ĐẠI HỌC KINH DOANH VÀ CÔNG NGHỆ HÀ NỘI

KHOA CNTT

# Big data & Clound

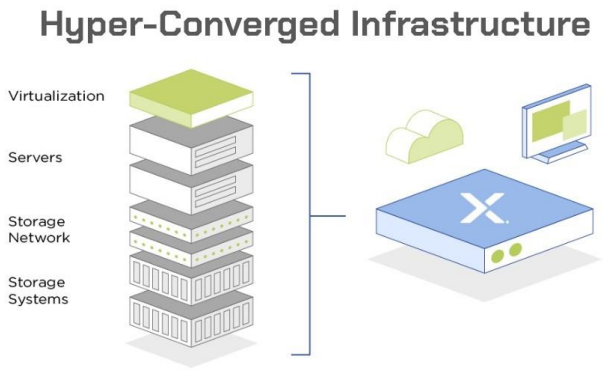
# CHƯƠNG 4: Hệ thống Dữ liệu lớn và Điện toán đám mây

- **Hạ tầng lưu trữ Đám mây**
- **Lợi ích vượt trội của Cloud:**
  - Triển khai hạ tầng nhanh chóng, linh hoạt
  - Tốc độ xử lý - Độ chính xác
  - Phân tích theo thời gian thực
  - Tối ưu chi phí duy trì hoạt động
  - Lợi thế cạnh tranh đối với Doanh nghiệp
- **Một số nền tảng Điện toán đám mây hỗ trợ Dữ liệu lớn:**
  - Google
  - Microsoft
  - Amazon
  - Cloudera

# Hạ tầng cho Bigdata

**Đáp ứng yêu cầu cao về lưu trữ và xử lý dữ liệu lớn có 2 mô hình hiện đại phù hợp nhất cho hệ thống Big Data:**

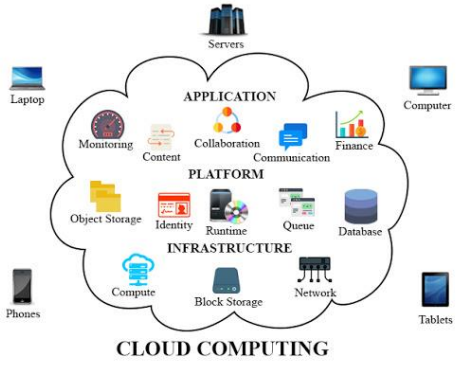
- **Hệ thống Siêu hội tụ (HCI),** kết hợp các máy chủ và hệ thống lưu trữ trong các node để dựng thành các **cluster** có khả năng mở rộng quy mô (scale-out).
- **Hạ tầng Điện toán đám mây (Cloud computing),** đang nhanh chóng được sử dụng rộng rãi, với cơ chế vận hành hybrid và multi-cloud ngày càng đưa ra các lựa chọn thiết thực hơn bao giờ hết.




# Big Data và Cloud Computing

- **Apache Hadoop -** công nghệ Big Data phổ biến nhất hiện nay, được xây dựng trên cơ sở nghiên cứu của Google (MapReduce 2003) và triển khai lần đầu tại Yahoo năm 2007 (với 1000 node cluster).
- **Google đưa ra thuật toán** để hỗ trợ trong việc biên soạn lập chỉ mục các trang Web – rất khó với hệ thống sẵn có lúc bấy giờ.
- Do yêu cầu cao về năng lực lưu trữ, xử lý nên **tiềm năng khổng lồ của Dữ liệu lớn** sẽ không được khai thác triệt để với các máy tính đơn thuần.
- **Cloud Computing** cho phép sử dụng cơ sở hạ tầng tiên tiến nhất, cũng như tiết kiệm, tối ưu hóa nguồn lực và kinh phí dành cho tác vụ phân tích dữ liệu.
- **Dữ liệu lớn và Điện toán đám mây: sự kết hợp hoàn hảo.**

Nếu không có dữ liệu lớn, chắc chắn số lượng ứng dụng điện toán đám mây sẽ ít hơn rất nhiều so với hiện tại → các ứng dụng này cũng được coi là nguồn quan trọng cho sự bùng nổ của dữ liệu lớn.

# Điện toán đám mây là gì?

"**Điện toán đám mây là việc phân phối các tài nguyên CNTT theo nhu cầu qua Internet với chính sách thanh toán theo mức sử dụng.** Thay vì mua, sở hữu và bảo trì các trung tâm dữ liệu và máy chủ vật lý, bạn có thể tiếp cận các dịch vụ công nghệ, như năng lượng điện toán, lưu trữ và cơ sở dữ liệu, khi cần thiết, từ nhà cung cấp dịch vụ đám mây" - **Amazone**

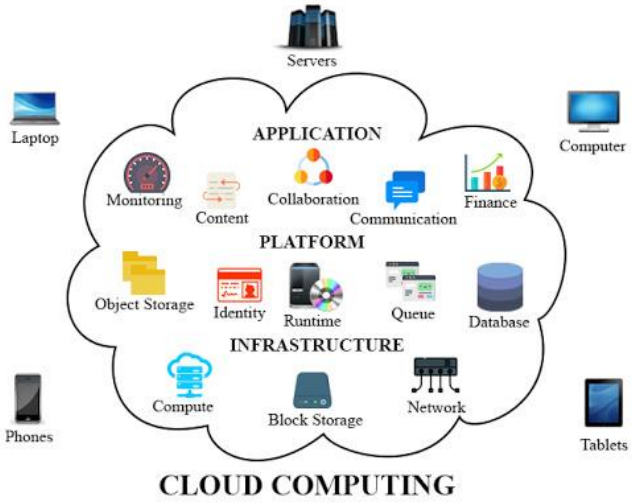

- **Cơ sở hạ tầng dưới dạng dịch vụ (IaaS)**
IaaS chứa các khối xây dựng cơ bản cho đám mây CNTT. IaaS thường cung cấp quyền truy cập vào các tính năng mạng, máy tính (ảo hoặc trên phần cứng chuyên dụng) và không gian lưu trữ dữ liệu.
- **Nền tảng dưới dạng dịch vụ (PaaS)**
PaaS giúp không cần quản lý cơ sở hạ tầng của tổ chức (thường là phần cứng và hệ điều hành) và cho phép tập trung vào công tác triển khai cũng như quản lý các ứng dụng
- **Phần mềm dưới dạng dịch vụ (SaaS)**
SaaS cung cấp sản phẩm hoàn chỉnh được nhà cung cấp dịch vụ vận hành và quản lý.

# Lợi ích vượt trội của Cloud:

**Điện toán đám mây đóng vai trò quan trọng trong thế giới Big Data, bằng cách cung cấp cơ sở hạ tầng được tối ưu hóa và mở rộng.**

**Điều đó hỗ trợ trong việc thực tế hóa Big Data.**

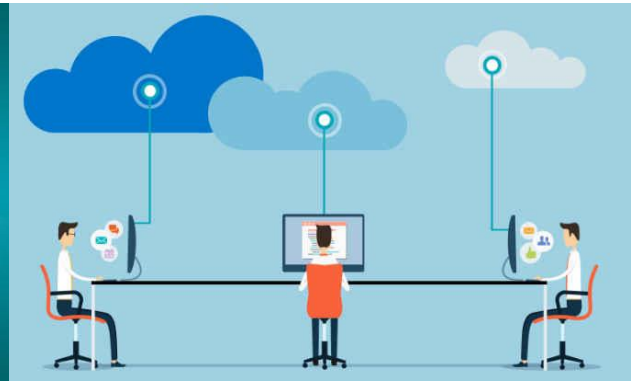
- ✓ **Triển khai hạ tầng nhanh chóng, linh hoạt**
- ✓ **Tốc độ xử lý - Độ chính xác**
- ✓ **Phân tích theo thời gian thực**
- ✓ **Tối ưu chi phí duy trì hoạt động**
- ✓ **Lợi thế cạnh tranh đối với Doanh nghiệp**




→ Trong các mô hình điện toán từ trước tới nay Điện toán đám mây sở hữu nhiều ưu điểm vượt trội hơn cả và chứa ít rủi ro, đặc biệt phù hợp ứng dụng cho các hệ thống hiện đại - yêu cầu cao như Bigdata

# Ứng dụng của đám mây trong Big Data

- **IAAS trong đám mây chung:** Sử dụng cơ sở hạ tầng của nhà cung cấp cho các dịch vụ Big Data, cho phép truy cập vô hạn vào kho lưu trữ và tính toán. IAAS có thể được sử dụng bởi các khách hàng doanh nghiệp để tạo ra các giải pháp CNTT hiệu quả hay để mở rộng quy mô doanh nghiệp.
- **PAAS trong đám mây riêng:** Các nhà cung cấp PAAS đang bắt đầu kết hợp các công nghệ Big Data như Hadoop và MapReduce vào các dịch vụ PAAS của họ, giúp loại bỏ sự phức tạp của việc quản lý các yếu tố phần mềm và phần cứng riêng lẻ.
- **SAAS trong đám mây lai:** Nhiều tổ chức cảm thấy cần phải phân tích ý kiến khách hàng, đặc biệt là trên phương tiện mạng xã hội. Các nhà cung cấp SAAS cung cấp nền tảng cho việc phân tích cũng như dữ liệu mạng xã hội.

# AWS, Microsoft Azure và Google Cloud

Figure 1. Magic Quadrant for Cloud Infrastructure and Platform Services


AWS vẫn đang dẫn đầu Thị trường đám mây. Azure và Google Cloud cũng đang phát triển không ngừng.

- **Amazon Web Services**: 33% thị phần
- **Microsoft Azure**: 18%
- **Google Cloud Platform**: 8% thị phần

| Nhà cung cấp | AWS | Azure | GCP |
|---|---|---|---|
| Số lượng dịch vụ | 212 | 100+ | 60+ |

**AWS** chỉ tăng 41% trong năm 2019. Trong khi **Google** Cloud và **Azure** lần lượt tăng 80% và 75%. Điều này cho thấy Azure và Google Cloud đang bắt kịp.

# AWS, Microsoft Azure và Google Cloud

**Dịch vụ lưu trữ**
Lưu trữ là một dịch vụ rất quan trọng khi nói đến Điện toán đám mây vì chỉ sau khi lưu trữ dữ liệu, mới có thể nghĩ đến các dịch vụ khác có thể giúp xử lý dữ liệu

| Services | Amazon Web Services | Microsoft Azure | Google Cloud |
|---|---|---|---|
| **Object Storage** | Amazon S3 | Azure Disk Storage | Google Cloud Storage |
| **Block Store** | Amazon EBS | Azure Blob Storage | Google Compute Engine (Persistent Disks) |
| **Archival/Cold Storage** | Amazon Glacier | Azure Archive Blob Storage | Google Nearline |
| **File System Storage** | Amazon EFS | Azure File Storage | Google ZFS/Avere |

# AWS, Microsoft Azure và Google Cloud

**Dịch vụ tính toán** (compute service), AWS cung cấp hệ thống EC2 rất phổ biến trên thị trường. Nó cũng hỗ trợ nhiều dịch vụ tính toán khác liên quan đến PaaS, container và thậm chí cả dịch vụ serverless. Azure và Google Cloud cũng có nhiều dịch vụ tương đồng với AWS trong các lĩnh vực này.

| Services | Amazon Web Services | Microsoft Azure | Google Cloud |
|---|---|---|---|
| **Infrastructure as a Service** | Amazon EC2 | Virtual Machines (VM) | Google Compute Engine |
| **Platform as a Service** | AWS Elastic Beanstalk | App Service and Cloud Services | Google App Engine |
| **Container Services** | Amazon Elastic Container Service | Azure Kubernetes Service or AKS | Google Kubernetes Engine |
| **Serverless Computing** | Amazon Lambda | Azure Functions | Cloud Functions |

# AWS, Microsoft Azure và Google Cloud

**Dịch vụ cơ sở dữ liệu**

| Services | Amazon Web Services | Microsoft Azure | Google Cloud |
|---|---|---|---|
| Relation DB | Amazon RDS | SQL DB | Google Cloud SQL |
| NoSQL DB: Key-value | Amazon DynamoDB | Table Storage | Google Cloud DatastoreGoogle Cloud Bigtable |
| NoSQL DB: With Indexing | Amazon SimpleDB | Azure Cosmos DB | Google Cloud Datastore |

# Google Cloud Platform


- **Google Cloud Platform** cung cấp giải pháp Big Data cho phép thu thập, xử lý, lưu trữ và phân tích dữ liệu của trong một nền tảng duy nhất.
- **Cloud Computing** cho phép tập chung vào dự án mà không cần lo lắng về quản lý hạ tầng

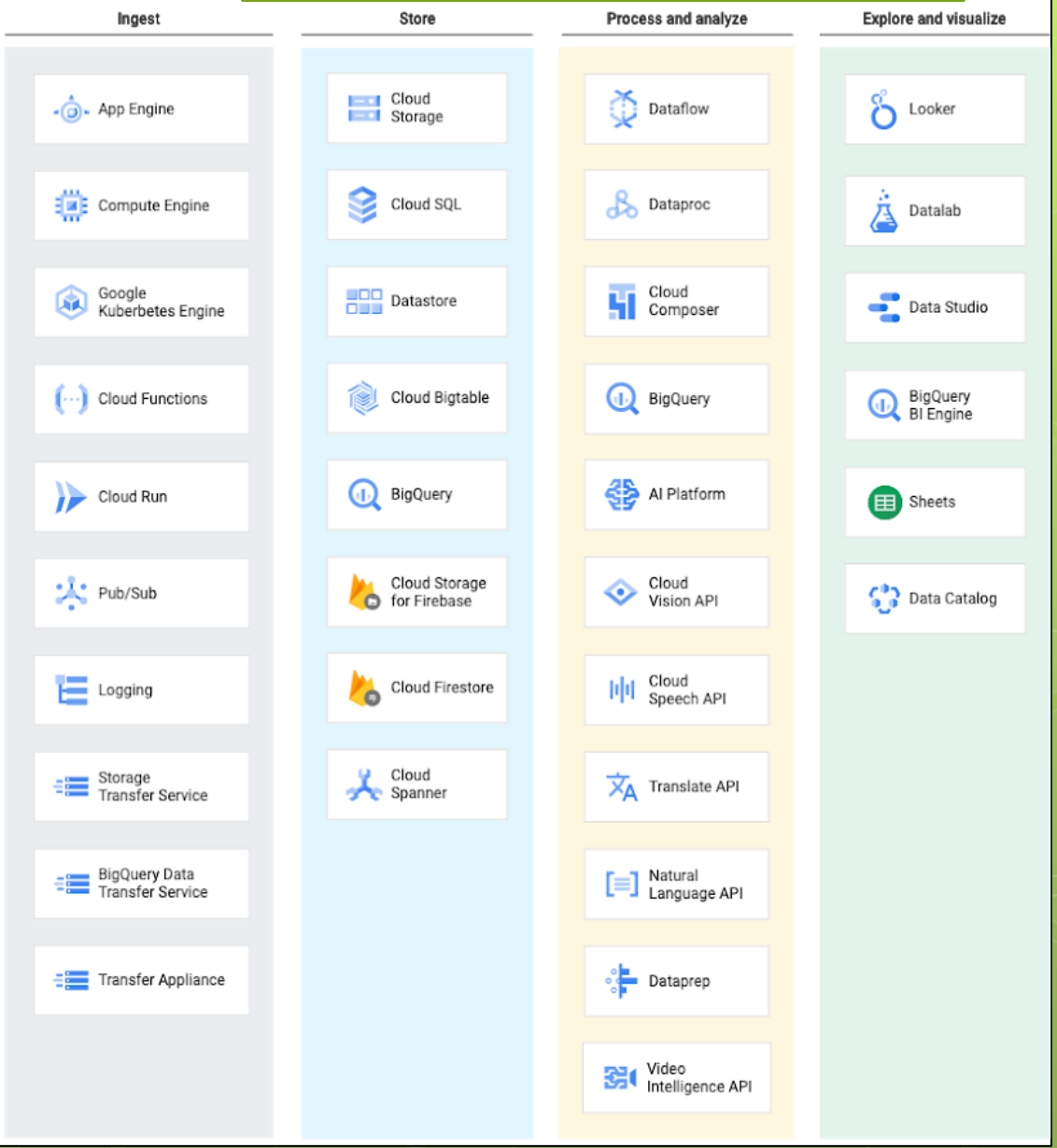
**Google Cloud cung cấp những sản phẩm chính sau đây:**

- **Big Data:** BigQuery, Cloud Dataproc, Cloud Dataflow…
- **Services:** Translate API, Prediction API…
- **Storage:** Cloud Storage, Cloud Datastore, Cloud SQL…
- **Compute:** App Engine, Compute Engine, …

Google Cloud Platform

# Google Cloud Platform

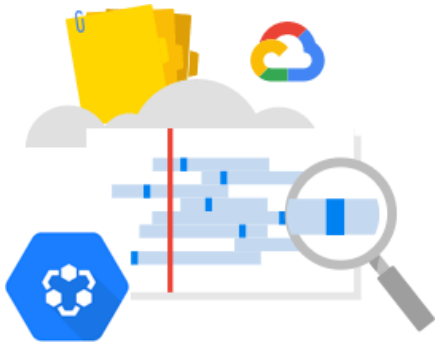
- Google **Cloud Data Fusion**
- Google **Cloud Pub / Sub**
- Google **Cloud Dataprep**
- Google **Cloud Dataproc**
- Google **Cloud Datalab**
- Google **Cloud Dataflow**
- Google **BigQuery**
- Google **Cloud Bigtable**
- Google **Data Catalog**
- Google **Data Studio**



# Google Cloud Platform

**Kết hợp dữ liệu đám mây, quản lý đầy đủ**

Google Cloud Data Fusion là là một dịch vụ tích hợp dữ liệu dựa trên đám mây được quản lý hoàn toàn, giúp người dùng xây dựng và quản lý hiệu quả các đường ống dữ liệu ETL / ELT. Với giao diện đồ họa và thư viện mã nguồn mở rộng gồm các trình kết nối và biến đổi được cấu hình sẵn.



**Dịch vụ nhắn tin và nhắn tin mở rộng**

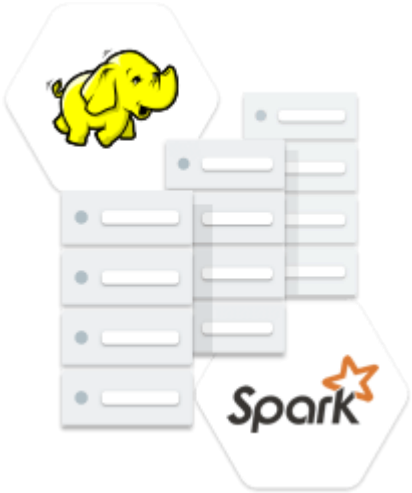
Google Cloud Pub / Sub là một nền tảng đơn giản, đáng tin cậy, có thể mở rộng để phân tích luồng và hệ thống máy tính hướng sự kiện. Bạn có thể gửi và nhận tin nhắn giữa các ứng dụng độc lập và dữ liệu cung cấp thông qua ứng dụng chạy trên nền tảng đám mây.

# Google Cloud Platform

**Quản lý Apache Spark và Apache Hadoop**

Cloud Dataproc là một dịch vụ đám mây để chạy các cụm Apache Spark và Apache Hadoop theo cách đơn giản hơn, tiết kiệm chi phí hơn.

Cloud Dataproc tích hợp với các dịch vụ lưu trữ, tính toán và giám sát trên các sản phẩm của Google Cloud, một nền tảng xử lý dữ liệu mạnh mẽ và đầy đủ.




**Chuẩn bị dữ liệu thông minh**

Google Cloud Dataprep là dịch vụ dữ liệu thông minh giúp khám phá, làm sạch và chuẩn bị dữ liệu có cấu trúc và không có cấu trúc để phân tích. Cloud Dataprep không hoạt động và hoạt động ở bất kỳ quy mô nào. Không có cơ sở hạ tầng để triển khai hoặc quản lý. Chuẩn bị dữ liệu dễ dàng với các nhấp chuột và không có viết mã

# Google Cloud Platform

**Tối ưu năng lực khai phá**

Google Cloud Datalab (dựa trên Jupyter) để khám phá, cộng tác, phân tích và hình dung dữ liệu. Nó được tích hợp với BigQuery và Google Cloud Machine Learning giúp bạn dễ dàng truy cập vào các dịch vụ xử lý dữ liệu quan trọng

**Xử lý dữ liệu hàng loạt và luồng**

Google Cloud Dataflow cung cấp mô hình lập trình thống nhất và dịch vụ được quản lý để thực hiện nhiều mẫu xử lý dữ liệu bao gồm phân tích luồng, ETL và tính toán hàng loạt.

Cloud Dataflow giúp người dùng không cần lo nghĩ như lập kế hoạch năng suất, quản lý tài nguyên và tối ưu hóa hiệu suất.

# Google Cloud Platform

**Analytics Data Warehouse**

Google BigQuery là kho dữ liệu không có máy chủ, chi phí thấp, được quản lý hoàn toàn của Google, phù hợp với nhu cầu năng lượng lưu trữ và tính toán của bạn.

BigQuery có thể phân tích terabyte đến petabyte dữ liệu với tốc độ nhanh.

BigQuery là nền tảng Phân tích dữ liệu mạnh mẽ được sử dụng phổ biến cho các tổ chức từ khởi nghiệp đến các công ty lớn

**Quản lý cơ sở dữ liệu NoQuery**

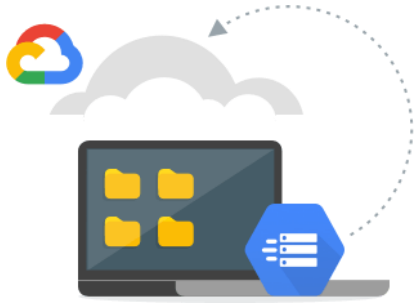
Cloud Bigtable cung cấp cơ sở dữ liệu NoQuery có khả năng mở rộng phù hợp với khối lượng công việc có độ trễ thấp và thông lượng cao.

Google Cloud Bigtable tích hợp dễ dàng với các công cụ dữ liệu lớn phổ biến như Hadoop và Spark, và nó hỗ trợ API HBase tiêu chuẩn, nguồn mở.

# Google Cloud Platform

**Quản lý đầy đủ danh mục**

Danh mục dữ liệu – Google Data Catalog là một dịch vụ quản lý siêu dữ liệu được quản lý đầy đủ và có thể mở rộng, cho phép các tổ chức nhanh chóng khám phá, quản lý và hiểu tất cả dữ liệu của họ trong Google Cloud. Nó cung cấp giao diện tìm kiếm đơn giản và dễ sử dụng để khám phá dữ liệu, hệ thống danh mục linh hoạt và mạnh mẽ để nắm bắt cả siêu dữ liệu kỹ thuật và kinh doanh.



**Dữ liệu tốt cho quyết định phù hợp**

Google Data Studio chuyển dữ liệu thành các trang tổng quan và báo cáo dễ đọc, biểu đồ, chia sẻ và tùy chỉnh được. Dễ dàng truy cập vào tất cả các nguồn dữ liệu trên Google Cloud bạn cần để hiểu về doanh nghiệp của bạn và đưa ra quyết định tốt hơn.

# Nền tảng hỗ trợ Dữ liệu lớn

| | Google | Microsoft | Amazon | Cloudera |
|---|---|---|---|---|
| *Big data storage* | Google cloud services | Azure | S3 | |
| *MapReduce* | AppEngine | Hadoop on Azure | Elastic MapReduce (Hadoop) | MapReduce YARN |
| *Big data analytics* | BigQuery | Hadoop on Azure | Elastic MapReduce (Hadoop) | Elastic MapReduce (Hadoop) |
| *Relational database* | Cloud SQL | SQL Azure | MySQL or Oracle | MySQL, Oracle, PostgreSQL |
| *NoSQL database* | AppEngine Datastore | Table storage | DynamoDB | Apache Accumulo |
| *Streaming processing* | Search API | Streaminsight | Nothing prepackaged | Apache Spark |
| *Data import* | Network | Network | Network | Network |
| *Data sources* | A few sample datasets | Windows Azure marketplace | Public Datasets | Public Datasets |

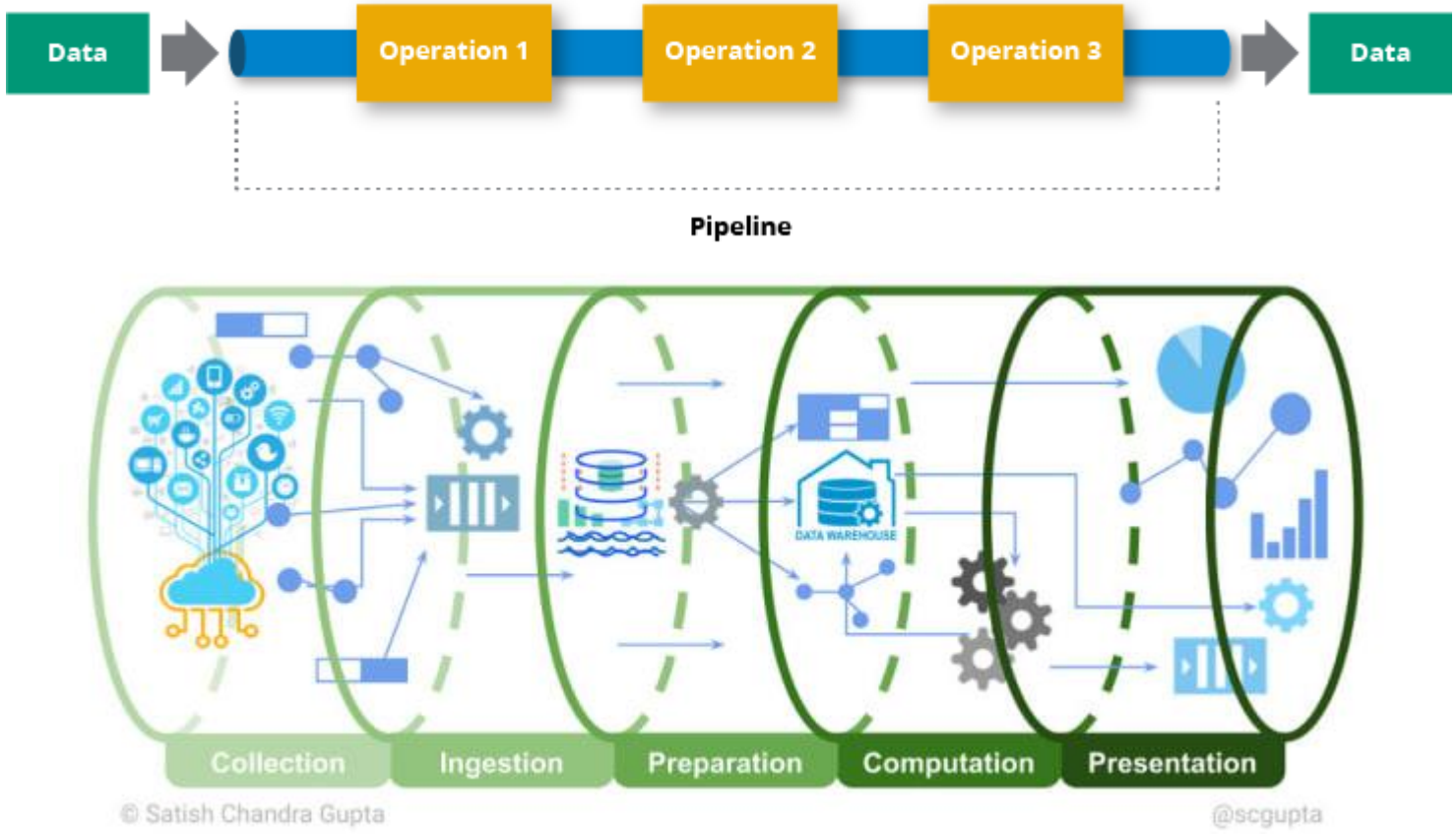
***Software Architecture for Big Data and the Cloud - 1st Edition - June 12, 2017***
*https://www.elsevier.com/books/software-architecture-for-big-data-and-the-cloud/mistrik/978-0-12-805467-3*

# Big Data – Cloudera Data Platform

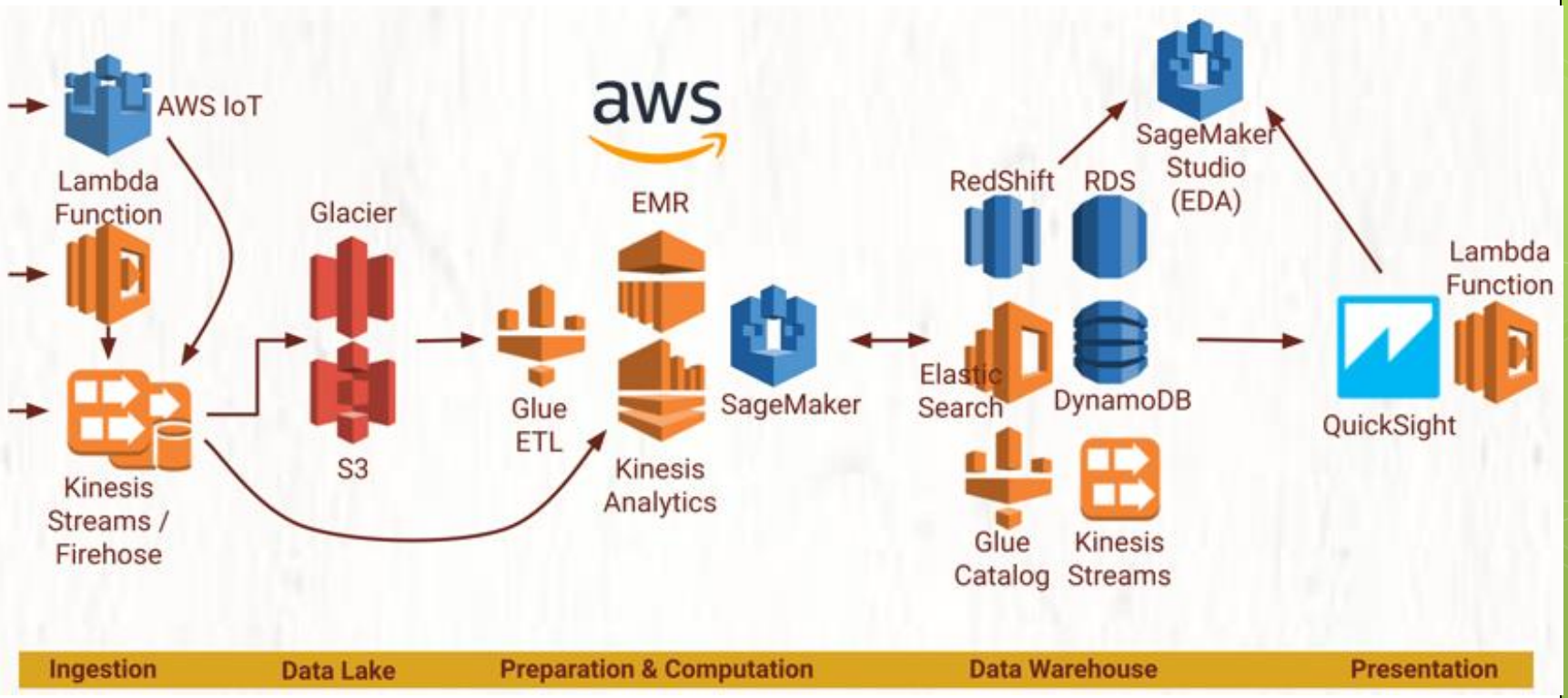
**Cloudera hiện là một trong những Công cụ Dữ liệu lớn nhanh nhất và an toàn nhất hiện có.** Khởi đầu phân phối Apache Hadoop mã nguồn mở dành cho việc triển khai cấp doanh nghiệp. Nền tảng linh hoạt này giúp việc thu thập dữ liệu từ bất kỳ hệ thống nào trở nên đơn giản.

- Cung cấp thông tin chi tiết và giám sát dữ liệu theo thời gian thực.
- Cloudera Enterprise có thể được triển khai trên AWS, Google Cloud và Microsoft Azure, trong số các nền tảng Đám mây khác.
- Cung cấp tùy chọn Đám mây kết hợp cho doanh nghiệp.

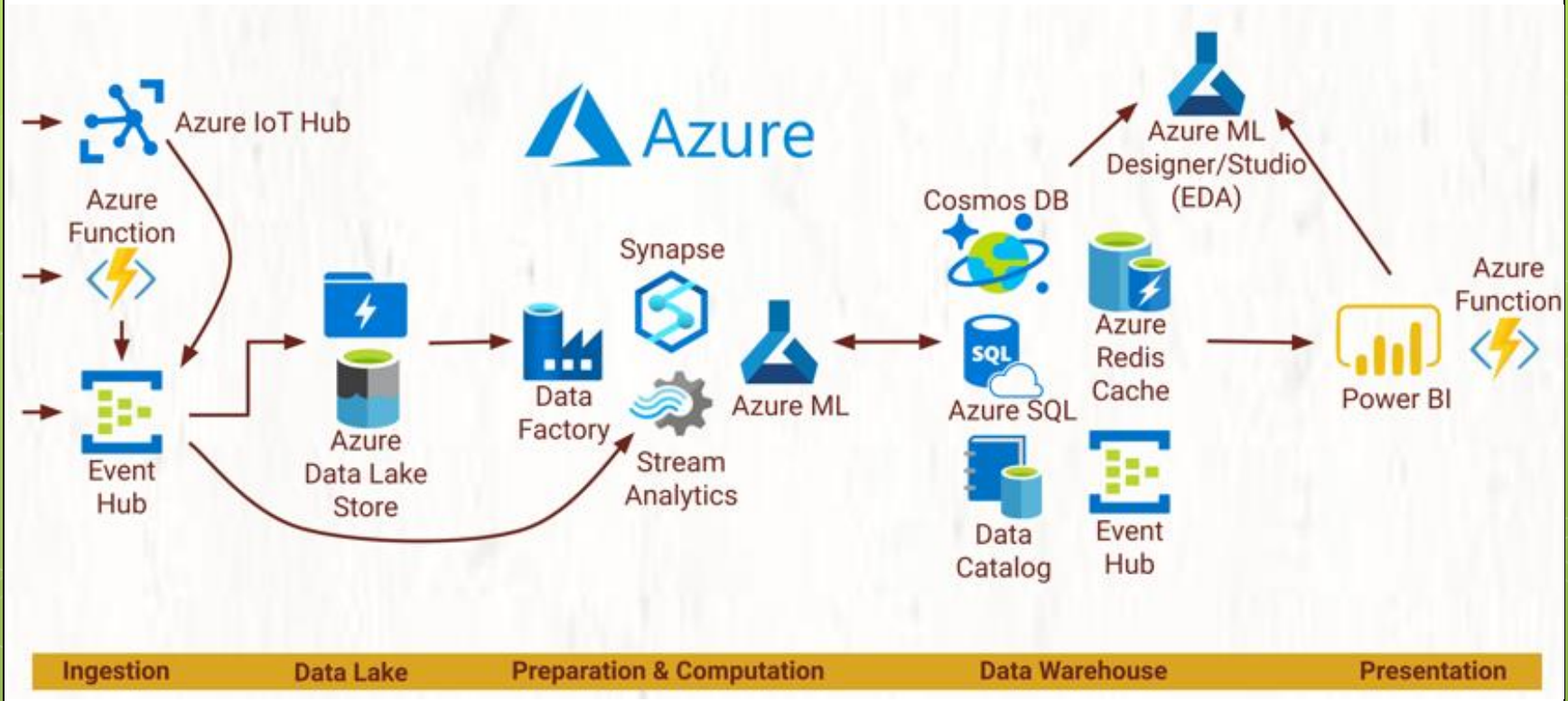
# Big Data pipeline end-to-end



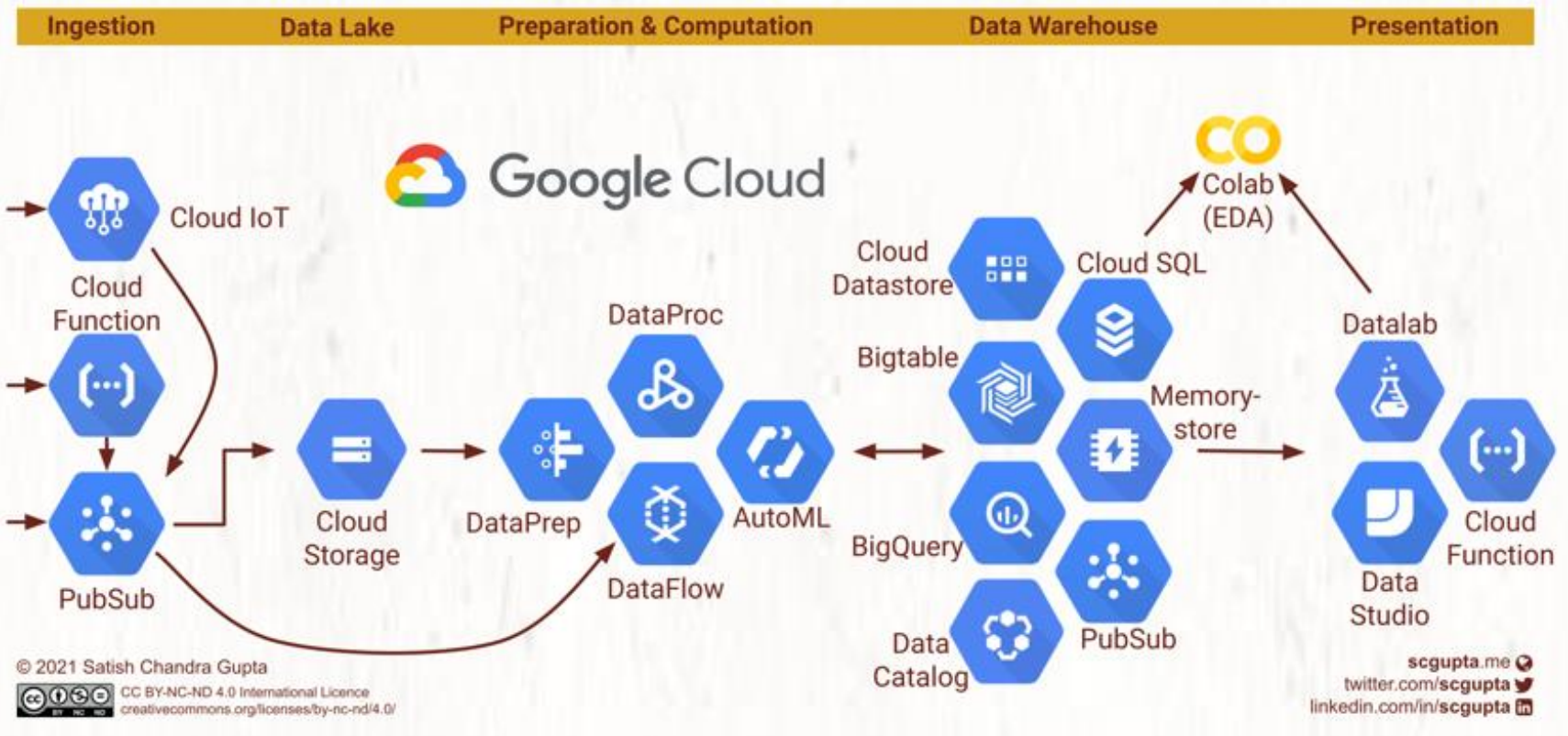
# Big Data Pipelines - Amazon



# Big Data Pipelines - Microsoft



# Big Data Pipelines - Google




Cảm ơn các bạn đã
lắng nghe!