Streaming SQL

Streaming machine learning

Apache kafka

Apache Strom

Apache Spark

AWS

Azure

NFR

* **Releiability** – (Fault , failure and Errors )should continue to work correctly even in case of failures
* **Scalability** – able to cope up with increase in load
* **Maintainability** (Operable , simplified , eXtensible)should be able to work currently and adaptative to changes in future
* **Extensibility**
* **Debuggability**
* **Low Latency 🡪** Both read and write response time should be as low as possible .

Scaling with Traditional Databases

* Client 🡪 Web Server 🡪 Queue 🡪 Database shard
* Database sharding is a technique for horizontal scaling of databases, where the data is split across multiple database instances, or shards, to improve performance and reduce the impact of large amounts of data on a single database.
* Rise of Big Data

Structured 🡪 Database

Semi Structured 🡪 XML

Unstructured 🡪 Web page Images

Terabyte , Peta Byte , ZettaByte of data

3v 🡪 hig Volume , High Variety , High Velocity

Processing 🡪 Codt Effective , Innovative forms of processing

Insights 🡪 Enhanced insight , enhanced decision making

3 properties of Data

* Rawness
* Immmutability
* Eternity – data is always pure and true

Data source 🡪 Data Storage 🡪 Batch processing 🡪 analytical Data Store 🡪 Analytics and Reporting

Big Data Solutions

* Batch Processing
* Real time applications
* Interactive exploration
* Predictive analytics and machine learning

BDS Advanctage

* Technology choices
* Performance through parallism
* Elastic Scale
* Interoperability with existing solutions
* Complexity
* Skillset
* Technology maturity

Real time Systems: Twitter Example., tracking flight at Delhi, Stock updates

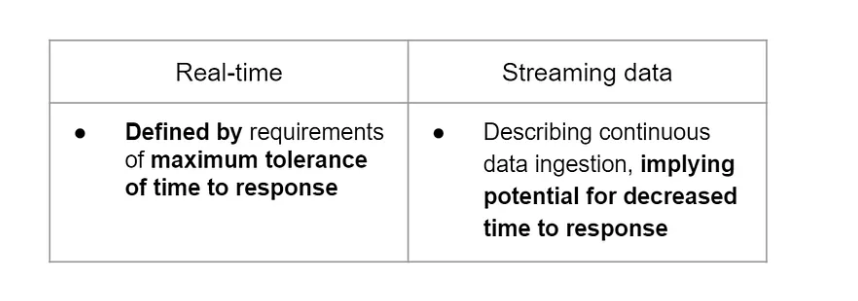
Classification

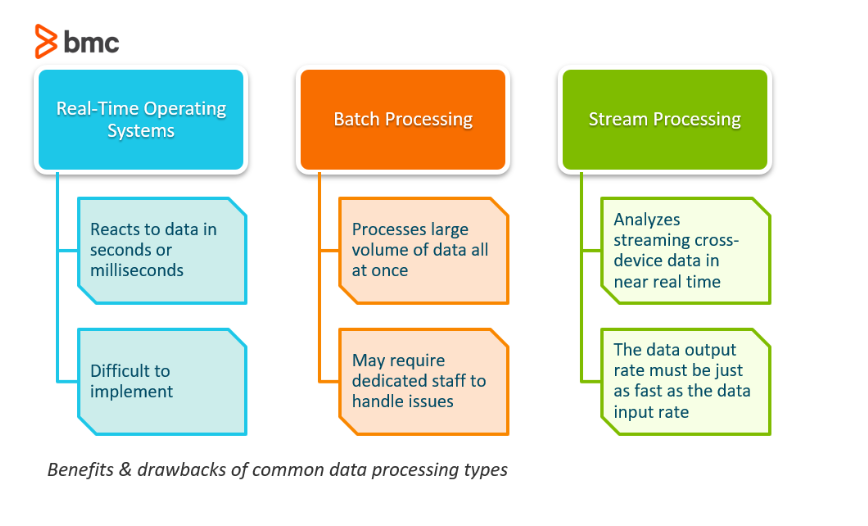
Hard 🡪 Pacemaker 🡪 Microseconds

Soft 🡪 Airline reservation system 🡪 Milliseconds

Near 🡪 Skype Seconds

Real time and streaming systems





Other situations where using real-time systems would be beneficial are:

* ATMs
* Air traffic control
* Anti-lock braking systems in your car

A go-to example of a batch processing job is all of the transactions a financial firm might submit over the course of a week. Batching can also be used in:

* Payroll processes
* Line item invoices
* Supply chain and fulfillment

|  |  |
| --- | --- |
| Batch Processing System | Stream processing |
| Group of transactions is collected over period of time |  |
| Hadoop focused on batch processing | Apache Strom , Sparek Stream processing are framework |
| Speerate program for i/p , process out putput | Immediate action |
| Hig volume of storage  Data Stored and processed sequencially  Sequencial jobs are executed in repeted manner  Payroll processing syste m. Exsmaple | No storage required 🡪 Customer transactions ,activites ,website vistis |

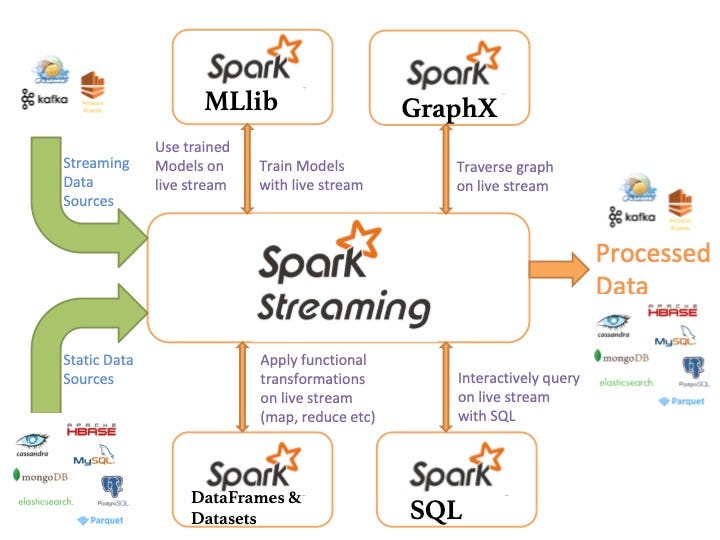
Examples of Stream processing

* Fraud detection application
* Stock Trading system
* Manufacturing system
* Intelligence system

Apache Strom

Apache Storm is**a distributed system for processing big data in real-time.** The Storm is designed to handle massive amounts of data in a fault-tolerant and scale-out manner. It is a streaming data frame that has the capability of the highest reception speed. It’s simple, and you can do all kinds of real-time data manipulations in parallel.

Spark Streaming



Flink

In the current generation, Apache Flink is the big giant tool that is nothing but 4G of Big Data. It’s the true stream processing framework. Flink’s kernel ( core) is a streaming runtime that provides distributed processing, fault tolerance. Flink processes events at a constantly high speed with low latency

Apache Samza

Samza allows you to build stateful applications that process data in real-time from multiple sources including Apache Kafka.

Streaming Data Sources

Operational Monitoring – Internet Speed

Web analytics

Online Advertising - Auctions

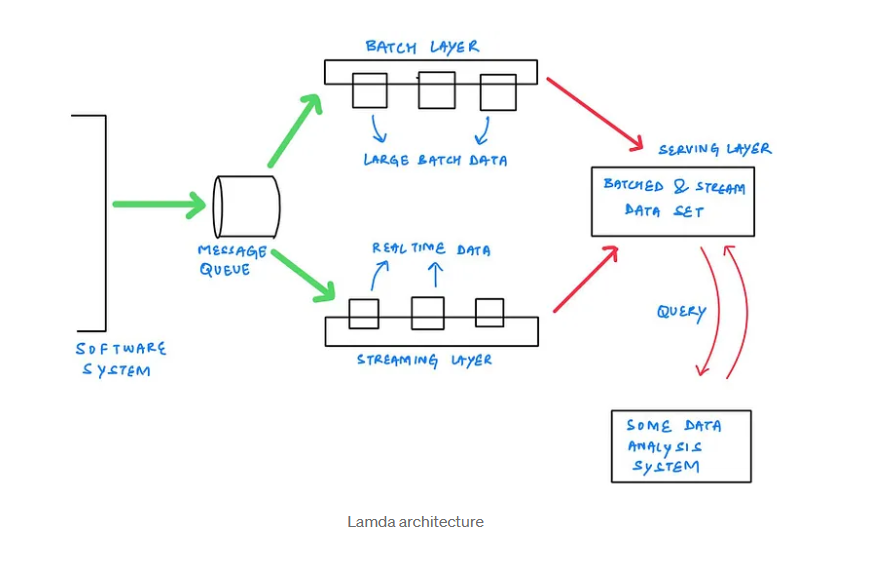
Socis Media

Streaming Data system 🡪 collection 🡪 Data Flow 🡪 Processing 🡪 Storage 🡪 Delivery

Data format 🡪 JSON , AVRO , Thrift are avaialbel

Delivery layer 🡪 HTML + CSS + Java Script + WebScokets usedto create interfaces and update them.

Lambda Architecture



Advantage :

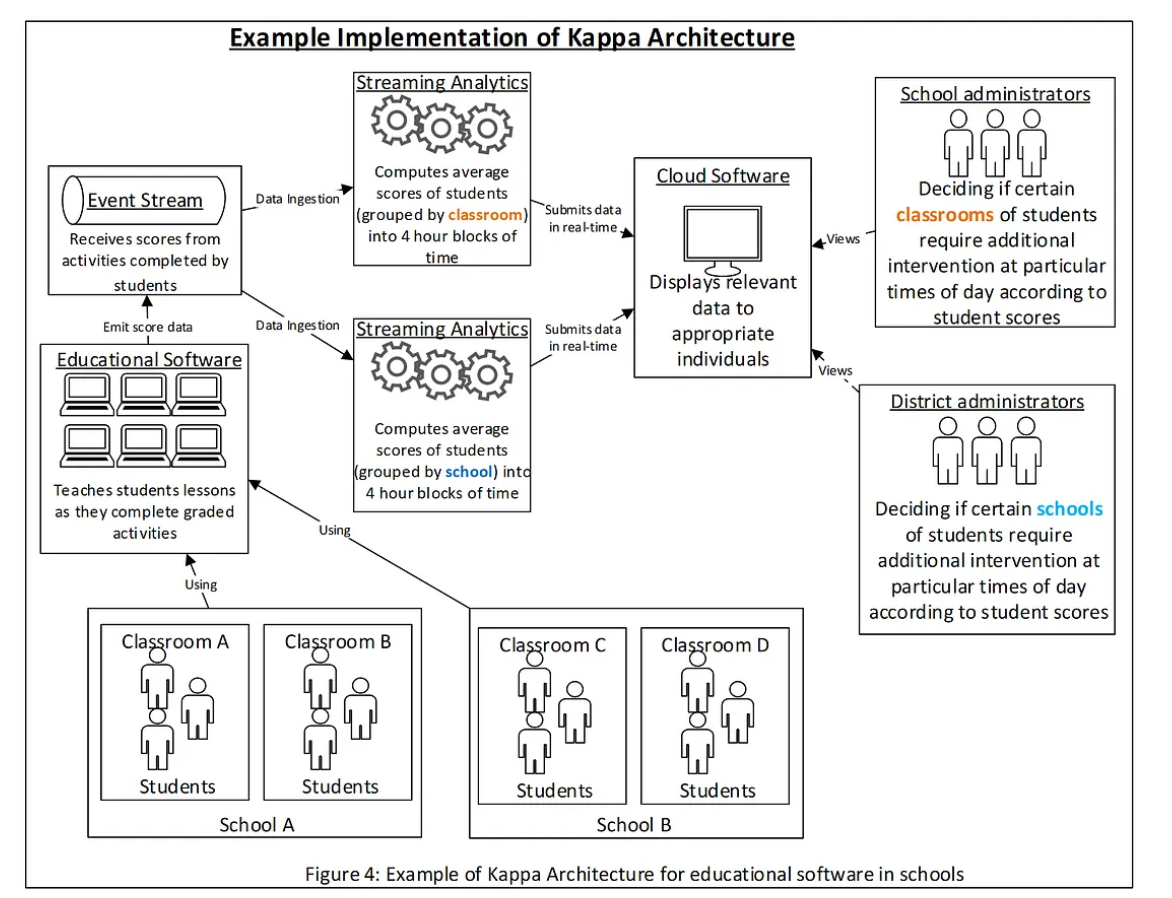
* 1. Manges historical data with faulttolerance
  2. Good blalance of speed and reliability
  3. Scalbe architecture for data processing

Disadvantage

1. Involvement of comprehensive processing
2. Re-rpcossing every batch sycel
3. Difficult for migrate ore reorganize

Kappa Architecture

Kappa Architecture[[1]](https://medium.com/@devin.bost/the-kappa-architecture-8105a3c10f98#_ftn1) utilizes event streaming, stream processing, and stream analytics to replace traditional patterns of data access, data retrieval, and application design. Kappa Architecture eliminates many of the performance and maintainability problems associated with legacy data architectures, especially when analytics are involved.



Adv :

Can be deployed with fixed memory

Fewer resources are required at machine learning on rela time basis .

**Real time system characters**

Distributed systems :

Client 🡪 Cluster ( Collection of Nodes )

Clock synchronization :

Lead to disordering the events

Apache Zookeeper

* Zookeeper is a distributed, open-source coordination service for distributed applications. It exposes a simple set of primitives to implement higher-level services for synchronization, configuration maintenance, and group and naming.

Zookeeper Workflow , CLI and Java Api

Benefits of Zookeeper

* Synchronization
* Serialization
* Reliability
* Atomicity

Types of nodes

* Persistence znodes - by default all znodes are persistent
* Ephemeral Znode – play important role in leader election
* Sequential znode - plays important role in locking and synchronization

How Zookeper works ?

How Zookeeper Works?

Zookeeper follows a client-server model for synchronization across the nodes in a cluster. The service provider nodes are referred to as servers while the service consumer nodes are considered as clients during the ZooKeeper cluster processes. The multiple of servers, called ZooKeeper ensemble, shares the status of the data with server nodes after starting the ensemble. Each server node keeps a copy of the status and transaction details in its local log file. Any change in a server is considered as successful unless it is written to at least half of the servers, which is referred to as quorum, in an ensemble. If any of the servers fail and join the ensemble back, it simply synchronizes its status with the rest of the servers. The ZooKeeper ensemble will offer its services unless more than half of the servers (quorum) fail.

Client nodes can connect to any of the ZooKeeper servers in the ensemble. Once a client is connected, the connection is verified with the client by sending and receiving the acknowledgment signals and a unique ID is assigned. After the election of the ensemble leader, followers synchronize their data with each other, and then clients can connect to servers.

Data Delivery semantics

* At most once delivery
  + Monitoring purpose
  + Inform admin about problems
  + Down sample the data to improve performance
* Atleast once delivery
  + Pushing the message handling semantics to the customer
  + Consumers are free to implement message handling without bother about other consumers
* Exactly once delivery
  + Achieved using queung system Active MQ , RabbitMQ
  + Financial system or advestising system

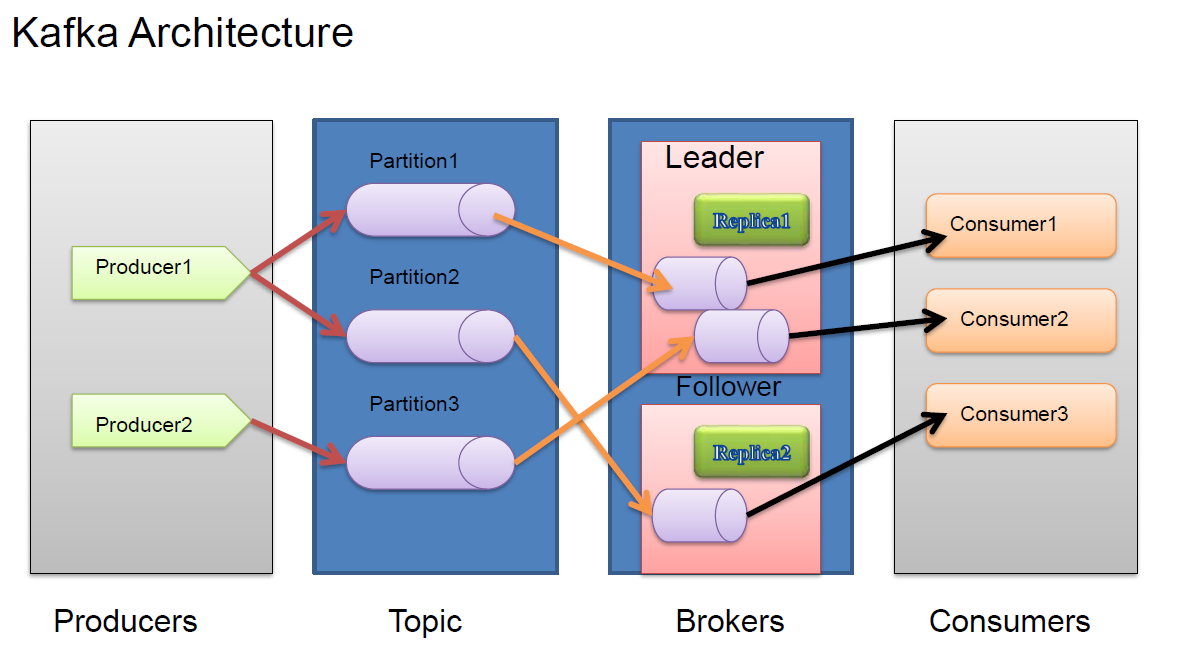
Apache Kafka

Messaging system

* System transfer from one application to another
* Point to Point system : Sender 🡪 Message Queue 🡪 Consumer
* Point to point message system : Sender 🡪 Message Queue 🡪 Consumer
* Fit for large scle data processing
* Integrate wit hStrom and Spark easily
* Suitable for bothe online and offline message transfer

Kafka benefits

* Releiability
* Scablity
* Durability
* Performance



Each message has a unique sequence id called “Offset”

Replica are backup

Replicate has not read or write 🡪 only for data loss recovery

Kafka Producer Api – Appliation intending to write into kafka topics

Kafka Consumer Api – Application intending to read messages from kafka topics

Kafka Streaming Api - Applciations intending to do computation on messages