Streaming vs Batching

Conundrum, May 2020

mark.andreev@gmail.com — Mark Andreev

Agenda

- About data processing pipeline
- Batching approach
- Streaming approach
- Architecture
- Tools
- Conclusion

https://clck.ru/VQL9W

Data pipeline

- Event := (timestamp, payload)
- Pipeline := F($Event_1$… $Event_n$; state)

Batch
Latency - Hours

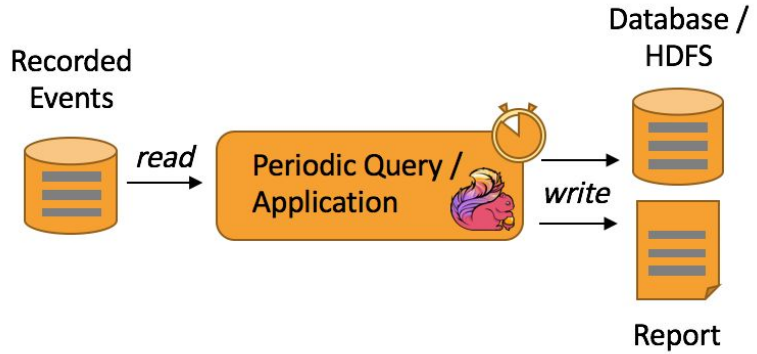
Streaming
Latency - Minutes/Seconds

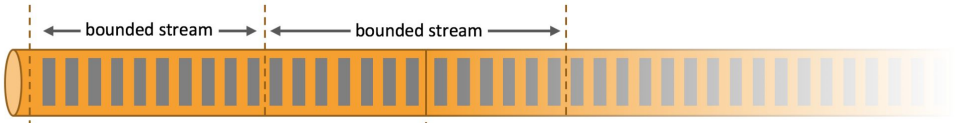
Examples:

- Web Banner CTR
- Forecast Ads Budget Consumption
- Fraud prevention
- Compute analytics over session windows

Batching approach

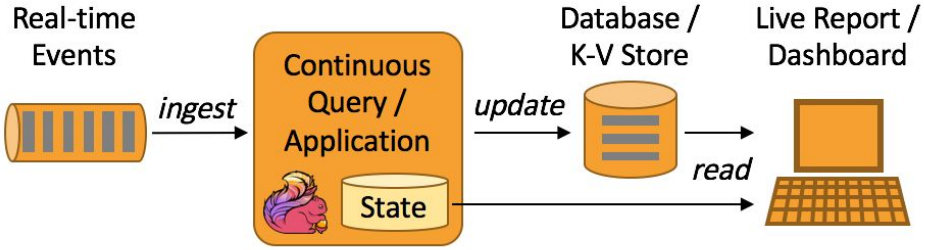
- Fetch periodically
- Bounded stream approach
- Easy to write (look like SQL select/insert)
- Scheduled by standalone tool (like Airflow)

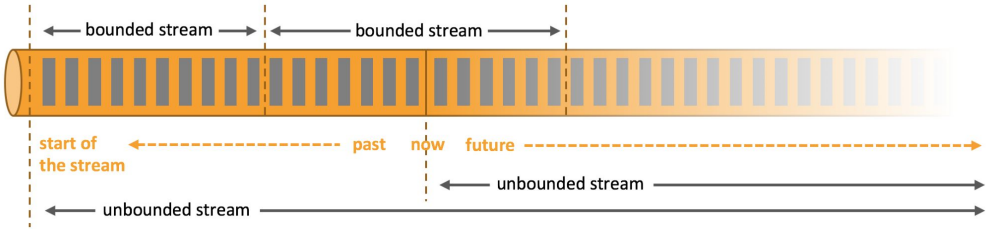




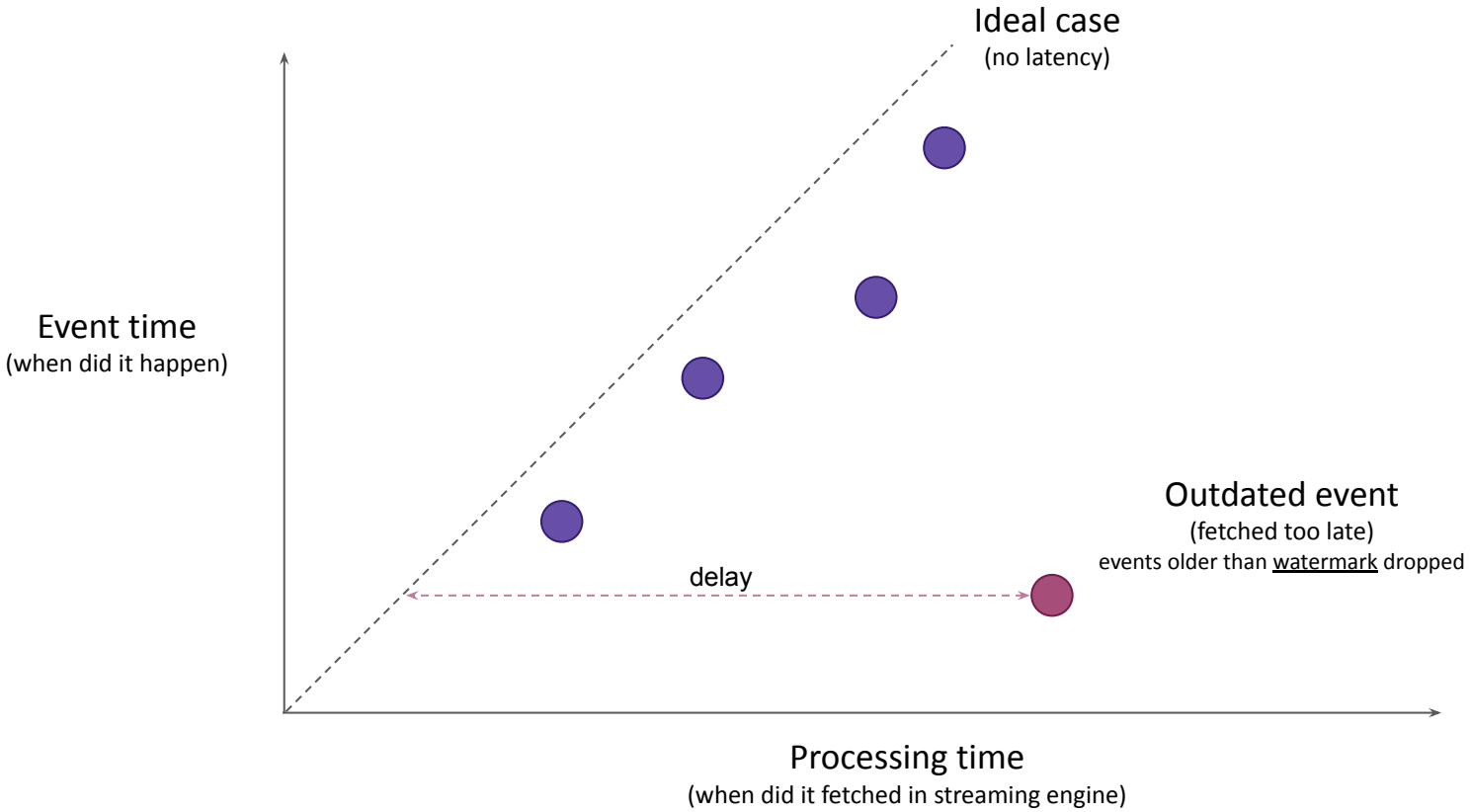
Streaming approach

- Fetch as soon as possible
- Unbounded stream approach
- Require KV storage for state

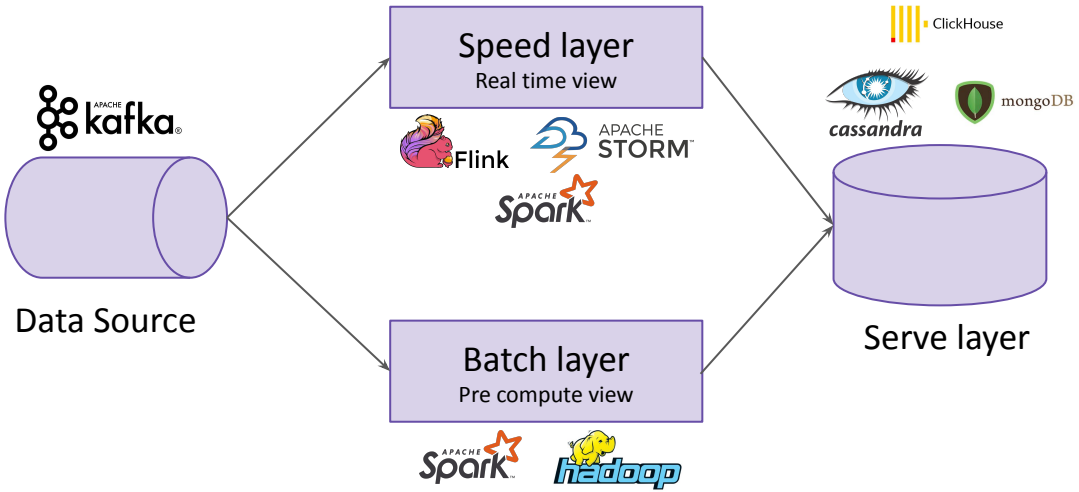




Streaming approach [Watermarks]

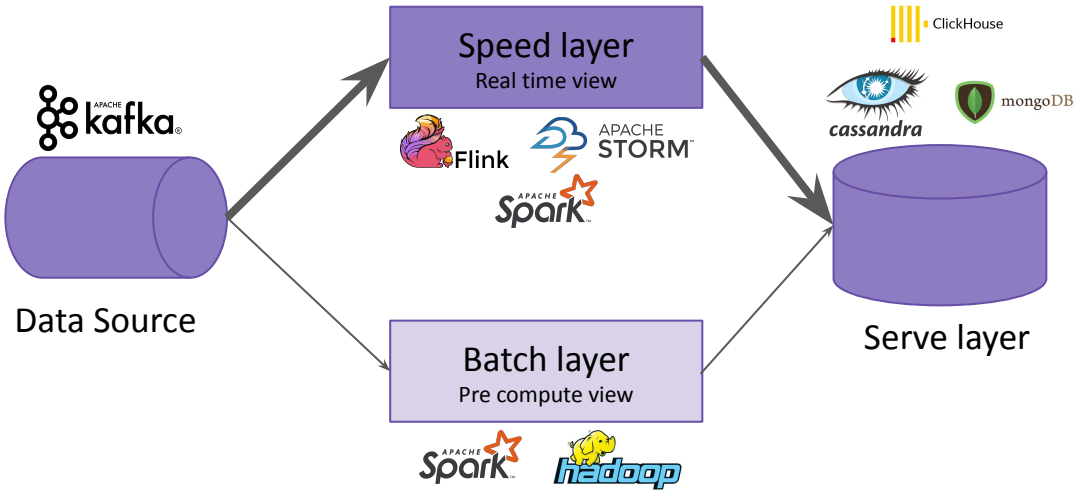


Architecture [Lambda Architecture]



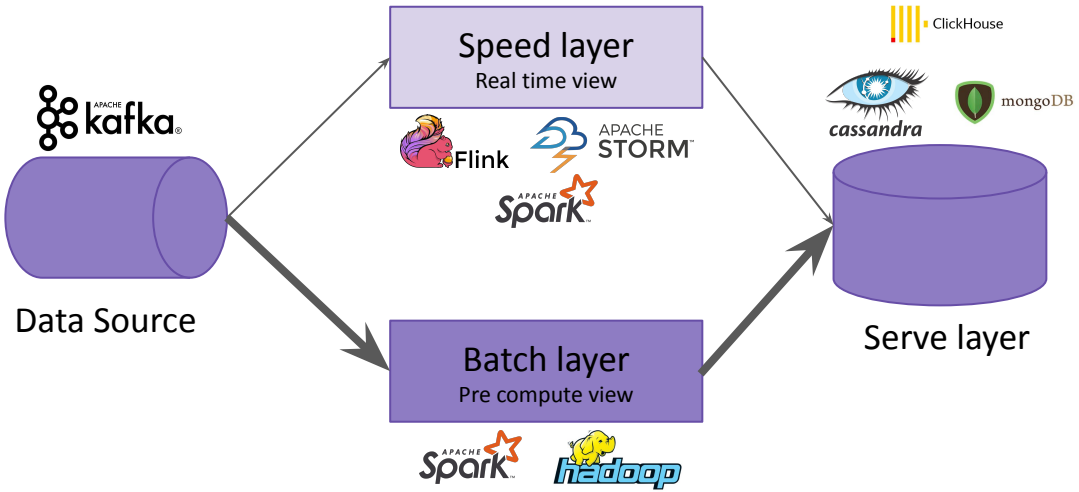
https://clck.ru/VR4Lg

Architecture [Lambda Architecture]



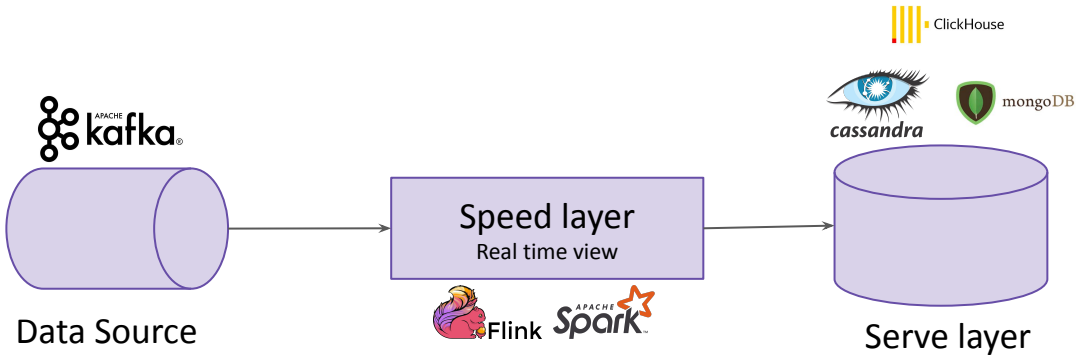
https://clck.ru/VR4Lg

Architecture [Lambda Architecture]



https://clck.ru/VR4Lg

Architecture [Kappa Architecture]



https://clck.ru/VR4Bm

Architecture [Lambda vs Kappa]

Lambda	Kappa
Batch + Streaming	Streaming
Two scripts for both approach	Single script
Query all data	Incremental algorithms on deltas
Batch is reliable Streaming is approximate	Streaming with consistency

Architecture [Lambda vs Kappa]

Lambda	Kappa
Batch + Streaming	Streaming
Two scripts for both approach	Single script
Query all data	Incremental algorithms on deltas
Batch is reliable Streaming is approximate	Streaming with consistency

Architecture [Lambda vs Kappa]

Lambda	Kappa
Batch + Streaming	Streaming
Two scripts for both approach	Single script
Query all data	Incremental algorithms on deltas
Batch is reliable Streaming is approximate	Streaming with consistency

Architecture [Lambda vs Kappa]

Lambda	Kappa
Batch + Streaming	Streaming
Two scripts for both approach	Single script
Query all data	Incremental algorithms on deltas
Batch is reliable Streaming is approximate	Streaming with consistency

References

- Flink Concepts [https://clck.ru/VQLMU]
- Spark Streaming [https://clck.ru/VQLPA]
- Streaming Systems: The What, Where, When, and How of Large-Scale Data Processing
- Stream Processing with Apache Flink: Fundamentals, Implementation, and Operation of Streaming Applications