Lambda Architecture in the Cloud with Azure Databricks

Andrei Varanovich, InSpark

#SAISDev6

Selfie

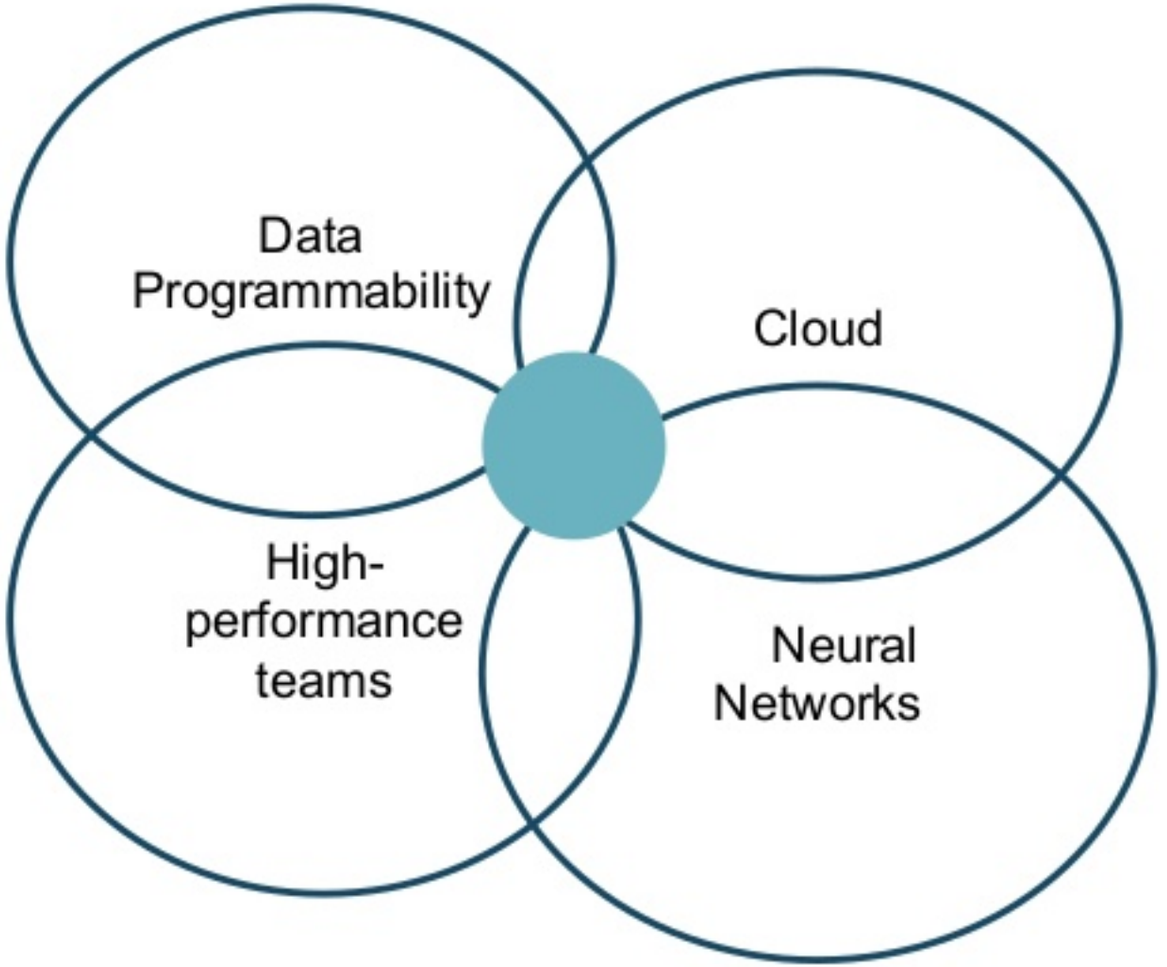
INSPARK

Data & AI Lead

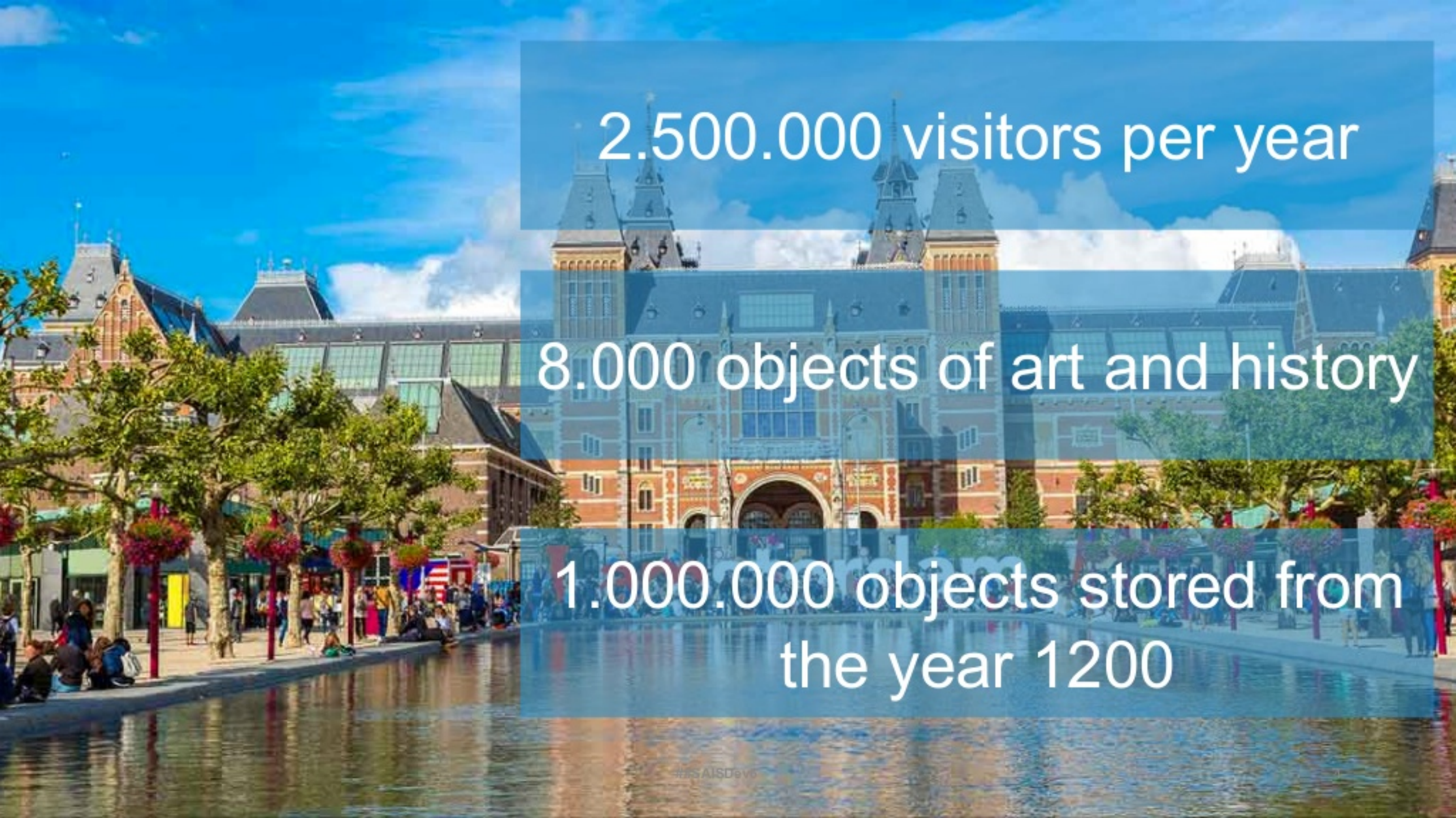
@DrGigabit

andrei.varanovich

andrei.varanovich@inspark.nl



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

Capacity Planning

Ticket Sales

Warehouse Management

Financial Performance

Marketing Performance

Sponsorship engagement

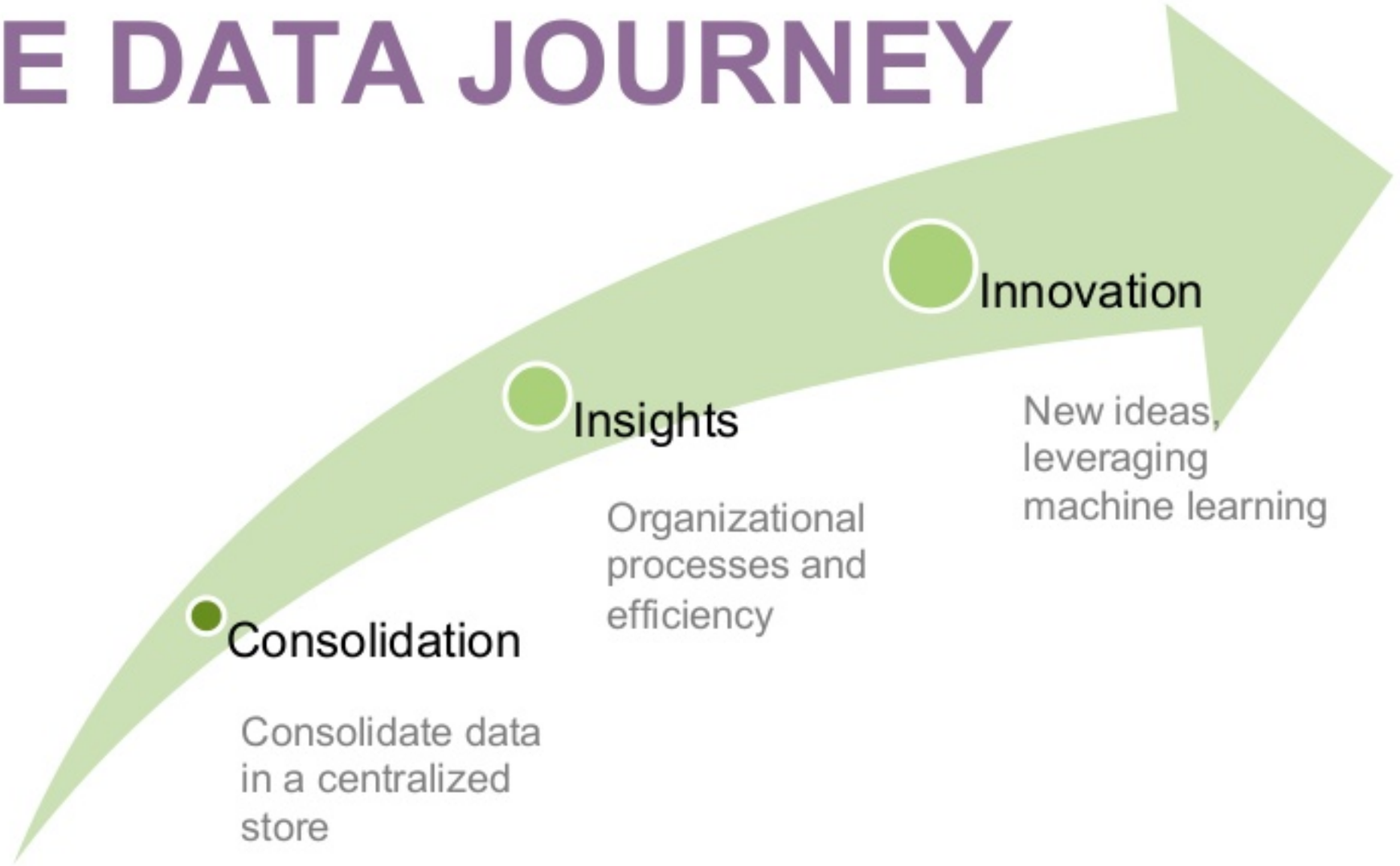
Service Management

Occupancy rate

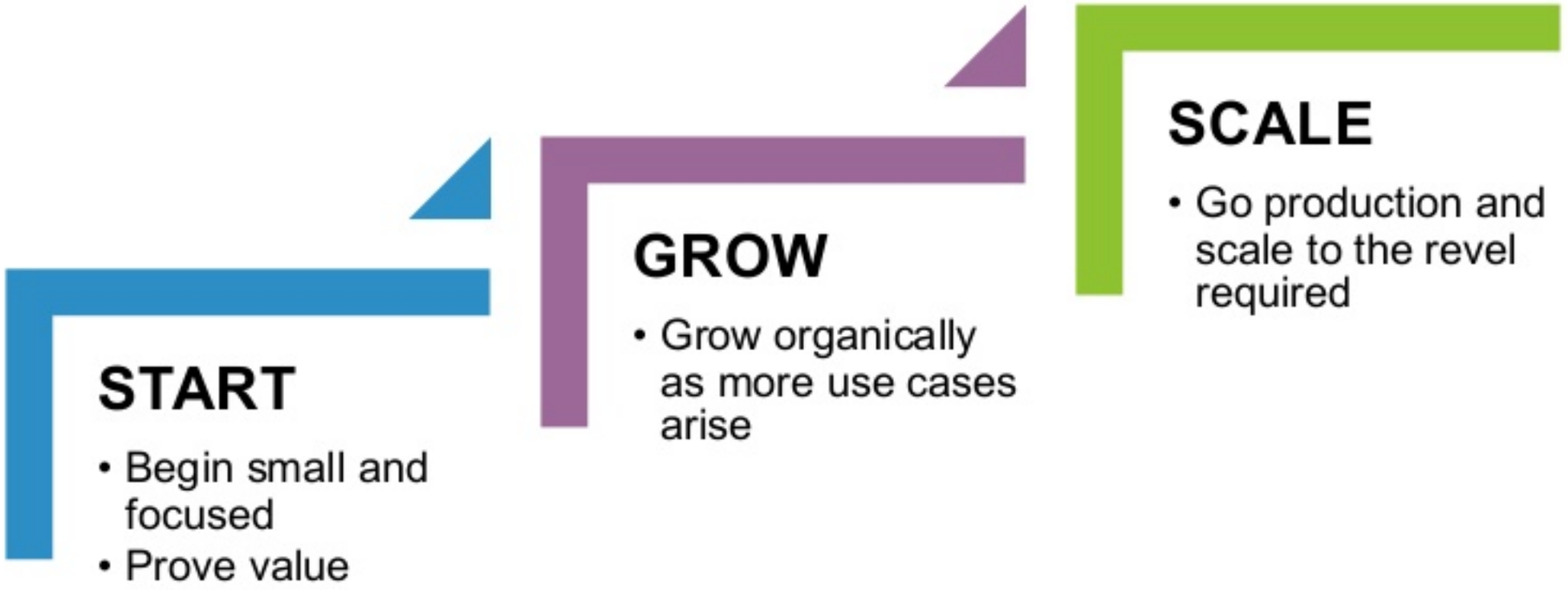
Retail

Incident Management

THE DATA JOURNEY

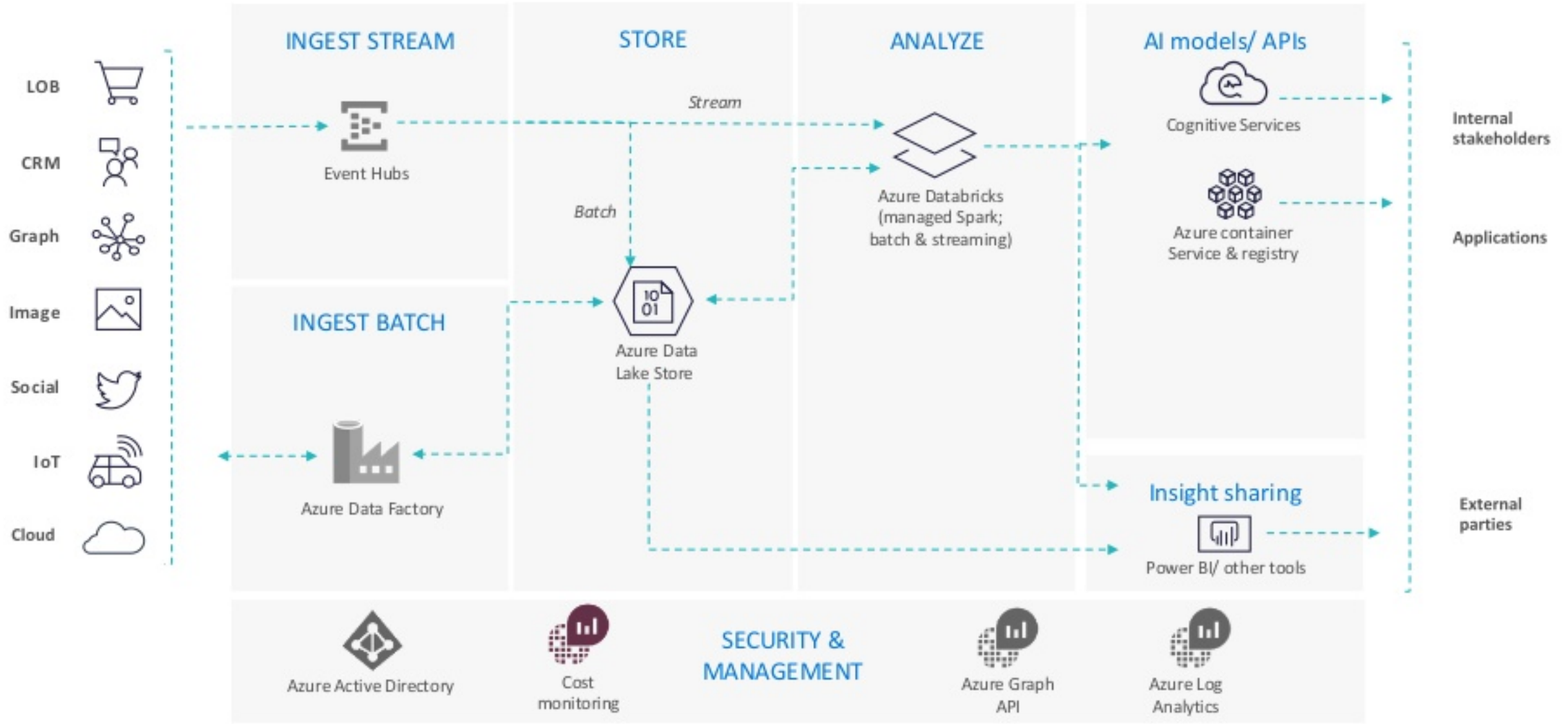


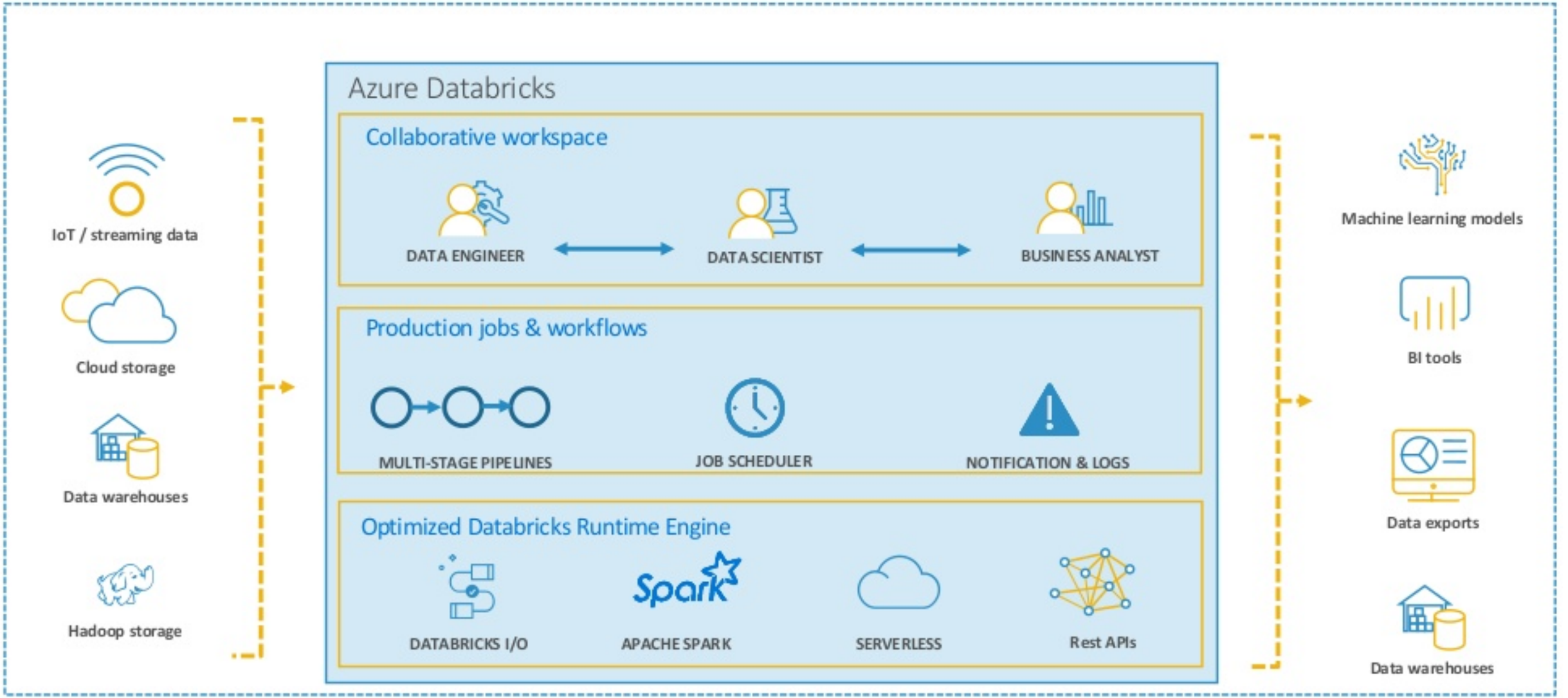
IN THE NEED FOR THE PLATFORM



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.

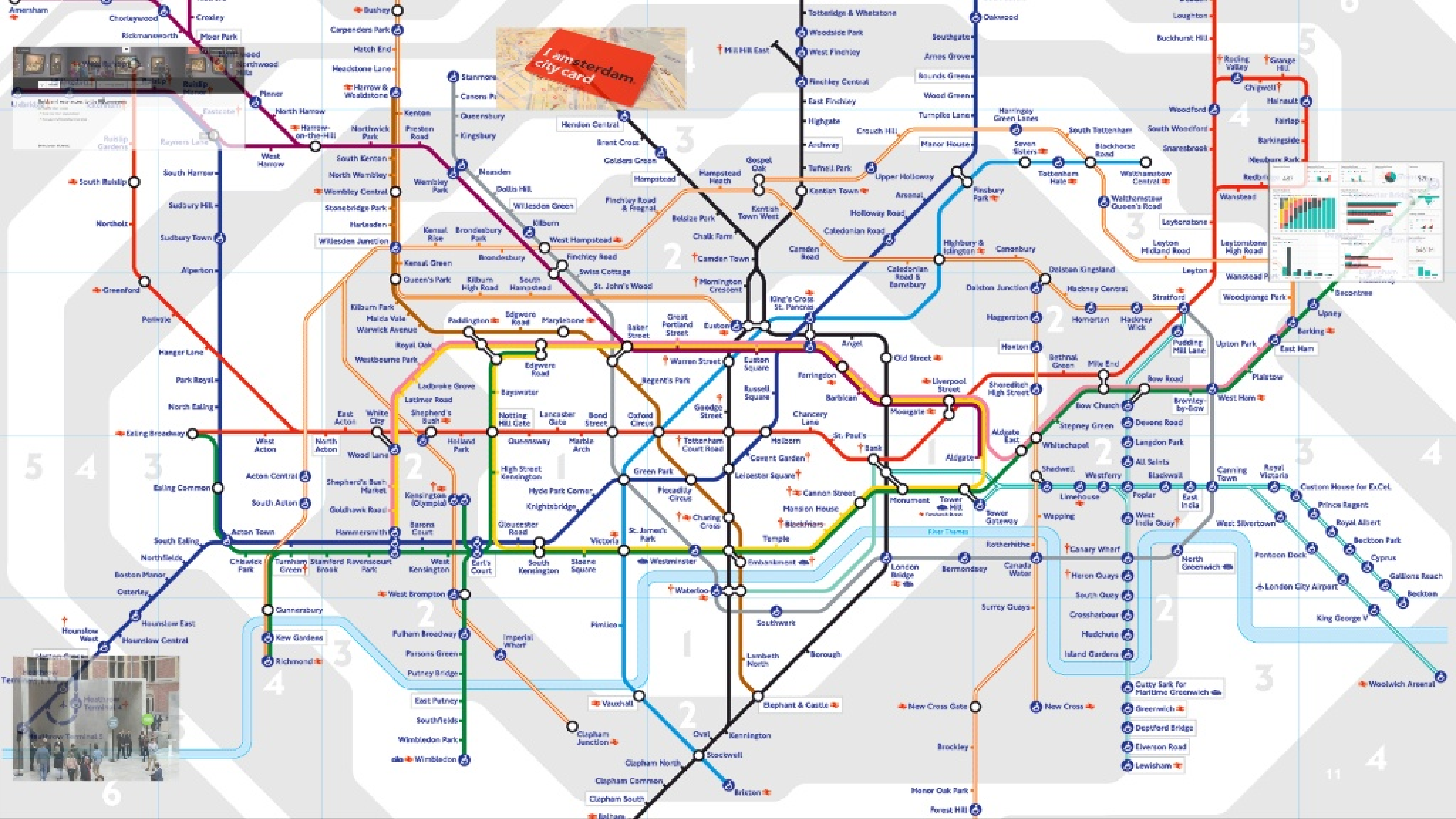
Lambda Architecture on Azure




IoT / streaming data
Cloud storage
Data warehouses
Hadoop storage
Azure Databricks
Collaborative workspace
DATA ENGINEER
DATA SCIENTIST
BUSINESS ANALYST
Production jobs & workflows
MULTI-STAGE PIPELINES
JOB SCHEDULER
NOTIFICATION & LOGS
Optimized Databricks Runtime Engine
DATABRICKS I/O
APACHE SPARK
SERVERLESS
Rest APIs
Machine learning models
BI tools
Data exports
Data warehouses

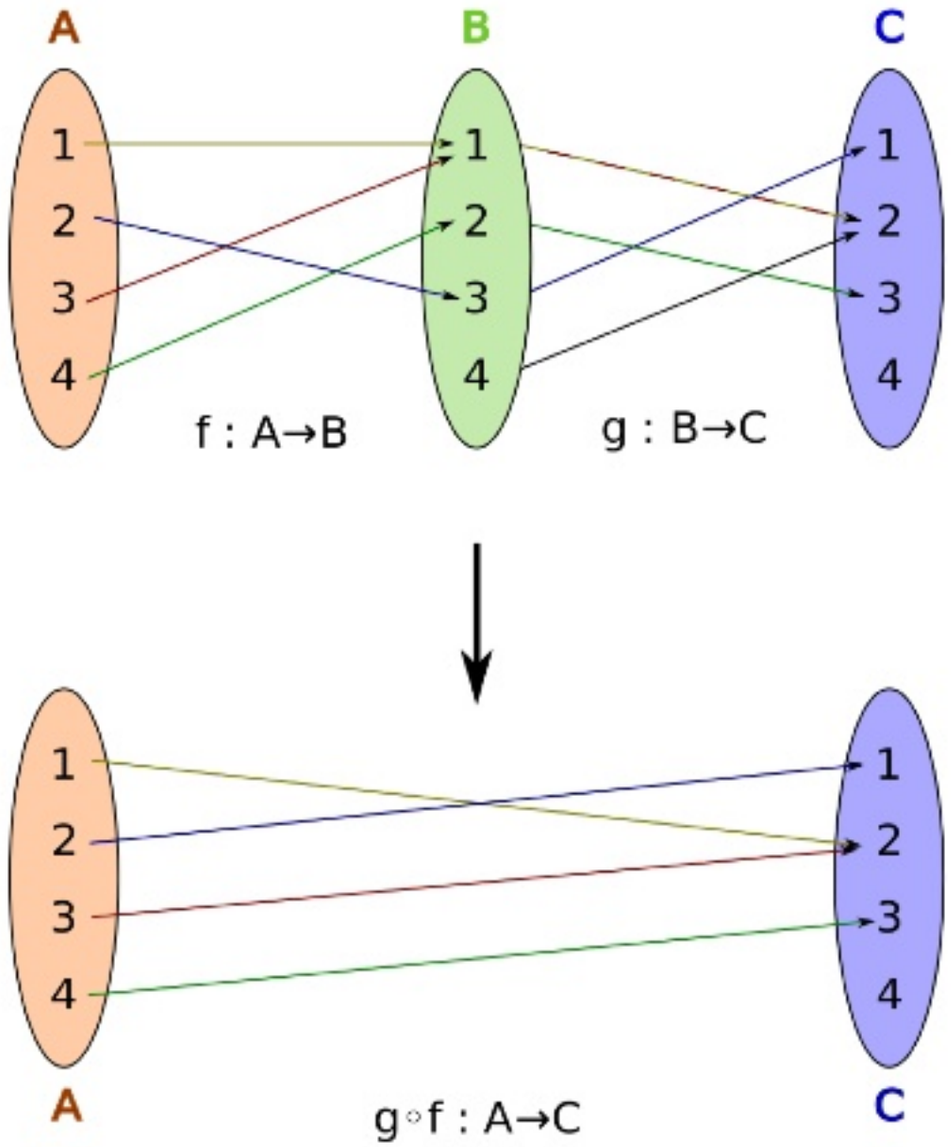
Simplicity is the ultimate sophistication

Leonardo da Vinci



Amersham
Chorleywood
Rickmansworth
Croxley
Moor Park
Northwood
Northwood Hills
Pinner
North Harrow
Harrow-on-the-Hill
West Harrow
Eastcote
Rayners Lane
Ruislip Gardens
South Ruislip
Northolt
Greenford
Perivale
Hanger Lane
Park Royal
North Ealing
Ealing Broadway
Ealing Common
Acton Town
South Ealing
Northfields
Boston Manor
Osterley
Hounslow East
Hounslow Central
Hounslow West
South Harrow
Sudbury Hill
Sudbury Town
Alperton
Bushey
Carpenders Park
Hatch End
Headstone Lane
Harrow & Wealdstone
Kenton
South Kenton
North Wembley
Wembley Central
Stonebridge Park
Harlesden
Willesden Junction
Kensal Green
Queen's Park
Kilburn Park
Maida Vale
Warwick Avenue
Paddington
Royal Oak
Westbourne Park
Ladbroke Grove
Latimer Road
Shepherd's Bush
White City
East Acton
West Acton
North Acton
Wood Lane
Acton Central
South Acton
Shepherd's Bush Market
Goldhawk Road
Hammersmith
Barons Court
Chiswick Park
Turnham Green
Stamford Brook
Ravenscourt Park
West Kensington
Gunnersbury
Kew Gardens
Richmond
West Brompton
Fulham Broadway
Parsons Green
Putney Bridge
East Putney
Southfields
Wimbledon Park
Wimbledon
Kensington (Olympia)
Holland Park
Notting Hill Gate
Queensway
Lancaster Gate
Marble Arch
Bond Street
Oxford Circus
Tottenham Court Road
Holborn
Chancery Lane
St. Paul's
Bank
High Street Kensington
Hyde Park Corner
Knightsbridge
Gloucester Road
South Kensington
Sloane Square
Earl's Court
Victoria
St. James's Park
Westminster
Embankment
Green Park
Piccadilly Circus
Leicester Square
Covent Garden
Charing Cross
Temple
Blackfriars
Mansion House
Cannon Street
Monument
Tower Hill
Tower Gateway
Waterloo
Southwark
London Bridge
Borough
Lambeth North
Elephant & Castle
Kennington
Oval
Stockwell
Brixton
Vauxhall
Pimlico
Imperial Wharf
Clapham Junction
Clapham North
Clapham Common
Clapham South
Balham
Stanmore
Canons Park
Queensbury
Kingsbury
Wembley Park
Neasden
Dollis Hill
Willesden Green
Kilburn
West Hampstead
Finchley Road
Swiss Cottage
St. John's Wood
Baker Street
Marylebone
Edgware Road
Bayswater
Northwick Park
Preston Road
Kensal Rise
Brondesbury Park
Brondesbury
Kilburn High Road
South Hampstead
Hendon Central
Brent Cross
Golders Green
Hampstead
Finchley Road & Frognal
Hampstead Heath
Gospel Oak
Belsize Park
Chalk Farm
Camden Town
Mornington Crescent
Euston
Euston Square
Warren Street
Great Portland Street
Regent's Park
Goodge Street
Russell Square
King's Cross St. Pancras
Angel
Old Street
Farringdon
Barbican
Moorgate
Liverpool Street
Aldgate
Aldgate East
Totteridge & Whetstone
Woodside Park
West Finchley
Mill Hill East
Finchley Central
East Finchley
Highgate
Archway
Tufnell Park
Kentish Town
Kentish Town West
Camden Road
Caledonian Road
Caledonian Road & Barnsbury
Holloway Road
Arsenal
Upper Holloway
Crouch Hill
Oakwood
Southgate
Arnos Grove
Bounds Green
Wood Green
Turnpike Lane
Manor House
Harringay Green Lanes
Finsbury Park
Highbury & Islington
Canonbury
Seven Sisters
South Tottenham
Tottenham Hale
Blackhorse Road
Walthamstow Central
Walthamstow Queen's Road
Leytonstone
Leyton Midland Road
Leyton
Dalston Kingsland
Dalston Junction
Haggerston
Hoxton
Hackney Central
Homerton
Hackney Wick
Stratford
Shoreditch High Street
Bethnal Green
Mile End
Bow Road
Bow Church
Stepney Green
Whitechapel
Devons Road
Langdon Park
All Saints
Bromley-by-Bow
West Ham
Pudding Mill Lane
Shadwell
Limehouse
Westferry
Poplar
Blackwall
East India
West India Quay
Canary Wharf
Heron Quays
South Quay
Crossharbour
Mudchute
Island Gardens
Cutty Sark for Maritime Greenwich
Greenwich
Deptford Bridge
Elverson Road
Lewisham
Wapping
Rotherhithe
Canada Water
Surrey Quays
Bermondsey
North Greenwich
Canning Town
Royal Victoria
Custom House for ExCeL
Prince Regent
Royal Albert
Beckton Park
Cyprus
Gallions Reach
Beckton
West Silvertown
Pontoon Dock
London City Airport
King George V
Woolwich Arsenal
New Cross Gate
New Cross
Brockley
Honor Oak Park
Forest Hill
Loughton
Buckhurst Hill
Roding Valley
Grange Hill
Chigwell
Hainault
Fairlop
Barkingside
Newbury Park
Redbridge
Woodford
South Woodford
Snaresbrook
Wanstead
Leytonstone High Road
Wanstead Park
Woodgrange Park
Upton Park
East Ham
Plaistow
Barking
Upney
Becontree
River Thames
Heathrow Terminal 5
Heathrow Terminal 4
I amsterdam city card

LAMBDA TO THE RESCUE

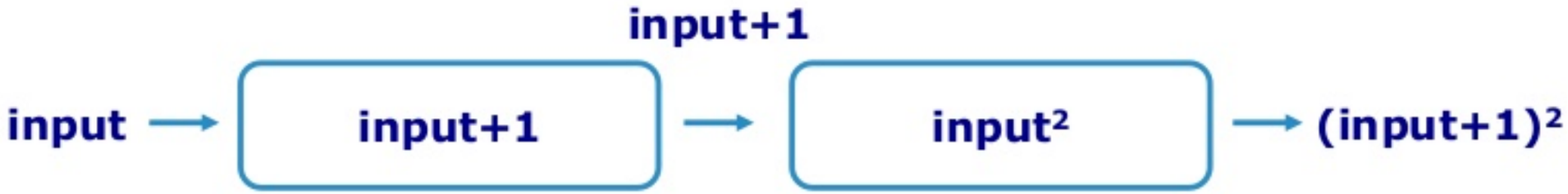


Composition of functions is applying one function to the result of another

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

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

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



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

Transformation pipeline as a series of transitions

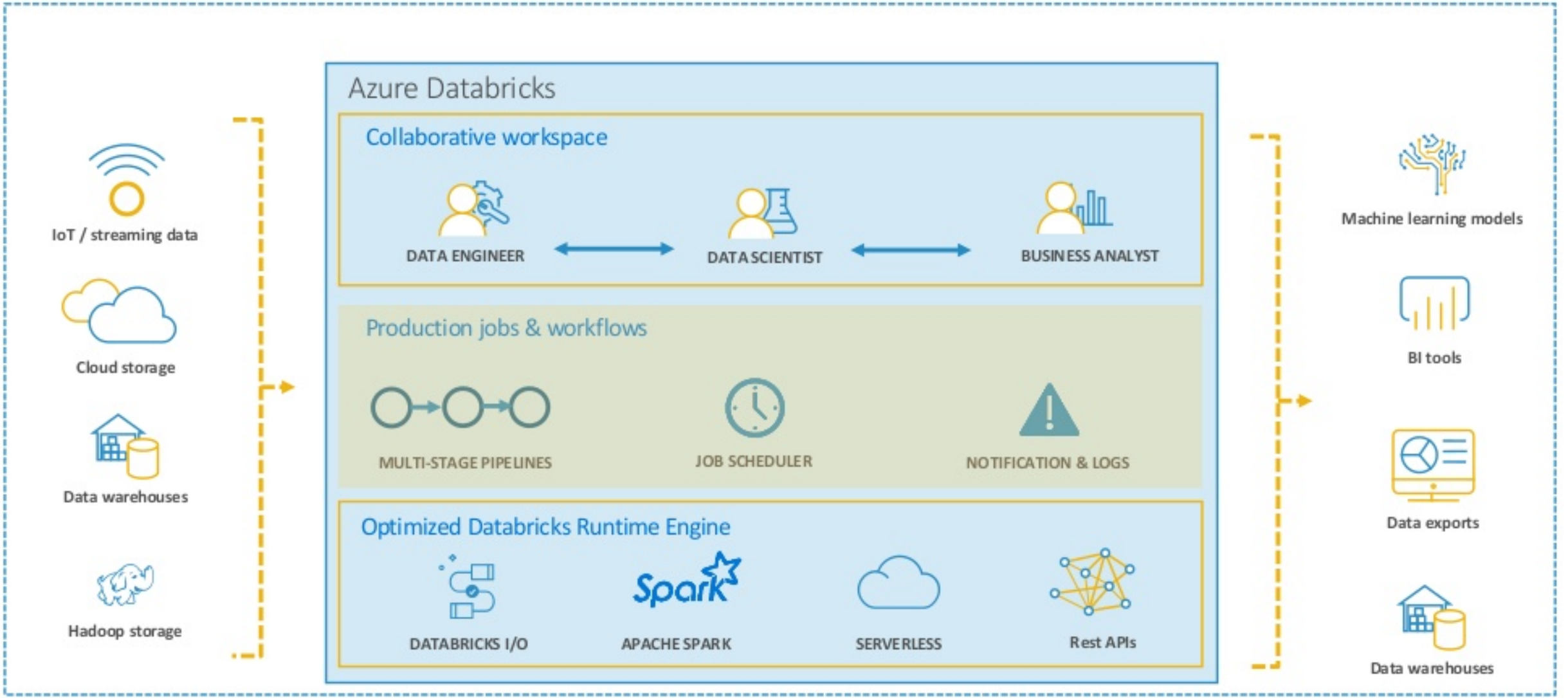
f1

f2

f3

s1

sum


IoT / streaming data
Cloud storage
Data warehouses
Hadoop storage
Azure Databricks
Collaborative workspace
DATA ENGINEER
DATA SCIENTIST
BUSINESS ANALYST
Production jobs & workflows
MULTI-STAGE PIPELINES
JOB SCHEDULER
NOTIFICATION & LOGS
Optimized Databricks Runtime Engine
DATABRICKS I/O
APACHE SPARK
SERVERLESS
Rest APIs
Machine learning models
BI tools
Data exports
Data warehouses

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?