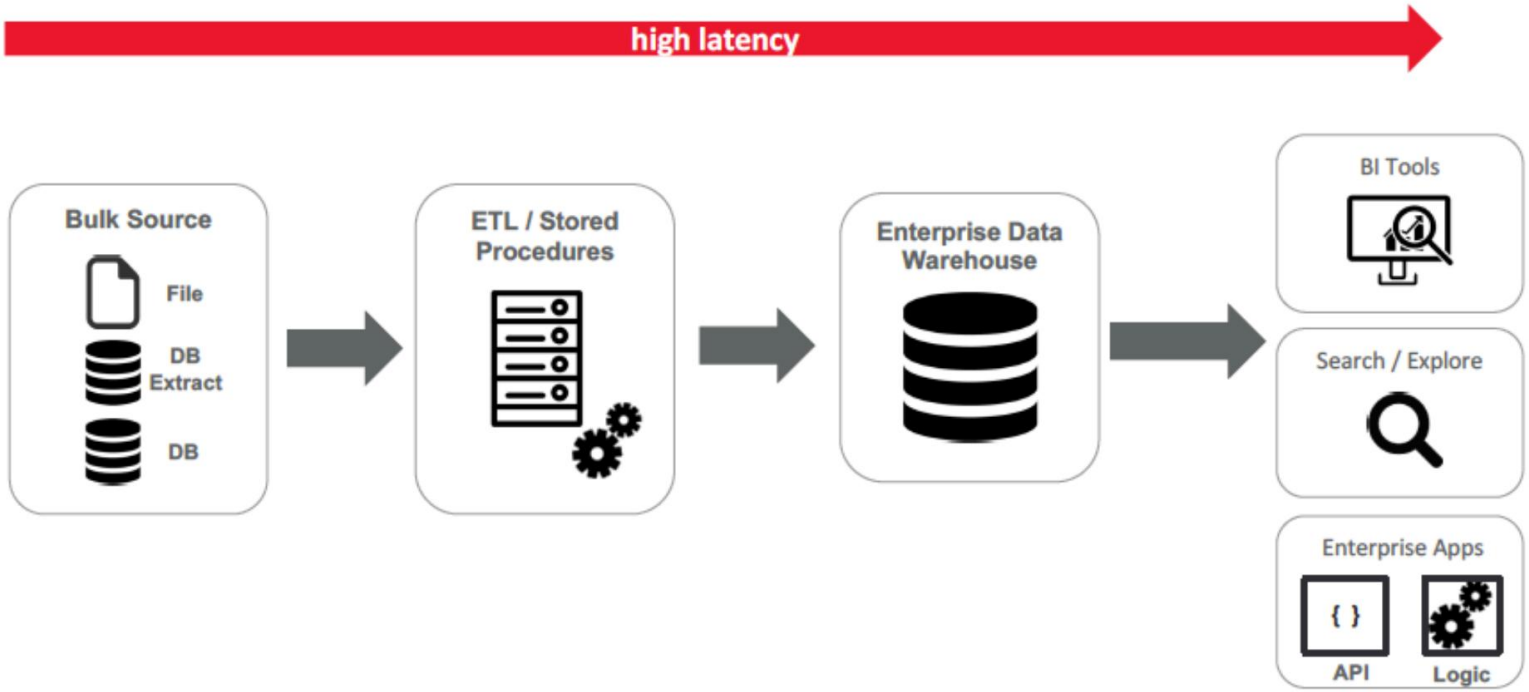
HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

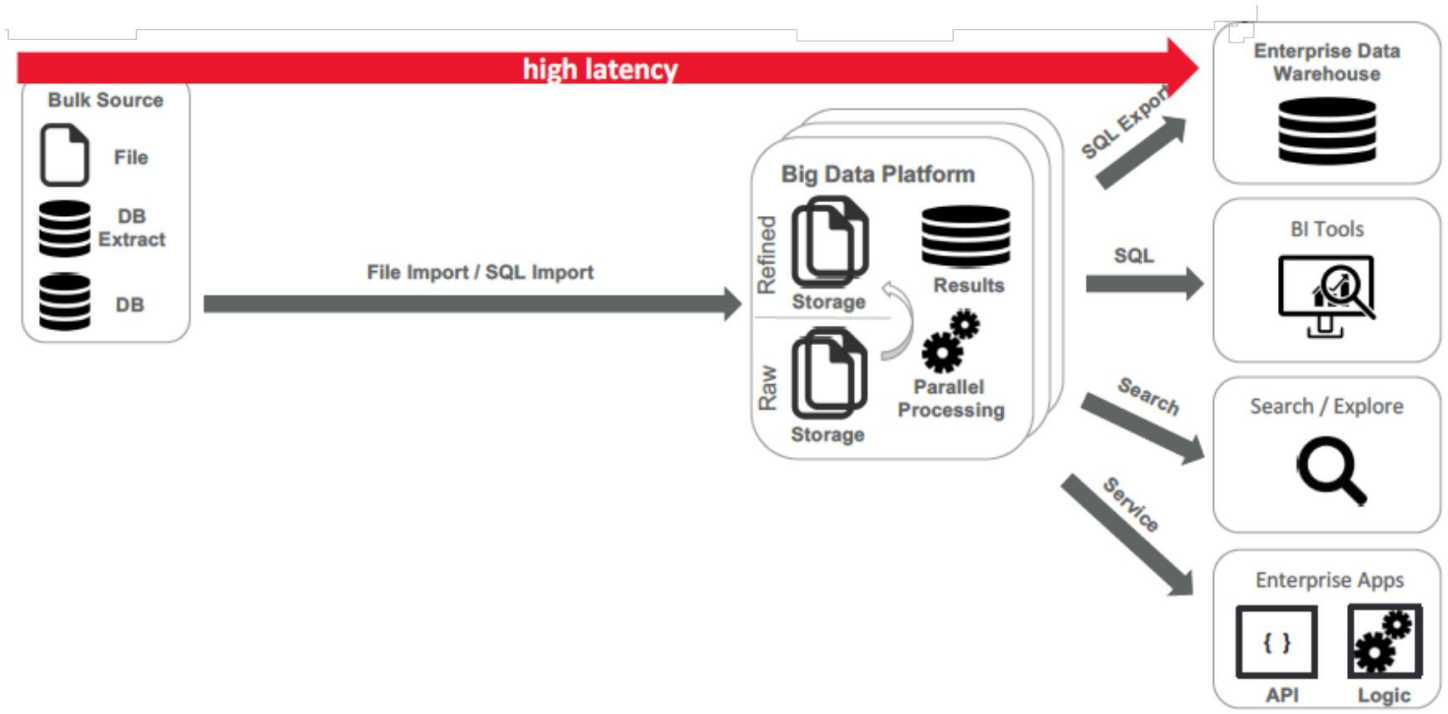
Chương 8

Kiến trúc dữ liệu lớn

Cơ sở hạ tầng BI truyền thống



Hadoop giải quyết vấn đề về Khối lượng và Sự đa dạng – không phải Tốc độ

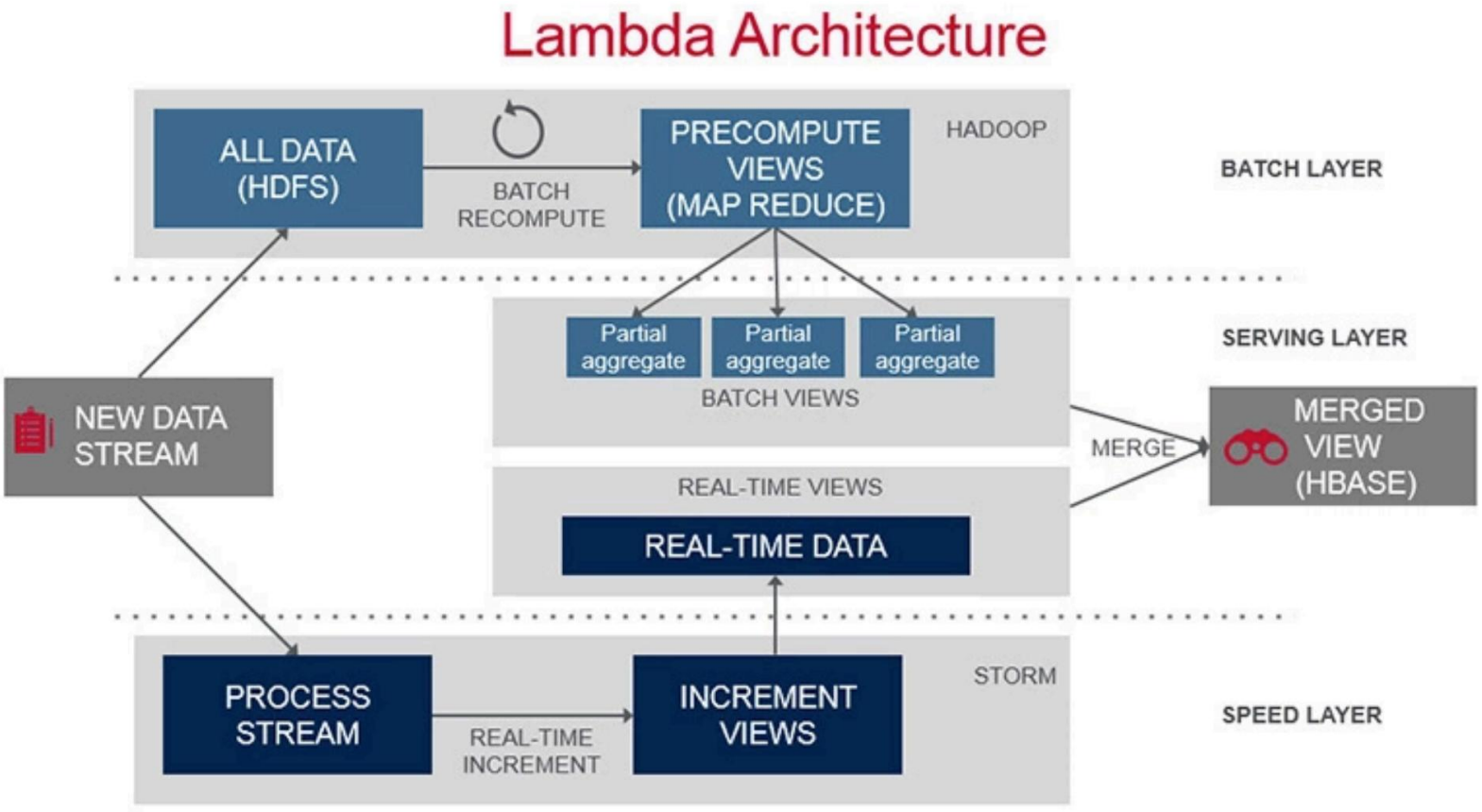


Kiến trúc Lambda

- Một kiến trúc xử lý dữ liệu được thiết kế để xử lý khối lượng dữ liệu lớn bằng cách tận dụng cả phương pháp xử lý theo lô và theo luồng.
- Spark là một trong số ít các khuôn khổ xử lý dữ liệu cho phép bạn tích hợp liền mạch xử lý hàng loạt và luồng
  - Của petabyte dữ liệu
  - Trong cùng một ứng dụng

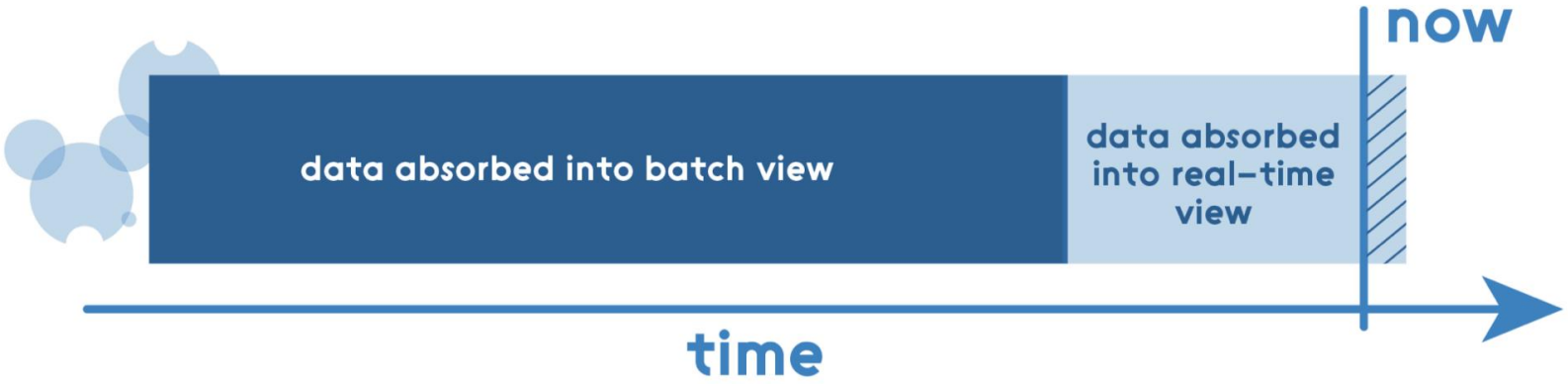
I need fast access to historical data on the fly for predictive modeling with real time data from the stream

Kiến trúc Lambda

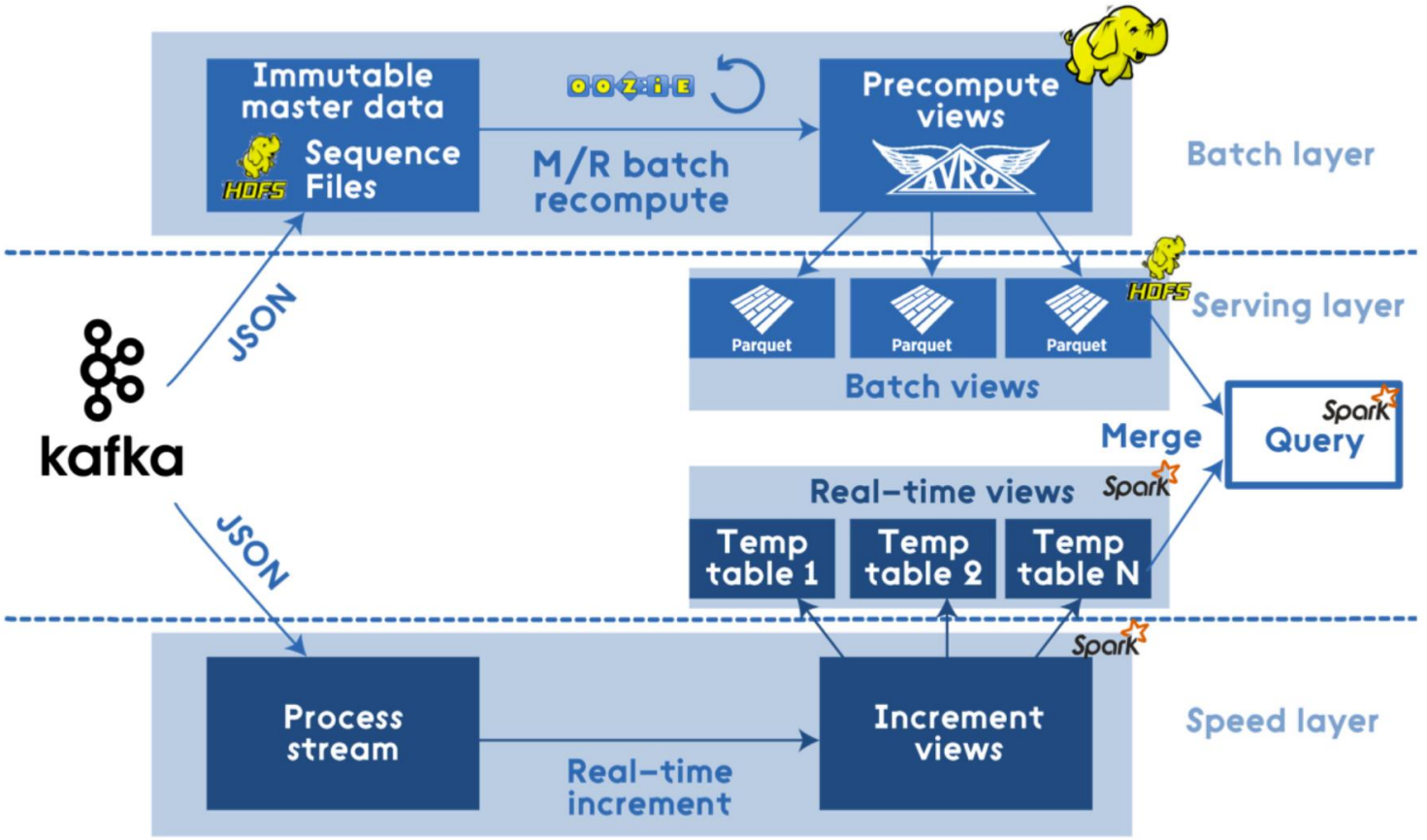


Sự liên quan của dữ liệu

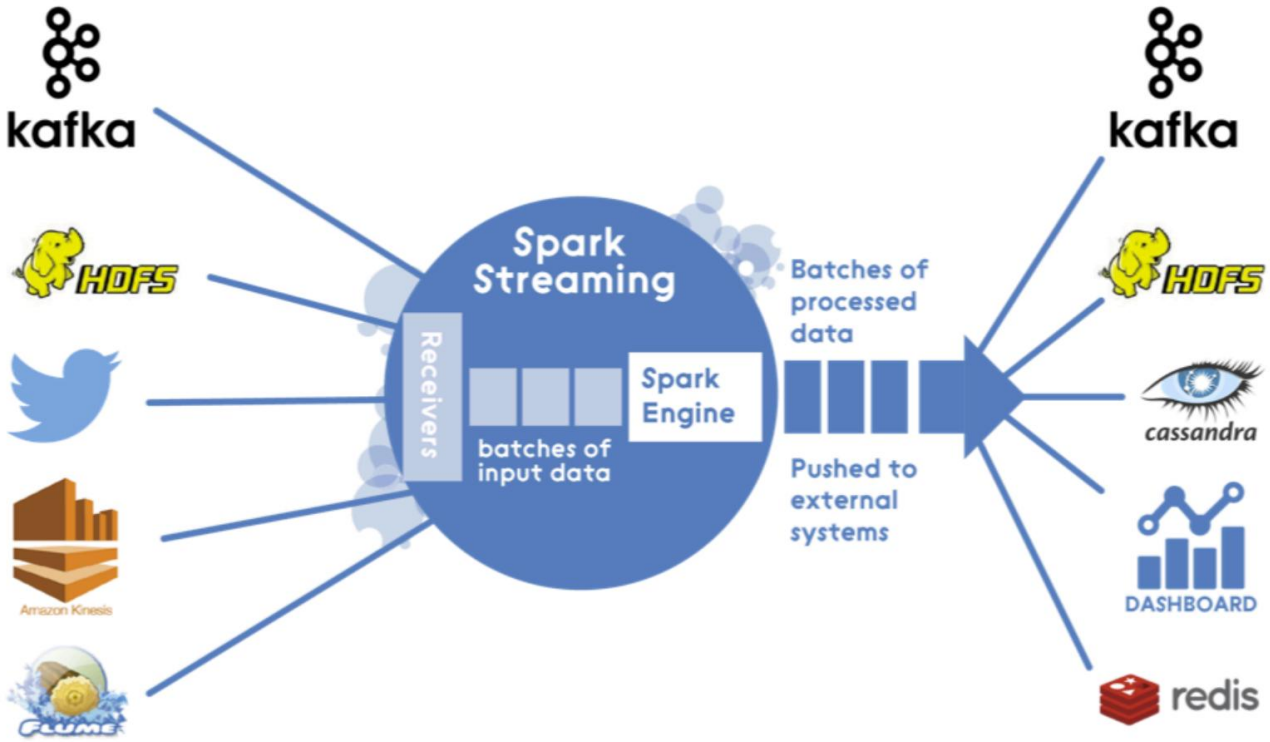
query = function(batch view, real time view)
real time view = function(real time view, new data)
batch view = function(all data)



Kiến trúc Lambda: một triển khai



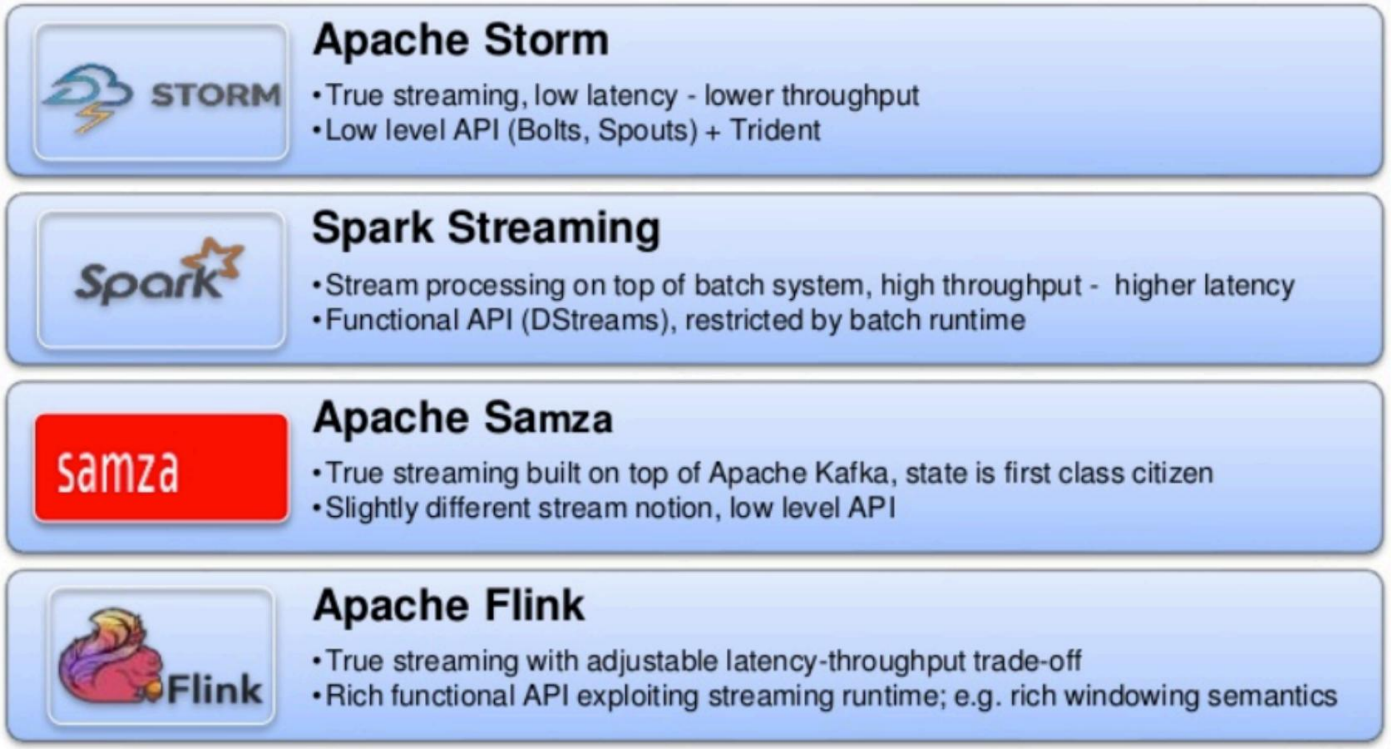
Spark phát trực tuyến



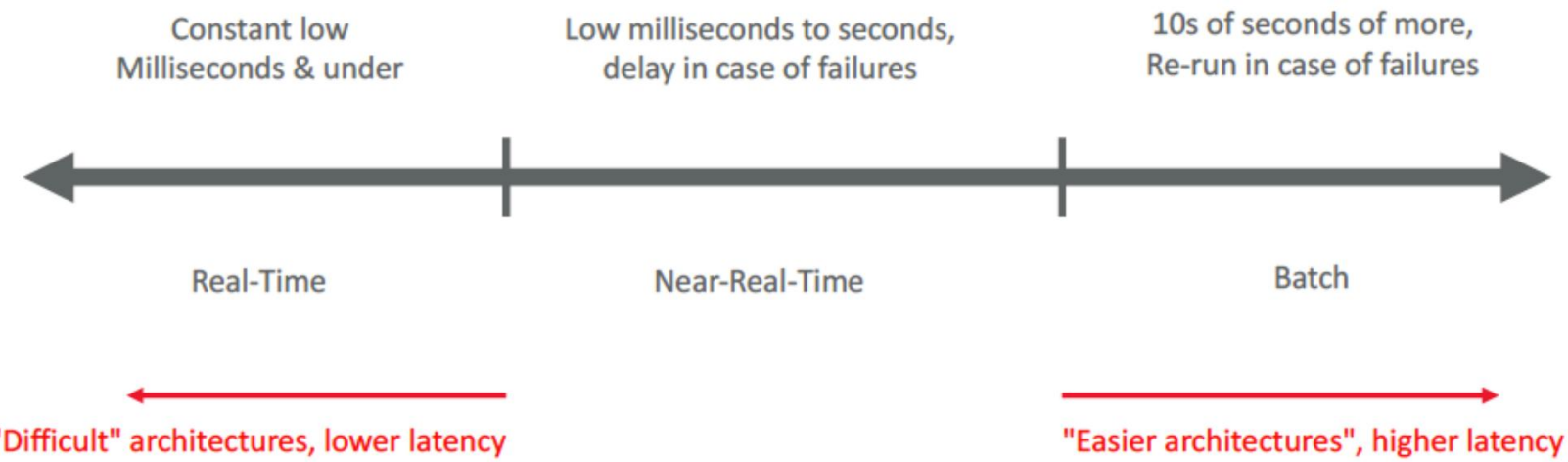
Spark phát trực tuyến

- Hệ thống xử lý luồng có khả năng mở rộng, chịu lỗi
- tính toán luồng như: một loạt các tác vụ hàng loạt rất nhỏ, xác định
  - Chia luồng trực tiếp thành các đợt X giây
  - Spark xử lý từng lô dữ liệu như RDD và xử lý chúng bằng cách sử dụng Hoạt động RDD
  - Cuối cùng, kết quả đã xử lý của các hoạt động RDD được trả về trong lô hàng

Phong cảnh phát trực tuyến



Xử lý luồng so với xử lý hàng loạt



Tài liệu tham khảo

- https://github.com/OryxProject/oryx
- https://github.com/MicrosoftDocs/azure-docs/blob/master/articles/cosmos-db/lambda-architecture.md
- https://github.com/apssouza22/lambda-arch
- https://github.com/knoldus/Lambda-Arch-Spark

ĐẠI HỌC BÁCH KHOA

25 YEARS ANNIVERSARY SOICT

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Cảm ơn sự chú ý của bạn!!!