


Инструменты визуализации при работе с Большими Данными

Андрей Кузнецов
22.10.2022



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. Approximate алгоритмы для больших данных
8. Поточковая обработка данных (Kafka, Spark Streaming, Flink)
9. Гостевая лекция VK
10. Гостевая лекция VK

План занятия

1. Apache Zeppelin
2. Polynote
3. Big Data Tools
4. Cloud Solutions
5. Workshop

Where we are?

Big Data platform

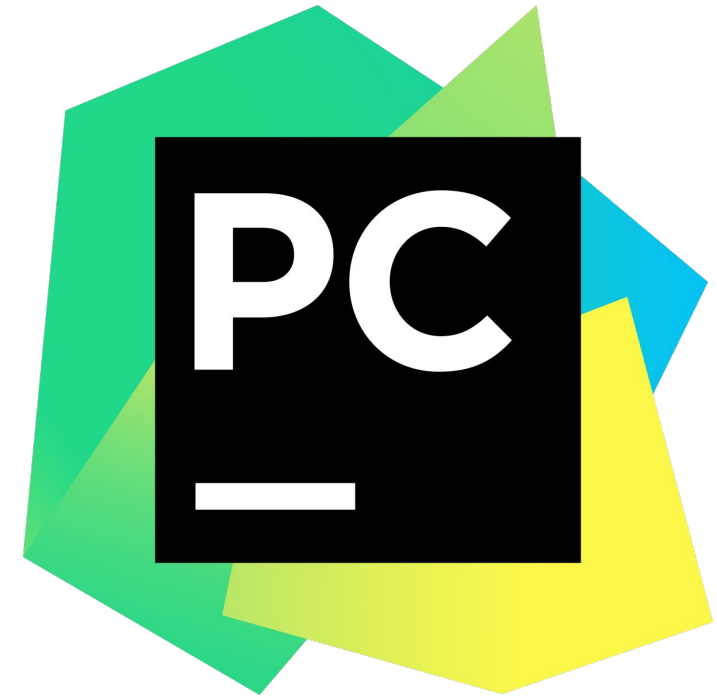
- Hadoop Distributed Filesystem (NM, DN)
- Apache Hadoop YARN (RM, NM)

Big Data applications

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

➞ **Tools to work with**

Classic small data stack









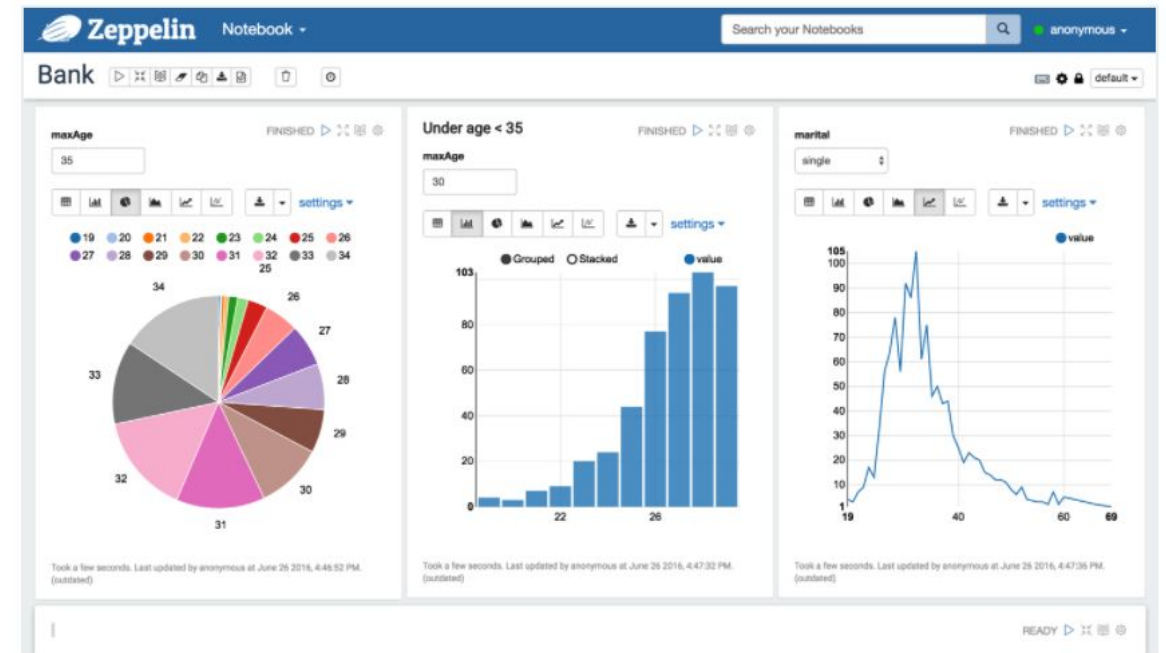
Apache Zeppelin

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)

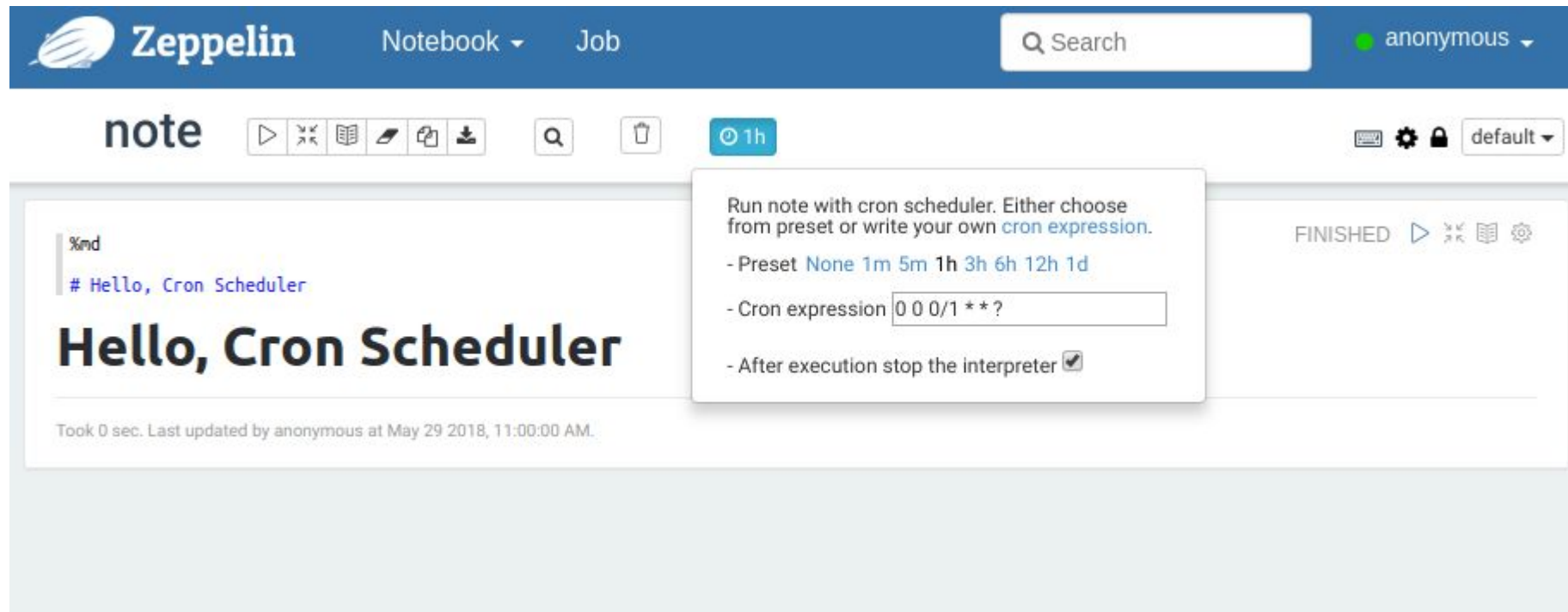
FINISHED    

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



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

Scheduling. Don't try on prod!



The image shows the Zeppelin Notebook interface. At the top is a blue header with the Zeppelin logo, 'Notebook' and 'Job' tabs, a search bar, and a user dropdown set to 'anonymous'. Below the header, the word 'note' is displayed next to a toolbar with icons for play, expand, view, edit, copy, and download. A '1h' refresh button is also present. On the right, there are icons for keyboard shortcuts, settings, a lock, and a 'default' dropdown. The main workspace contains a code editor with the following content:

```
%md
# Hello, Cron Scheduler
```

Hello, Cron Scheduler

Took 0 sec. Last updated by anonymous at May 29 2018, 11:00:00 AM.

A modal dialog box is open in the center, titled 'Run note with cron scheduler. Either choose from preset or write your own cron expression.' It contains the following options:

- Preset: [None](#) [1m](#) [5m](#) [1h](#) [3h](#) [6h](#) [12h](#) [1d](#)
- Cron expression:
- After execution stop the interpreter ☒

On the right side of the workspace, the status 'FINISHED' is shown next to icons for play, expand, view, and settings.

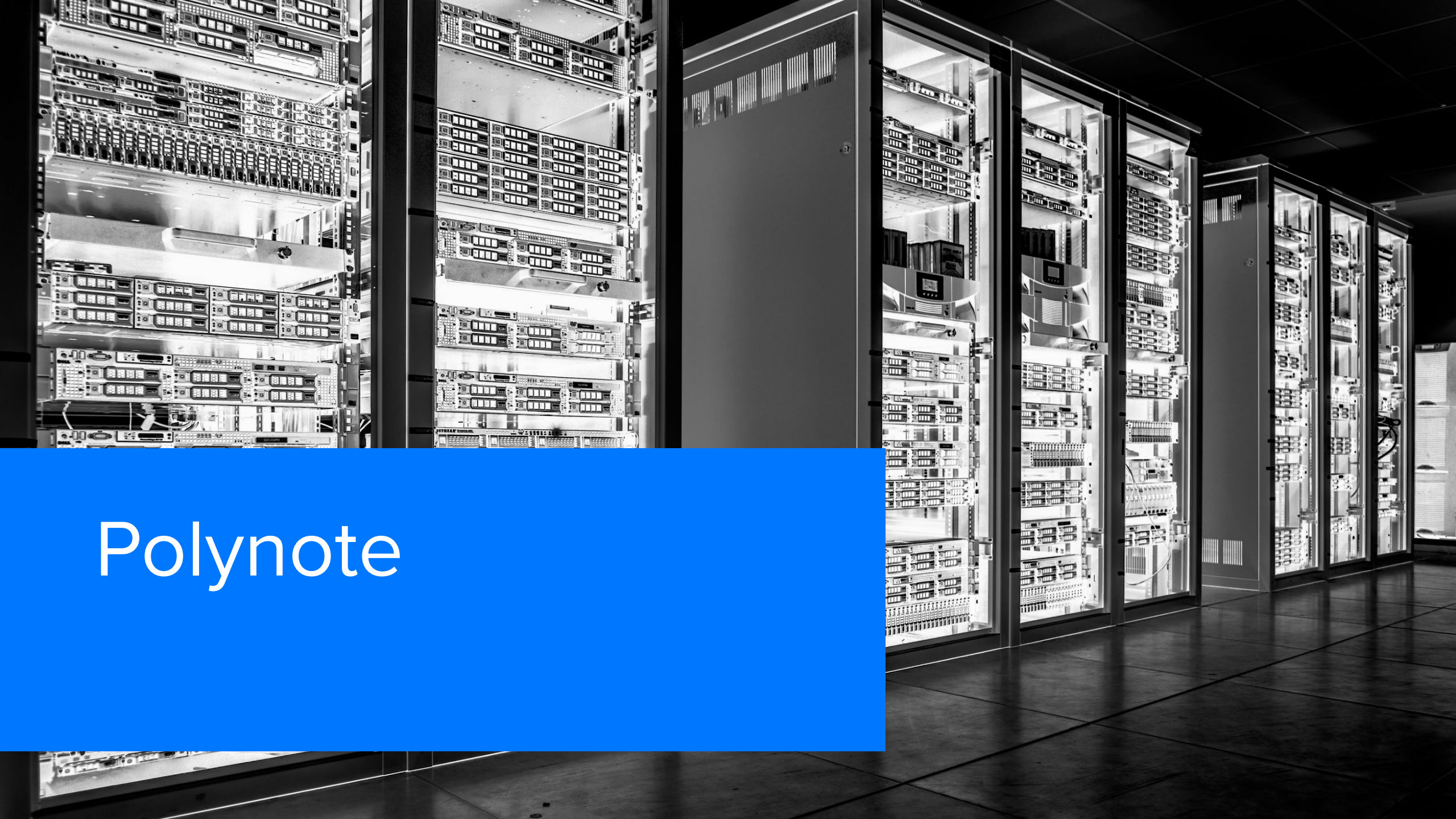
Notebook storages

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/>

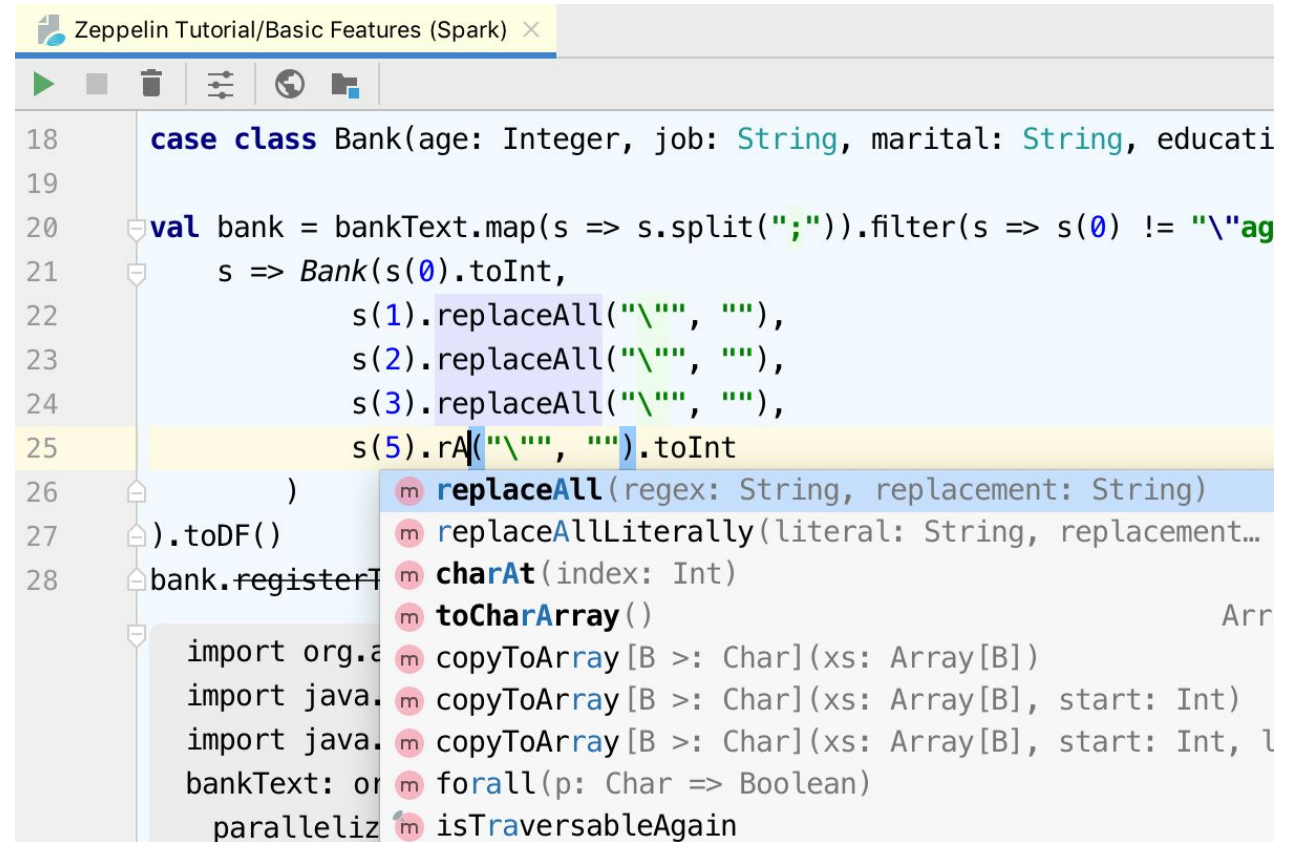




Big Data Tools

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



```
Zeppelin Tutorial/Basic Features (Spark) x
▶ ■ ⌵ ⌵ ⌵ ⌵ ⌵
18 case class Bank(age: Integer, job: String, marital: String, educati
19
20 val bank = bankText.map(s => s.split(";")).filter(s => s(0) != "\"ag
21   s => Bank(s(0).toInt,
22         s(1).replaceAll("\\'", "'"),
23         s(2).replaceAll("\\'", "'"),
24         s(3).replaceAll("\\'", "'"),
25         s(5).replaceAll("\\'", "'").toInt
26   )
27   ).toDF()
28   bank.registerToTable("bank")

import org.apache.spark.sql._
import java.util._
import java.util.concurrent._
bankText: org.apache.spark.rdd.RDD[String]
parallelize

m replaceAll(regex: String, replacement: String)
m replaceAllLiterally(literal: String, replacement: String)
m charAt(index: Int)
m toCharArray()
m copyToArray[B >: Char](xs: Array[B])
m copyToArray[B >: Char](xs: Array[B], start: Int)
m copyToArray[B >: Char](xs: Array[B], start: Int, length: Int)
m forall(p: Char => Boolean)
m isTraversableAgain
```



More vis tools

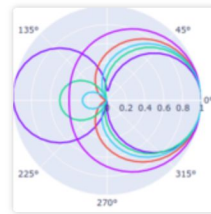
Plotly. Multilang vis tool

Fundamentals

[More Fundamentals »](#)



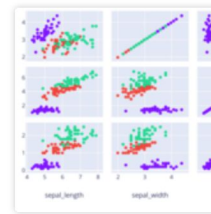
The Figure Data Structure



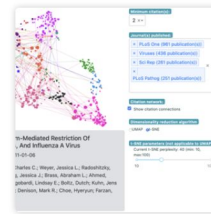
Creating and Updating Figures



Displaying Figures



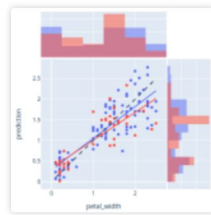
Plotly Express



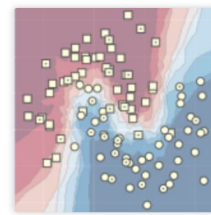
Analytical Apps with Dash

Artificial Intelligence and Machine Learning

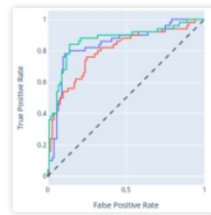
[More AI and ML »](#)



ML Regression



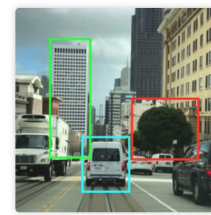
kNN Classification



ROC and PR Curves



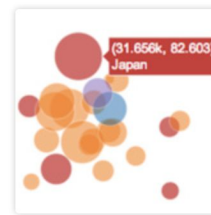
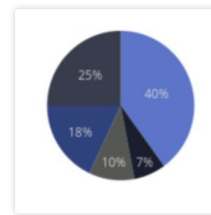
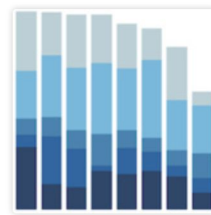
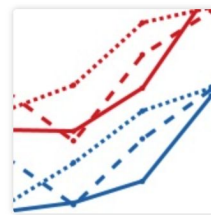
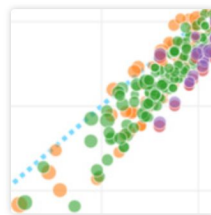
PCA Visualization



AI/ML Apps with Dash

Basic Charts

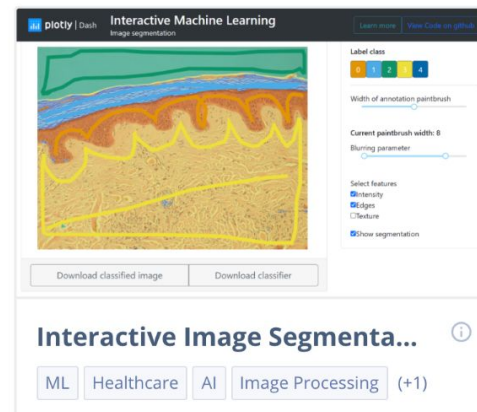
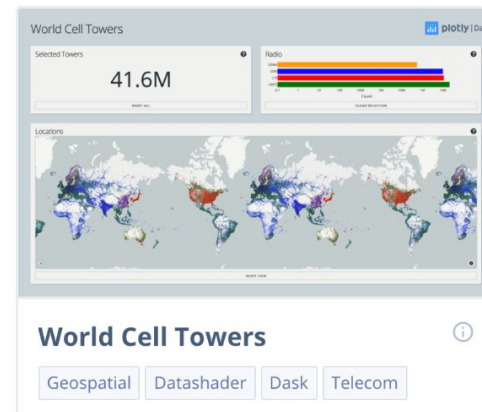
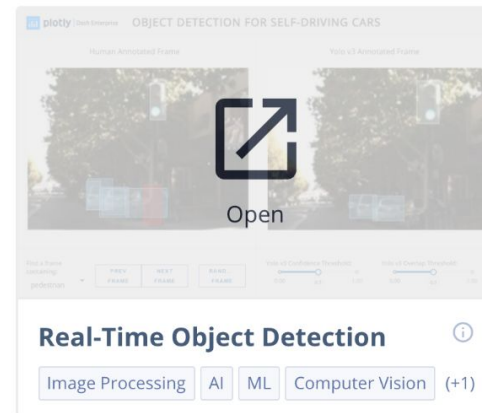
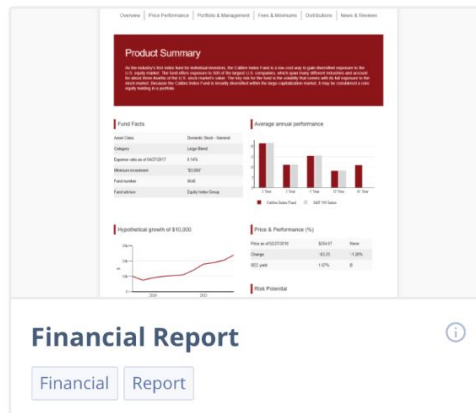
[More Basic Charts »](#)



[Plotly Python Graphing Library | Python](#)

Dash. Visual apps at web

All Apps (110)



<https://dash.gallery/Portal/>

Tableau. Expensive industrial standart

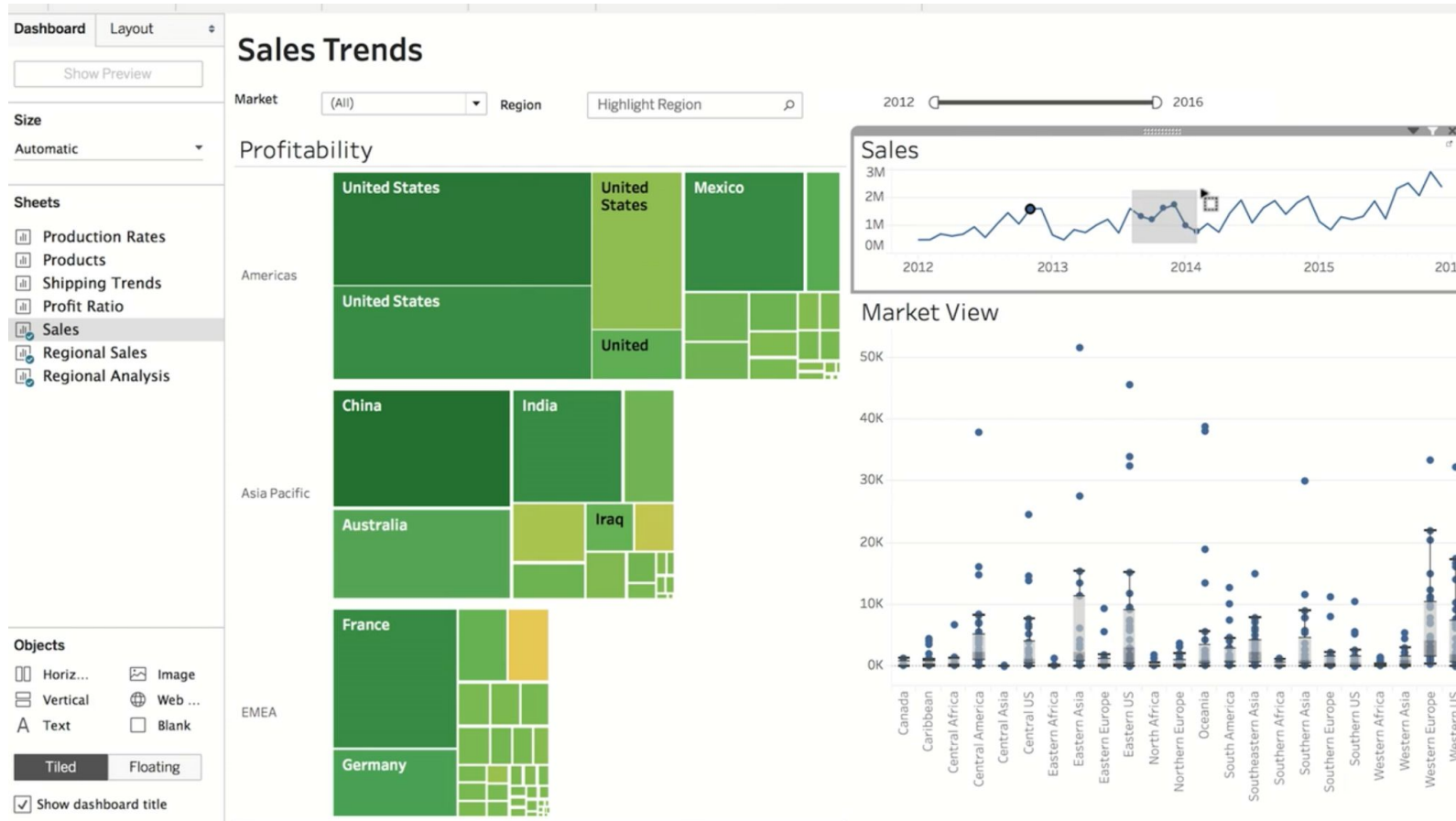
Tableau native data connectors



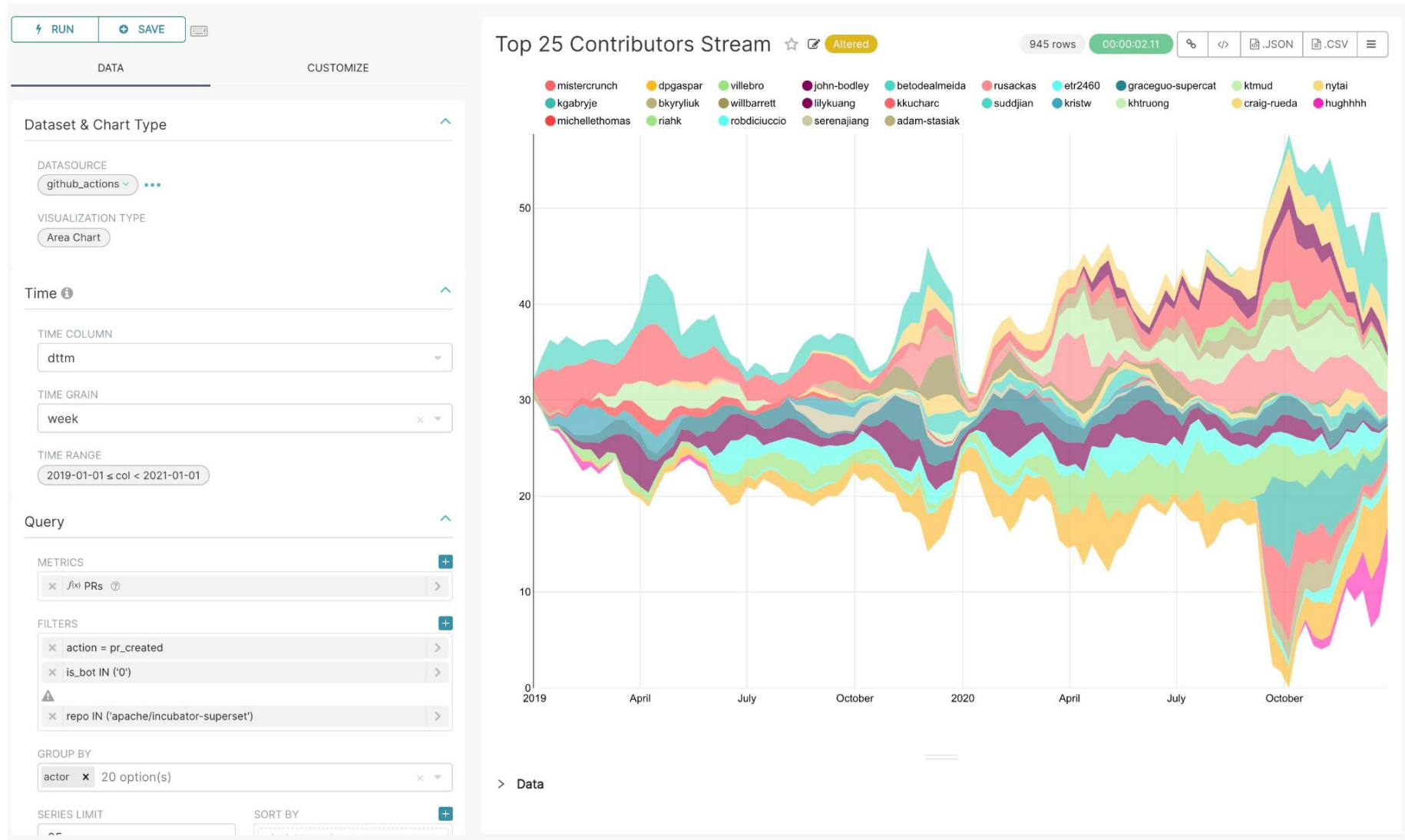
Connect to all of your data—no matter where it resides.

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



SuperSet

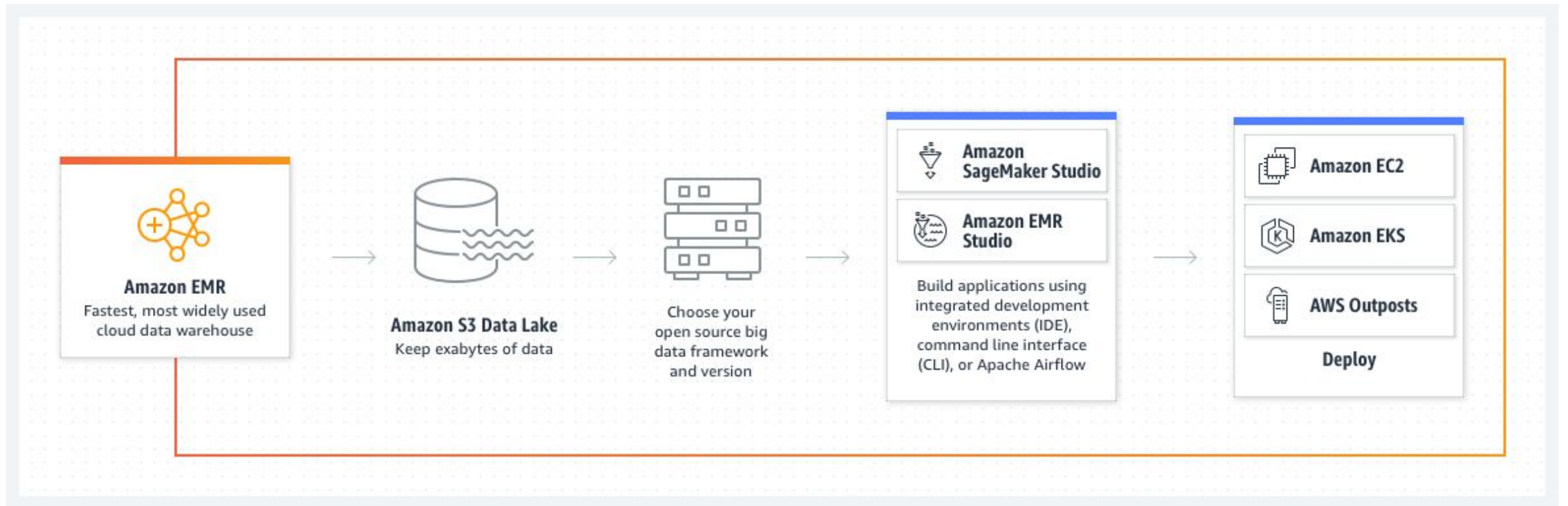




Cloud ecosystems for Big Data

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.

VK Cloud Solutions

Mail.ru Cloud Solutions

Облачные вычисления

Виртуальные сети

Объектное хранилище

Контейнеры

Базы данных

Аналитические БД

Магазин приложений

Большие данные

Кластеры

Графические адаптеры

Машинное обучение

Специальные сервисы

Управление доступами

Баланс

cloud big data

Помощь

Большие данные

Облачные Big Data сервисы MCS обеспечивают быстрое решение задач обработки больших данных, событий и realtime-аналитики на базе Hadoop и Spark


Кластеры

Быстрое создание кластеров на базе Hadoop Hortonworks

Создать кластер

S3-совместимое хранилище и кластеры Spark до 16 TB RAM

Создать кластер



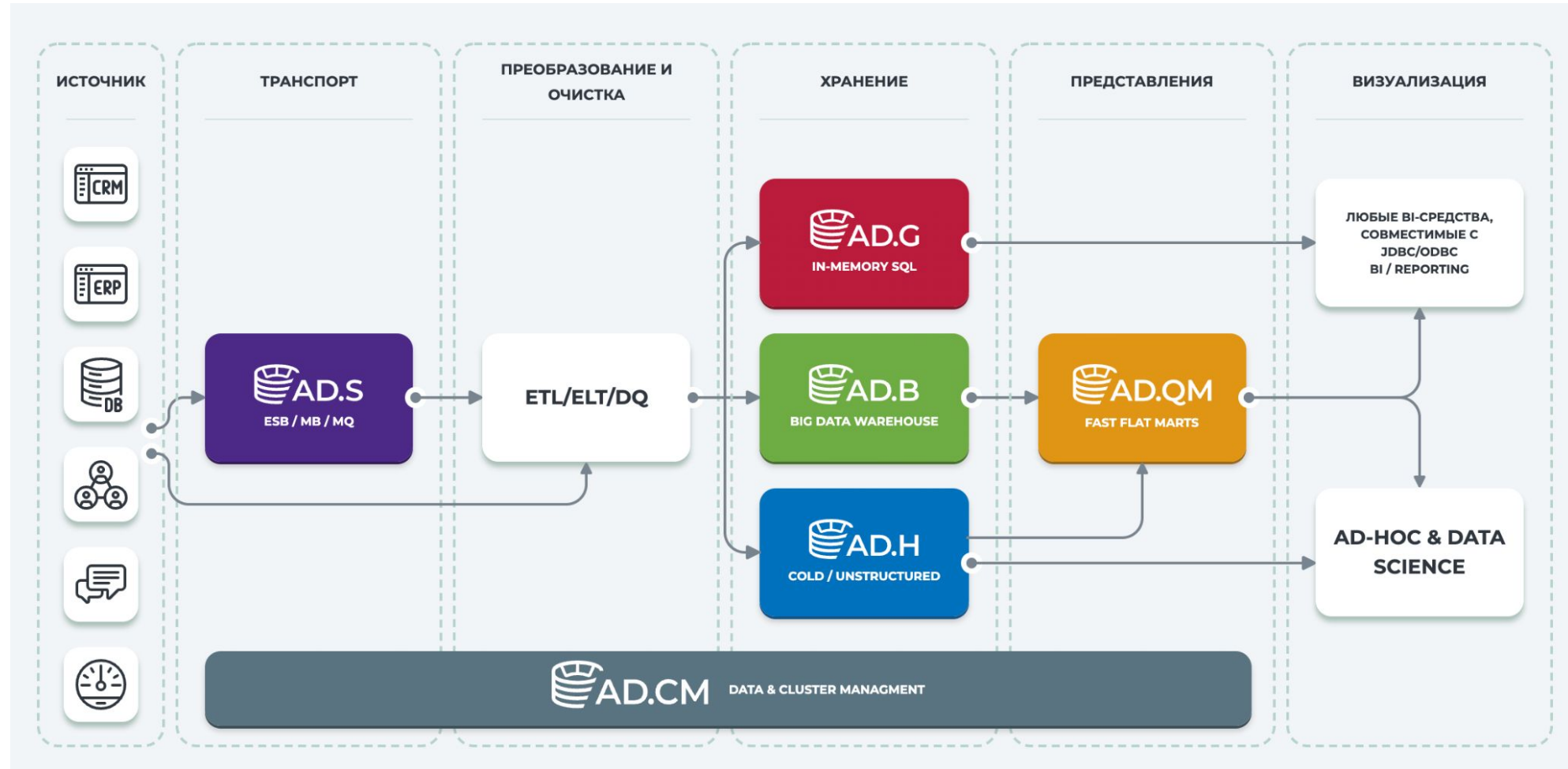
- ✓ Динамическое масштабирование до сотен узлов при пиковых нагрузках
- ✓ Посекундная тарификация, отсутствие затрат на закупку оборудования

Если у вас возникли вопросы, вы можете ознакомиться с [документацией](#)

f

ArenaData

<https://arenadata.tech/>



Snowflake

**WHERE YOUR DATA CLOUD EXPERIENCE BEGINS:
ONE PLATFORM, MANY WORKLOADS, NO DATA SILOS**



Recommended links and literature

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