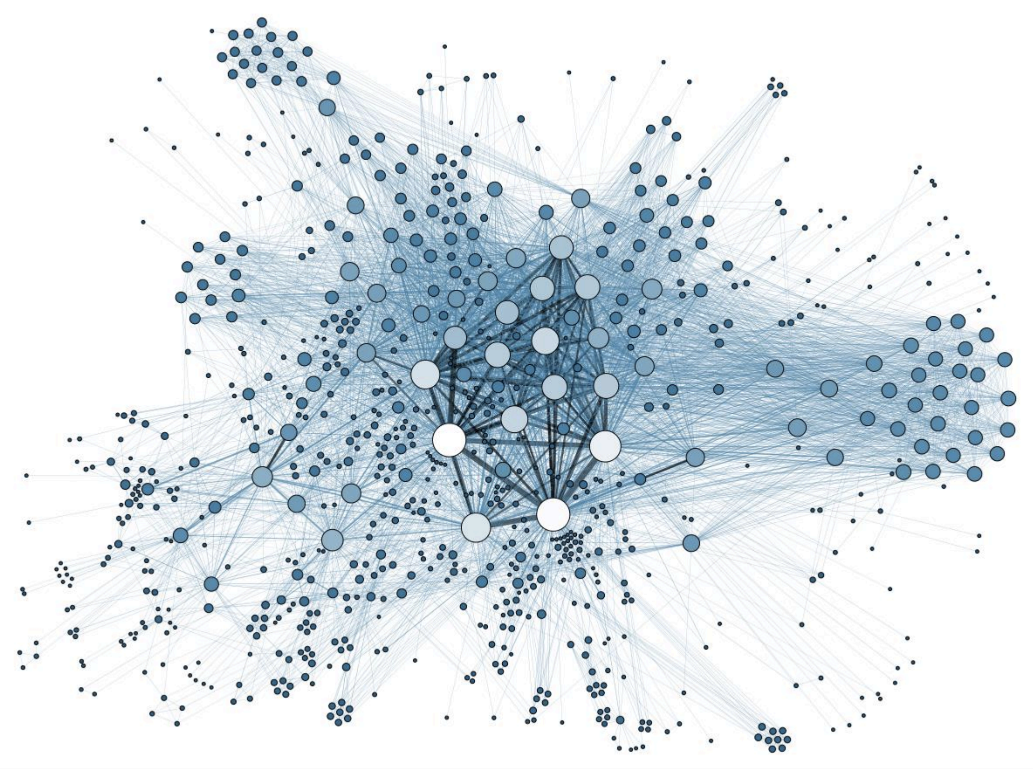
Why is Big Data Processing Different?



challenges of ingesting and processing

After this video you will be able to..

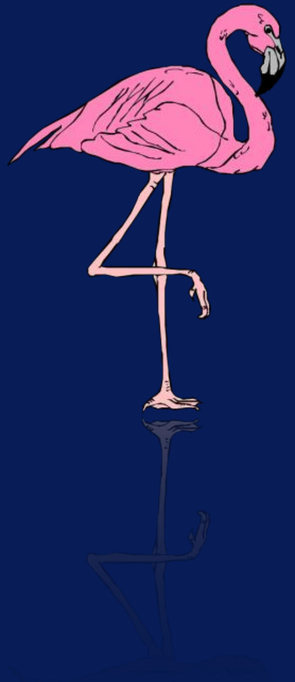
- Summarize the requirements of programming models for big data and why you should care about them

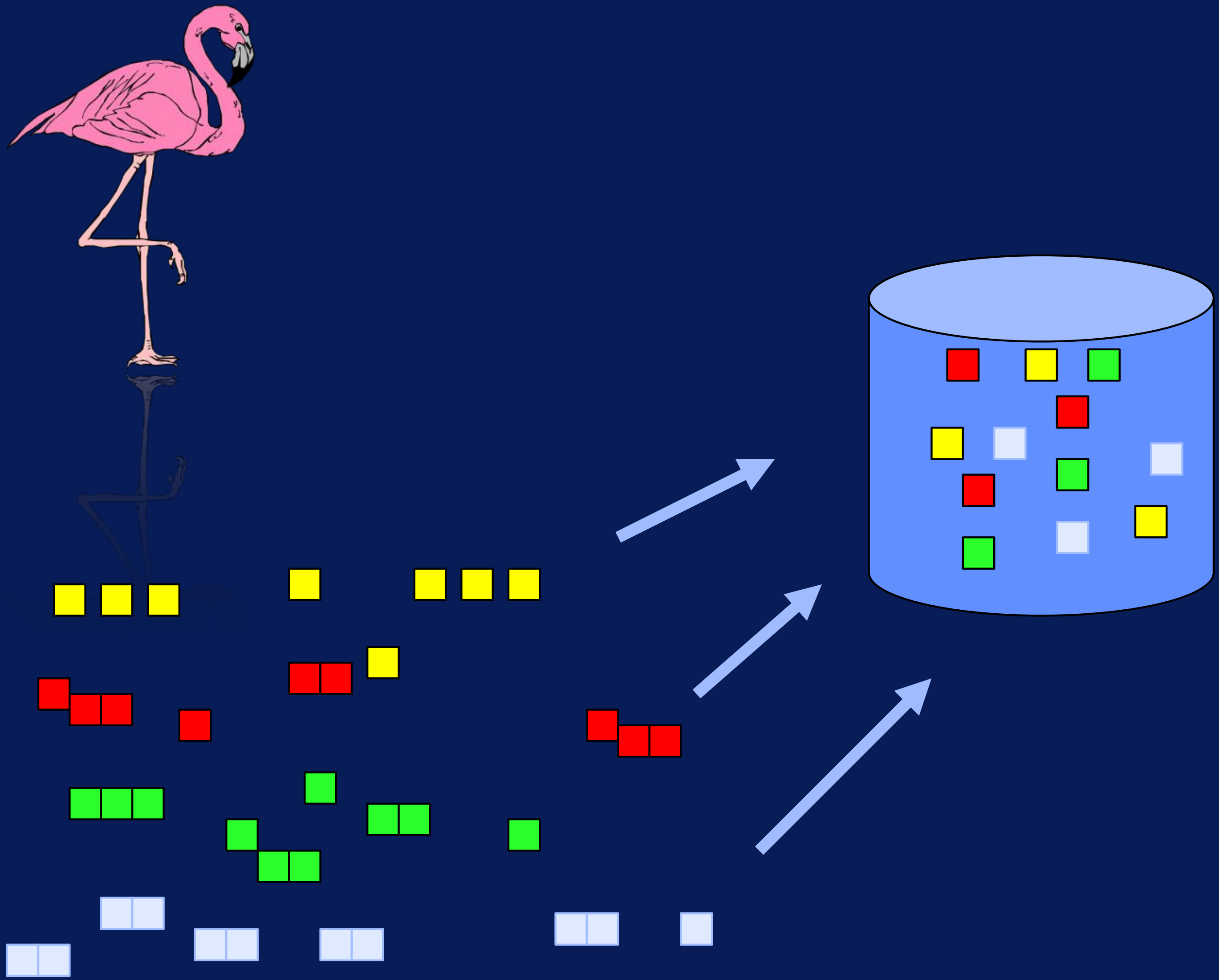
- Explain how the challenges of big data related to its variety, volume and velocity affects its processing

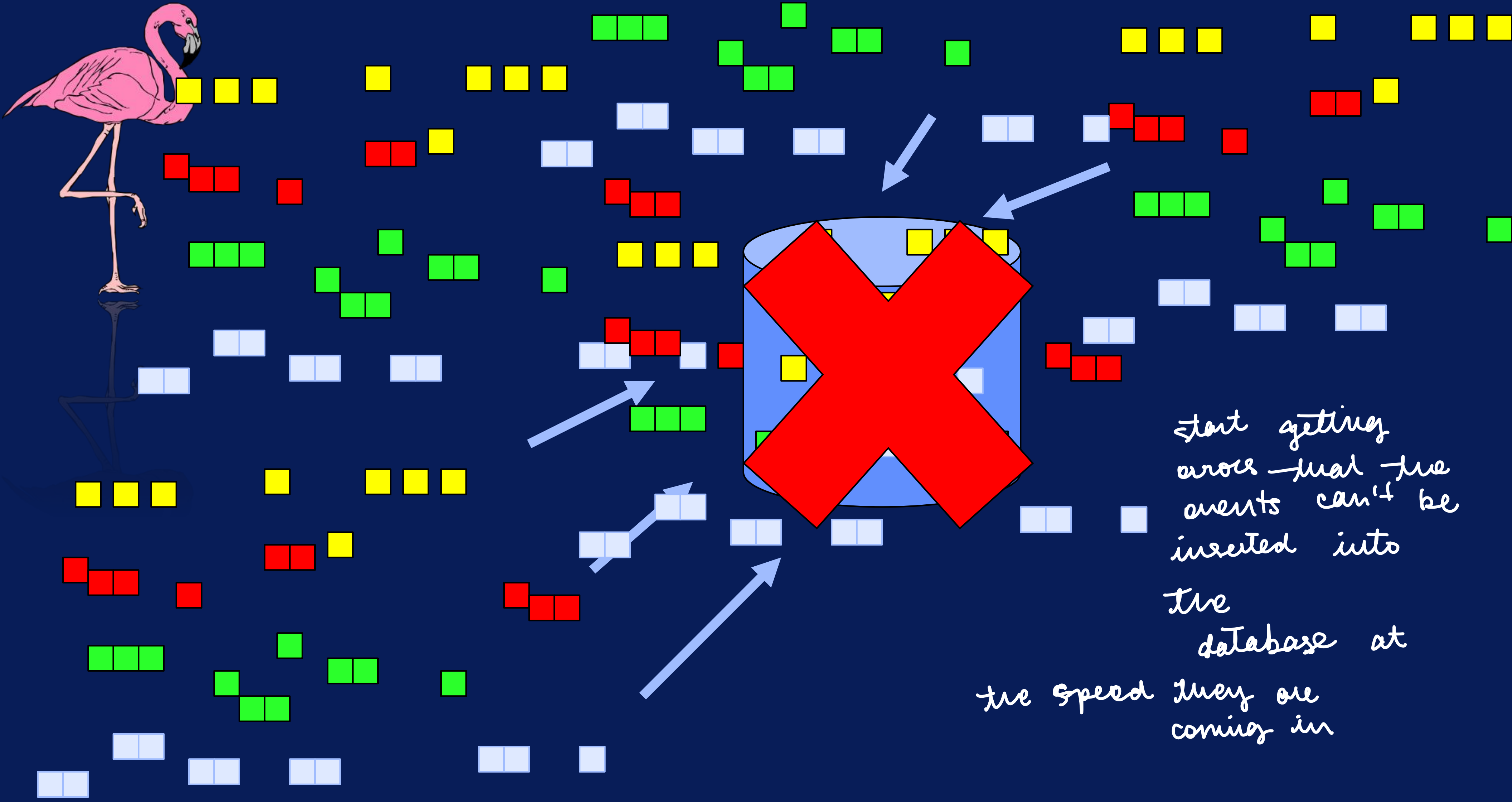
Requirements for Big Data Systems

* online gaming use case

A Big Data System for an Online Game



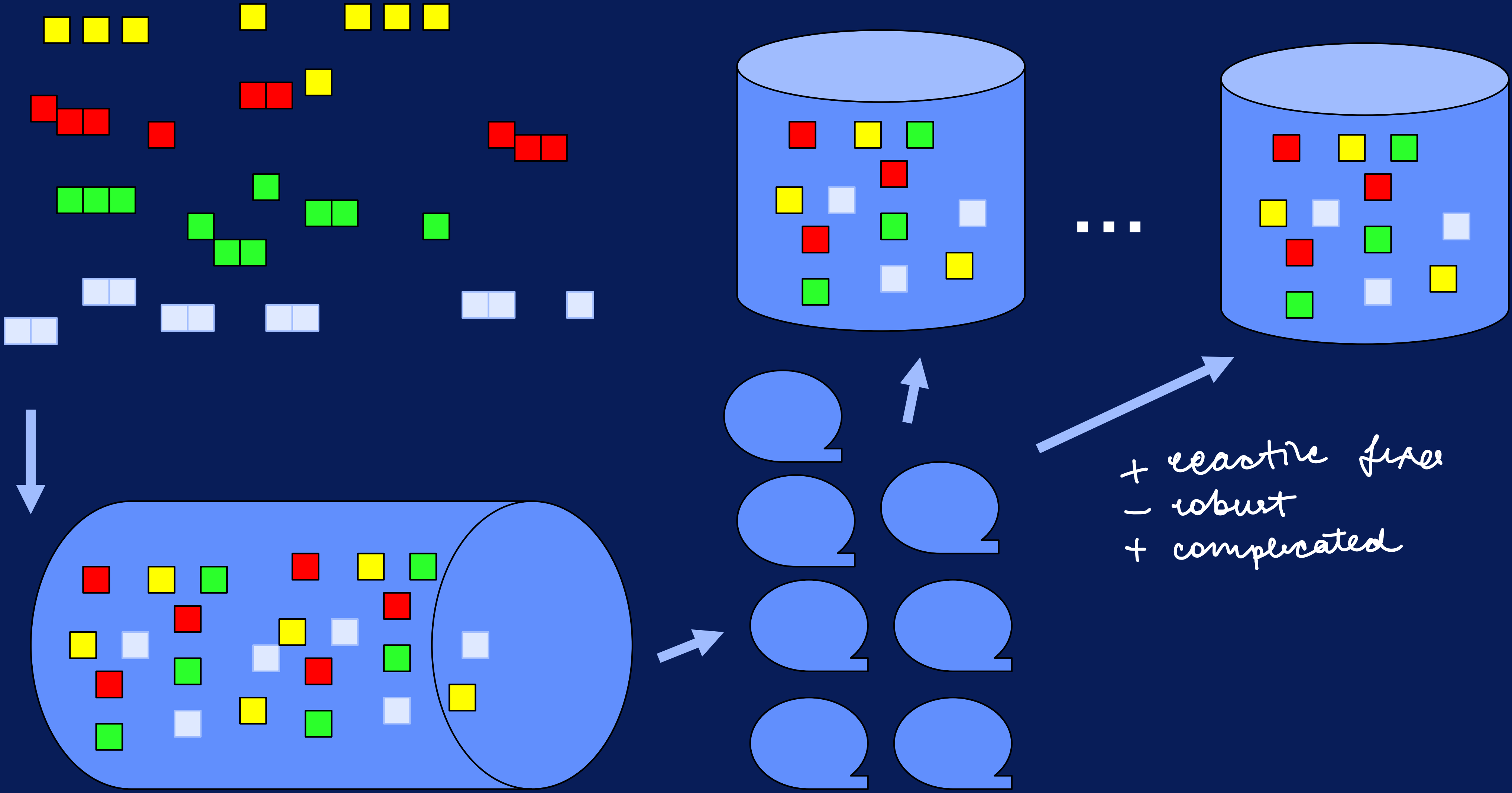




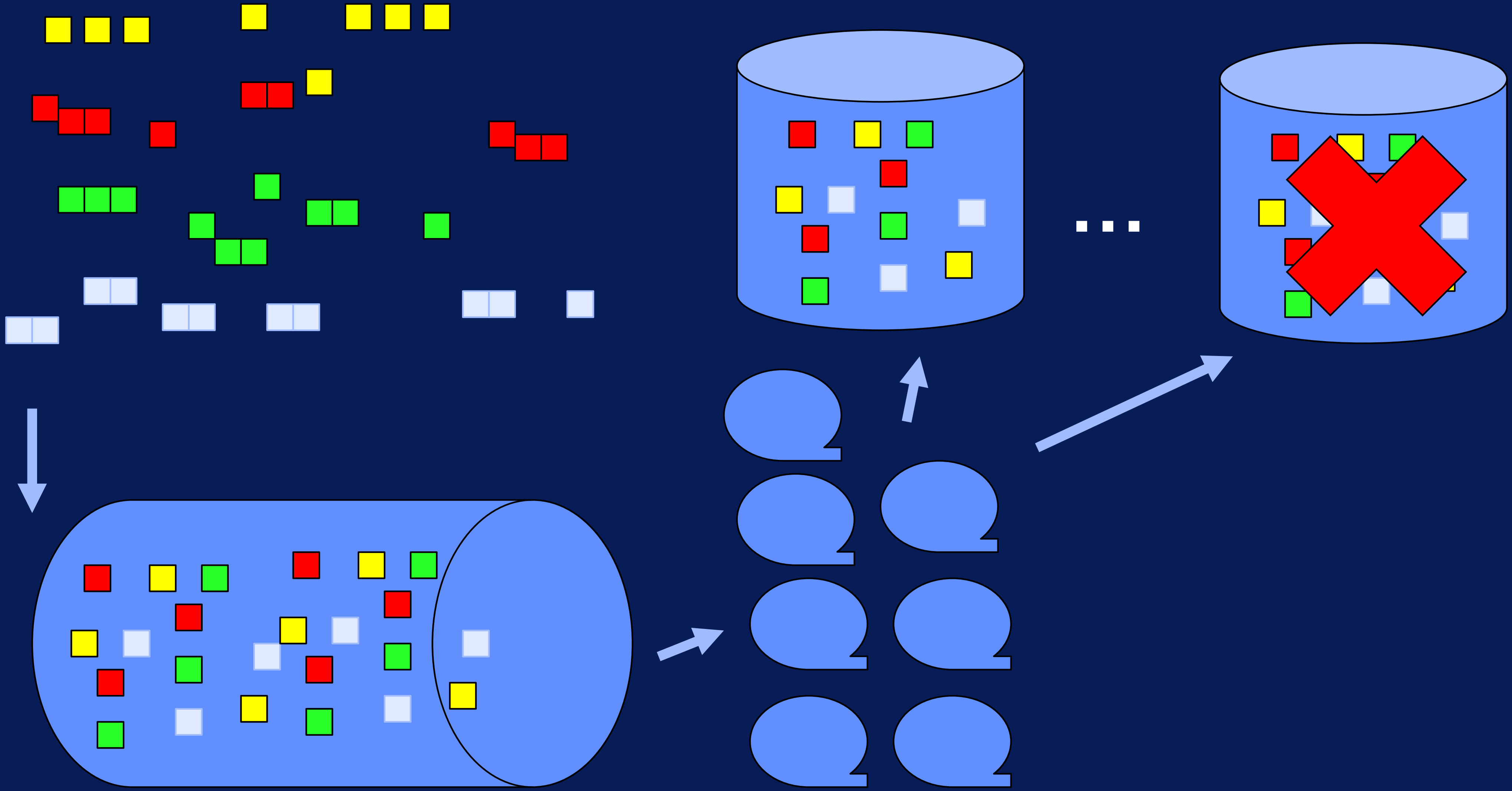
start getting errors that the events can't be inserted into the database at the speed they are coming in processing then to be organized in windows of time

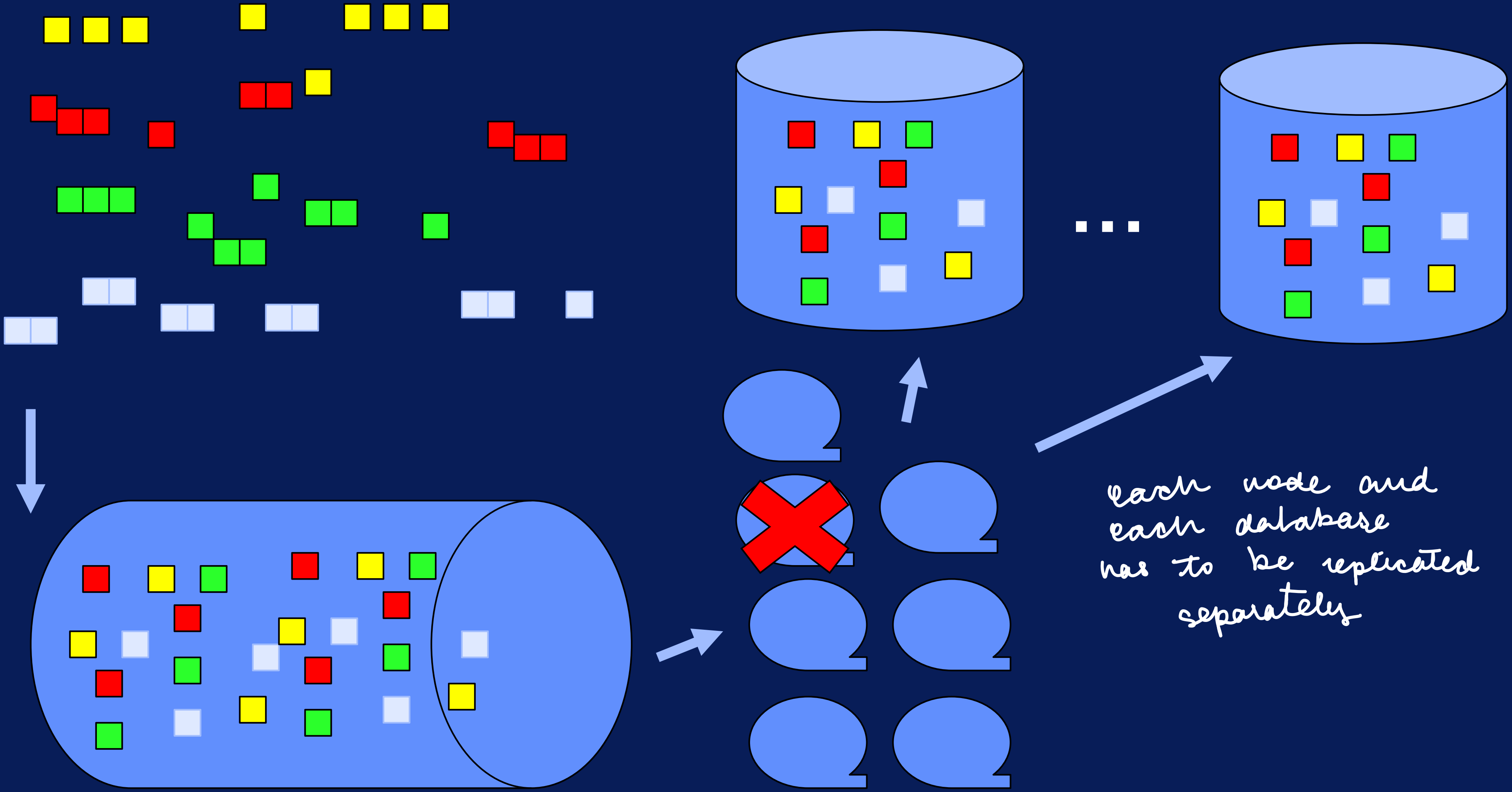
Processing node

buffer or queue to process advancing chunks

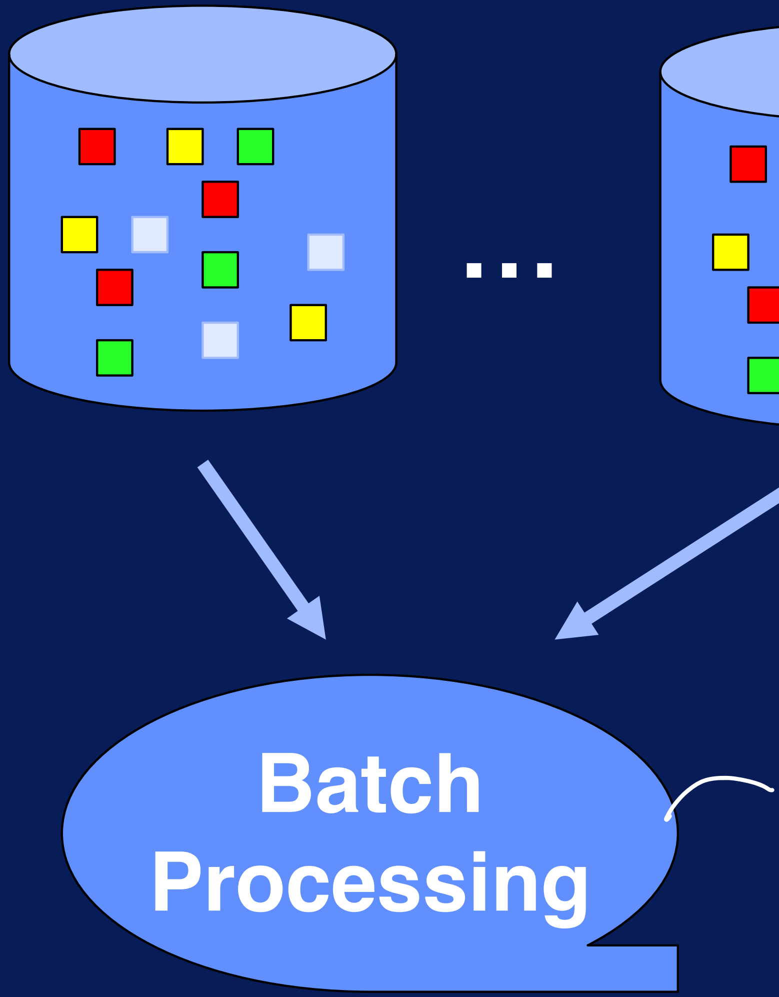


- + reactive fixes
- – robust
- + complicated



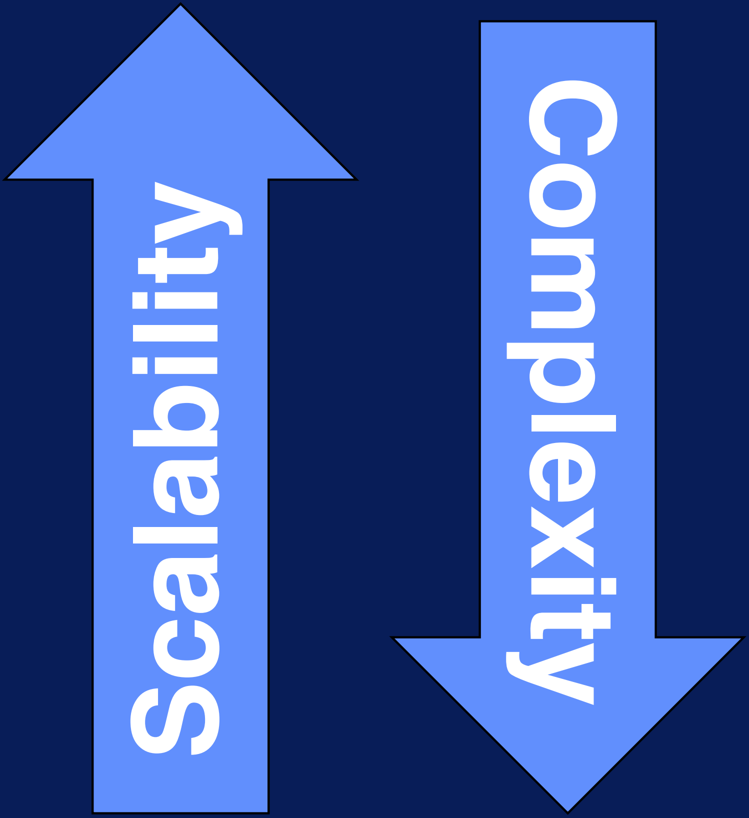


each node and each database has to be replicated separately



need to access
and maintain we
of the data separately

- slow
- costly

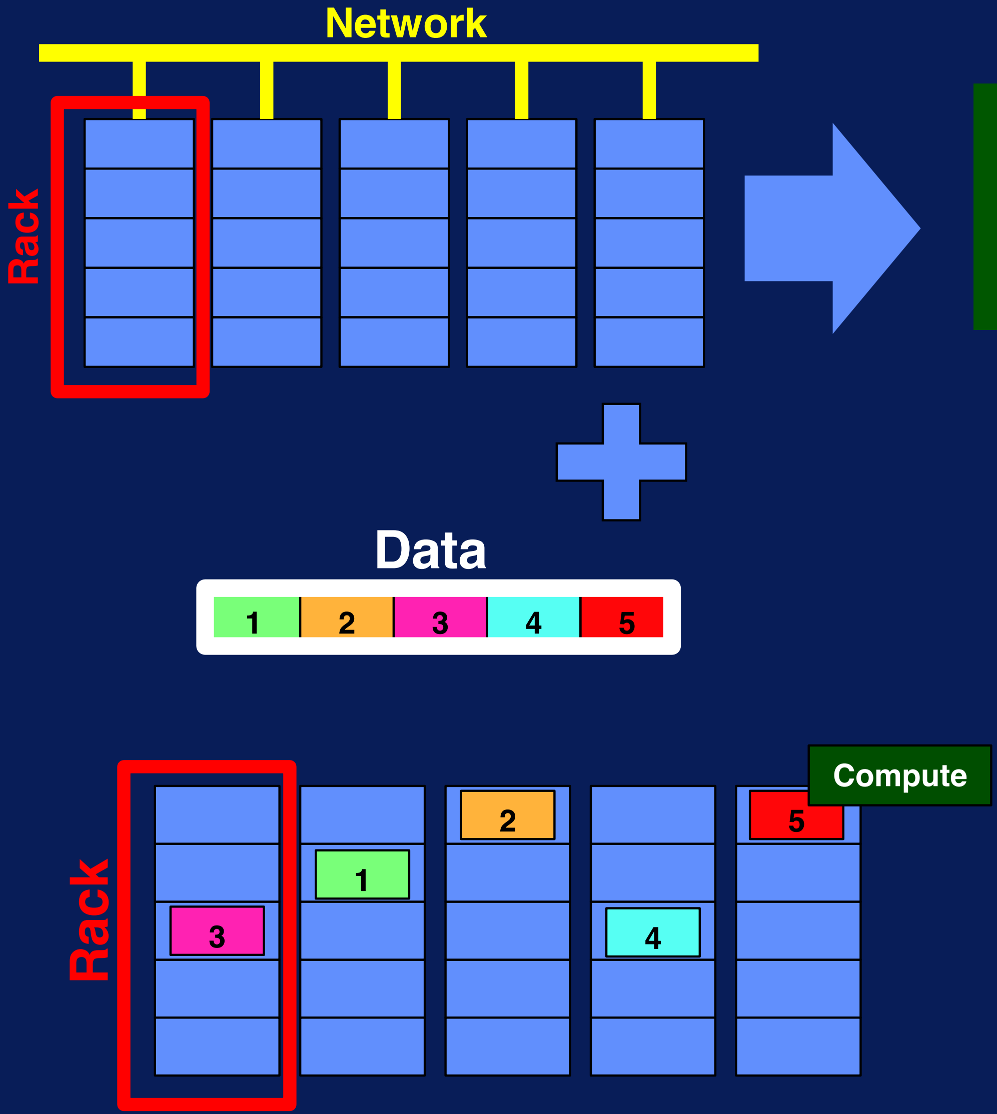


Scalability

Complexity

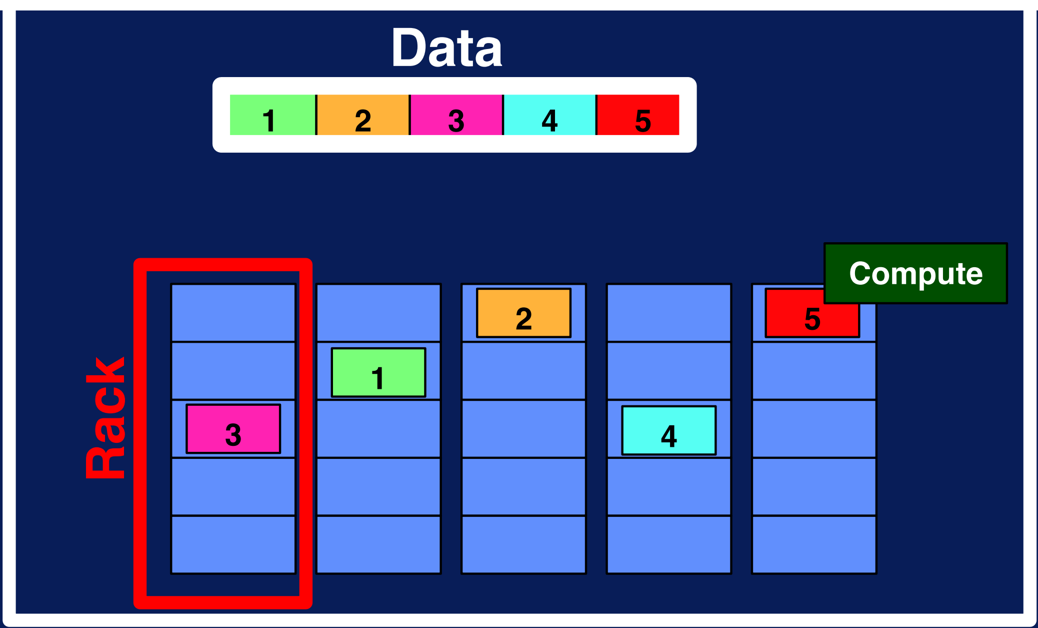
complexities
- failing servers
- breaking compute nodes

Data-parallel scalability



Programming Model = abstractions

Runtime Libraries + Programming Languages



Requirements for Big Data Systems

1. Support Big Data Operations

manage and

Split volumes of data : partitioning and placement of data in and out of computer memory

1. Support Big Data Operations

Split volumes of data

Access data fast

1. Support Big Data Operations

Split volumes of data

Access data fast

Distribute computations to nodes

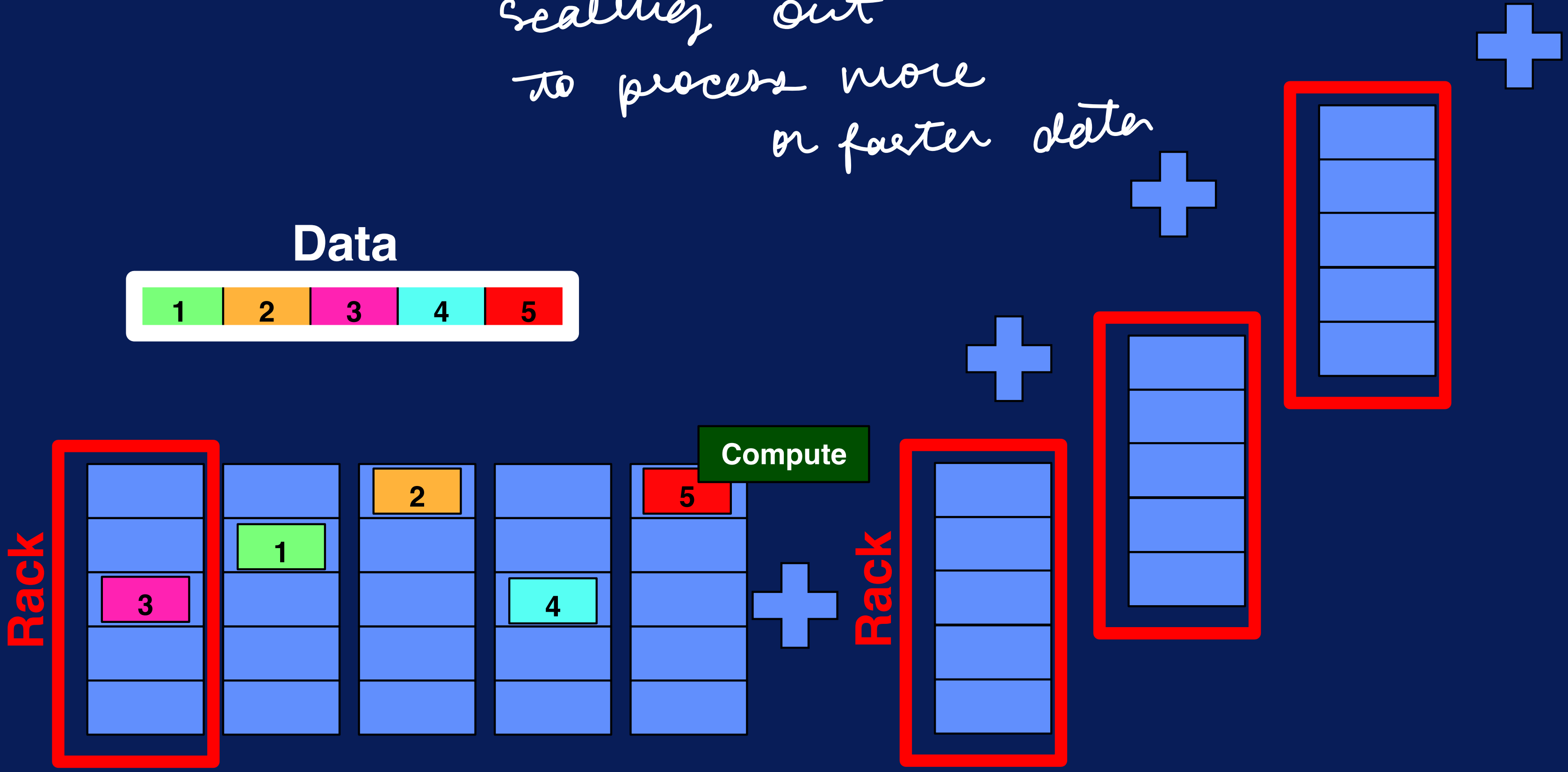
2. Handle Fault Tolerance

Replicate data partitions

Recover files when needed

3. Enable Adding More Racks

scaling out
to process more
or faster data



4. Optimized and extensible for many data types

Document

Table

Key-value

Graph

Multimedia

Stream

enable
processing
of variety of data

optimize
handling each
type
separately
and together

5. Enable both streaming and batch processing

how fast data is being processed

quantification of delay in the processing of the streaming data

Low latency processing of streaming data

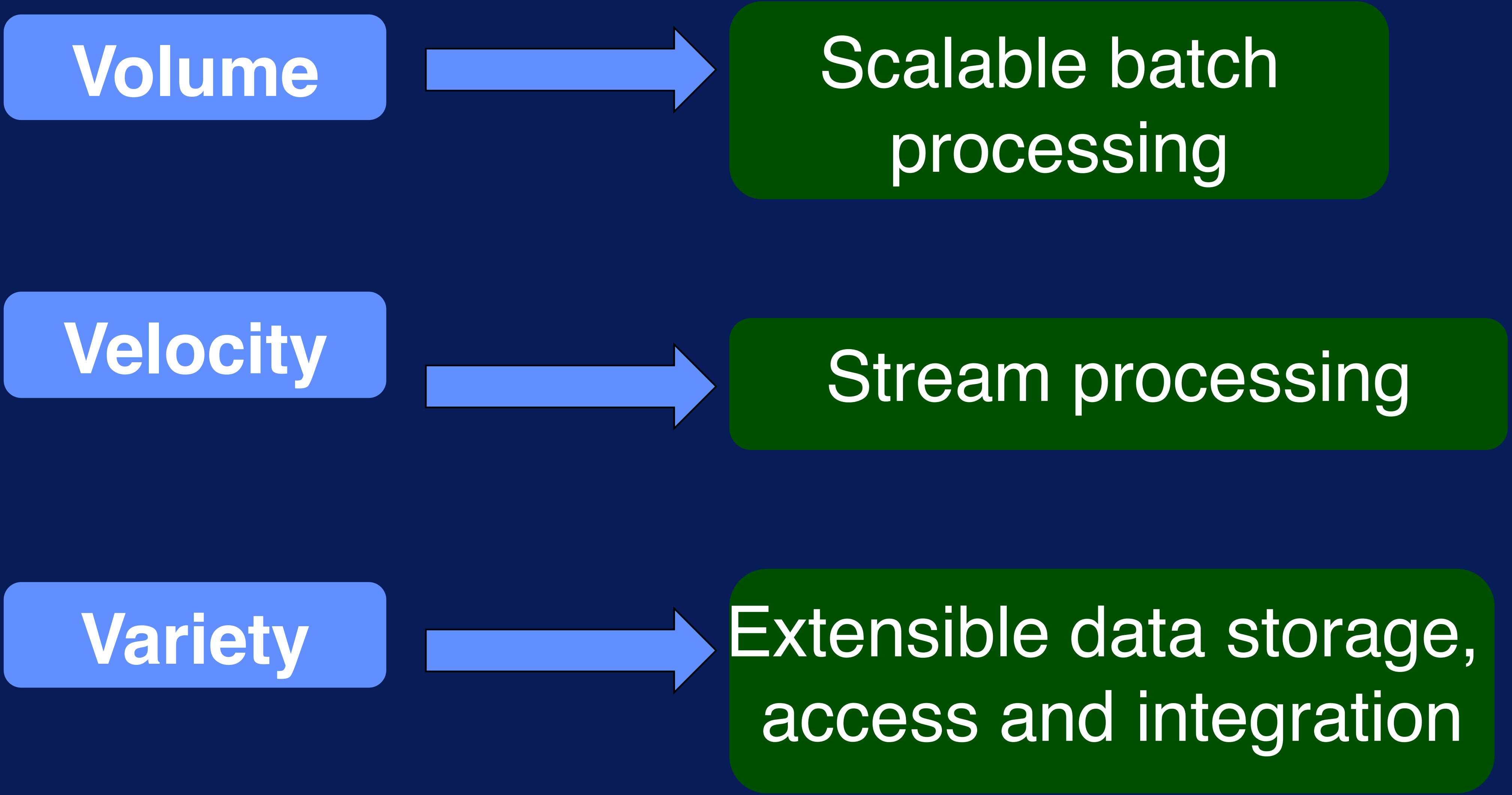
* hadoop is not great for low latency

Accurate processing of all available data

debuggable and extensible

handle operations at small chunks of data streams with minimal delay

* all through same system architecture



Volume

Scalable batch processing

Velocity

Stream processing

Variety

Extensible data storage, access and integration