



# Using data to support every learner from classroom to career

Open Education Analytics (OEA) is an open source community coordinated by Microsoft Education. We collaborate with education systems across the world to develop modern data intelligence capabilities.


Get started with the modern data estate using the OEA reference architecture, open source modules, training, and use cases. Ensure responsible data and AI practices with our OEA toolkit. Join our community and contribute education data solutions.

[Get Started](#)





# OEA repo update

 Search or jump to...

microsoft / OpenE


docs


framework


modules

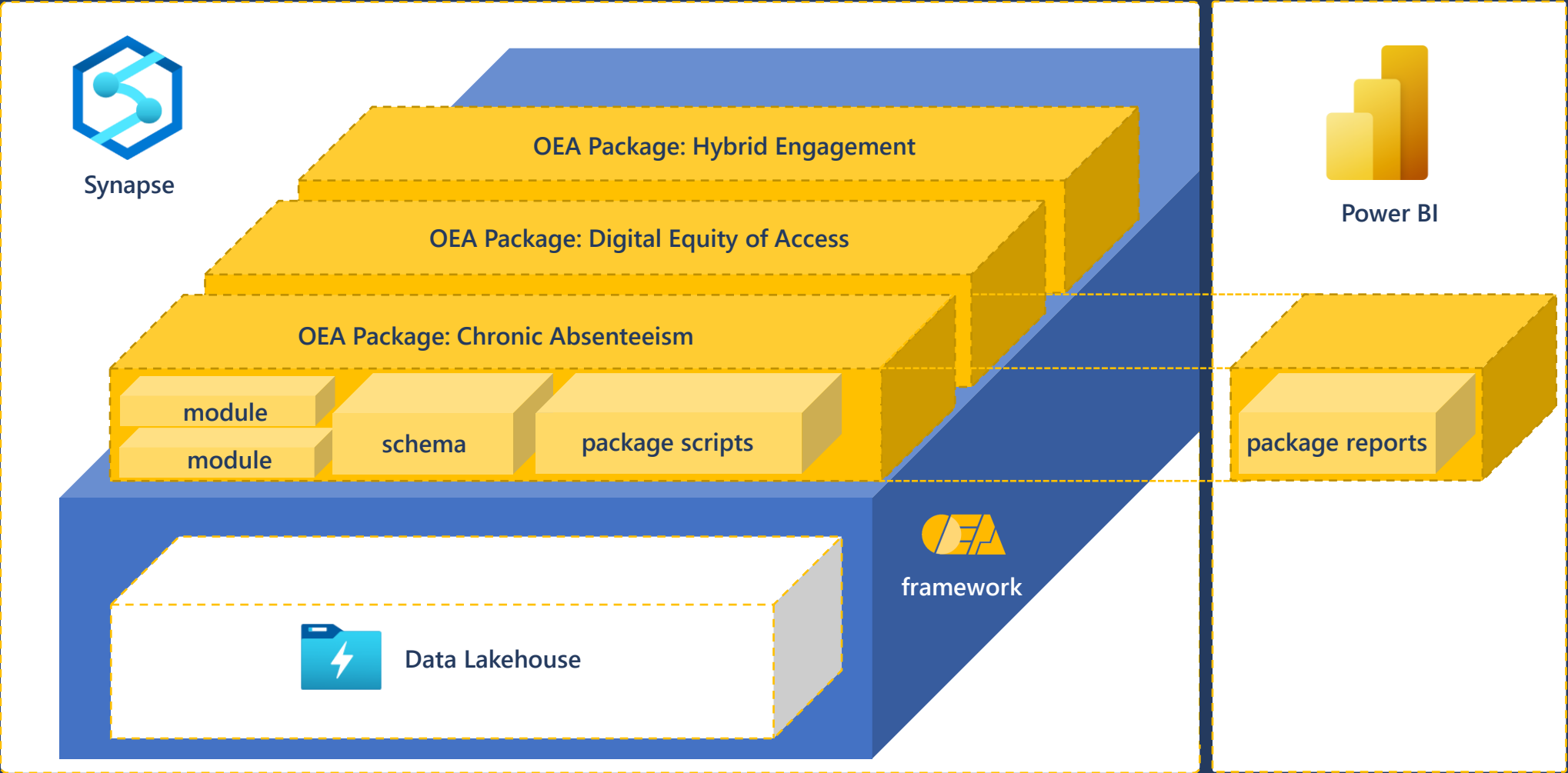
packages

schemas

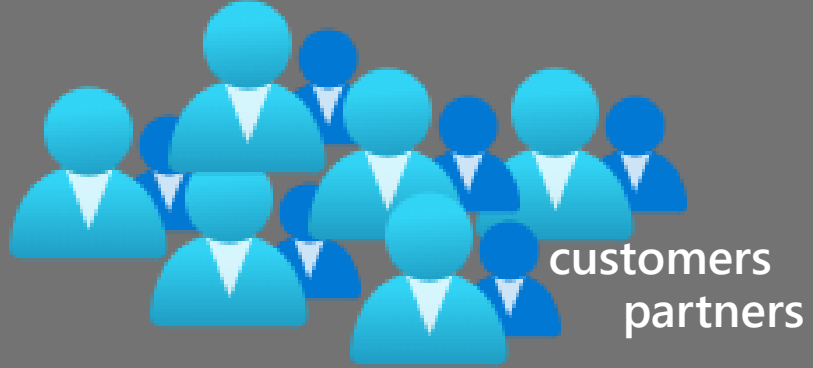
 .gitignore

 README.md

 setup.sh



# The Ecosystem



technical  
guidance

feedback

prod mgrs  
developers

# The Tech

Custom Assets



Modules

Packages

Framework, Conventions, Standards

- refine
- standardize
- commoditize

Platform

# Inspire Opening

ON DEMAND

Wednesday, July 14 | 11:30 AM - 12:11 PM Eastern Daylight Time Duration 41 minutes

Speaker:  **Satya Nadella** | Microsoft Corporation

As we've seen over the past year, digital adoption curves are accelerating across every industry and business function. In this context, Satya will talk about what differentiates the Microsoft Cloud and the investments and innovation we are making across every layer to unlock new opportunity for partners to help our mutual customers transform in this new era.



-  Modern work
-  Digital & app innovation
-  Business applications
-  Data & AI
-  Infrastructure
-  Security

02:28 / 41:45

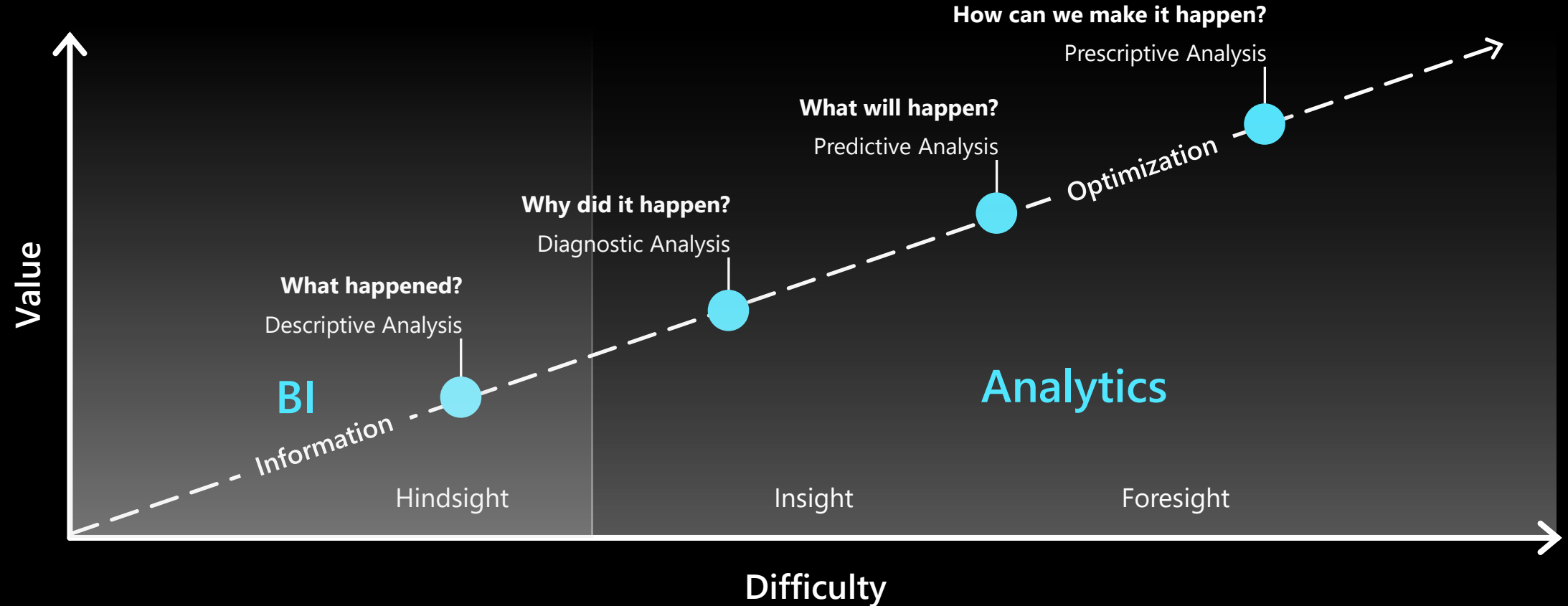
"Our ambition is to foster foundational innovation that creates entire ecosystems even greater than the platforms themselves. That's what the Microsoft cloud uniquely enables."

## Microsoft Cloud

Trust in technology  
and business model  
alignment

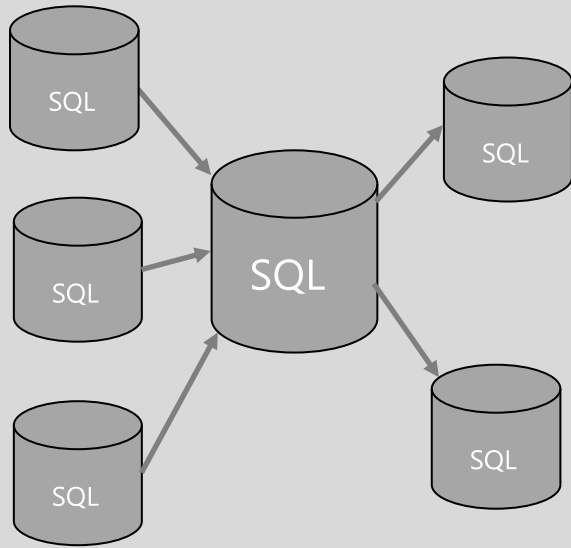
Integrated  
and open across  
every layer

# Modern Analytics Journey

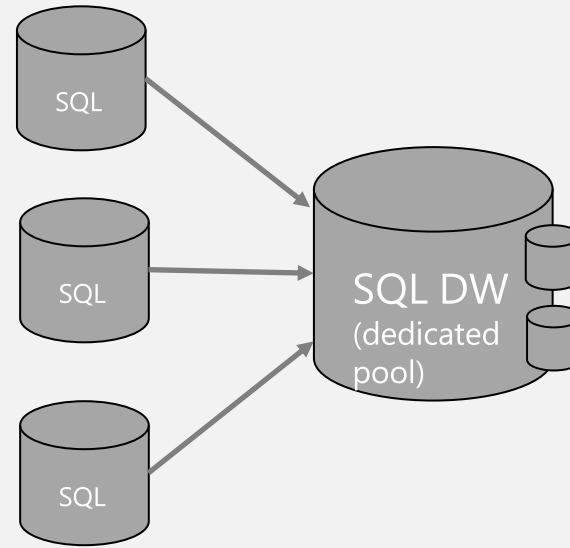


# Data warehousing evolution

on-prem, datacenter, IaaS



PaaS (cloud-based, managed services)

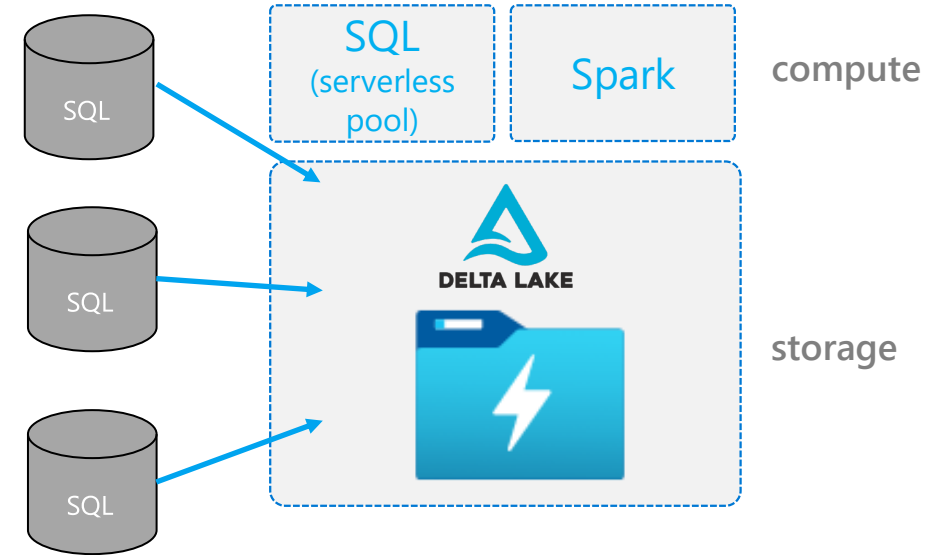


transactional

analytical

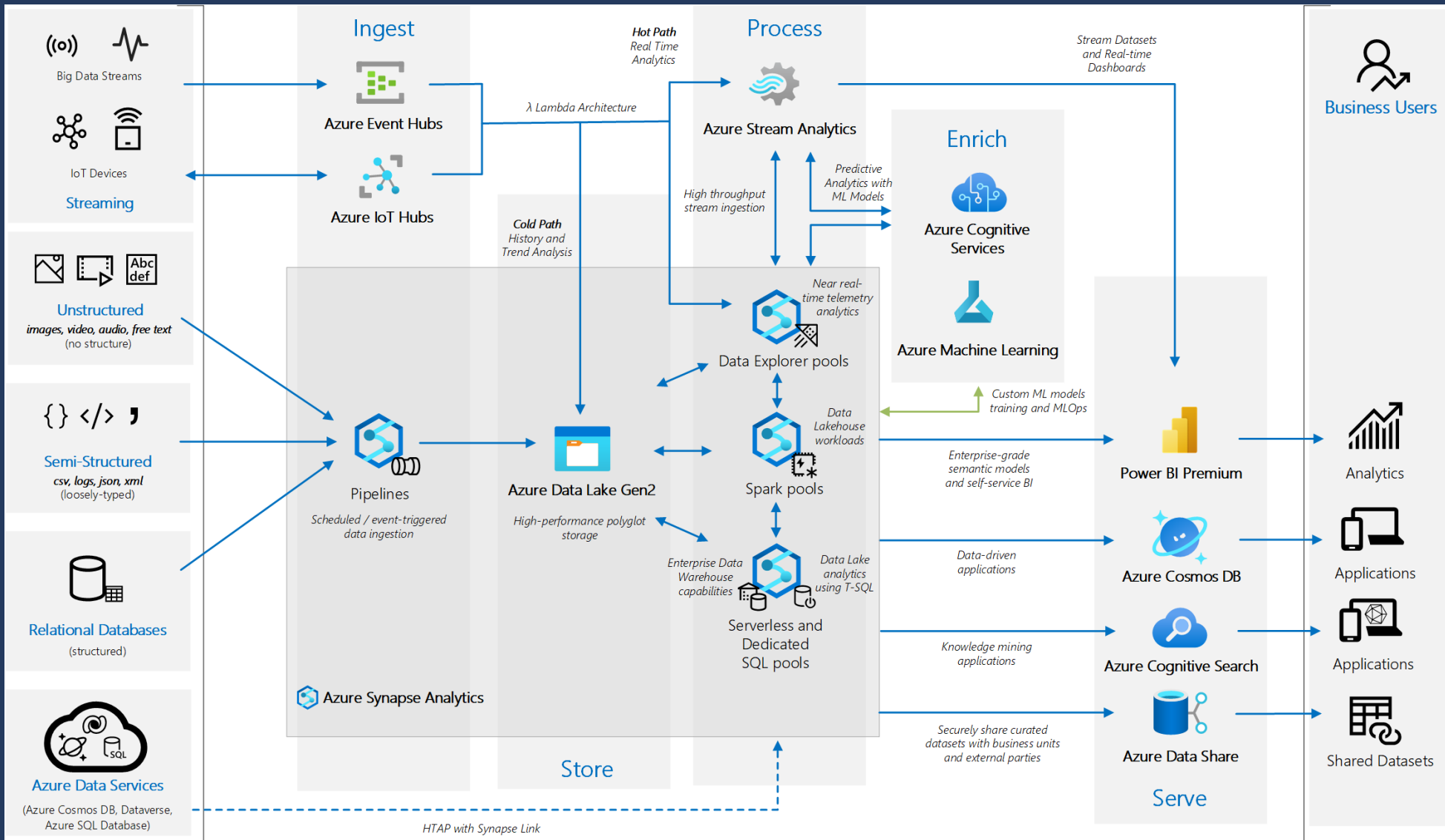


Lakehouse



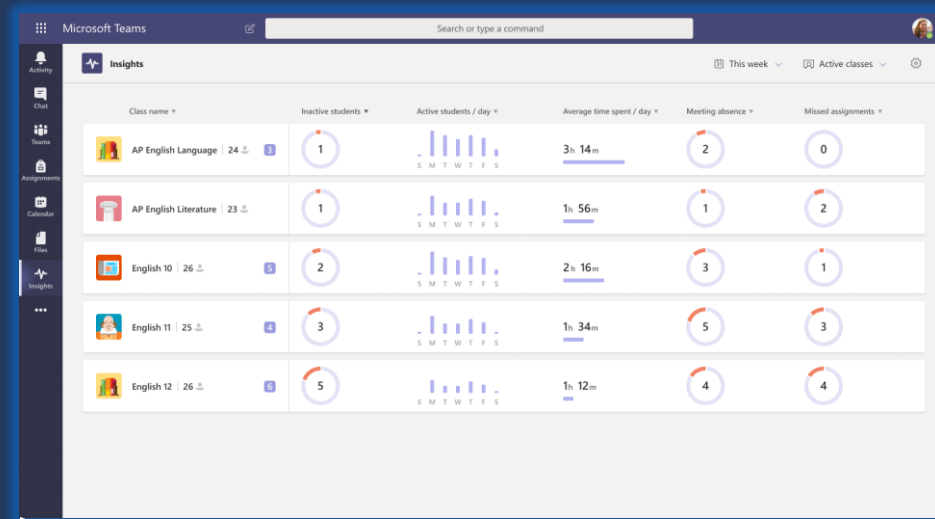
transactional

analytical

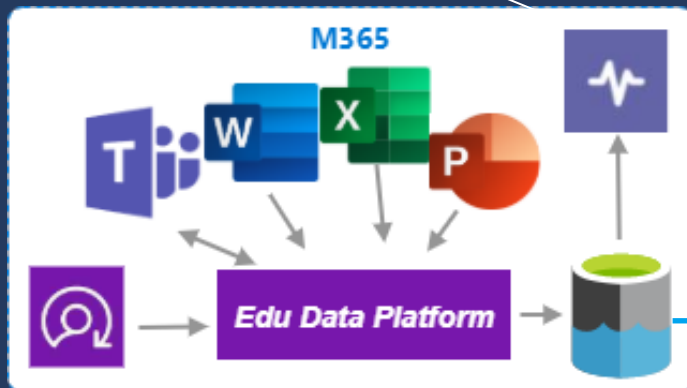




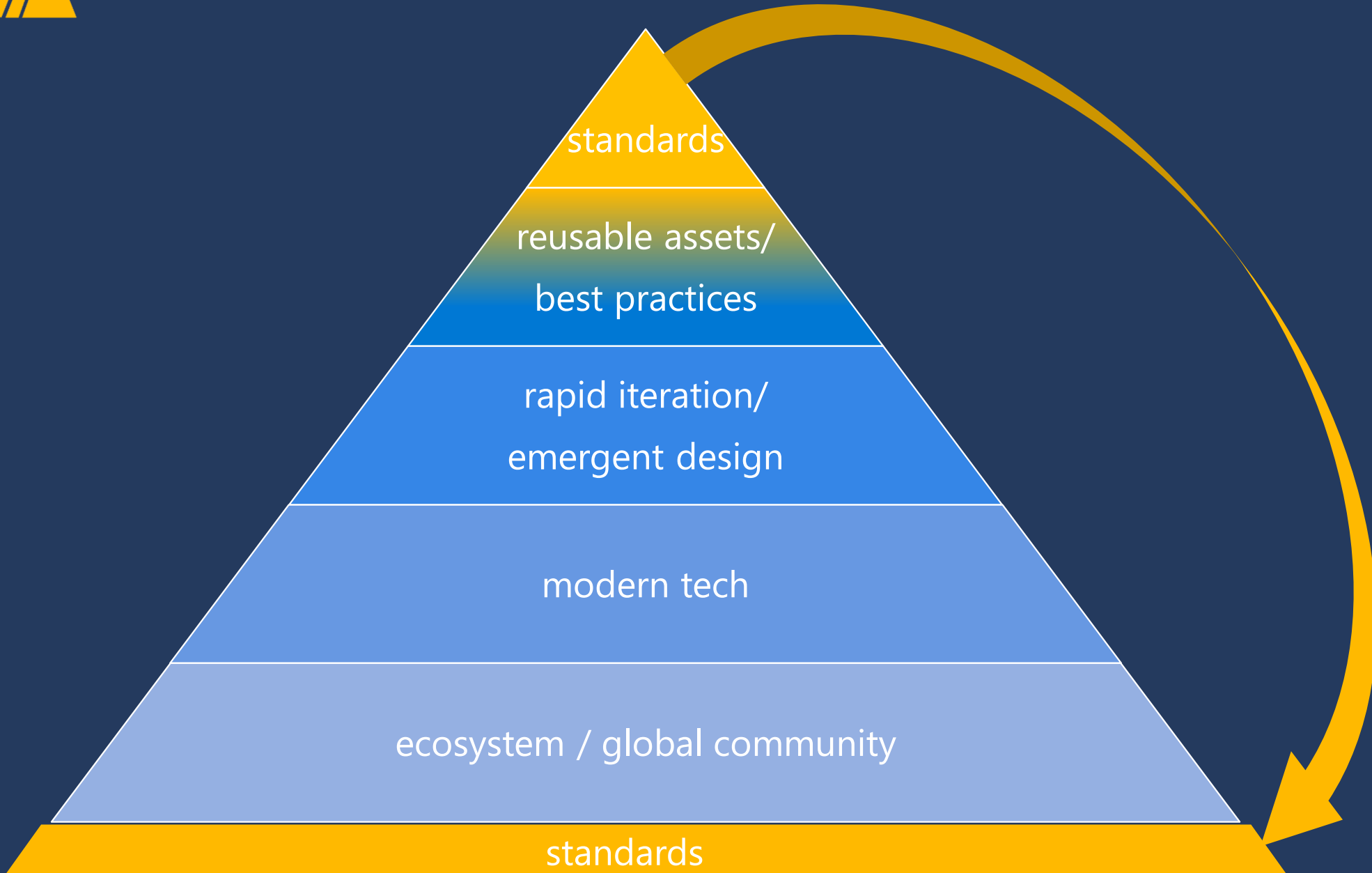
# Education Analytics and Insights



SIS



Customer's Azure Subscription



  
Data Engineers

  
Data Scientists

  
Analysts



## Open Education Analytics

Azure Synapse Analytics



Power BI



  
Administrators

  
Teachers,  
Faculty

  
Students,  
Parents

1) Collect

2) Ingest

3) Analyze

4) Publish

5) Inform

  
Data Engineers

  
Data Scientists

  
Analysts



## Open Education Analytics

data processing



Synapse Analytics

storage



Data Lake Storage

data governance



Purview

authentication/authorization



Azure Active Directory

reporting



Power BI



Administrators




Teachers,  
Faculty



Students,  
Parents

  
on-prem data

  
cloud data

  
SaaS data

1) Collect

2) Ingest

3) Analyze

4) Publish

5) Inform



Data Engineers



Data Scientists



Analysts

on-prem data

cloud data

SaaS data

modules, schemas, packages



framework



data processing



Synapse Analytics

storage



Data Lake Storage

data governance



Purview

authentication/authorization



Azure Active Directory

reporting



Power BI

Administrators



Teachers,  
Faculty



Students,  
Parents

1) Collect

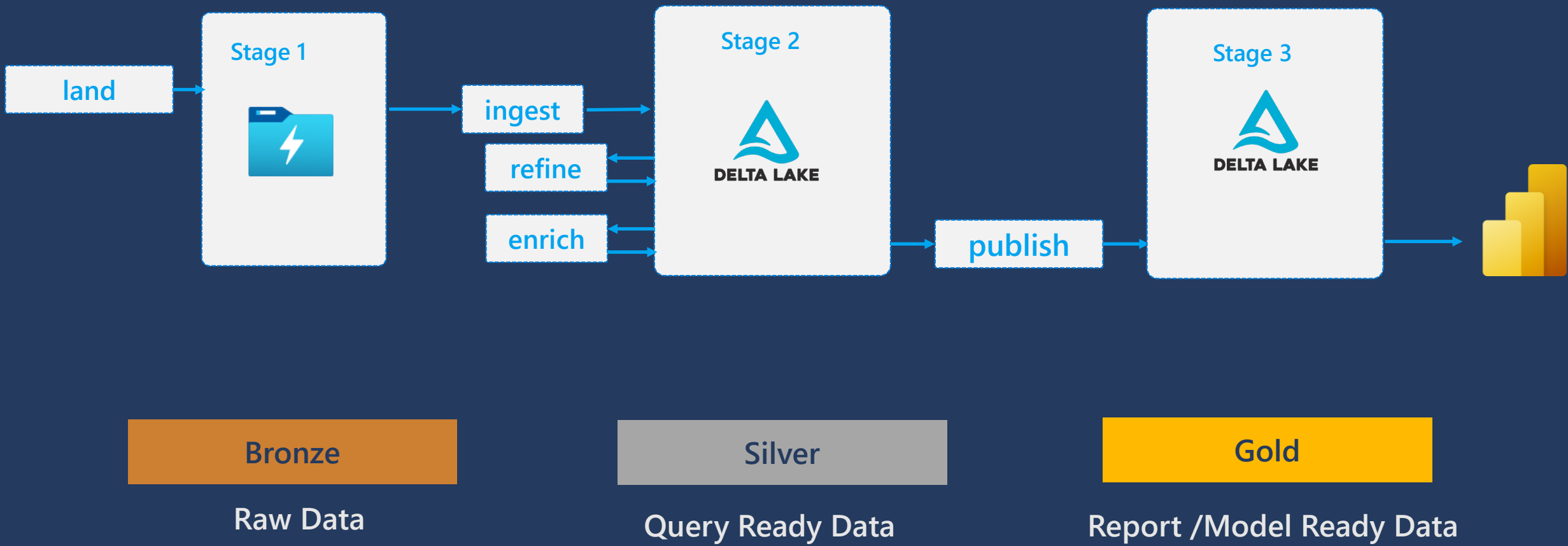
2) Ingest

3) Analyze

4) Publish

5) Inform

# Data lake structure



# Processing Steps Defined

## 1) land

- extract from source system
- land in original format, in batches
- no validation done at this point, no transformations

## 2) ingest

- process inbound batch files (parse and read into a dataframe)
- sanity check (validation only at the batch level to ensure it's a reasonable dataset to be merged in)
- add unique id at the entity level if needed (when merging data from multiple db's, eg EdFi)
- no data loss (no data is dropped on the floor)
- merge into delta lake (rename columns to remove spaces to satisfy parquet requirements)

## 3) refine

- include schema info (data types)
- transformations (json flattening, generate hkey values)
- pseudonymization

## 4) enrich

- validation
- transformations, aggregations
- additional enrichment from other data sources
- map to standard schemas

## 5) publish

- publish to data products
- transformations to build star schema
- compute aggregations

## Stage 1

1) land

### Transactional

```
| -EdFi
|--v5.3
|---DistrictId=255901
|----SchoolYear=2017
|-----courses
|-----delta_batch_data
|-----upserts
|-----rundate=2022-09-13
|----->data-d782.json
|-----deletes
|-----rundate=2022-09-13
|----->data-ab34.json
```

```
| -GraphAPI
|--v1.3
|---SchoolYear=2017
|-----userlogins
|-----additive_batch_data
```



Raw Data

2) ingest

3) refine

4) enrich

## Stage 2

### Ingested

```
| -EdFi
|--v5.3
|---courses
|----_delta_log
|----SchoolYear=2017
|-----DistrictId=255901
|----->part-000.parquet
|---students
```

### Refined

```
| -EdFi
|--v5.3
|---Sensitive
|----courses_lookup
|-----_delta_log
|-----SchoolYear=2017
|-----DistrictId=255901
|----->part-000.parquet
|---General
|----courses
|-----_delta_log
|-----SchoolYear=2017
```

### Enriched

```
| -EdFiAnalytics
|--v0.1
|---Sensitive
|----students_lookup
|---General
|----students
```



Enriched/Query ready Data  
(enriched datasets)

5) publish

## Stage 3

### EdFiAMT

```
| -v1.2
|--General
|---students
|----_delta_log
```

### EdFiByEdGraph

```
| -v2.3
```



Report /Model Ready Data  
(data products)





## Stage 1

### Transactional

```
|-EdFi
|--v5.3
|---DistrictId=255901
|----SchoolYear=2017
|-----courses
|-----delta_batch_data
|-----upserts
|-----rundate=2022-09-13
|----->data-d782.json
|-----deletes
|-----rundate=2022-09-13
|----->data-ab34.json

|-GraphAPI
|--v1.3
|---SchoolYear=2017
|----userlogins
|-----additive_batch_data
```

1) land

modules

2) ingest

3) refine

4) enrich

## Stage 2

### Ingested

```
|-EdFi
|--v5.3
|---courses
|----_delta_log
|----SchoolYear=2017
|-----DistrictId=255901
|----->part-000.parquet
|---students
```

### Refined

```
|-EdFi
|--v5.3
|---Sensitive
|----courses_lookup
|-----_delta_log
|-----SchoolYear=2017
|-----DistrictId=255901
|----->part-000.parquet
|---General
|----courses
|-----_delta_log
|-----SchoolYear=2017
```

### Enriched

```
|-EdFiAnalytics
|--v0.1
|---Sensitive
|----students_lookup
|---General
|----students
```

schemas

5) publish

## Stage 3

### EdFiAMT

```
|-v1.2
|--General
|---students
|----_delta_log
```

### EdFiByEdGraph

```
|-v2.3
```

Raw Data

Enriched/Query ready Data  
(enriched datasets)

Report /Model Ready Data  
(data products)

