академия больших данных

mail.ru group

Инструменты визуализации при работе с Большими Данными

Андрей Кузнецов

09.10.2021

Структура курса

1. Введение в Большие Данные
2. Hadoop экосистема и MapReduce
3. SQL поверх больших данных
4. Инструменты визуализации при работе с Большими Данными
5. Введение в Scala
6. Модель вычислений Spark: RDD
7. Распараллеливание алгоритмов ML
8. Spark Pipelines
9. Approximate алгоритмы для больших данных
10. Spark для оптимизации гиперпараметров
11. Потоковая обработка данных (Kafka, Spark Streaming, Flink)
12. Архитектуры в продакшене

План занятия

1. Apache Zeppelin
2. Polynote
3. Big Data Tools
4. Cloud Solutions
5. Workshop:
    a. cloud-based work
    b. zeppelin
    c. polynote
    d. homework discussion

Where we are?

Big Data platform

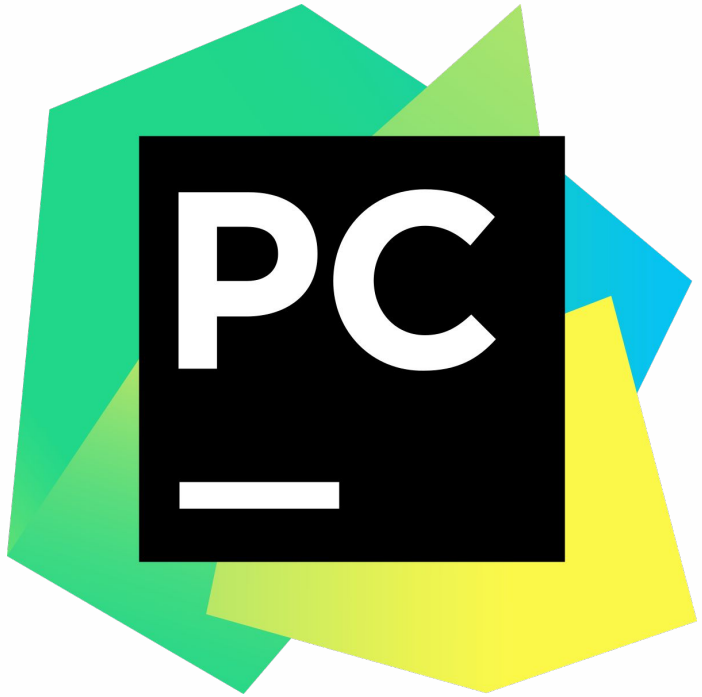
- Hadoop Distributed Filesystem (NM, DN)
- Apache Hadoop YARN (RM, NM)

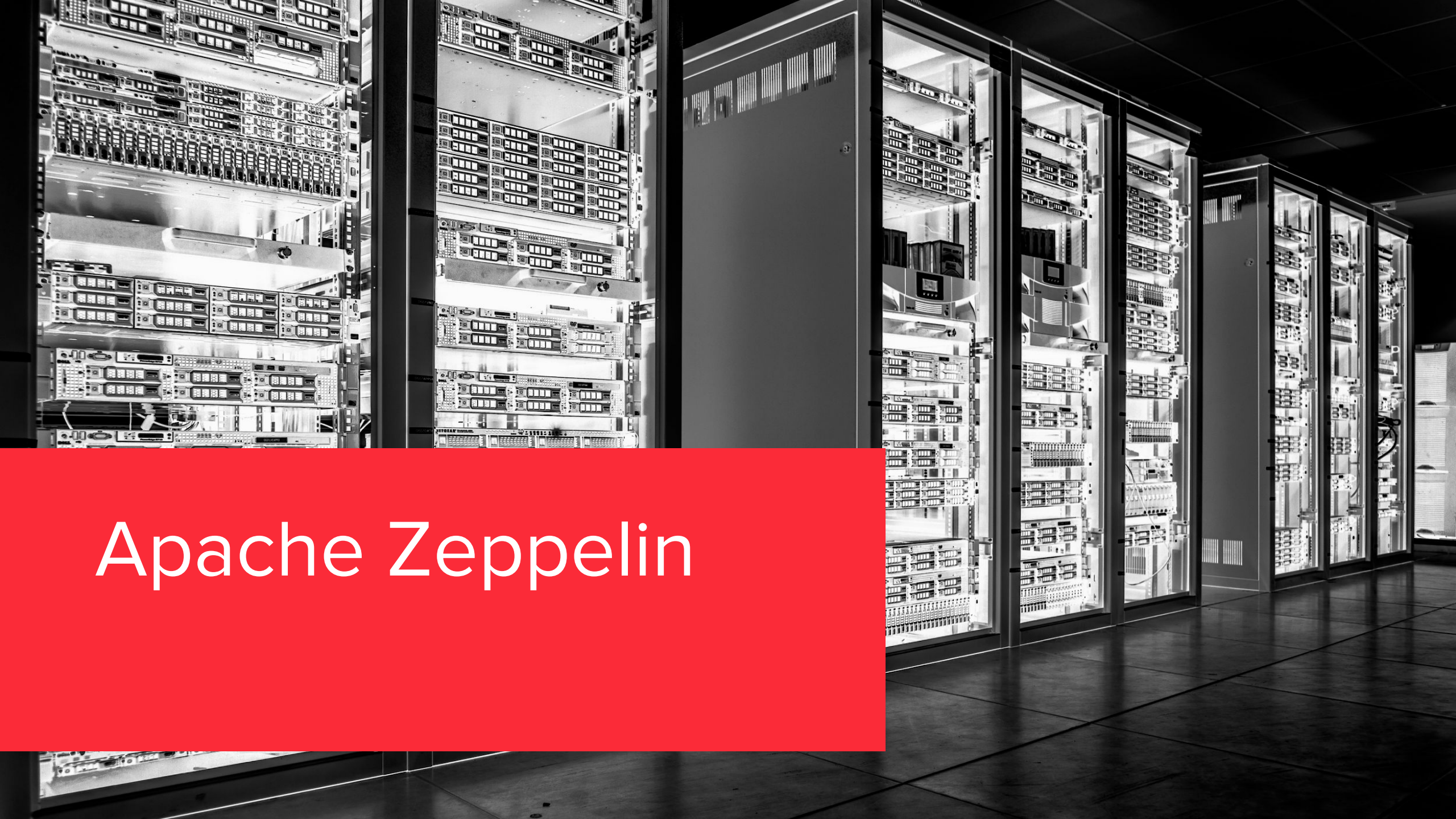
Big Data applications

- Hadoop MapReduce
- SQL-like processing frameworks
- Apache Spark
- Stream processing frameworks + Apache Kafka

Tools to work with

Classic small data stack





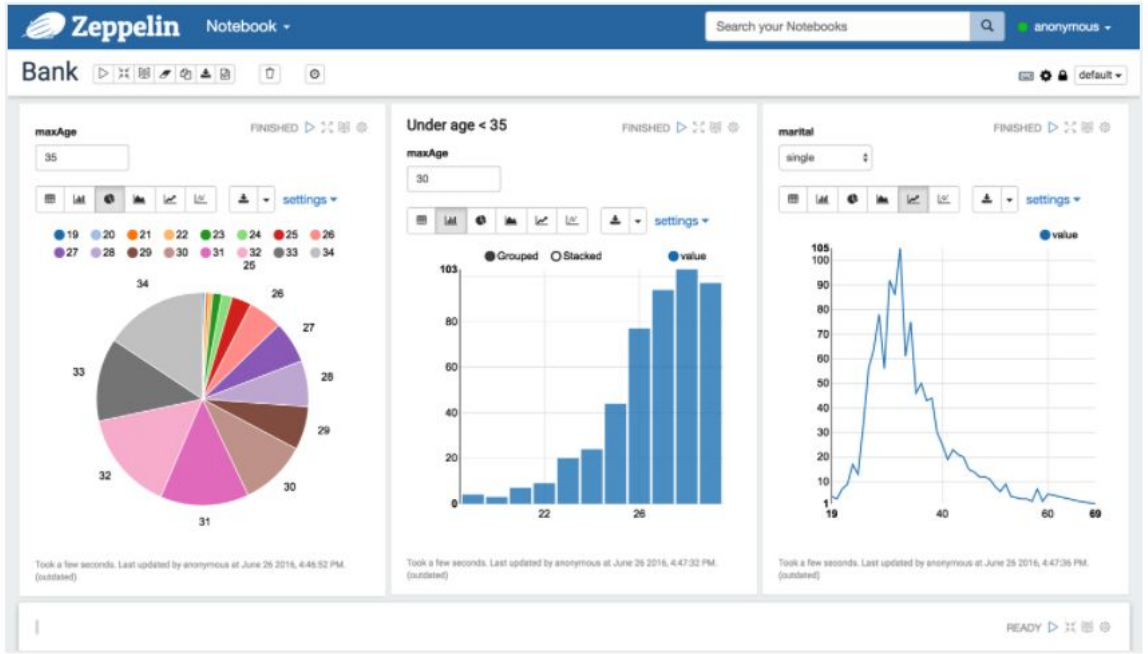
Apache Zeppelin

Apache Zeppelin in a nutshell

Multi-purpose Notebook

The Notebook is the place for all your needs

- Data Ingestion
- Data Discovery
- Data Analytics
- Data Visualization & Collaboration



Apache Zeppelin in a nutshell

- Хочет быть как jupyter notebook / jupyterlab
- умеет нативно работать со Scala/Java
- Hive/Spark удобно конфигурировать и работать
- Неплохие визуализации

Zep context. One interface to rule them all

```
%spark                                                    FINISHED

// This is just for demo, it could be done via running
    a spark job
z.put("maxAge", 83)
```

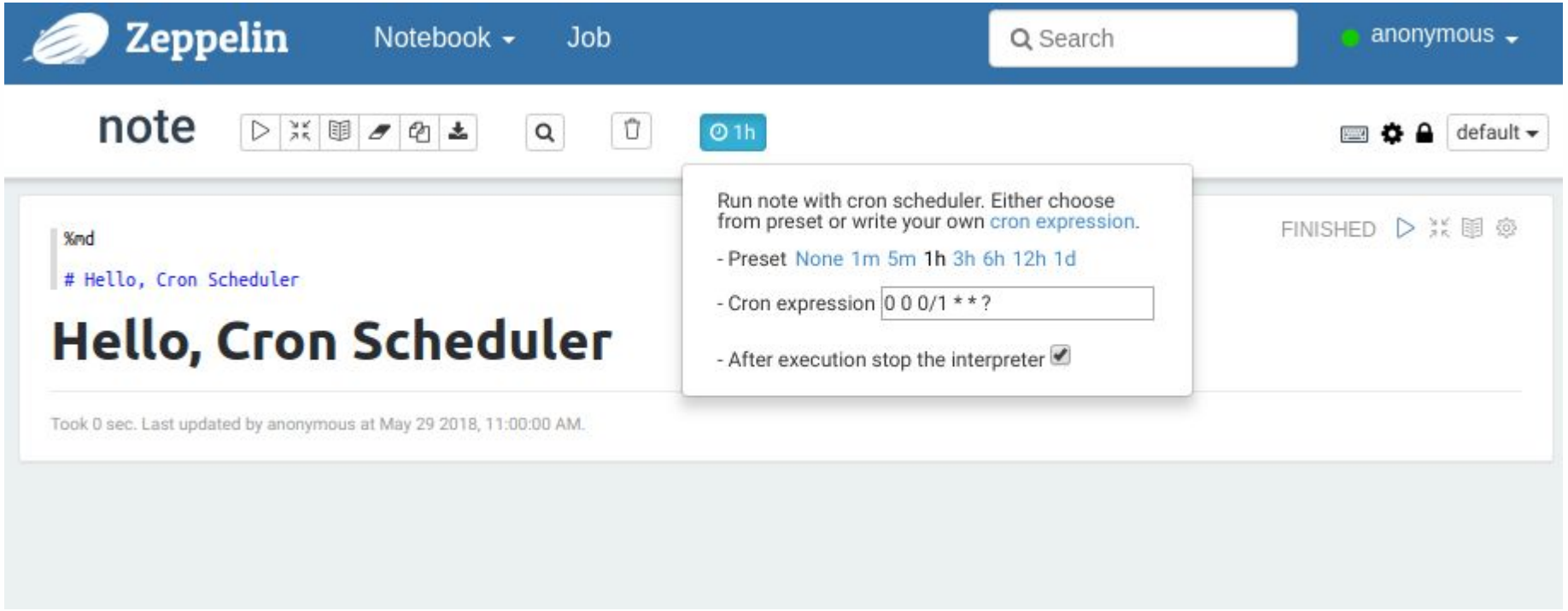
```
%jdbc(interpolate=true)                                   FINISHED

select * from bank where age = {maxAge}
```

settings

age	job	marital
83	retired	married
83	retired	married
83	retired	divorced
83	retired	divorced

Scheduling. Don’t try on prod!



Notebook storages

1. (default) use local file system and version it using local Git repository - GitNotebookRepo
2. all notes are saved in the notebook folder in your local File System - VFSNotebookRepo
3. all notes are saved in the notebook folder in hadoop compatible file system - FileSystemNotebookRepo
4. storage using Amazon S3 service - S3NotebookRepo
5. storage using Azure service - AzureNotebookRepo
6. storage using Google Cloud Storage - GCSNotebookRepo
7. storage using Aliyun OSS - OSSNotebookRepo
8. storage using MongoDB - MongoNotebookRepo
9. storage using GitHub - GitHubNotebookRepo

Interpreters. JDBC

- Postgresql - JDBC Driver
- Mysql - JDBC Driver
- MariaDB - JDBC Driver
- Redshift - JDBC Driver
- Apache Hive - JDBC Driver
- Presto/Trino - JDBC Driver
- Impala - JDBC Driver
- Apache Phoenix itself is a JDBC driver
- Apache Drill - JDBC Driver
- Apache Tajo - JDBC Driver

Polynote

Polynote in a nutshell

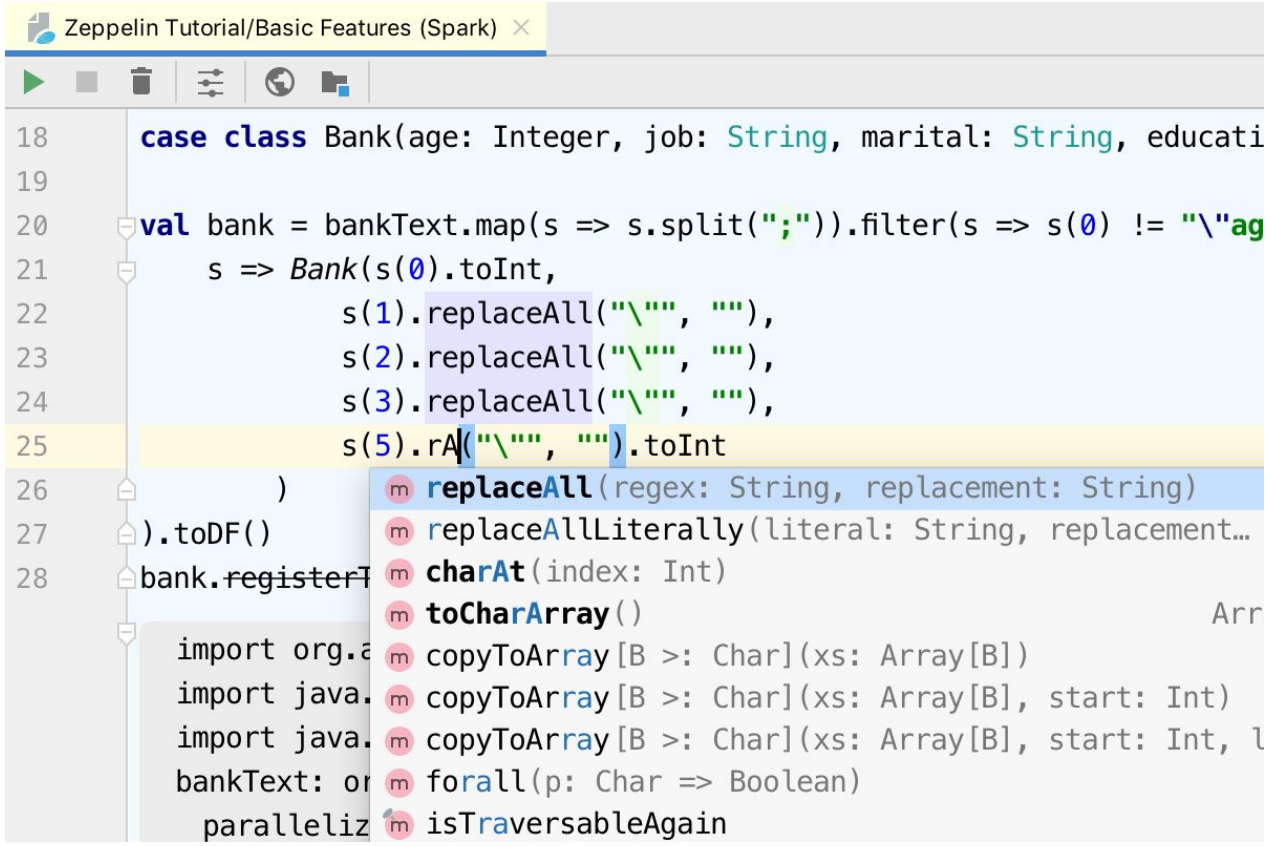
- Умеет варить объекты в одном кернели и конвертить из Scala в Python
- Сырой, но хоть какая-то альтернатива Zeppelin / Jupyter

https://polynote.org/

Big Data Tools

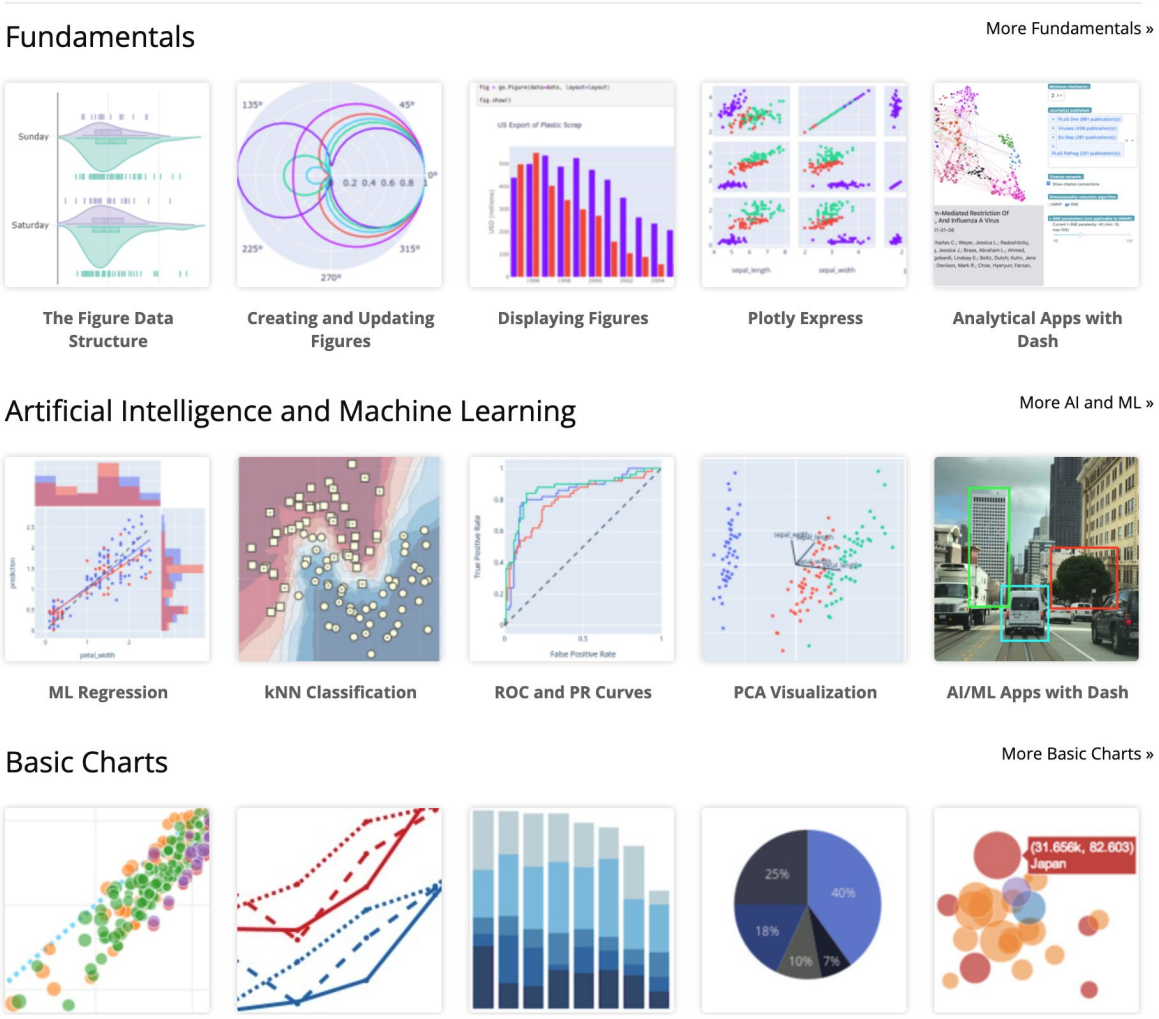
Big Data Tools in a nutshell

- Интеллектуальная поддержка Zeppelin notebooks
- Интеграция с инструментами Spark и Hadoop
- Распределенные файловые системы и столбцовые форматы
- Инструменты для работы с таблицами и диаграммами



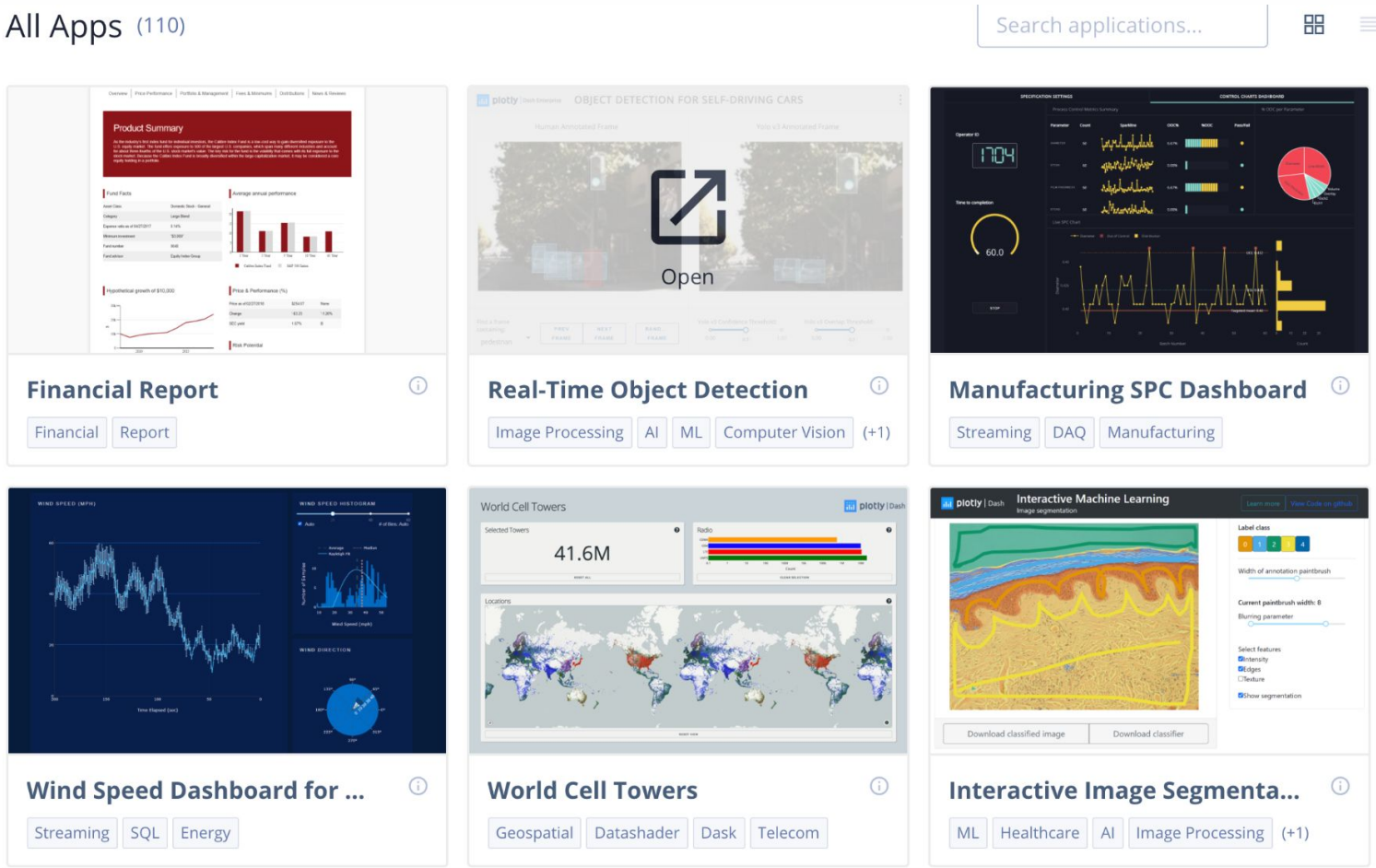
More vis tools

Plotly. Multilang vis tool



https://plotly.com/python/

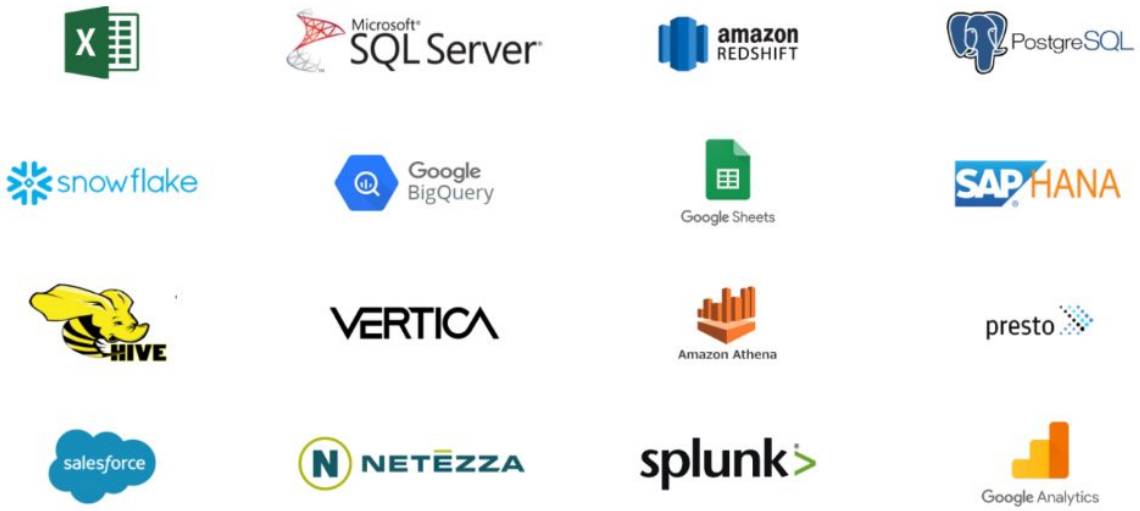
Dash. Visual apps at web



https://dash.gallery/Portal/

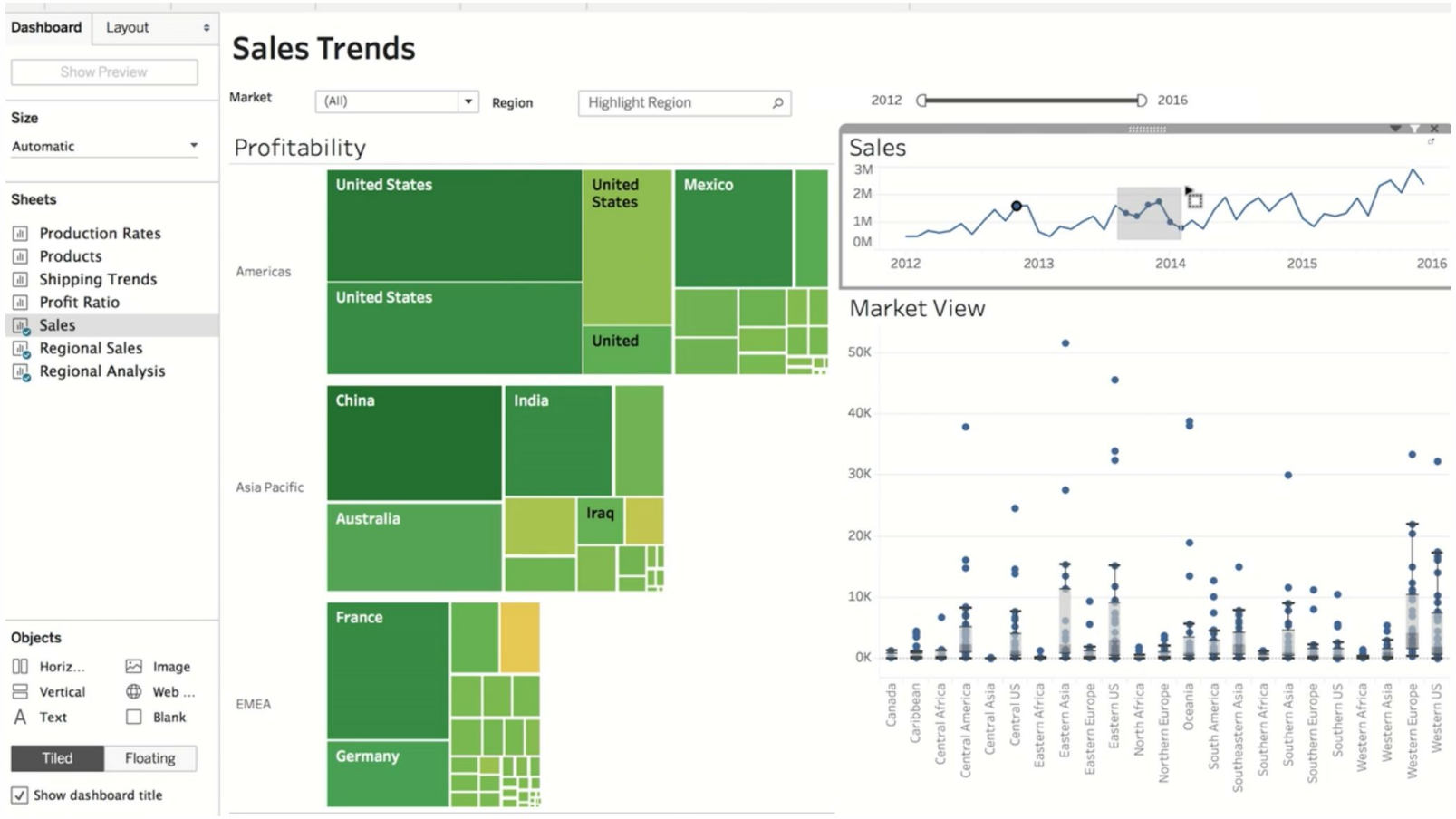
Tableau. Expensive industrial standart

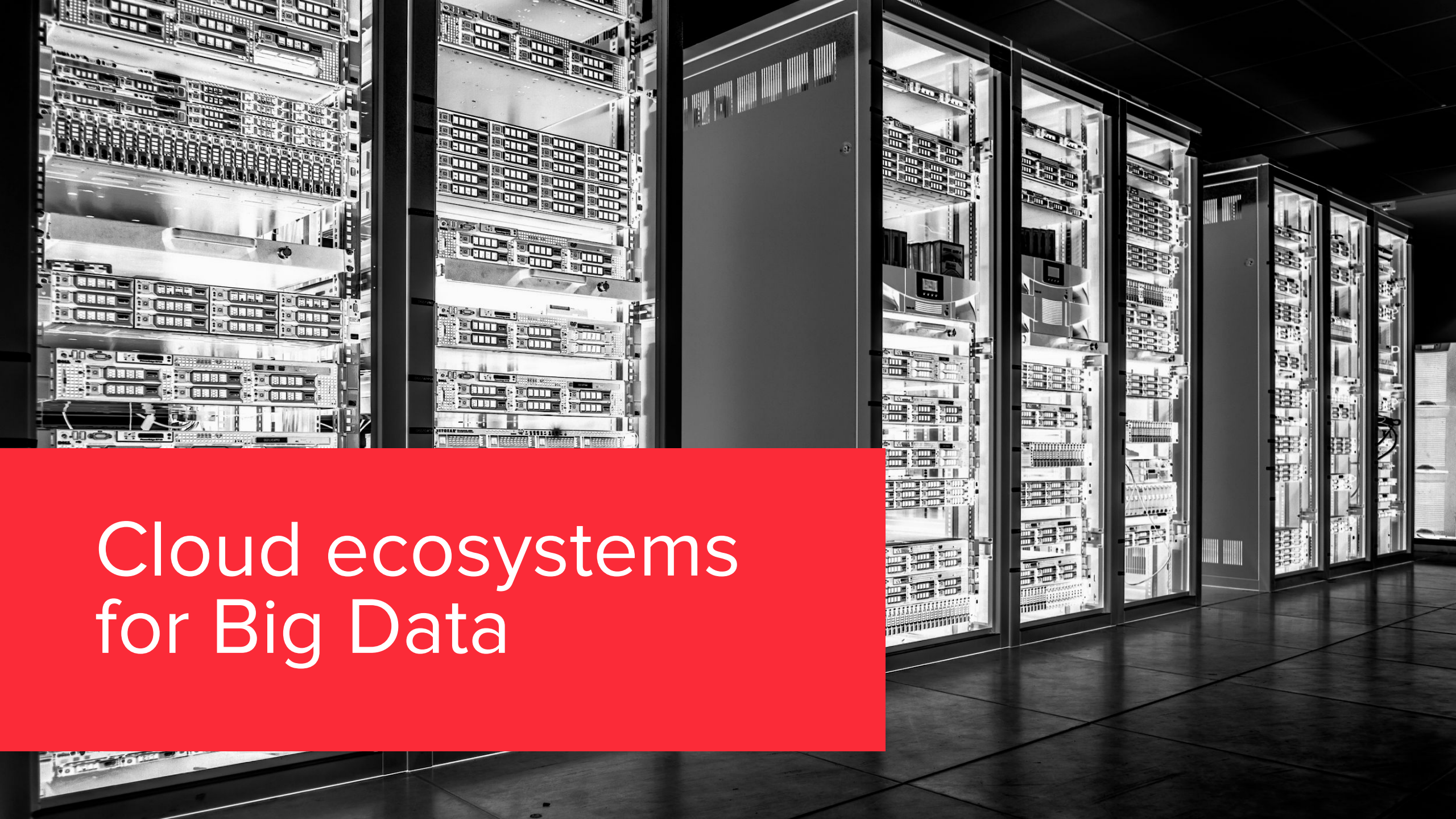
Tableau native data connectors



Connect to all of your data—no matter where it resides. Tableau offers native connectors built and optimized for many databases and files—from spreadsheets and PDFs to big data, cube, and relational databases on-premises or in the cloud, even application data or data on the web.

Tableau. Interface

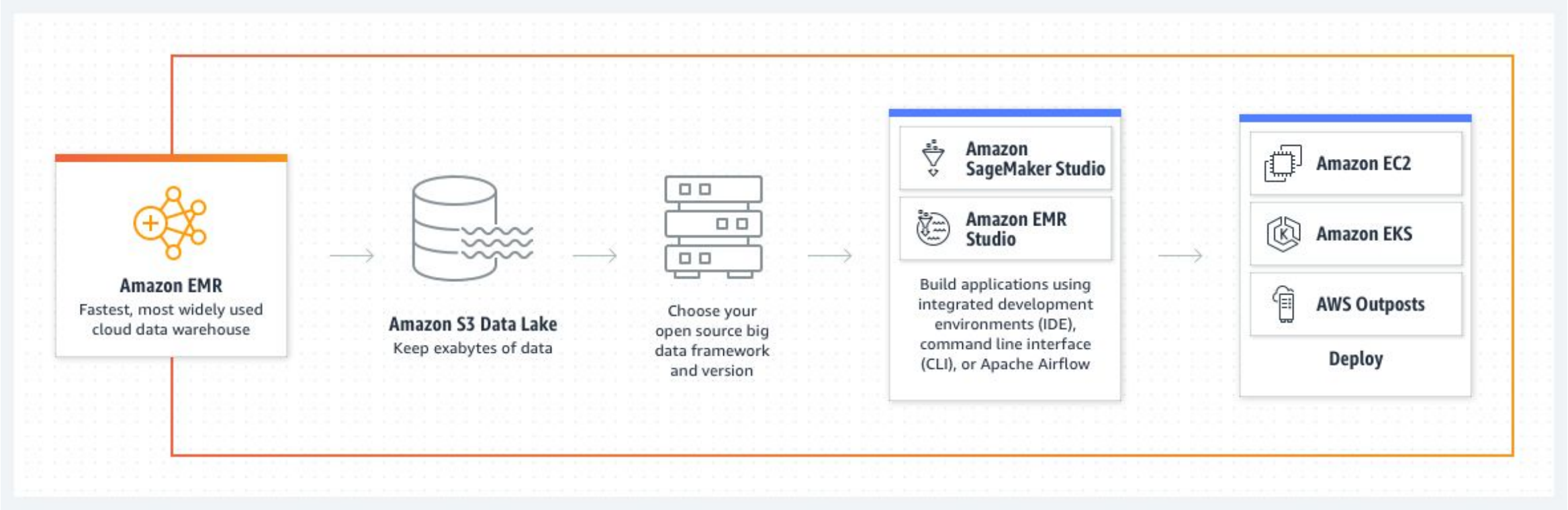




Cloud ecosystems for Big Data

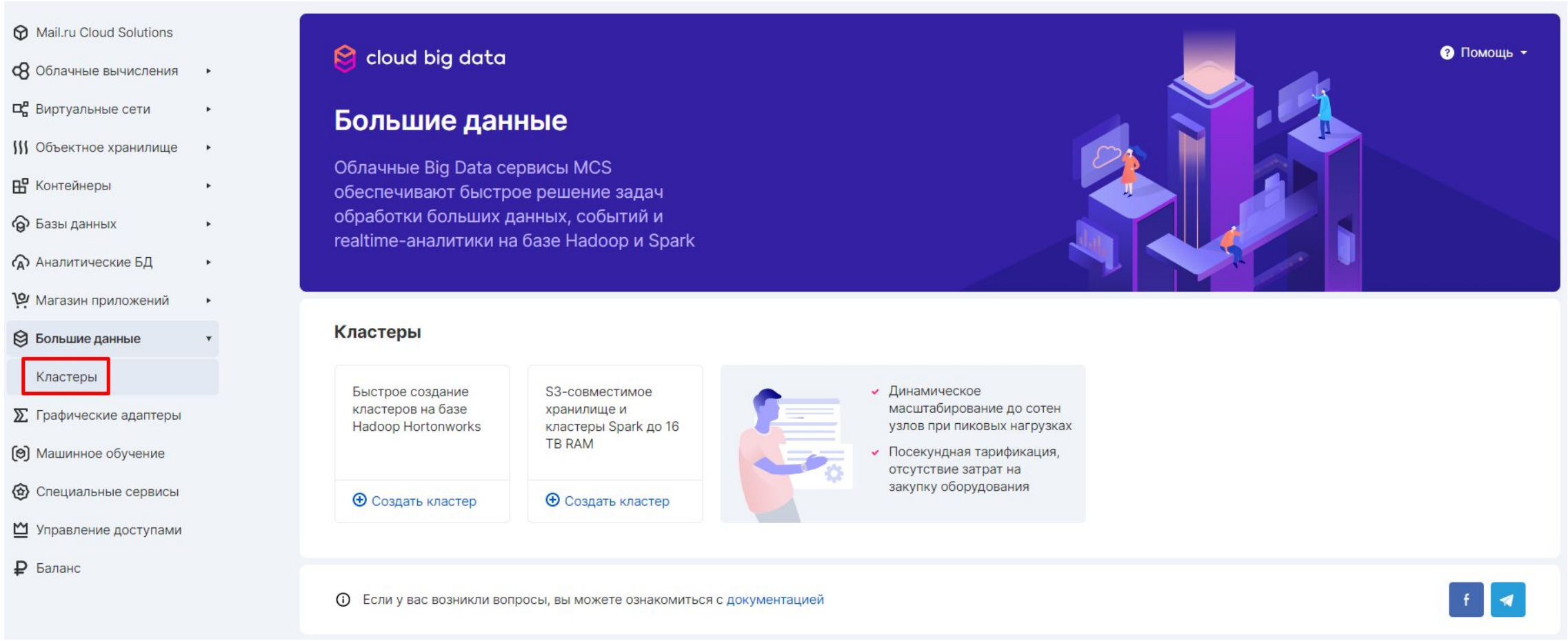
Amazon EMR

https://aws.amazon.com/ru/emr/features/?nc=sn&loc=2&dn=1



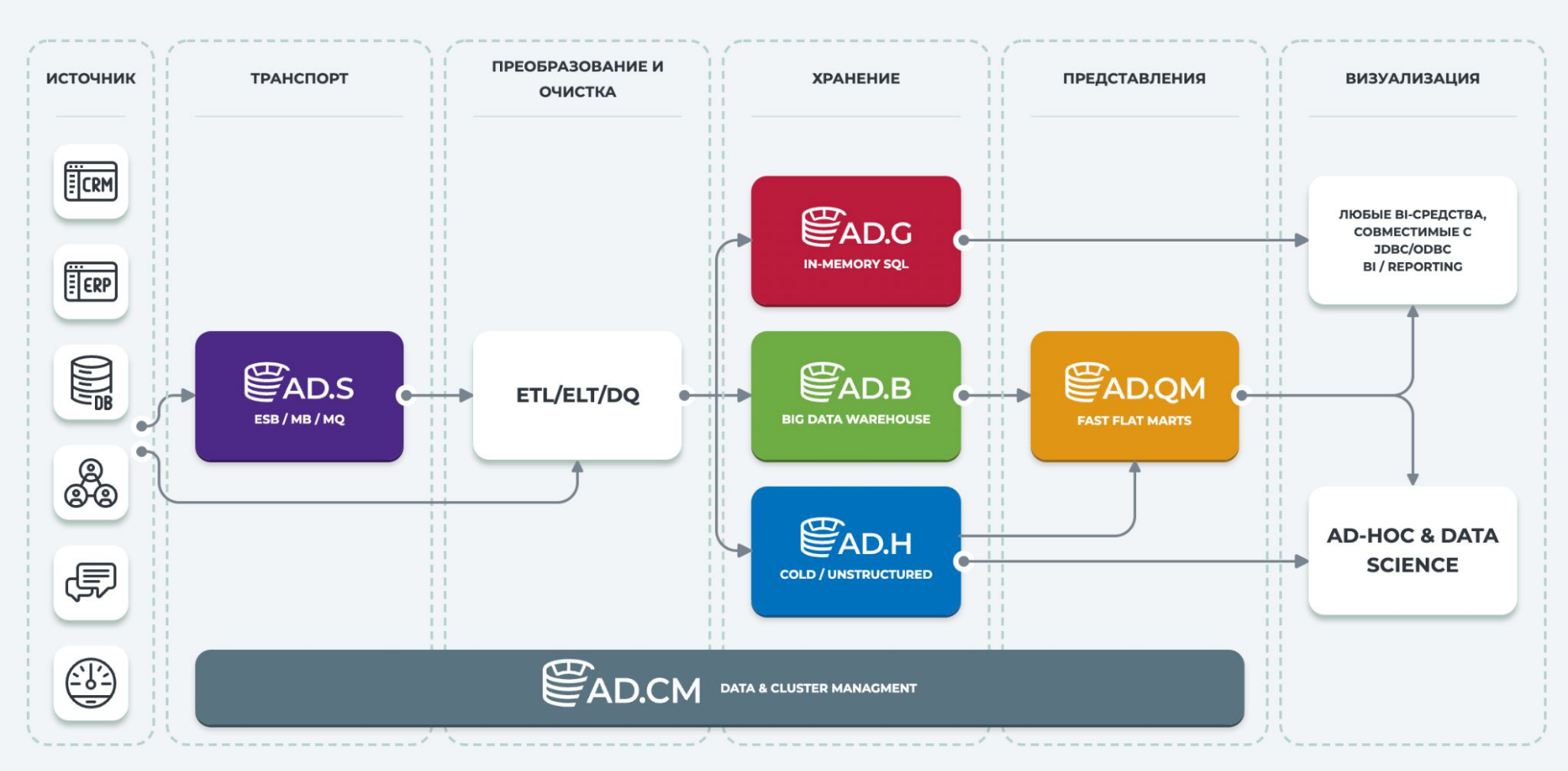
Amazon EMR – это платформа для быстрой обработки, анализа и работы с большими данными с помощью машинного обучения (ML), использующая платформы с открытым исходным кодом.

Mail.ru Cloud Solutions



ArenaData

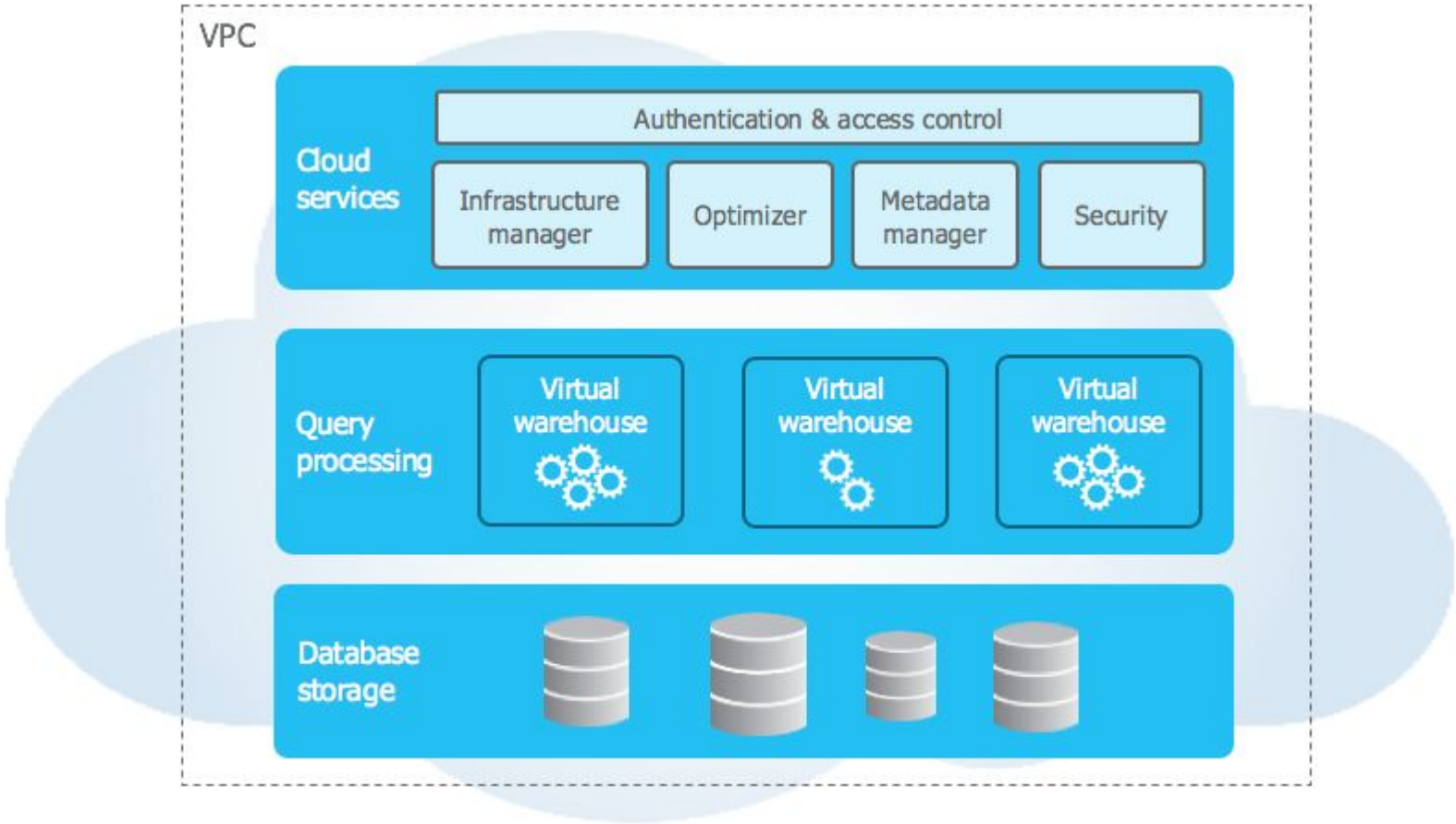
https://arenadata.tech/



Snowflake


WHERE YOUR DATA CLOUD EXPERIENCE BEGINS:
ONE PLATFORM, MANY WORKLOADS, NO DATA SILOS
DATA ENGINEERING
DATA LAKE
DATA WAREHOUSE
DATA SCIENCE
DATA APPLICATIONS
DATA SHARING
DATA SOURCES
DATA CONSUMERS
snowflake®

Snowflake


VPC
Cloud services
Authentication & access control
Infrastructure manager
Optimizer
Metadata manager
Security
Query processing
Virtual warehouse
Virtual warehouse
Virtual warehouse
Database storage

Recommended links and literature

1) **https://polynote.org/**
2) **https://zeppelin.apache.org/**
3) **https://docs.aws.amazon.com/emr/index.html**