DISTRIBUTED DATA SCIENCE

KENSU

Data Science for enterprises

STREAMING DATA

IN THE FLOW…

DATA AS A STREAM

- With distributed systems, you want to work with immutable data representations (no update).
- But… it is very likely that the data you want to exploit is not static.
- Thus most datasets can be viewed as the accumulation of data streams.
- E.g. logs, events, …

REAL TIME COMPUTING (?)

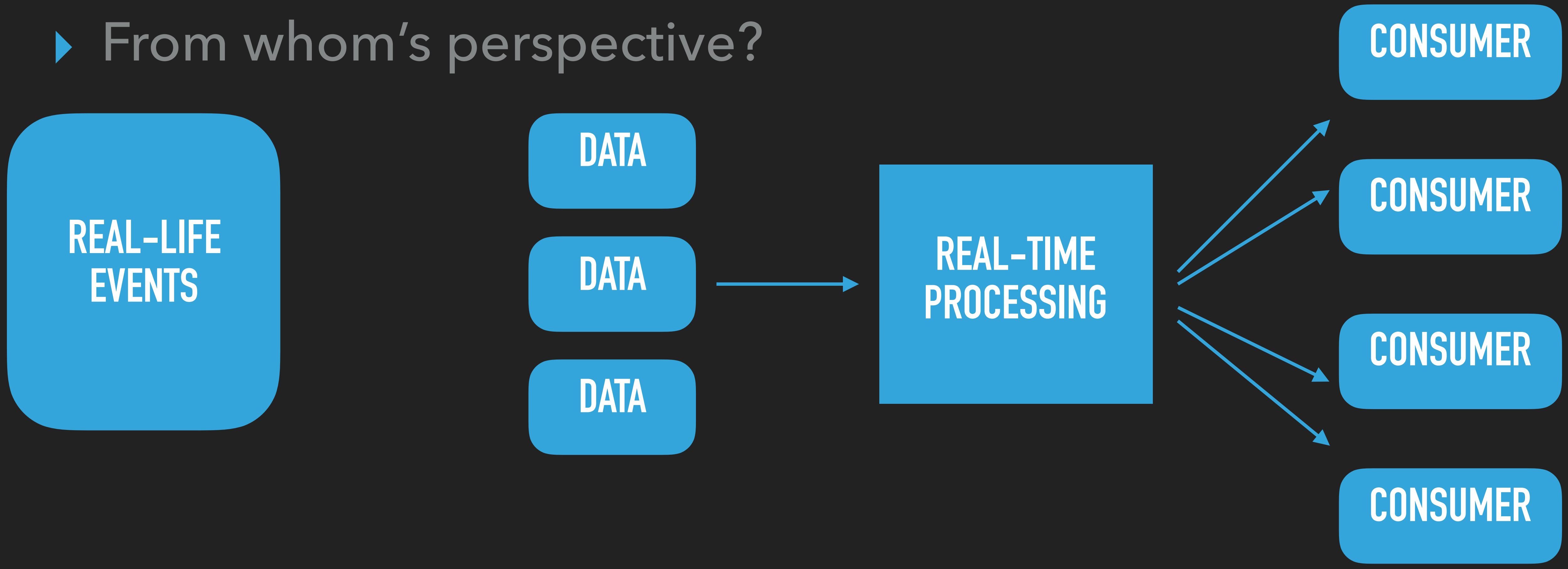
- What makes a system real-time?
- Its total correctness depends also on the time in which operations are performed. Real-time systems have 3 possible levels:
  - Hard: a missed deadline is a total system failure.
  - Firm: few missed deadlines are tolerable but may degrade the overall quality of service, results produced after the deadline are not useful at all.
  - Soft: results usefulness degrade after deadline.

REAL TIME COMPUTING (?)

- If you really need hard real-time computing, then it is very likely that:
  - the computation engine will be embedded
  - dedicated resources are not shared with other processes
  - the consumer of the result is unique and permanently wired
- Doesn’t sound like the king of applications deployed on distributed systems with their inherent latency (network, IO, cpu) and flexibility with data consumption (multiple consumers).

REAL TIME COMPUTING (?)

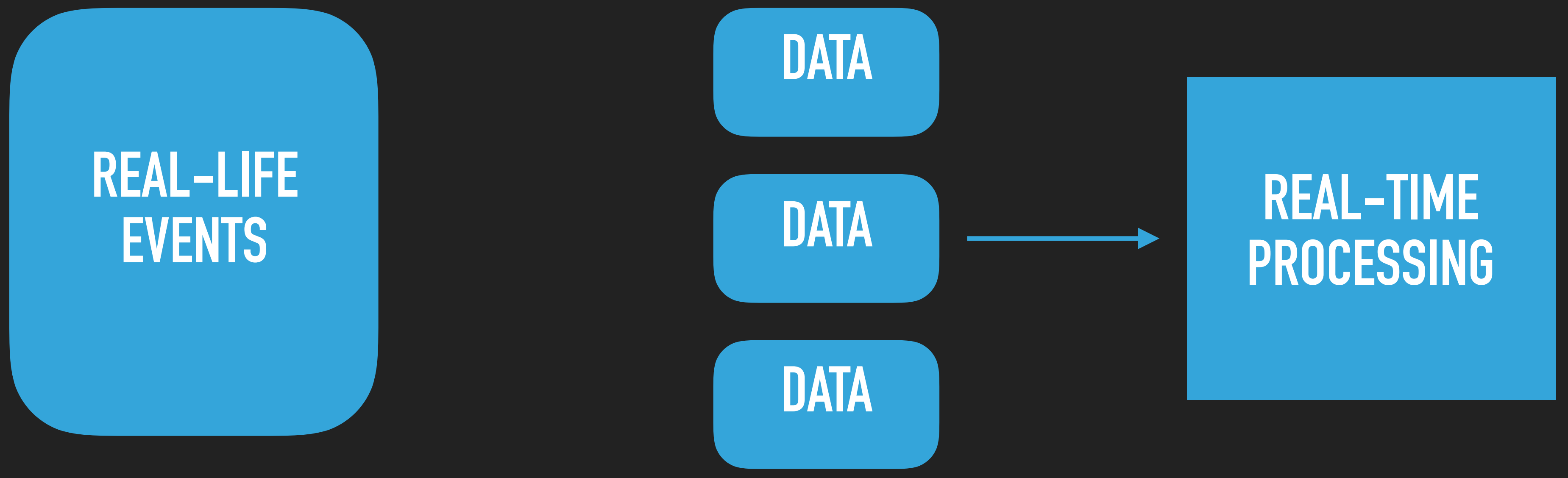
- It can be confusing to classify real-time computing.
  - From whom's perspective?



    - A single consumer? What about the producer?

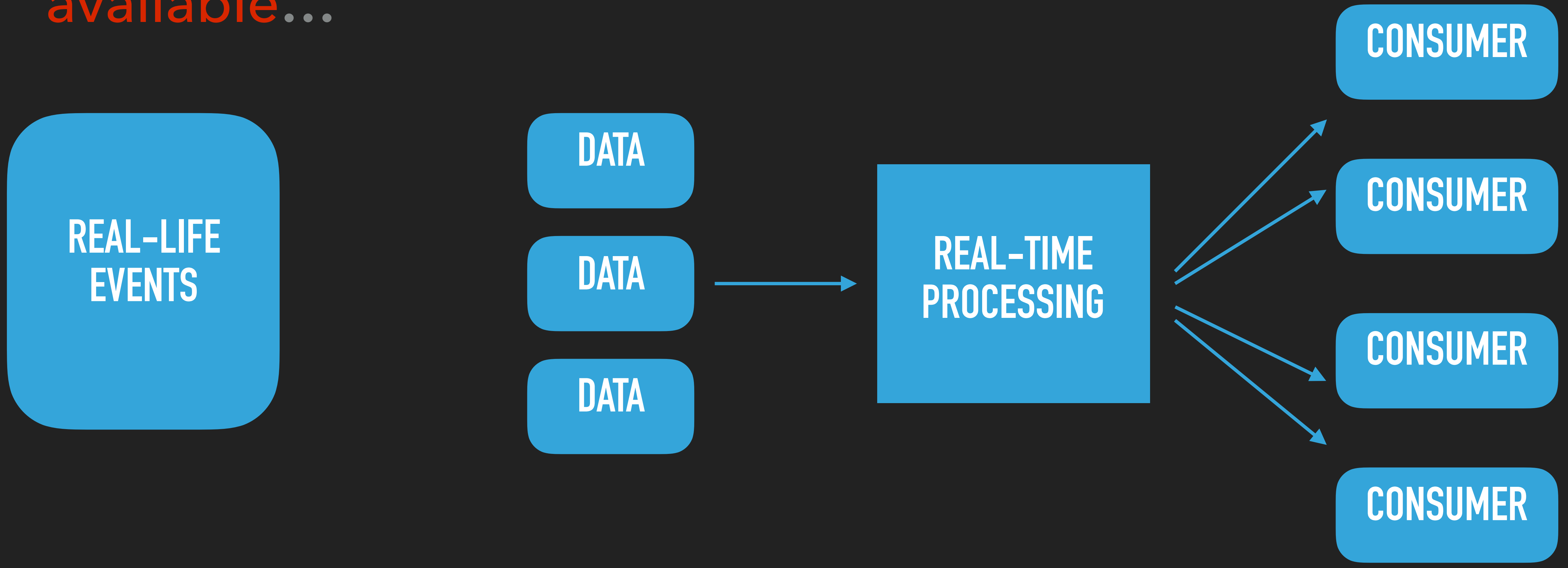
REAL TIME COMPUTING (?)

- Non hard real-time service making computations available…



REAL TIME COMPUTING (?)

- Non hard real-time service making computations available…



- … and consumption happens only when needed.

STREAMING

- A stream is a sequence of data to be made available over time.
  - Potentially unlimited
  - Functions applied to streams are different from the ones applied to batch data.
    - E.g. how to average a value in a stream?

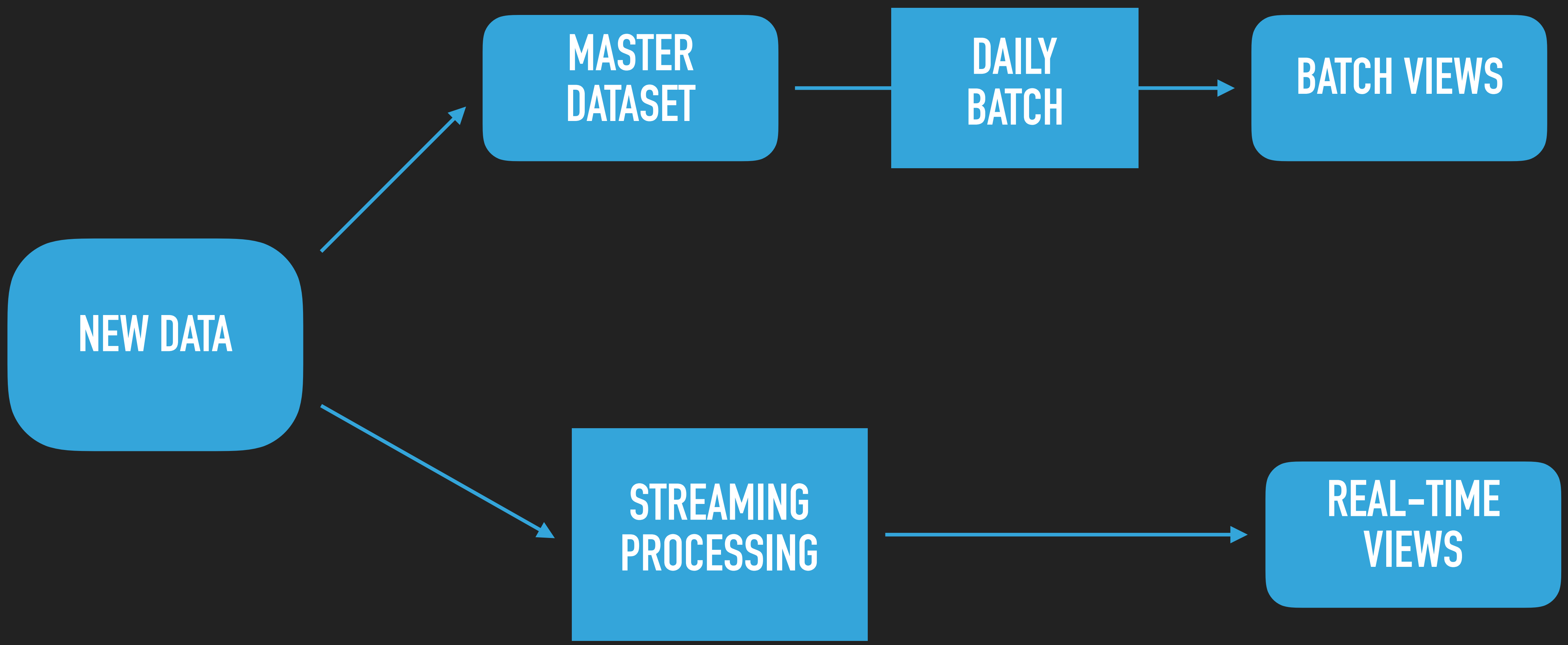
SPARK STREAMING

MICRO-BATCHING AT SCALE.

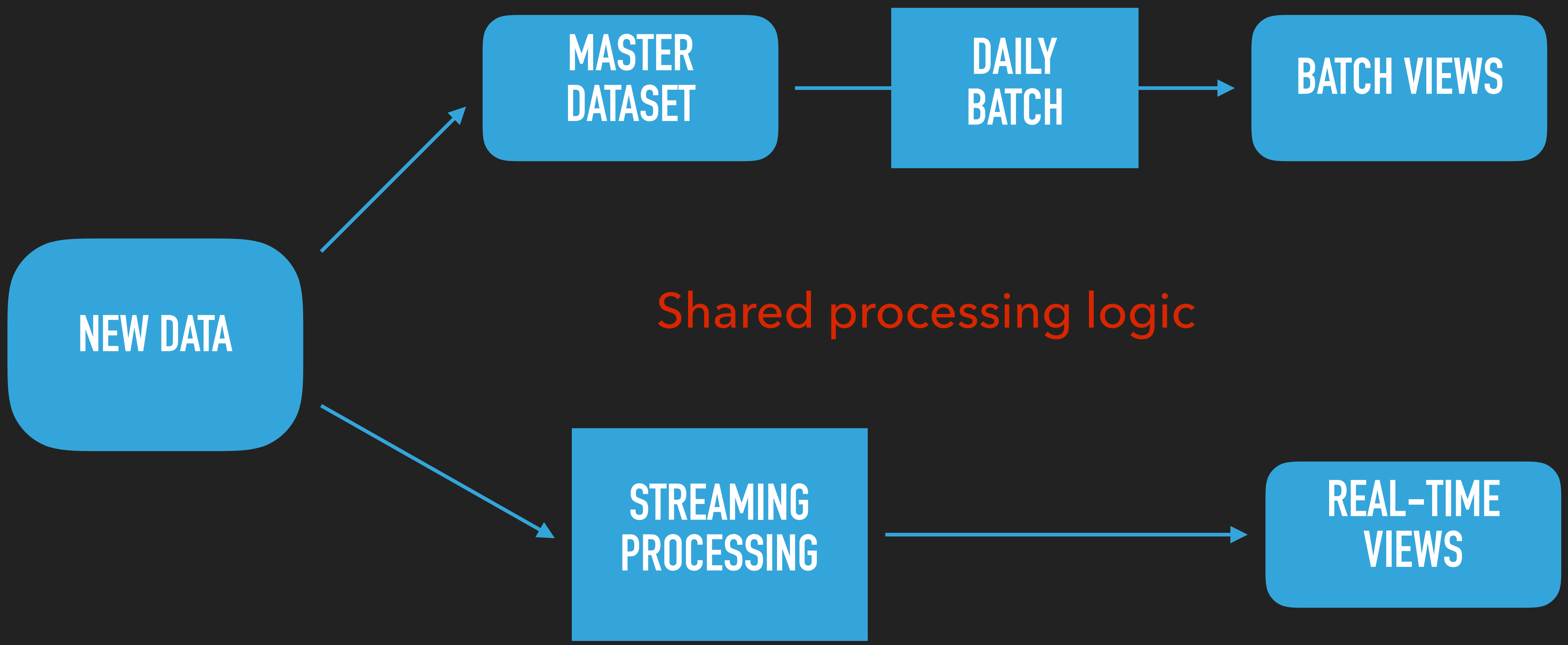
WHY BATCH?

- It can be an advantage to press events immediately at arrival (event processing)
- But… allocating computing resources to a single event can be costly while the deadline requirements may not be that high.
- Spark is a batch processing engine:
  - Spark Streaming leverages the batch engine, hence the code written for batch processing can be reused (cf. lambda architecture)

LAMBDA ARCHITECTURE



LAMBDA ARCHITECTURE

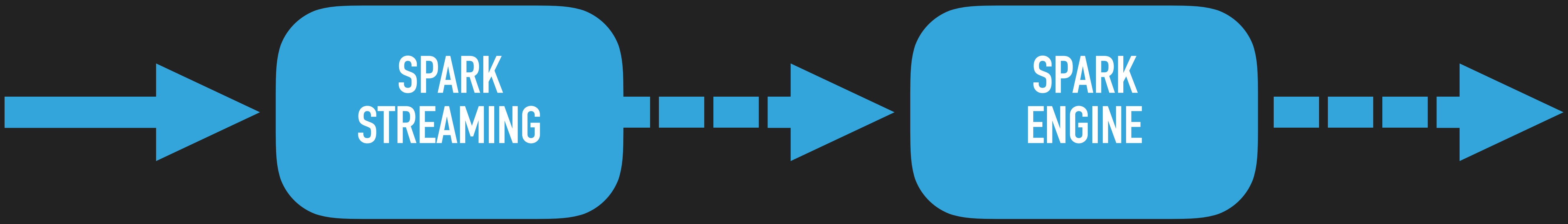


SPARK STREAMING

- Resilient Distributed Datasets (RDDs)
  - Define computations on partitioned datasets
  - Very similar to Collection API
- Discretized Stream (DStream)
  - Infinite sequence of RDDs
  - Same API as RDDs for streaming transformations
  - Batch data stored as RDDs
  - Fixed time interval per batch (0.5 to 60+ sec.)

SPARK STREAMING

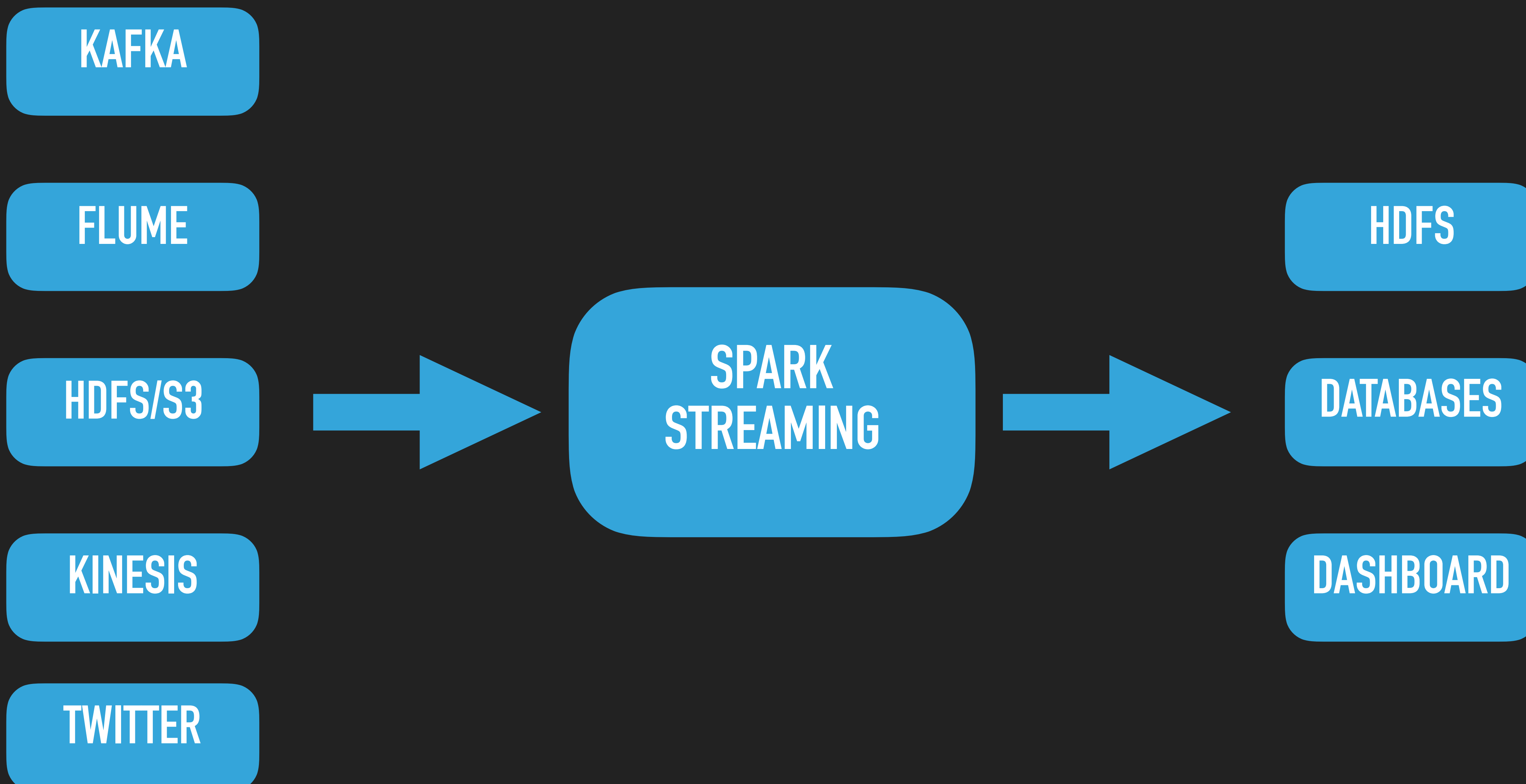
- Discretized Stream (DStream)



Spark Streaming documentation:

http://spark.apache.org/docs/latest/streaming-programming-guide.html

SPARK STREAMING



Spark Streaming documentation:

http://spark.apache.org/docs/latest/streaming-programming-guide.html

NOTEBOOK

CONSUME STREAMING...

IN-MEMORY DATA

FA(S)T STORAGE

IN-MEMORY

- Data is consumed, results are produced:
  - these results will be consumed
- Do we know in advance all use of these results?
- No! Therefore, we need flexibility because exposing all potential results or transformations from the original data is not possible with fast or large datasets.
- Hence the need to prepare views for further detailed/specialized analysis. These views should be fast.

CASSANDRA

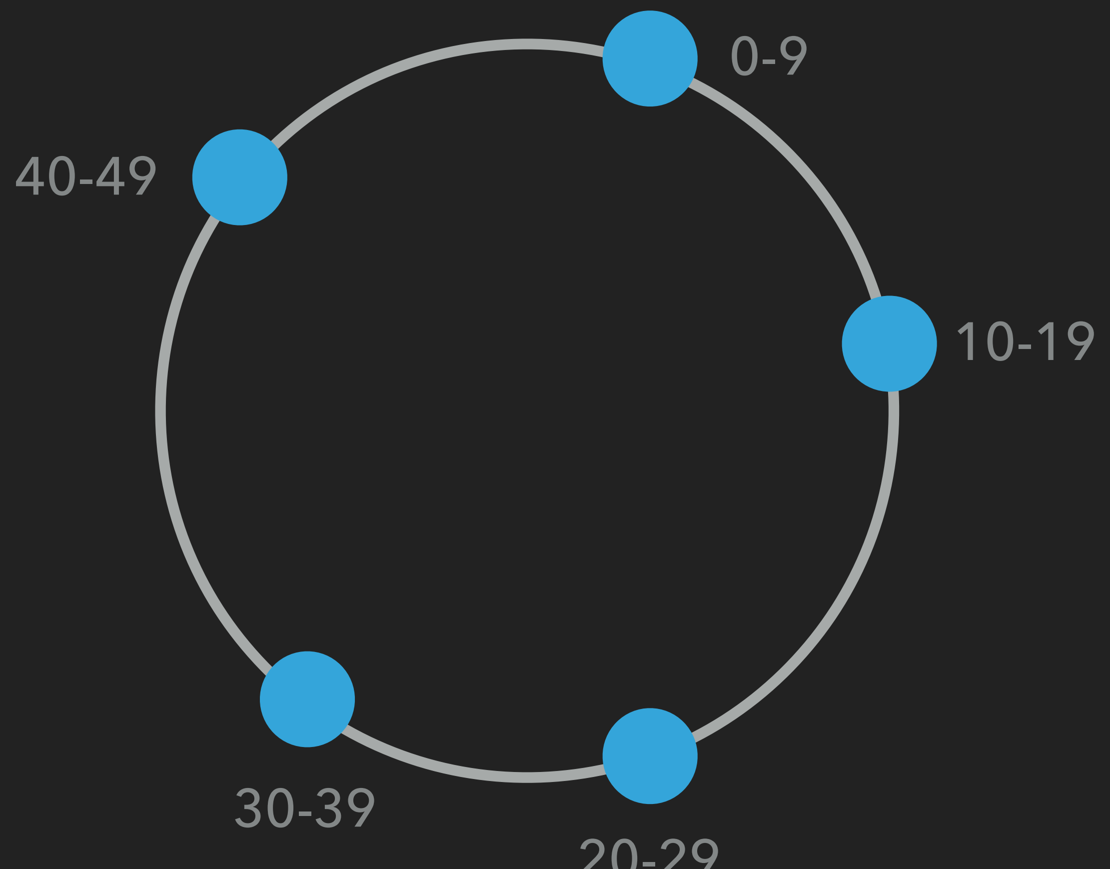
DISTRIBUTED DATABASE

CASSANDRA

- Distributed column-oriented NoSQL database.
  - See Google Big Table, Amazon Dynamo, …
  - Continuous availability and resilience/multi-site.
  - Linear scaling: add nodes for performance and size.
  - Works with commodity hardware, on-premise or in the cloud.
  - True peer-to-peer architecture (one node type)

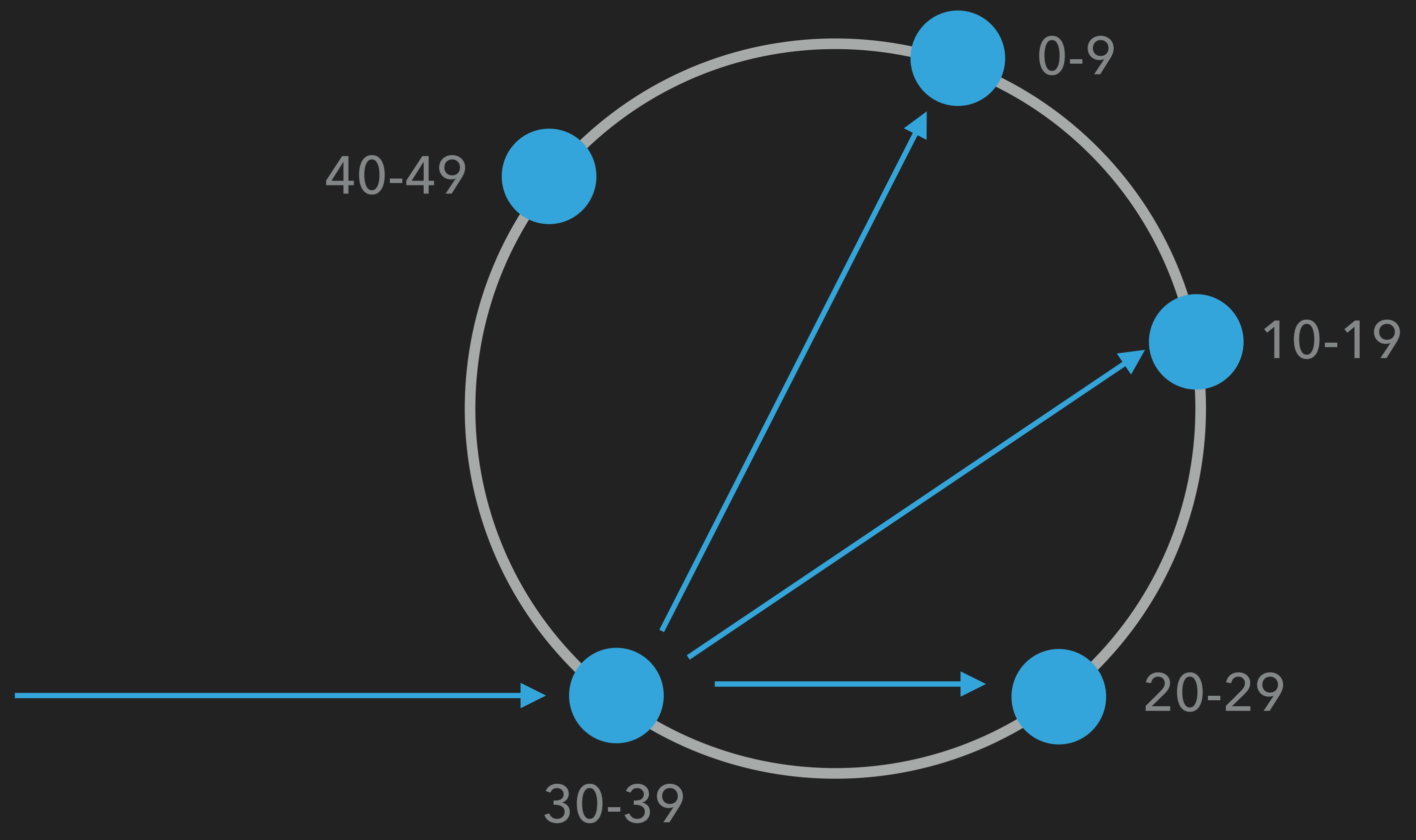
KEY-HASHING

- Data is addressed by keys.
- Consistent key hashing allows every user to know where to find data.



REPLICATION AND CONSISTENCY

- Data is replicated with network topology awareness.
- Consistency tunable:
  - one
  - quorum
  - all



KEYS

```
CREATE TABLE quotes (
  symbol text,
  ts timestamp,
  price double
  PRIMARY KEY (symbol, ts)
  );
```

KEY	Columns				
Symbol	ts:1	ts:2	ts:3	ts:4	ts:5
BAC	10.3	10.1	9.98	9.87	9.78

- Efficient at writing because data is written in sequence, and with range queries, the system knows which nodes to query to get a complete range because columns are sorted.

NOTEBOOK

STORING THE STREAM

STATISTICS, ML & CO.

DATA ANALYSIS

DATA ANALYSIS

- Most likely, you want to do something with this data…
- Model a facet of the data to answer specific questions like:
  - What are the odds to loose a customer in the next period?
  - What is the best recommendation we can give?
  - Is there an arbitrage in the current price of an asset?
  - What are the odds that this transaction is fraudulent?

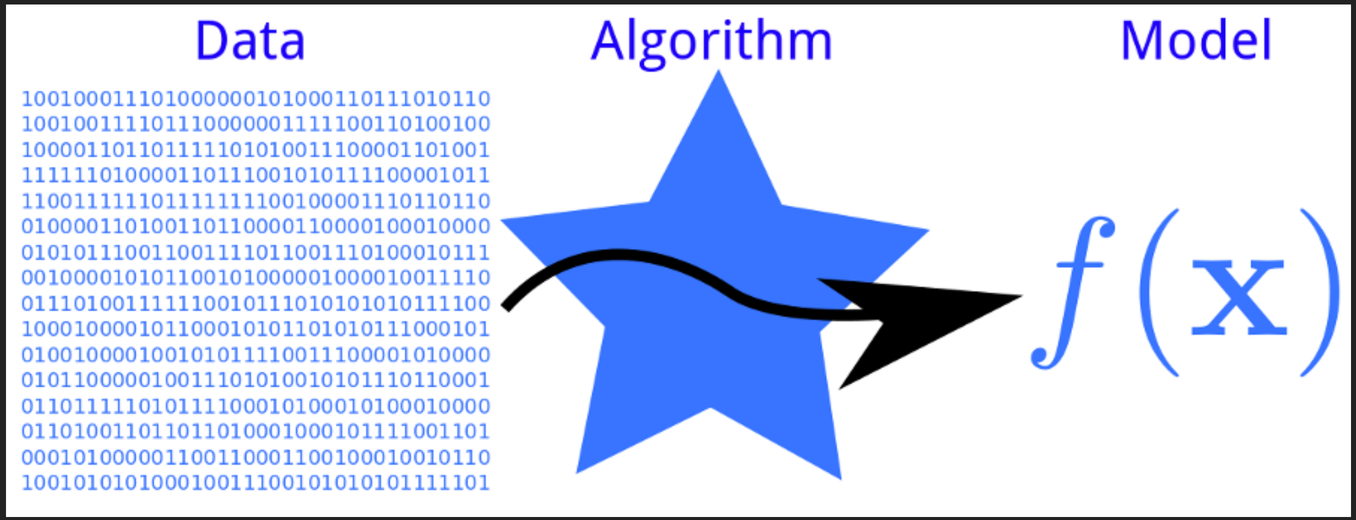
DATA ANALYSIS

- To answer these questions, we need several things:
  - Access to the data.
  - Data cleaning tools.
  - Data transformation tools (e.g. features engineering)
  - Descriptive statistics
  - Machine learning (training models and predicting)

- All this better be done within a single interactive environment…

MACHINE LEARNING IN A NUTSHELL

- ML is the process of learning a function from the data only



- The learning process is an optimization procedure to minimize expected risk (error) of unseen samples.

COMPLEXITY CONTROL

- An infinitely complex function could perfectly fit any finite training sample.
- But… it would capture specifics of the dataset (noise) and not more general features of the process generating the data…
  - this is over-fitting.
- We avoid this by setting constraints on the function complexity.

COMPLEXITY CONTROL

- A too simple function is not able to the characteristics of the generated data…
  - this is under-fitting.
- Procedure to have model complexity in balance are available:
  - regularization (i.e penalty on complexity)
- Tuning the penalty can be done with resampling.

STATISTICS AND MACHINE LEARNING AT SCALE

- New (and old) algorithms are designed to compute statistics and models efficiently on distributed datasets.
- They are based on partial computations on batches (partitions) that can be aggregated in a single statistics (model)
- We need this because we want to work on the dataset, not on a sample in a local mode with lots of data transfer and information loss risk

SPARK AND MLLIB

ML AT SCALE…

MLLIB - DATA STRUCTURES AND LINEAR ALGEBRA

- Distributed implementation of common ML algorithms.
- Data represented as RDDs.
- Individual samples (row) use the LabeledPoint type (for supervised learning) and Breeze library vectors.
- Manipulation of distributed matrices.

All the required linear algebra…

Spark ML documentation:

http://spark.apache.org/docs/latest/ml-guide.html

MLLIB - BASIC STATISTICS

- Summary statistics (e.g. mean, variance, …)
- Correlations
- Hypothesis testing
- Random data generation
- Kernel density estimation

MLLIB - MACHINE LEARNING

- Classification and regression
  - Linear models, naive Bayes
  - Decision trees and random forests
- Collaborative filtering (for recommender systems)
- Dimensionality reduction
  - SVD, PCA, …
- Feature extraction and transformation:
  - NLP methods (TF-IDF, Word2Vec, etc).
  - Scalers
- Evaluation metrics

DISTRIBUTED MACHINE LEARNING

- Many libraries for distributed machine learning are available.
- Example of libraries compatible with Spark:
  - Spark ML (identical to MLlib but uses Dataframes as inputs)
  - KeystoneML by Amplab (similar to Spark ML)
  - H2O: open source ML library (GLM, Deep Learning, Random forests, K-means, PCA, etc). In-memory implementation of map-reduce.
  - DL4J: open-source, distributed, Deep Learning library for the JVM
  - …

NOTEBOOK

STOCK MARKET MODELING

OUR MODEL

- We have time series data for a bunch of stocks
  - they certainly correlate and one can be inferred from the others (?).
- We want a model predicting JPM pricing from the others, thus building a « proxy » for JPM stock price.
  - We will use a supervised learning from data hosted in a database.
  - A linear regression model should make the cut…