The most powerful open source data science technologies in your browser.

Yves Hilpisch
I. The Market and The Problem
   II. How We Solve The Problem
   III. Market Size and Facts
   IV. Strategic Opportunities
Mega Trends

**Mega trends that influence data science**

- **Github**: Social Coding
  - Today’s standard is “open source”, even for key technologies.

- **Quandl**: 
  - More and more data sets are “open and free”.

- **IP[y]**: IPython Interactive Computing
  - Complex analytics work flows are coded in the browser.

- **Meetup**: 
  - Dynamic communities shape the way knowledge is transmitted.

- **DigitalOcean**: 
  - Infrastructure is a standardized commodity, billed by the hour.

- **DropBox**: 
  - Individuals and institutions store more and more data in the cloud.
Open Source Software Revolution
OSS revolutionizes data science both in the front & back end

<table>
<thead>
<tr>
<th>FRONT END</th>
<th>BACK END</th>
</tr>
</thead>
<tbody>
<tr>
<td>In the front end, OSS revolutionizes how data scientists and developers work on a daily basis.</td>
<td>In the back end, OSS revolutionizes how analytics workflows and data applications are deployed and scaled.</td>
</tr>
</tbody>
</table>

```
“DigitalOcean is a simple and fast cloud hosting provider built for developers. Customers can create a cloud server in 55 seconds, and pricing plans start at only $5 per month for 512MB of RAM, 20GB SSD, 1 CPU, and 1TB Transfer.”
```
The Problem
Obstacles to using OSS for data science

- **Open Source**
  - fast changing environment

- **Vendors & Partners**
  - almost no vendors that provide help & support

- **Libraries**
  - huge amount of libraries to manage

- **Tools**
  - multitude of useful standalone tools

- **Deployment**
  - complex, lengthy, costly, risky

- **Maintenance**
  - how to update, maintain infrastructure?

- **Diverse End Users**
  - computer & data scientists as well as domain experts

- **Training**
  - how to train and re-train people?

- **Start**
  - where and how to start, who to talk to?
I. The Market and The Problem

II. How We Solve The Problem

III. Market Size and Facts

IV. Strategic Opportunities
Open source data science technologies in your browser

Tools and technologies data scientists know and love.
Browser-based Data Science
datapark capitalizes on new Web technologies and tools

1. **Generation: Move Data Around** — data analytics started by moving data from one place to another, analyzing it locally and moving results back to the remote data source

2. **Generation: Move Code Around** — moving tons of data is costly and time consuming; moving small code sets is faster and less costly

3. **Generation: Don't Move Anything** — the Browser and Web technologies allow to work directly and in real-time on the infrastructure where data and code are stored (replacing e.g. remote ssh access)
Feature Rich datapark is essentially a data scientist’s wish list

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Browser</td>
<td>you only need your browser to use the full fledged Data Science environment</td>
</tr>
<tr>
<td>Analyze</td>
<td>interactive notebooks for explorative data analytics with e.g. Python, R, Julia</td>
</tr>
<tr>
<td>Visualize</td>
<td>easily visualize your data – both statically and interactively (D3.js)</td>
</tr>
<tr>
<td>Edit &amp; Develop</td>
<td>edit all typical code files within the browser (e.g. Python, HTML, CSS)</td>
</tr>
<tr>
<td>Your Data</td>
<td>easily upload, download and work with your data, files, etc.</td>
</tr>
<tr>
<td>Integrate</td>
<td>integrate with your code and data sources, like Github, Google Drive or Dropbox</td>
</tr>
<tr>
<td>Collaborate</td>
<td>define projects, collaborate within your team and with others on datapark.io</td>
</tr>
<tr>
<td>Share &amp; Publish</td>
<td>share &amp; publish your documents &amp; files, deploy your Web applications</td>
</tr>
<tr>
<td>Get Social</td>
<td>let others know what you have been working on</td>
</tr>
</tbody>
</table>
The Platform
Bringing the best of Open Source together in the browser

“Absorb what is useful, discard what is not, and add what is uniquely your own.”
—Bruce Lee
Natural Evolution
From Python for Finance to Open Source for Data Analytics
I. The Market and The Problem
II. How We Solve The Problem
III. Market Size and Facts
IV. Strategic Opportunities
Data Scientists and Engineers
There are about 10mn people in technical computing

Source: diverse Web resources; in mn people
Data Analytics

Data analytics is a top priority of almost any organisation.

“Companies will spend an average of $7.4M on data-related initiatives over the next twelve months, with enterprises investing $13.8M, and small & medium businesses (SMBs) investing $1.6M.

80% of enterprises and 63% of small & medium businesses (SMBs) already have deployed or are planning to deploy big data projects in the next twelve months.

83% of organizations are prioritizing structured data initiatives as critical or high priority in 2015, and 36% planning to increase their budgets for data-driven initiatives in 2015.”

Source: http://www.forbes.com
Open Source Data Science
OS languages dominate data science these days

Data Science Languages

Open Source Data Science

R, Python and SQL dominate OS data science

Poll data from August 2014; usage in %. Source: http://www.kdnuggets.com
Vendor Criteria in Data Analytics

Integration, security, ease of use & scalability important

Source: 2015 Big Data Analytics Survey (Summary Slides)
Platform Competitors ...

... trying to solve the platform problem for data scientists

- **sense.io**: Proprietary Notebook solution, closed platform.
- **modeanalytics.com**: SQL focus, closed platform.
- **wakari.io**: Python focus, cloud version not maintained.
Major Competitor

The major competitor is Jupyter deployed in the cloud.

DigitalOcean droplet for 5 USD p.m., Jupyter with Python 3.4, deployed via Docker for 20+ users.

The MVP: http://jupyter.quant-platform.com

How to maintain, how to ensure security, how to share, how to control?
I. The Market and The Problem
II. How We Solve The Problem
III. Market Size and Facts
IV. Strategic Opportunities
Use Cases for datapark.io
From teaching to data science to a social app store

- Teaching Programming & Data Science
- Data Science Platform in Institutions and Corporations
- Analytics-as-a-Service for Open Source Projects and Proprietary Data and Code
- Market Place for Ideas, Projects, Apps etc. ("Social Data Science")
Establishing a Standard
Building critical mass & social components, improve scalability

- Goal of 100,000 users to learn from
- Building out Social Components
- Improving Deployment and Scalability
- Becoming the Github for Data Science
How do we want to reach our goals
Making usage as simple as possible based on standards

Sign-up
Two fields only — 30 seconds, immediate full fledged functionality

Infrastructure
Well established components — Ubuntu, Anaconda, Docker, ...

Tools
All that you know & love — IPython Notebook, ACE, Shell (Git, Vim), ...

Open
Using standards only — IPYNBs, Linux FS, Dropbox, Drive (easy in/out)
Just try it.

http://datapark.io

Give us feedback.

team@datapark.io
Dr. Yves J. Hilpisch

datapark.io | team@datapark.io | @dataparkio

The Python Quants GmbH