DPTO
Dobre Praktyki
Tworzenia Oprogramowania

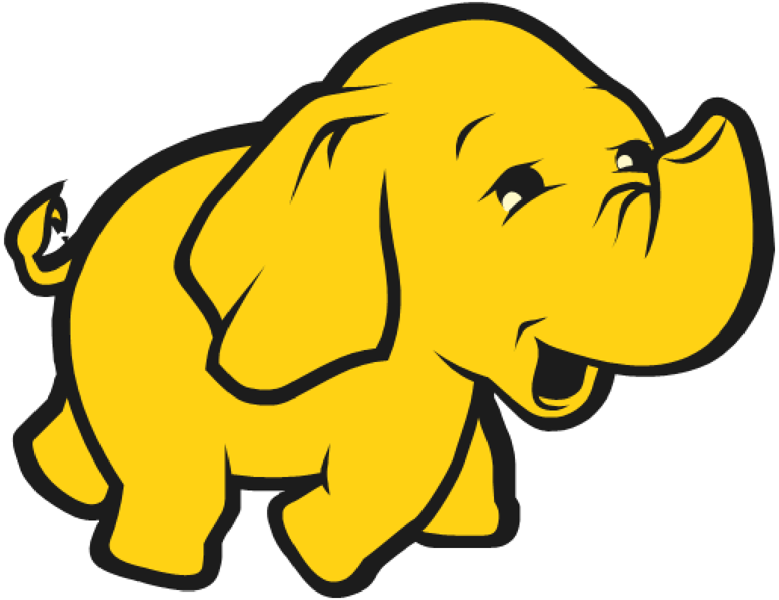
Data Processing

Tomasz Krawczyk

Future Processing

Odwiedź nas: WWW.DPTO.PL

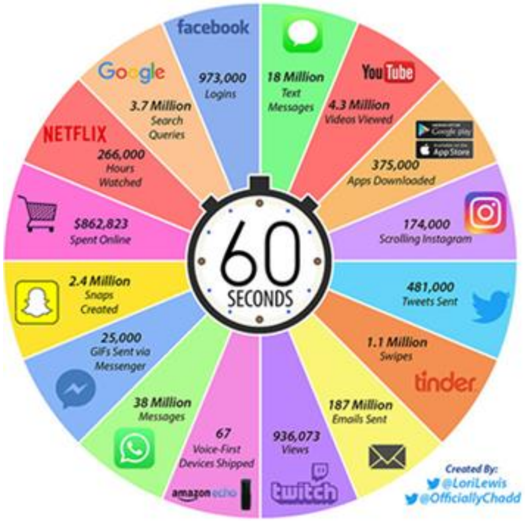
DPTO
Dobre Praktyki
Tworzenia Oprogramowania



?

Internet Minute

- **Google - 3.7 Million Search Queries**
- **Twitter – 481.000 Tweets Sent**
- **18 Million Text Messages**
- **187 Million Emails Sent**

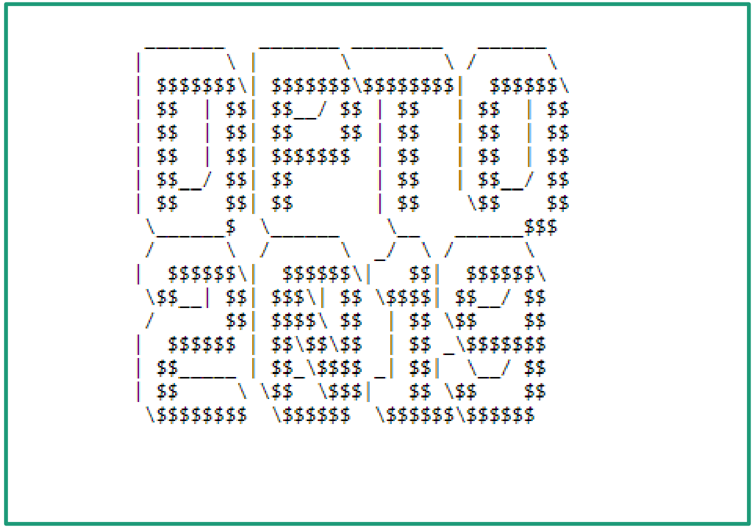


Source:https://www.visualcapitalist.com/internet-minute-2018/

40 Zetta bytes by 2020
163 Zetta bytes by 2025

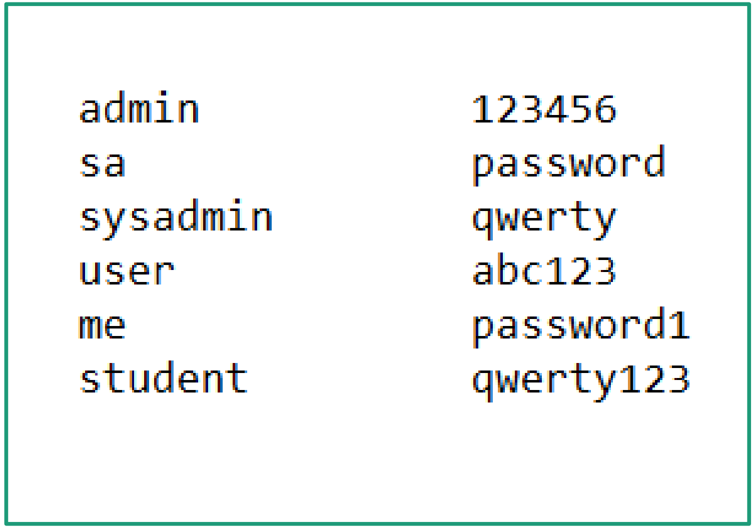
Byte	One grain of rice
Kilobyte	Cup of rice
Megabyte	8 bags of rice
Gigabyte	**3 semi trucks**
Terabyte	**2 container ships**
Petabyte	**Blankets Manhattan**
Exabyte	**Blankets west coast states**
Zettabyte	**Fills the Pacific Ocean**
Yottabyte	**As earth-sized rice ball**

Value of Data



```
 _______   _______ ________   ______
|       \ |       \        \ /      \
| $$$$$$$\| $$$$$$$\$$$$$$$$|  $$$$$$\
| $$  | $$| $$__/ $$ | $$   | $$  | $$
| $$  | $$| $$    $$ | $$   | $$  | $$
| $$  | $$| $$$$$$$  | $$   | $$  | $$
| $$__/ $$| $$       | $$   | $$__/ $$
| $$    $$| $$       | $$    \$$    $$
 \______$  \______    \__   ______$$$
 /      \  /      \  _/  \ /      \
|  $$$$$$\|  $$$$$$\|   $$|  $$$$$$\
 \$$__| $$| $$$\| $$ \$$$$| $$__/ $$
 /      $$| $$$$\ $$   | $$ \$$    $$
|  $$$$$$ | $$\$$\$$  | $$ _\$$$$$$$
| $$_____ | $$_\$$$$ _| $$|  \__/ $$
| $$      \ \$$  \$$$|   $$ \$$    $$
 \$$$$$$$$  \$$$$$$  \$$$$$$\$$$$$$
```

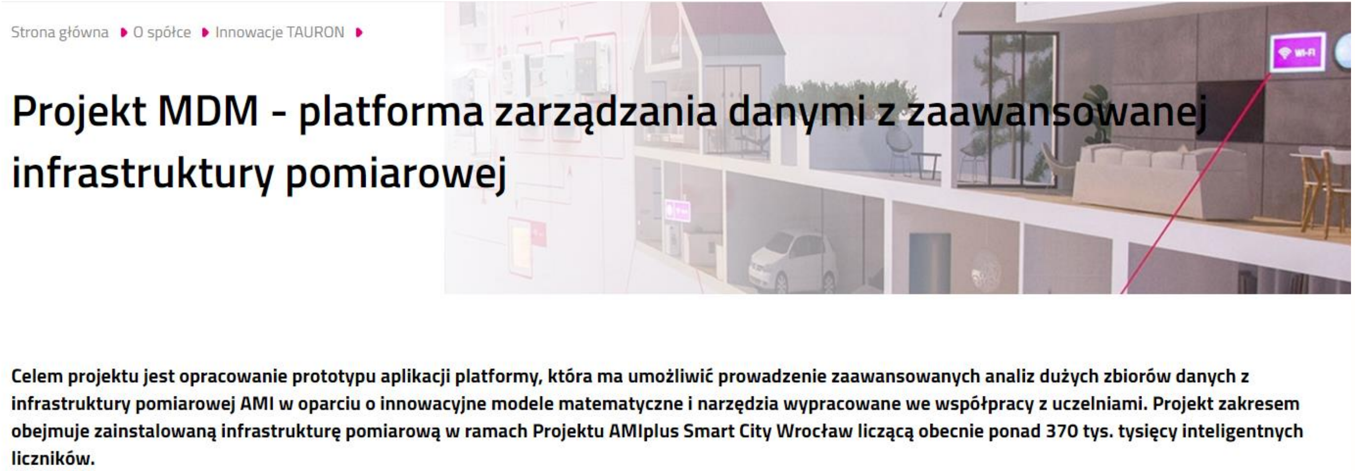
FILE 1



```
admin       123456
sa          password
sysadmin    qwerty
user        abc123
me          password1
student     qwerty123
```

FILE 2

Do you know…



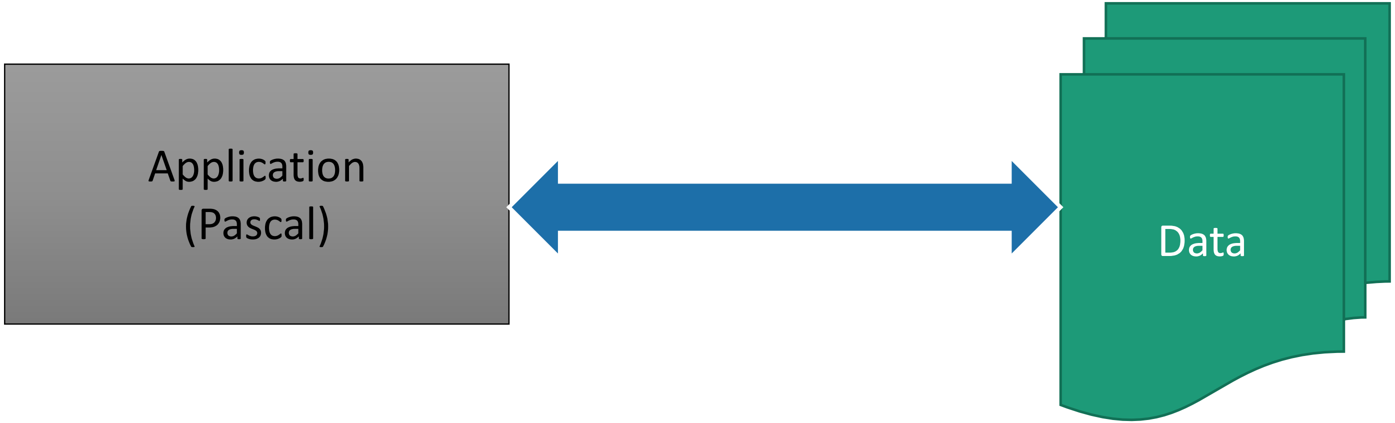
Strona główna ▸ O spółce ▸ Innowacje TAURON ▸

Projekt MDM - platforma zarządzania danymi z zaawansowanej infrastruktury pomiarowej

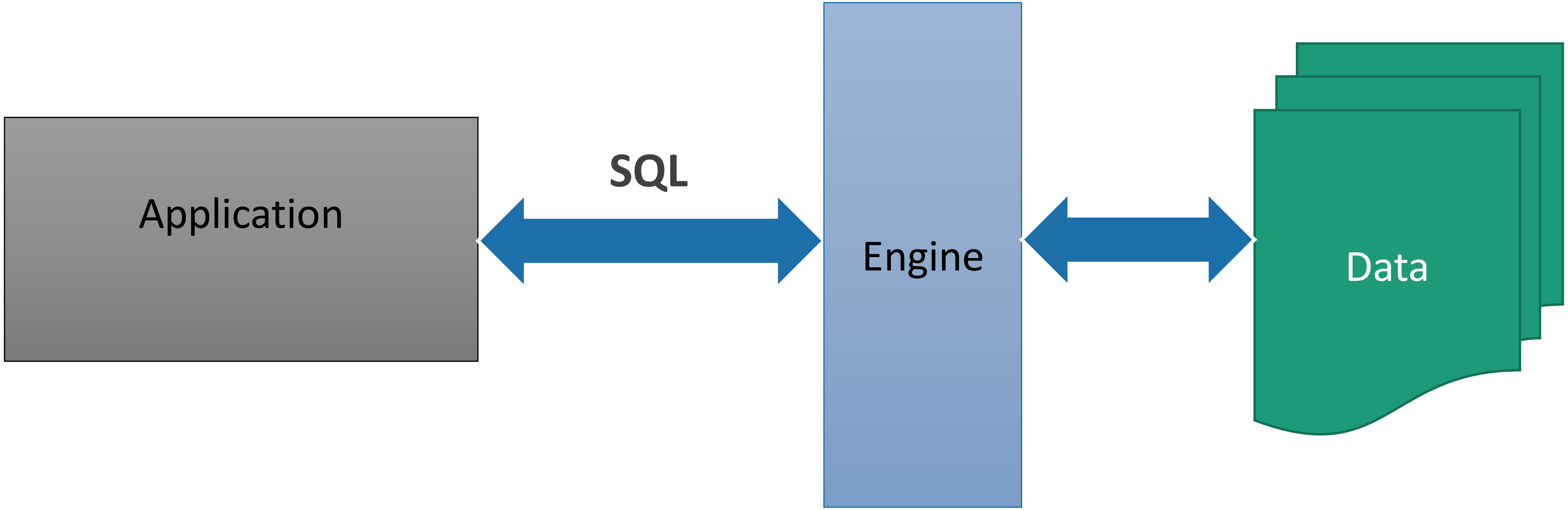
Celem projektu jest opracowanie prototypu aplikacji platformy, która ma umożliwić prowadzenie zaawansowanych analiz dużych zbiorów danych z infrastruktury pomiarowej AMI w oparciu o innowacyjne modele matematyczne i narzędzia wypracowane we współpracy z uczelniami. Projekt zakresem obejmuje zainstalowaną infrastrukturę pomiarową w ramach Projektu AMIplus Smart City Wrocław liczącą obecnie ponad 370 tys. tysięcy inteligentnych liczników.

History…

My First Application



My First Application



Relation Model

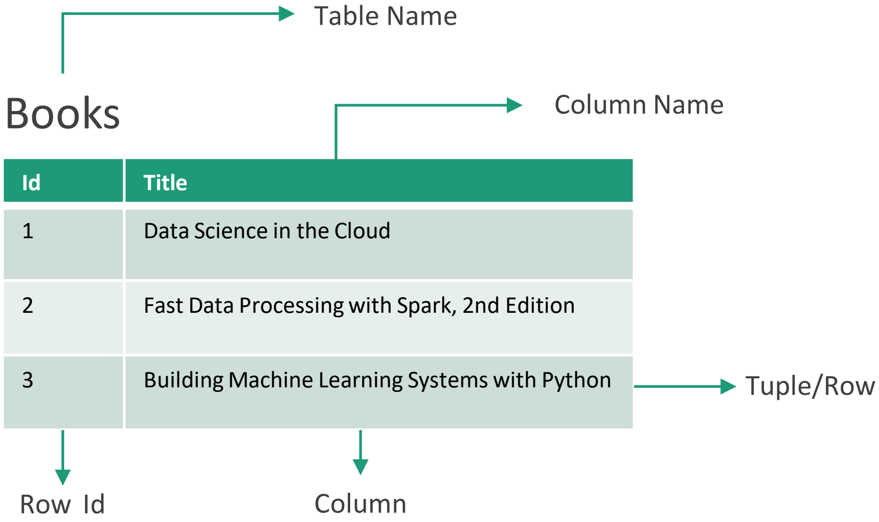


Table Name

Books

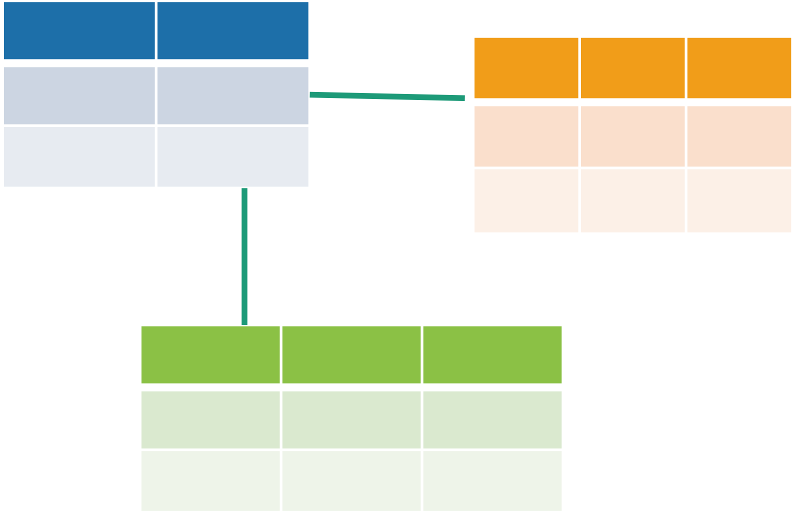
Column Name

Id	Title
1	Data Science in the Cloud
2	Fast Data Processing with Spark, 2nd Edition
3	Building Machine Learning Systems with Python

Tuple/Row

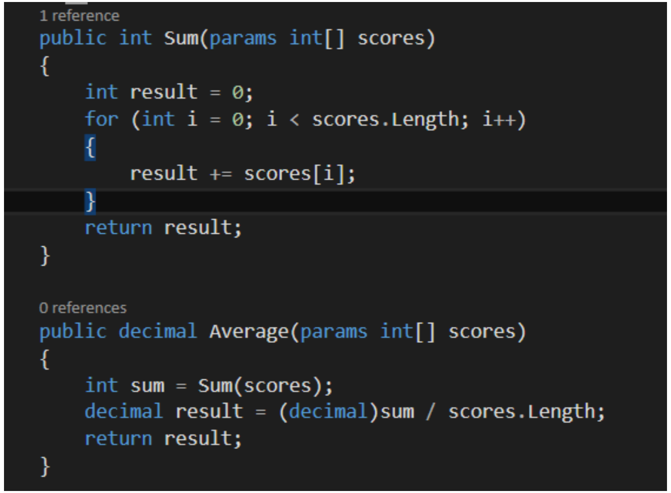
Row Id

Column



SQL

BookId	Score
1	★★★★
2	★★
1	★★★★
2	★★★★



```
1 reference
public int Sum(params int[] scores)
{
    int result = 0;
    for (int i = 0; i < scores.Length; i++)
    {
        result += scores[i];
    }
    return result;
}

0 references
public decimal Average(params int[] scores)
{
    int sum = Sum(scores);
    decimal result = (decimal)sum / scores.Length;
    return result;
}
```

```
SELECT AVG(Score) AS AvgScore FROM Scores WHERE BookId = 1
```

SQL

BookId	Score
1	★★★★
2	★★
1	★★★★
2	★★★★

```
SELECT BookId ,AVG(Score) AS AvgScore
FROM Scores GROUP BY BookId
ORDER BY AvgScore DESC
```

Relational Model -Challenges

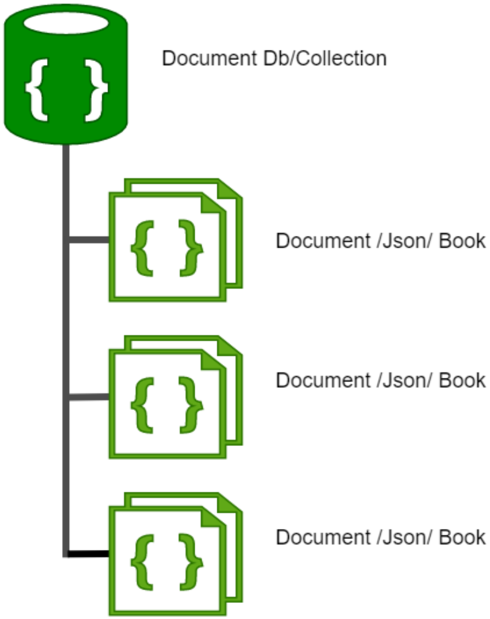
Books

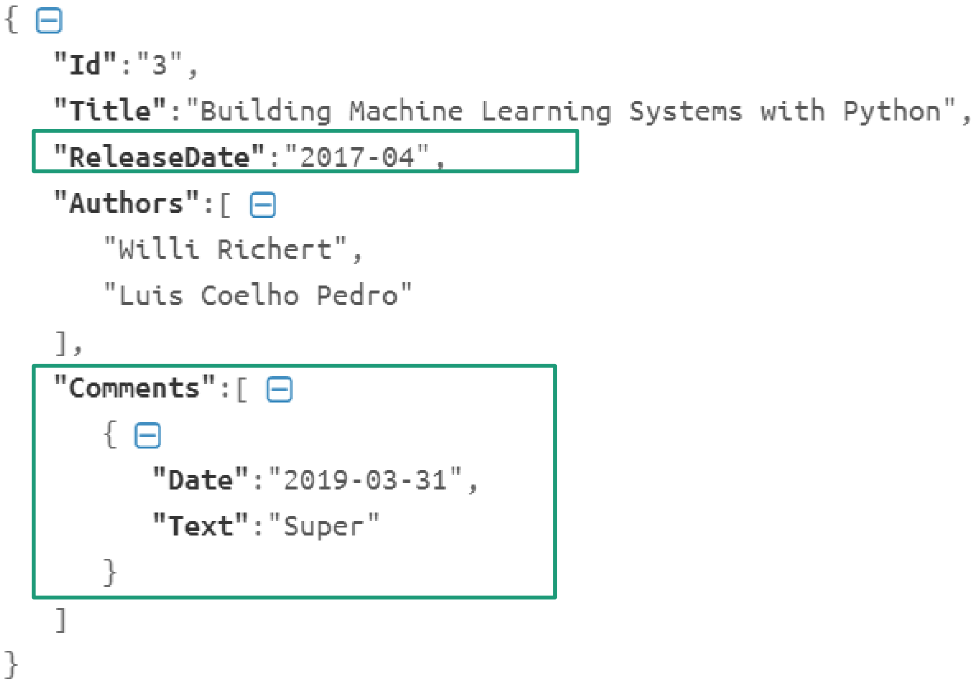
Id	Title	Release date
1	Data Science in the Cloud	
2	Fast Data Processing with Spark, 2nd Edition	
3	Building Machine Learning Systems with Python	2017-04

BookId	Author
1	Stephen F. Elston
2	Krishna Santar
3	Willi Richert
3	Luis Coelho Pedro

BookId	Comment
1	
1	Ok
3	Super

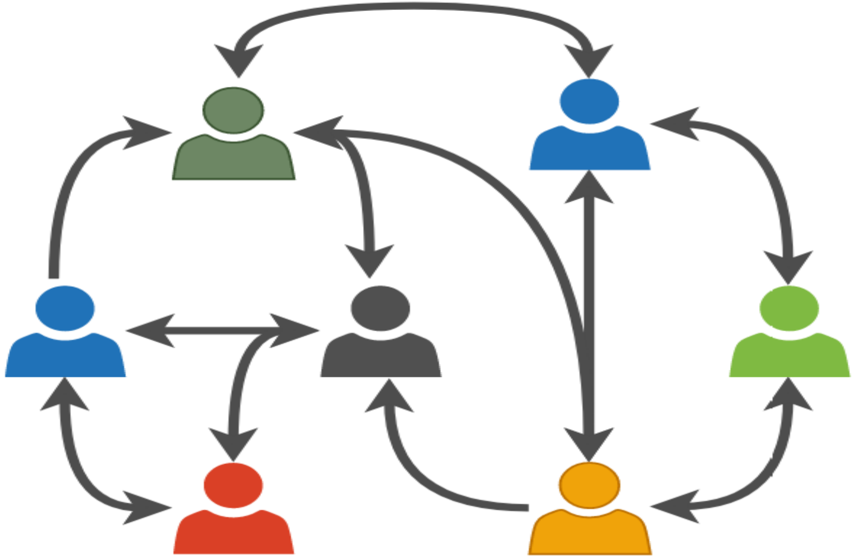
Document Db

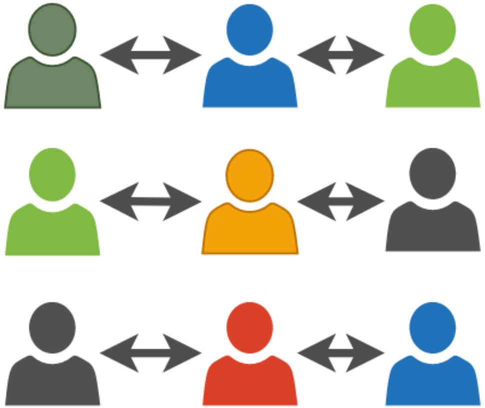




```
{ ⊟
  "Id":"3",
  "Title":"Building Machine Learning Systems with Python",
  "ReleaseDate":"2017-04",
  "Authors":[ ⊟
     "Willi Richert",
     "Luis Coelho Pedro"
  ],
  "Comments":[ ⊟
     { ⊟
        "Date":"2019-03-31",
        "Text":"Super"
     }
  ]
}
```

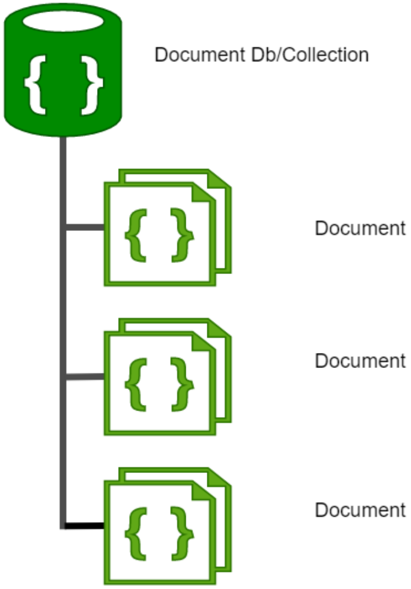
New Challenges



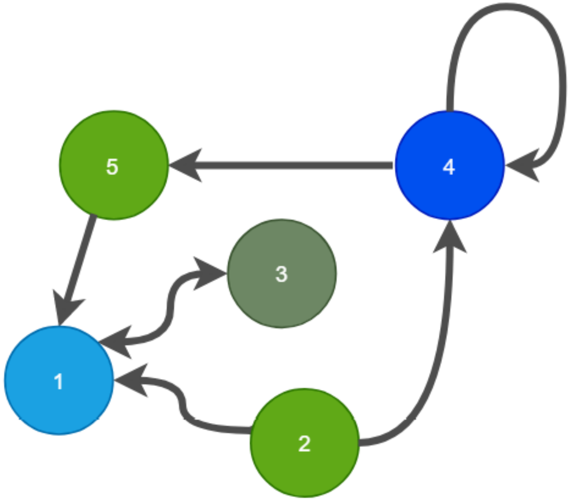


NoSQL –Not only SQL

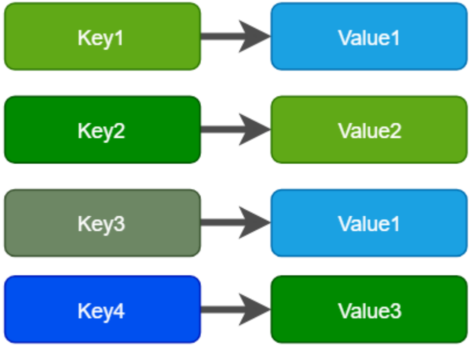
Document Db



Graph Db



Key Value Db



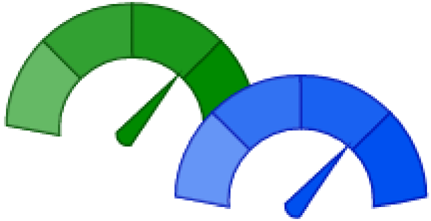
Key1	Value1
Key2	Value2
Key3	Value1
Key4	Value3

Big Data (3V)

Byte	One grain of rice
Kilobyte	Cup of rice
Megabyte	8 bags of rice
Gigabyte	**3 semi trucks**
Terabyte	**2 container ships**
Petabyte	**Blankets Manhattan**
Exabyte	**Blankets west coast states**
Zettabyte	**Fills the Pacific Ocean**
Yottabyte	**As earth-sized rice ball**

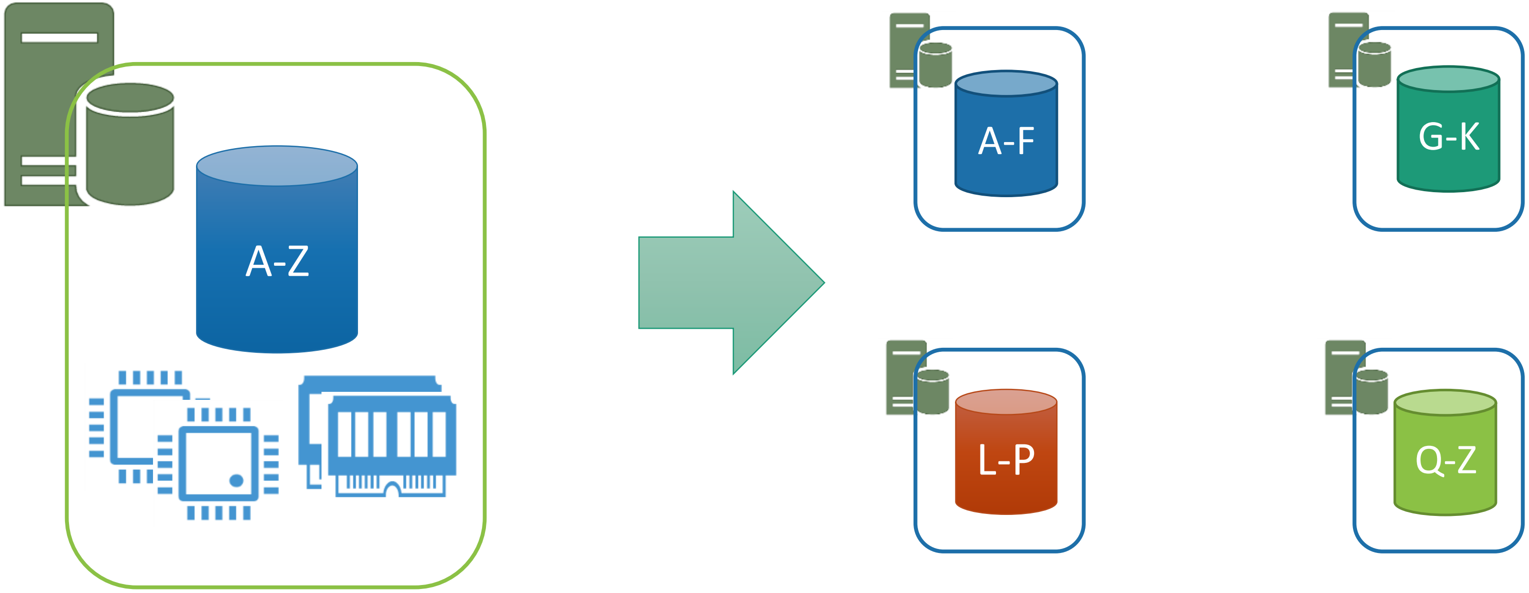
Data Volume

Data Variety

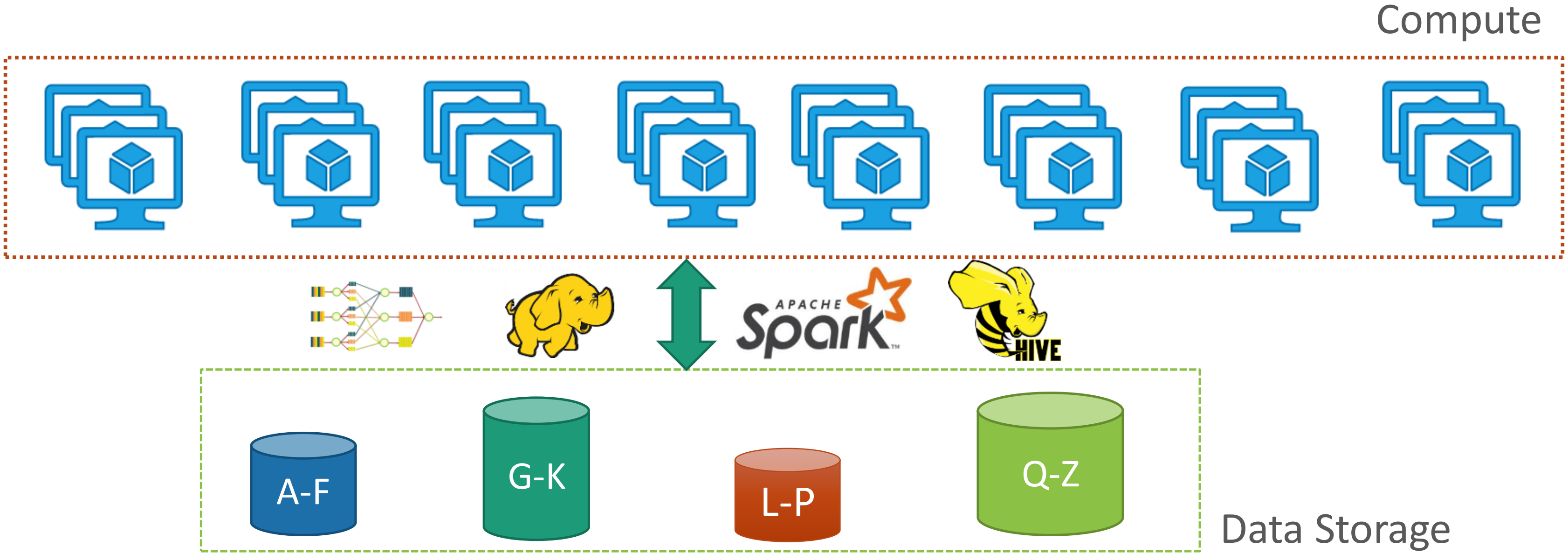


Data Velocity

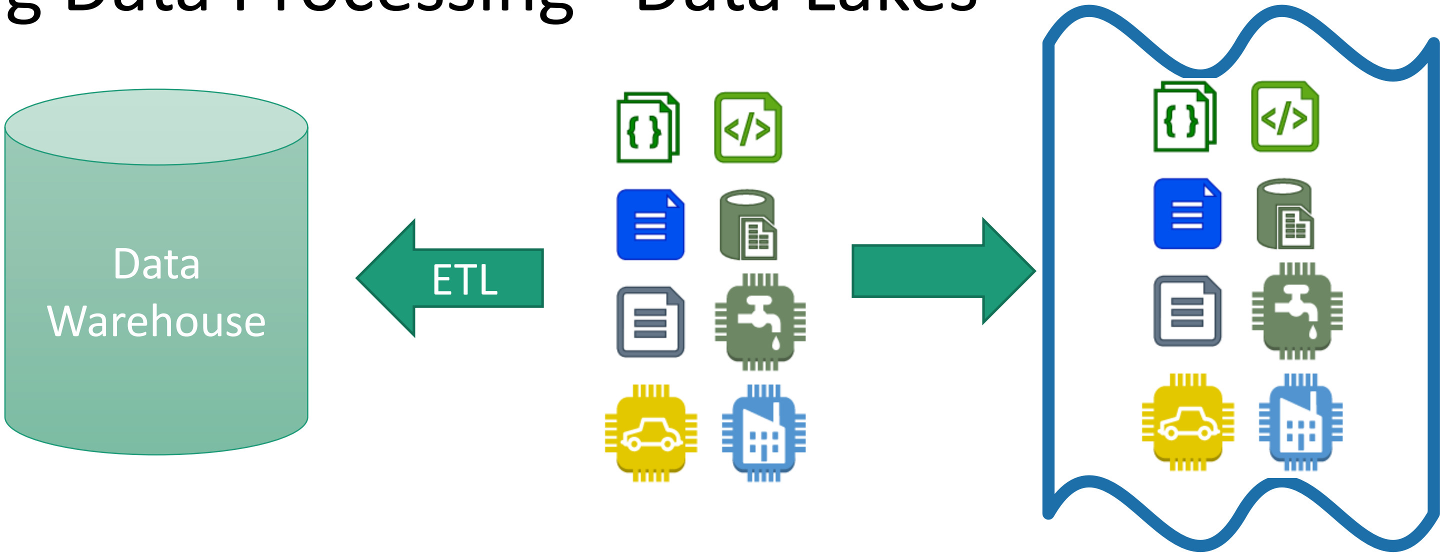
Big Data Processing



Big Data Processing



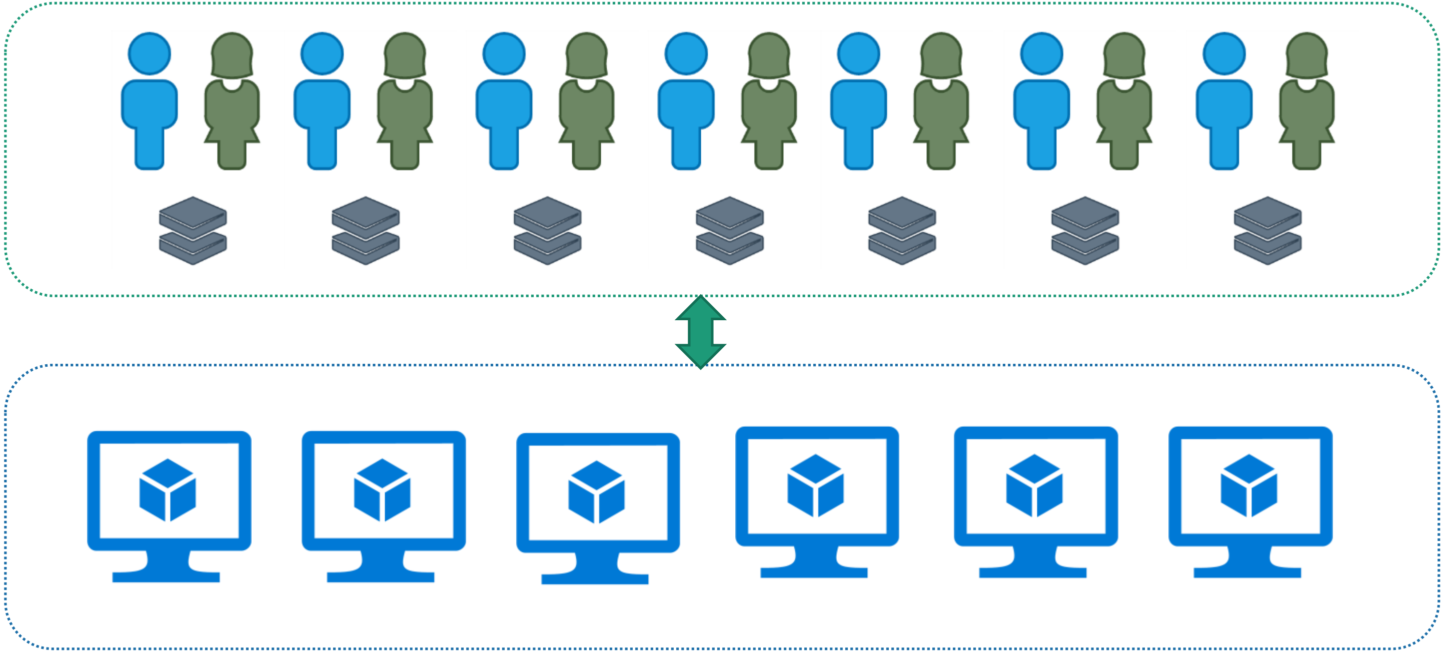
Big Data Processing –Data Lakes



I(ngest) S(tore) A(nalyse) S(urface) A(ct)

Make Me More Money

Scalable runtime



Cloud

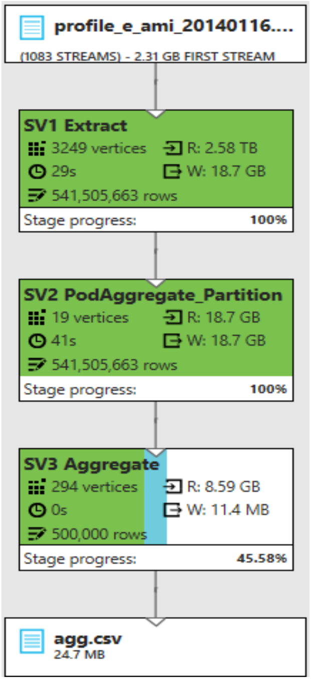
amazon
web services

Google
Cloud Platform

Alibaba Cloud

Microsoft
Azure

Cloud -Example



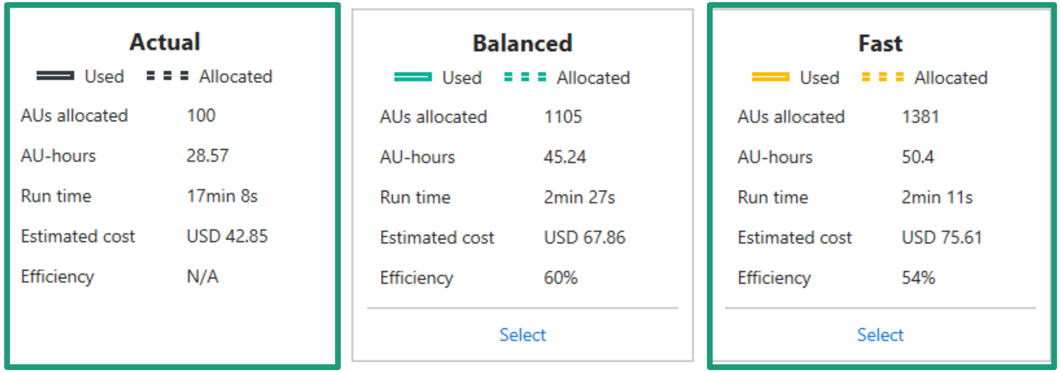
profile_e_ami_20140116....
(1083 STREAMS) - 2.31 GB FIRST STREAM

SV1 Extract
3249 vertices R: 2.58 TB
29s W: 18.7 GB
541,505,663 rows
Stage progress: 100%

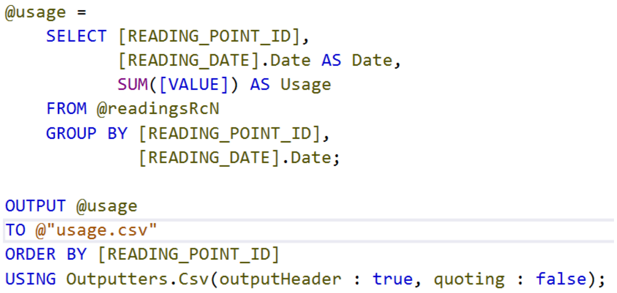
SV2 PodAggregate_Partition
19 vertices R: 18.7 GB
41s W: 18.7 GB
541,505,663 rows
Stage progress: 100%

SV3 Aggregate
294 vertices R: 8.59 GB
0s W: 11.4 MB
500,000 rows
Stage progress: 45.58%

agg.csv
24.7 MB

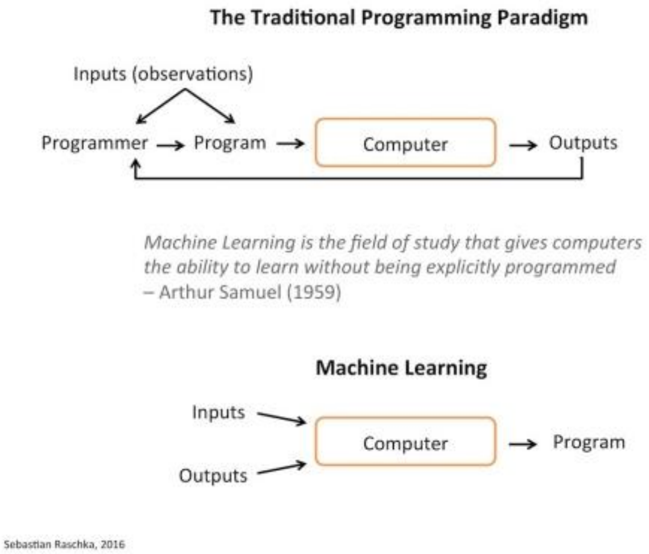


	Actual	Balanced	Fast
	▬ Used ▪▪▪ Allocated	▬ Used ▪▪▪ Allocated	▬ Used ▪▪▪ Allocated
AUs allocated	100	1105	1381
AU-hours	28.57	45.24	50.4
Run time	17min 8s	2min 27s	2min 11s
Estimated cost	USD 42.85	USD 67.86	USD 75.61
Efficiency	N/A	60%	54%
		Select	Select



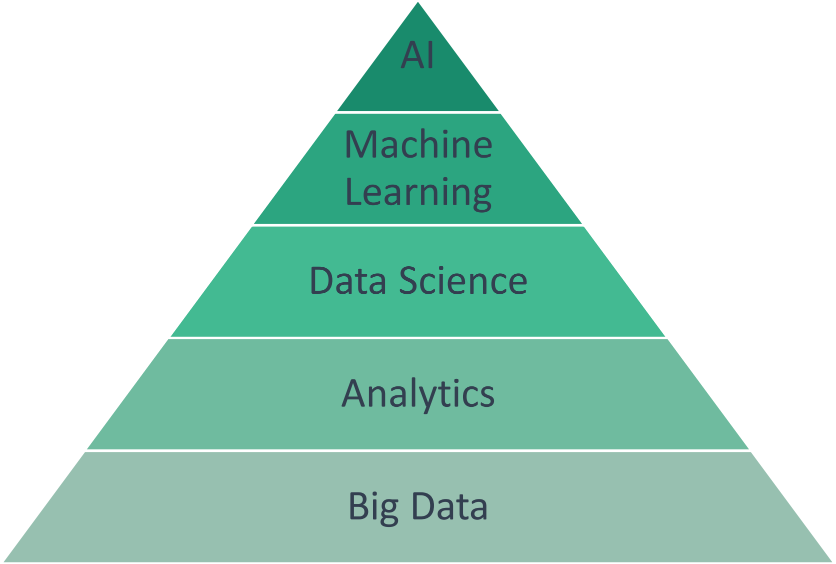
```
@usage =
    SELECT [READING_POINT_ID],
           [READING_DATE].Date AS Date,
           SUM([VALUE]) AS Usage
    FROM @readingsRcN
    GROUP BY [READING_POINT_ID],
             [READING_DATE].Date;

OUTPUT @usage
TO @"usage.csv"
ORDER BY [READING_POINT_ID]
USING Outputters.Csv(outputHeader : true, quoting : false);
```

AI and Machine Learning



Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed
– Arthur Samuel (1959)

Sebastian Raschka, 2016



Summary

Data Scientist
also known as Data Managers, statisticians.

Data Engineers
also known as database administrators and data architects.

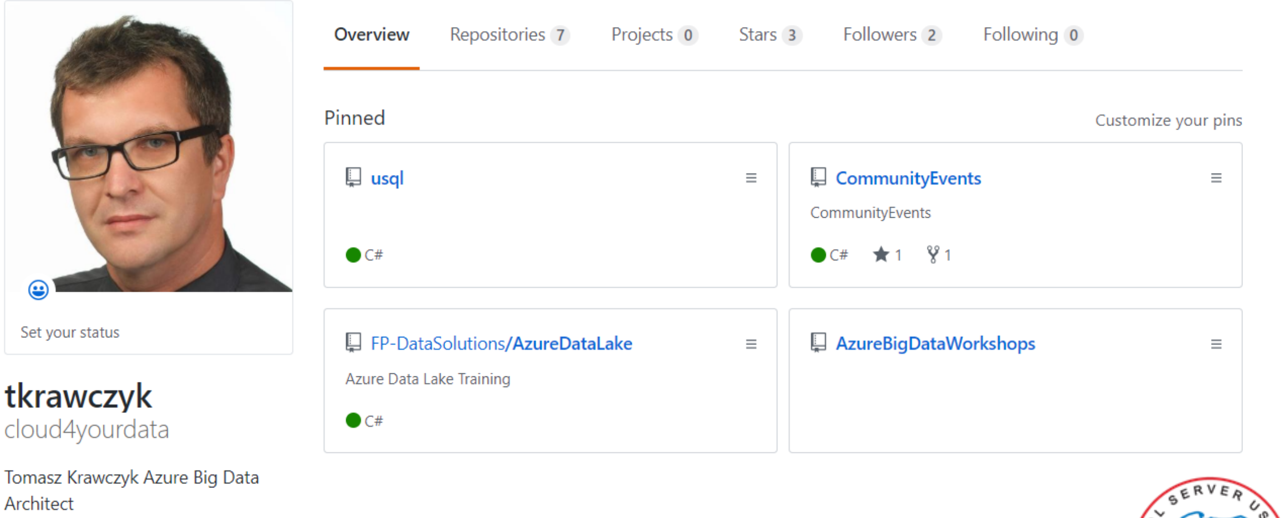
Data Analysts
also known as business Analysts.

SQL - Structured Query Language

R language is a golden child of machine learning

Python is a king of machine learning

About Me



Overview Repositories 7 Projects 0 Stars 3 Followers 2 Following 0

Set your status

tkrawczyk
cloud4yourdata

Tomasz Krawczyk Azure Big Data Architect

Pinned — Customize your pins

usql ● C#	CommunityEvents CommunityEvents ● C# ★ 1 1
FP-DataSolutions/AzureDataLake Azure Data Lake Training ● C#	AzureBigDataWorkshops

https://github.com/cloud4yourdata

https://github.com/cloud4yourdata/CommunityEvents

https://github.com/FP-DataSolutions/AzureBigDataWorkshops

DPTO
Dobre Praktyki
Tworzenia Oprogramowania

Q & A

Future Processing

Odwiedź nas: WWW.DPTO.PL

DPTO
Dobre Praktyki
Tworzenia Oprogramowania

KONTAKT@DPTO.PL

Future Processing

Odwiedź nas: WWW.DPTO.PL