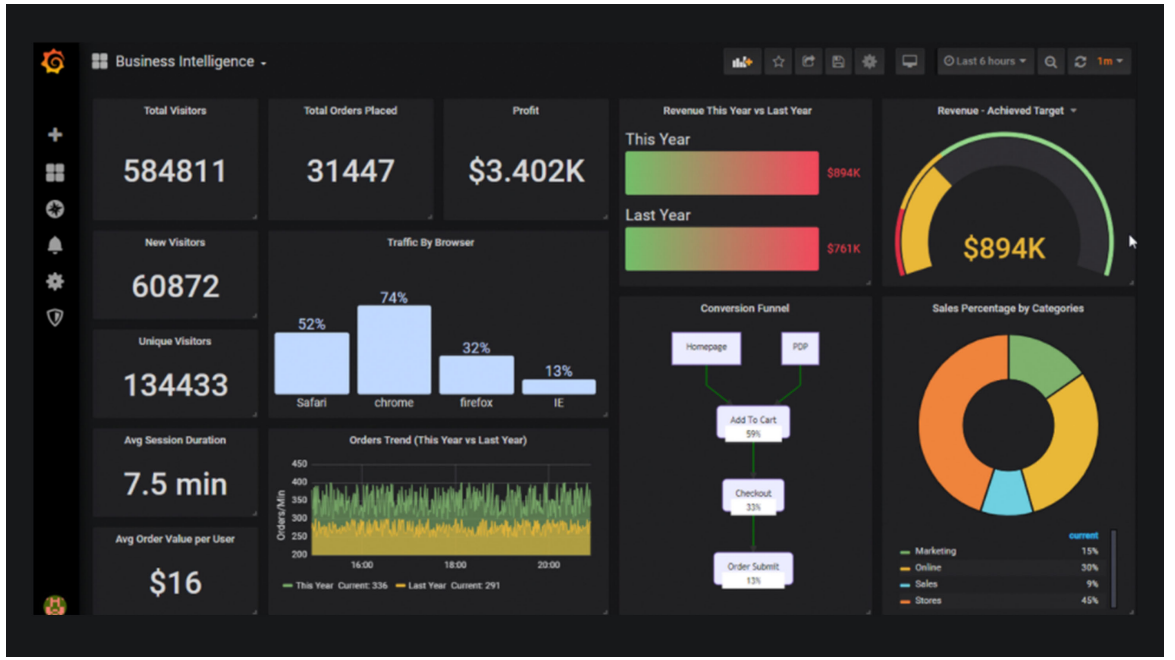
https://superset.apache.org/

Визуализация – Grafana, Kibana

kibana

Grafana





+ Исследование данных

https://www.elastic.co/kibana/

+ Временные ряды

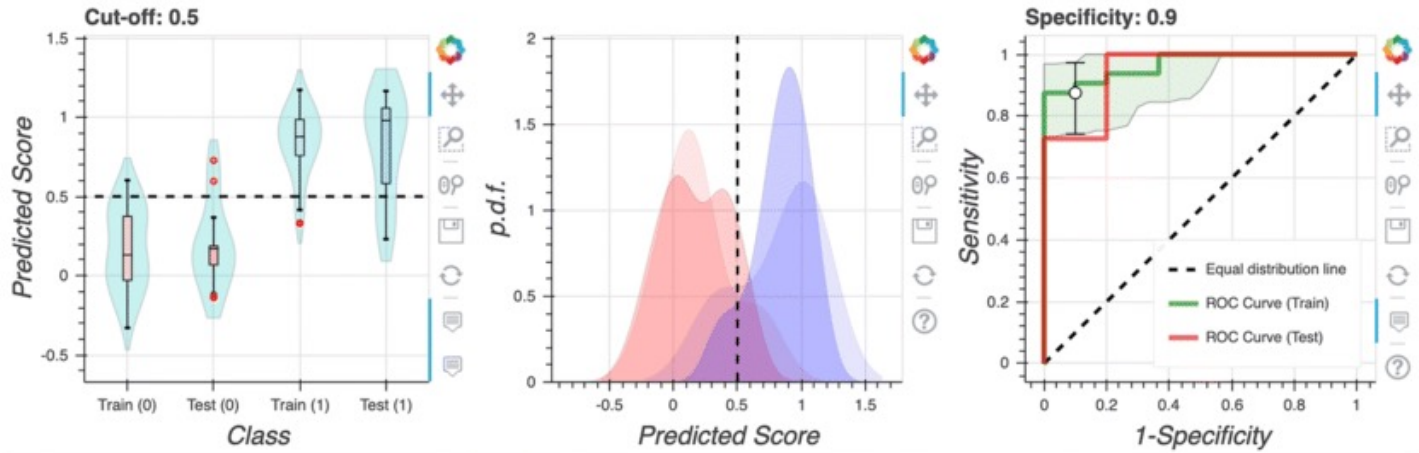
https://github.com/grafana/grafana

Исследование данных, визуализация (I)

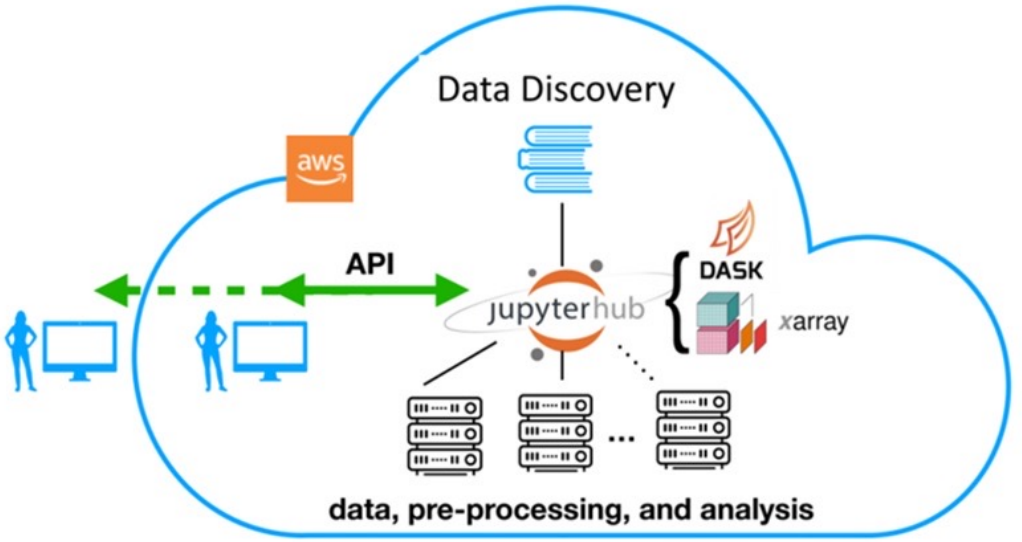
```
# Calculate Ypredicted score using modelPLS.test
YVpred = modelPLS.test(XVknn)

# Evaluate Ypred against Ytest
evals = [Ytest, YVpred]    # alternative formats: (Ytest, Ypred) or np.array([Ytest, Ypred])
#modelPLS.evaluate(evals, specificity=0.9)
modelPLS.evaluate(evals, cutoffscore=0.5)
```

BokehJS 1.1.0 successfully loaded.

Score cut-off fixed to: 0.5

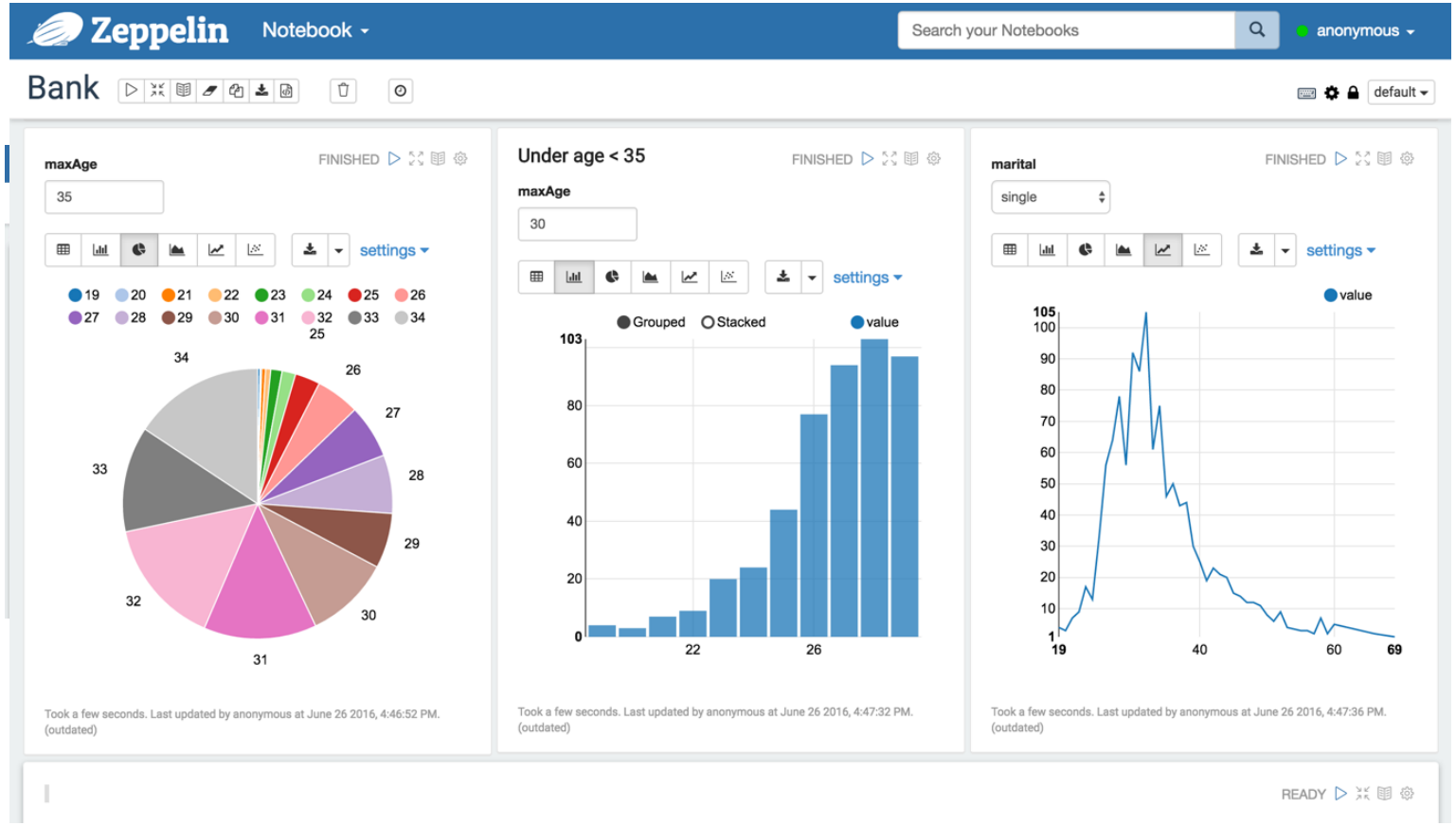


#	Evaluate	MW-U Pvalue	R2	AUC	Accuracy	Precision	Sensitivity	F1score
0	Train	2.66e-10	0.67 (0.5, 0.77)	0.97 (0.91, 1.0)	0.89 (0.81, 0.94)	0.9 (0.82, 0.95)	0.88 (0.74, 0.97)	0.89 (0.8, 0.95)
1	Test	6.37e-04	0.52	0.95	0.76	0.8	0.73	0.76

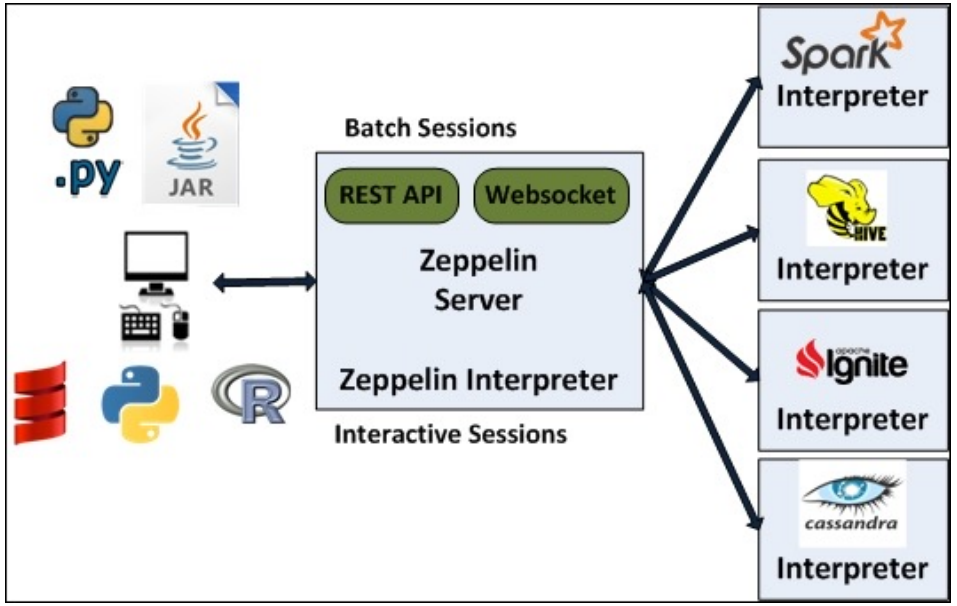


https://jupyter.org/hub

Исследование данных, визуализация (II)

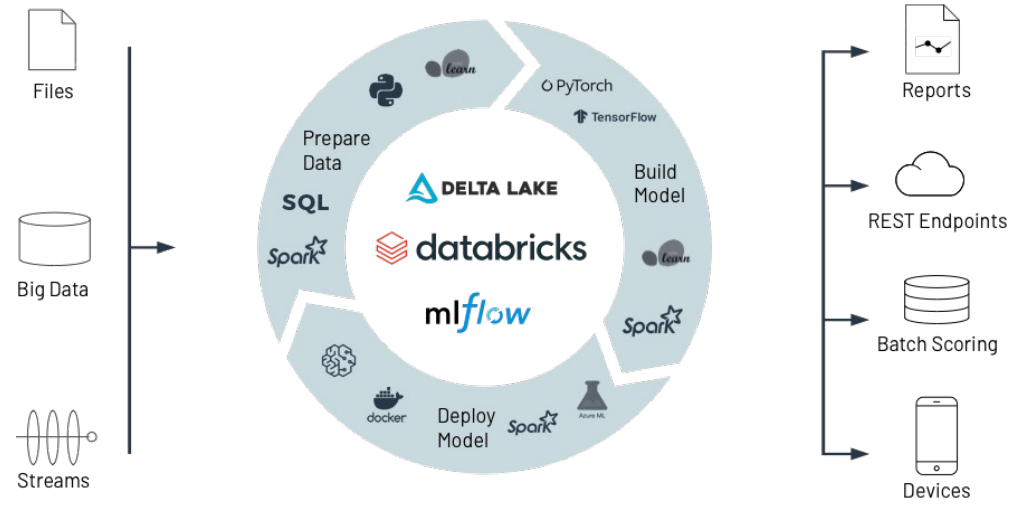


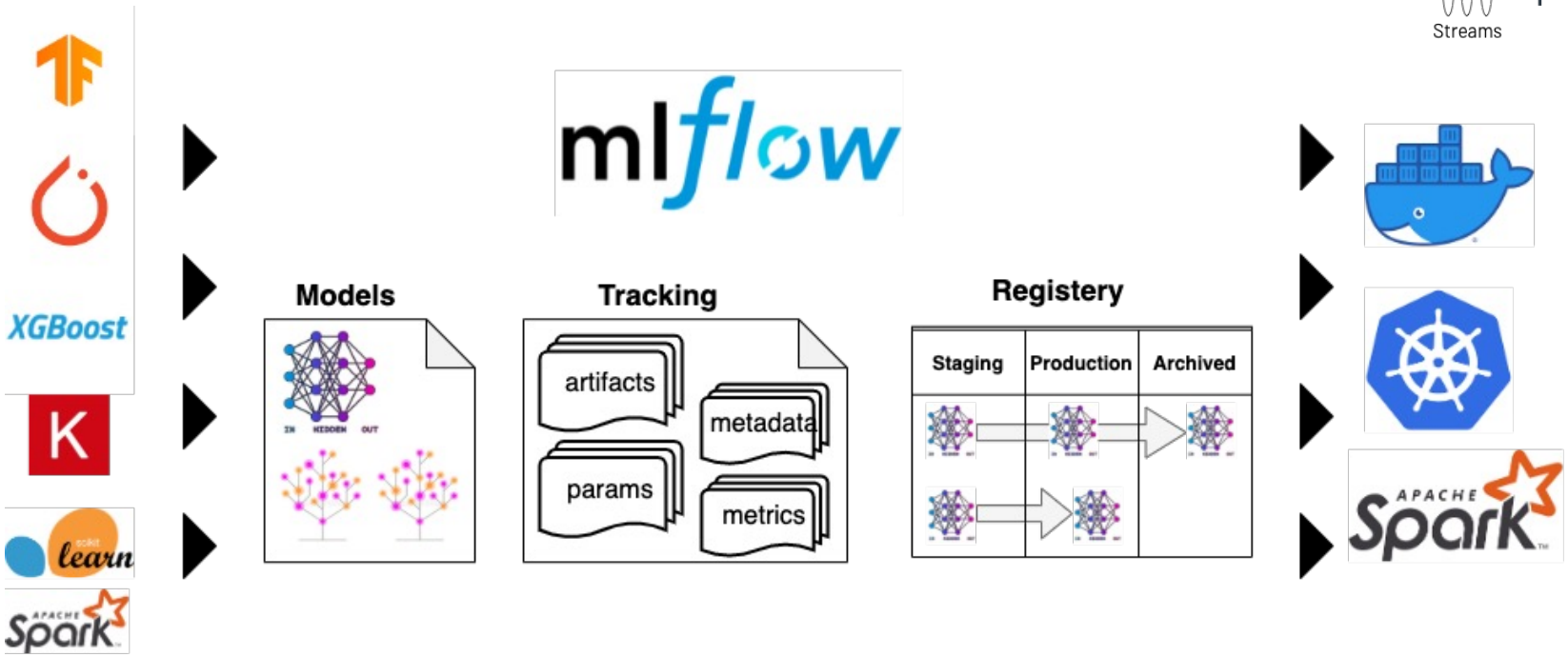
Apache Zeppelin



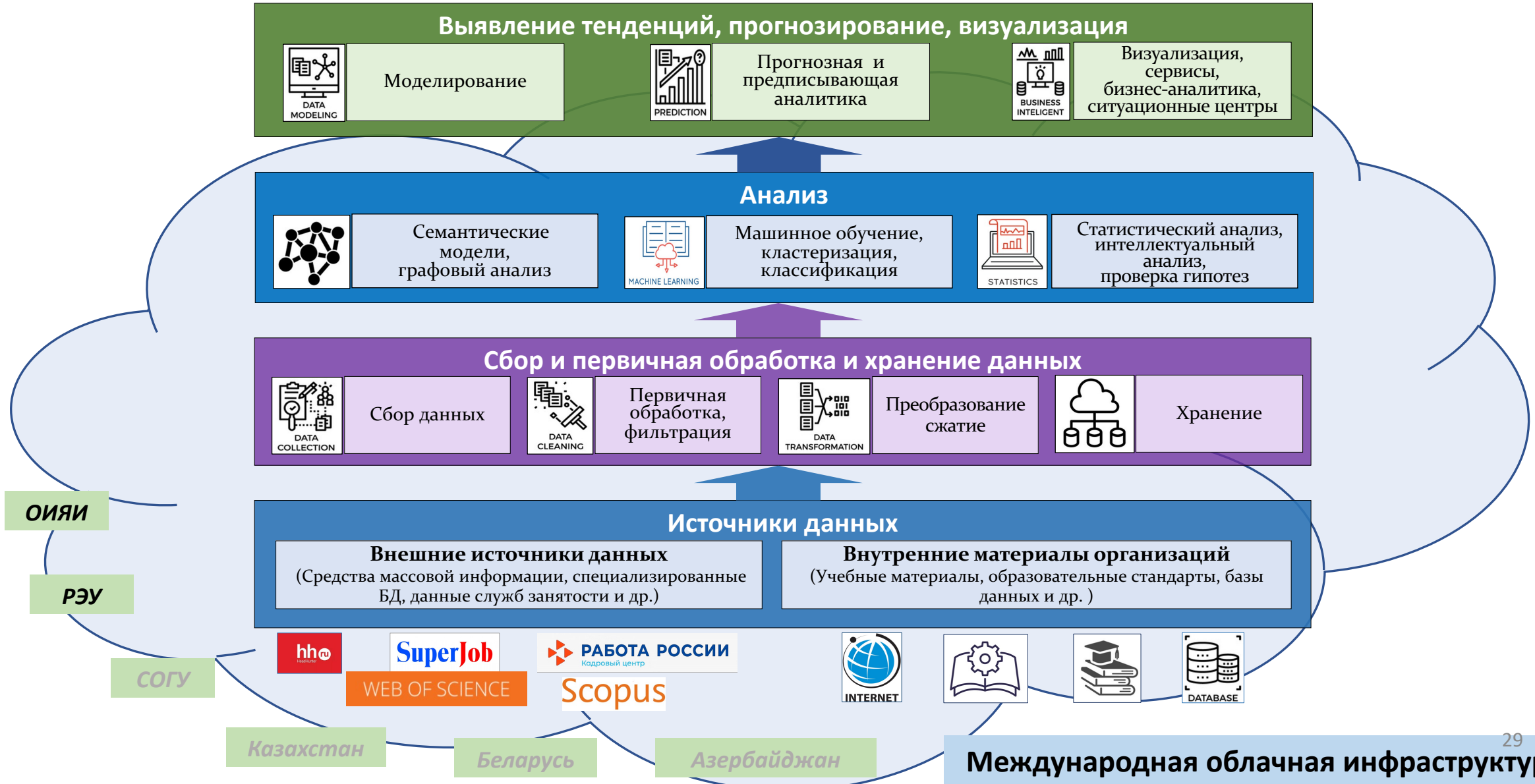
https://zeppelin.apache.org/

Вычислительные эксперименты и машинное обучение – MLflow

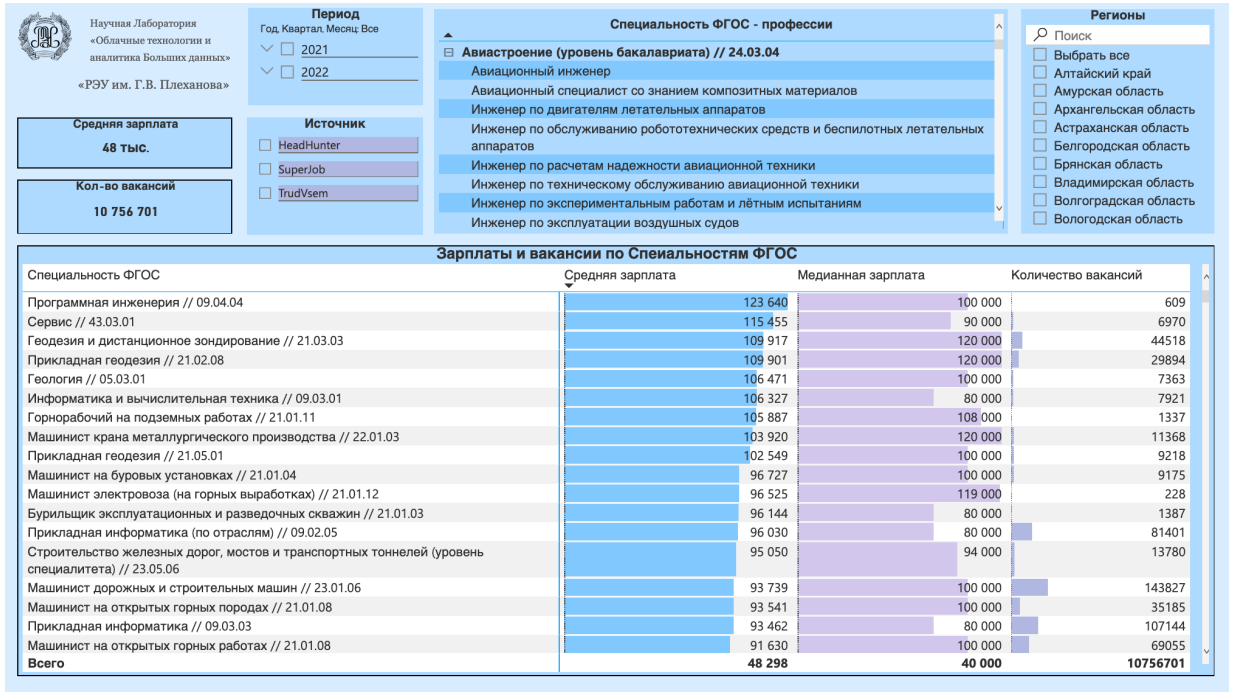


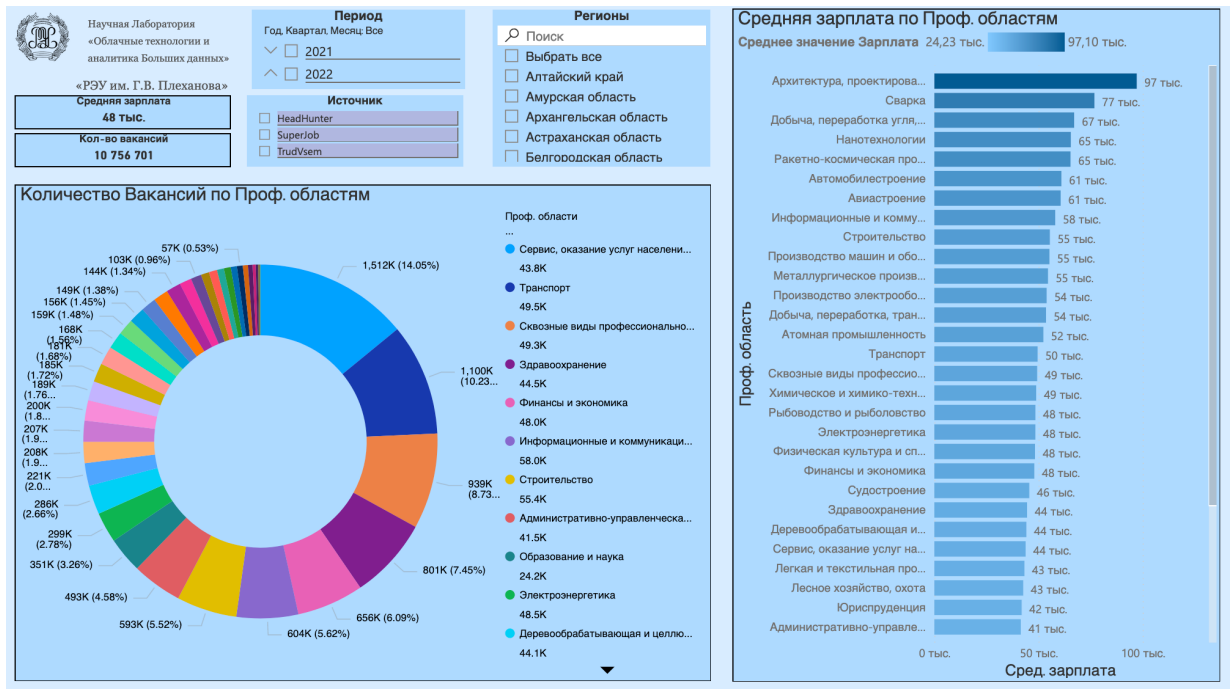


Пример реализации платформы Больших данных



Приложение: анализ рынка труда





Приложение: анализ рынка труда

Экосистема программных компонентов

Уровень	Программные пакеты
Визуализация и интерфейсы для доступа к системе	• Zeppelin, Jupyter (пользовательский интерфейс) • Graphana (создание отчетов и графическое представление результатов) • KrakenD (организация программных шлюзов для различных компонентов)
Распределенная аналитика Больших данных	• Apache Kylin
Вычислительные эксперименты в области машинного обучения	• MLflow
Вычисления в памяти	• Apache Spark, Dask, Hadoop
Организация процесса управлением потоками данных и сбором данных	• Apache Kafka, Apache Flume, • Apache Airflow, Celery, Scrapy
Хранилища и специализированные базы данных	• CEPH, NFS (хранение и доступ к файлам) • Elasticsearch (хранение и анализ структурированных данных) • Apache Ignite (база с хранением данных в памяти для быстрого доступа и кеширования) • Российское «озеро данных» • Apache Calcite (интеграция хранилищ)
Аутентификация и сквозная авторизация, безопасность	• Free IPA, Vault
Компьютерная инфраструктура, управление ресурсами	• OpenNebula, Kubernetes, • Docker, Puppet, Git

Спасибо за внимание!