



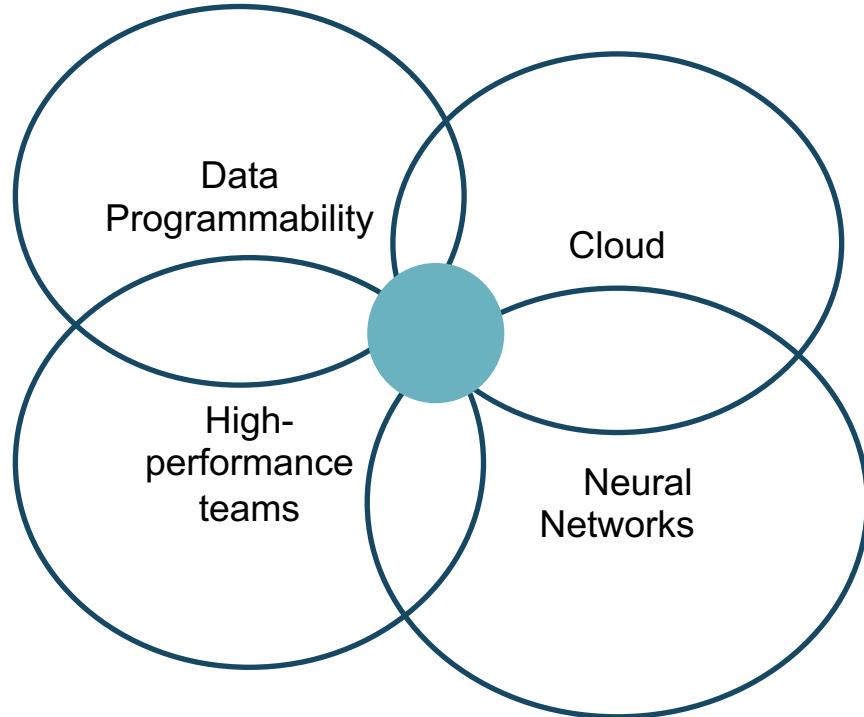
Lambda Architecture in the Cloud with Azure Databricks

Andrei Varanovich, InSpark

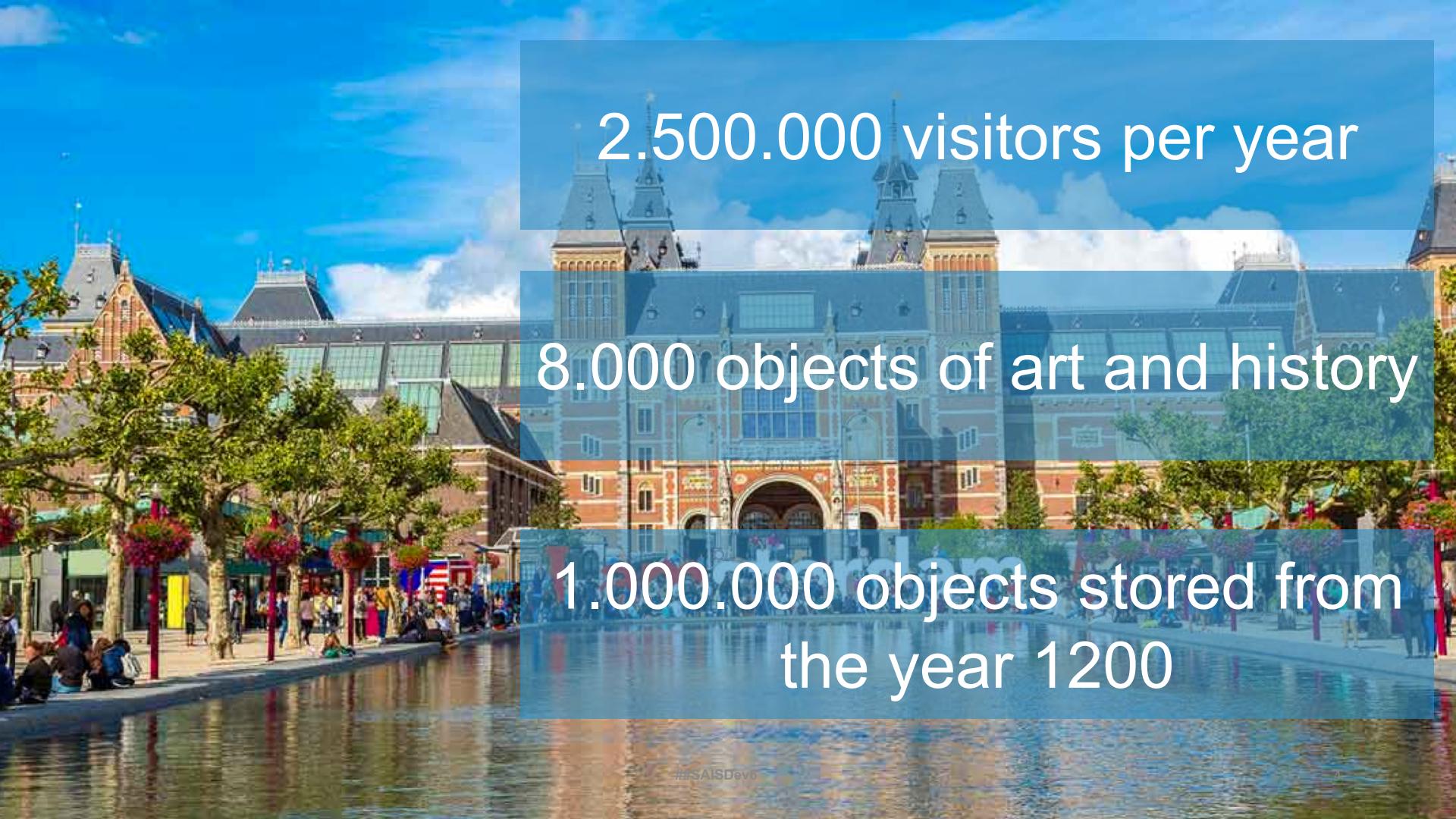
#SAISDev6



Selfie



**Big Data problem is many
small data problems**

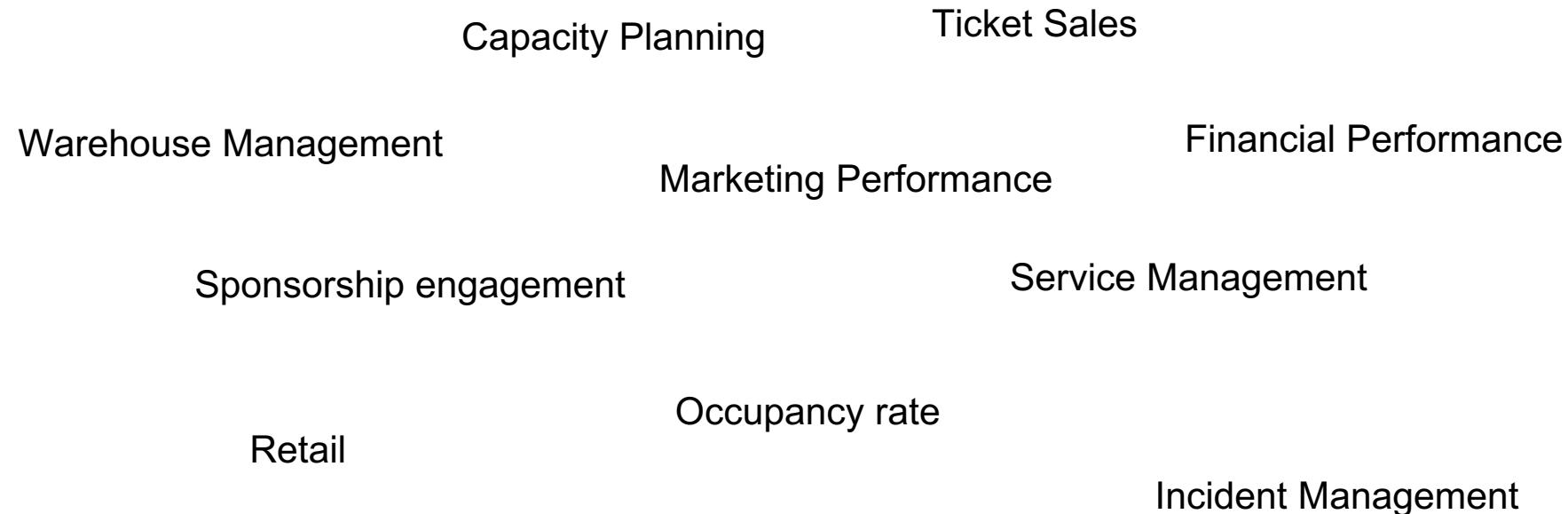


2.500.000 visitors per year

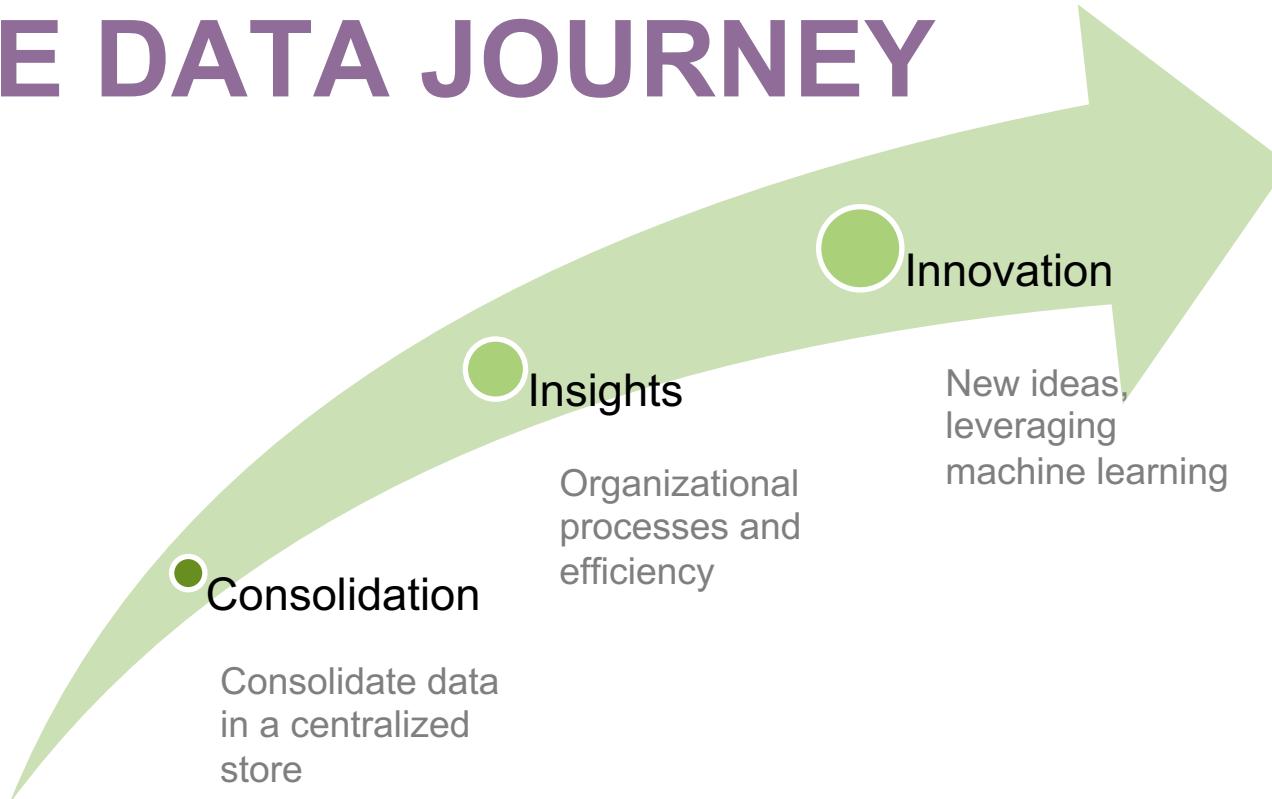
8.000 objects of art and history

1.000.000 objects stored from
the year 1200

Under the hood



THE DATA JOURNEY

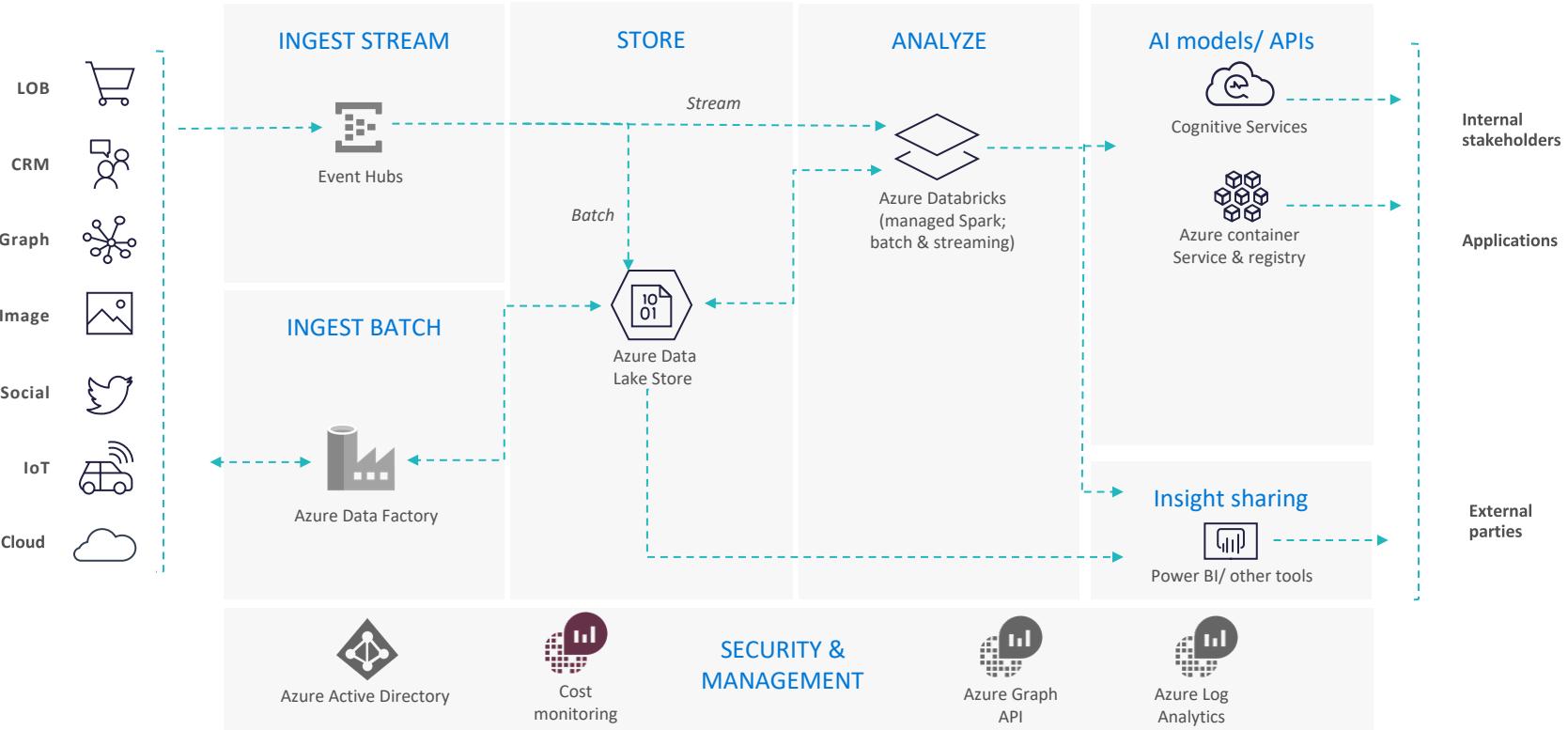


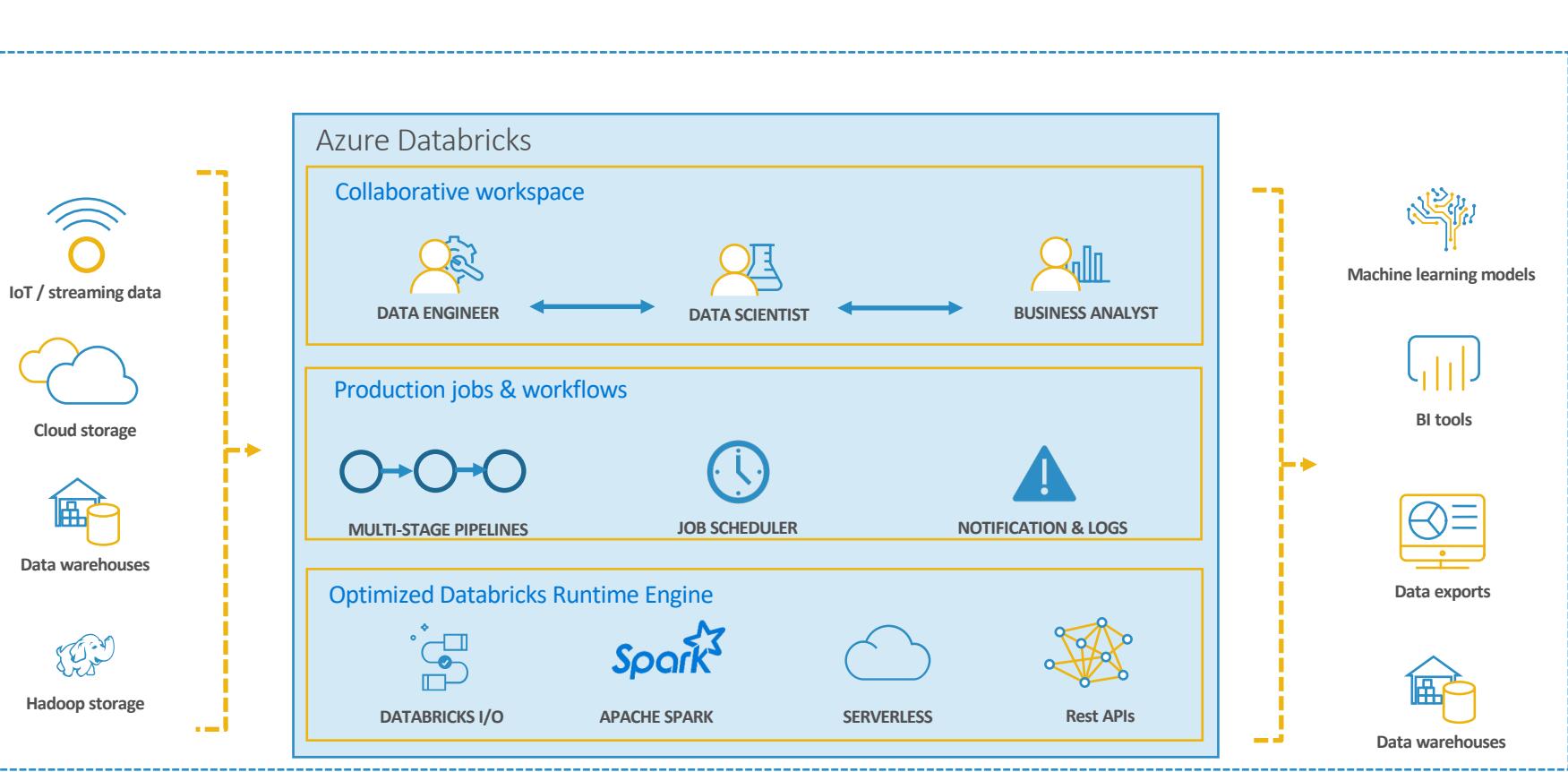
IN THE NEED FOR THE PLATFORM



We are in the need of the truly elastic data platform, to avoid any upfront planning, deployment and operations expenses, and put business value discovery first. The platform should support the [big]data projects in any stage, without the need to reengineer the whole solution.

Lambda Architecture on Azure





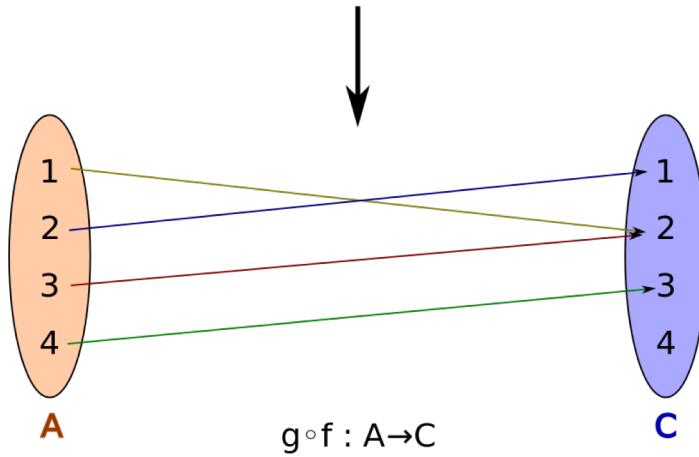
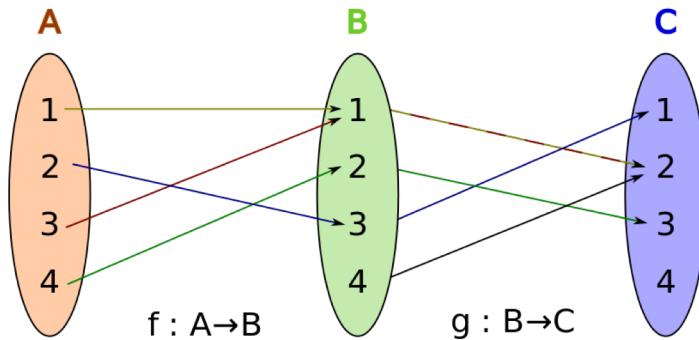
Simplicity is the ultimate sophistication

Leonardo da Vinci



LAMBDA TO THE RESCUE



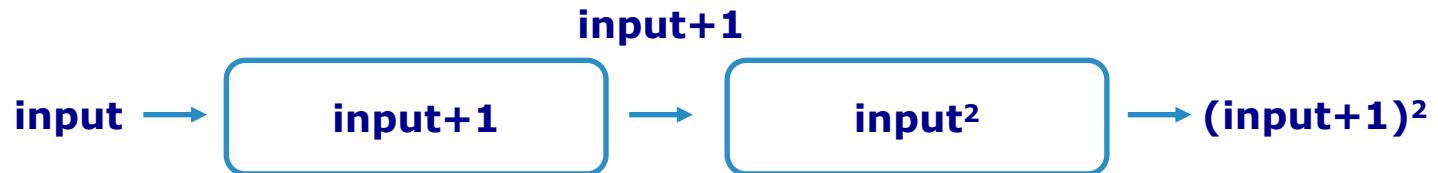


Composition of
functions is applying
one function to the
result of another

$$f(x) = x+1$$

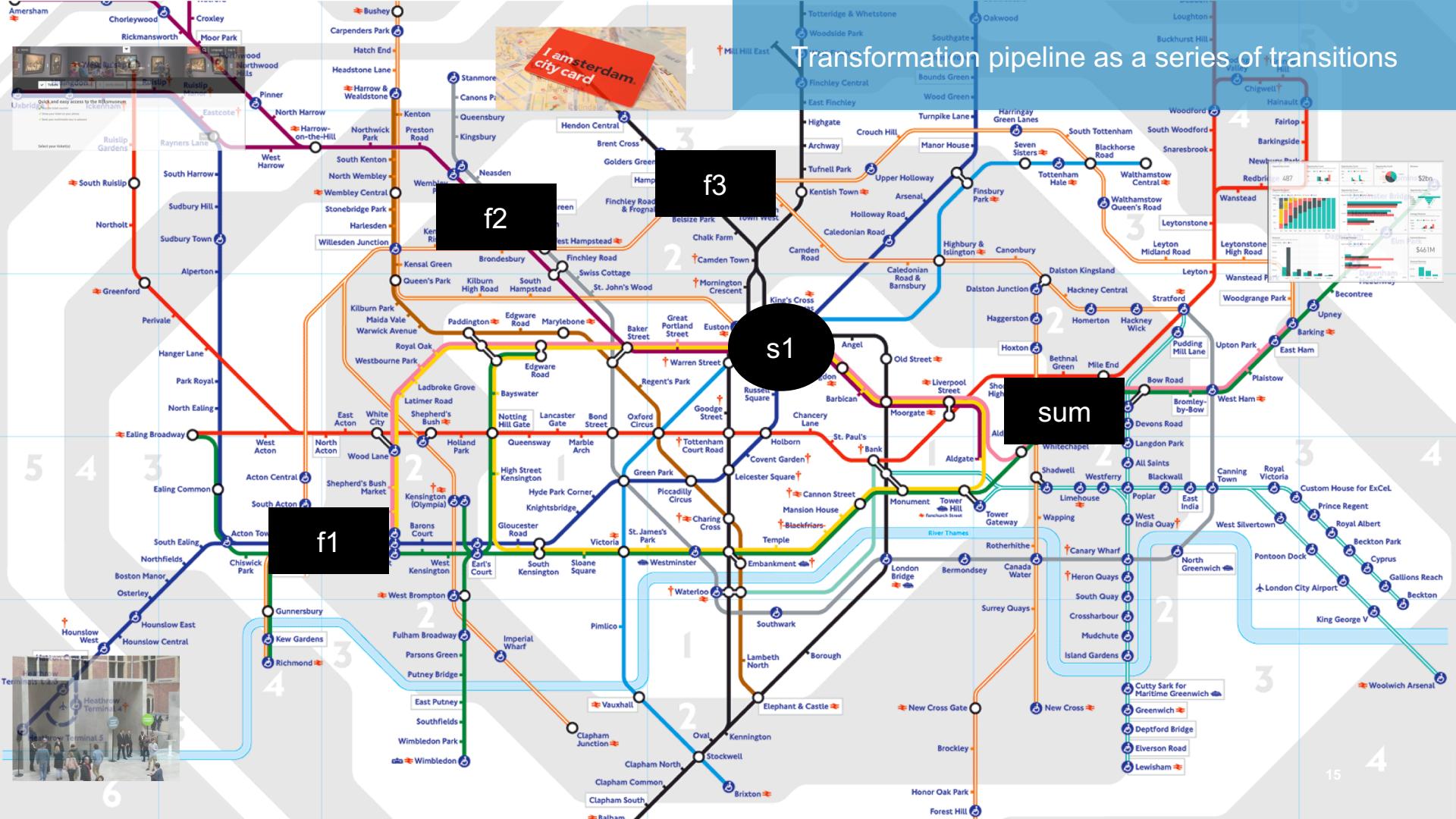
$$g(x) = x^2$$

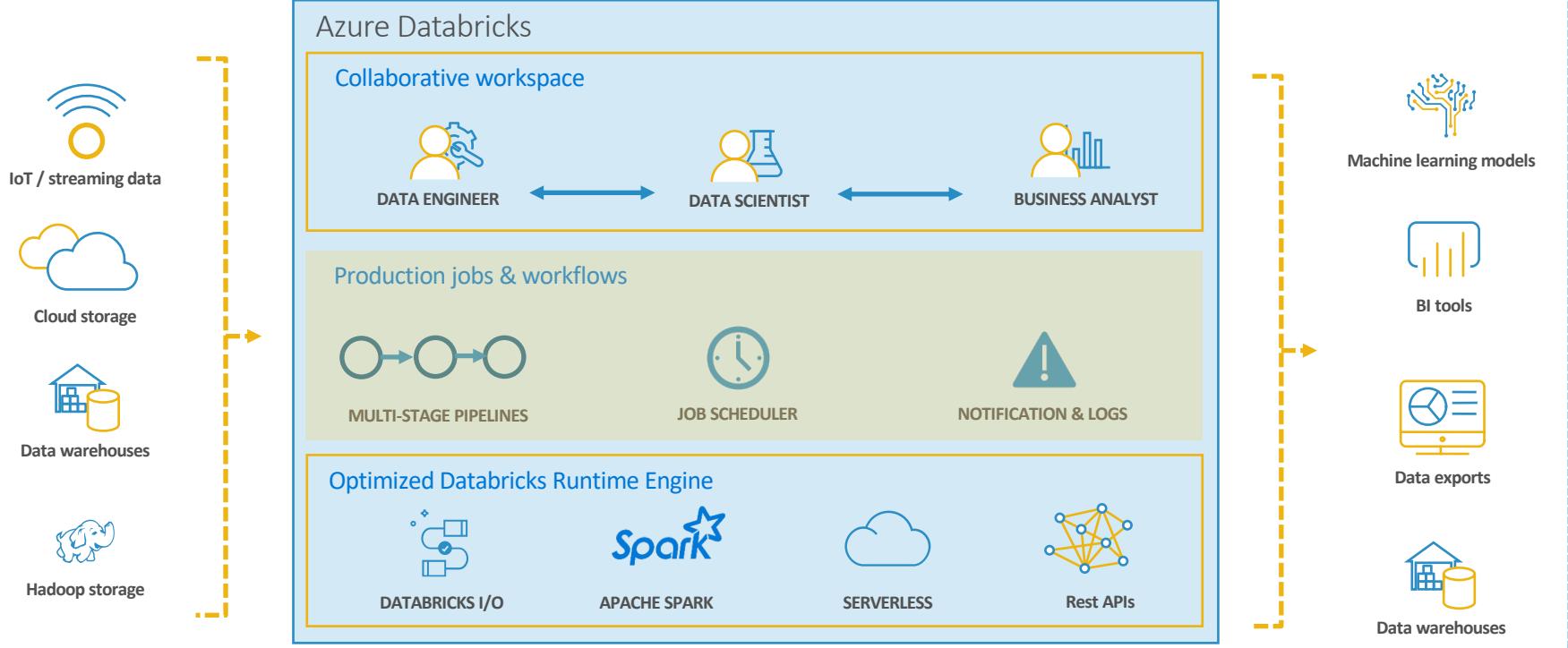
$$(g \circ f)(x) = g(f(x))$$



$$(g \circ f)(x) = (x+1)^2$$

Transformation pipeline as a series of transitions





Conclusions

... with proper design, the features come cheaply. This approach is arduous, but continues to succeed.

—Dennis Ritchie

- Standardization on Apache Spark allows us to move forward without introducing extra complexity.
- 100% PaaS offering is important – no need to maintain the infrastructure. All components we use offered as PaaS on Azure.
- Data pipelines as function composition allows us to ensure end-to-end consistency and spot the errors quickly.
- Saving intermediate states allows to quickly inspect the data sets.

Thank you!

Questions?