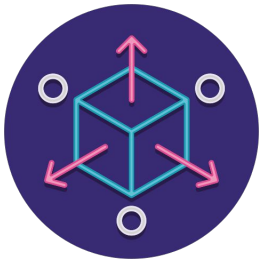
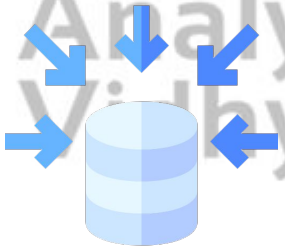
Spark Streaming

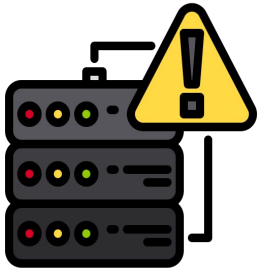
Spark Streaming is an extension of the core Spark API



Scalable



High-Throughput



Fault Tolerance

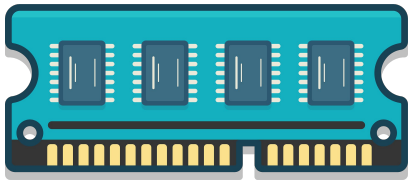
Spark Streaming

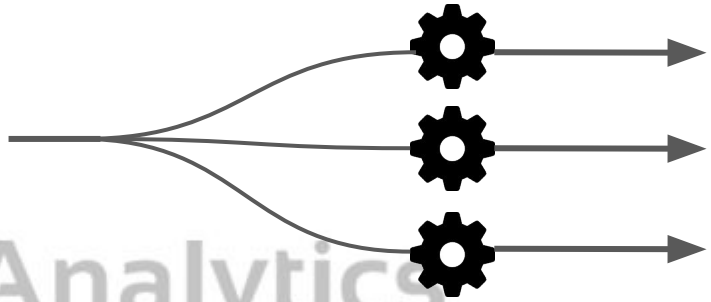
Analytics Vidhya

Spark Streaming

Spark Ecosystem

RAM

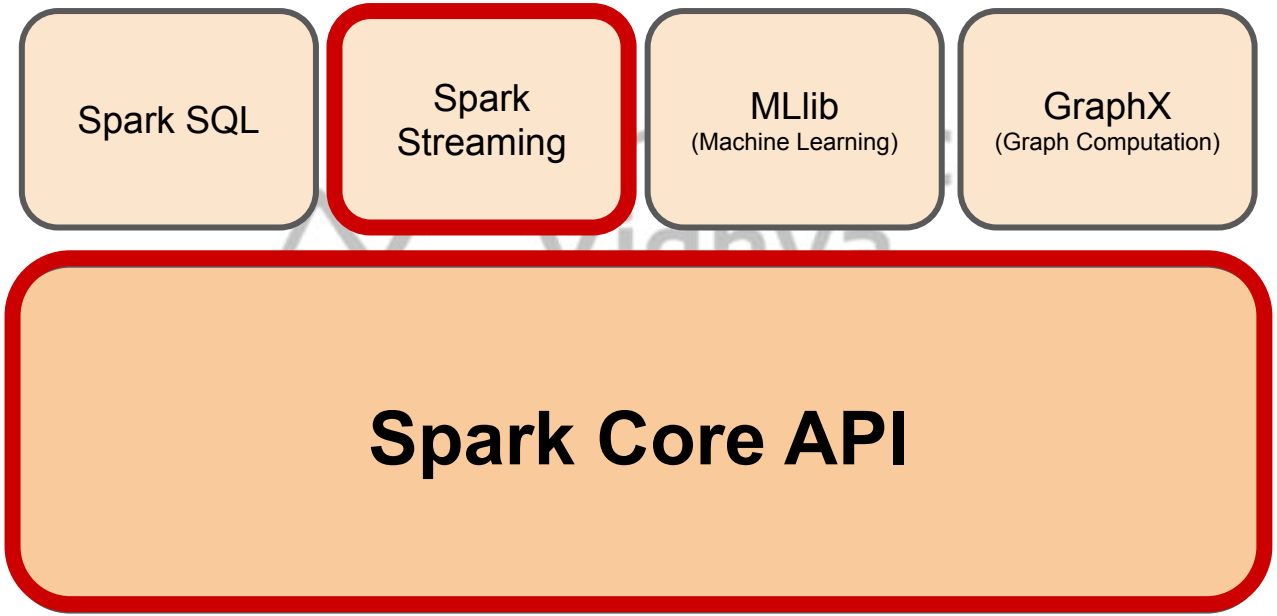




df = df.map(lambda x: x*2)

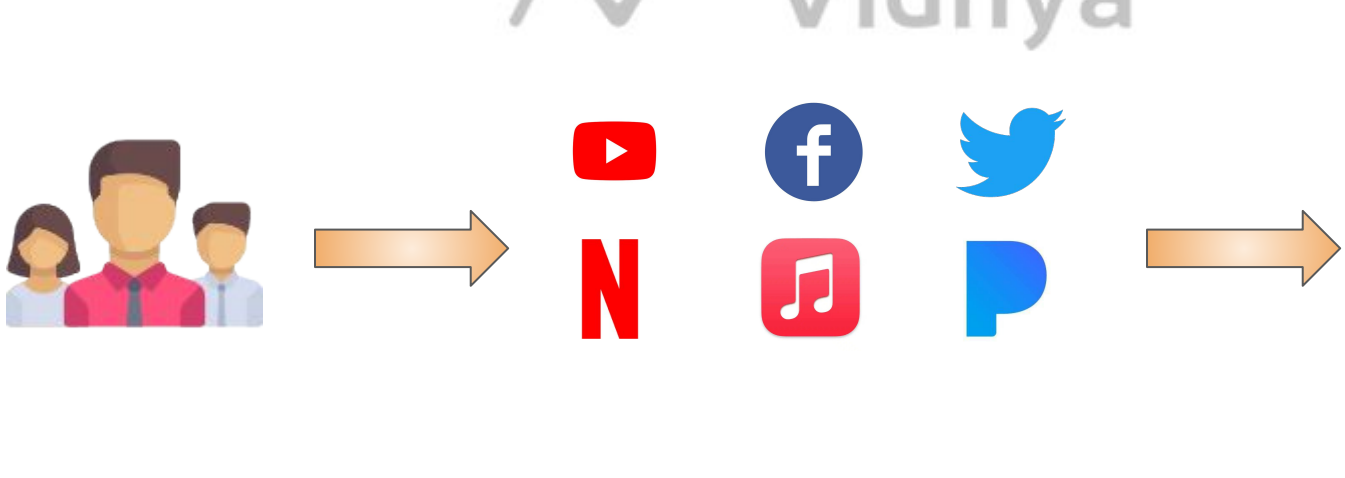
Spark Core API

Spark Ecosystem

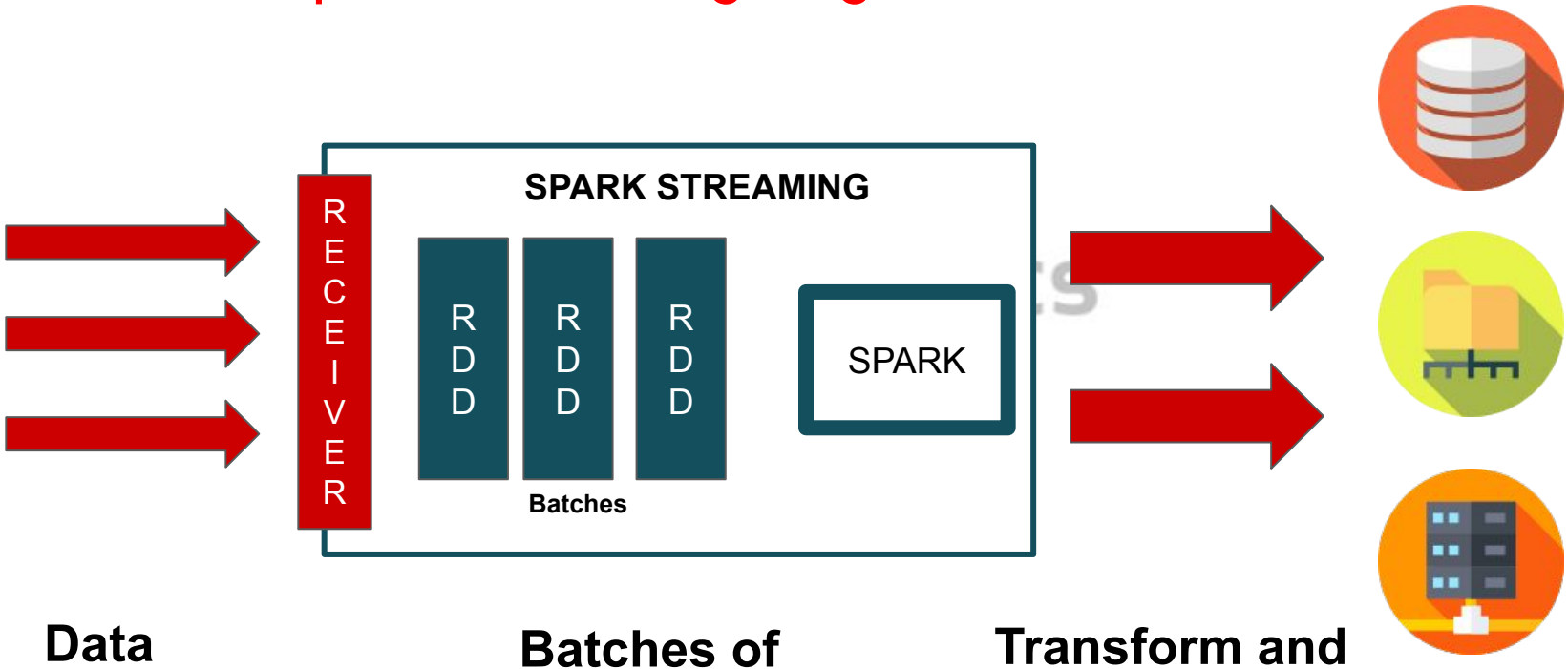


Spark Streaming

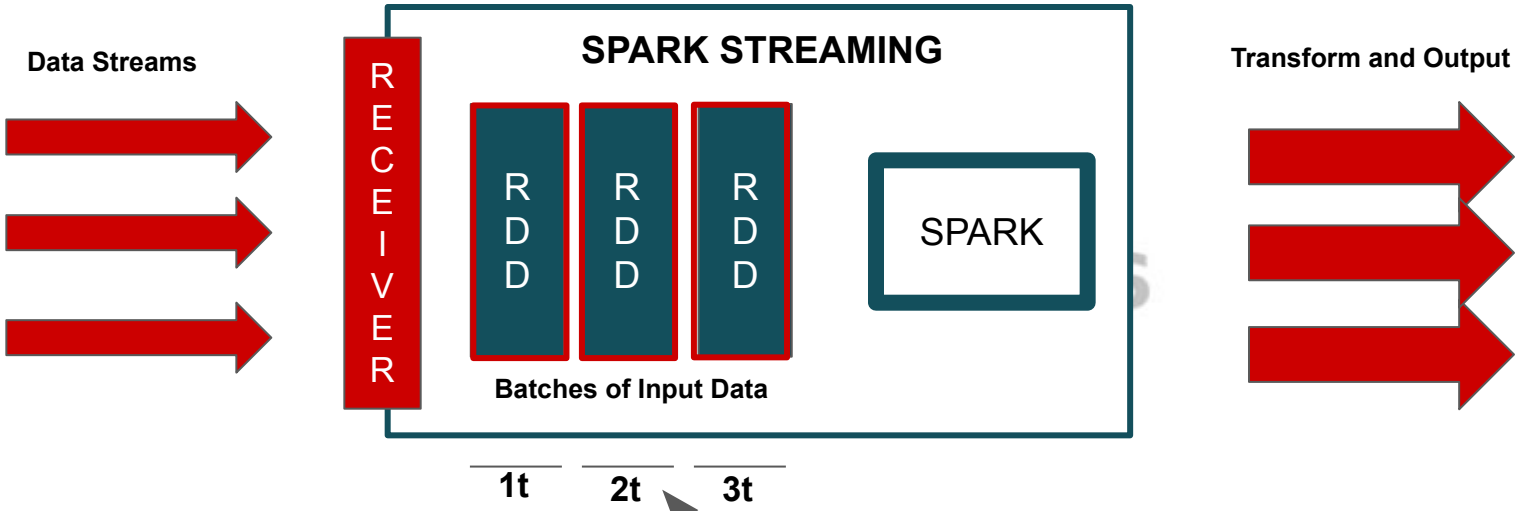
- **Spark Streaming** is an extension of the core **Spark** API that
- Allows data engineers and data scientists to process real-time data
- Can process data from various sources including (but not limited to) Kafka, Flume, and Amazon Kinesis.
- This processed data can be pushed out to file systems, databases, and live dashboards.



Spark Streaming: High Level View



Micro-Batch architecture: Apache Spark



- The stream is treated as a series of batches of data
- New batches are created at regular time intervals
- The size of the time intervals is called the **batch interval**
- The batch interval is typically between 500 ms and several seconds

Spark Streaming

Batch

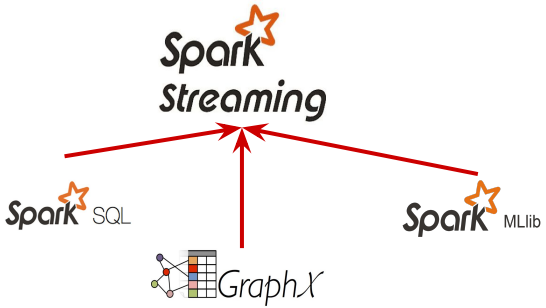
Streaming Data

Spark Streaming - Benefits

Fast Recovery

Load Balancing

Combining Streaming Data with static datasets

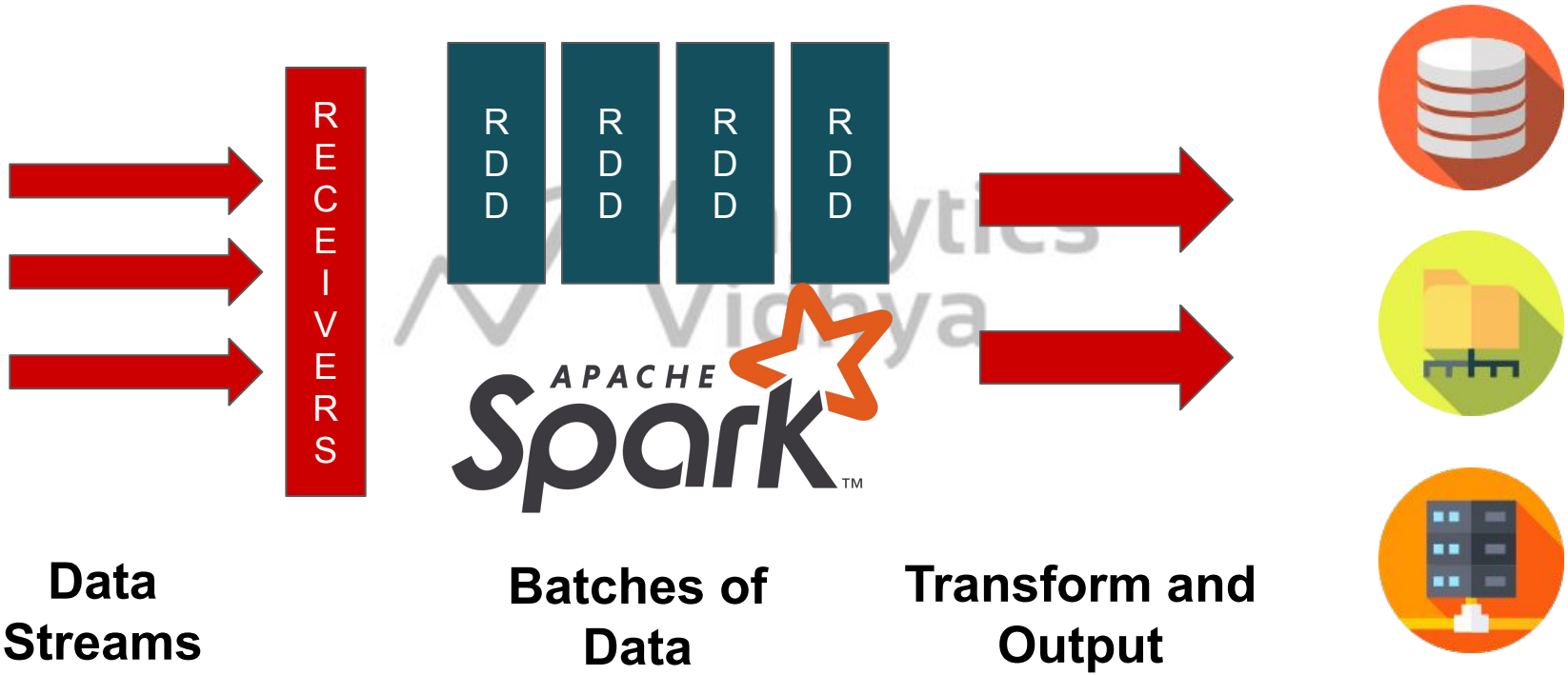


Integration with advanced libraries

Thank You

Analytics
Vidhya

Spark Streaming: High Level View



Micro-Batch architecture: Apache Spark

- The stream is treated as a series of batches of data
- New batches are created at regular time intervals
- The size of the time intervals is called the batch interval
- The batch interval is typically between 500 ms and several seconds