Microsoft

Insert session code here

Building The Modern Data Warehouse with Azure SQL DWH, Spark & Power BI

Catalin Esanu
Cloud Solution Architect
Microsoft Israel

cesanu@Microsoft.com

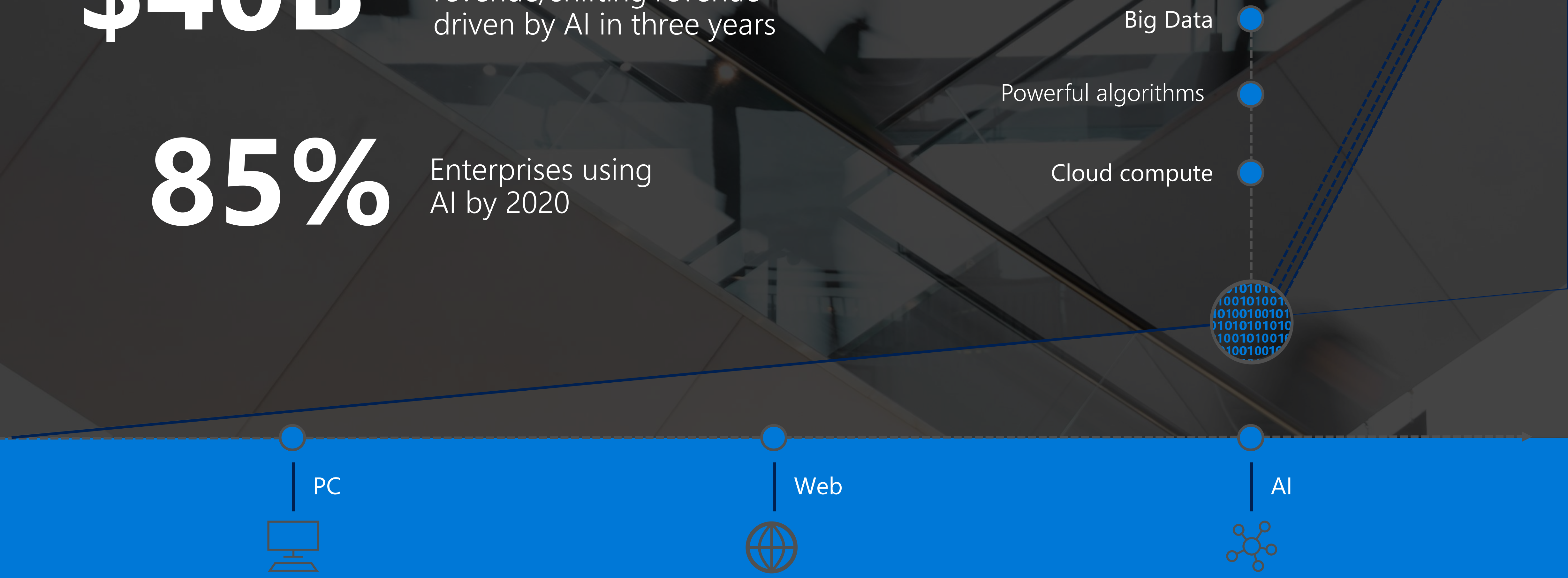
Who am I?

Use UserVoice and help us be better!

Why now?

$40B estimated additional revenue/shifting revenue driven by AI in three years

85% Enterprises using AI by 2020



Big Data

Powerful algorithms

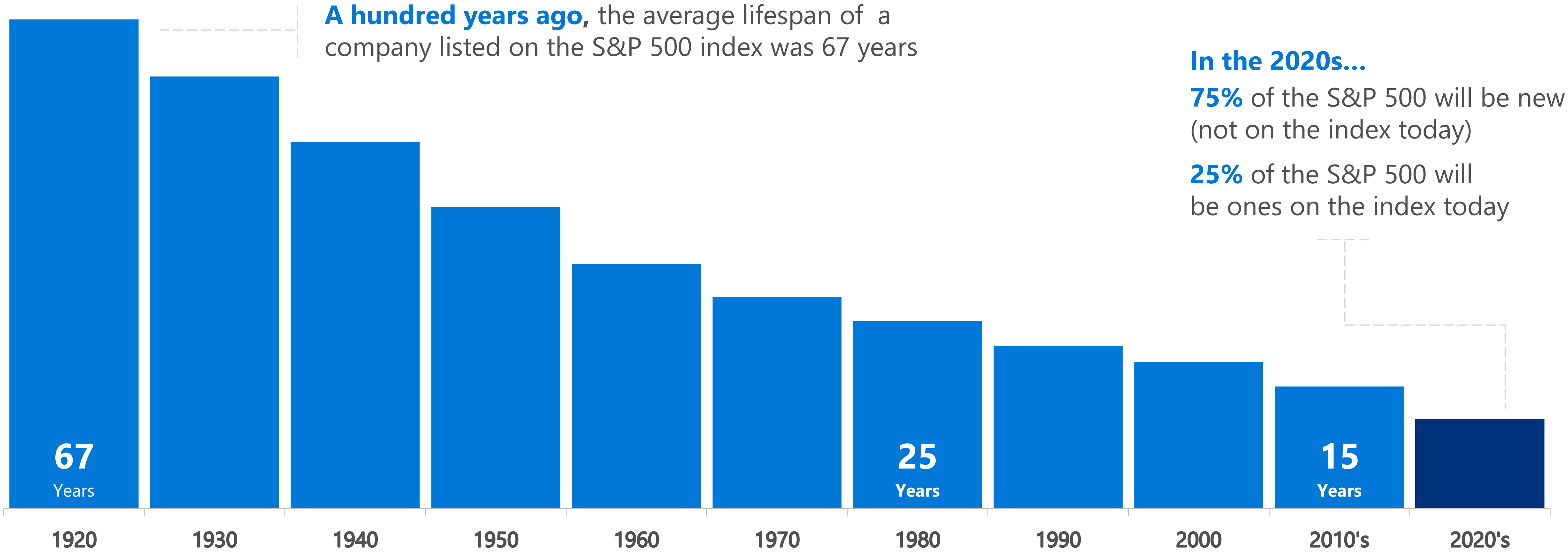
Cloud compute

PC

Web

AI

The time to adapt to disruptions is shrinking



A hundred years ago, the average lifespan of a company listed on the S&P 500 index was 67 years

In the 2020s…

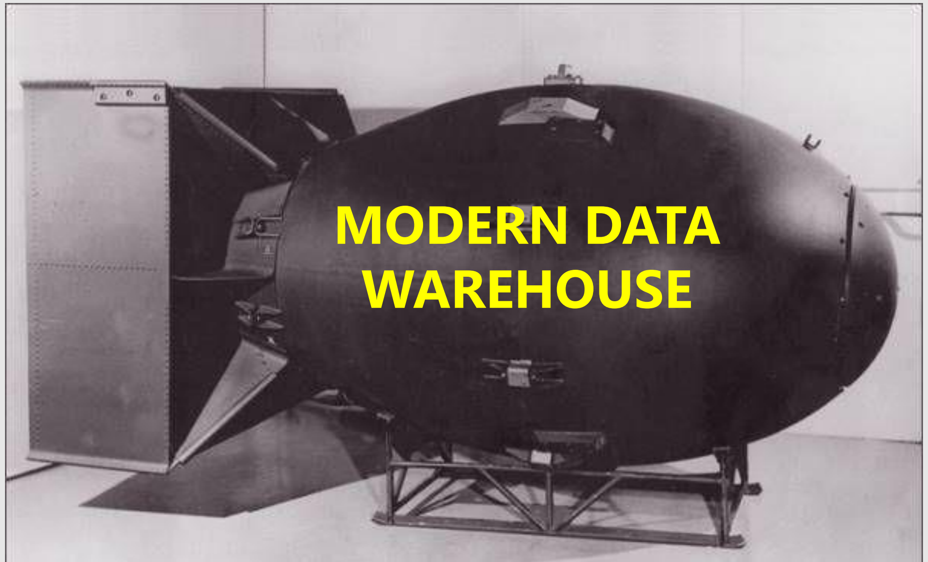
75% of the S&P 500 will be new (not on the index today)

25% of the S&P 500 will be ones on the index today

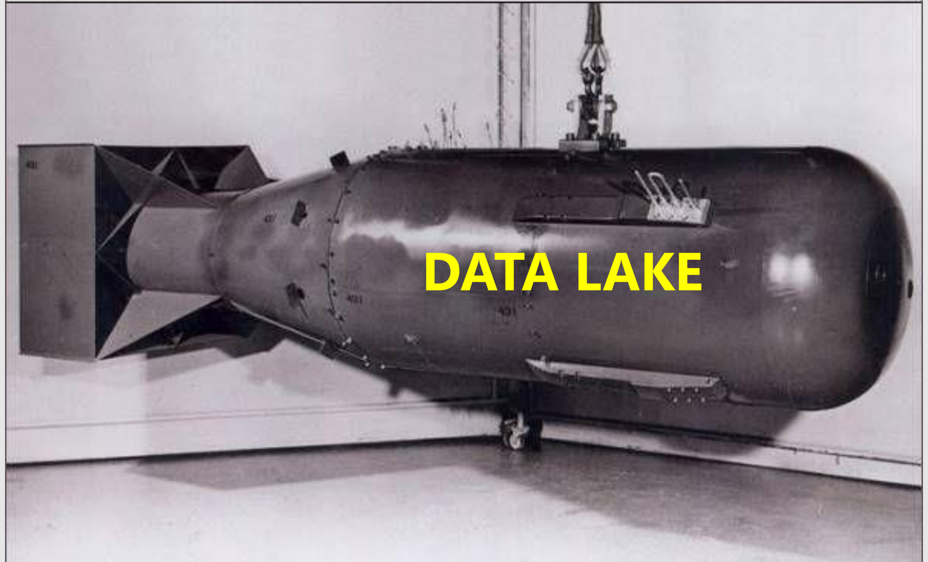
Source: BBC

Microsoft

BUZZWORD BOMBS!



Mockup of the original "Fat Man" weapon



A post-war "Little Boy" model.

What is a modern data warehouse?

Integrated Data Platform

Near Real-Time

Advanced Analytics

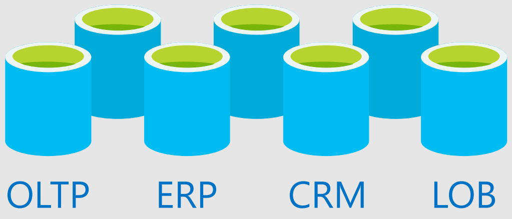
Multi-Structured Data

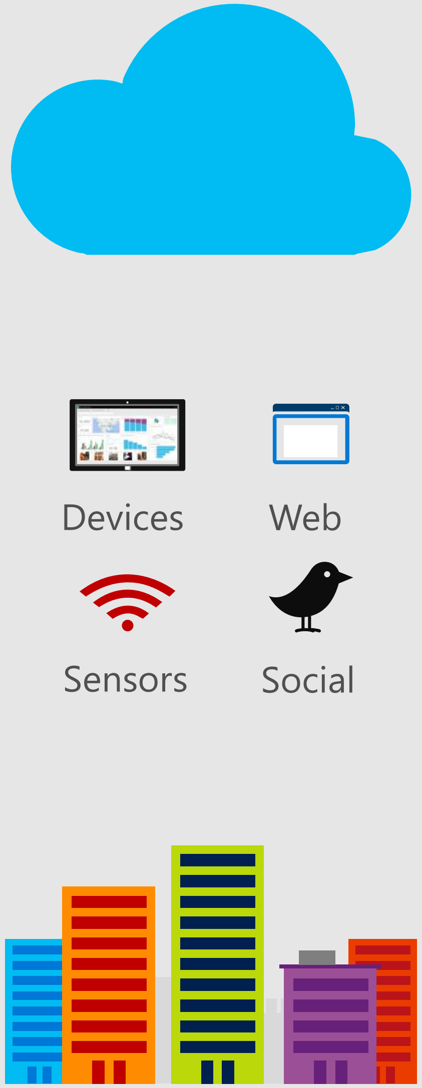
Performance

Scalable (Dynamic)

Customer Challenges in Data Warehousing

- Increased data types and volumes
- Varied data sources
- Added complexity and cost
- Technology confusion!

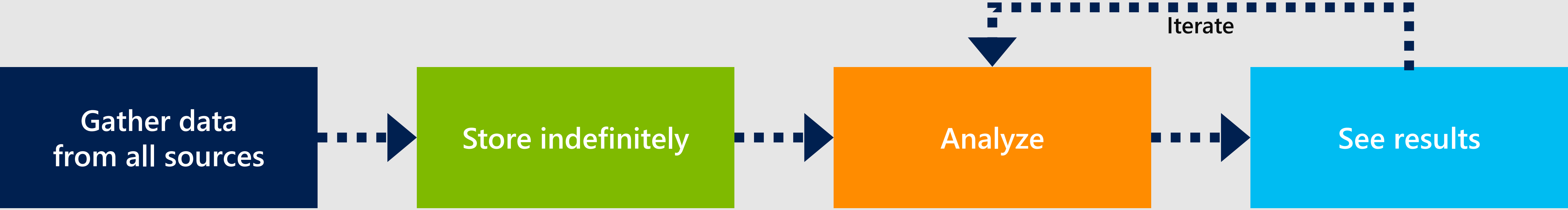

OLTP ERP CRM LOB


Devices Web

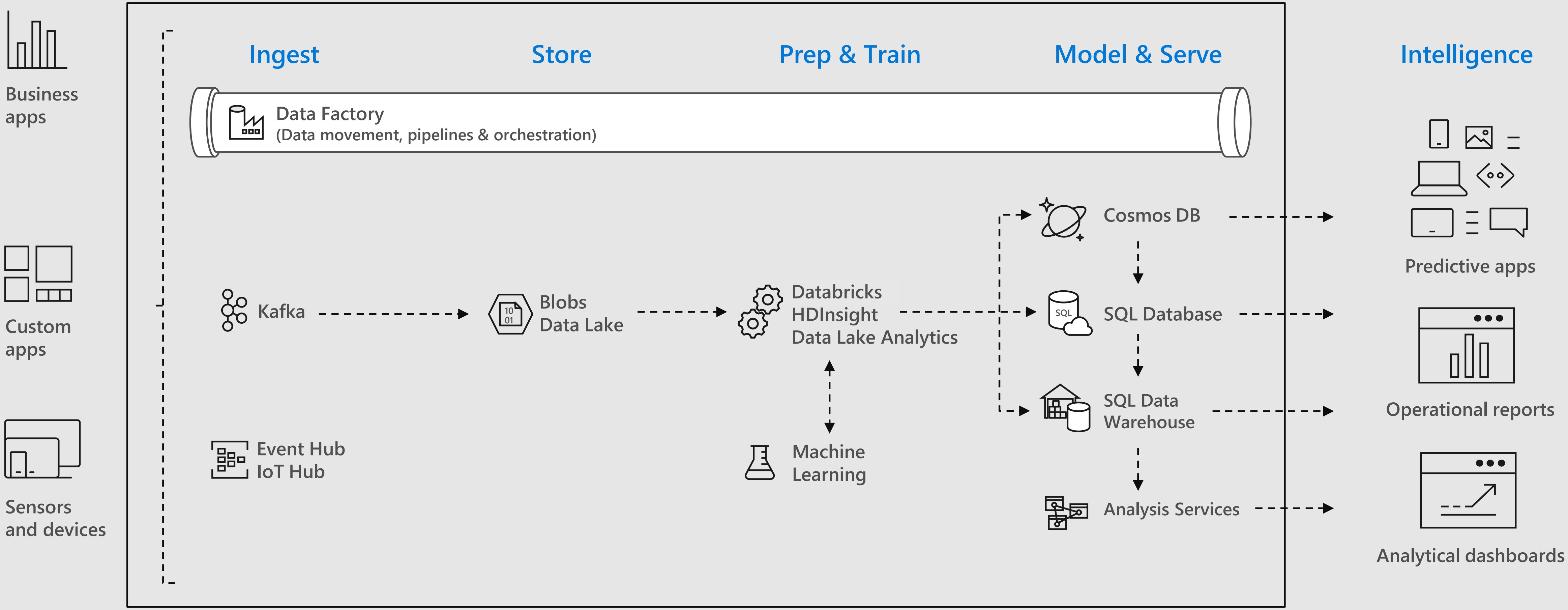
Sensors Social

New data thinking: all data has value!

- All data has potential value
- Data hoarding
- No defined schema—stored in native format
- Schema is imposed and transformations are done at query time *(schema-on-read).*
- Apps and users interpret the data as they see fit



BIG DATA & ADVANCED ANALYTICS AT A GLANCE

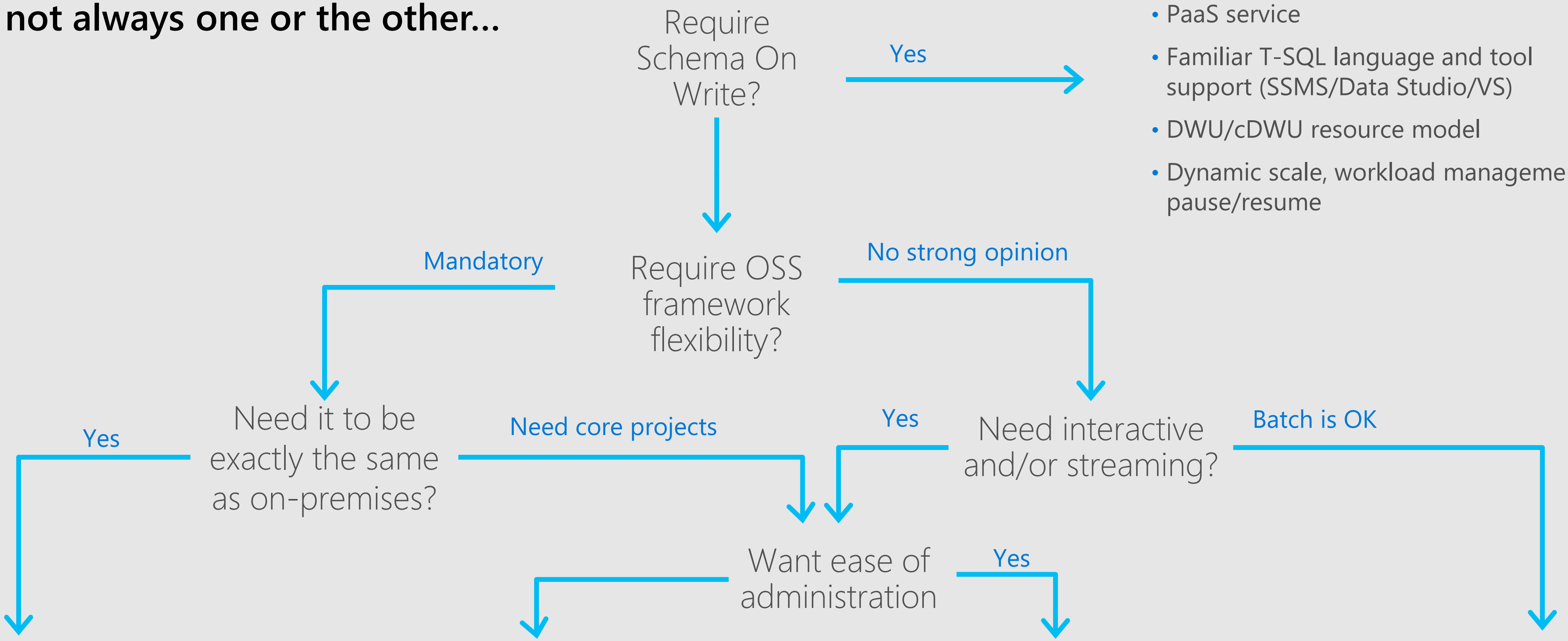


Demo 1: BRING IT!

Copy Data With Azure Data Factory

Use the right solution for the job

But it is not always one or the other…



Require Schema On Write?

Yes

Require OSS framework flexibility?

Mandatory

No strong opinion

Need it to be exactly the same as on-premises?

Yes

Need core projects

Need interactive and/or streaming?

Yes

Batch is OK

Want ease of administration

Yes

Azure SQL Data Warehouse

- PaaS service
- Familiar T-SQL language and tool support (SSMS/Data Studio/VS)
- DWU/cDWU resource model
- Dynamic scale, workload management, pause/resume

Azure Marketplace (IaaS)

- All workloads like on-premises
- Cloudera/MapR
- Hortonworks Sandbox or Standard
- Install variety of Apache frameworks

Azure HDInsight

- Variety of cluster types (Hadoop, Spark, IQ)
- Fully managed by Microsoft
- ADLS support
- Apache Ranger integration

Azure Databricks

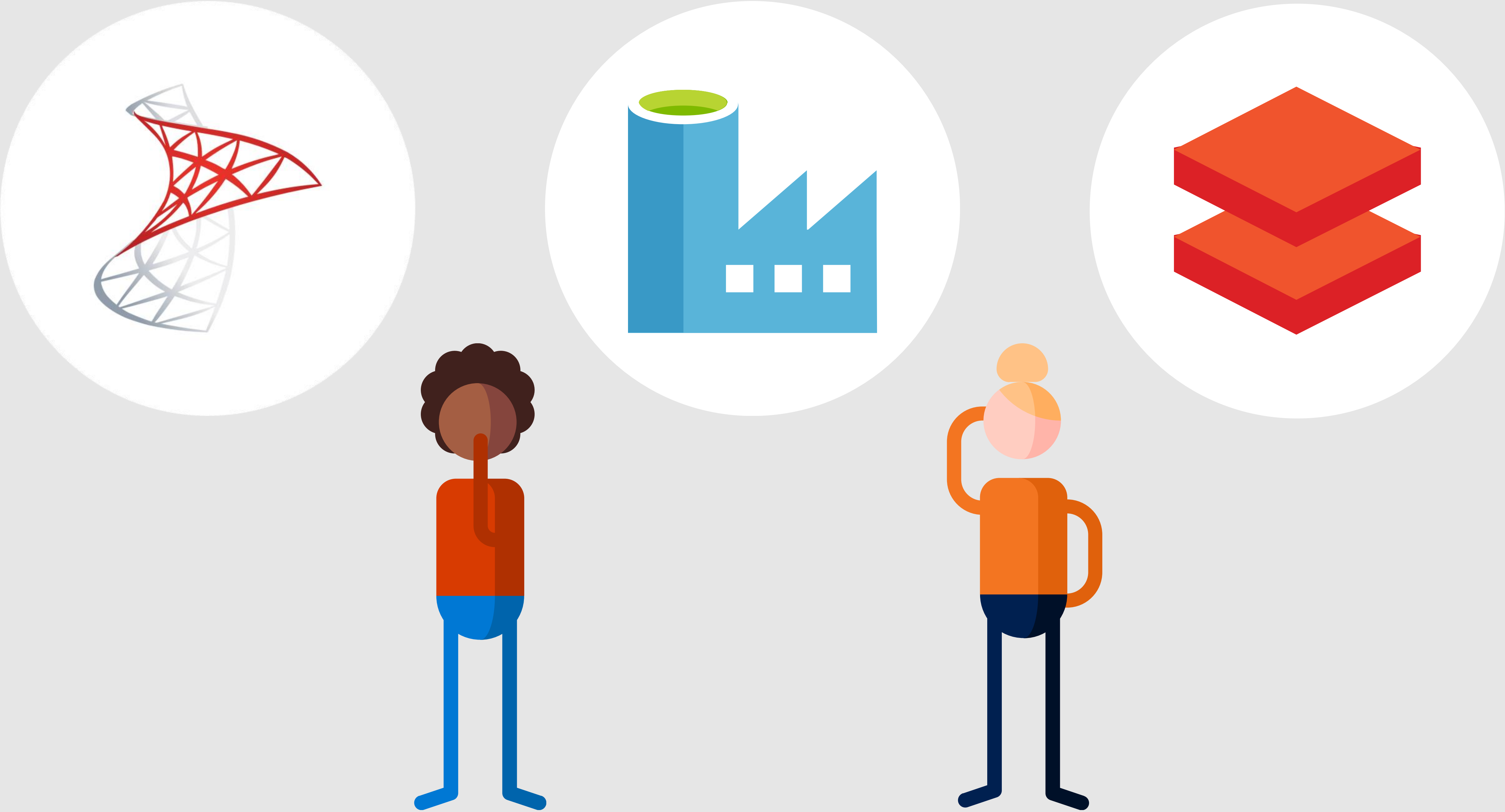
- Apache Spark-based engine
- Autoscale
- ADLS support
- First class integration with many data sources/destinations and BI tools

Azure Data Lake

- Easiest experience for admin: no sense of clusters, instant scale per job
- Easiest experience for developers: Visual Studio/U-SQL (C#+SQL)
- Batch workloads only

Technology Choices

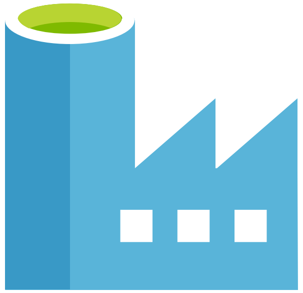
Ingestion



Data Integration

Extract, Transform, Load (ETL)

On-Premises



Data Movement & Orchestration

Extract, Load, Transform (ELT)

Hybrid

Apache Spark-based Analytics Service

Collaborative Notebooks

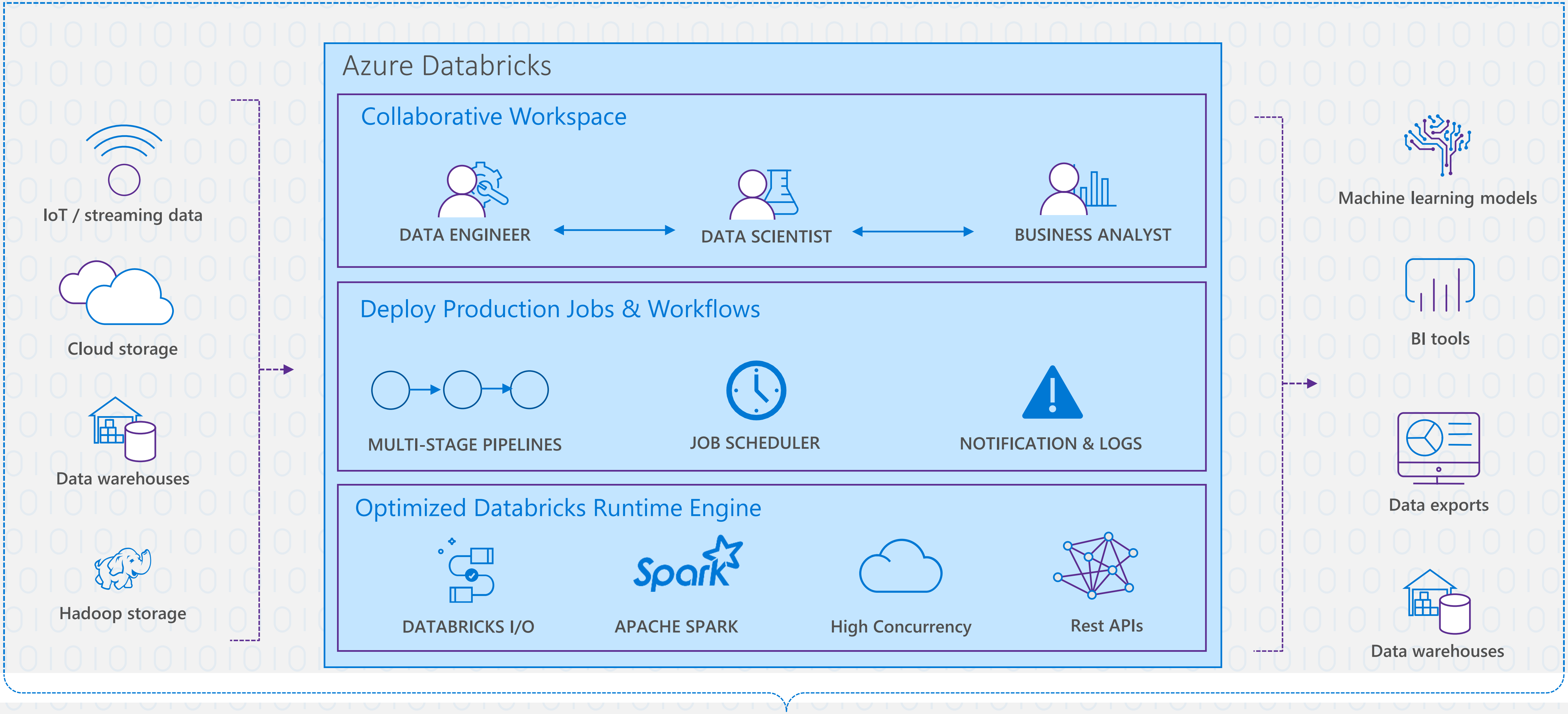
Cloud

Prep & Transform

AZURE DATABRICKS

- Azure Databricks is a **first party** service on Azure.
  - Unlike with other clouds, it is not an Azure Marketplace or a 3rd party hosted service.
- Azure Databricks is integrated seamlessly with Azure services:
  - Azure Portal: Service an be launched directly from Azure Portal
  - Azure Storage Services: Directly access data in Azure Blob Storage and Azure Data Lake Store
  - Azure Active Directory: For user authentication, eliminating the need to maintain two separate sets of uses in Databricks and Azure.
  - Azure SQL DW and Azure Cosmos DB: Enables you to combine structured and unstructured data for analytics
  - Apache Kafka for HDInsight: Enables you to use Kafka as a streaming data source or sink
  - Azure Billing: You get a single bill from Azure
  - Azure Power BI: For rich data visualization
- Eliminates need to create a separate account with Databricks.

AZURE DATABRICKS



IoT / streaming data

Cloud storage

Data warehouses

Hadoop storage

Azure Databricks

Collaborative Workspace

DATA ENGINEER ↔ DATA SCIENTIST ↔ BUSINESS ANALYST

Deploy Production Jobs & Workflows

MULTI-STAGE PIPELINES

JOB SCHEDULER

NOTIFICATION & LOGS

Optimized Databricks Runtime Engine

DATABRICKS I/O

APACHE SPARK

High Concurrency

Rest APIs

Machine learning models

BI tools

Data exports

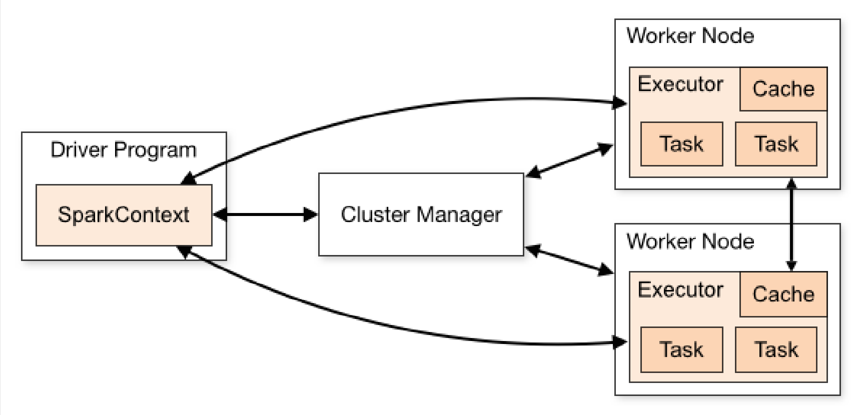
Data warehouses

Enhance Productivity

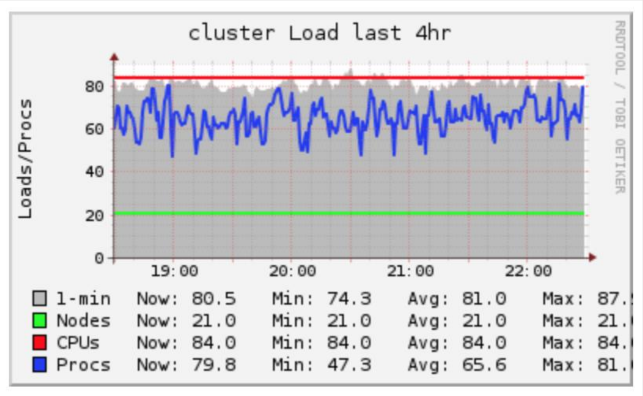
Build on secure & trusted cloud

Scale without limits

Infinite Scale, Lower Cost, Zero Management

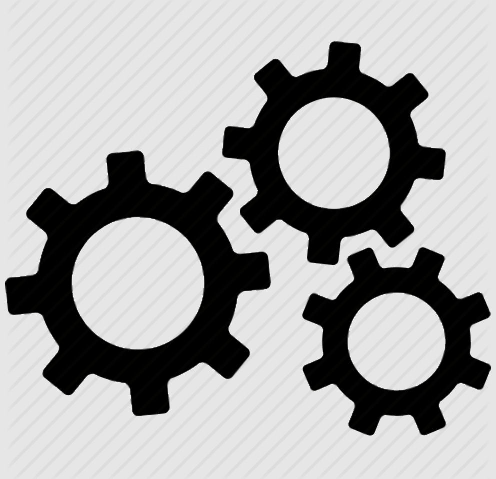


1 to 1000s of Worker Nodes



▨ 1-min	Now: 80.5	Min: 74.3	Avg: 81.0	Max: 87.
🟩 Nodes	Now: 21.0	Min: 21.0	Avg: 21.0	Max: 21.
🟥 CPUs	Now: 84.0	Min: 84.0	Avg: 84.0	Max: 84.
🟦 Procs	Now: 79.8	Min: 47.3	Avg: 65.6	Max: 81.

Auto-scale
Compute & Storage



Auto-Recovery & Upgrade

Demo 2: Massage Your Data

Process semi-structured data with Databricks

Azure Databricks

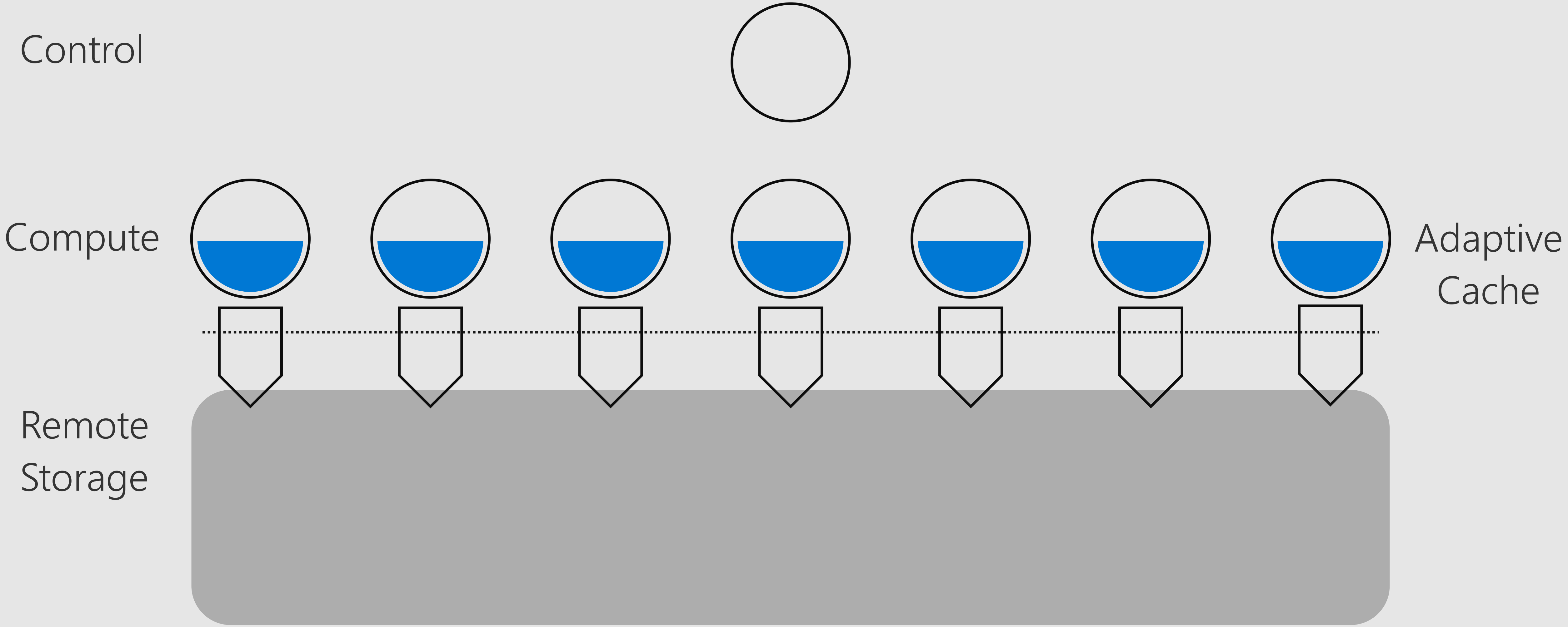
Performance Considerations

- Land data in Blob Store/ADLS partitioned into separate directories
- For best query performance use a Delta table. Alternatively, use a regular Spark table backed by Parquet
- Avoid small files. File size 100s MB – 1GB preferred
    - Delta supports compaction `OPTIMIZE events WHERE date >= '2017-01-01’`
    - Improve the speed of read queries from a table by coalescing small files into larger ones
- Use Secrets (AKV or DB backed)

Model & Serve

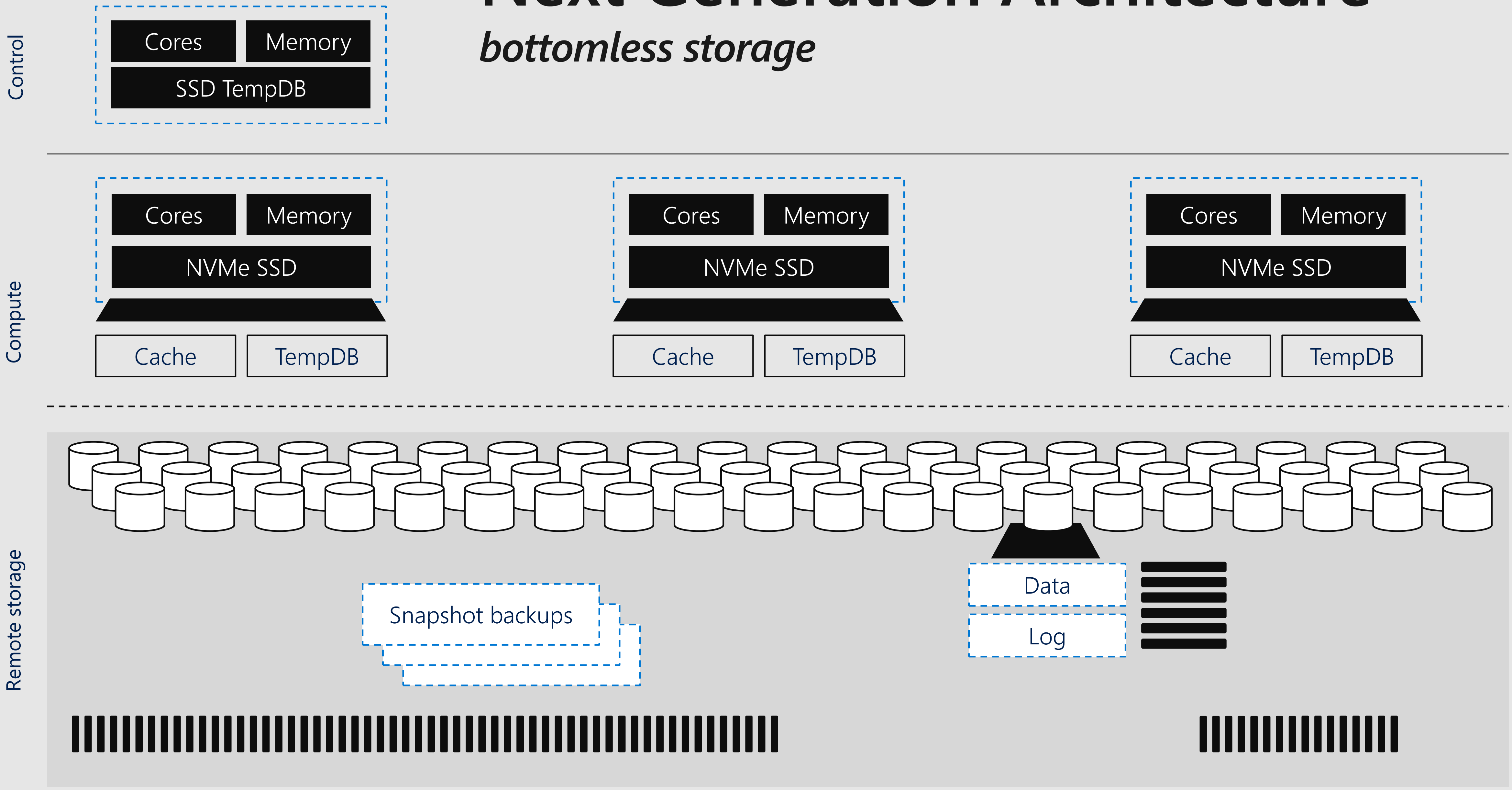
Next Generation Architecture

Adaptive caching



Next Generation Architecture

bottomless storage

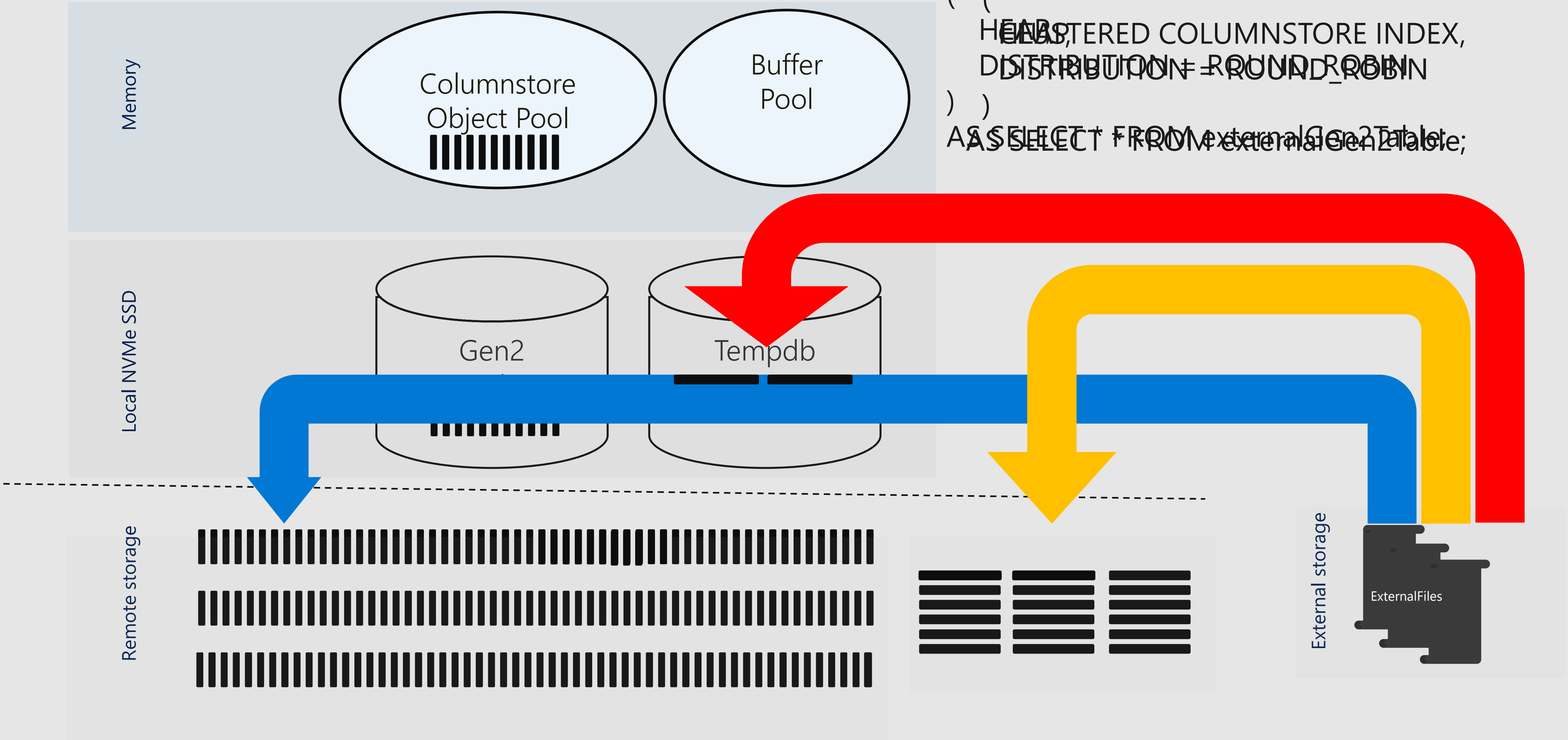


Loading into Gen2

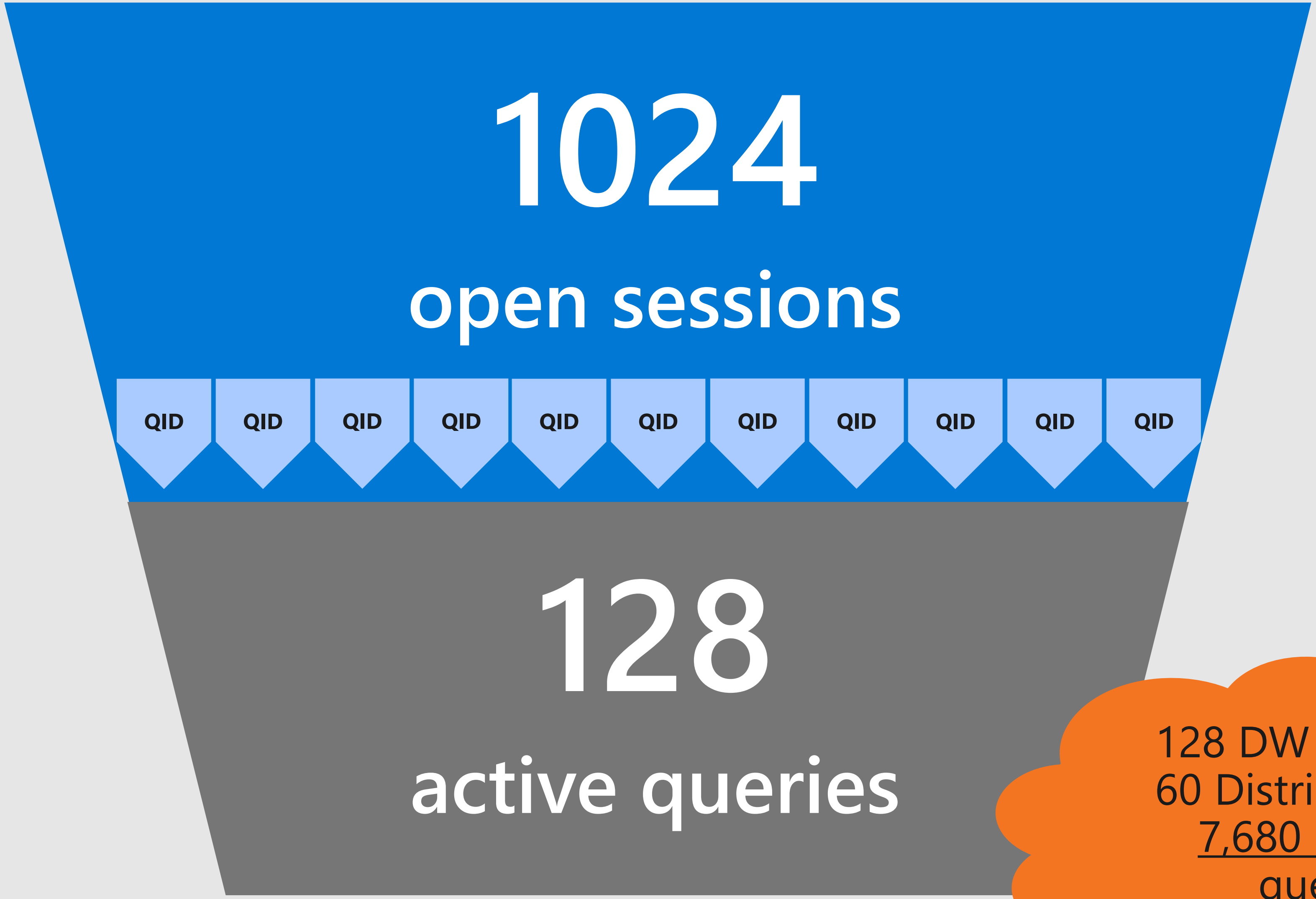
```
SELECT  SUM(TotalCost)
FROM    gen2TableHeap
```

```
CREATE TABLE gen2TableHEAP
WITH
(
    HEAP,
    DISTRIBUTION = ROUND_ROBIN
)
AS SELECT * FROM externalGen2Table;
```

```
 CREATE TABLE #gen2Table_Heap
 WITH #gen2TableHeap
  (
    CLUSTERED COLUMNSTORE INDEX,
    DISTRIBUTION = ROUND_ROBIN
  )
 AS SELECT * FROM externalGen2Table;
```



Gen2 Query Concurrency



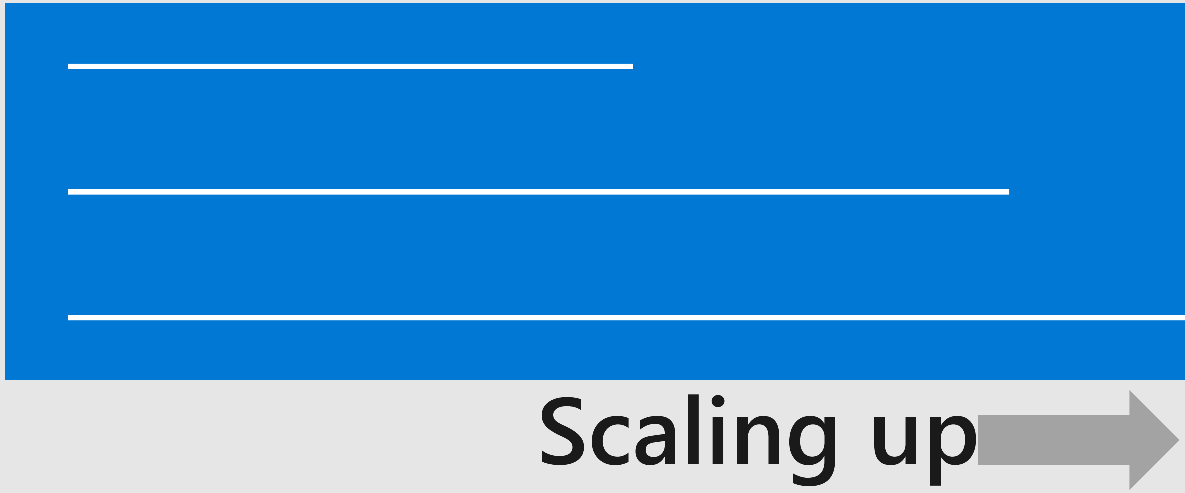
128 DW Queries x
60 Distributions =
7,680 SQL DB
queries

Resource Classes – Dynamic

Allocates variable amounts of memory depending on the scale of the DW instance.

✓ Beneficial for variable sized workloads that scale to meet demand.

🚫 There is no increase in concurrency with scaling. Should be avoided.

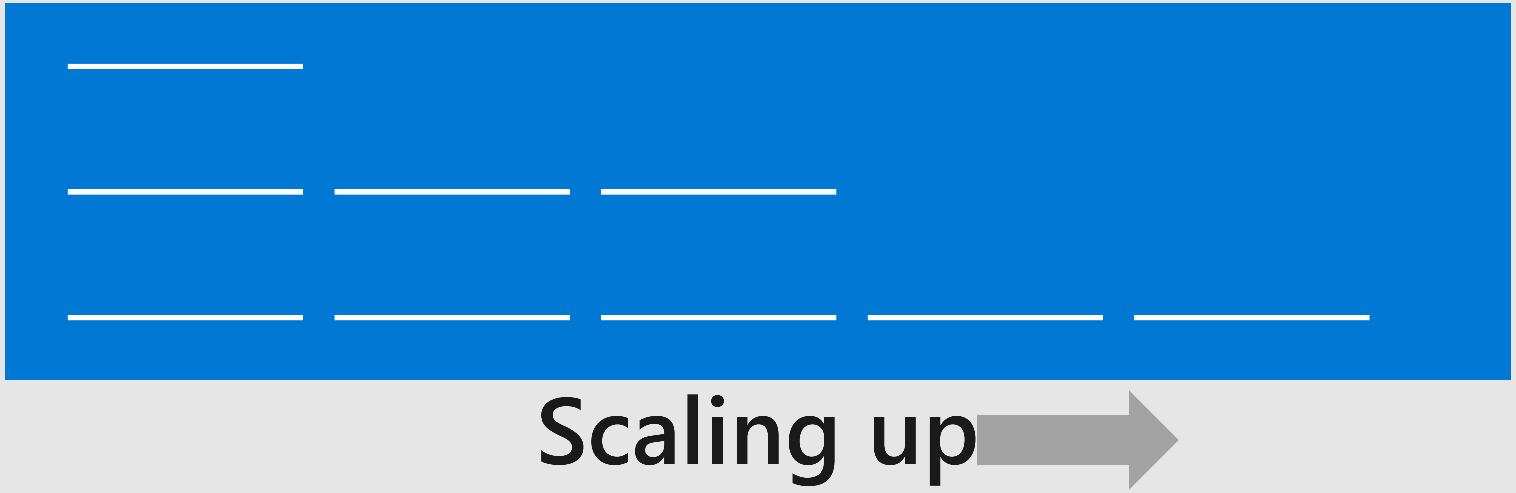


Resource Classes – Static

Allocates a fixed amount of memory regardless of the scale level.

✓ Essential for high query concurrency workloads.

🚫 Queries may run the same regardless of the scale unit.



Demo 3: Load Your Data

Load data to SQL DW with Polybase

Azure SQL Data Warehouse

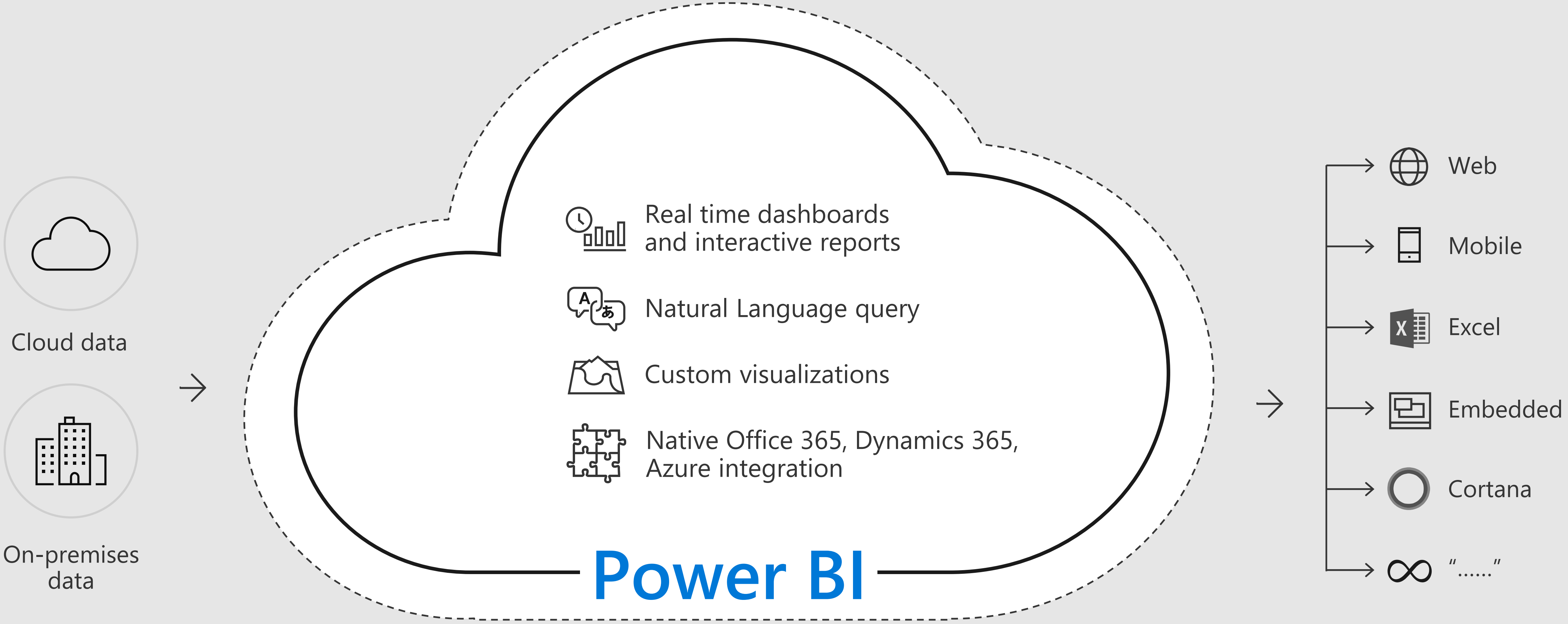
Performance Considerations

- Manage table statistics
- Use PolyBase to load data but do not use external tables for queries
- Use distributed tables and do not over partition them
- Data movement is a common cause of bad performance
- Monitor for data skew
- Leverage HEAP for initial load
  - Faster load time
  - Prevent cache population/eviction
- Assign appropriate resource class for optimal CCI compression

Visualize & Consume

Power BI: experience your data

Any data, any way, anywhere



Cloud data

On-premises data

Real time dashboards and interactive reports

Natural Language query

Custom visualizations

Native Office 365, Dynamics 365, Azure integration

Power BI

Web

Mobile

Excel

Embedded

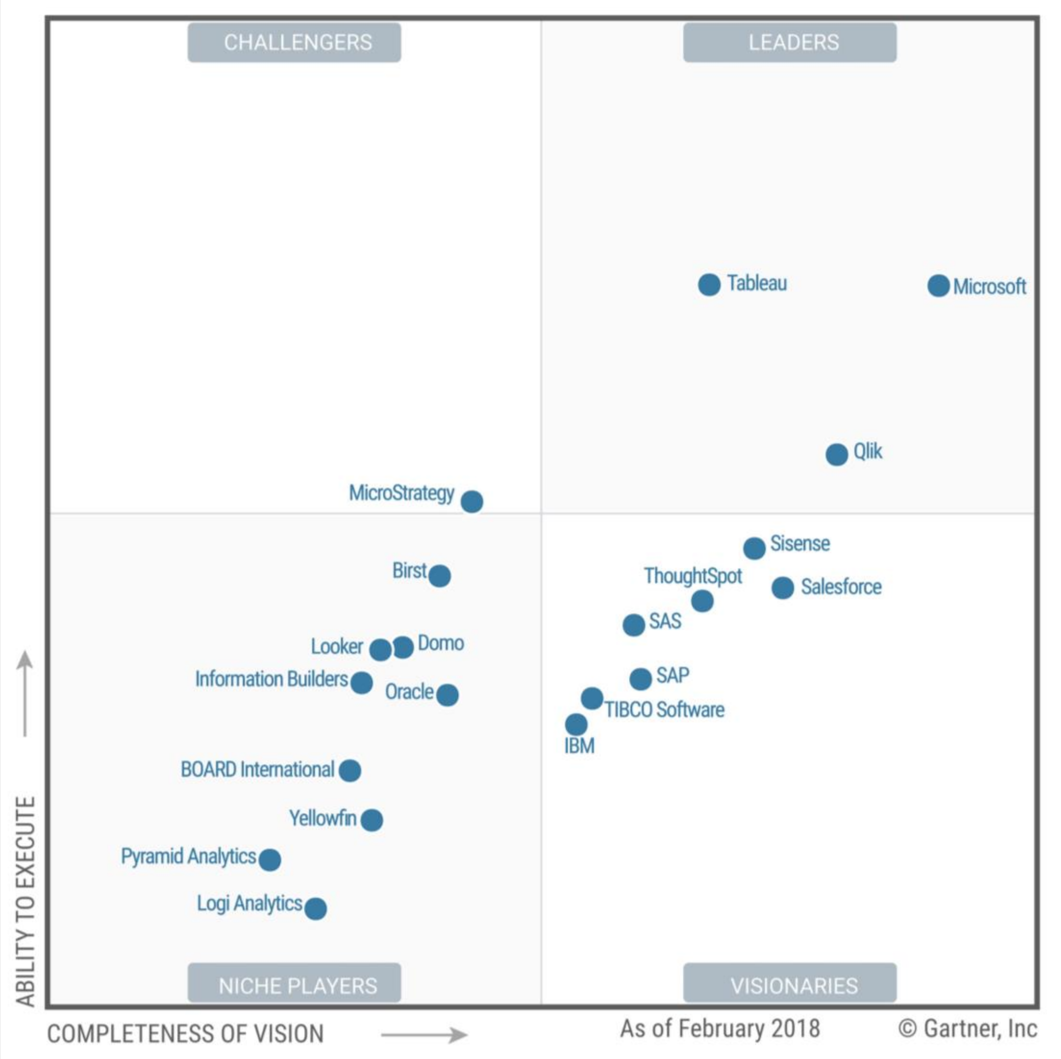
Cortana

“……”

Gartner®

February 2018

A Leader in Analytics & BI Platforms*

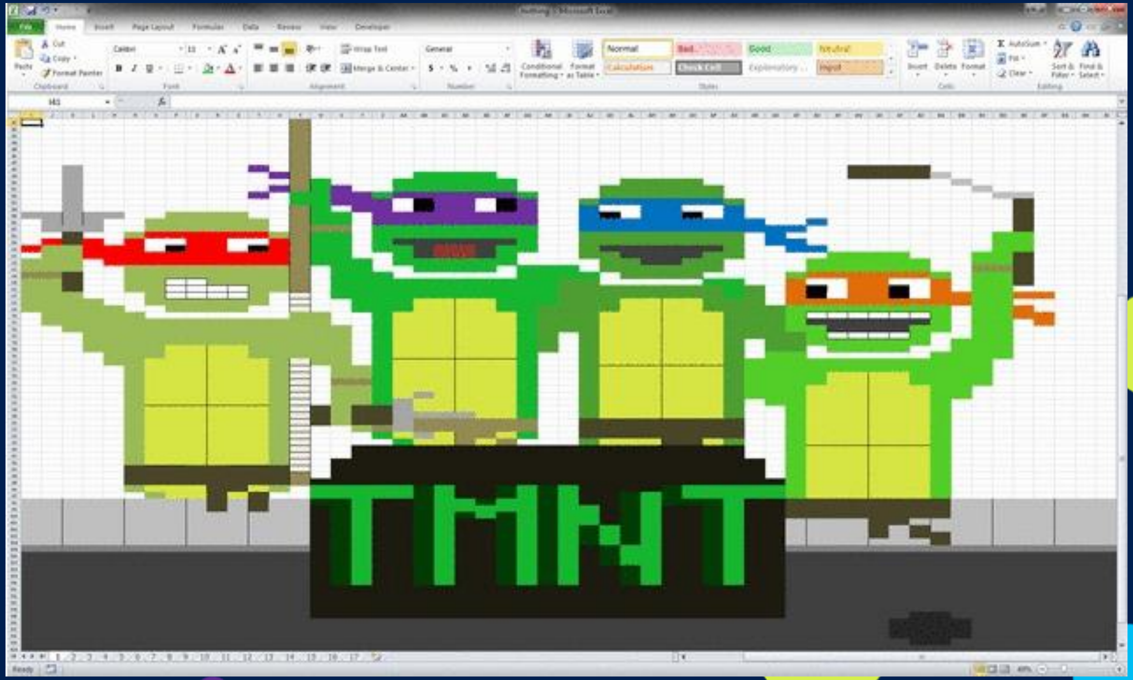


Source: Gartner (February 2018)

***Gartner "Magic Quadrant for Analytics and Business Intelligence Platforms," by Cindi Howson, Rita L. Sallam, James Laurence Richardson, Joao Tapadinhas, Carlie J. Idoine, Alys Woodward, February 2018**
The above graphics were published by Gartner, Inc. as part of a larger research document and should be evaluated in the context of the entire document. The Gartner document is available upon request from Microsoft. Gartner does not endorse any vendor, product or service depicted in its research publications, and does not advise technology users to select only those vendors with the highest ratings or other designation. Gartner research publications consist of the opinions of Gartner's research organization and should not be construed as statements of fact. Gartner disclaims all warranties, expressed or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose.

Demo 4: Visualize Your Data

Connect to your data sources with Power BI



Key Takeaways

What did we see today?

- What is a modern data warehouse?
- Building a data pipeline in Azure:
    - Use Data Factory to Orchestrate
        - Data movement from Cloud/OnPrem to Cloud/On Prem
        - Crunch and prep large datasets with Databricks
        - Load data to the serving layer
    - Visualize with Power BI

Key Takeaways

Plan your data strategy

Use case driven approach, evolve the target state

Use data warehouse services which meet the needs of the organization

Cloud is the obvious choice

Accelerate time to value and global service availability

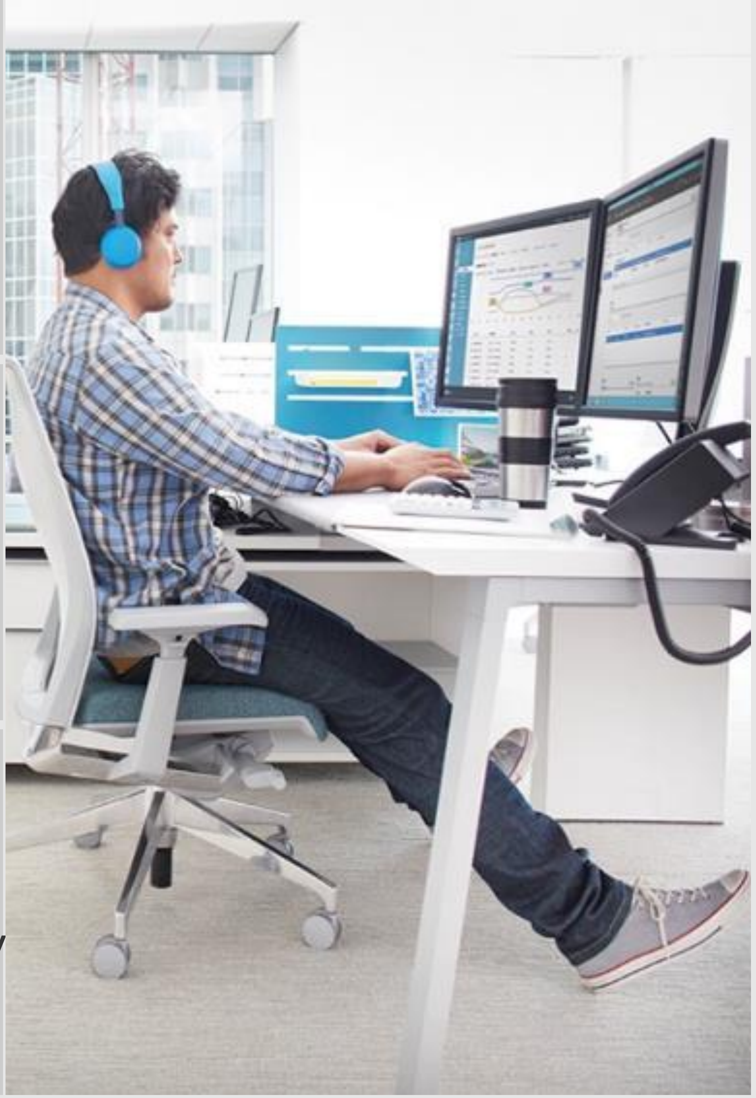
“No longer waiting six months to procure a server”

Multiple implementation options and considerations

Consider non-technical challenges

Technology is the easy part. Greater challenge with people and process

Example: organizational capabilities, evolution of procedures, data privacy and security concerns



Q&A

If you have questions please proceed to the Q&A MICROPHONE located in your session room (or is it?)

Microsoft

Thank You.

Catalin Esanu
Cloud Solution Architect, STU
Microsoft Israel
cesanu@microsoft.com