Data Science lives in the Cloud

SPEAKING

CONSTANTINO CASADO - Data Engineering Principal
Constantino Casado

A Journey:  
Science & Data → Data Science

PhD in Physics, IESE PDD, IT & industry analytics and material science.

Data Principal @ CAPSiDE

Certifications
- EMC-DataScience
- Azure Cloud
- Google Cloud Data Engineering
CAPSiDE DNA Keys

- Automation (Smart Infra)
- Open Tools
- Reliability
- Public Cloud
- Agility & DevOps
- Infra & service as code
Our Journey today

The Global Picture

The Local Picture

The Practical Approach
Is strong AI inevitable?
Evolution or Revolution
What makes a Company?
Business building blocks

- People
- Tools
- Processes
Data Driven company

Data → Discover → Model → Evaluate → Operate
BUSINESS VALUE
Business building blocks

Data <-> People

Tools <-> Processes
The challenges

1. The data headline
2. Hybrid knowledge (Blend data and expertise)
3. Data is about *Discovery*
Make data available + train users

\[
\text{Rydberg constant} \quad E_n = R_0 \frac{1}{n^2} \quad \text{(atomic units)}
\]

\[
\text{Transmutation rates} \quad N_{1+1} = N_1 e^{-\lambda t}
\]

\[
\text{Mean life time} \quad \tau = \frac{1}{\lambda} = \frac{1}{\frac{dN}{dt}}
\]

\[
\text{Population spectrum} \quad P = e^{-\lambda t} \frac{dN}{dt}
\]

\[
\text{Nuclear Physics} \quad \text{Stability against } Z \text{- decay}
\]

\[
\text{Bragg-Gray mass spectrometer} \quad q = \frac{B}{m} \quad \text{m} = 8.55 \times 10^{-6}
\]

\[
\text{Energy scaling } E_0 = T_0 \left(1 + \frac{M}{M_0}\right)
\]

\[
\text{Cherednyk time (nuclear)} \quad Q = M \left(2 - Q/A - Q/A\right)
\]

\[
\text{Decay energy of nucleus} \quad M < Z m + N m
\]

\[
\text{β-decay} \quad n \rightarrow p + e^- + \nu
\]
The anti-hybrid Pattern

Data

Business know-how

Tools

People

Proceses

Business know-how
The data value
Why did you come to that conclusión my dear slide rule?
The ghost in the shell

Exceptions
High dimensional problems
Abstract modeling
Emotions and creative
Blind spot perception
Sparse data
Biased by emotions

NLP
Vision
Medicine
Car Drive

Automated standard process
Limited dimensionality
Perception Biased
Limited learning abilities
Non creative
Massive calculation capability
Business building blocks

Data

People

Tools

Processes

KNOWLEDGE
The ghost in the machine
Advanced Chess

“perfect tactical play and highly meaningful strategic plans”

DEEP THINKING by Garry Kasparov
Self Service Machine Learning?

**CLOUD AUTOML VISION**

Upload and label images  |  Train your model  |  Evaluate

- Handbag
- Shoes
- Hat

Cloud AutoML Vision
DATA + OPS = DataOps

TPS = Quality is not a step at the end of the process

DEVOPS = Operation of software is not an after thought

DataOps = Data Insights are not an after thought
Data and Cloud technology?
What does it take?
DATA
INDUSTRIALIZATION
Frequent Story
SW Development

LOW uncertainty

Engineering = Build
DS Development

HIGH uncertainty

Research = More experimentation
Experimentation is THE opportunity
Cheap experimentation is key
Trial and error
In the End
ON PREMISES

- Invest
- Almost a week to configure
- Rigid fixed scale
- Too hard to try for fun
• Pay per use
• 5’ up and running
• Scalable
• Easy to turn on, try and off again

+ ALL KIND OF TOOLS
<table>
<thead>
<tr>
<th></th>
<th>Local</th>
<th>Google</th>
<th>AWS</th>
<th>Azure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Cleaning</td>
<td>Trifacta Wrangler</td>
<td>DataPrep</td>
<td>Glue</td>
<td>MLStudio</td>
</tr>
<tr>
<td>SandBox/Notebook</td>
<td>Jupiter, ...</td>
<td>DataLab</td>
<td>JupyterHub</td>
<td>MLStudio</td>
</tr>
<tr>
<td>Hadoop grid</td>
<td>Cloudera / HortonWorks ...</td>
<td>DataProc</td>
<td>EMR</td>
<td>HDInsights</td>
</tr>
<tr>
<td>Ingest</td>
<td>Kafka</td>
<td>Pub/Sub</td>
<td>Kinesis</td>
<td>Event Hub</td>
</tr>
<tr>
<td>Stream Processing</td>
<td>Apache Flume/ETL</td>
<td>DataFlow</td>
<td>Data Pipeline</td>
<td>Data Factory</td>
</tr>
<tr>
<td>Neural Networks</td>
<td>TensorFlow</td>
<td>Machine Learning Engine + APIs + AutoML</td>
<td>TensorFlow on AWS</td>
<td>MLStudio</td>
</tr>
<tr>
<td>DWh</td>
<td>Teradata Sybase IQ</td>
<td>BigQuery</td>
<td>Redshift</td>
<td>AzureDB DWH</td>
</tr>
<tr>
<td>Others</td>
<td>MongoDB, Cassandra, ...</td>
<td>Bigtable, Spanner</td>
<td>Athena, S3, ...</td>
<td>CosmosDB</td>
</tr>
</tbody>
</table>
INDUSTRIALIZATION
Takeaways

Velocity  Elasticity  Tools

Cheap experimentation  Industrialization
DATA SCIENCE ♥ CLOUD
Thank you!