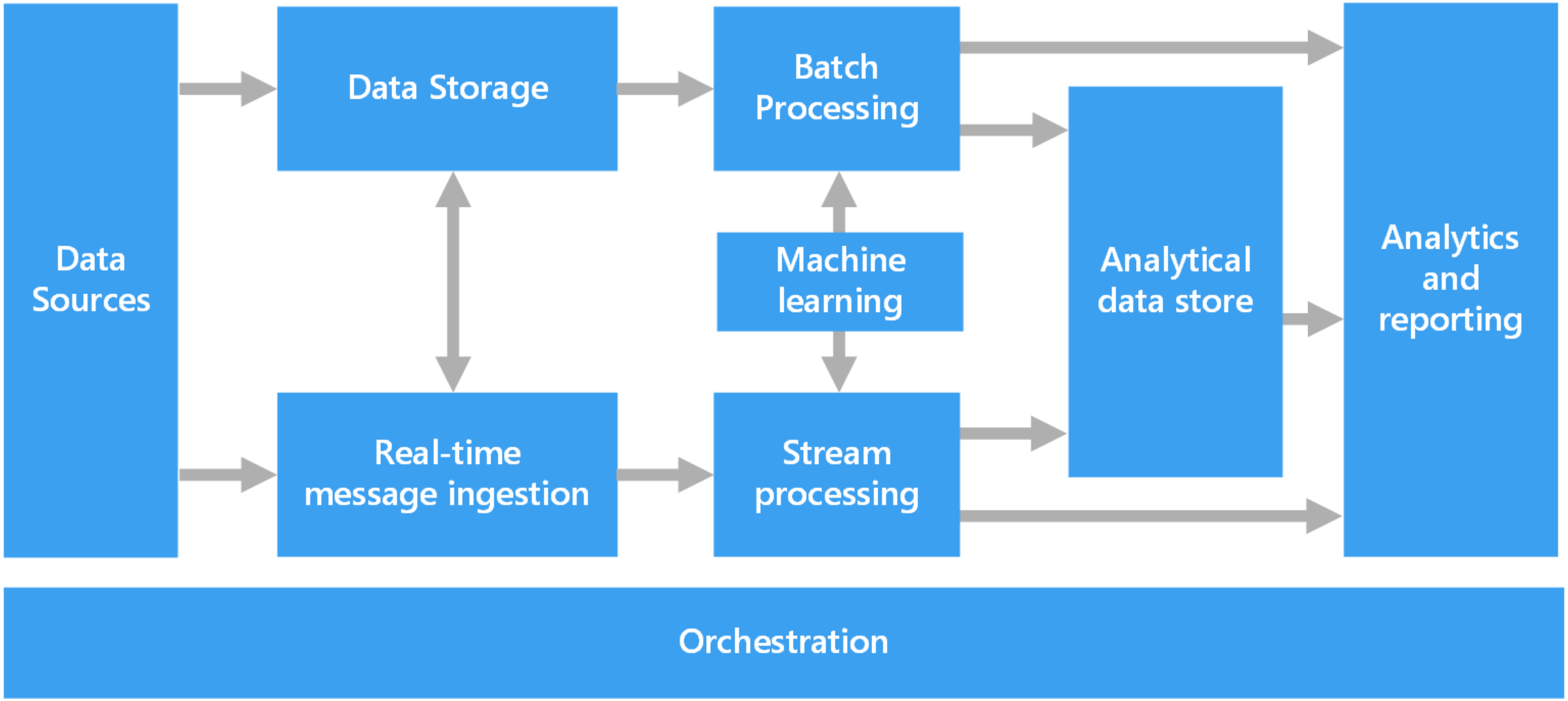
Microsoft Azure
TLV Cloud Workshops

18.11.2018 | Hilton Tel-Aviv

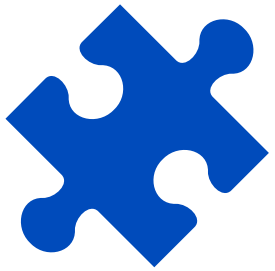
Microsoft

Lambda architecture simplified with Azure Cosmos DB and Azure Databricks

Oshik Avioz
Cloud Solution Architect | Data & AI
osavioz@microsoft.com
@OshikAvioz
www.linkedin.com/oshikavioz
https://github.com/oavioz



https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/



Requirements

The tolerance to human errors

The tolerance to hardware crashes

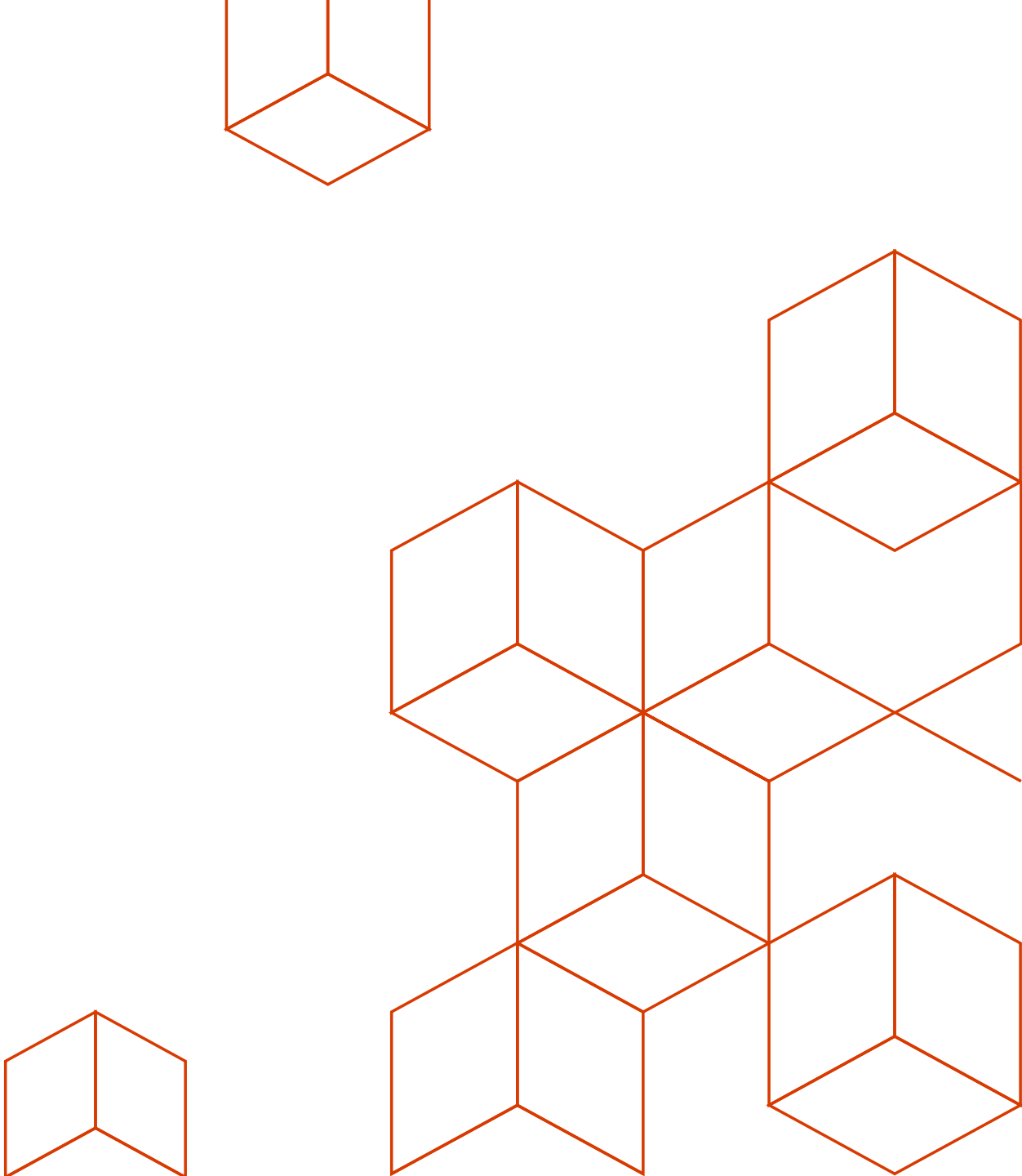
Scalability and quick response time

Solutions

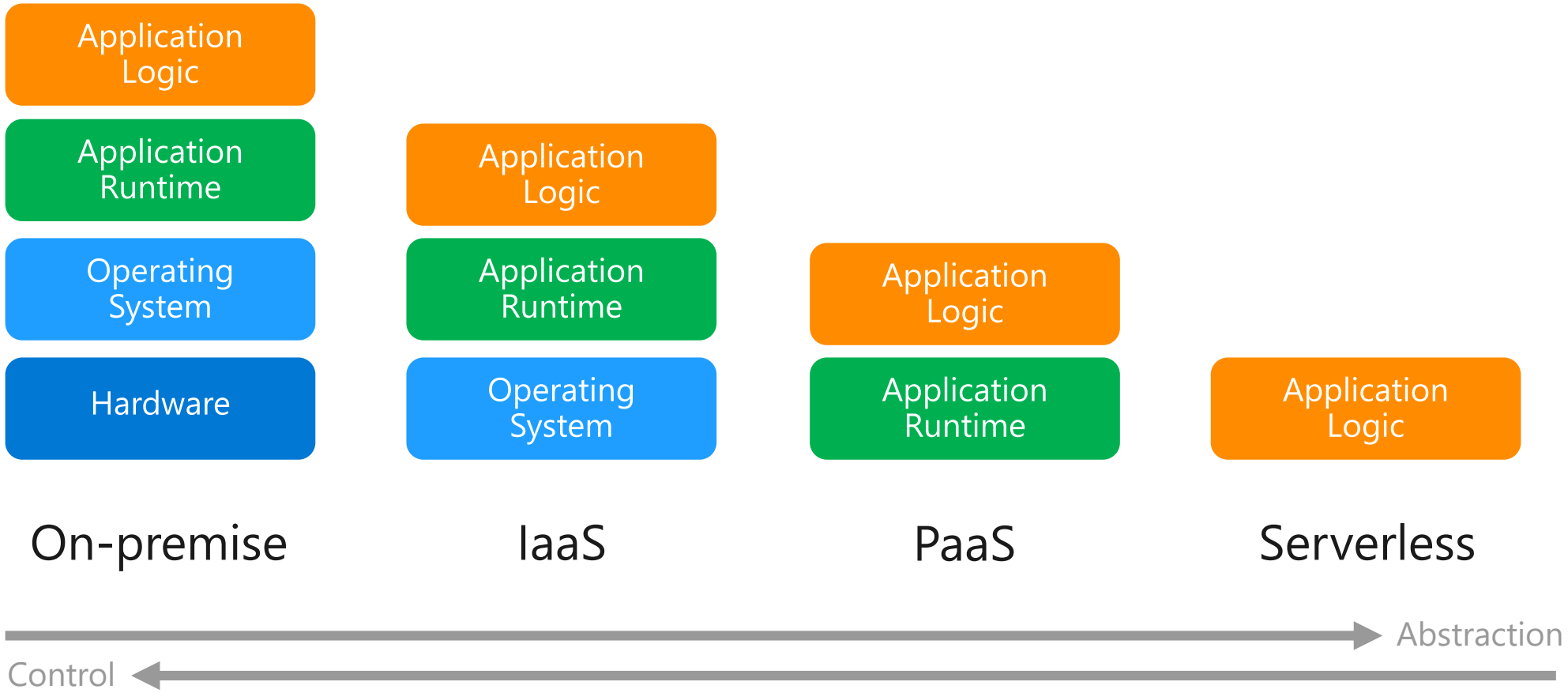
- ✓ Automated data processes
- ✓ Managed cloud systems
- ✓ Serverless architecture

https://www.talend.com/blog/2016/05/31/the-lambda-architecture-and-big-data-quality/

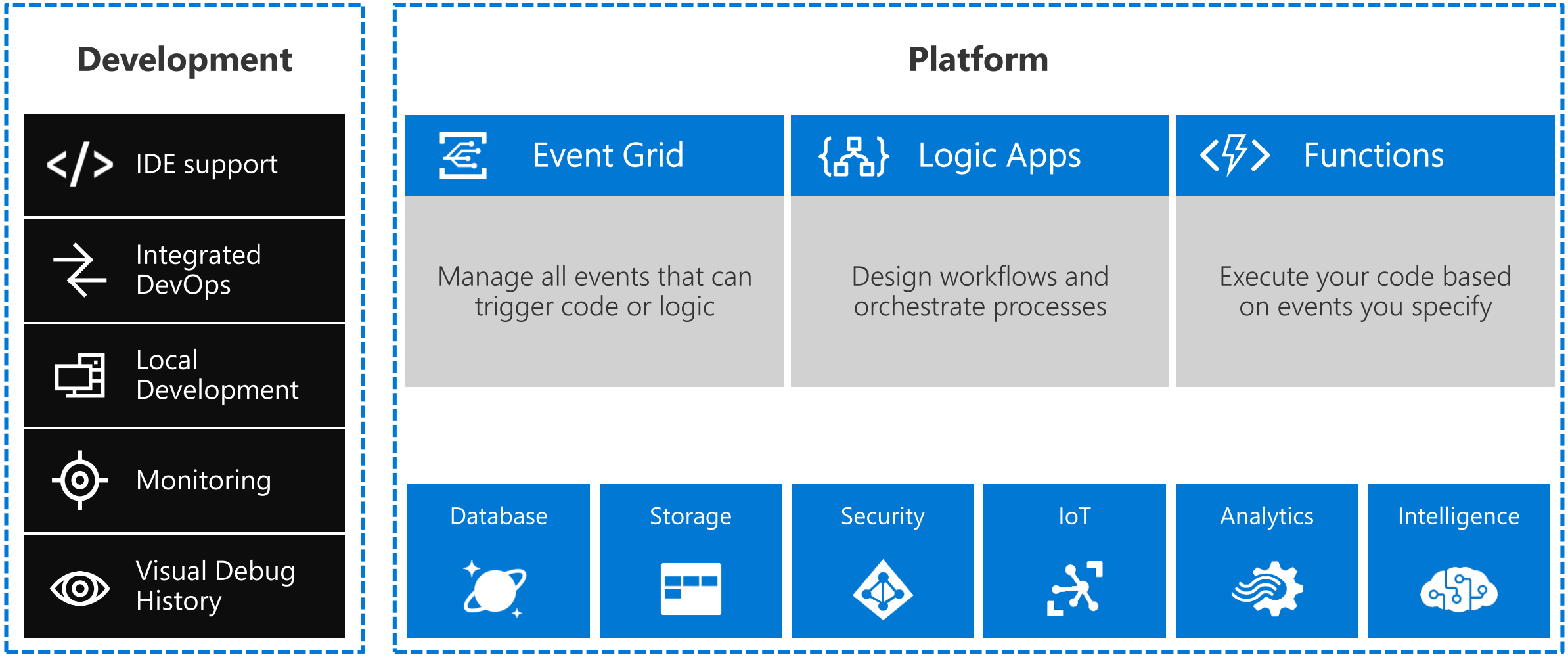
Using the cloud to solve manageability challenges



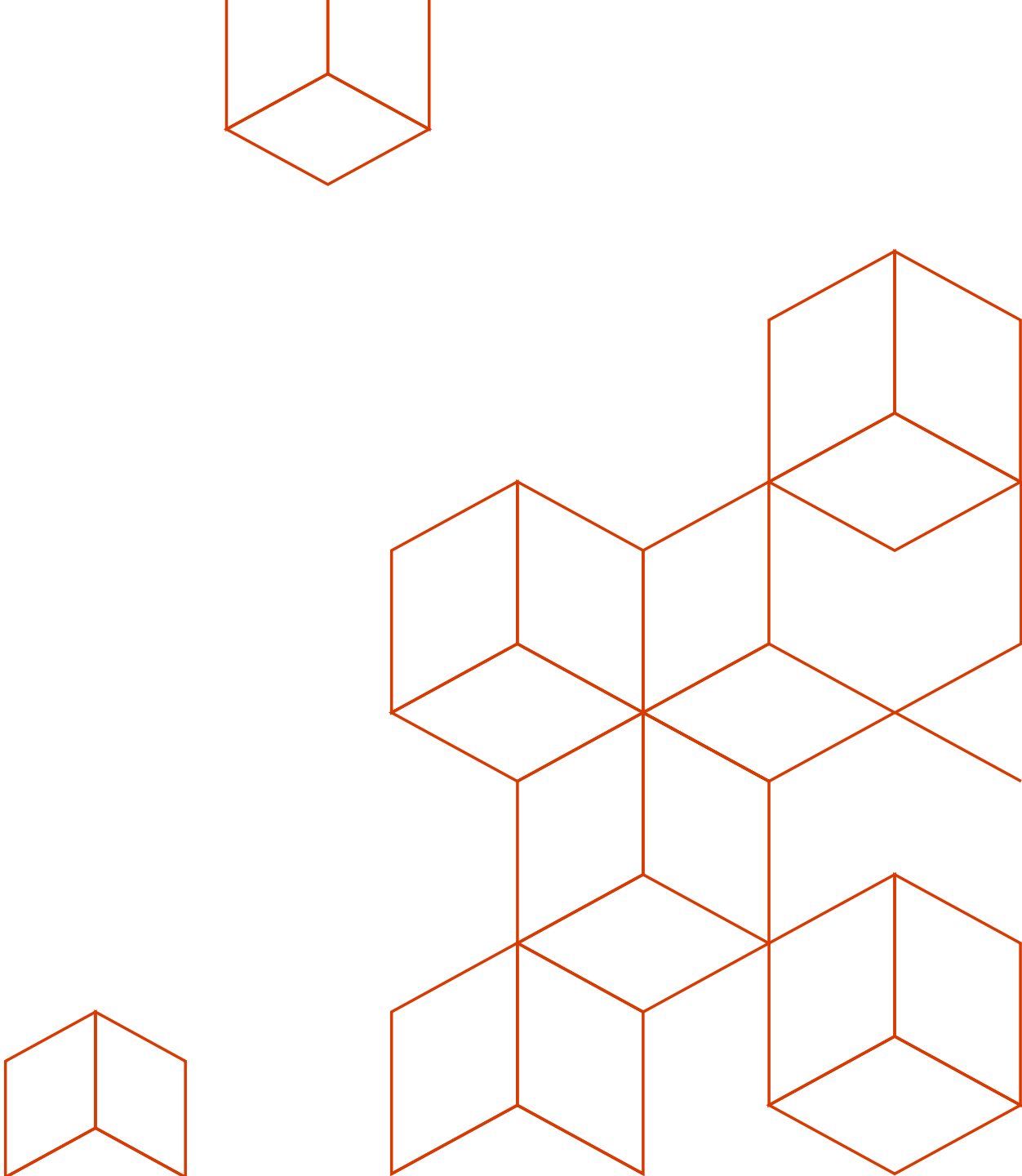
Management Requirements



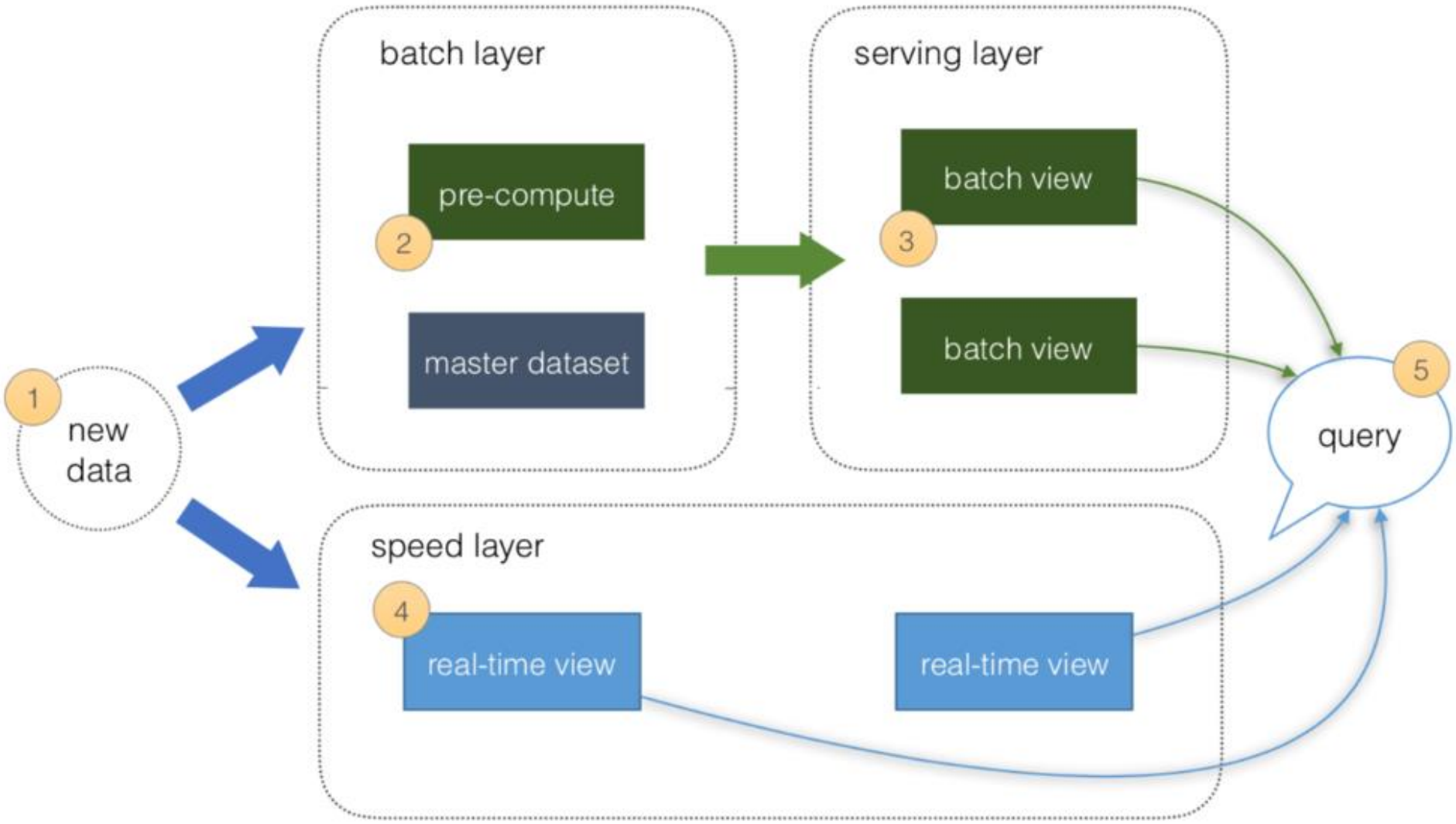
Serverless application platform components



What about the data architecture?



What is a Lambda Architecture?



Needs for a lambda architecture

- Both real-time and aggregated data processing.
- A scalable architecture with tunable components.
- Database(s) with workloads of different nature:

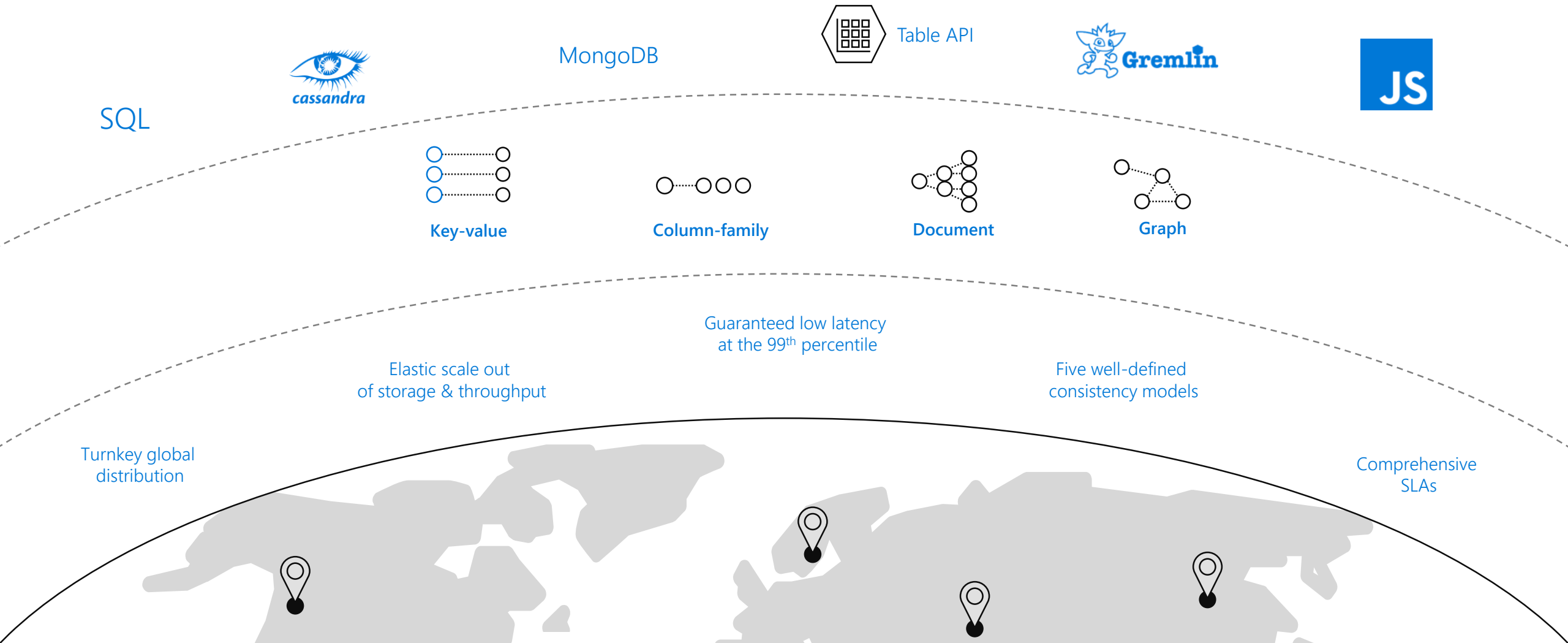
Component	Read frequency	Write frequency	Storage size
Speed layer	Real-time	Real-time	Minimal/Stream
Batch layer	Periodical	Periodical	Batch size
Master dataset	Both	Both	Batch size + historical

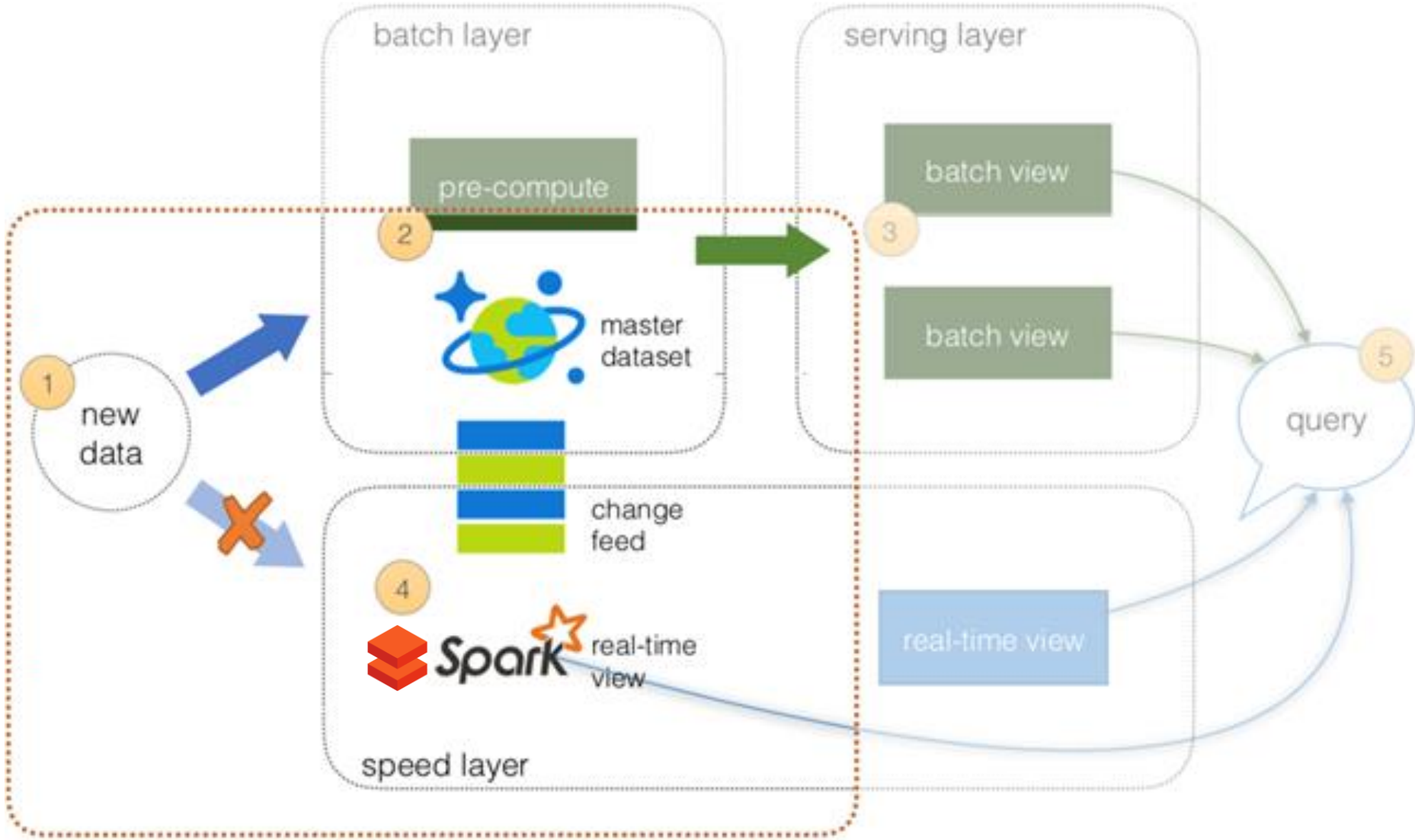
Cosmos DB can address these scenarios by providing tunable data querying in different collections.

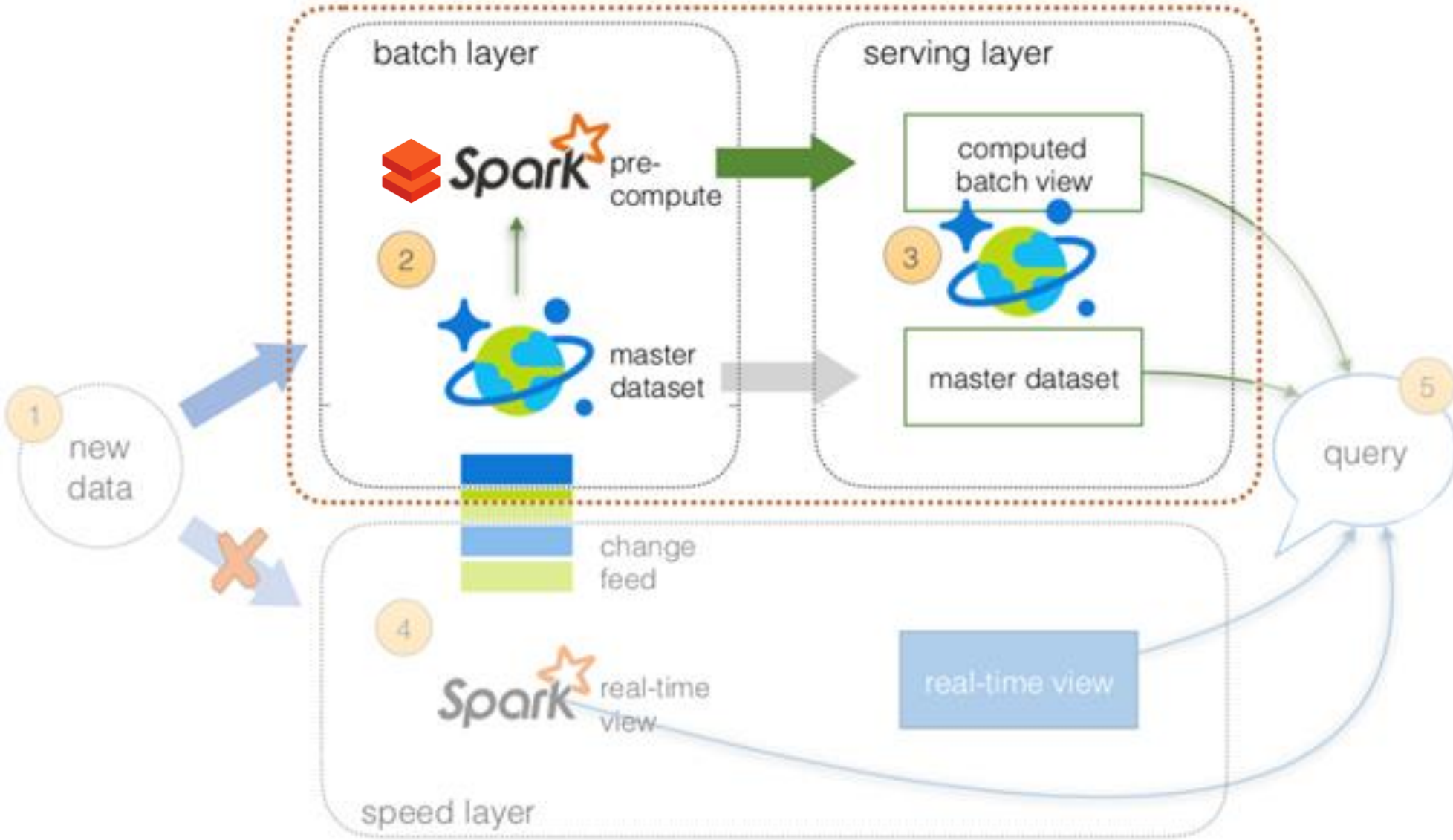
Simplifying the lambda architecture with Cosmos DB

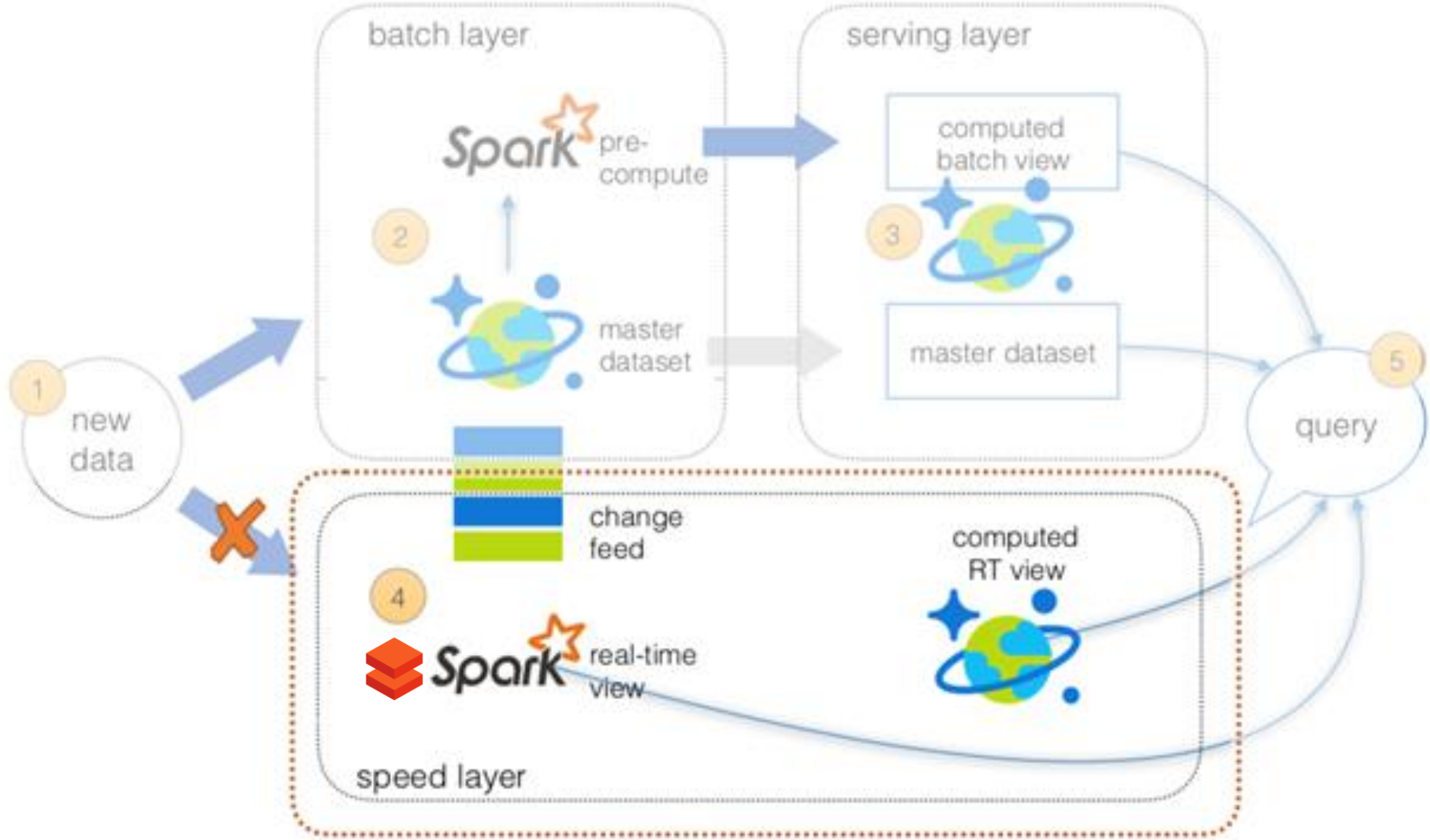
What is Azure Cosmos DB

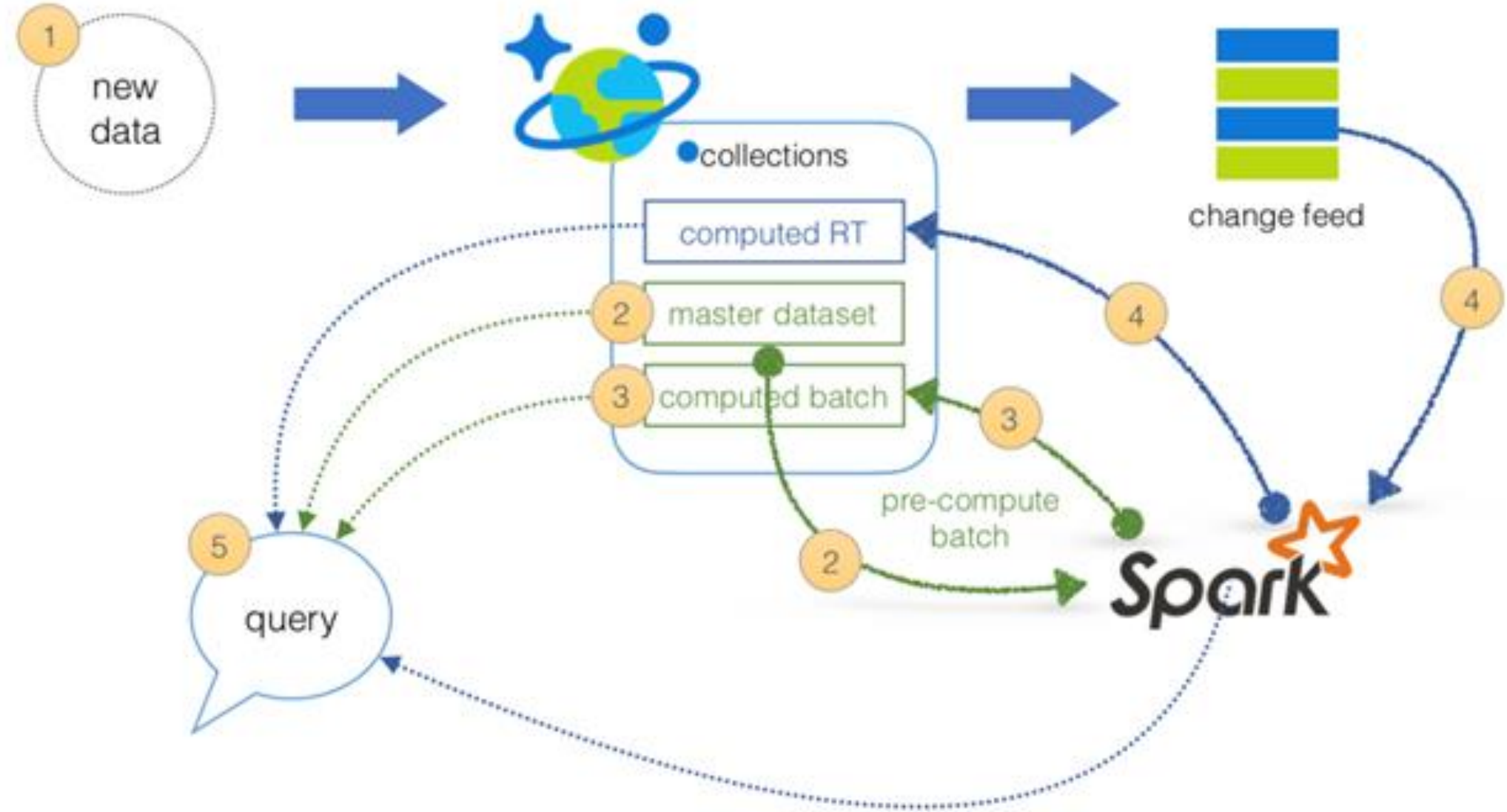
A globally distributed, massively scalable, multi-model database service



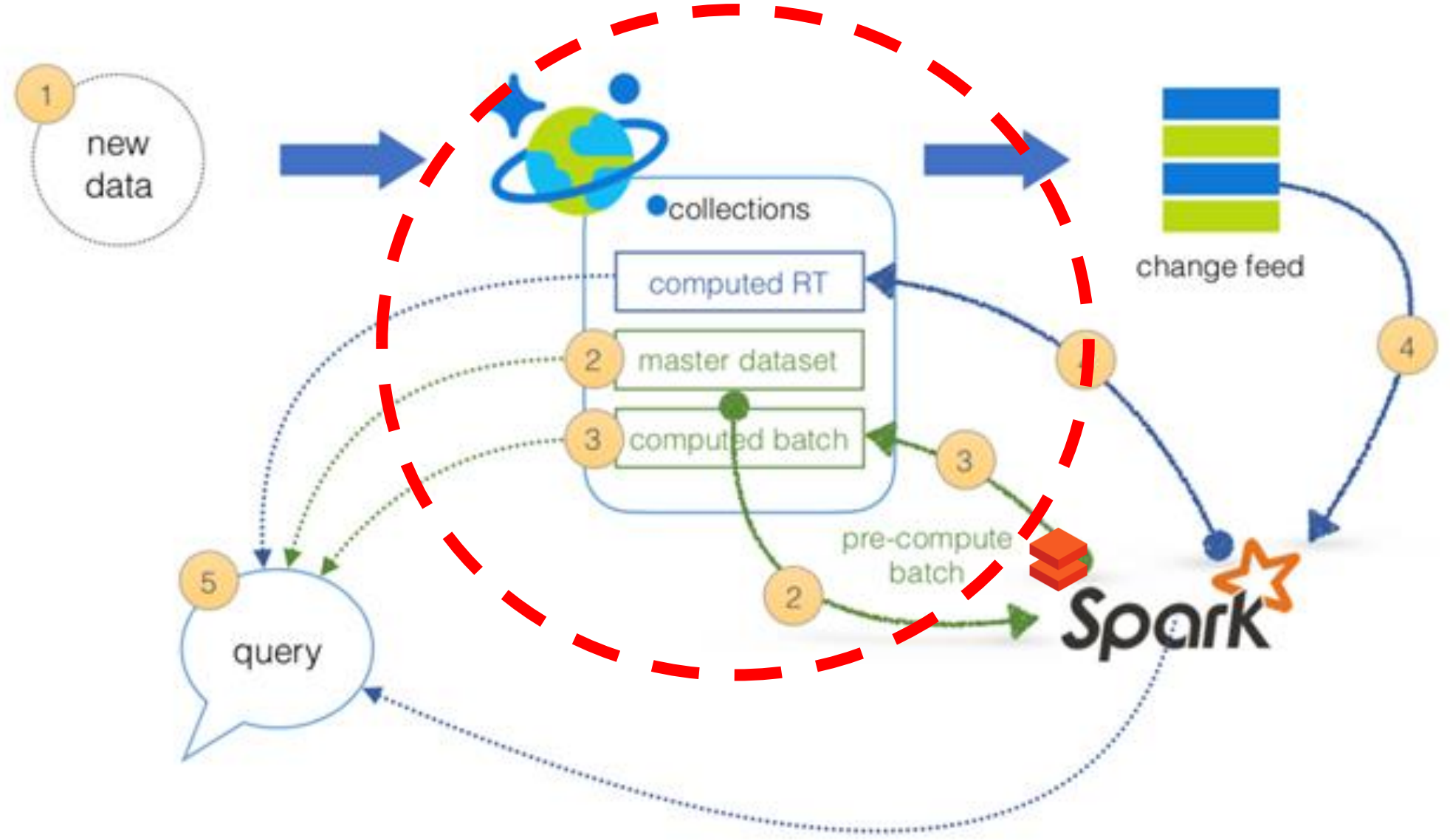

batch layer
serving layer
pre-compute
2
3
batch view
batch view
master
dataset
1
new
data
5
query
change
feed
4
Spark
real-time
view
real-time view
speed layer


batch layer
serving layer
Spark
pre-
compute
computed
batch view
2
3
master
dataset
master dataset
1
new
data
5
query
change
feed
4
Spark
real-time
view
real-time view
speed layer

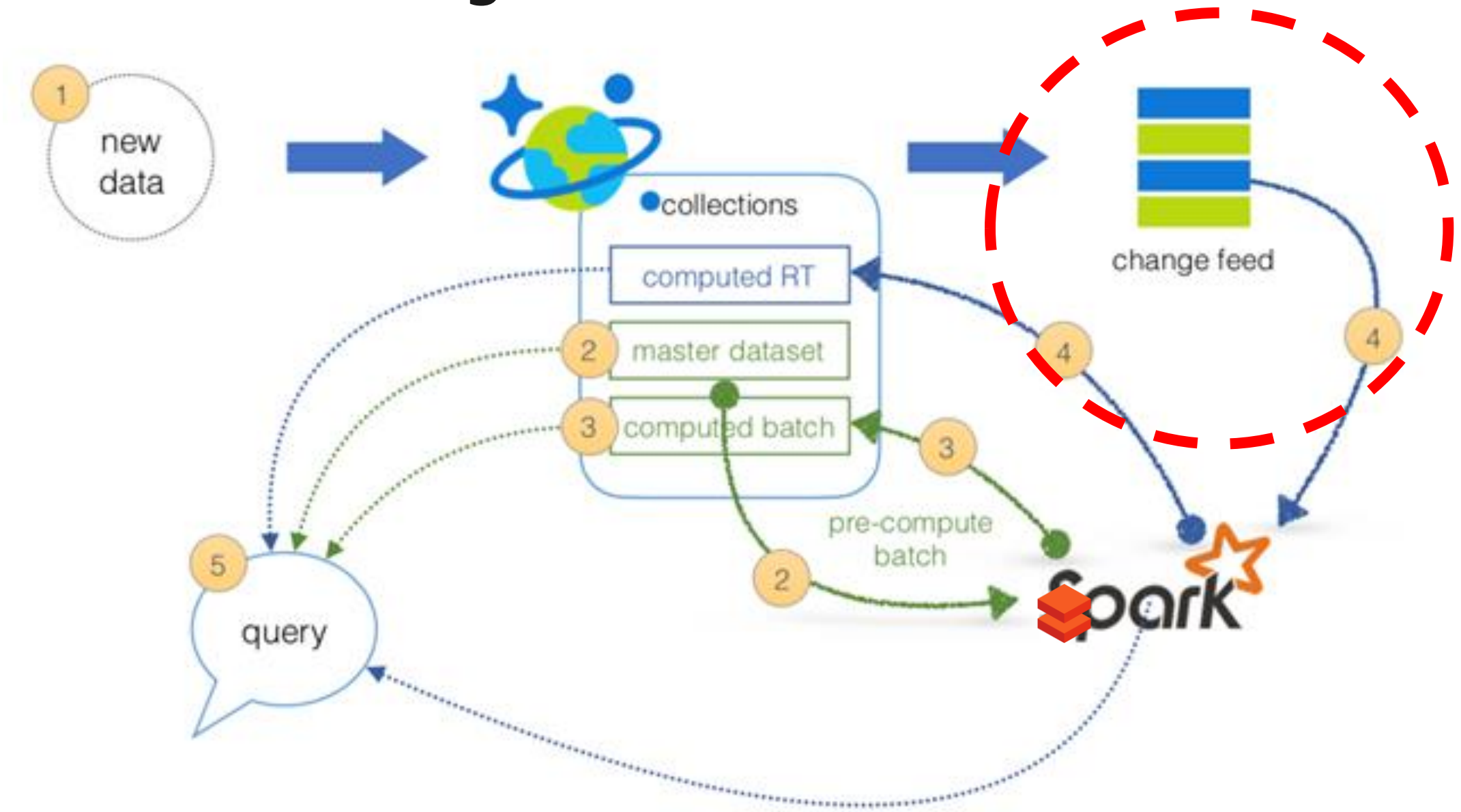

batch layer
serving layer
Spark
pre-
compute
computed
batch view
2
3
master
dataset
master dataset
1
new
data
5
query
change
feed
computed
RT view
4
Spark
real-time
view
speed layer


1
new
data
collections
computed RT
2
master dataset
3
computed batch
change feed
4
4
3
pre-compute
batch
2
5
query
Spark

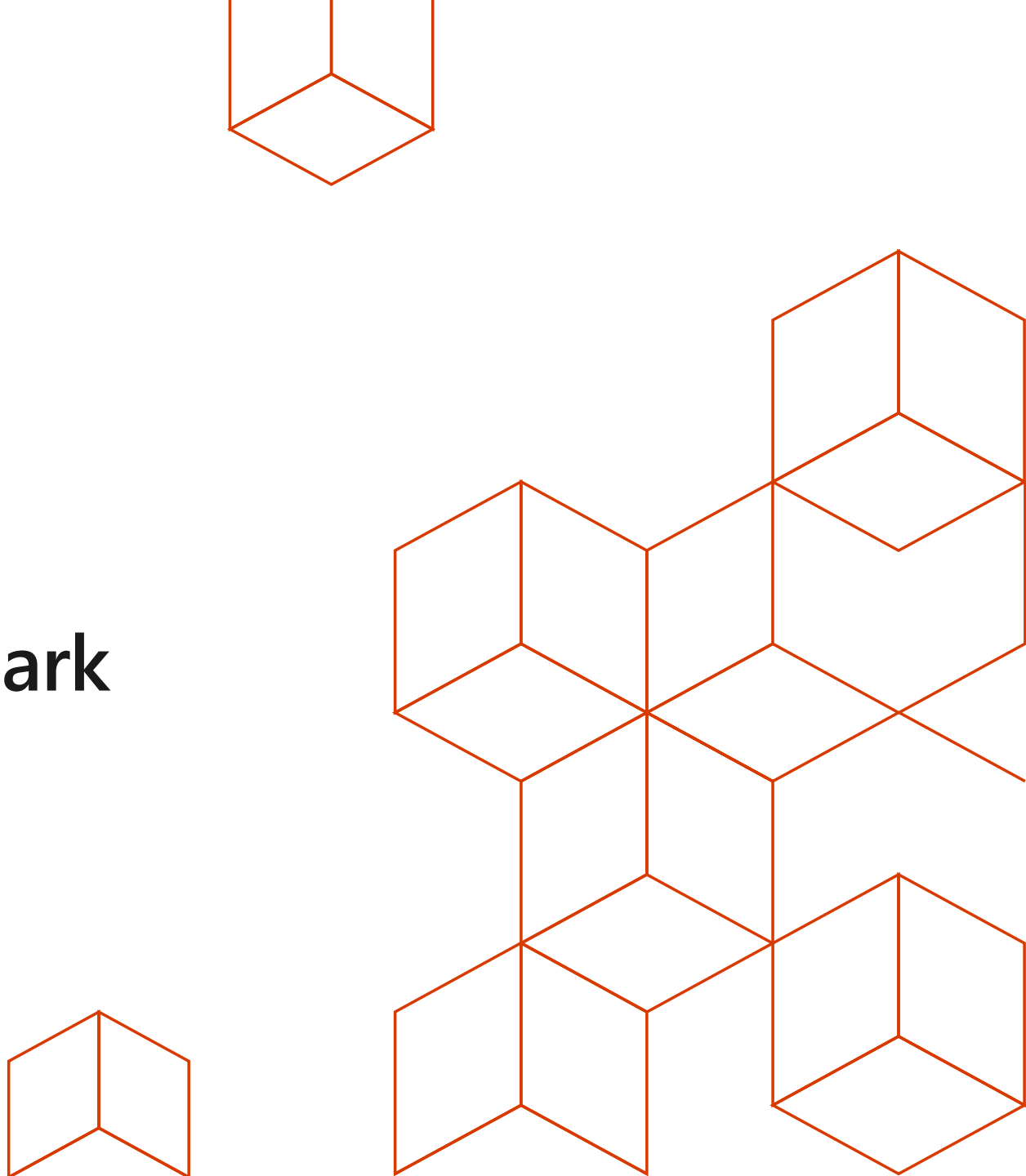
Cosmos DB for the master dataset, computed storage and serving layer



Cosmos DB Change Feed



Cosmos DB and Apache Spark
through Azure Databricks



What is Apache Spark?

Open-source cluster computing framework.

Scalable data processing layer with implicit parallelism and fault-tolerance.

Simple programmability with extensible libraries.

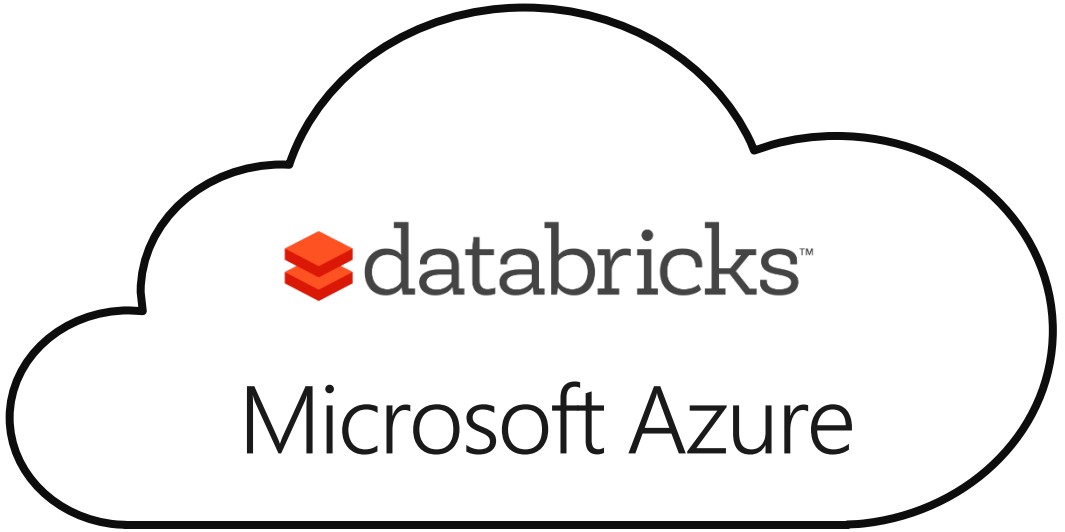
Uses scale-out/horizontal provisioning of nodes.

Provides a standard platform for processing jobs that require powerful computation.

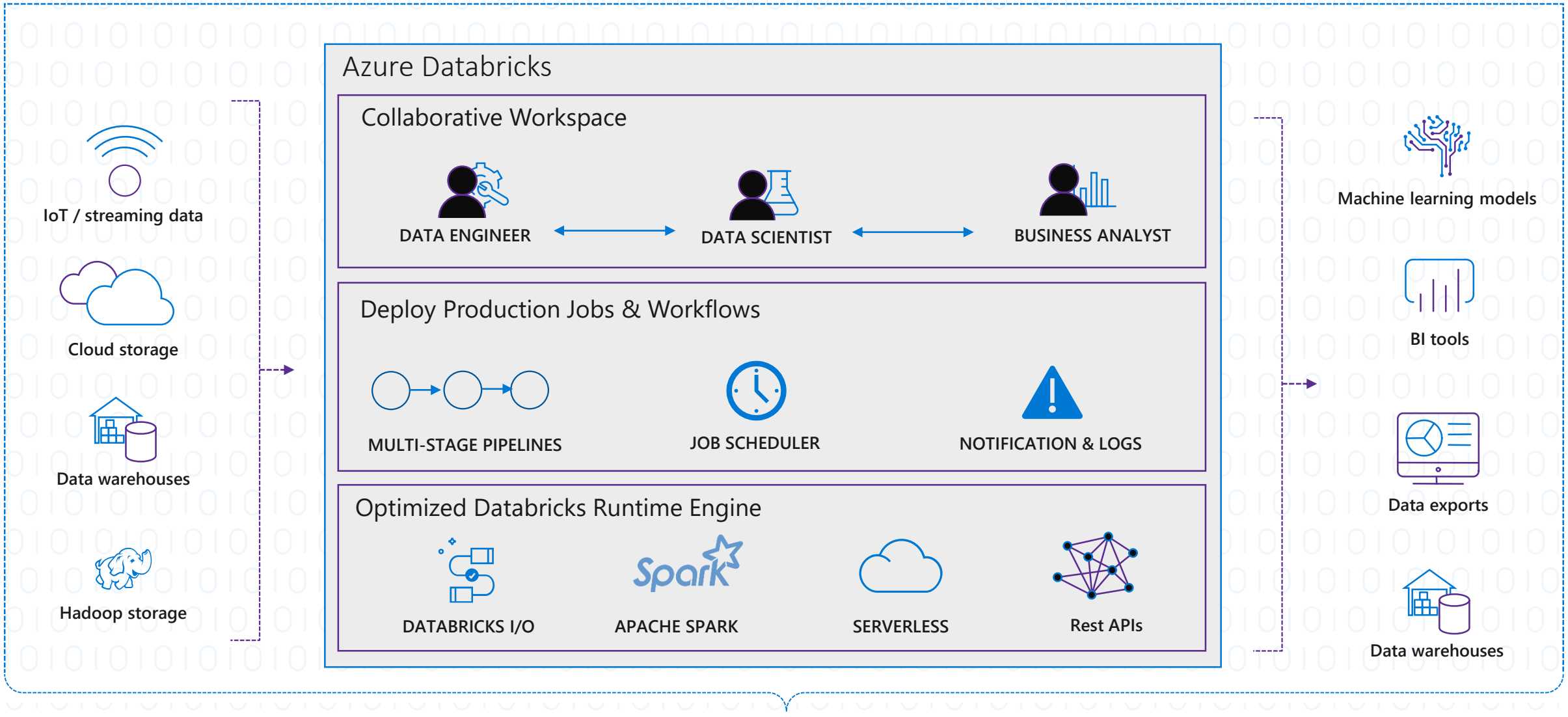
What is Azure Databricks?

Managed cloud solution for Apache Spark.

- Uses an optimized version of Spark.
- Provides abstracted control of clusters/nodes with cloud resources.
- Easy to use interface similar to jupyter notebooks.
- Provides Python, SQL and Scala environments.
- Highly configurable for library and runtime environment versions.



AZURE DATABRICKS

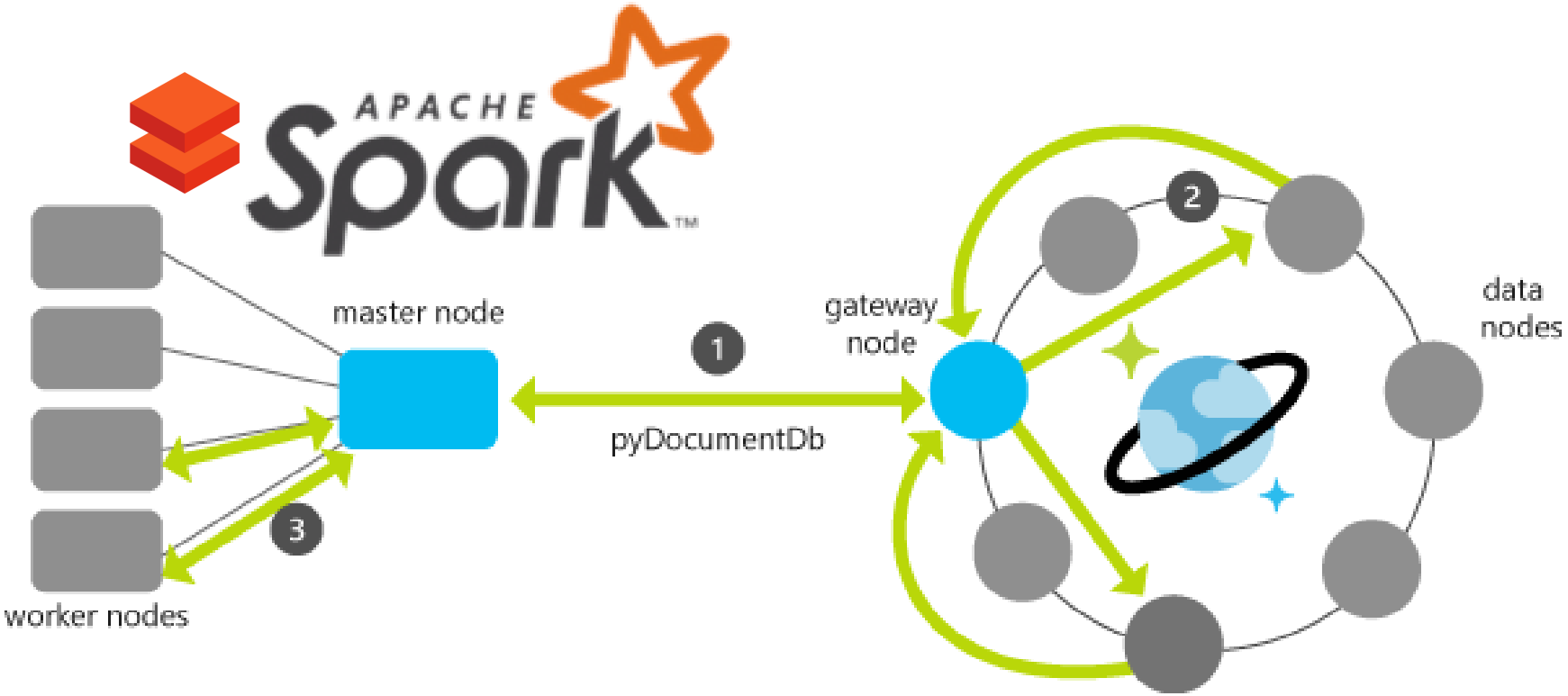


Enhance Productivity

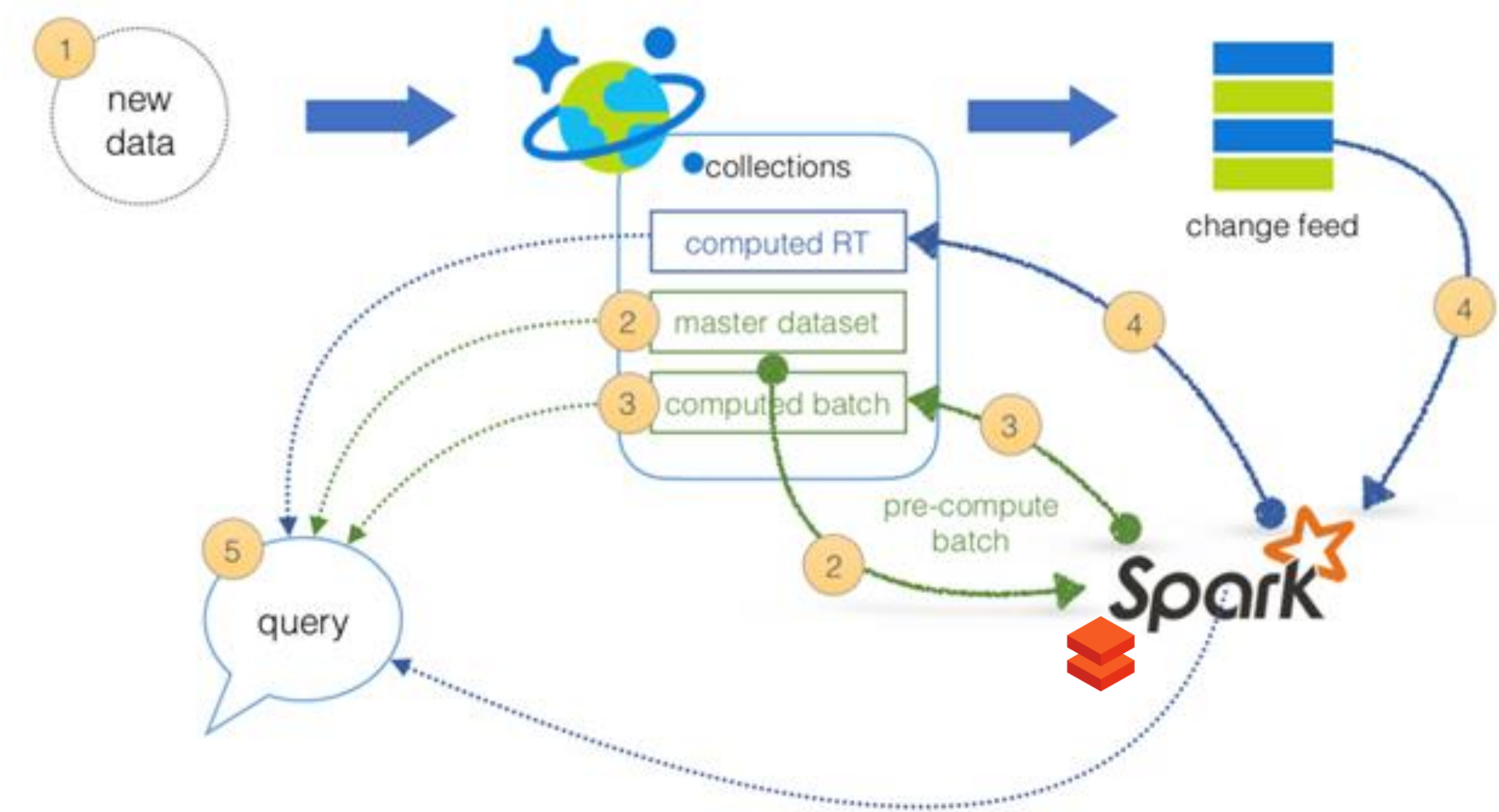
Build on secure & trusted cloud

Scale without limits

Cosmos DB Native Integration with Spark

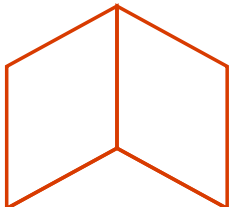


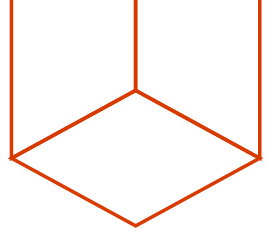
Available for both Python and Scala/Java

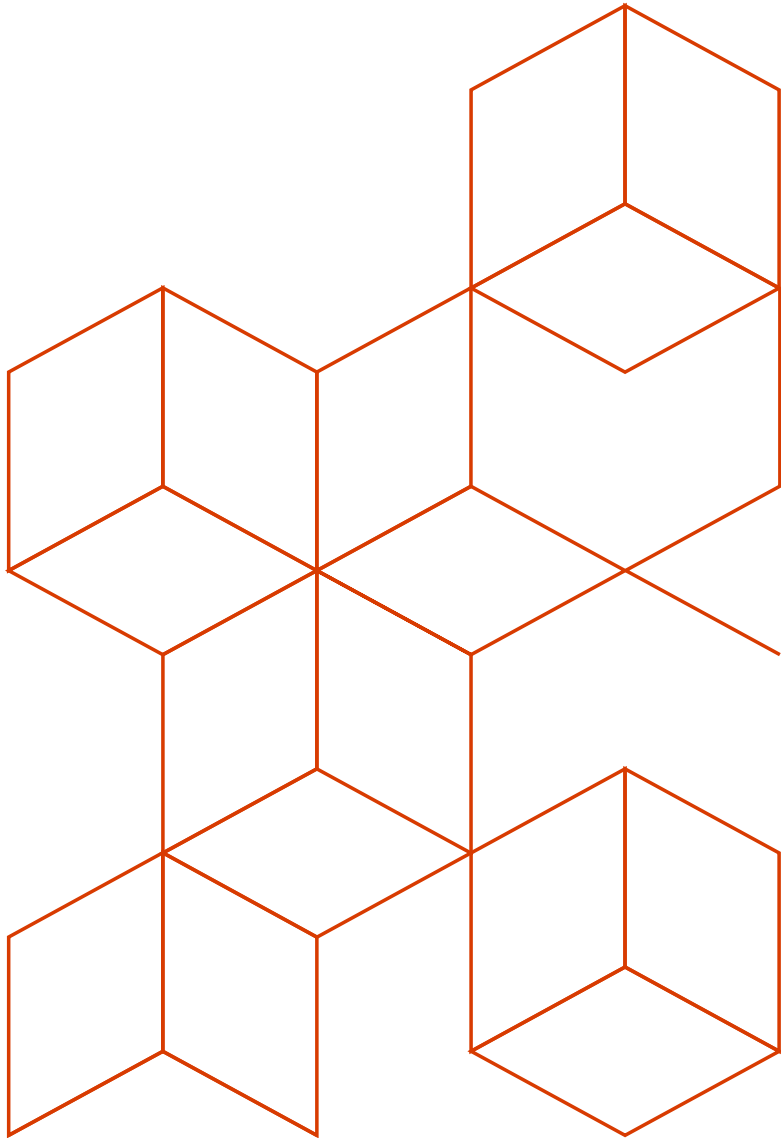

1
new
data
collections
computed RT
2
master dataset
3
computed batch
change feed
4
4
3
pre-compute
batch
2
5
query
Spark

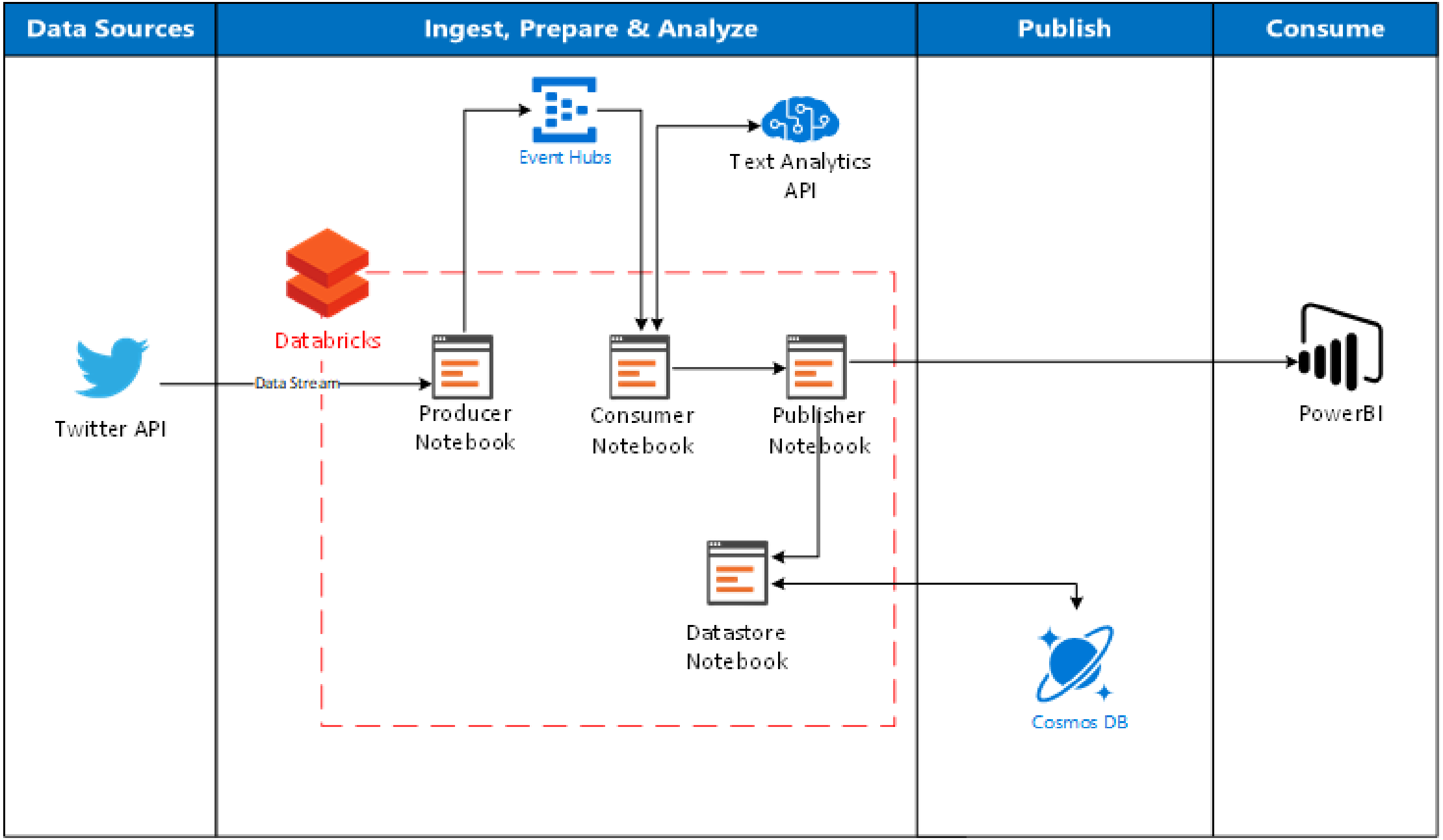
Demo

A lambda architecture using Cosmos DB and Azure Databricks








Data Sources
Ingest, Prepare & Analyze
Publish
Consume
Event Hubs
Text Analytics API
Databricks
Twitter API
Data Stream
Producer Notebook
Consumer Notebook
Publisher Notebook
PowerBI
Datastore Notebook
Cosmos DB

Demo:
Streaming at scale in CosmosDB & Azure Databricks

Q&A