WHY Kafka:

Kafka is often used in real-time streaming data architectures to provide real-time analytics

Kafka is FAST, SCALABLE, DURABLE

Kafka support massive message streams for low-latency follow-up analysis

Kafka Use cases

* Kafka is used for stream processing, website activity tracking, metrics collection and monitoring, log aggregation, real-time analytics,
* Uses Kafka to track activity data and operational metrics

Structure:

* Sử dụng mô hình truyền thông: public – subscribe.
* Public (producer)
* Subscribe (consumer) ---- nhận dữ liệu theo topic.
* -----------------------------------------------------------------------------------------
* Topic: (chủ đề) các dữ liệu liên quan/tương tự được nhóm trong các chủ đề. Mỗi chủ đề là một nguồn dữ liệu riêng biệt. => trong Kafka dữ liệu được truyền theo topic. => truyền dữ liệu khác nhau bằng cách tạo ra các topic.
* Partition: Topic được chia thành nhiều partition – (nơi lưu trữ dữ liệu cho 1 topic). Partition có id (offset)

Hadoop (framework): Phân tích dữ liệu

Zookeeper: (xử lý các máy con theo cơ chế master – node / cluster)

Map-reduce (xử lý, phân tích dữ liệu – cơ chế.).

Kafka: Hệ thống hang đợi dữ liệu (producer-subscribe). – hỗ trợ mô hình phân tán (distribute) – phân chia (partition), đồng bộ (replicate) -> Hệ thống logging, lưu lại trạng thái hệ thống phòng tránh mất thông tin.(track dữ liệu)

=== thường so sánh kafka – spark.

Spark (Streaming data). Phân tích dữ liệu. đốc độ xử lý nhanh (in-memory processing)

