

Инструменты аналитика

Сбор данных

- Scraping-Bot
- Scrapeworks
- Diggernaut
- ScrapingBee
- Scraper API

Обзор инструментов

- Excel
- SAS, SPSS
- MATLAB, Octave
- Онлайн
платформы
- Hadoop, Spark
- Python, R
- Ноутбуки
- Библиотеки



Джон
Форман

МНОГО ЦИФР

Анализ больших данных
при помощи Excel



Знания — это сила, а знания, полученные
из больших данных, — большая сила.



MATLAB

File Edit Debug Desktop Window Help

Shortcuts How to Add What's New

Workspace

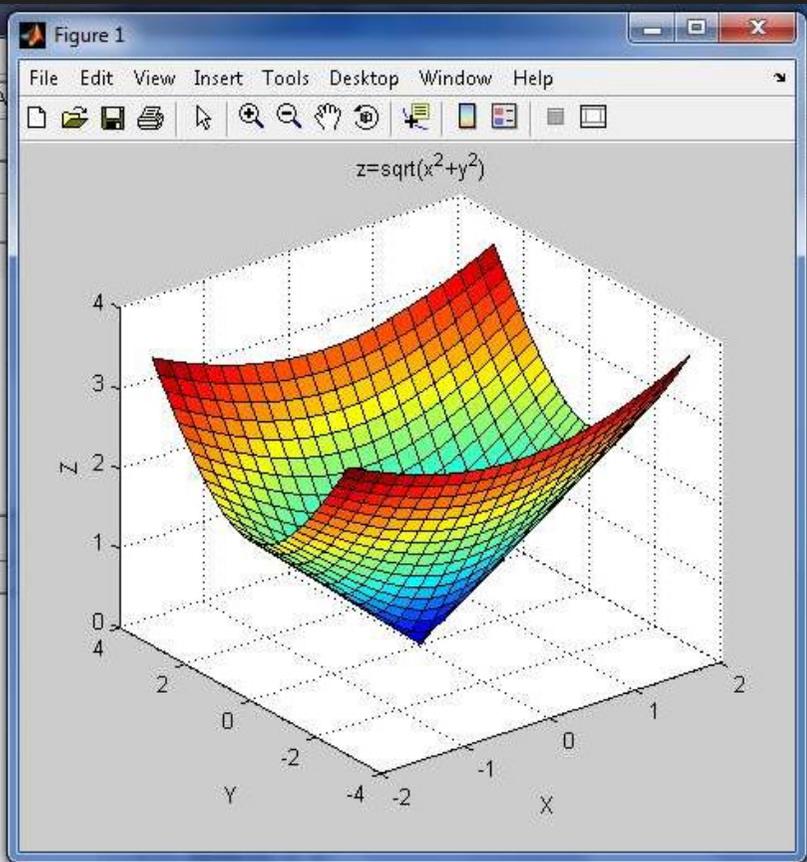
N...	Value	Class
a	3	double
phi	<1x401 double>	double
ro_1	<1x401 double>	double
ro_2	<1x401 double>	double
t	<1x101 double>	double
x	<31x21 double>	double
y	<31x21 double>	double

Current Directory Workspace

```

y=a*t.*(t.^2-1)./(t.^2+1);
plot(x,y,'c-');
grid on;
title('(x+a)a^2+(x-a)y^2=0, a=3');
xlabel('X');
ylabel('Y');
[x y]=meshgrid(-2:0.5:2, -3:0.5:3)
[x y]=meshgrid(-2:2, -3:3)
z=y.^2-x.^2
mesh(x,y,z)
[x y]=meshgrid(-2:0.2:2, -3:0.2:3);
z=sqrt(x.^2+y.^2);
surf(x,y,z);
grid on;
title('z=sqrt(x^2+y^2)');
xlabel('X');
ylabel('Y');
zlabel('Z');

```



```

>> [x y]=meshgrid(-2:0.2:2, -3:0.2:3);
>> z=sqrt(x.^2+y.^2);
>> surf(x,y,z);
>> grid on;
>> title('z=sqrt(x^2+y^2)');
>> xlabel('X');
>> ylabel('Y');
>> zlabel('Z');
>>

```

Облака

- Amazon AWS
- Microsoft Azure
- IBM Watson Analytics

The screenshot displays the Microsoft Azure Machine Learning Studio interface. The top bar shows the title "Microsoft Azure Machine Learning Studio" and various utility icons. The left sidebar contains a search bar and a list of experiment items, including "Saved Datasets", "Trained Models", "Data Format Conversions", "Data Input and Output", "Data Transformation", "Feature Selection", "Machine Learning", "OpenCV Library Modules", "Python Language Modules", "R Language Modules", "Statistical Functions", "Text Analytics", "Time Series", "Web Service", and "Deprecated".

The main workspace is titled "Auto ARIMA" and shows a workflow diagram. The workflow starts with an "Enter Data Manually" module (marked with a '1' and a green checkmark), which connects to an "Execute R Script" module. The "Execute R Script" module has two inputs: "Web service input" and "Web service output". The "Execute R Script" module also has two outputs: "Web service input" and "Web service output".

The bottom bar contains several action buttons: "NEW", "RUN HISTORY", "SAVE", "SAVE AS", "DISCARD CHANGES", "RUN", "DEPLOY WEB", and "PUBLISH TO".



Getting Started

Manage Data

Open Workbook

Welcome to Watson Analytics!
Explore our solutions by role →



MARKETING



SALES



FINANCE



OPERATIONS



HR



IT

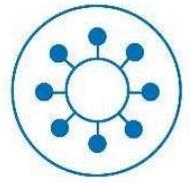
Enter a keyword to filter the list below, or to ask Watson a question about your data!



Start from Data

TOOL

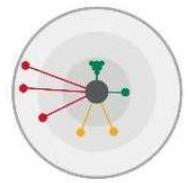
EXPLORE YOUR DATA



The easy, beautiful way to find the stories in your data

TOOL

PREDICT AND EXPLAIN



Discover the drivers of behavior and results

TOOL

FORECAST



Start from a Story

TUTORIAL

GETTING STARTED WITH WATSON ANALYTICS

Take a tour of Watson Analytics!

MARKETING

IMPROVE CAMPAIGN EFFECTIVENESS

Understand the drivers of campaign success

HR

RETAIN YOUR TEAM

Identify high risk employees

TUTORIAL

WORKING WITH DATA

Get more from your data with Watson Analytics

HR

PREVENTING EMPLOYEE ATTRITION

Identify the causes of attrition before its too late

SALES

FIND PATTERNS IN WINS AND LOSSES

What combination of factors leads to a win?

HR

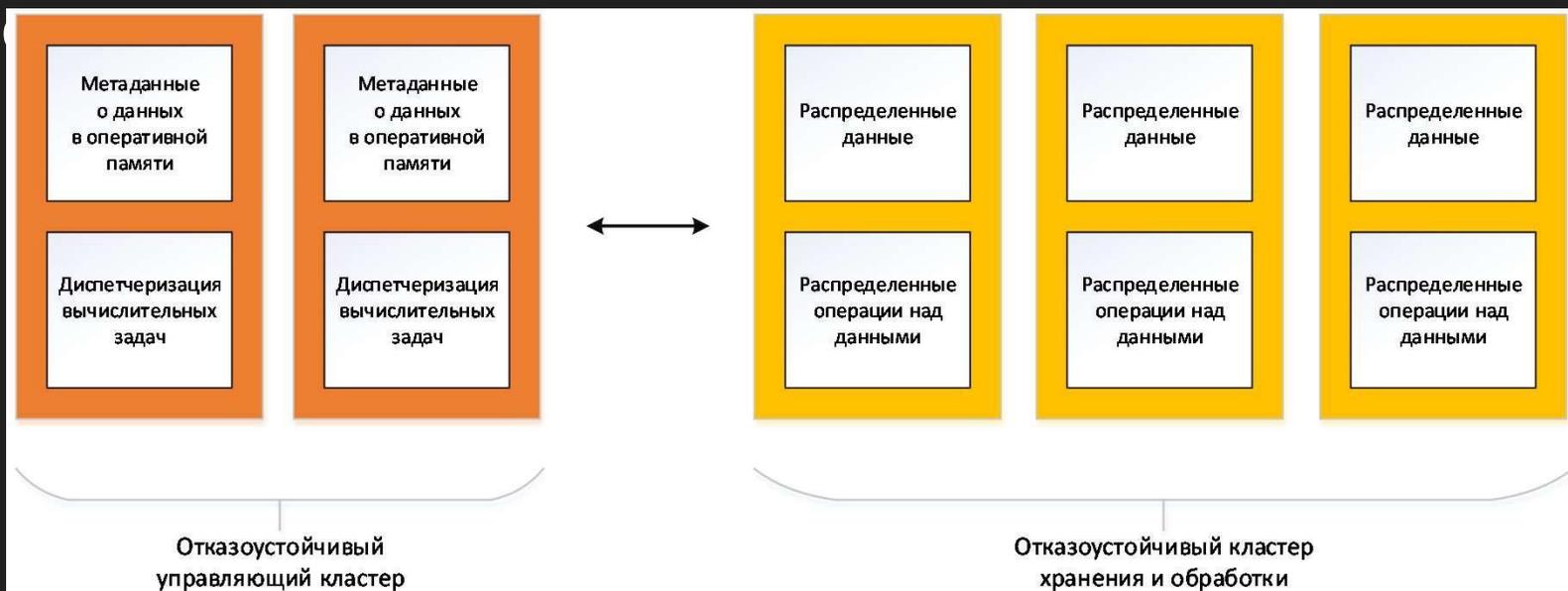
MARKETING

SALES

Зачем нужен Hadoop

- В 10 раз дешевле СХД
- Вычисления и данные — в одном месте
- Вместе с удешевлением HDD устроил

рев



Вычисления на кластере

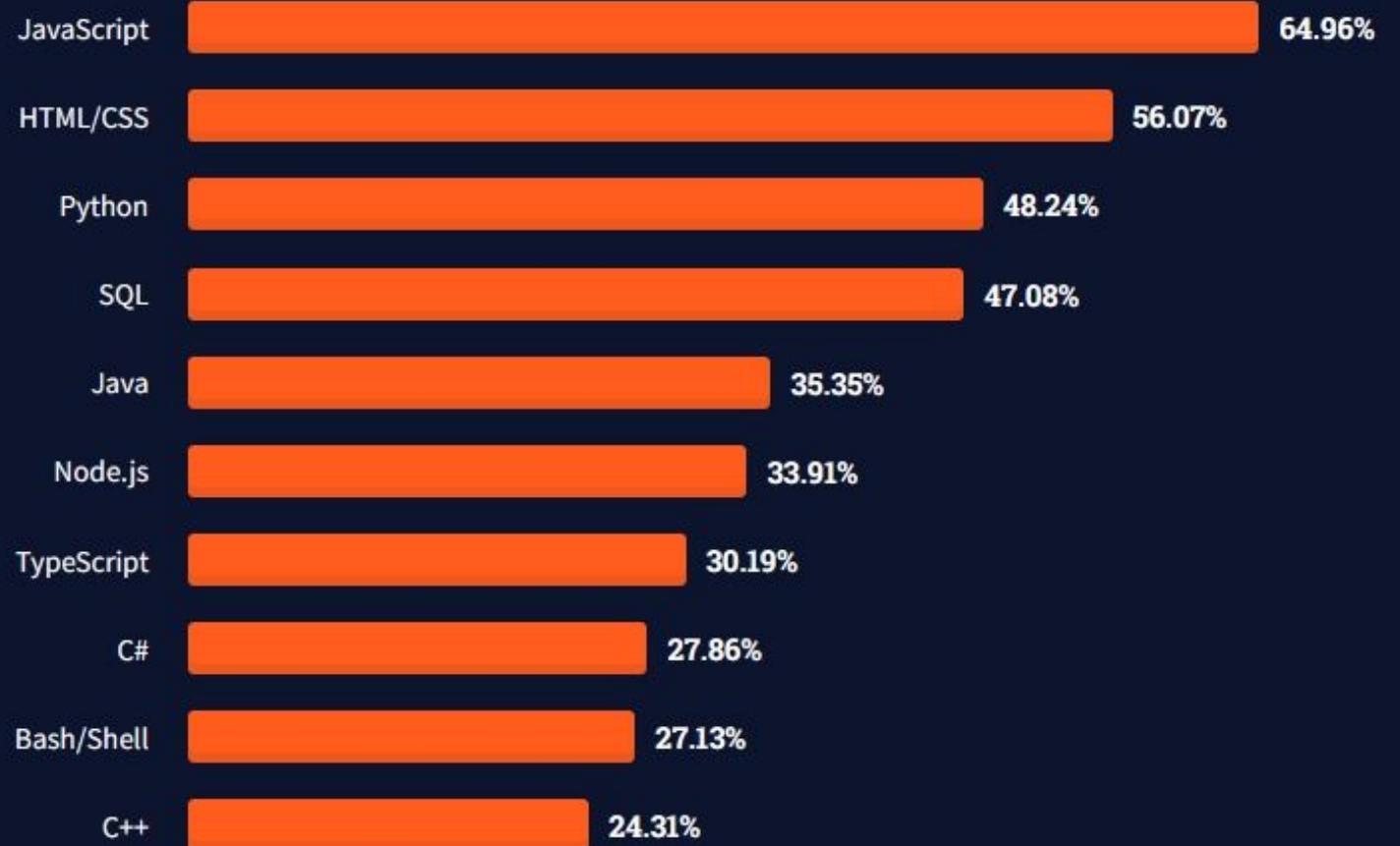
- Единый шедуллер, разделяющий ресурсы между всеми
- Типичные потребители:
 - Базы данных (HBASE)
 - SQL-like инструменты (HIVE)
 - Map Reduce операции
 - Spark (in-memory)
- Крайне важно грамотно организовать параллельность

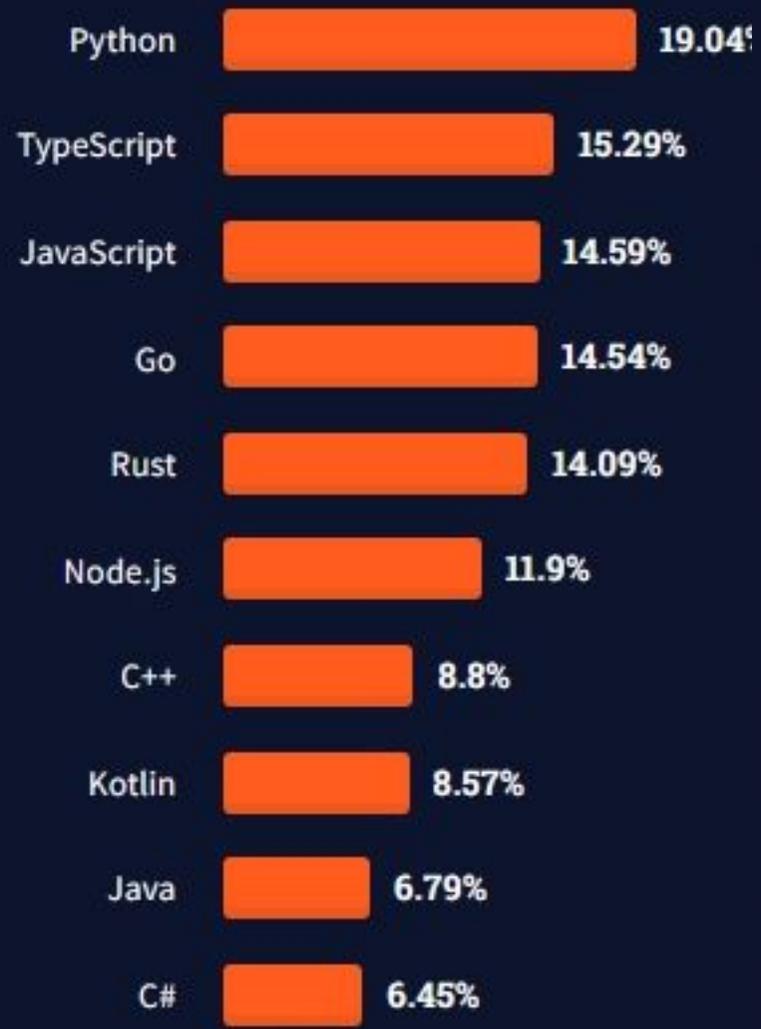
Python

- Все становятся программистами
- Преимущества Python
 - Простота и удобство, легкость в освоении
 - Расширяемость, огромное количество библиотек и примеров
 - Data Science стек, парсинг сайтов, веб-сервисы...
- Недостатки Python
 - Медленнее компилируемых языков e.g. C++, Java
 - Не подходит для мобильной разработки
 - Не всегда лучшее решение для enterprise

All Respondents

Professional Developers





Рейтинг языков программирования по количеству вакансий

Язык программирования	Количество вакансий для языка программирования
SQL	28729
JavaScript	15903
Python	15027
Java	11362
C#	6697
PHP	6539
C++	5612
TypeScript	4227
Kotlin	3012
Golang	1560
Swift IOS	1457
Ruby	929
Scala	867
Objective-C	686
ABAP	631
Delphi	615
Perl	585
Assembler	357
Dart	246
Rust	165
Elixir	98
Erlang	93
Haskell	43
F#	15

О

ЯЗЫКЕ

- 1991 год рождения, Нидерланды
- Основан на ABC, который основан на SETL, 1969

C++ – 1983, C – 1973

R – 1993, S – 1976

JavaScript – 1995

SQL – 1979

Красивое лучше, чем уродливое. Явное лучше, чем неявное.

Простое лучше, чем сложное.

Сложное лучше, чем запутанное. Плоское лучше, чем вложенное. Разреженное лучше, чем плотное.

Читаемость имеет значение.

Особые случаи не настолько особые, чтобы нарушать правила.

При этом практичность важнее

безупречности. Ошибки никогда не должны замалчиваться.

Если они не замалчиваются явно.

Встретив двусмысленность, отбрось искушение угадать. Должен существовать один и, желательно, только один очевидный способ сделать это.

Хотя он поначалу может быть и не очевиден, если вы не голландец.

Сейчас лучше, чем никогда.

Хотя никогда зачастую лучше, чем прямо сейчас. Если реализацию сложно объяснить — идея плоха.

Если реализацию легко объяснить — идея,

возможно

Основные свойства

- Python – интерпретируемый язык
- CPython – основная реализация интерпретатора, написан на C
- Динамическая типизация
- «white space» играет роль
- Установка модулей через пакетный менеджер (pip, conda)
- 130 000 различных модулей (март 2021)

Hello, world!

C++

```
#include <iostream>
using namespace std;

int main()

{
    // print output to user
    cout << "Hello, world!" << endl;
    return 0;
}
```

Python

```
print("Hello world!")
```

Python

- Numpy
- Scipy
- Pandas
- Matplotlib
- Scikit-learn

- Ну и тысячи других

Модели, фреймворки

- Градиентный бустинг

- XGBoost
- Catboost
- LightGBM

- Нейросети

- Keras
- Caffe
- TensorFlow
- Theano
- PyTorch

- Обёртки для языков
- Применение из консоли
- Параллельное обучение
- Параллельное применение
- На одной машине и на кластере
- Исполнение на CPU, GPU
- Поддержка Windows/Mac/Linux
- Поддержка ARM

[Don't Miss AnacondaCon Apr 8-11 Austin TX!](#)

Download Anaconda Distribution

Version 5.1 | Release Date: February 15, 2018

Download For:   

IPython и Jupyter

- IPython — интерактивная консоль python' а
- Jupyter — популярный аналитический ноутбук

Jupyter spectrogram (autosaved)

File Edit View Insert Cell Kernel Help

Save + Copy Paste Undo Redo Run Stop Refresh Markdown CellToolbar

Simple spectral analysis

An illustration of the [Discrete Fourier Transform](#)

$$X_k = \sum_{n=0}^{N-1} x_n \exp\left(\frac{-2\pi i}{N} kn\right) \quad k = 0, \dots, N-1$$

```
In [2]: from scipy.io import wavfile
rate, x = wavfile.read('test_mono.wav')
```

And we can easily view it's spectral structure using matplotlib's builtin specgram routine:

```
In [5]: fig, (ax1, ax2) = plt.subplots(1,2,figsize(16,5))
ax1.plot(x); ax1.set_title('Raw audio signal')
ax2.specgram(x); ax2.set_title('Spectrogram');
```

Raw audio signal

Spectrogram

20

Облака

- Google Cloud Platform

The screenshot shows the Google Cloud Platform interface for a billing account. The top navigation bar includes the Google Cloud Platform logo, a search bar, and a menu icon. Below the navigation bar, the page title is "Overview" and "My Billing Account". There are two action buttons: "RENAME BILLING ACCOUNT" and "CLOSE BILLING ACCOUNT". The main content area is divided into two tabs: "Billing account overview" (selected) and "Payment overview". The "Billing account overview" tab shows the billing account ID: 0027A8-81D87A-037C4B. Below this, there is a section for "Credits" with a table showing one credit: "Free Trial" with an expiration date of "May 8, 2018", a promotion value of "\$300.00", and an amount remaining of "\$119.55". At the bottom, there is a section for "Projects linked to this billing account" with a table showing one project: "My First Project" with a project ID of "bamboo-position-167008".

Google Cloud Platform

Overview My Billing Account [RENAME BILLING ACCOUNT](#) [CLOSE BILLING ACCOUNT](#)

Billing account overview Payment overview

Billing account ID: 0027A8-81D87A-037C4B

Credits

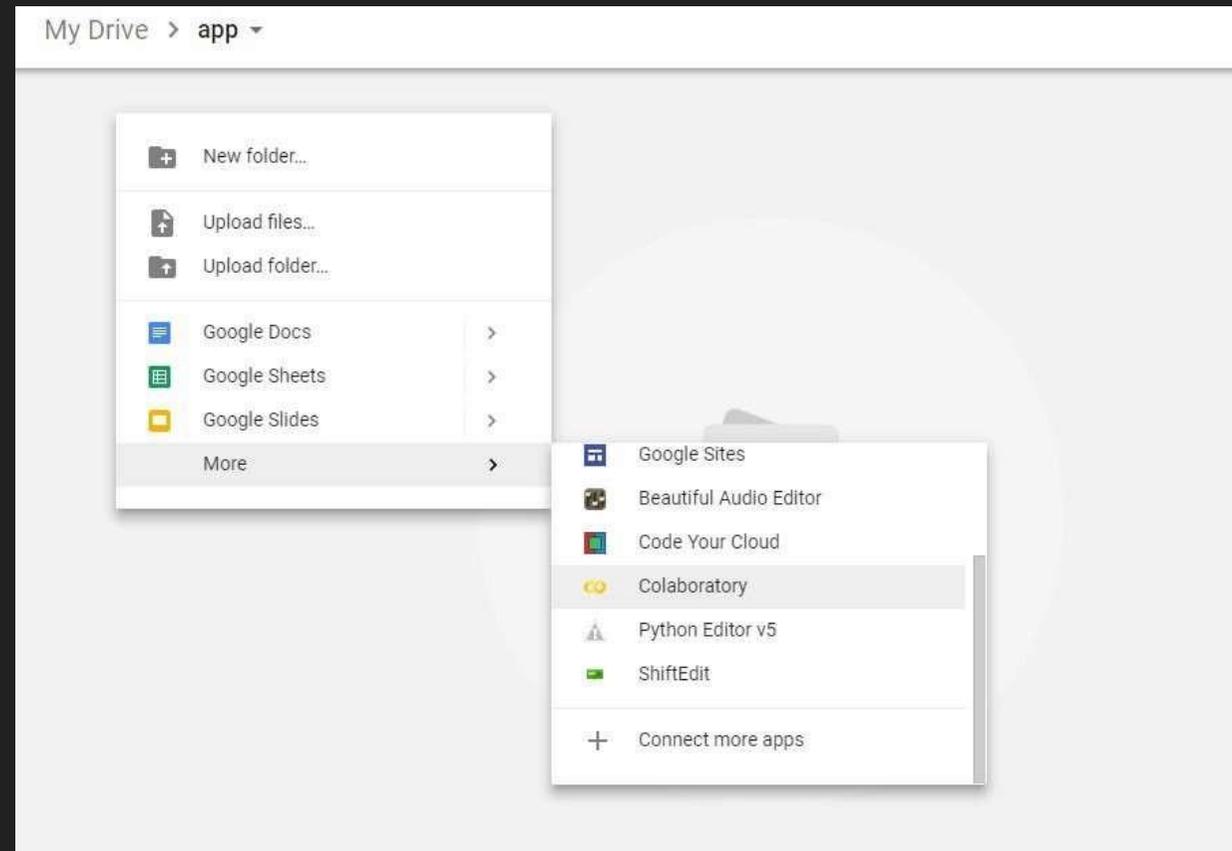
Promotion ID	Expires ^	Promotion value	Amount remaining
Free Trial	May 8, 2018	\$300.00	\$119.55

Projects linked to this billing account

Project name	Project ID
My First Project	bamboo-position-167008

Облака

- Google Colab



Tableau, Power BI, Google Data Studio: простая онлайн-визуализация без

The screenshot displays a Tableau dashboard titled "Traffic Fatalities Report" for the period of January 2015 to December 2015. The dashboard includes a table of fatalities by location, a pie chart showing the distribution of incidents, a line chart of total fatalities over time, and a bar chart of the most unlucky states. A settings panel on the right is open, showing options for layout and theme.

Traffic Fatalities Report

January 2015 - December 2015

relation_to_traffic...	number_of_fatalities
1. On Roadway	21,195
2. On Roadside	10,826
3. On Median	1,362
4. Outside Trafficway	692
5. On Shoulder	519
6. Gore	177
7. Separator	96
8. Off Roadway - Locatio...	77
9. In Parking Lane/Zone (...)	65
10. Unknown	56

Where do incidents occur?

state_name

- On Roadway (60.4%)
- On Roadside (30.9%)
- On Median
- Outside Trafficway
- others

Total number of fatalities

number_of_fatalities

Jan 1, Feb 10, Mar 22, May 1, Jun 10, Jul 20, Aug 29, Oct 8, Nov 17, Dec 27

What are the most unlucky states?

1 - 10 / 12

- Texas
- California
- Florida
- Georgia
- North Carolina

1.5K

number_of_drunk_drivers, number_of_forms_submitted_for_persons_...

900

Layout and Theme

LAYOUT | THEME

View Mode

Header visibility

Visible Auto hide

Navigation position

Left Top

Display mode

Fit to width Actual size

Has margin

Canvas Size

US letter (4:3) - Landscape

Width (px) 1200 Height (px) 900

Программа на Python

```
problems_test.py x
1 # coding=utf-8
2 # Copyright 2018 The Tensor2Tensor Authors.
3 #
4 # Licensed under the Apache License, Version 2.0 (the "License");
5 # you may not use this file except in compliance with the License.
6 # You may obtain a copy of the License at
7 #
8 # http://www.apache.org/licenses/LICENSE-2.0
9 #
10 # Unless required by applicable law or agreed to in writing, software
11 # distributed under the License is distributed on an "AS IS" BASIS,
12 # WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13 # See the License for the specific language governing permissions and
14 # limitations under the License.
15
16 """tensor2tensor.problems test."""
17
```

Программа на Python

```
21  
22     # Dependency imports  
23  
24     from tensor2tensor import problems  
25  
26     import tensorflow as tf  
27
```

Базовые алгоритмические конструкции

```
a = 1
```

```
b = 2
```

```
c = a + b
```

```
d = a - b
```

```
print(c)
```

```
print(d)
```

Оператор УСЛОВИЯ

if $a > b$:

$c = a$

else:

$c = b$

Оператор УСЛОВИЯ

if a > 0:

 c = a

elif a == 0:

 c = b

else:

 c =

 d

ЦИКЛЫ

```
# while
```

```
m = 0
```

```
while m < 10:
```

```
    m = m + 1
```

```
    print(m)
```

```
# for
```

```
for n in range(1, 10):
```

```
    print(n)
```

ФУНКЦИИ

пример определения и вызова

функции `def time(hour, minute=0):`

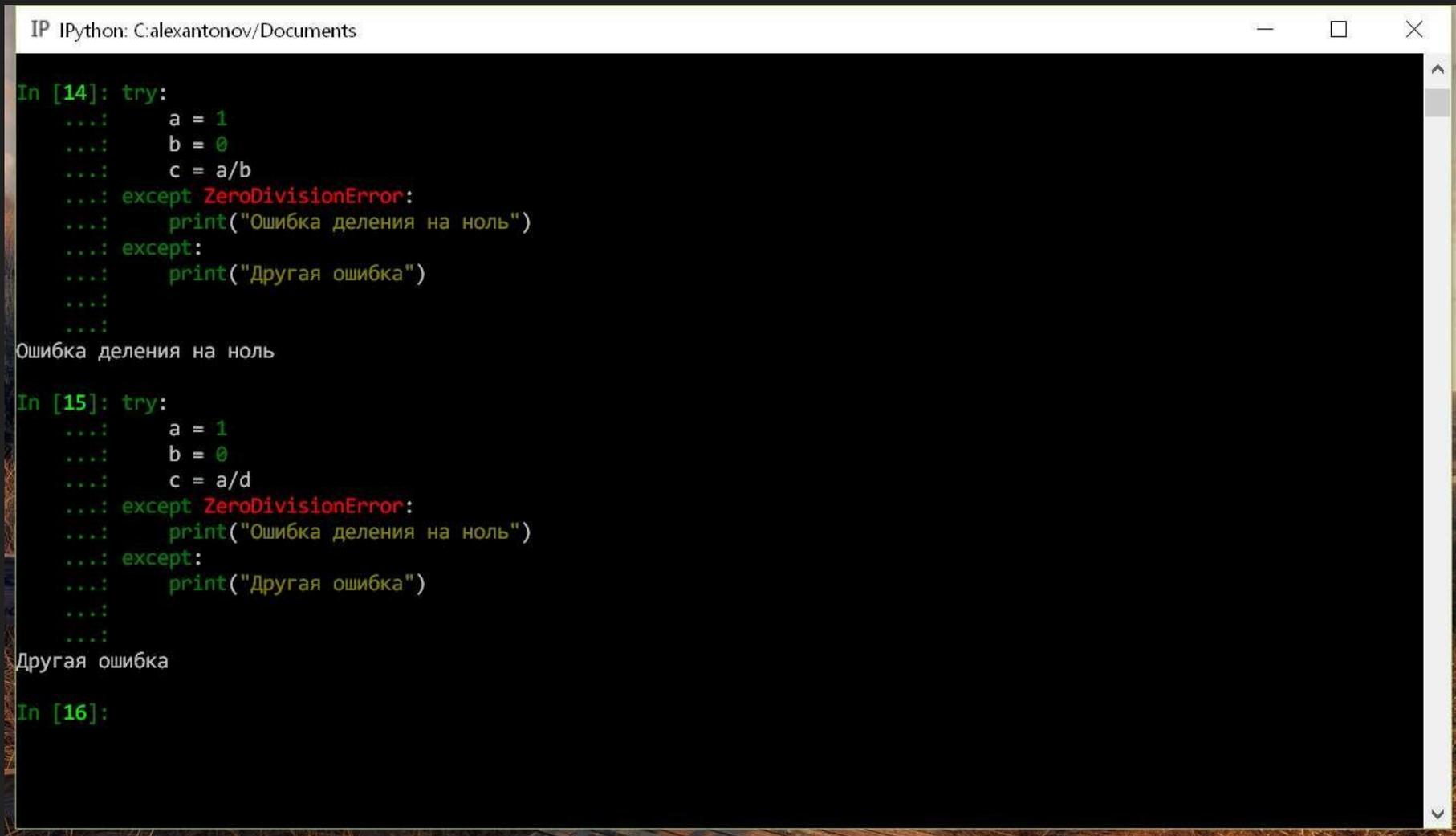
`return("Время: %i часов %i минут" % (hour,
 minute))`

`time(8)`

`time(9, 20)`

`time(minute=5, hour=10)`

Исключени я



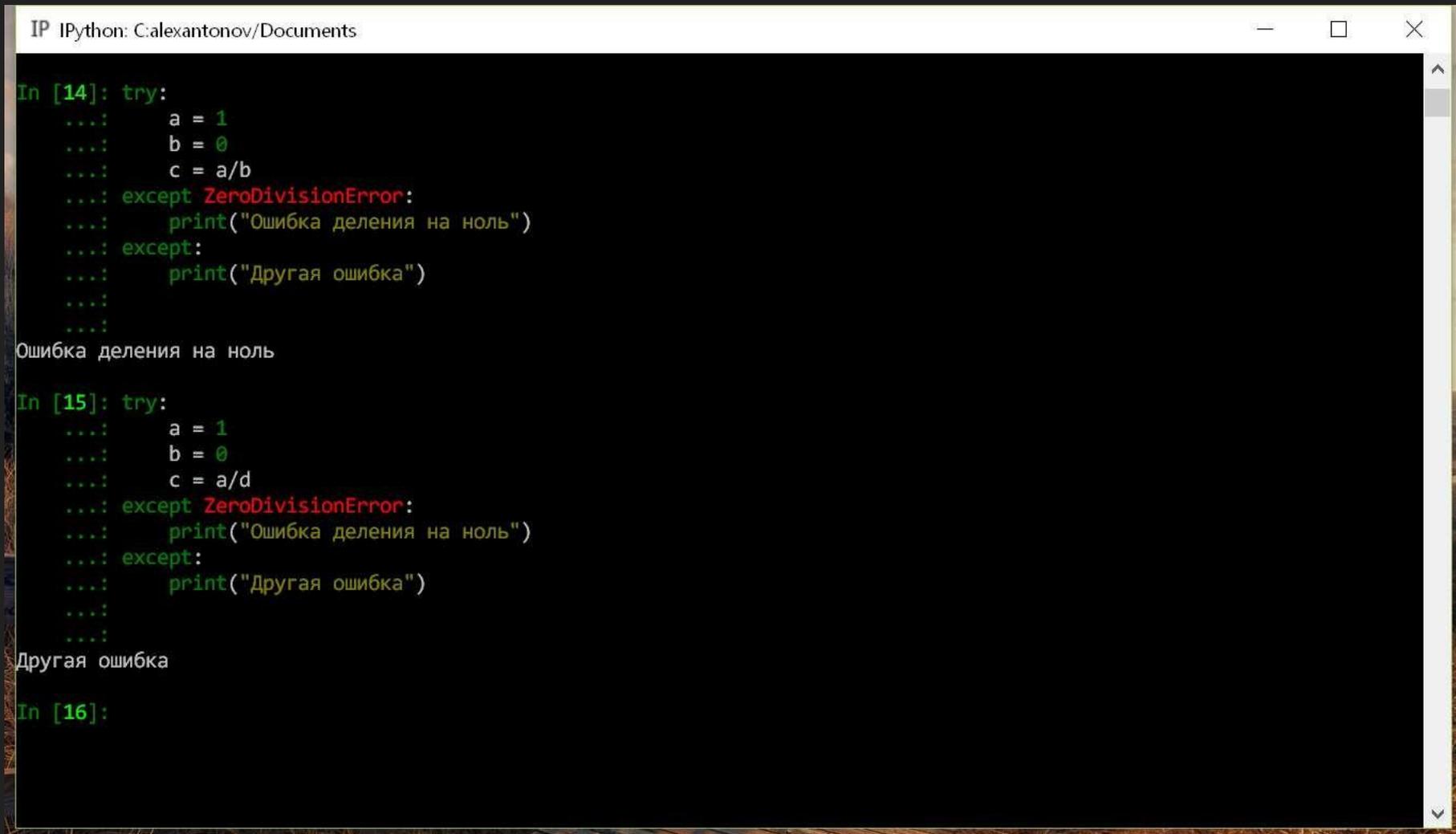
```
IPython: C:\alexantonov\Documents

In [14]: try:
...:     a = 1
...:     b = 0
...:     c = a/b
...: except ZeroDivisionError:
...:     print("Ошибка деления на ноль")
...: except:
...:     print("Другая ошибка")
...:
...:
Ошибка деления на ноль

In [15]: try:
...:     a = 1
...:     b = 0
...:     c = a/d
...: except ZeroDivisionError:
...:     print("Ошибка деления на ноль")
...: except:
...:     print("Другая ошибка")
...:
...:
Другая ошибка

In [16]:
```

Исключени я



```
IPython: C:\alexantonov\Documents

In [14]: try:
...:     a = 1
...:     b = 0
...:     c = a/b
...: except ZeroDivisionError:
...:     print("Ошибка деления на ноль")
...: except:
...:     print("Другая ошибка")
...:
...:
Ошибка деления на ноль

In [15]: try:
...:     a = 1
...:     b = 0
...:     c = a/d
...: except ZeroDivisionError:
...:     print("Ошибка деления на ноль")
...: except:
...:     print("Другая ошибка")
...:
...:
Другая ошибка

In [16]:
```

Исключени я

```
IPython: C:\alexantonov/Documents

In [29]: a = 1
In [30]: b = 2
In [31]: c = a + b
In [32]: assert c == a + b
In [33]: c = c + 1
In [34]: assert c == a + b
-----
AssertionError                                Traceback (most recent call last)
<ipython-input-34-c66e68564792> in <module>()
----> 1 assert c == a + b

AssertionError:

In [35]:
```

Типы данных

- целое число `int` и `long`
- число с плавающей точкой
`float`
- логический `bool` (`True` или
`False`)
- строка `string`

Типы данных

- кортеж `tuple` `p = 1, "Winter", True`
- список `list` `l = [1, 2, 3, 4]`
- словарь `dict` `d = {1: 'one', 2: 'two', 3: 'three', 4: 'four'}`