Data Analytics on Conference Session Catalog

using Jupyter Notebooks

handson workshop

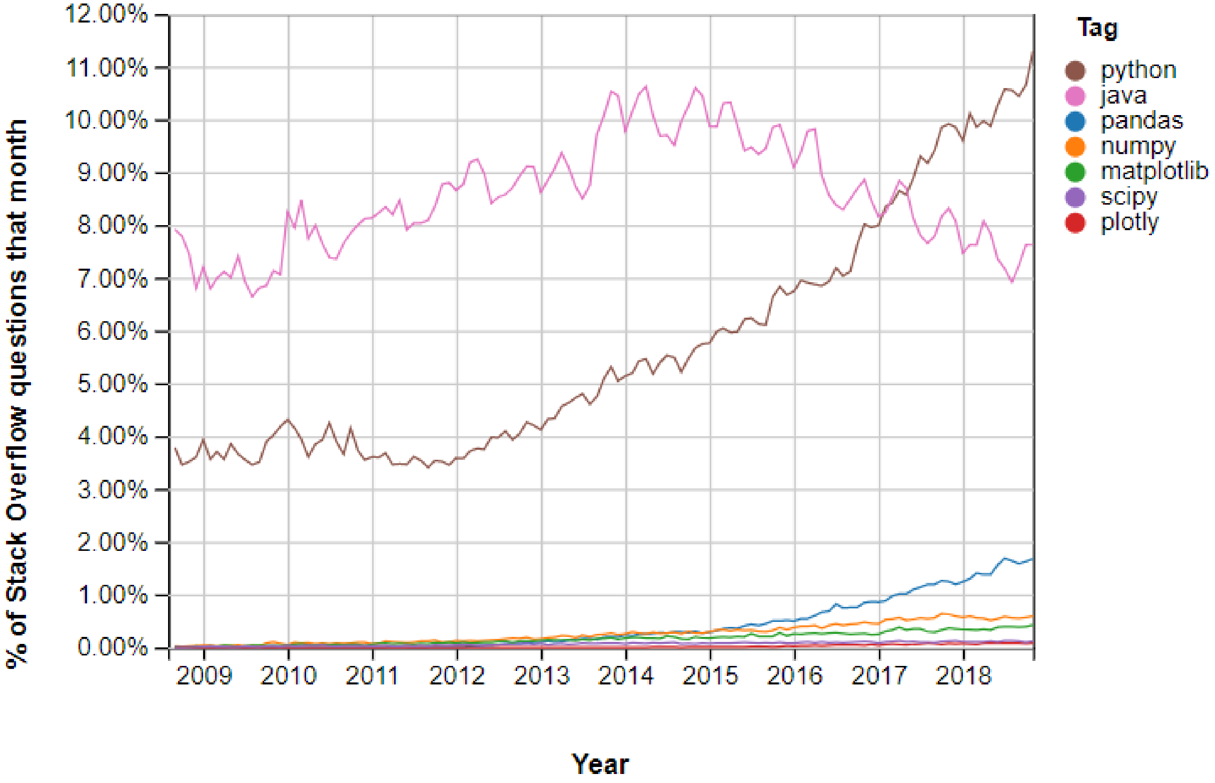
Lucas Jellema
Conclusion Machine Learning Gilde
21 februari 2019

Hands On met Jupyter Notebooks in vier stappen

- Een werkende Jupyte Notebook Server omgeving
  - Lokaal – op basis van Docker container
  - Cloud – in een KataCoda omgeving
- Hello World Notebook
  - Eerste stappen met Notebook, Markdown, Python & Pandas
- Casus Oracle OpenWorld 2018 Session Catalog
  - Gather
  - Wrangle
  - Analyze
  - Visualize
- Doe Het Zelf Notebook met Titanic Data Set

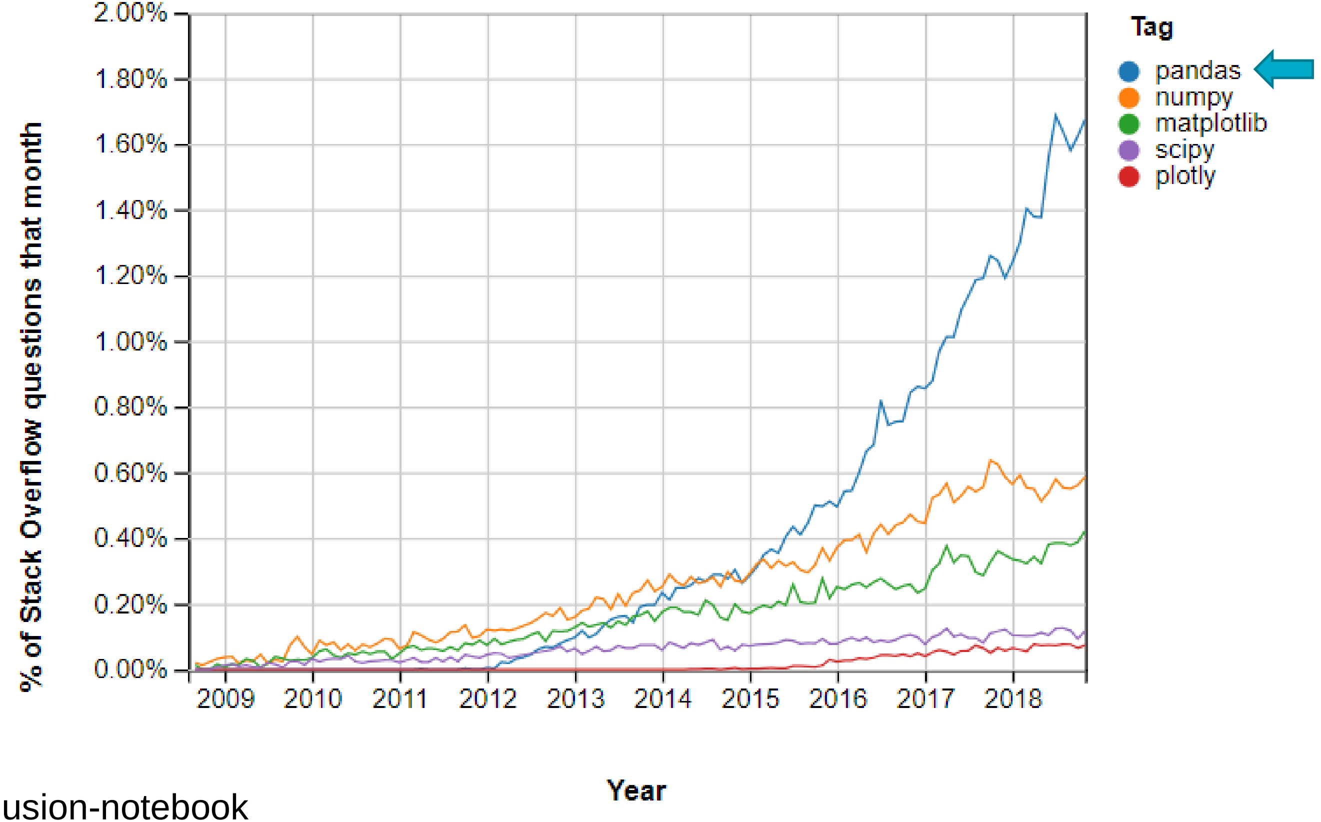
Resources: bit.ly/conclusion-notebook

Groeiende belangstelling in Python…

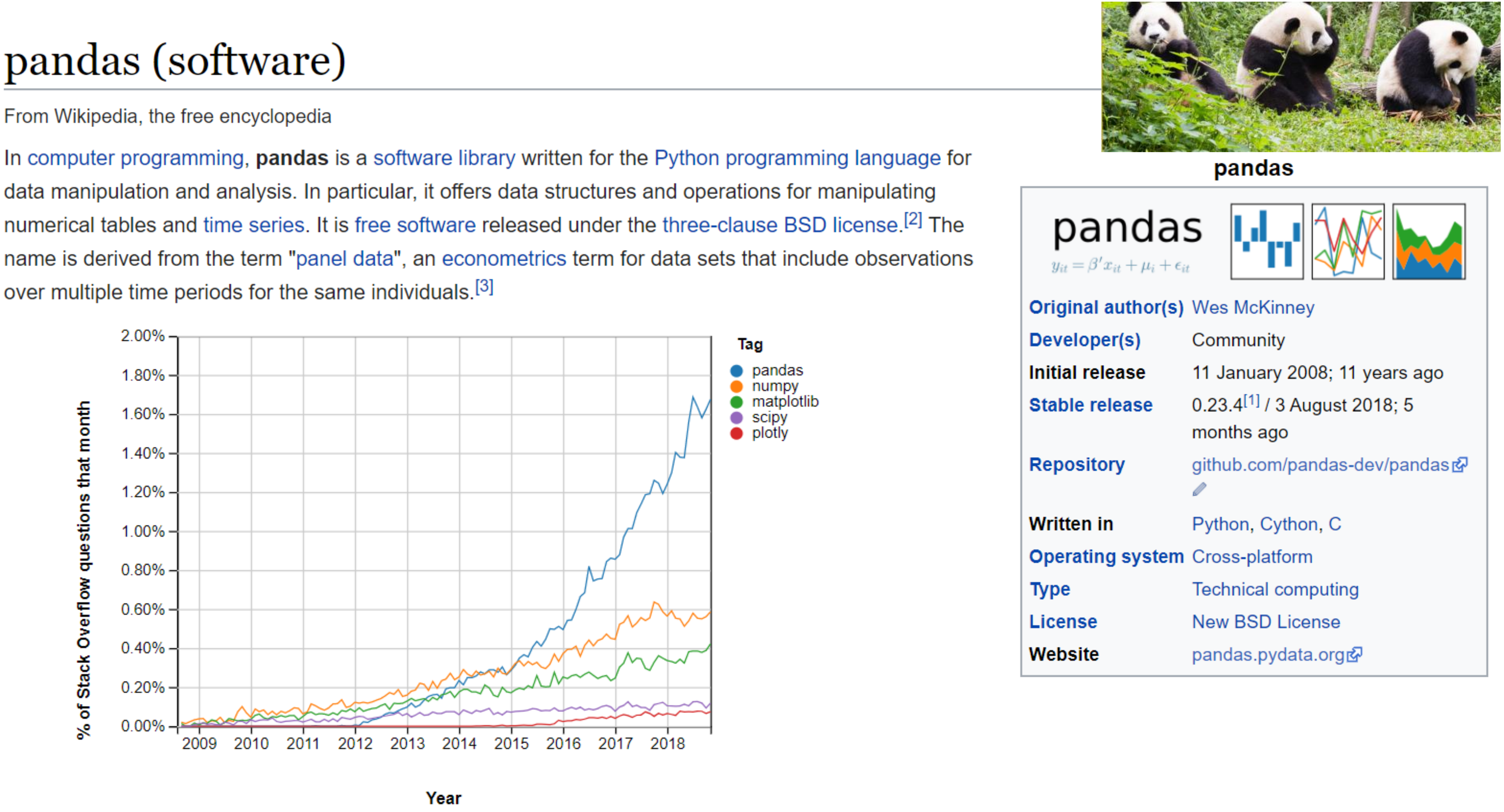


Resources: bit.ly/conclusion-notebook

.. en in een specifieke Python Library in het bijzonder



Resources: bit.ly/conclusion-notebook

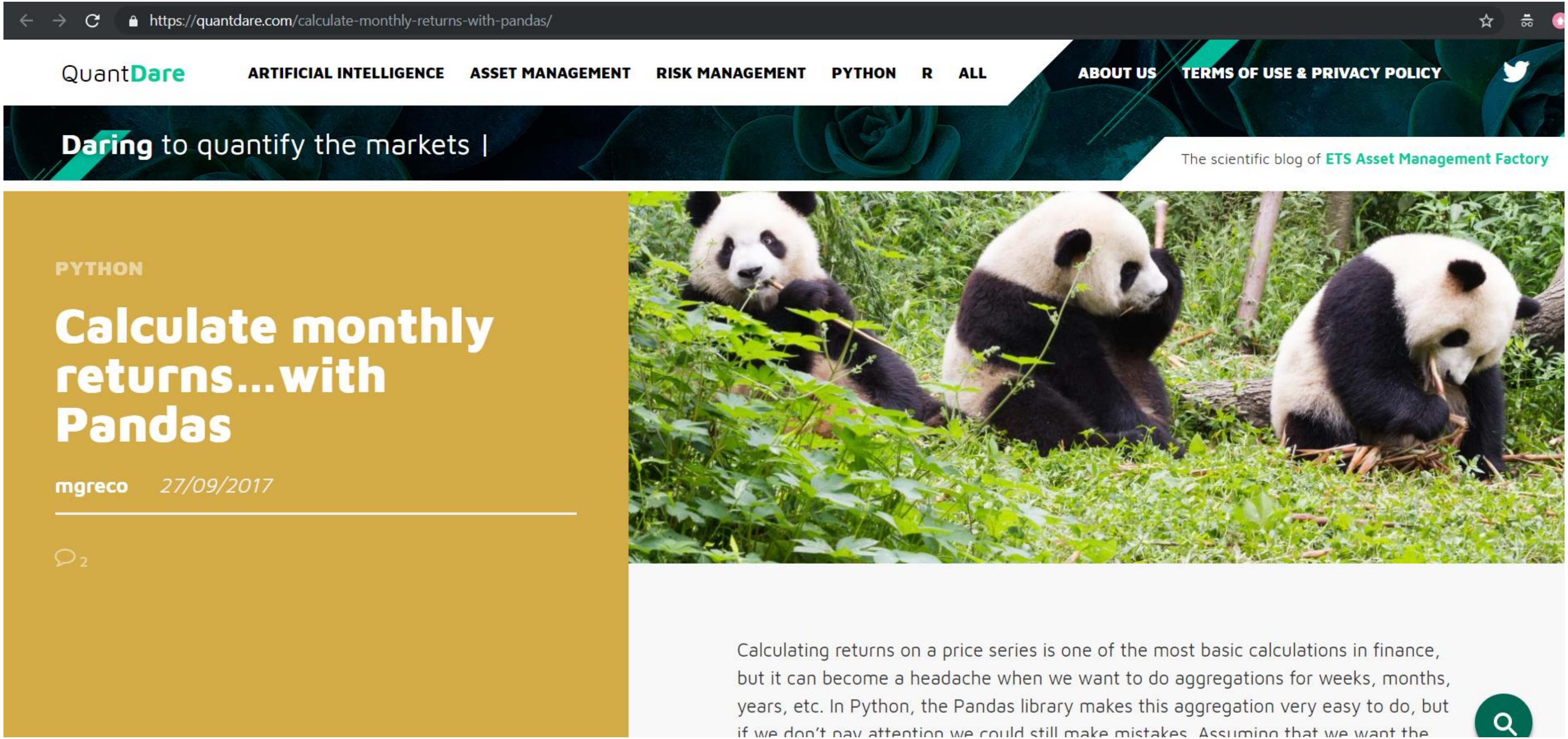


pandas (software)

From Wikipedia, the free encyclopedia

In computer programming, **pandas** is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.[2] The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals.[3]

Original author(s)	Wes McKinney
Developer(s)	Community
Initial release	11 January 2008; 11 years ago
Stable release	0.23.4[1] / 3 August 2018; 5 months ago
Repository	github.com/pandas-dev/pandas
Written in	Python, Cython, C
Operating system	Cross-platform
Type	Technical computing
License	New BSD License
Website	pandas.pydata.org


https://quantdare.com/calculate-monthly-returns-with-pandas/
QuantDare
ARTIFICIAL INTELLIGENCE
ASSET MANAGEMENT
RISK MANAGEMENT
PYTHON
R
ALL
ABOUT US
TERMS OF USE & PRIVACY POLICY

Daring to quantify the markets |

The scientific blog of **ETS Asset Management Factory**

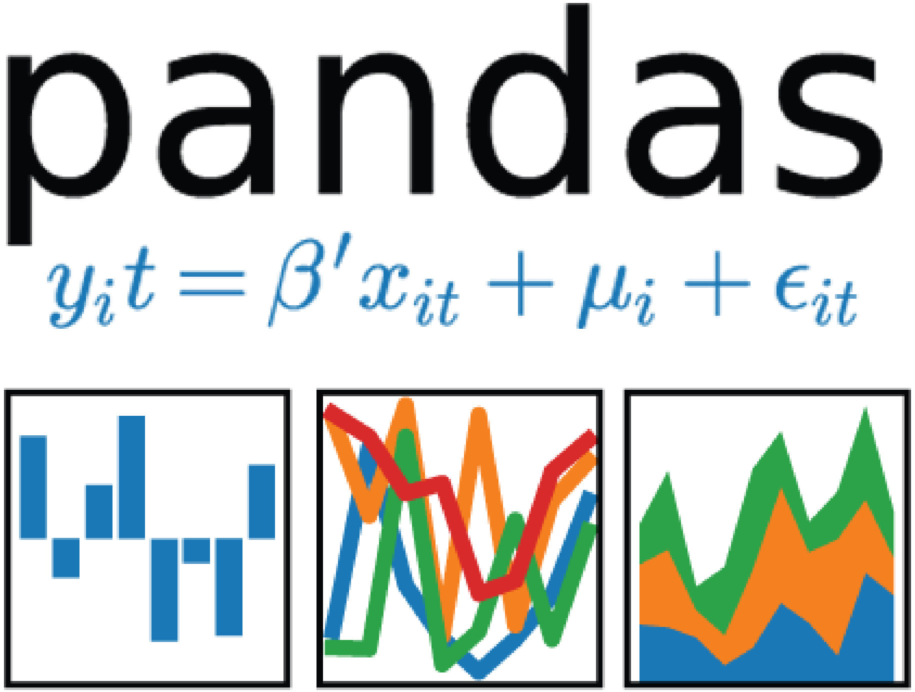
PYTHON

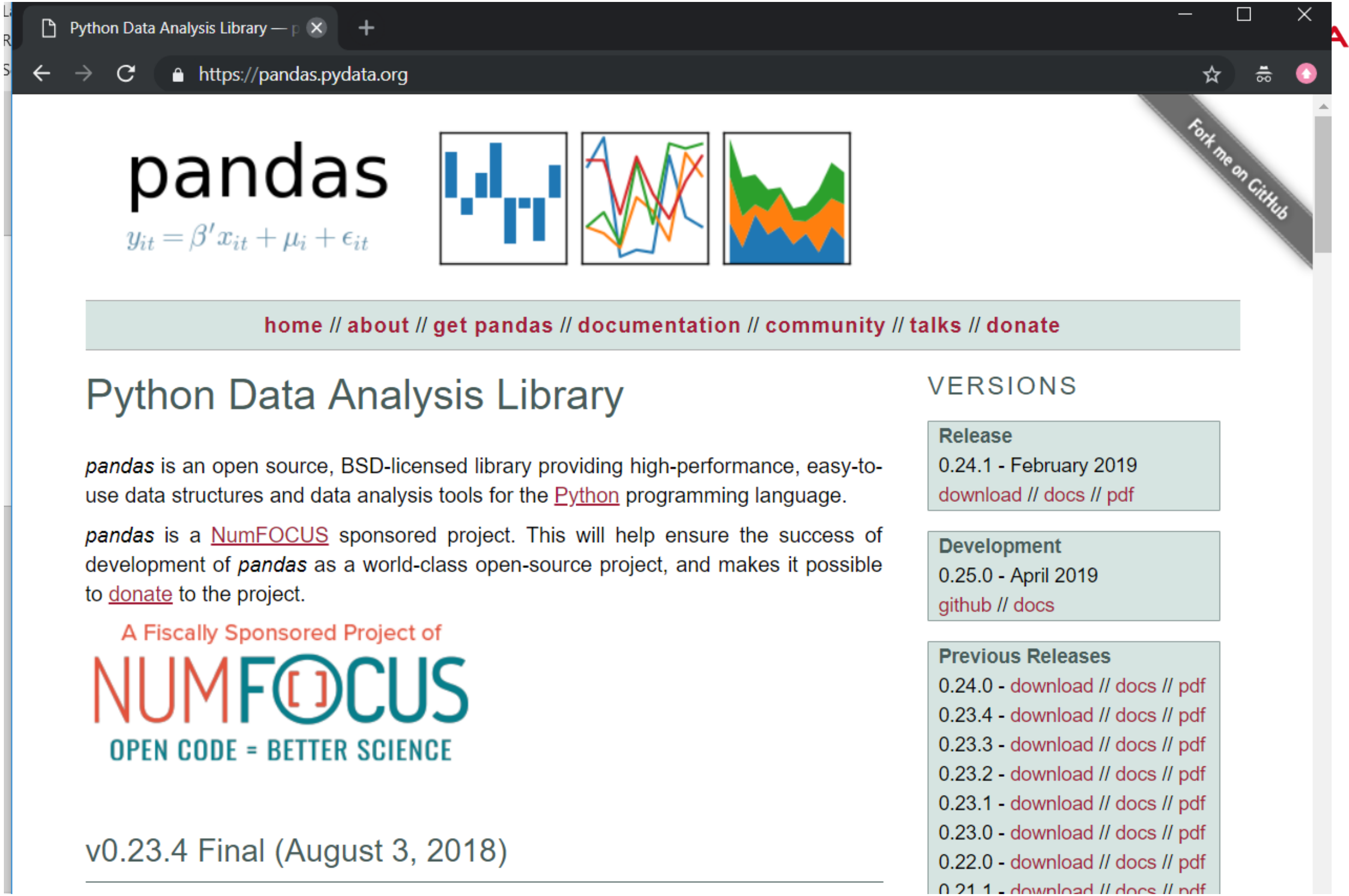
Calculate monthly returns…with Pandas

mgreco *27/09/2017*

Calculating returns on a price series is one of the most basic calculations in finance, but it can become a headache when we want to do aggregations for weeks, months, years, etc. In Python, the Pandas library makes this aggregation very easy to do, but if we don't pay attention we could still make mistakes. Assuming that we want the

Pandas


pandas
$y_it = \beta' x_{it} + \mu_i + \epsilon_{it}$



Python Data Analysis Library — p

https://pandas.pydata.org

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

home // about // get pandas // documentation // community // talks // donate

Python Data Analysis Library

pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

pandas is a NumFOCUS sponsored project. This will help ensure the success of development of *pandas* as a world-class open-source project, and makes it possible to donate to the project.

A Fiscally Sponsored Project of

NUMFOCUS

OPEN CODE = BETTER SCIENCE

v0.23.4 Final (August 3, 2018)

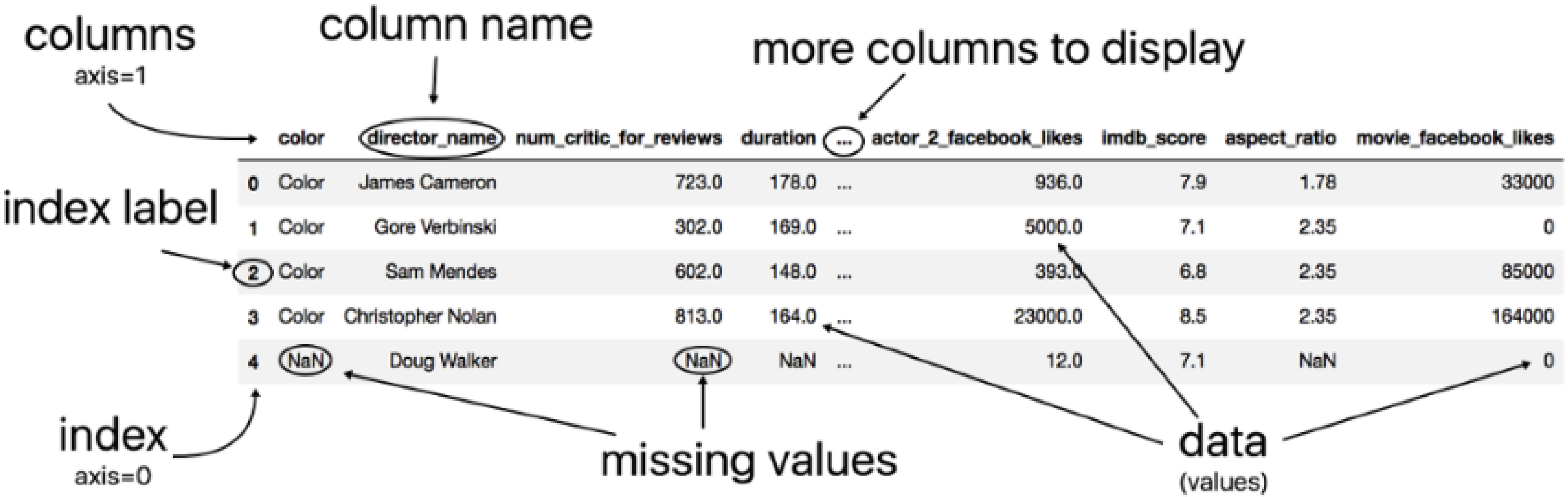
VERSIONS

Release
0.24.1 - February 2019
download // docs // pdf

Development
0.25.0 - April 2019
github // docs

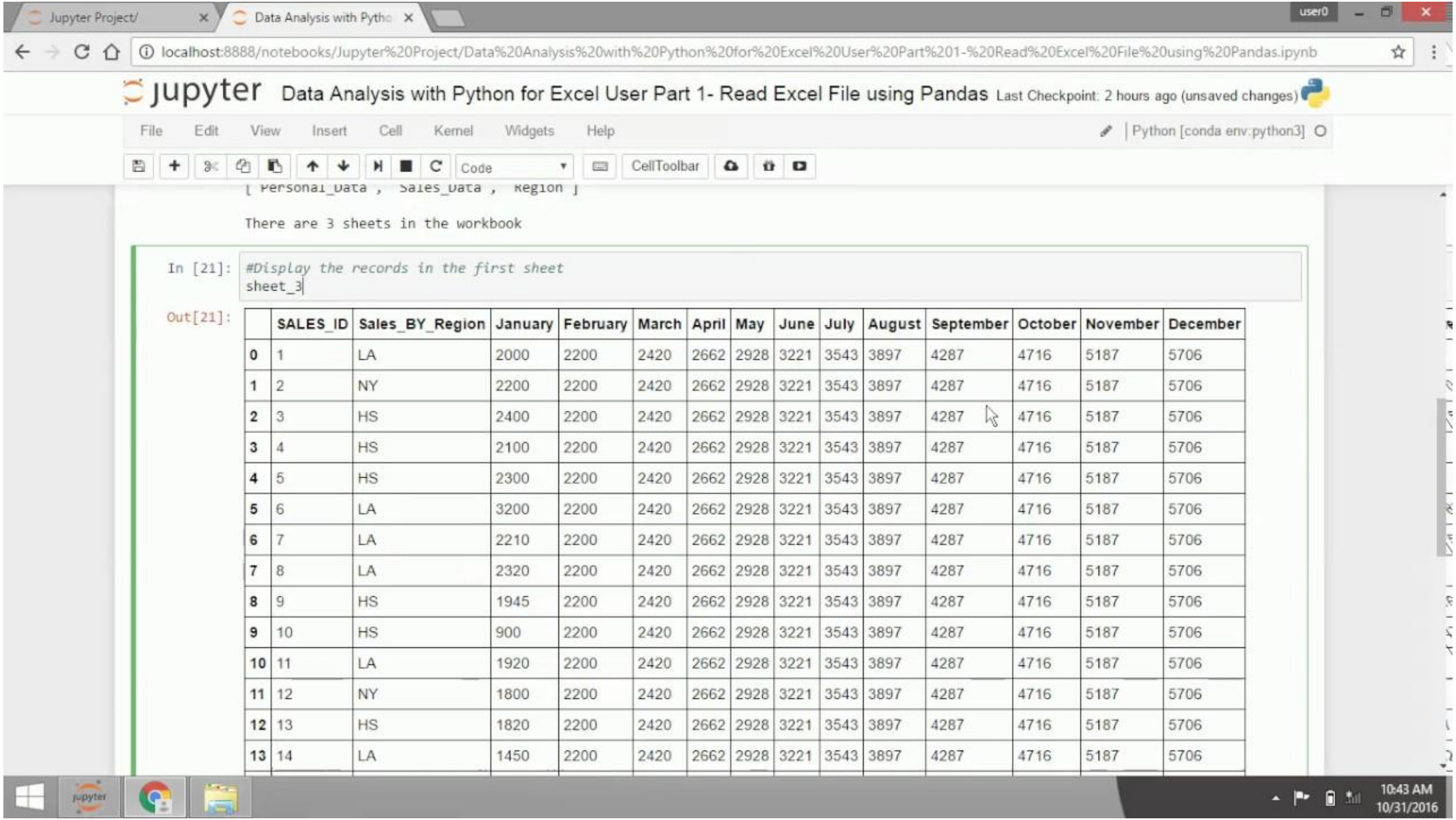
Previous Releases
0.24.0 - download // docs // pdf
0.23.4 - download // docs // pdf
0.23.3 - download // docs // pdf
0.23.2 - download // docs // pdf
0.23.1 - download // docs // pdf
0.23.0 - download // docs // pdf
0.22.0 - download // docs // pdf
0.21.1 - download // docs // pdf

Pandas = Panel Data



	color	director_name	num_critic_for_reviews	duration	...	actor_2_facebook_likes	imdb_score	aspect_ratio	movie_facebook_likes
0	Color	James Cameron	723.0	178.0	...	936.0	7.9	1.78	33000
1	Color	Gore Verbinski	302.0	169.0	...	5000.0	7.1	2.35	0
2	Color	Sam Mendes	602.0	148.0	...	393.0	6.8	2.35	85000
3	Color	Christopher Nolan	813.0	164.0	...	23000.0	8.5	2.35	164000
4	NaN	Doug Walker	NaN	NaN	...	12.0	7.1	NaN	0

Anatomy of a DataFrame


Jupyter Project/
Data Analysis with Pytho
user0
localhost:8888/notebooks/Jupyter%20Project/Data%20Analysis%20with%20Python%20for%20Excel%20User%20Part%201-%20Read%20Excel%20File%20using%20Pandas.ipynb
jupyter
Data Analysis with Python for Excel User Part 1- Read Excel File using Pandas
Last Checkpoint: 2 hours ago (unsaved changes)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Python [conda env:python3]
Code
CellToolbar
10:43 AM
10/31/2016

```
['Personal_Data', 'Sales_Data', 'Region']

There are 3 sheets in the workbook
```

In [21]:

```
#Display the records in the first sheet
sheet_3
```

Out[21]:

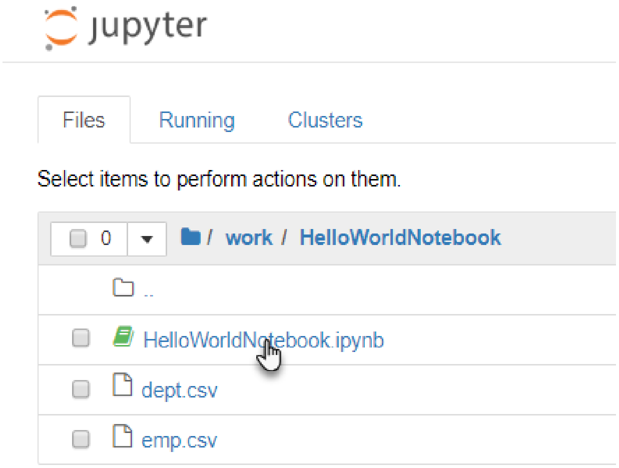
	SALES_ID	Sales_BY_Region	January	February	March	April	May	June	July	August	September	October	November	December
0	1	LA	2000	2200	2420	2662	2928	3221	3543	3897	4287	4716	5187	5706
1	2	NY	2200	2200	2420	2662	2928	3221	3543	3897	4287	4716	5187	5706
2	3	HS	2400	2200	2420	2662	2928	3221	3543	3897	4287	4716	5187	5706
3	4	HS	2100	2200	2420	2662	2928	3221	3543	3897	4287	4716	5187	5706
4	5	HS	2300	2200	2420	2662	2928	3221	3543	3897	4287	4716	5187	5706
5	6	LA	3200	2200	2420	2662	2928	3221	3543	3897	4287	4716	5187	5706
6	7	LA	2210	2200	2420	2662	2928	3221	3543	3897	4287	4716	5187	5706
7	8	LA	2320	2200	2420	2662	2928	3221	3543	3897	4287	4716	5187	5706
8	9	HS	1945	2200	2420	2662	2928	3221	3543	3897	4287	4716	5187	5706
9	10	HS	900	2200	2420	2662	2928	3221	3543	3897	4287	4716	5187	5706
10	11	LA	1920	2200	2420	2662	2928	3221	3543	3897	4287	4716	5187	5706
11	12	NY	1800	2200	2420	2662	2928	3221	3543	3897	4287	4716	5187	5706
12	13	HS	1820	2200	2420	2662	2928	3221	3543	3897	4287	4716	5187	5706
13	14	LA	1450	2200	2420	2662	2928	3221	3543	3897	4287	4716	5187	5706

Een paar operaties op een Pandas Data Frame

- *df.dtypes* – alle data types van alle kolommen
- *df['column_name'].value_counts* – telling van aantal verschillende waarden
- *df.head(5)* – toon de eerste vijf rijen van het data frame
  - df['column_name'].head(5) (of .tail(10))
  - df[['column_name', 'column_name_2']].head(5)
- *pd.crosstab(df['column_name'], df['column_name_2'])* – kruistabel
- *df['column_name3'] = 2 * df['column_name']* – voeg een kolom toe
  - df['column_name3'] = df['column_name'].apply(len) – bepaal waarde van nieuwe kolom door toepassen van een functie op een bestaande kolom
  - df['column_name3'] = df.apply(lambda row: 'Y' if row['column_name'] == 'HOT' else 'N') – bepaal waarde van nieuwe kolom op basis van conditie
- *df.drop('column_name3')* – verwijder kolom

Hello World Notebook

- Als je omgeving draait
  - Lokaal of Katacoda
- Open dan het Notebook
  HelloWorldNotebook.ipynb
  in folder /work/HelloWorldNotebook
- Stap door de cellen
  lees de instructie
  voer de opdrachten uit



Select items to perform actions on them.

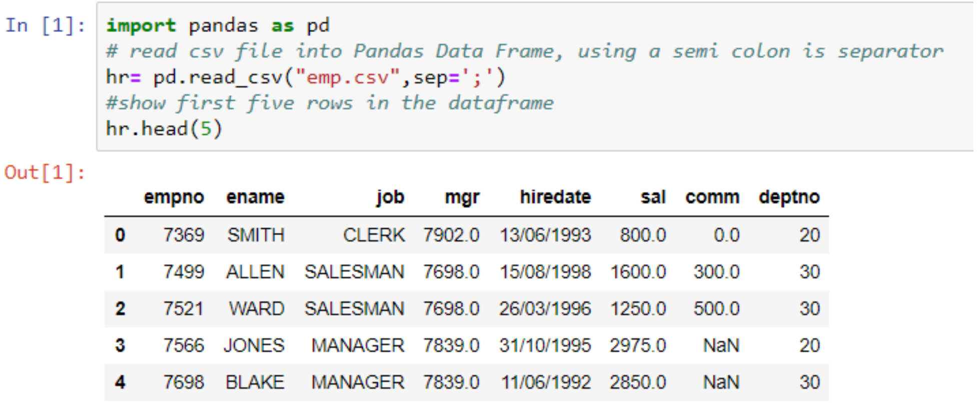
/ work / HelloWorldNotebook

..

HelloWorldNotebook.ipynb

dept.csv

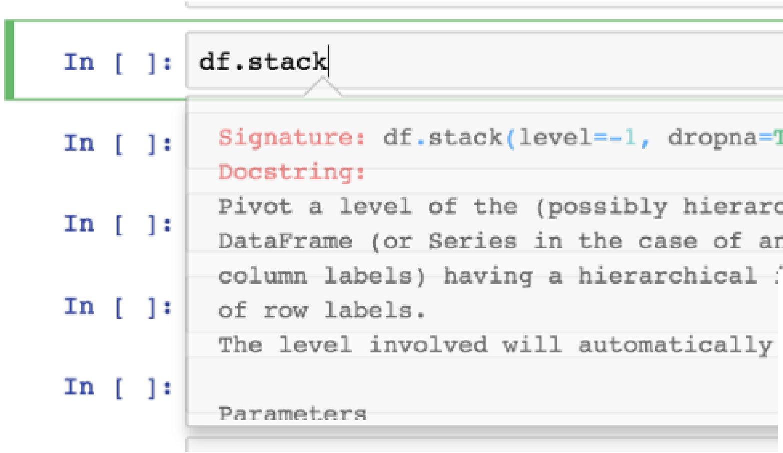
emp.csv



```
In [1]: import pandas as pd
        # read csv file into Pandas Data Frame, using a semi colon is separator
        hr= pd.read_csv("emp.csv",sep=';')
        #show first five rows in the dataframe
        hr.head(5)
```

```
Out[1]:
     empno  ename        job     mgr    hiredate     sal   comm  deptno
  0   7369  SMITH      CLERK  7902.0  13/06/1993   800.0    0.0      20
  1   7499  ALLEN   SALESMAN  7698.0  15/08/1998  1600.0  300.0      30
  2   7521   WARD   SALESMAN  7698.0  26/03/1996  1250.0  500.0      30
  3   7566  JONES    MANAGER  7839.0  31/10/1995  2975.0    NaN      20
  4   7698  BLAKE    MANAGER  7839.0  11/06/1992  2850.0    NaN      30
```

Press shift + tab + tab to get Help in a Jupyter Notebook



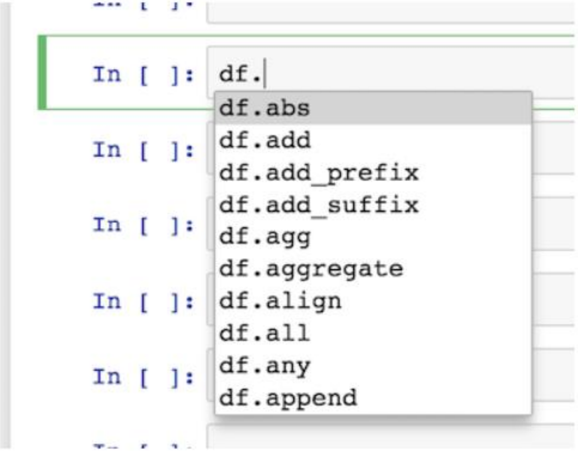
```
In [ ]: df.stack

Signature: df.stack(level=-1, dropna=True)
Docstring:
Pivot a level of the (possibly hierarchical) column labels, returning a
DataFrame (or Series in the case of an object with a single level of
column labels) having a hierarchical i
of row labels.
The level involved will automatically

Parameters
```

Pressing shift + tab + tab to reveal the docun

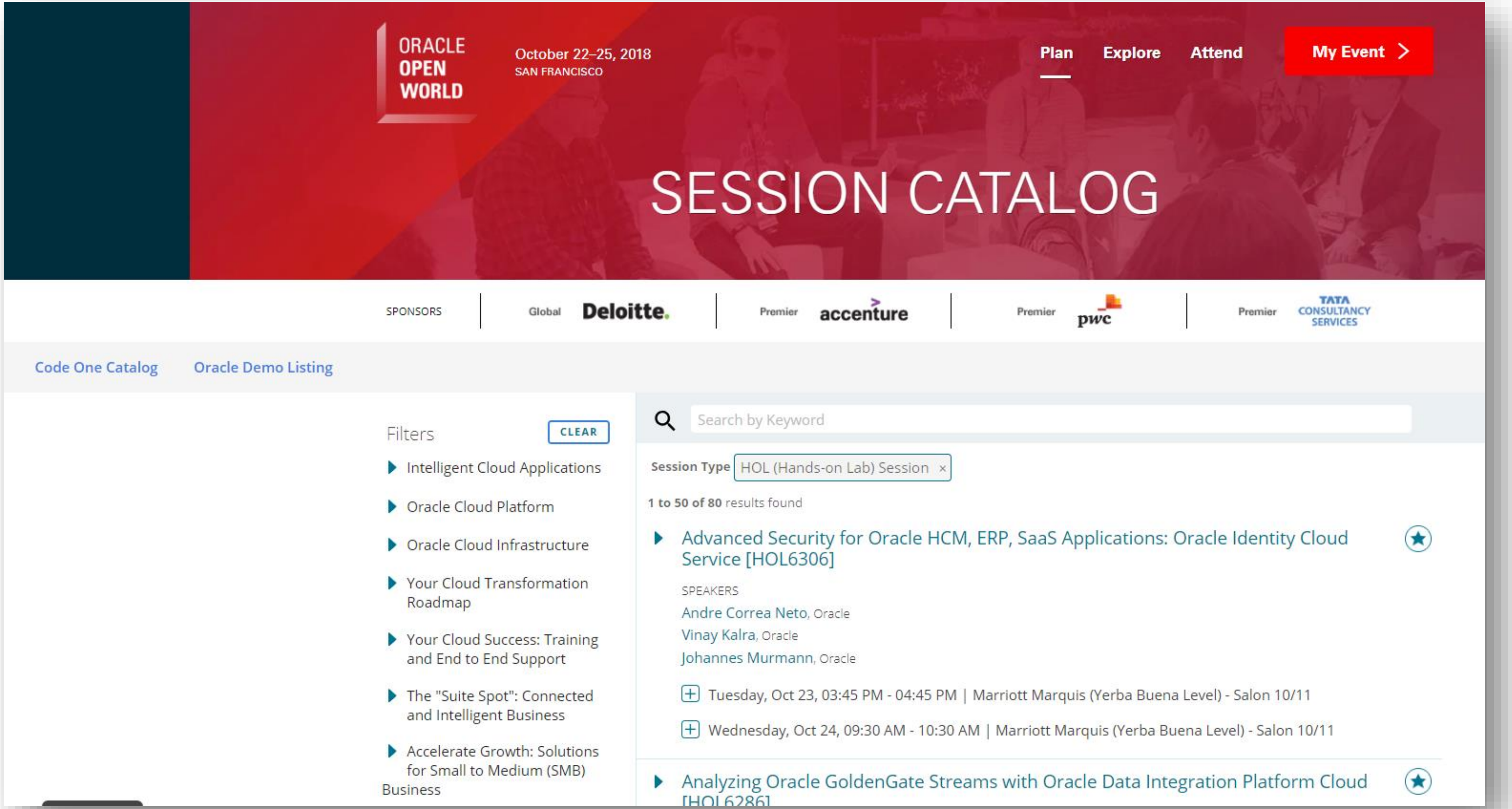
You can also press `tab` directly following a dot to have a dropdown menu



```
In [ ]: df.
```

Pressing tab following a DataFrame lists the 200+ available objects

Data Analytics on Oracle OpenWorld 2018 Session Details



Filters

- Intelligent Cloud Applications
- Oracle Cloud Platform
- Oracle Cloud Infrastructure
- Your Cloud Transformation Roadmap
- Your Cloud Success: Training and End to End Support
- The "Suite Spot": Connected and Intelligent Business
- Accelerate Growth: Solutions for Small to Medium (SMB) Business

Session Type: HOL (Hands-on Lab) Session ×

1 to 50 of 80 results found

Advanced Security for Oracle HCM, ERP, SaaS Applications: Oracle Identity Cloud Service [HOL6306]

SPEAKERS
Andre Correa Neto, Oracle
Vinay Kalra, Oracle
Johannes Murmann, Oracle

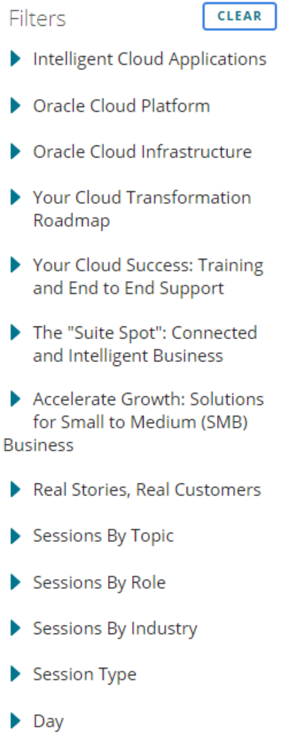
Tuesday, Oct 23, 03:45 PM - 04:45 PM | Marriott Marquis (Yerba Buena Level) - Salon 10/11

Wednesday, Oct 24, 09:30 AM - 10:30 AM | Marriott Marquis (Yerba Buena Level) - Salon 10/11

Analyzing Oracle GoldenGate Streams with Oracle Data Integration Platform Cloud [HOL6286]

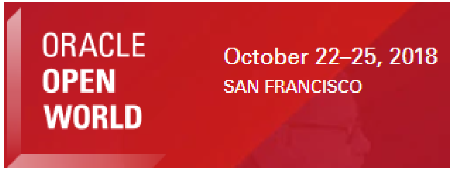
High level overview

- Two events: Oracle OpenWorld 2018 and Oracle CodeOne 2018
- Over 2000 sessions
- Over 2500 speakers
- Various dimensions:



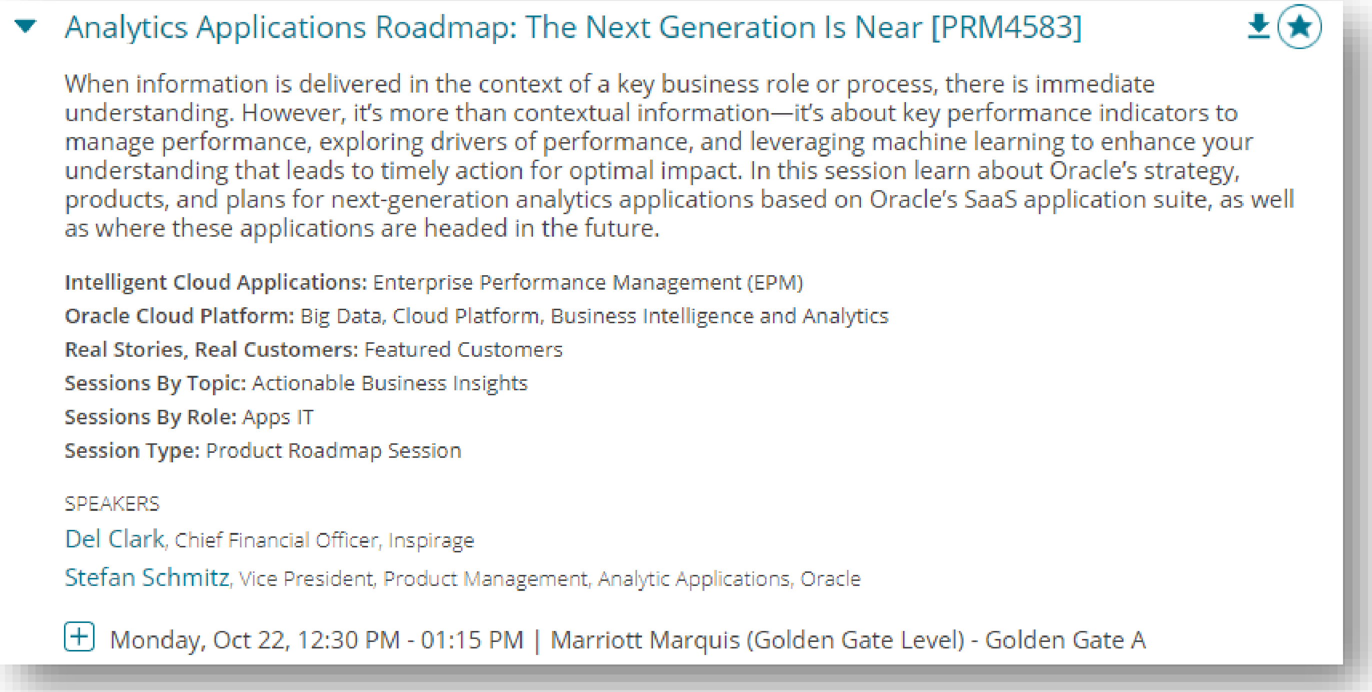
Filters CLEAR

- Intelligent Cloud Applications
- Oracle Cloud Platform
- Oracle Cloud Infrastructure
- Your Cloud Transformation Roadmap
- Your Cloud Success: Training and End to End Support
- The "Suite Spot": Connected and Intelligent Business
- Accelerate Growth: Solutions for Small to Medium (SMB) Business
- Real Stories, Real Customers
- Sessions By Topic
- Sessions By Role
- Sessions By Industry
- Session Type
- Day



High level overview

- Details per session:



Analytics Applications Roadmap: The Next Generation Is Near [PRM4583]

When information is delivered in the context of a key business role or process, there is immediate understanding. However, it's more than contextual information—it's about key performance indicators to manage performance, exploring drivers of performance, and leveraging machine learning to enhance your understanding that leads to timely action for optimal impact. In this session learn about Oracle's strategy, products, and plans for next-generation analytics applications based on Oracle's SaaS application suite, as well as where these applications are headed in the future.

Intelligent Cloud Applications: Enterprise Performance Management (EPM)
Oracle Cloud Platform: Big Data, Cloud Platform, Business Intelligence and Analytics
Real Stories, Real Customers: Featured Customers
Sessions By Topic: Actionable Business Insights
Sessions By Role: Apps IT
Session Type: Product Roadmap Session

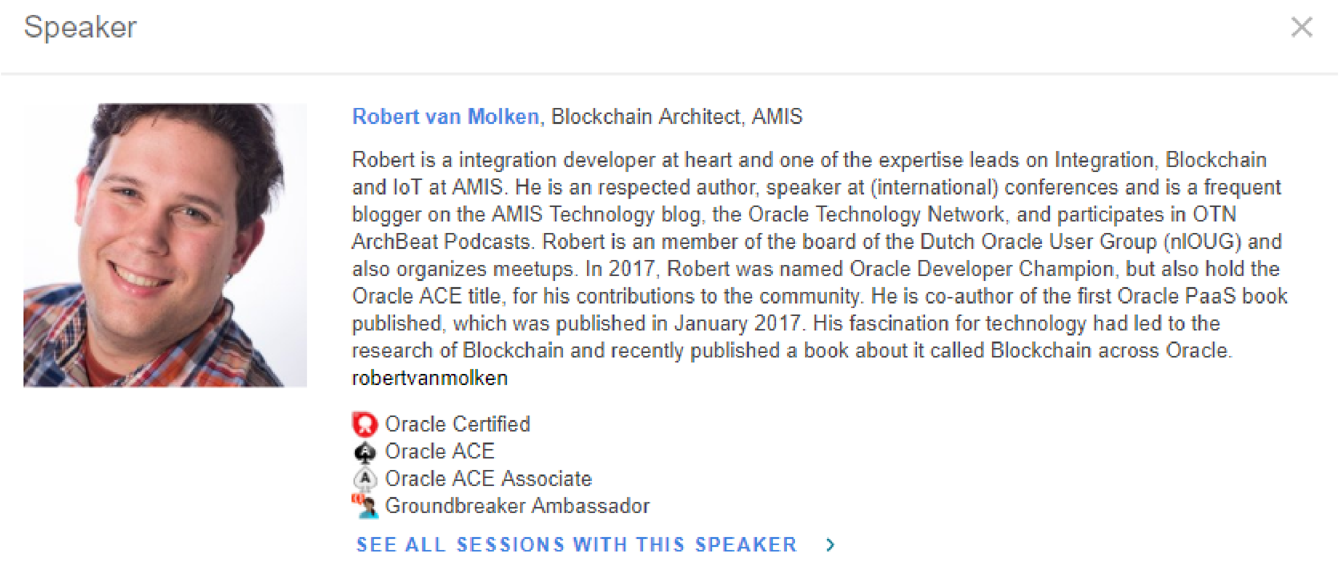
SPEAKERS

Del Clark, Chief Financial Officer, Inspirage
Stefan Schmitz, Vice President, Product Management, Analytic Applications, Oracle

⊞ Monday, Oct 22, 12:30 PM - 01:15 PM | Marriott Marquis (Golden Gate Level) - Golden Gate A

High level overview

- Details per speaker:



Speaker

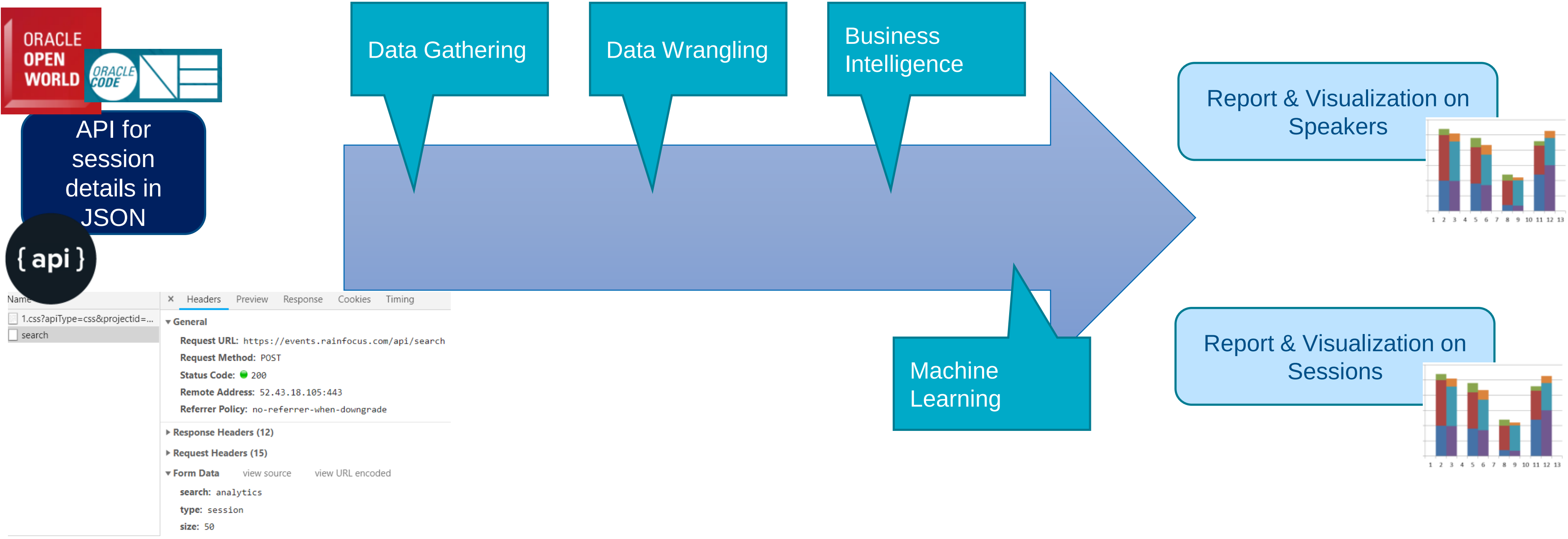
Robert van Molken, Blockchain Architect, AMIS

Robert is a integration developer at heart and one of the expertise leads on Integration, Blockchain and IoT at AMIS. He is an respected author, speaker at (international) conferences and is a frequent blogger on the AMIS Technology blog, the Oracle Technology Network, and participates in OTN ArchBeat Podcasts. Robert is an member of the board of the Dutch Oracle User Group (nlOUG) and also organizes meetups. In 2017, Robert was named Oracle Developer Champion, but also hold the Oracle ACE title, for his contributions to the community. He is co-author of the first Oracle PaaS book published, which was published in January 2017. His fascination for technology had led to the research of Blockchain and recently published a book about it called Blockchain across Oracle.

robertvanmolken

- Oracle Certified
- Oracle ACE
- Oracle ACE Associate
- Groundbreaker Ambassador

Flow for Oracle OpenWorld 2018 Session Data Analytics



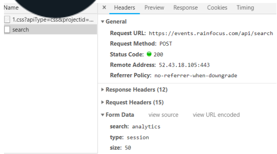
```
Name                              ×  Headers  Preview  Response  Cookies  Timing
1.css?apiType=css&projectid=...   ▼ General
search                               Request URL: https://events.rainfocus.com/api/search
                                     Request Method: POST
                                     Status Code: ● 200
                                     Remote Address: 52.43.18.105:443
                                     Referrer Policy: no-referrer-when-downgrade
                                  ▶ Response Headers (12)
                                  ▶ Request Headers (15)
                                  ▼ Form Data     view source     view URL encoded
                                     search: analytics
                                     type: session
                                     size: 50
```

Flow for Oracle OpenWorld 2018 Session Data Analytics

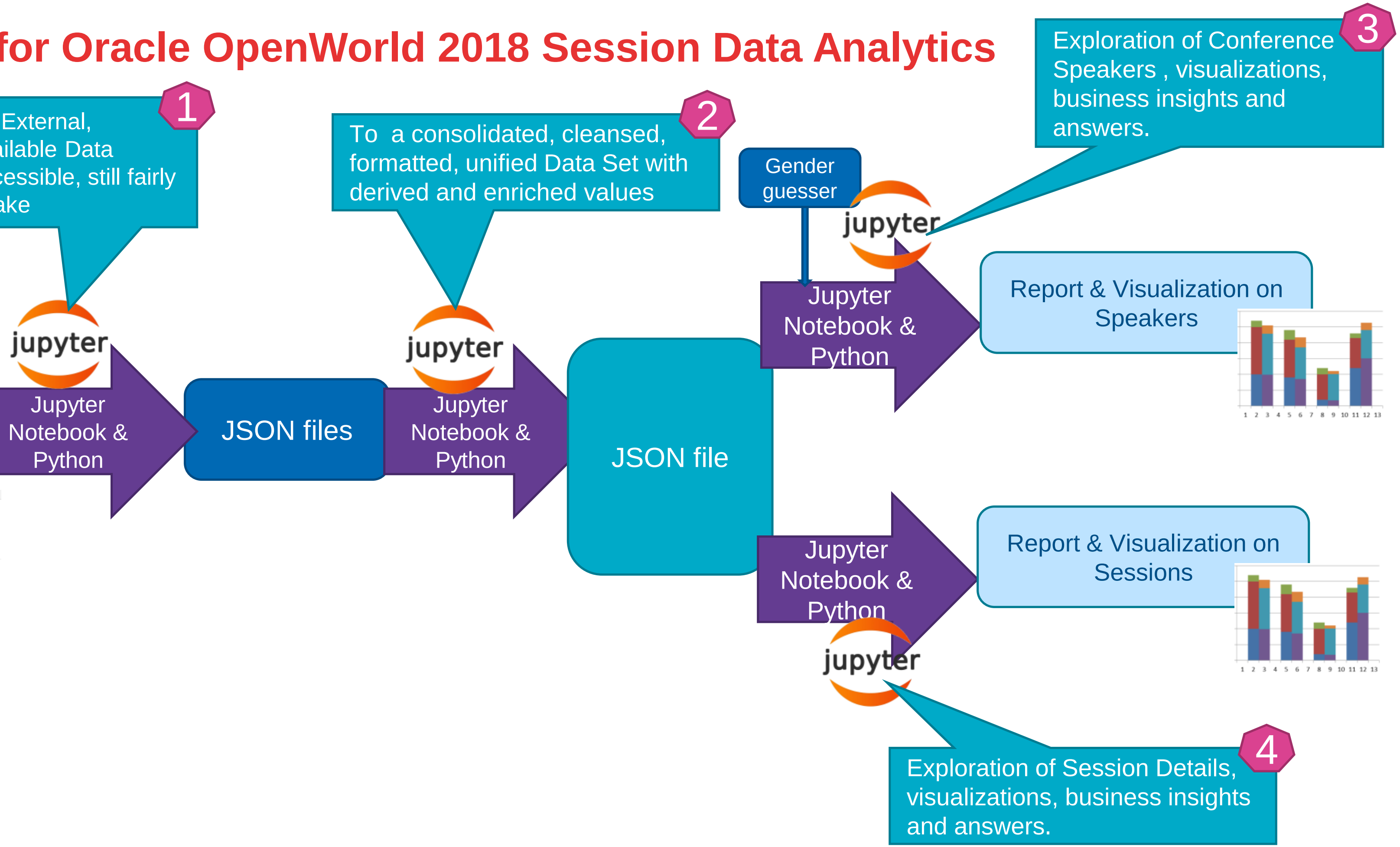
ORACLE OPEN WORLD

ORACLE CODE

API for session details in JSON



```
▾ General
    Request URL: https://events.rainfocus.com/api/search
    Request Method: POST
    Status Code: ● 200
    Remote Address: 52.43.18.105:443
    Referrer Policy: no-referrer-when-downgrade
▸ Response Headers (12)
▸ Request Headers (15)
▾ Form Data    view source    view URL encoded
    search: analytics
    type: session
    size: 50
```



1 From Raw, External, Slow/Unavailable Data to local, accessible, still fairly raw Data Lake

Jupyter Notebook & Python

JSON files

2 To a consolidated, cleansed, formatted, unified Data Set with derived and enriched values

Jupyter Notebook & Python

JSON file

Gender guesser

Jupyter Notebook & Python

3 Exploration of Conference Speakers , visualizations, business insights and answers.

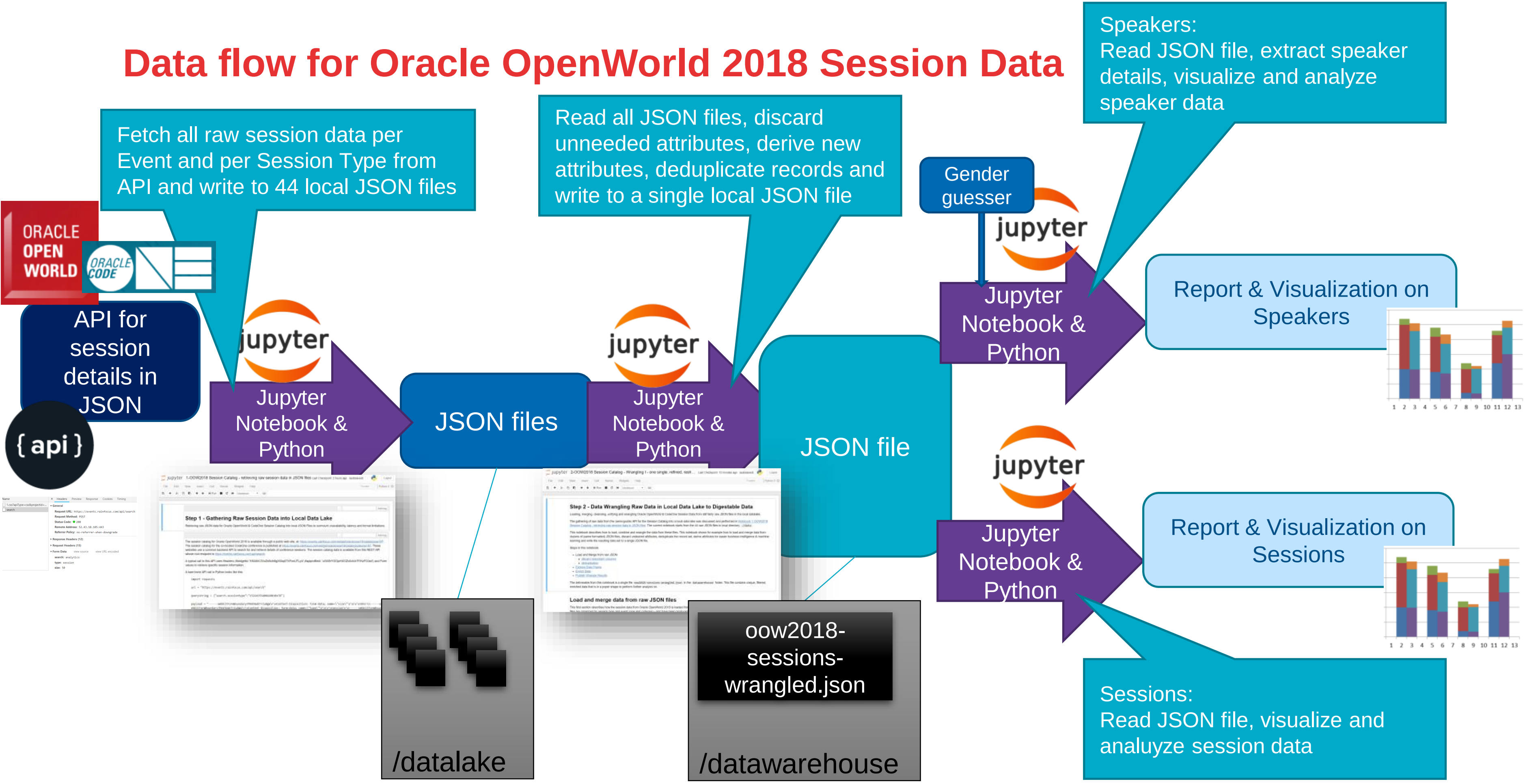
Report & Visualization on Speakers

Jupyter Notebook & Python

Report & Visualization on Sessions

4 Exploration of Session Details, visualizations, business insights and answers.

Data flow for Oracle OpenWorld 2018 Session Data



Fetch all raw session data per Event and per Session Type from API and write to 44 local JSON files

Read all JSON files, discard unneeded attributes, derive new attributes, deduplicate records and write to a single local JSON file

Speakers:
Read JSON file, extract speaker details, visualize and analyze speaker data

Sessions:
Read JSON file, visualize and analuyze session data

ORACLE OPEN WORLD
ORACLE CODE
API for session details in JSON
{ api }
jupyter
Jupyter Notebook & Python
JSON files
jupyter
Jupyter Notebook & Python
JSON file
Gender guesser
jupyter
Jupyter Notebook & Python
jupyter
Jupyter Notebook & Python
Report & Visualization on Speakers
Report & Visualization on Sessions
1 2 3 4 5 6 7 8 9 10 11 12 13
/datalake
oow2018-sessions-wrangled.json
/datawarehouse

jupyter 1-OOW2018 Session Catalog - retrieving raw session data in JSON files

Step 1 - Gathering Raw Session Data into Local Data Lake

jupyter 2-OOW2018 Session Catalog - Wrangling I - one single, refined, result...

Step 2 - Data Wrangling Raw Data in Local Data Lake to Digestable Data

Load and merge data from raw JSON files