Extra Notes- Big Data

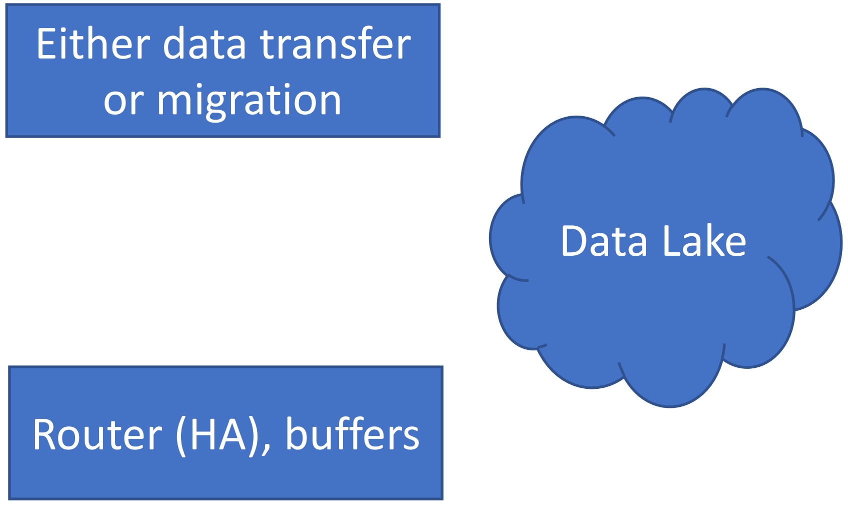
7 Aug 2023

Data-at-rest

SQL, NoSQL databases
Folders and files
Logs and crash reports
Dump files
Extracts, backups

Data-in-motion

Games, videos, CCTVs,
IoT, sensors/actuators



Either data transfer or migration

Data Lake

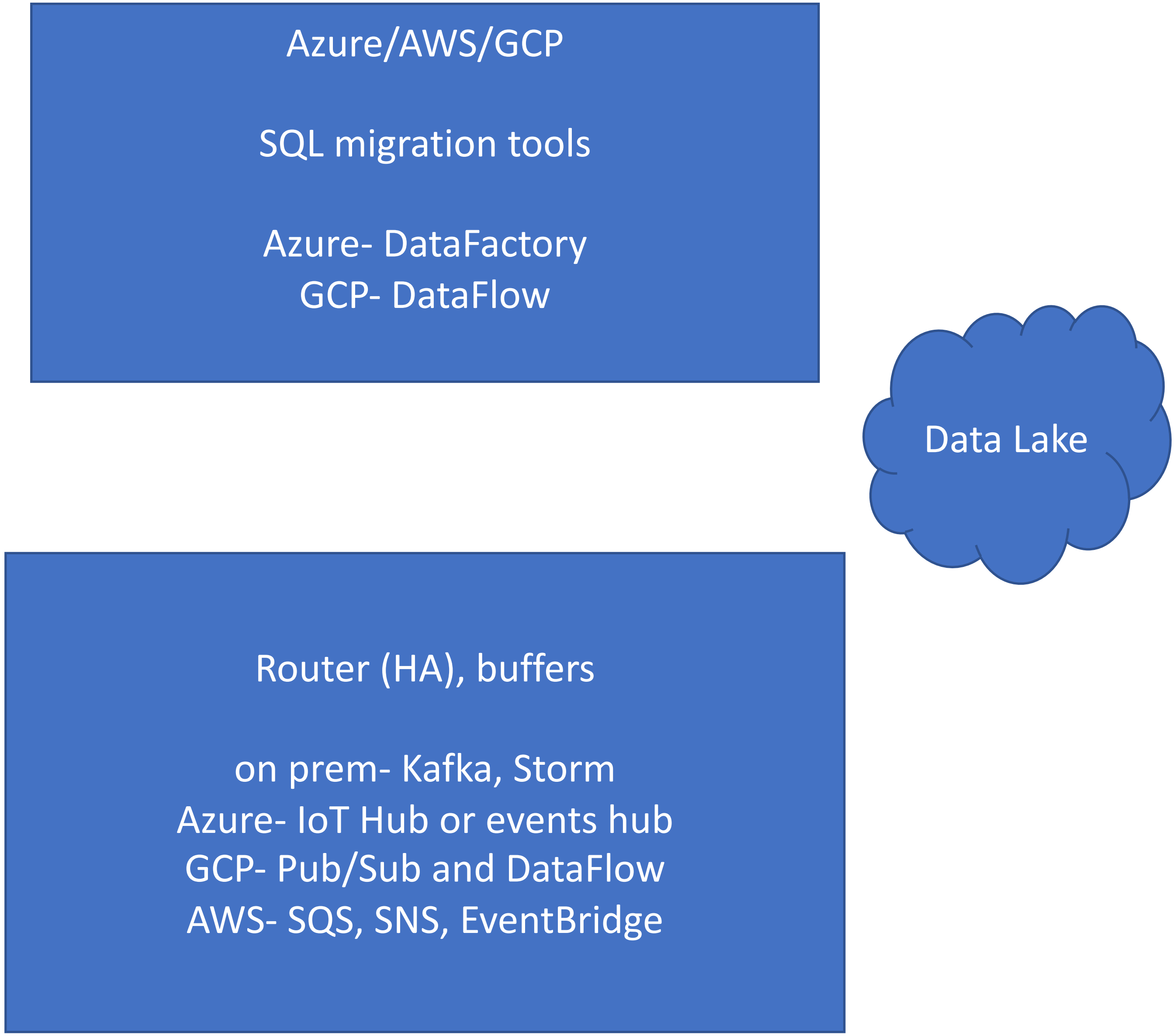
Router (HA), buffers

Data-at-rest (batch)

SQL, NoSQL databases
Folders and files
Logs and crash reports
Dump files
Extracts, backups

Data-in-motion (stream)

Games, videos, CCTVs,
IoT, sensors/actuators



Azure/AWS/GCP

SQL migration tools

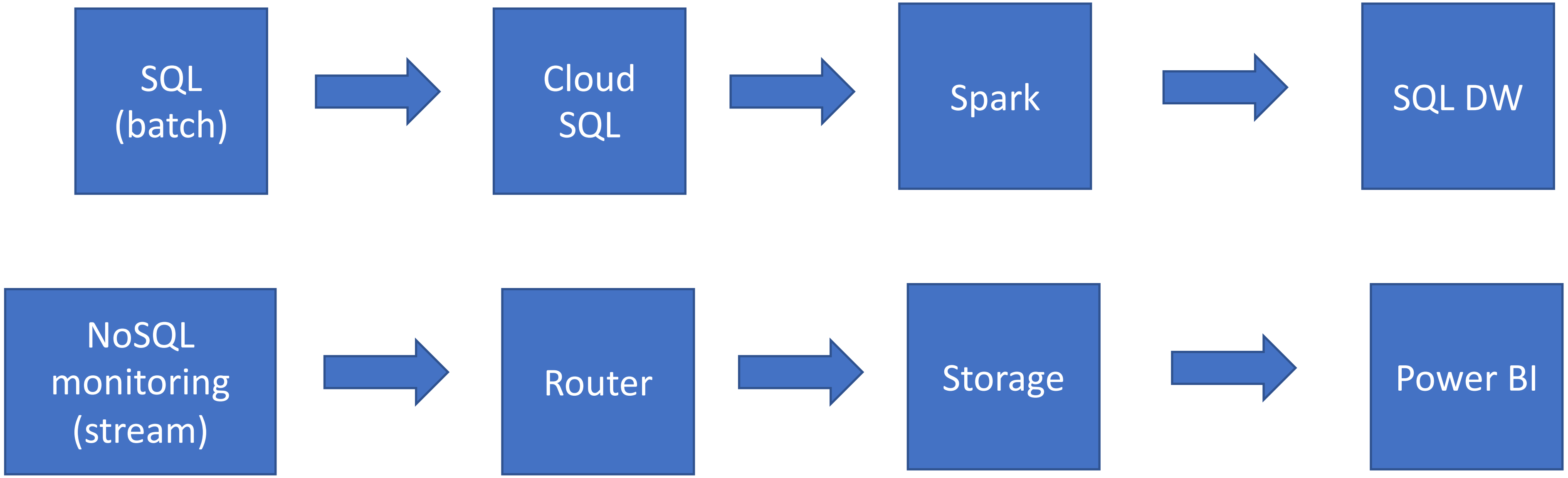
Azure- DataFactory
GCP- DataFlow

Data Lake

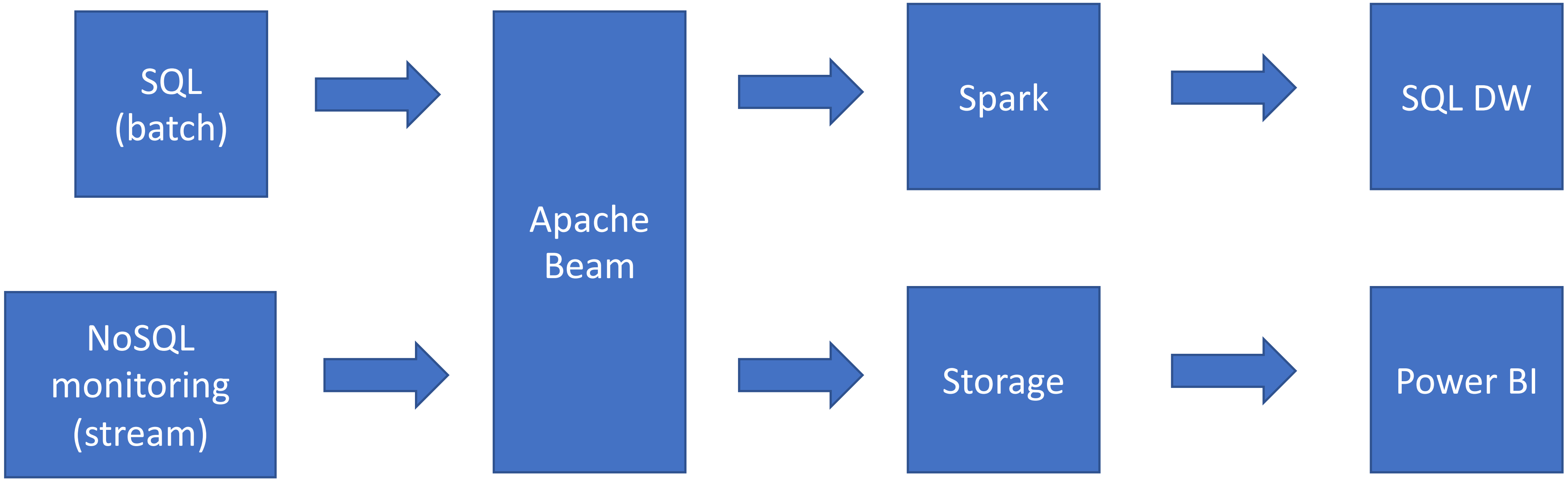
Router (HA), buffers

on prem- Kafka, Storm
Azure- IoT Hub or events hub
GCP- Pub/Sub and DataFlow
AWS- SQS, SNS, EventBridge

Lambda

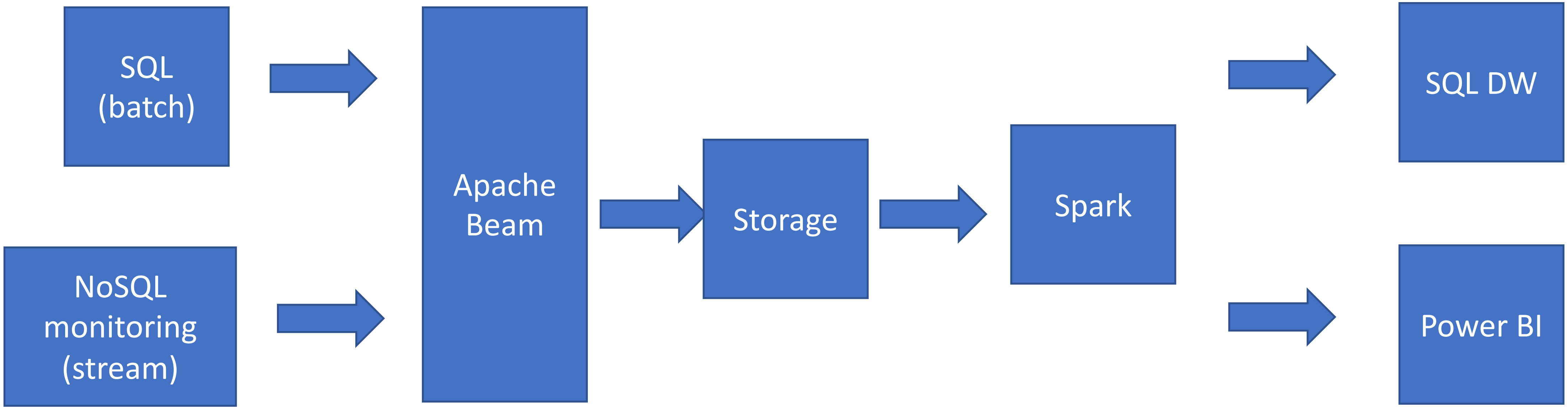

SQL (batch)
Cloud SQL
Spark
SQL DW
NoSQL monitoring (stream)
Router
Storage
Power BI

Kappa



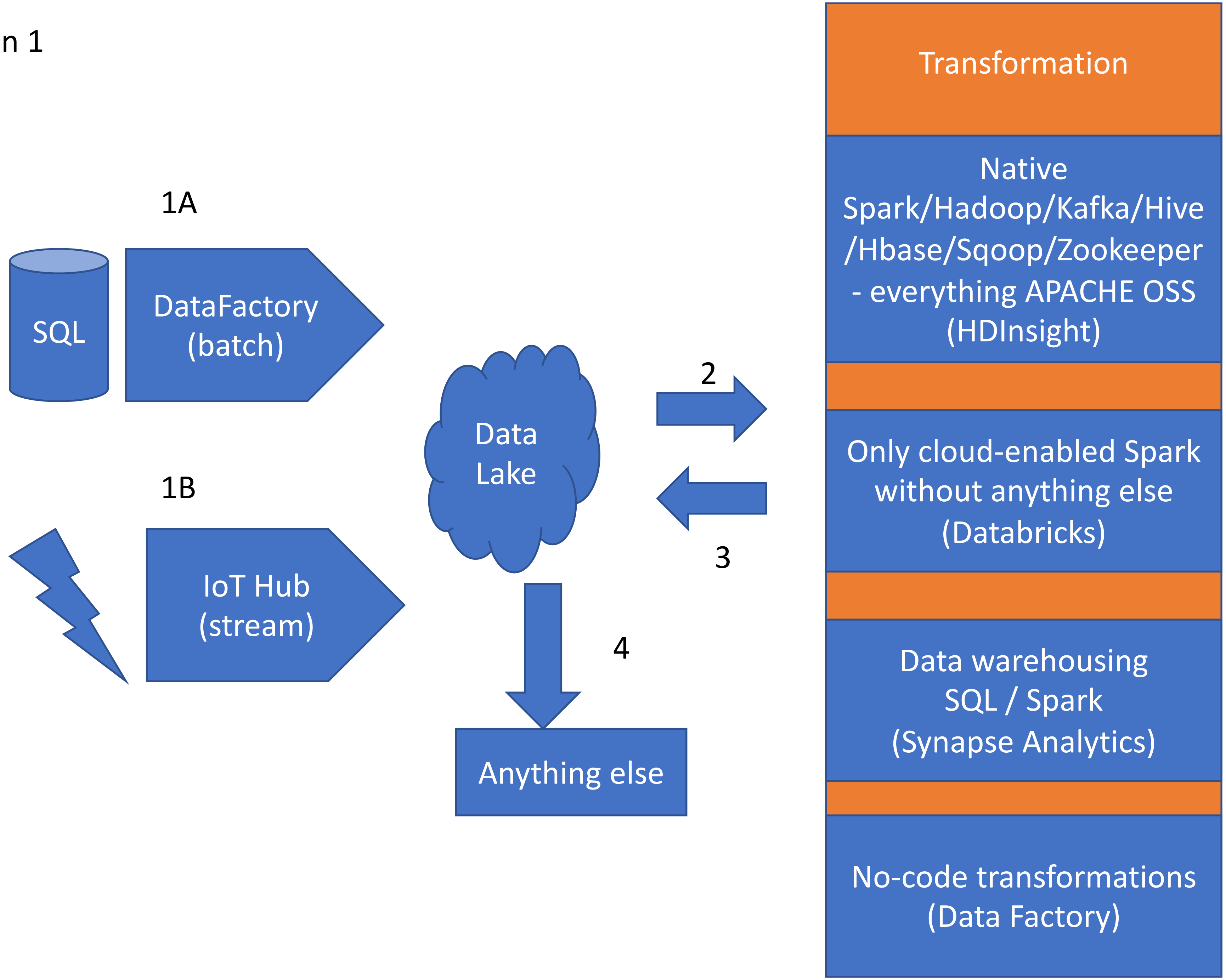
Beam = Batch + stEAM

Kappa



Beam = Batch + stEAM

Option 1



1A

SQL

DataFactory (batch)

1B

IoT Hub (stream)

Data Lake

2

3

4

Anything else

Transformation

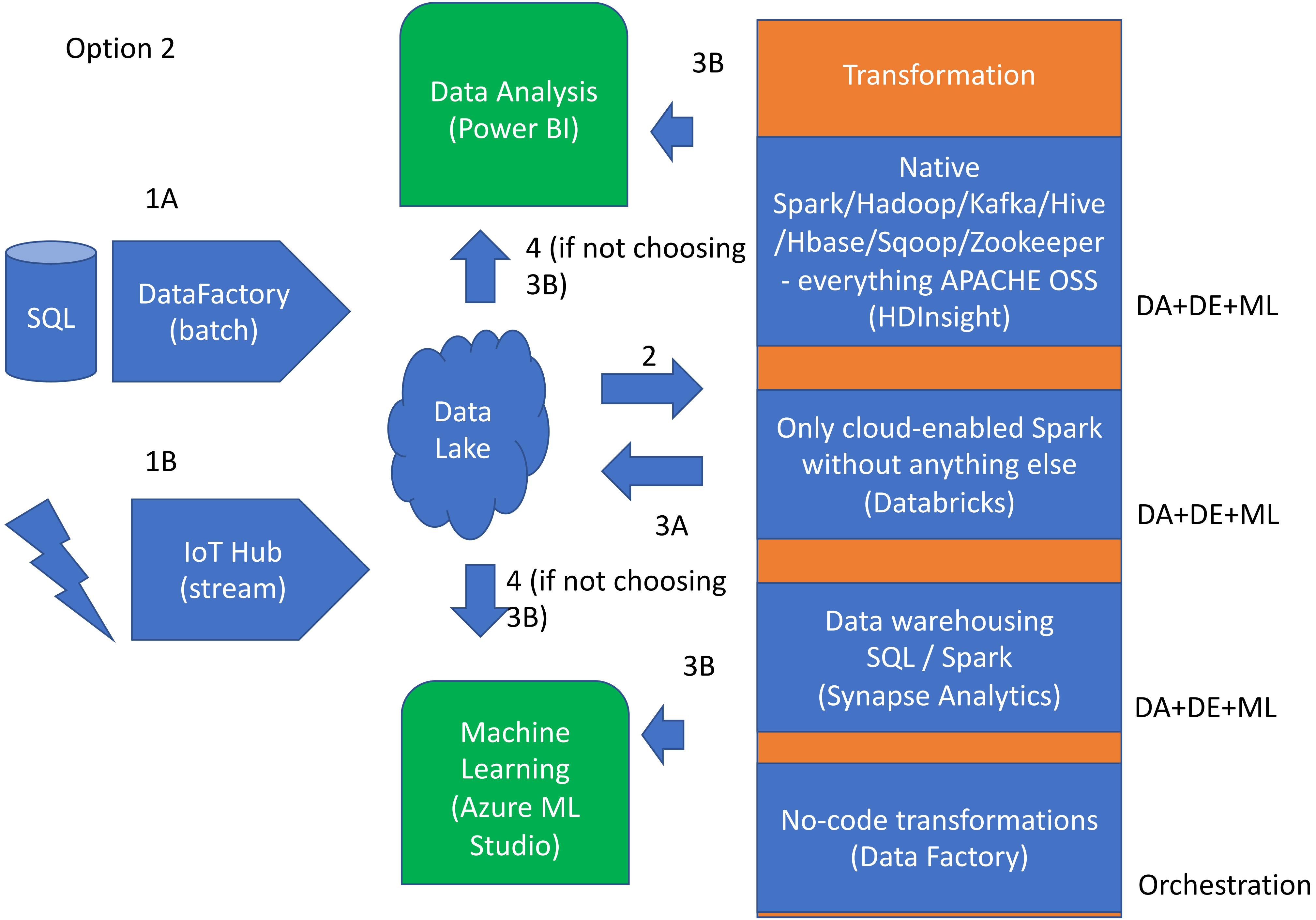
Native Spark/Hadoop/Kafka/Hive /Hbase/Sqoop/Zookeeper - everything APACHE OSS (HDInsight)

Only cloud-enabled Spark without anything else (Databricks)

Data warehousing SQL / Spark (Synapse Analytics)

No-code transformations (Data Factory)

Option 2



1A

SQL

DataFactory (batch)

1B

IoT Hub (stream)

Data Lake

Data Analysis (Power BI)

4 (if not choosing 3B)

2

3A

4 (if not choosing 3B)

Machine Learning (Azure ML Studio)

3B

3B

Transformation

Native Spark/Hadoop/Kafka/Hive /Hbase/Sqoop/Zookeeper - everything APACHE OSS (HDInsight)

DA+DE+ML

Only cloud-enabled Spark without anything else (Databricks)

DA+DE+ML

Data warehousing SQL / Spark (Synapse Analytics)

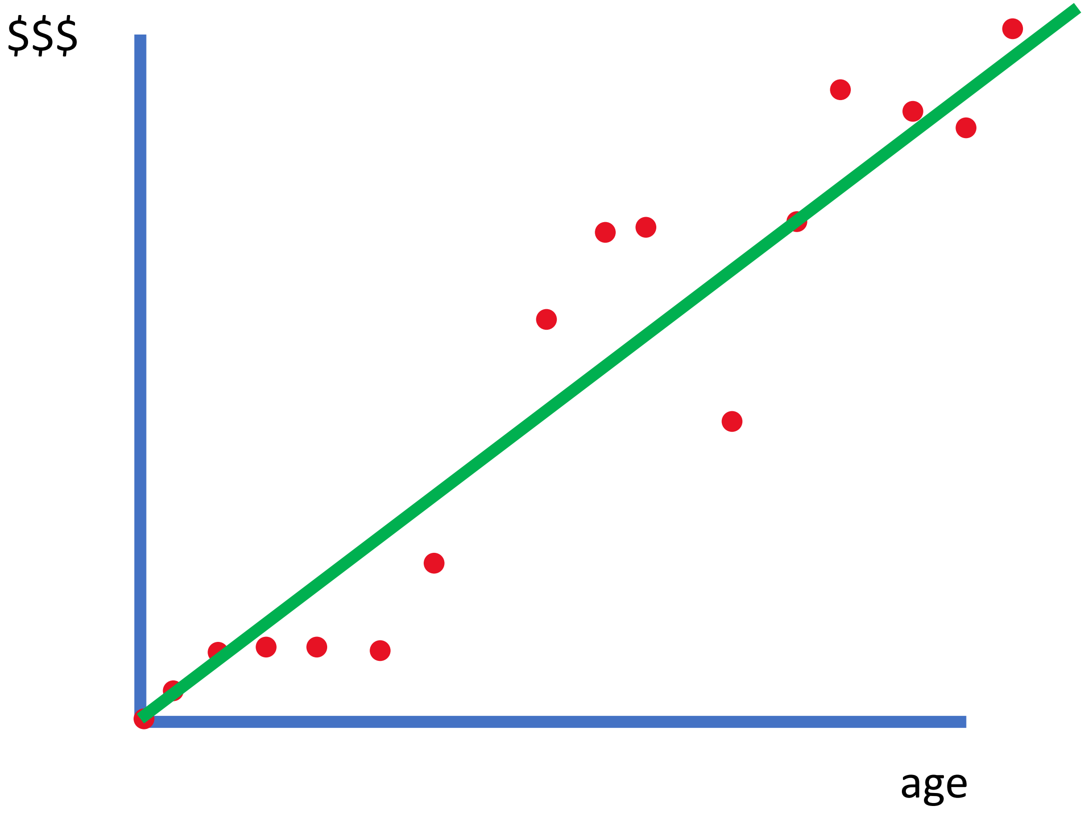
DA+DE+ML

No-code transformations (Data Factory)

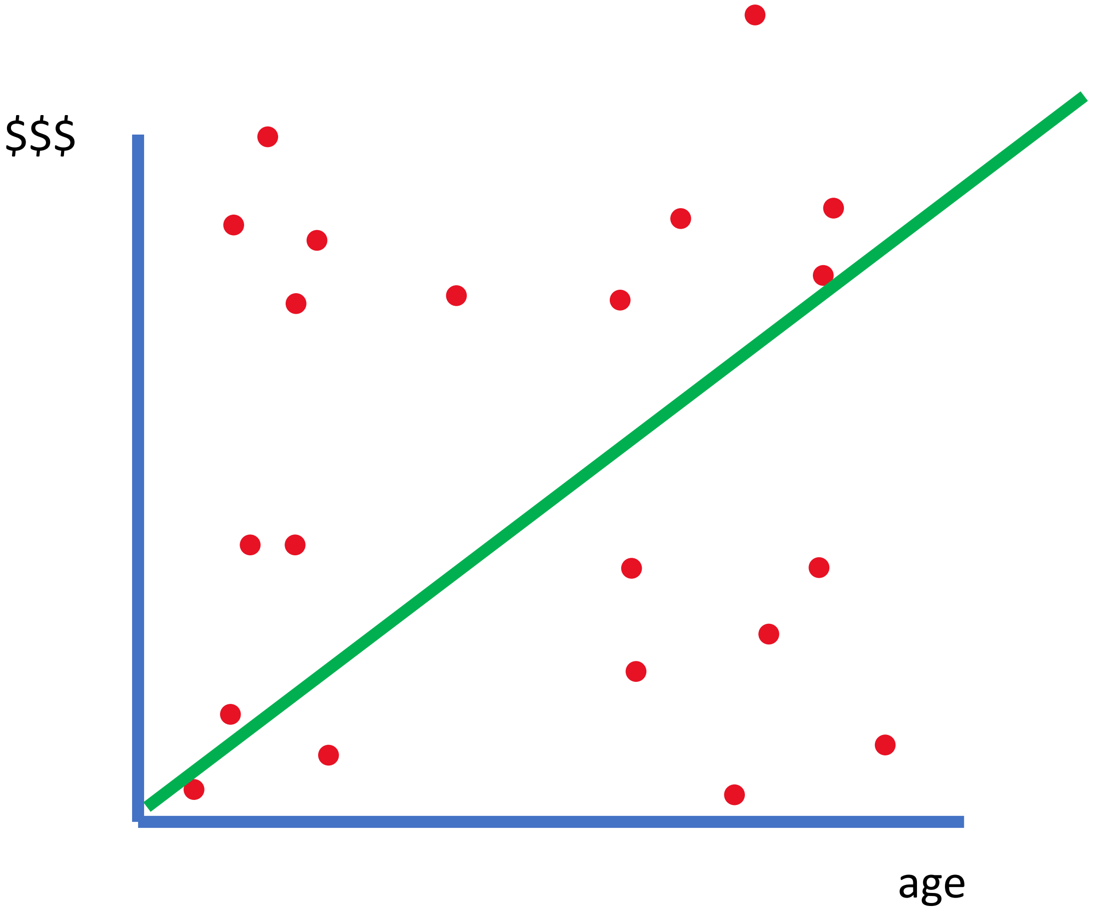
Orchestration

Intro to ML

This green line- this is ML!
ML is a BOUNDARY, an equation that can either propagate or classify

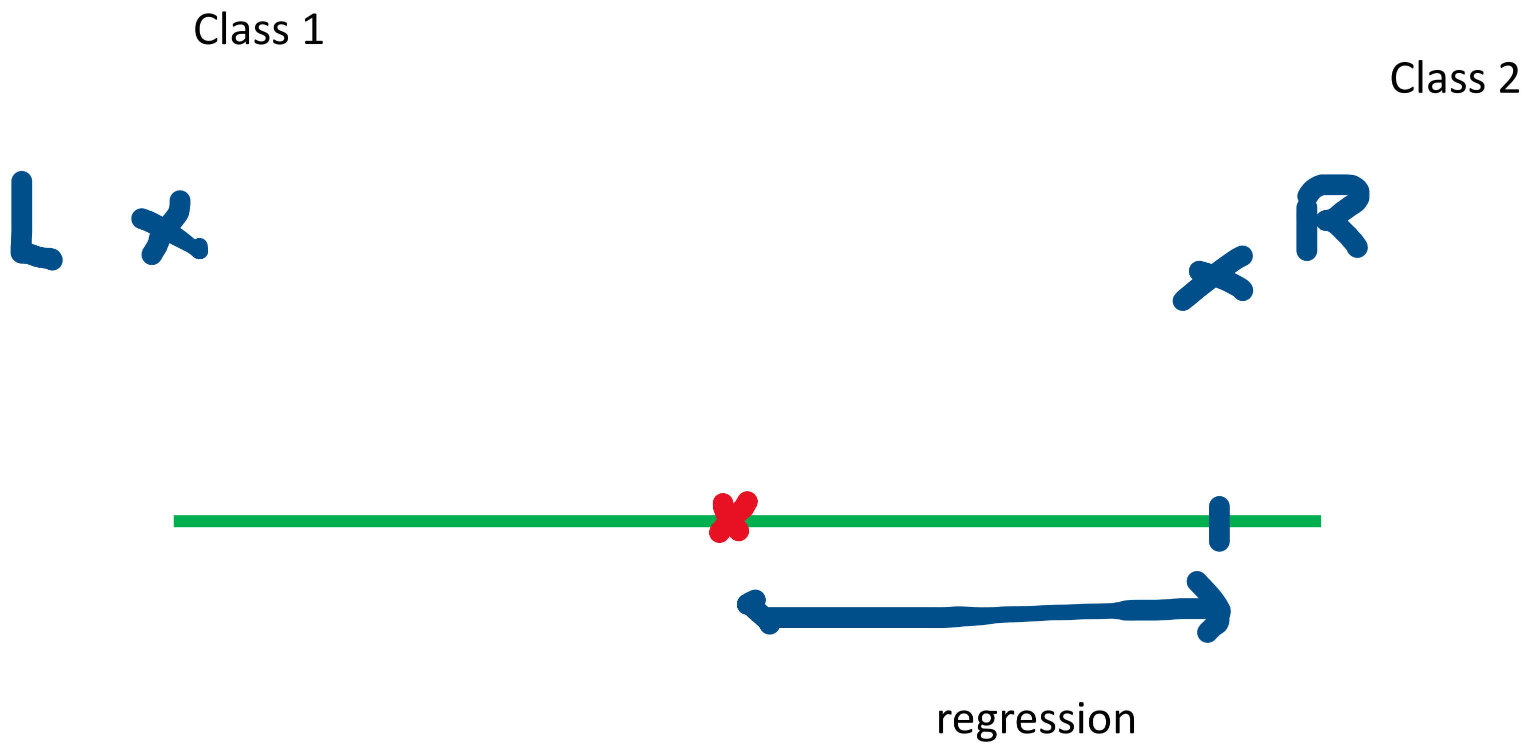


You can tell me $$$ for any age
$Y = mx + c$



Rich and Poor People
$Y1 > mx + c$
$Y2 < mx + c$

1-D

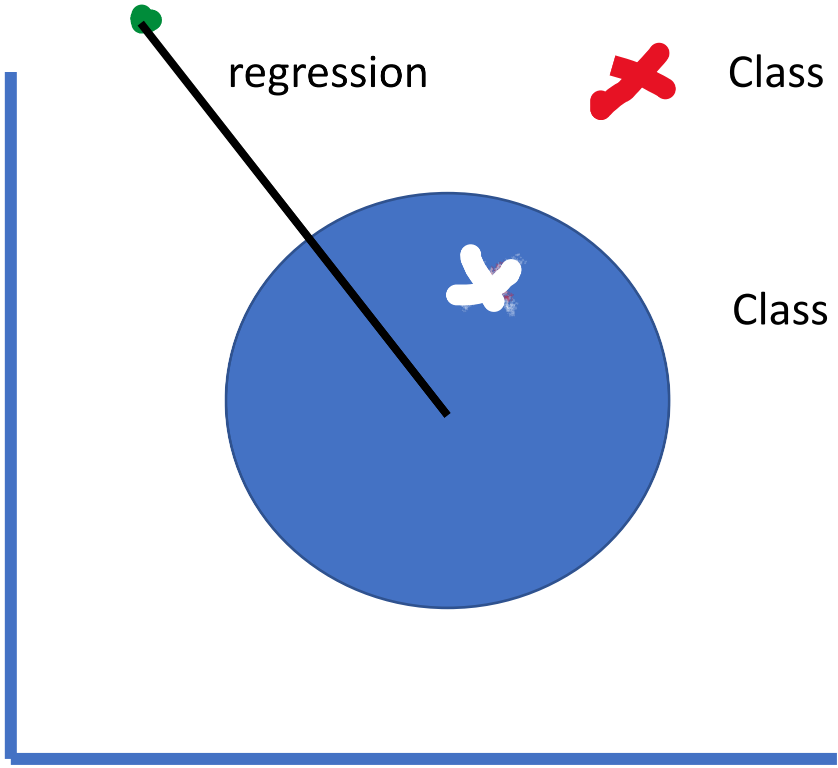


Class 1

Class 2

regression

2-D

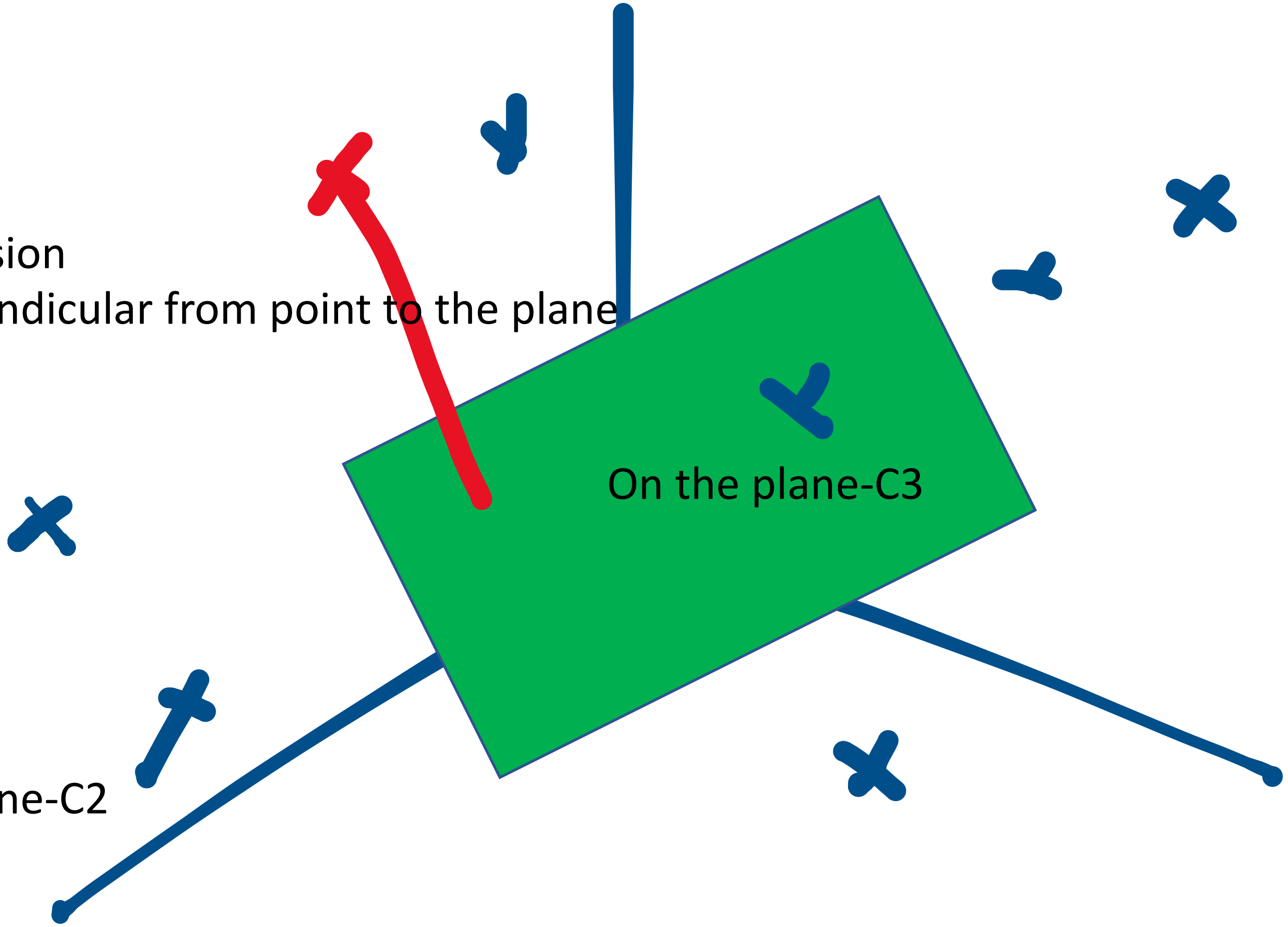


regression

Class 2- outside the circle

Class 1- inside the circle

3-D



Above the plane-C1

Regression
- Perpendicular from point to the plane

On the plane-C3

Below the plane-C2

Terminology

- TENSOR- flow of data
  - 0 dimension tensors -> SCALARS-> data not flowing anywhere!
  - 1 dimension tensors-> VECTORS-> data flowing in a direction!
  - 2 dimension tensors-> MATRICES -> data flowing in 2 directions (PLANE)
  - 3 dimension tensors-> CUBES -> Array of matrices
  - …… N dimension tensors
  - Tesseracts!
    - Higher dimensions can see lower dimensions
    - Lower dimension cannot visualize the higher dimensions
    - A bacteria which lives in a ~ 2D (tending to 2-D in our perspective- as flat as possible)- no understanding of what a human looks like- and even inside a human
    - BUT THE MOMENT WE INCREASE A DIMENSION- you see data like a never before- a new perspective to the DATA- and problems typically become much easier
    - https://www.youtube.com/watch?v=3liCbRZPrZA

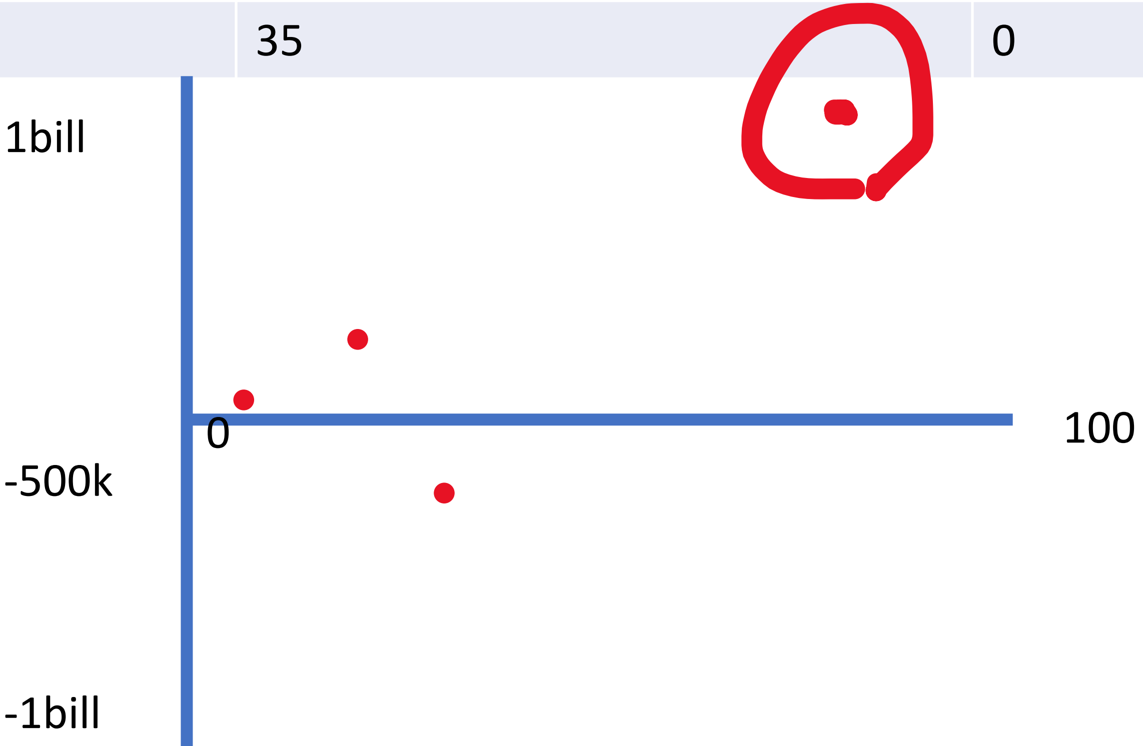
Image

- Space imaging- bandwidths – we could not even see!
  - Infrared – imaging technique – medical science, underwater, satellite and spaces (y) -> where y -> [0,255]
    - Once infrared is collect-> Z-Scoring -> data normalization on a scale that humans can see!

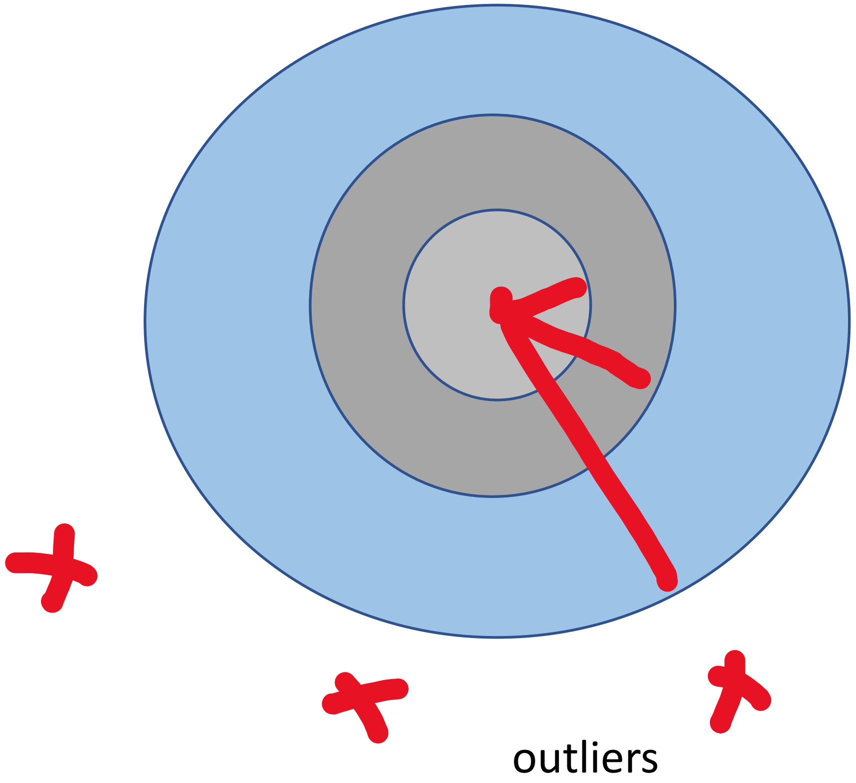
- Image is a 2-D array of pixels, where each pixel is built of (R, G, B)
  - Pixel-> (0,0,0) (255,255,255)
      - 2 types:
        - Light colors- if all light colors combine -> WHITE
        - Solid colors – if all solid colors combine -> BLACK

Z-Score

Money	Age	Happiness
$1,000,000,000	70	0
$100	5	1
$150,000	25	1
-$500,000	35	0



Innermost circle radius- standard deviation (sigma)
Centre – average value (mu)



Z-Score =.   How far away are you from center? And what's It's ratio with standard deviation?

=.   Any given point X,

Z-score $= \frac{mu - x}{sigma}$

Z-SCORE EXAMPLE

Age	Money	AgeNorm	MoneyNorm	mu1	50.8888889	sigma1	41.6906598
0	1000	1.220630453	0.33446663	mu2	1114471318	sigma2	3332082236
10	1E+10	0.980768573	-2.6666595				
100	150	-1.17798834	0.33446689				
25	6666666	0.620975754	0.33246618				
25	99999	0.620975754	0.33443692				
26	23423232	0.596989566	0.32743732				
72	23232	-0.50637508	0.33445996				
99	4354	-1.15400215	0.33446562				
101	23232	-1.20197453	0.33445996				

Max-abs scaling

- Formula:

$$\text{Max Abs Scalaing} = \frac{data\ -min}{\max - min}$$

Images-> min – 0, max – 255.0

$$\text{Scaling for images} = \frac{pixel\ -0}{255\ -0}$$