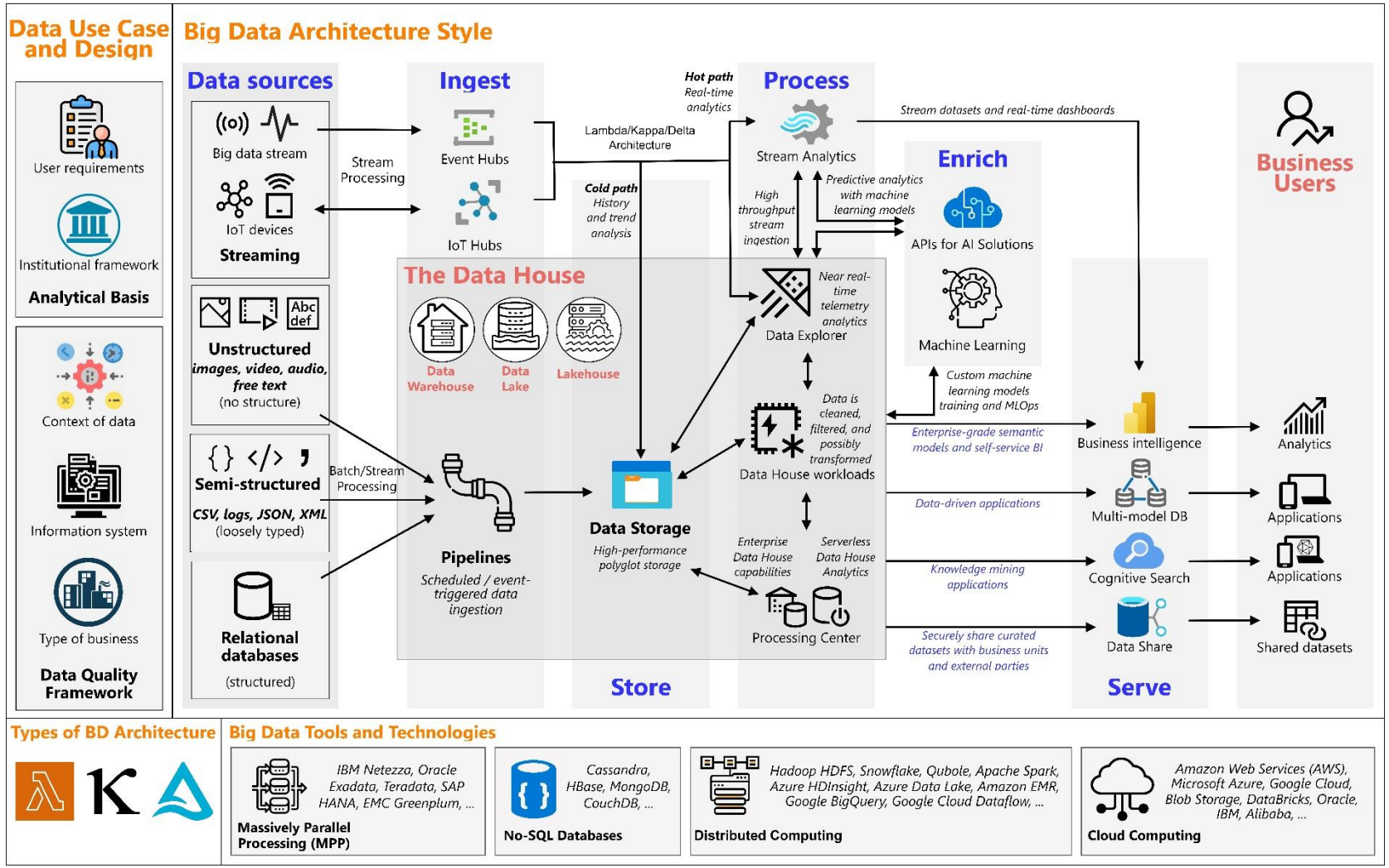
ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

CHƯƠNG 11
Bức tranh tổng quan về dữ liệu lớn trong doanh nghiệp

Biên soạn: ThS. Nguyễn Thị Anh Thư

Nội dung

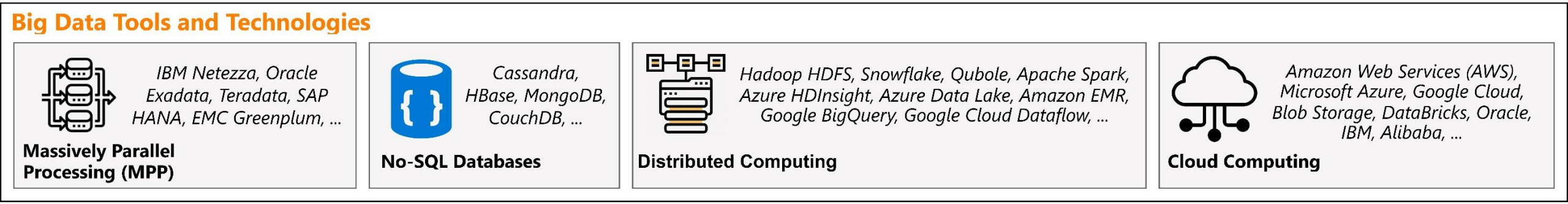
1. Giới thiệu
2. Công nghệ dữ liệu lớn
3. Thách thức và xu hướng
4. Tổng kết


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
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
Business Users
The Data House
Data Warehouse
Data Lake
Lakehouse
Near real-time telemetry analytics
Data Explorer
Custom machine learning models training and MLOps
Data is cleaned, filtered, and possibly transformed
Data House workloads
Pipelines
Scheduled / event-triggered data ingestion
Data Storage
High-performance polyglot storage
Enterprise Data House capabilities
Serverless Data House Analytics
Processing Center
Store
Enterprise-grade semantic models and self-service BI
Business intelligence
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

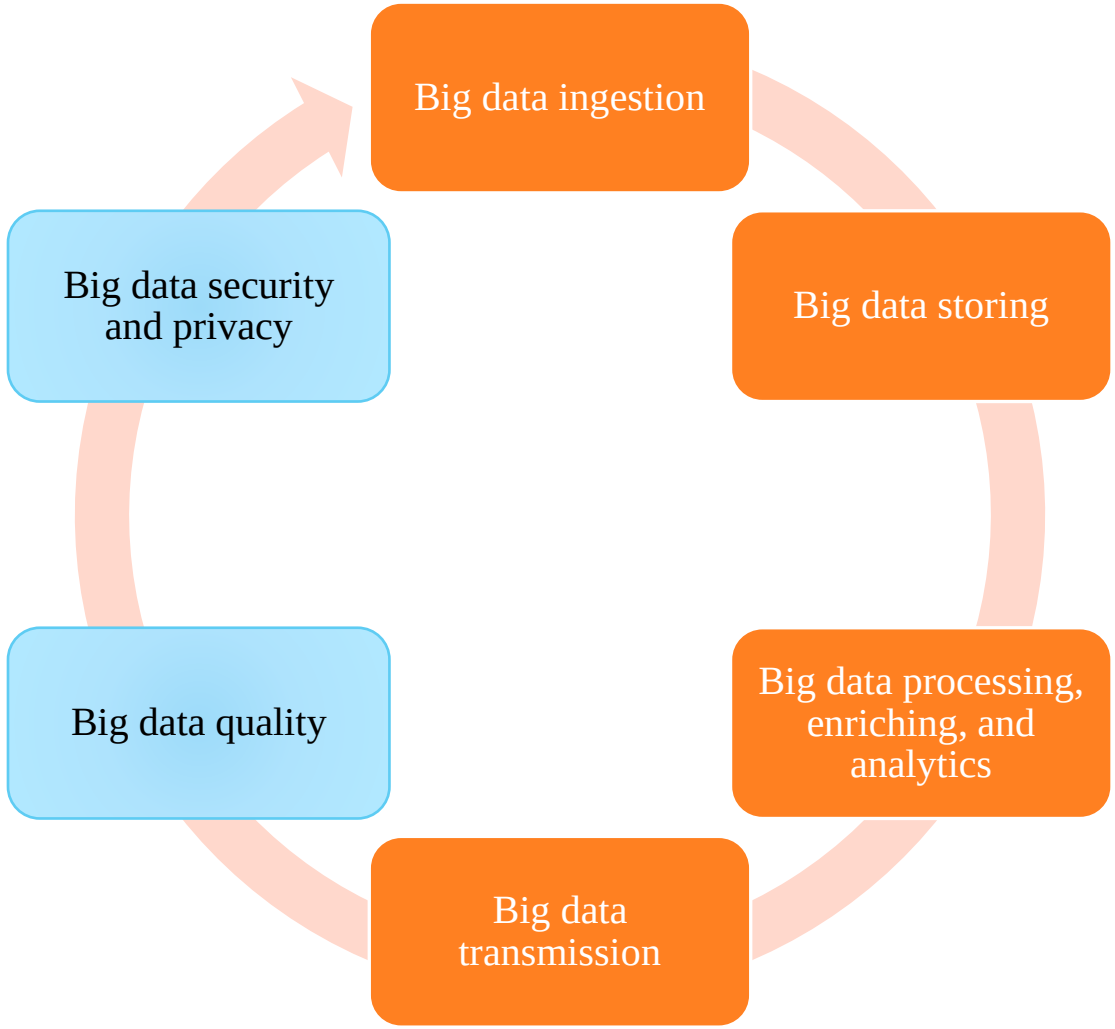
1. Giới thiệu

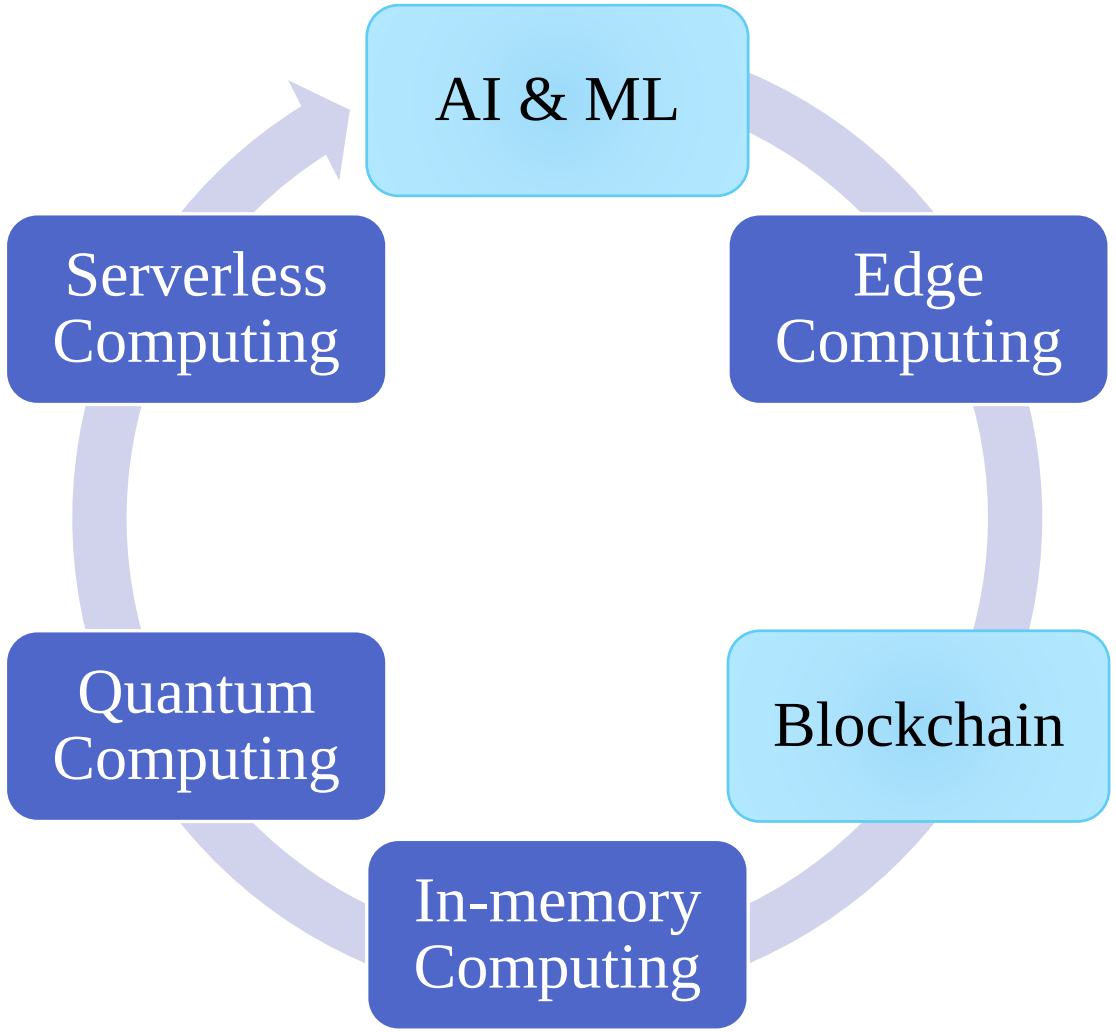
2. Công nghệ dữ liệu lớn

- **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. Thách thức và xu hướng





Xu hướng tương lai

AI & ML

- Những công nghệ này có thể tự động hóa việc nhập, làm sạch và chuyển đổi dữ liệu, nâng cao hiệu quả và độ chính xác.
- Thuật toán ML có thể được sử dụng để phát hiện sự bất thường, kiểm tra chất lượng dữ liệu và thậm chí là sắp xếp dữ liệu thông minh trong hệ thống lưu trữ.
- Phân tích dự đoán được hỗ trợ bởi AI có thể dự đoán nhu cầu lưu trữ và tối ưu hóa việc phân bổ tài nguyên.

Edge Computing

- Việc xử lý dữ liệu gần nguồn hơn, ở rìa mạng, có thể giảm bớt gánh nặng xử lý từ các máy chủ trung tâm và cải thiện khả năng xử lý dữ liệu theo thời gian thực.
- Điều này đặc biệt có thể áp dụng trong các tình huống liên quan đến khối lượng lớn dữ liệu cảm biến hoặc phân tích thời gian thực.

Xu hướng tương lai

Blockchain

- Công nghệ sổ cái phân tán an toàn này có thể được sử dụng để theo dõi nguồn gốc dữ liệu, đảm bảo tính toàn vẹn và tin cậy của dữ liệu trong hệ sinh thái dữ liệu lớn.
- Nó có khả năng tăng cường bảo mật dữ liệu và tạo điều kiện chia sẻ dữ liệu an toàn giữa các tổ chức.

In-memory Computing

- Lưu trữ dữ liệu trong RAM thay vì hệ thống lưu trữ truyền thống có thể cải thiện đáng kể tốc độ xử lý cho các ứng dụng và phân tích thời gian thực yêu cầu truy cập dữ liệu nhanh.
- Công nghệ này có thể đặc biệt có lợi cho các dịch vụ tài chính, phát hiện gian lận và các tình huống ra quyết định theo thời gian thực.

Xu hướng tương lai

Quantum Computing

- Mặc dù vẫn còn ở giai đoạn đầu, điện toán lượng tử có tiềm năng to lớn trong việc xử lý các vấn đề dữ liệu lớn phức tạp mà điện toán cổ điển hiện không thể giải quyết được.
- Khả năng xử lý các tập dữ liệu khổng lồ và thực hiện các phép tính phức tạp có thể cách mạng hóa nhiều lĩnh vực khác nhau, từ khoa học vật liệu và khám phá thuốc đến mô hình tài chính và phân tích rủi ro.

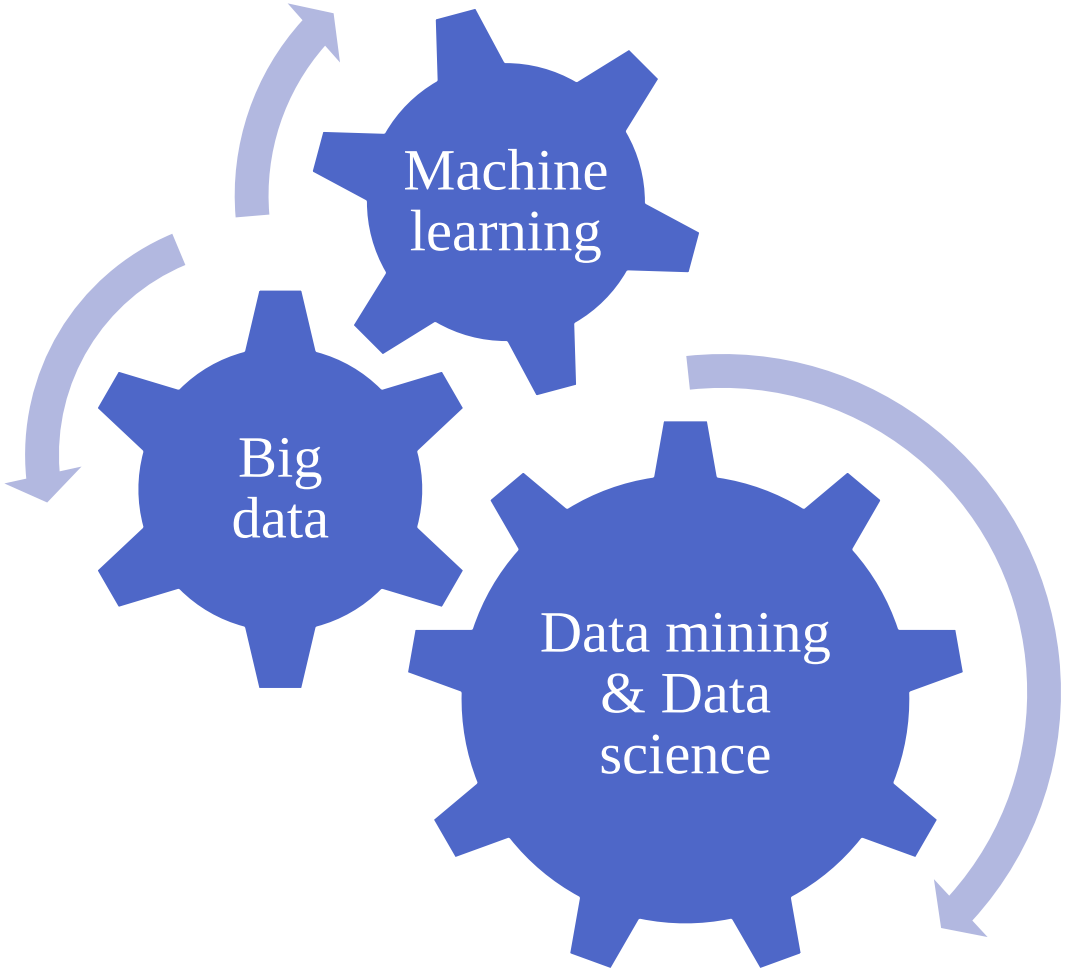
Serverless Computing

- Mô hình dựa trên đám mây này loại bỏ nhu cầu quản lý cơ sở hạ tầng máy chủ, cho phép các tổ chức tập trung vào các ứng dụng và phân tích dữ liệu mà không phải lo lắng về việc cung cấp và mở rộng quy mô máy chủ.
- Nó cung cấp giải pháp tiết kiệm chi phí và có thể mở rộng để xử lý dữ liệu lớn, đặc biệt đối với khối lượng công việc không liên tục hoặc không thể đoán trước.

Hệ thống ứng dụng

Lĩnh vực khai phá dữ liệu	Phân loại (Classification)
	Phân cụm (Clustering)
	Luật kết hợp (Association Rule)
	Hồi quy (Regression)
	Chuỗi thời gian (Time series)

4. Tổng kết



Question & Answer