ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

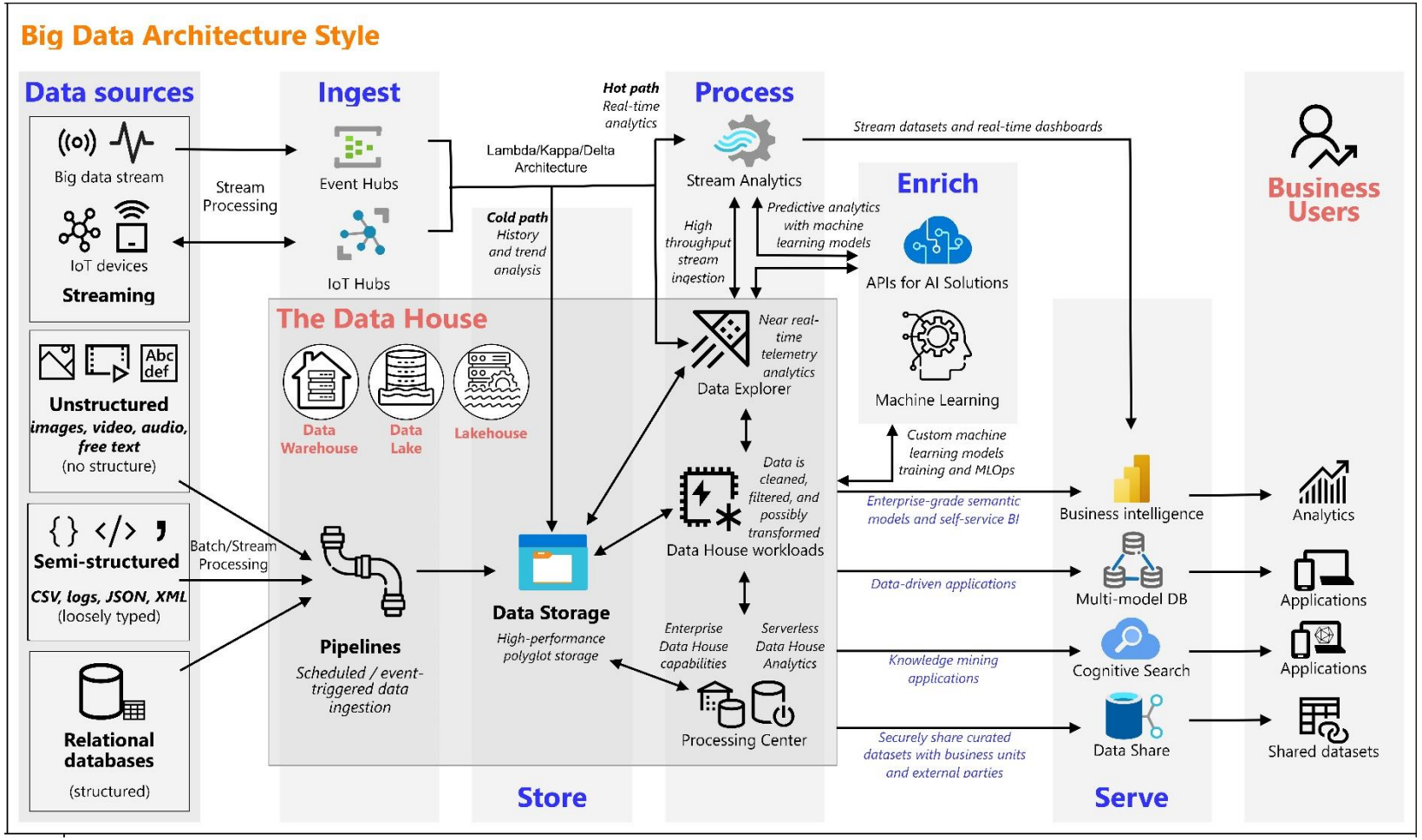
CHƯƠNG 2
Kiến trúc dữ liệu lớn trong doanh nghiệp

Biên soạn: ThS. Nguyễn Thị Anh Thư

Nội dung

1. Tổng quan
2. Thành phần và cách thức hoạt động
3. Công nghệ triển khai
4. Kiến trúc dữ liệu lớn nổi bật
5. Vấn đề và xu hướng
6. Tổng kết

1. Tổng quan



Kiến trúc dữ liệu lớn là một khuôn khổ toàn diện để quản lý, lưu trữ, xử lý và phân tích hiệu quả các tập dữ liệu lớn và phức tạp.

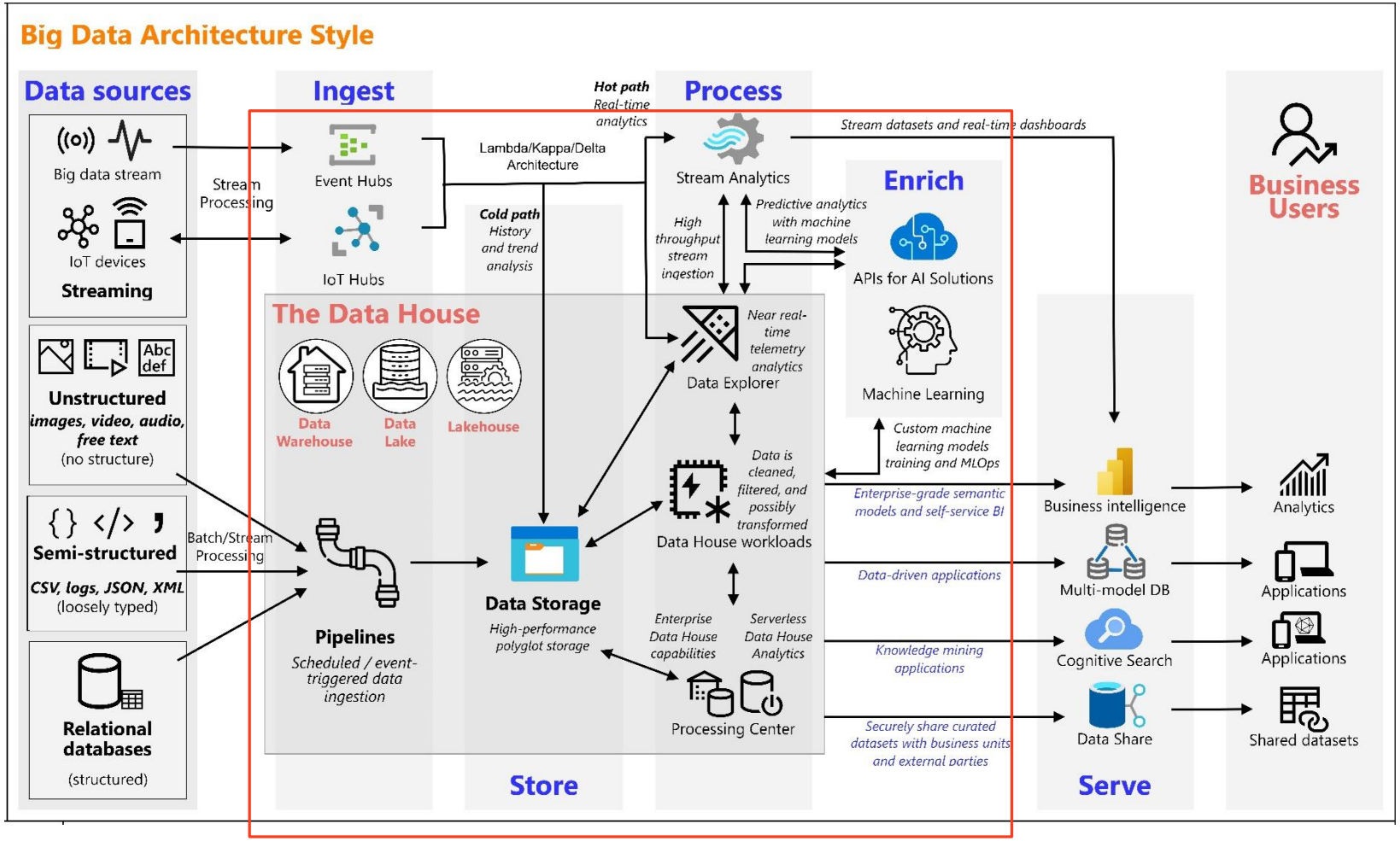
Bao gồm nhiều thành phần, công cụ và kỹ thuật khác nhau giúp các tổ chức xử lý lượng dữ liệu khổng lồ một cách hiệu quả.

1. Tổng quan

1. **Thu thập dữ liệu (Ingest):** Dữ liệu có thể đến từ nhiều nguồn khác nhau.
2. **Lưu trữ dữ liệu (Store):** Dữ liệu được lưu trữ trong các hệ thống lưu trữ phù hợp.
3. **Xử lý dữ liệu (Process):** Dữ liệu được xử lý để chuyển đổi, làm sạch và chuẩn bị cho việc phân tích.
4. **Phân tích dữ liệu (Enrich):** Dữ liệu được phân tích để trích xuất thông tin chi tiết và đưa ra quyết định.

➢**Điều phối và quản trị dữ liệu:** Chỉ huy và tự động hóa các bước khác nhau trong kiến trúc dữ liệu lớn. Quản trị dữ liệu đảm bảo dữ liệu được bảo mật, chất lượng và tuân thủ các quy định.

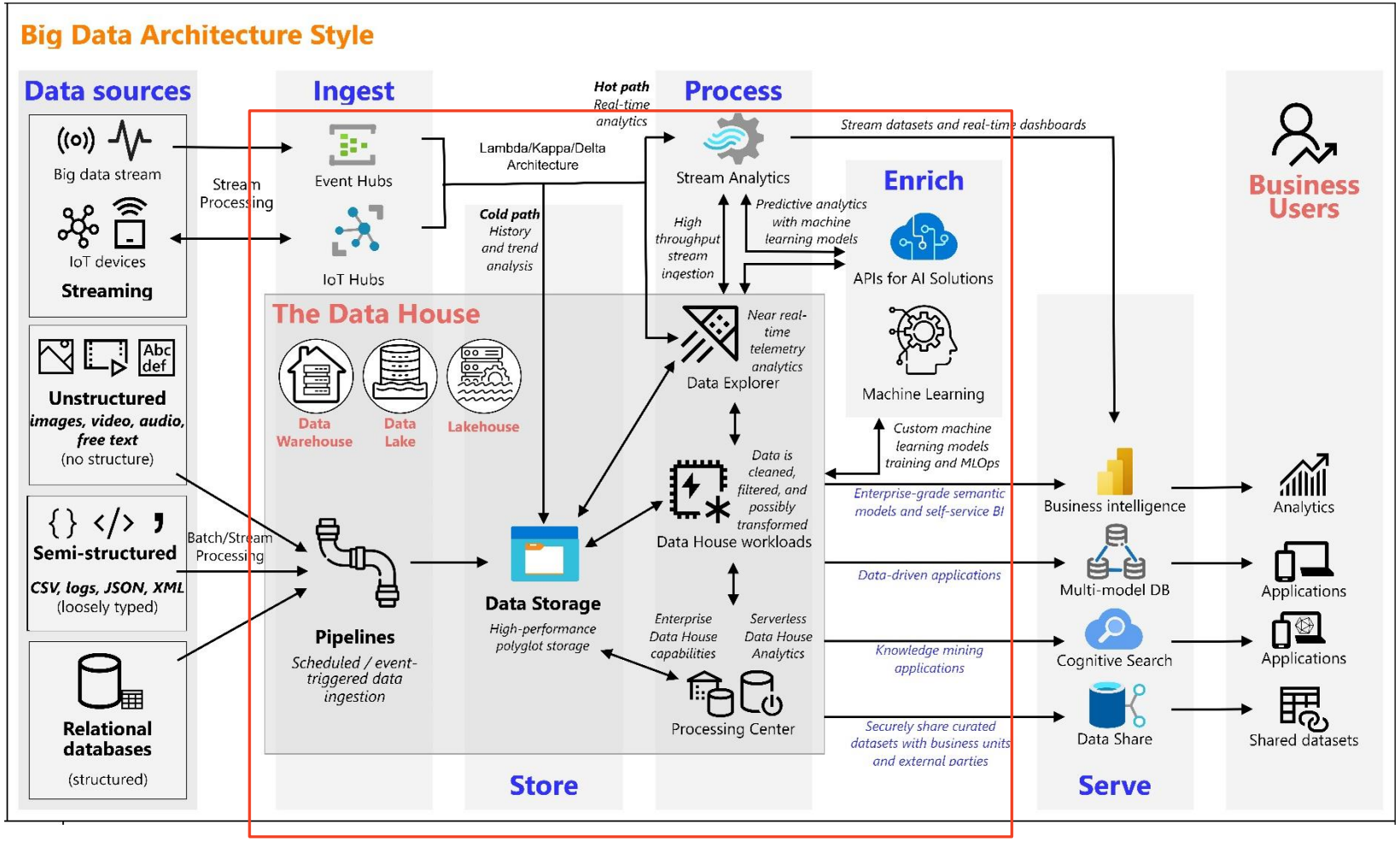
1. Tổng quan



Vòng đời dữ liệu lớn xác định các giai đoạn mà dữ liệu trải qua từ khi tạo đến khi được sử dụng để tạo ra giá trị.

Kiến trúc dữ liệu lớn cung cấp nền tảng cho việc quản lý và xử lý dữ liệu trong suốt vòng đời của dữ liệu.

1. Tổng quan



Quản trị dữ liệu: Quản trị dữ liệu đảm bảo dữ liệu được bảo mật, chất lượng và tuân thủ các quy định.

- **Bảo mật dữ liệu**: Bảo vệ dữ liệu khỏi truy cập trái phép
- **Chất lượng dữ liệu**: Đảm bảo dữ liệu chính xác và đầy đủ
- **Tuân thủ quy định**: Đảm bảo dữ liệu tuân thủ các quy định

2. Thành phần và cách thức hoạt động

Thu thập dữ liệu (Ingest)

Dữ liệu nội bộ

- Dữ liệu giao dịch, dữ liệu khách hàng, dữ liệu hoạt động, … từ các hệ thống thông tin

Dữ liệu bên ngoài

- Dữ liệu thị trường, dữ liệu mạng xã hội, dữ liệu cảm biến, …

Tính chất	Structured data	Semi-structured data	Unstructured data
Cấu trúc	Được định nghĩa trước, dạng bảng	Có tổ chức nhưng không cố định	Không có cấu trúc
Ví dụ	Dữ liệu khách hàng, bảng tính	JSON, XML, email có tag	Email, văn bản, hình ảnh, video
Ưu điểm	Dễ dàng tìm kiếm, phân tích	Linh hoạt, dễ lưu trữ, chia sẻ	Giàu thông tin, đa dạng
Nhược điểm	Ít linh hoạt, thiết kế phức tạp	Cần xử lý để trích xuất nghĩa	Khó tìm kiếm, phân tích
Lưu trữ	Cơ sở dữ liệu quan hệ	File, database NoSQL	File, database NoSQL
Công cụ phân tích	SQL, công cụ BI truyền thống	Công cụ phân tích chuyên biệt	Công cụ phân tích chuyên biệt

2. Thành phần và cách thức hoạt động

Thu thập dữ liệu (Ingest)

- Dựa trên cách thức tổ chức hoặc cấu trúc dữ liệu được phân loại thành ba dạng chính.

Tính năng	Xử lý Theo Mảng (Batch Processing)	Xử lý Dòng (Stream Processing)
Kiểu xử lý	Theo nhóm (batch)	Liên tục theo dòng (stream)
Thời gian xử lý	Sau khi thu thập dữ liệu	Ngay khi dữ liệu được tạo ra
Độ trễ (latency)	Cao	Thấp (thời gian thực)
Phù hợp cho	Phân tích phức tạp, cần độ chính xác cao	Phản hồi nhanh, nhận dạng thay đổi
Ví dụ	Báo cáo tài chính, phân tích nhật ký	Phát hiện gian lận, đề xuất sản phẩm

2. Thành phần và cách thức hoạt động

Thu thập dữ liệu (Ingest)

Phương pháp phổ biến để xử lý dữ liệu lớn (big data) theo thời gian thực:

- Lambda architecture
- Kappa architecture
- Delta architecture

2. Thành phần và cách thức hoạt động

Thu thập dữ liệu (Ingest)

Kiến trúc	Mô tả	Cách tiếp cận	Ưu điểm	Nhược điểm
Lambda	Kiến trúc phân tán dữ liệu theo batch	Lớp Batch: Xử lý dữ liệu theo từng đợt (batch) để có kết quả chính xác nhất. Lớp Speed: Xử lý dữ liệu theo thời gian thực để cung cấp kết quả nhanh chóng. Lớp Serving: Cung cấp kết quả cho các truy vấn.	Dễ triển khai, chi phí thấp	Khó xử lý dữ liệu thời gian thực
Kappa	Kiến trúc phân tán dữ liệu theo stream	Xử lý dữ liệu theo từng luồng (stream) riêng biệt, mỗi luồng có cấu trúc và ngữ nghĩa riêng.	Xử lý dữ liệu thời gian thực hiệu quả	Phức tạp hơn Lambda, chi phí cao hơn
Delta	Kiến trúc kết hợp Lambda và Kappa	Xử lý dữ liệu theo từng luồng (stream) riêng biệt, nhưng tất cả các luồng đều có cấu trúc và ngữ nghĩa thống nhất.	Cân bằng giữa hiệu quả và chi phí	Phức tạp hơn Lambda, khó triển khai hơn Kappa

2. Thành phần và cách thức hoạt động

Lưu trữ dữ liệu (Store)

Tính năng	Data Warehouse	Data Lake	Lakehouse
Dữ liệu	Kho lưu trữ tập trung dữ liệu được tổ chức, tối ưu hóa cho truy vấn và phân tích. Dữ liệu thường được xử lý trước (làm sạch, chuyển đổi, tải) và lưu trữ theo cấu trúc cố định.	Kho lưu trữ tập trung dữ liệu ở dạng thô, chưa được xử lý. Dữ liệu có thể được lưu trữ ở bất kỳ định dạng nào, có cấu trúc hoặc không có cấu trúc.	Sự kết hợp giữa Data Warehouse và Data Lake. Nó cung cấp sự linh hoạt và khả năng mở rộng của Data Lake cùng với hiệu suất truy vấn và khả năng phân tích của Data Warehouse.
Cấu trúc	Cố định	Linh hoạt	Linh hoạt
Truy vấn	Nhanh	Chậm	Nhanh
Phân tích	Dễ dàng	Khó	Dễ dàng
Linh hoạt	Thấp	Cao	Cao
Khả năng mở rộng	Thấp	Cao	Cao
Chi phí	Cao	Thấp	Trung bình
Kỹ năng	Trung bình	Cao	Cao

2. Thành phần và cách thức hoạt động

Xử lý dữ liệu (Process)

Chuyển đổi dữ liệu

Chuẩn hóa và chia tỷ lệ

Tạo đặc trưng

Làm sạch dữ liệu

Xử lý dữ liệu thiếu

Phát hiện và xử lý ngoại lệ

Xóa trùng lặp

Xác thực dữ liệu

Chuẩn bị dữ liệu

Định dạng dữ liệu

Lấy mẫu dữ liệu

Chia tách dữ liệu

Hiểu rõ dữ liệu

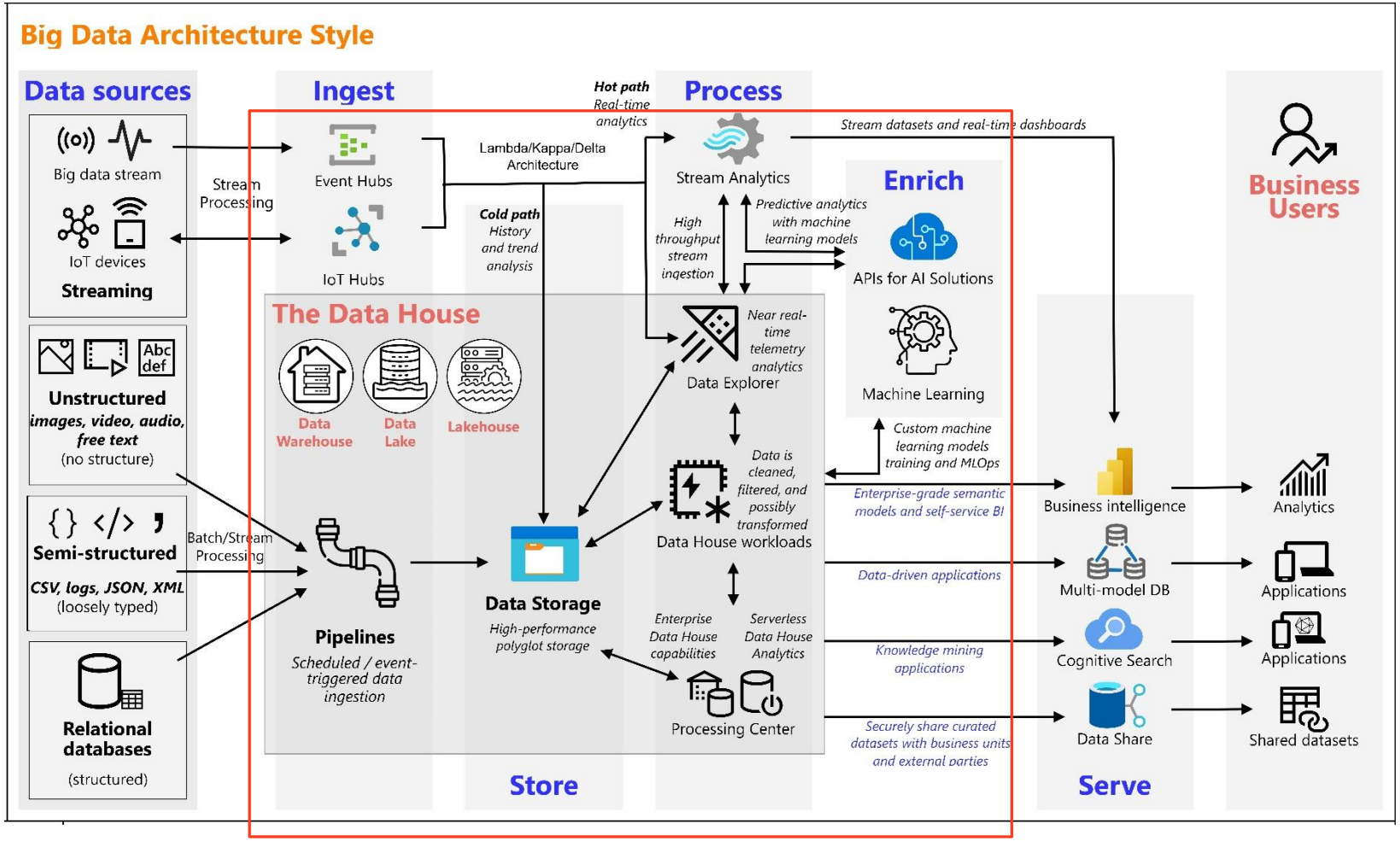
Tài liệu

Lặp lại

2. Thành phần và cách thức hoạt động

Phân tích dữ liệu (Enrich)

- **Thống kê mô tả**: Tóm tắt các đặc điểm chính của tập dữ liệu.
- **Thống kê suy luận**: Để đưa ra kết luận về tổng thể dựa trên một mẫu dữ liệu.
- **Phân tích hồi quy**: Dự đoán về giá trị của biến phụ thuộc dựa trên giá trị của các biến độc lập.
- **Phân tích dữ liệu chuỗi thời gian**: Dự báo các giá trị trong tương lai dựa trên dữ liệu trong quá khứ.
- **Phân tích tổ hợp**: So sánh các nhóm cá nhân theo thời gian.
- **Phân tích cụm**: Nhóm các điểm dữ liệu thành các cụm dựa trên sự giống nhau của chúng.
- **Phân tích văn bản**: Trích xuất ý nghĩa từ dữ liệu văn bản.
- …



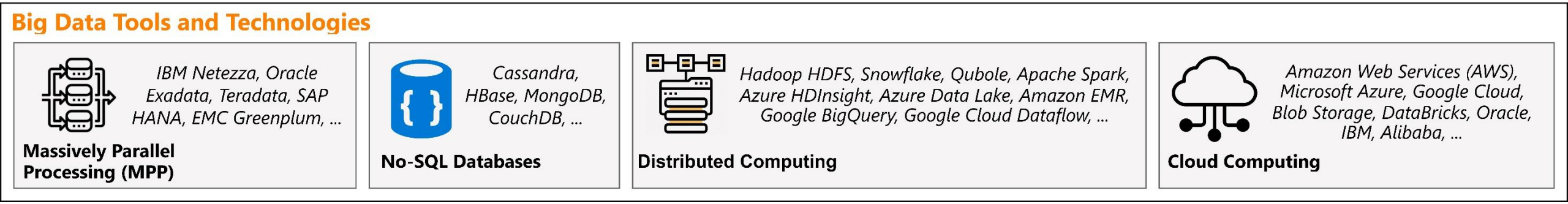
2. Thành phần và cách thức hoạt động

Điều phối (Orchestration)

- Đóng vai trò chỉ huy, điều phối và tự động hóa việc thực hiện các bước khác nhau liên quan đến vòng đời dữ liệu lớn.

3. Công nghệ triển khai

- **Massively parallel processing** (MPP) là một kiến trúc máy tính sử dụng hàng trăm hoặc hàng nghìn bộ xử lý riêng biệt để thực hiện các phép tính song song trên dữ liệu khổng lồ.
- **NoSQL database** là một loại cơ sở dữ liệu lưu trữ dữ liệu theo cách khác với mô hình quan hệ truyền thống sử dụng bảng và mối quan hệ giữa các bảng.
- **Distributed Computing** là một mô hình kiến trúc nhiều máy tính riêng biệt được kết nối với nhau thông qua mạng để cùng nhau thực hiện một tác vụ chung.
- **Cloud Computing** là mô hình cung cấp dịch vụ máy tính như máy chủ, lưu trữ, mạng, và ứng dụng thông qua internet.



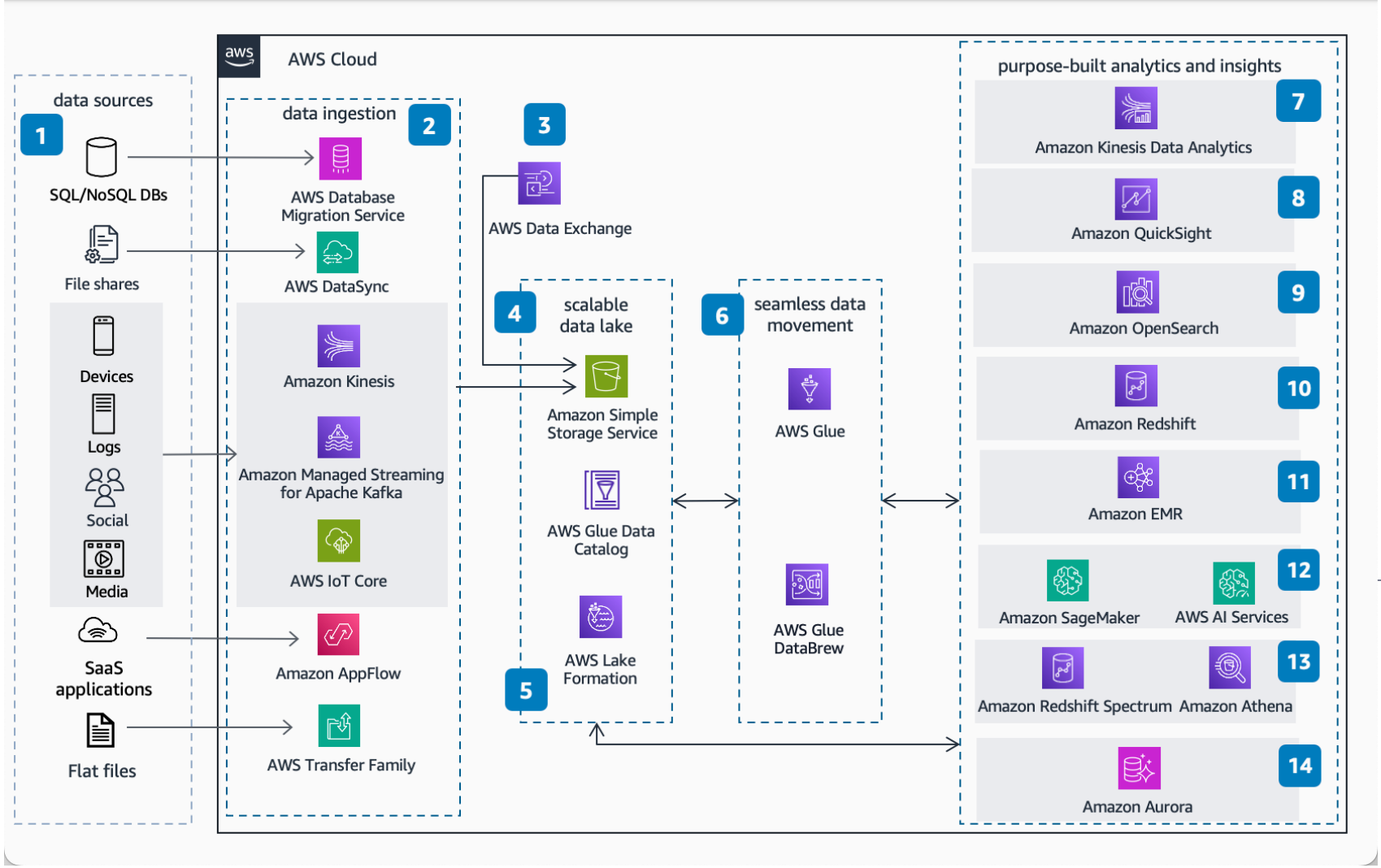
3. Công nghệ triển khai

Cloud Computing

- *Azure*
- *Google Cloud*

Ngôn ngữ lập trình

- *Python*: Có nhiều thư viện và công cụ dành cho khoa học dữ liệu, học máy và phân tích dữ liệu lớn.
- *R*: Ngôn ngữ chuyên dụng cho thống kê và phân tích dữ liệu.

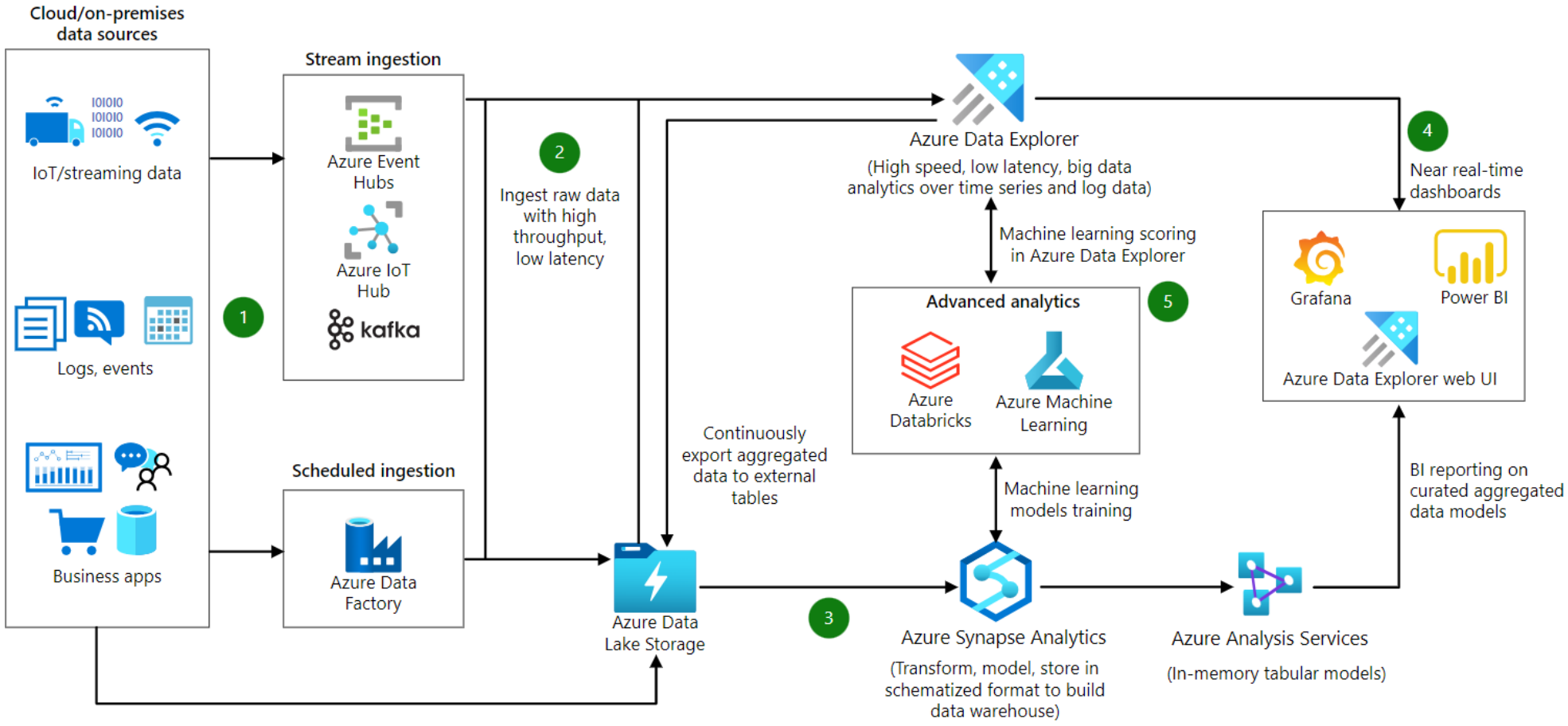


4. Kiến trúc dữ liệu lớn nổi bật

AWS

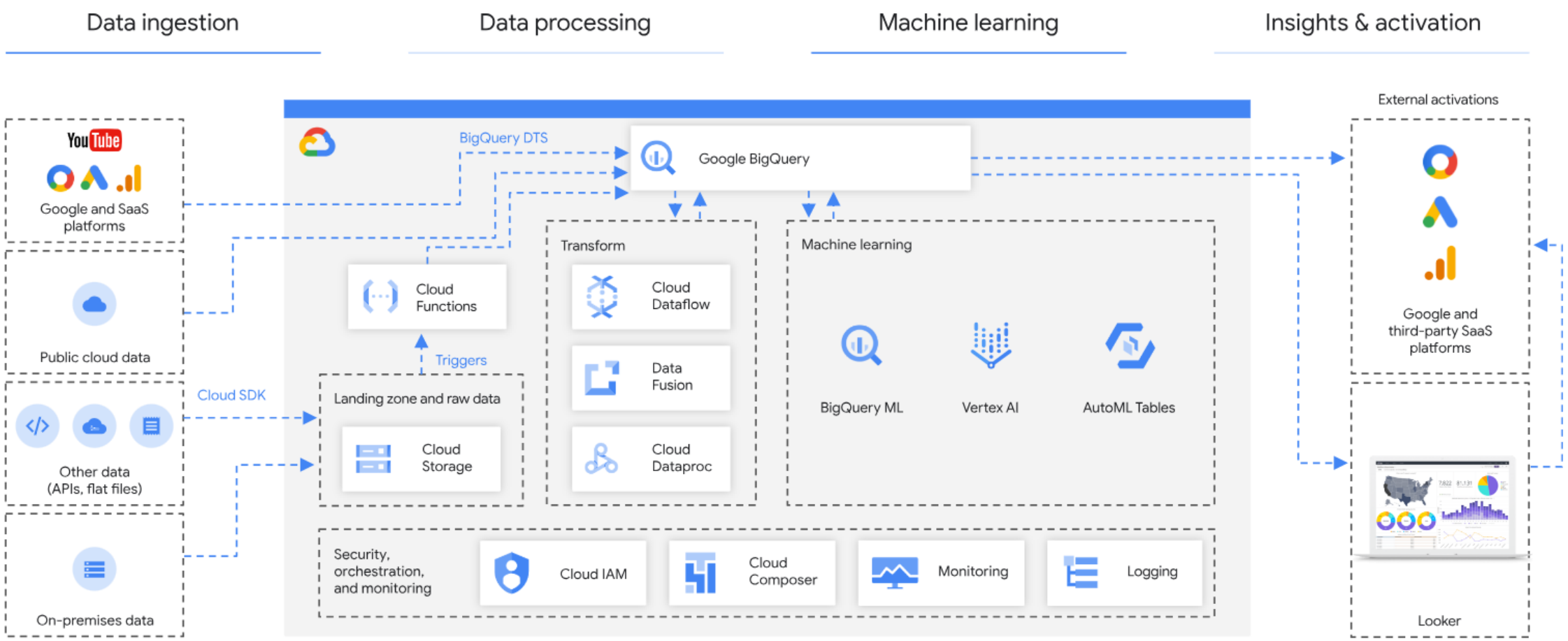
4. Kiến trúc dữ liệu lớn nổi bật

Azure Data Explorer

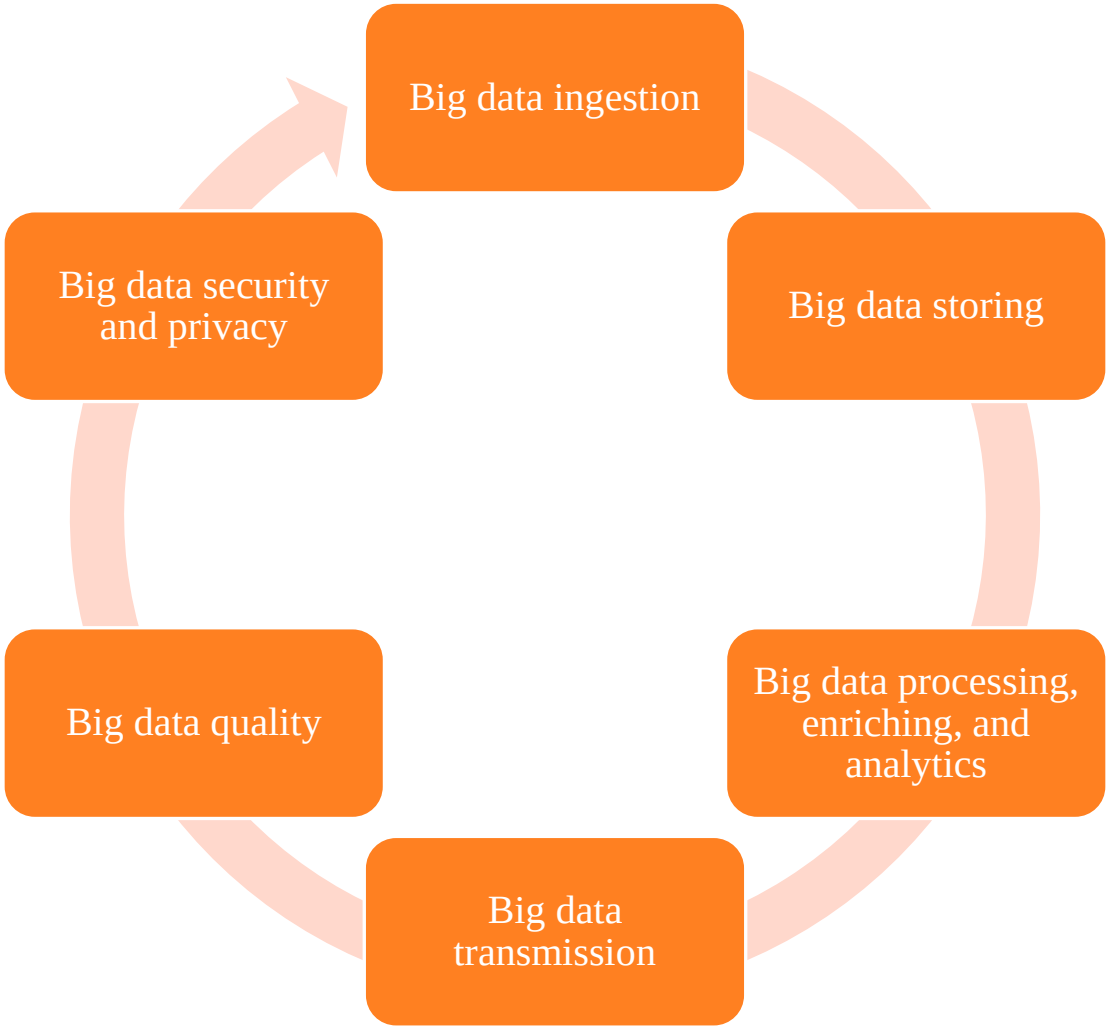


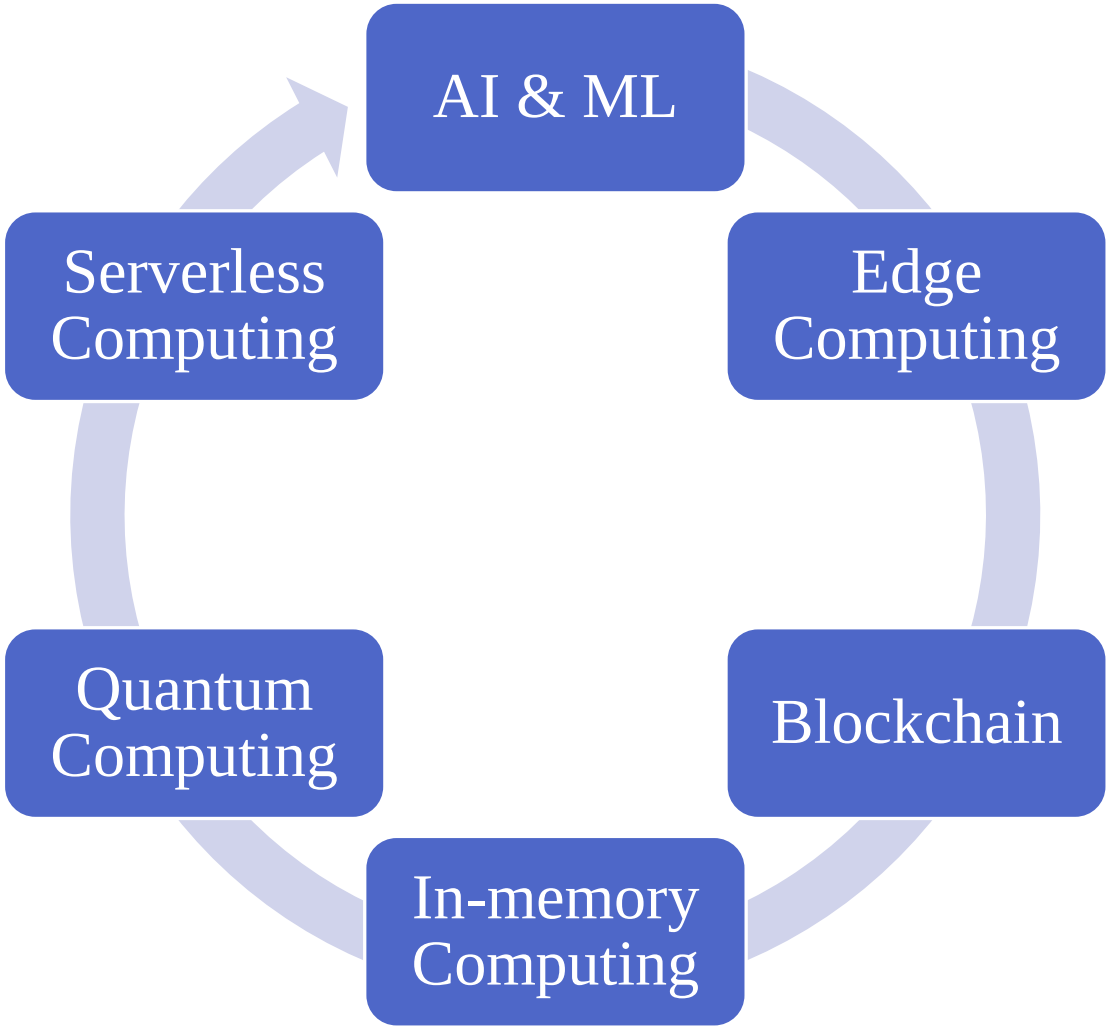
4. Kiến trúc dữ liệu lớn nổi bật

Google Cloud

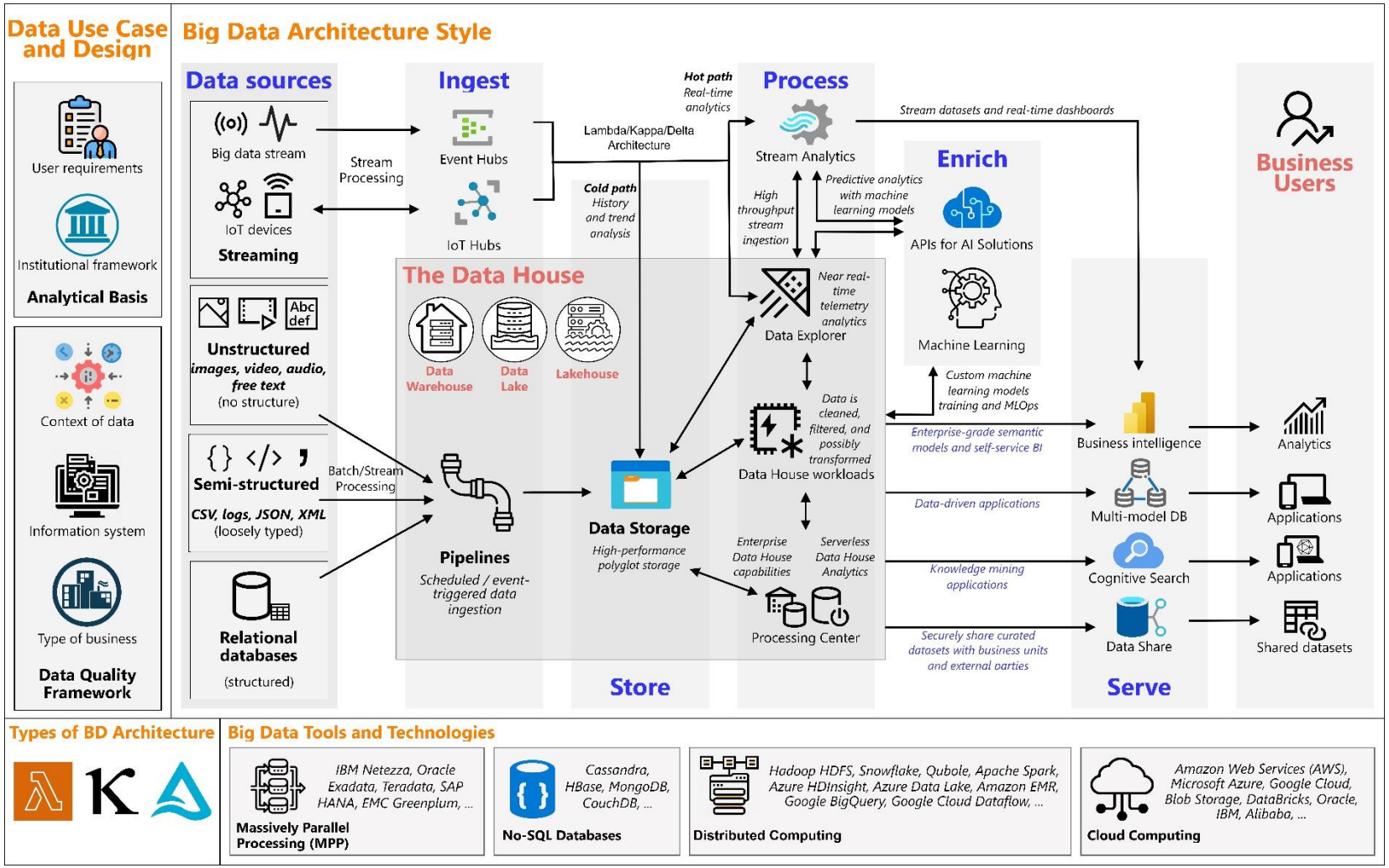


5. Vấn đề và xu hướng





6. Tổng kết


Data Use Case and Design
User requirements
Institutional framework
Analytical Basis
Context of data
Information system
Type of business
Data Quality Framework
Big Data Architecture Style
Data sources
Big data stream
IoT devices
Streaming
Unstructured
images, video, audio, free text
(no structure)
Semi-structured
CSV, logs, JSON, XML
(loosely typed)
Relational databases
(structured)
Stream Processing
Batch/Stream Processing
Ingest
Event Hubs
IoT Hubs
Hot path
Real-time analytics
Lambda/Kappa/Delta Architecture
Cold path
History and trend analysis
The Data House
Data Warehouse
Data Lake
Lakehouse
Pipelines
Scheduled / event-triggered data ingestion
Data Storage
High-performance polyglot storage
Store
Process
Stream Analytics
Stream datasets and real-time dashboards
High throughput stream ingestion
Predictive analytics with machine learning models
Enrich
APIs for AI Solutions
Machine Learning
Near real-time telemetry analytics
Data Explorer
Data is cleaned, filtered, and possibly transformed
Data House workloads
Custom machine learning models training and MLOps
Enterprise Data House capabilities
Serverless Data House Analytics
Processing Center
Business Users
Business intelligence
Enterprise-grade semantic models and self-service BI
Analytics
Data-driven applications
Multi-model DB
Applications
Knowledge mining applications
Cognitive Search
Applications
Securely share curated datasets with business units and external parties
Data Share
Shared datasets
Serve
Types of BD Architecture
Big Data Tools and Technologies
IBM Netezza, Oracle Exadata, Teradata, SAP HANA, EMC Greenplum, ...
Massively Parallel Processing (MPP)
Cassandra, HBase, MongoDB, CouchDB, ...
No-SQL Databases
Hadoop HDFS, Snowflake, Qubole, Apache Spark, Azure HDInsight, Azure Data Lake, Amazon EMR, Google BigQuery, Google Cloud Dataflow, ...
Distributed Computing
Amazon Web Services (AWS), Microsoft Azure, Google Cloud, Blob Storage, DataBricks, Oracle, IBM, Alibaba, ...
Cloud Computing

Question & Answer