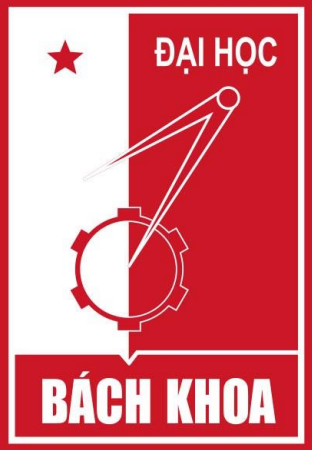


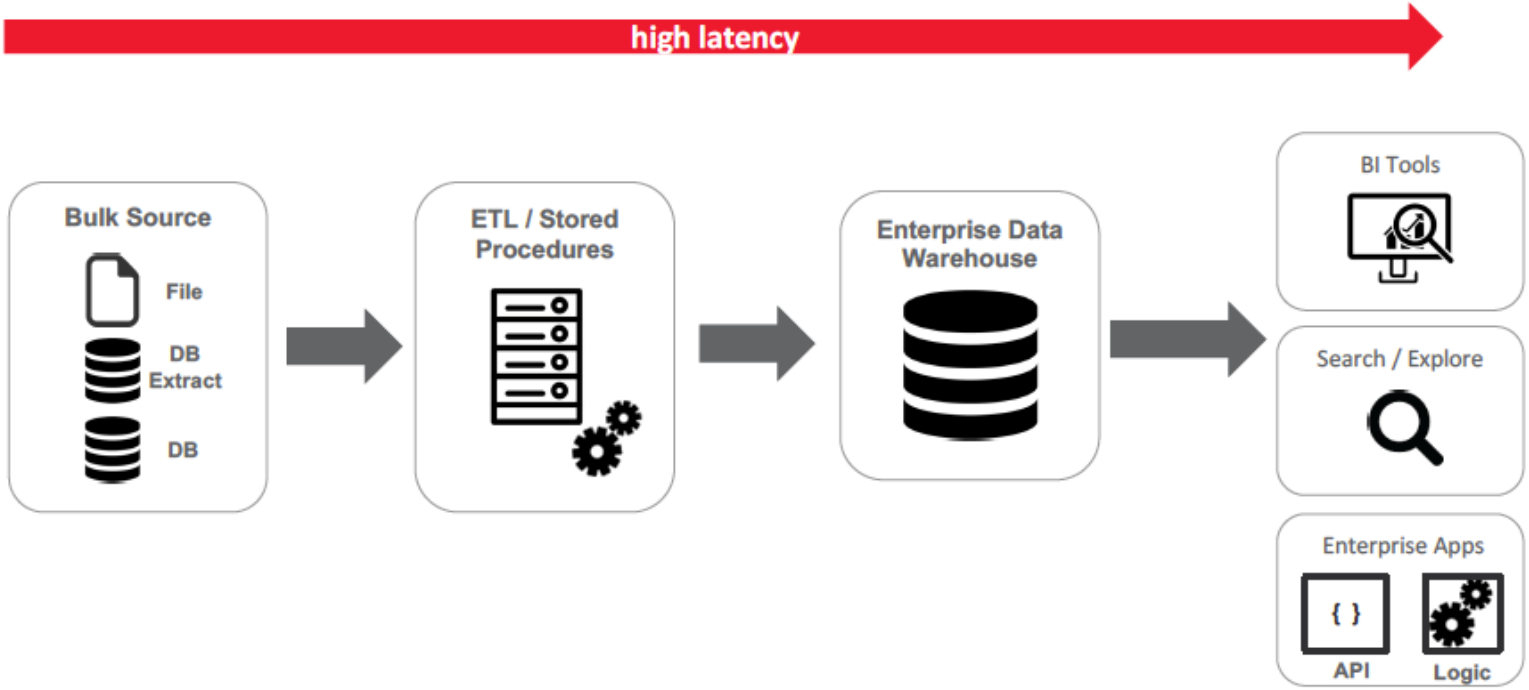
HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

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

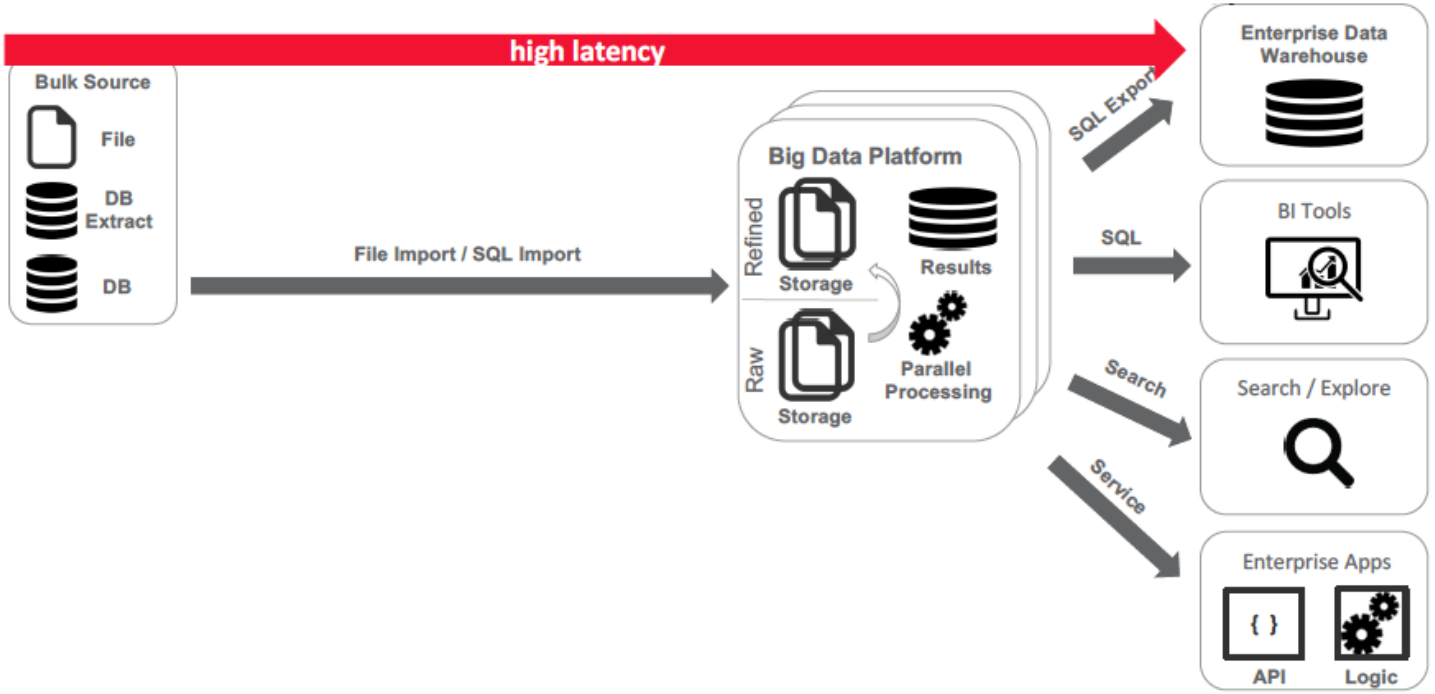
Chapter 8

Big data architecture

Traditional BI infrastructures



Hadoop solves Volume and Variety – not Velocity

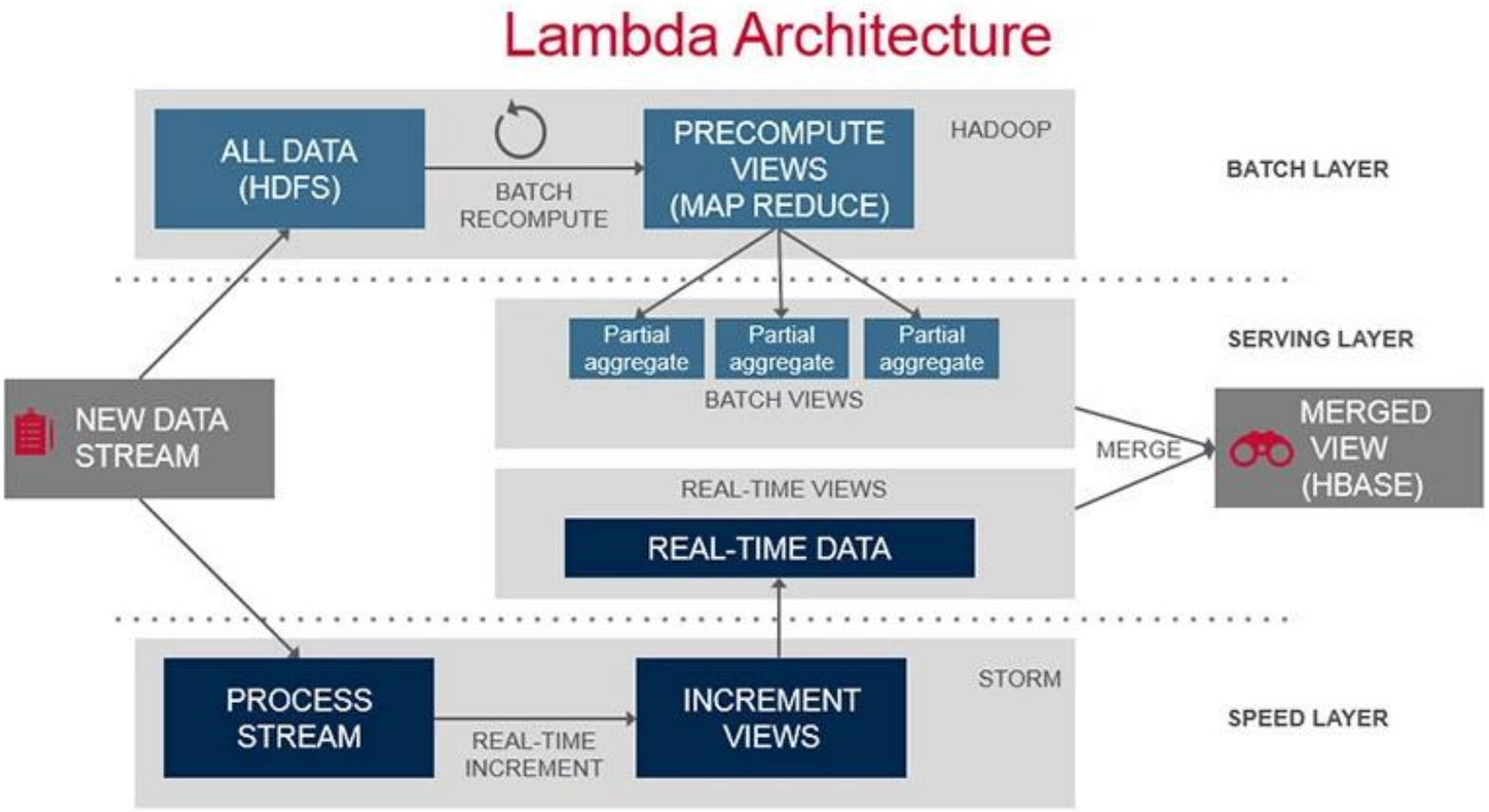


Lambda Architecture

- A data-processing architecture designed to handle massive quantities of data by taking advantage of both batch and stream processing methods.
- Spark is one of the few data processing frameworks that allows you to seamlessly integrate batch and stream processing
  - Of petabytes of data
  - In the same application

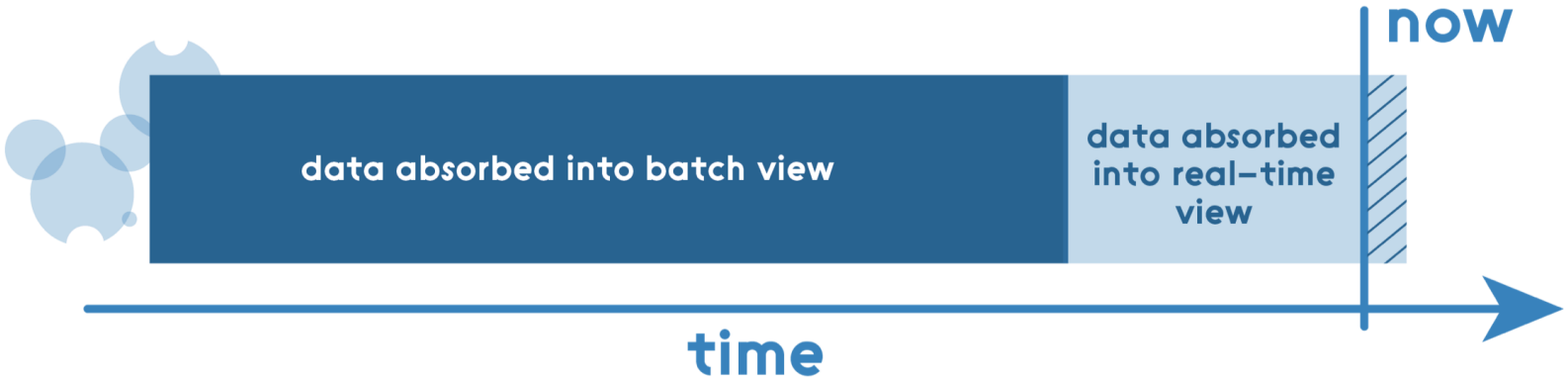
I need fast access to historical data on the fly for predictive modeling with real time data from the stream

Lambda architecture

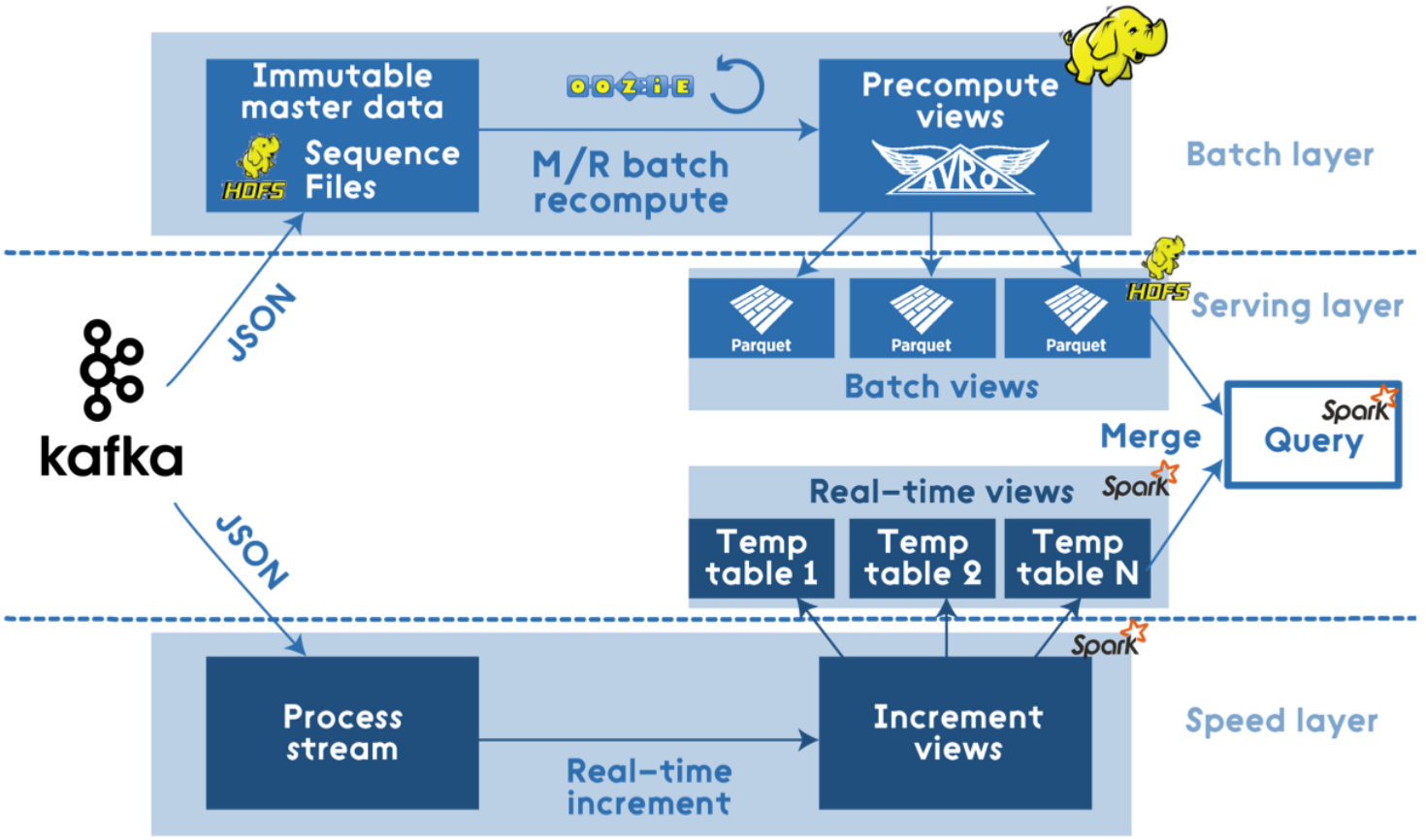


Relevance of data

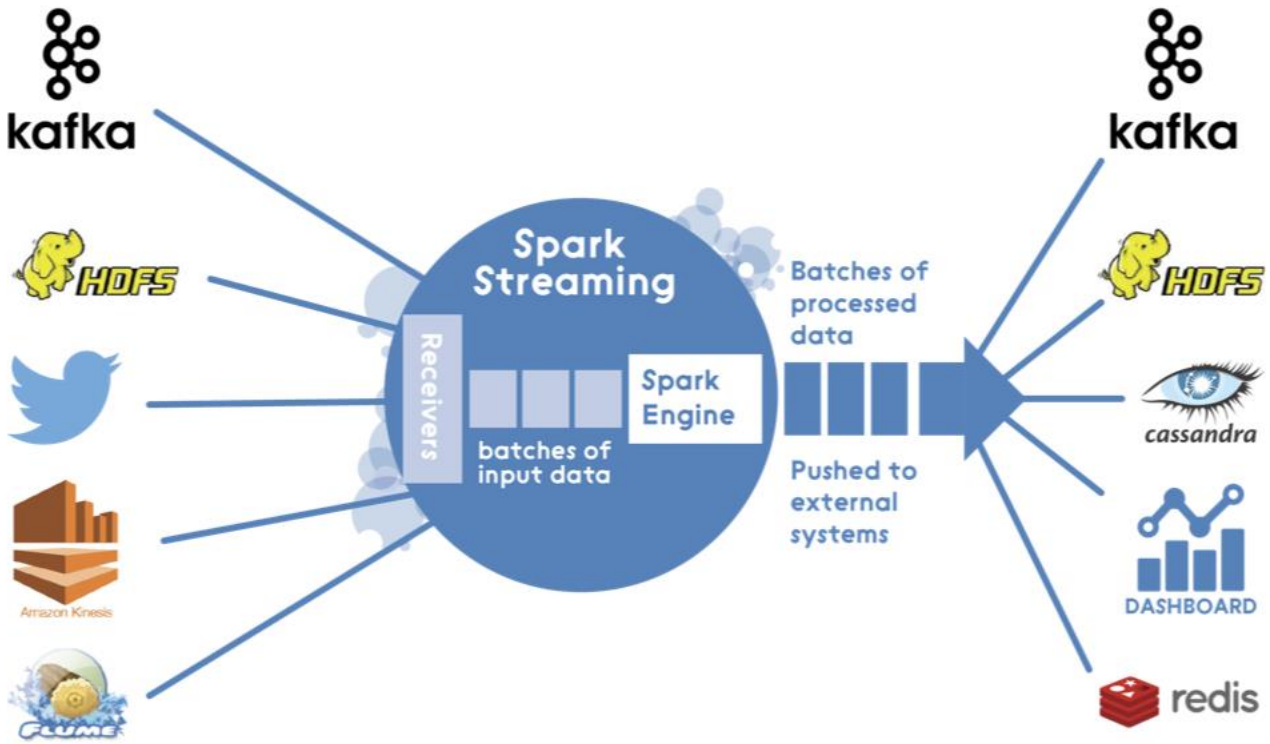
query = function(batch view, real time view)
real time view = function(real time view, new data)
batch view = function(all data)



Lambda architecture: one implementation



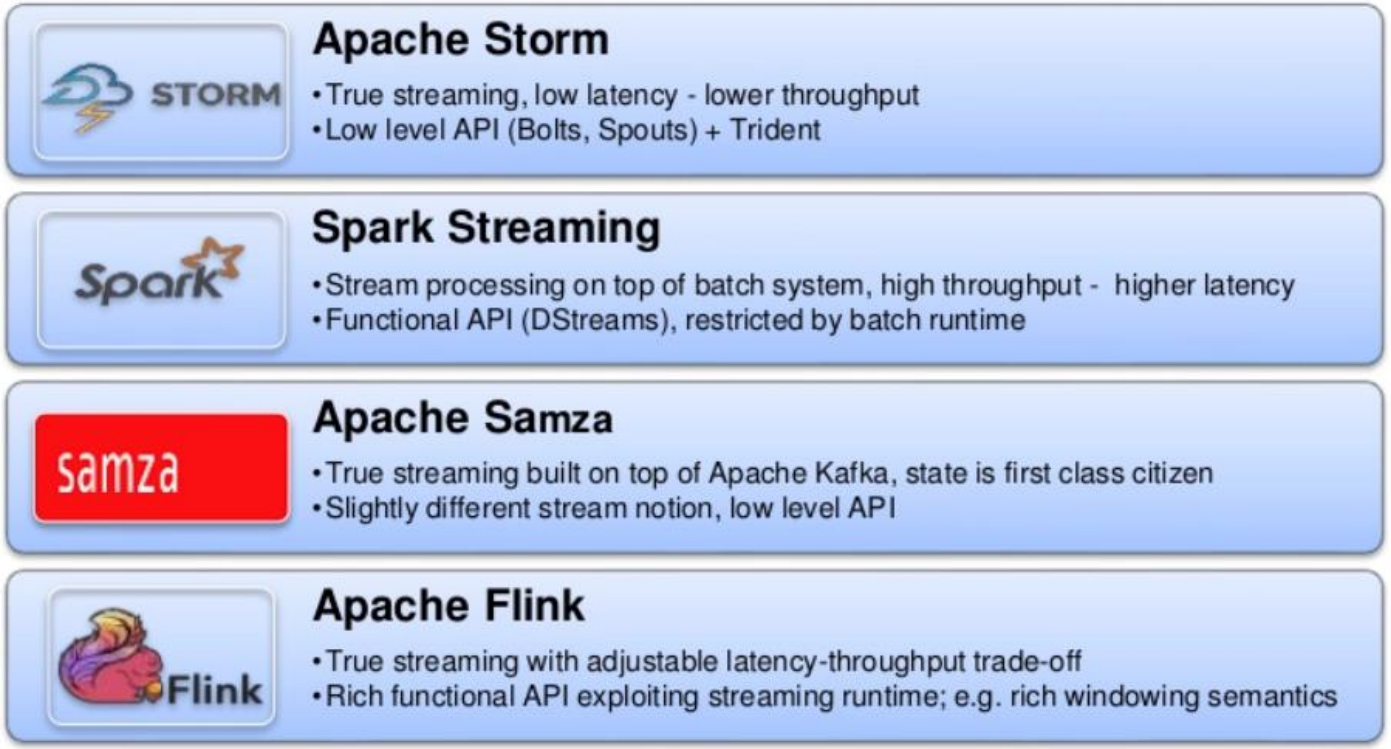
Spark streaming



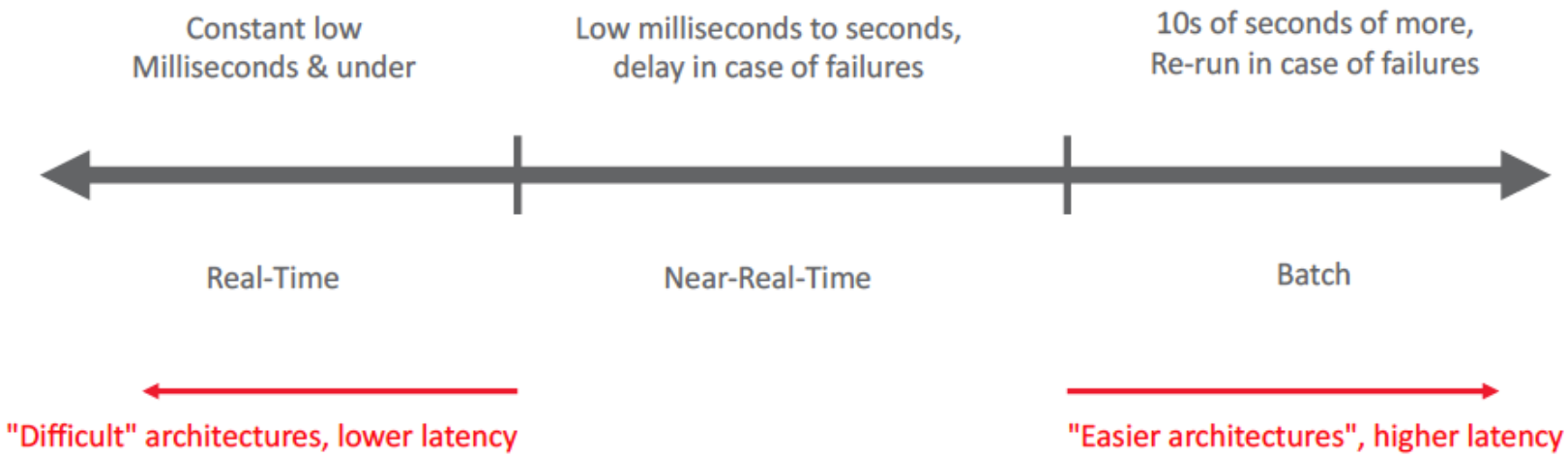
Spark streaming

- Scalable, fault-tolerance stream processing system
- a streaming computation as: a series of very small, deterministic batch jobs
  - Chop up the live stream into batches of X seconds
  - Spark treats each batch of data as RDDs and processes them using RDD operations
  - Finally, the processed results of the RDD operations are returned in batches

Streaming landscape



Stream vs. Batch processing



References

- 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

Thank you for your attention!!!

soict.hust.edu.vn/ fb.com/groups/soict