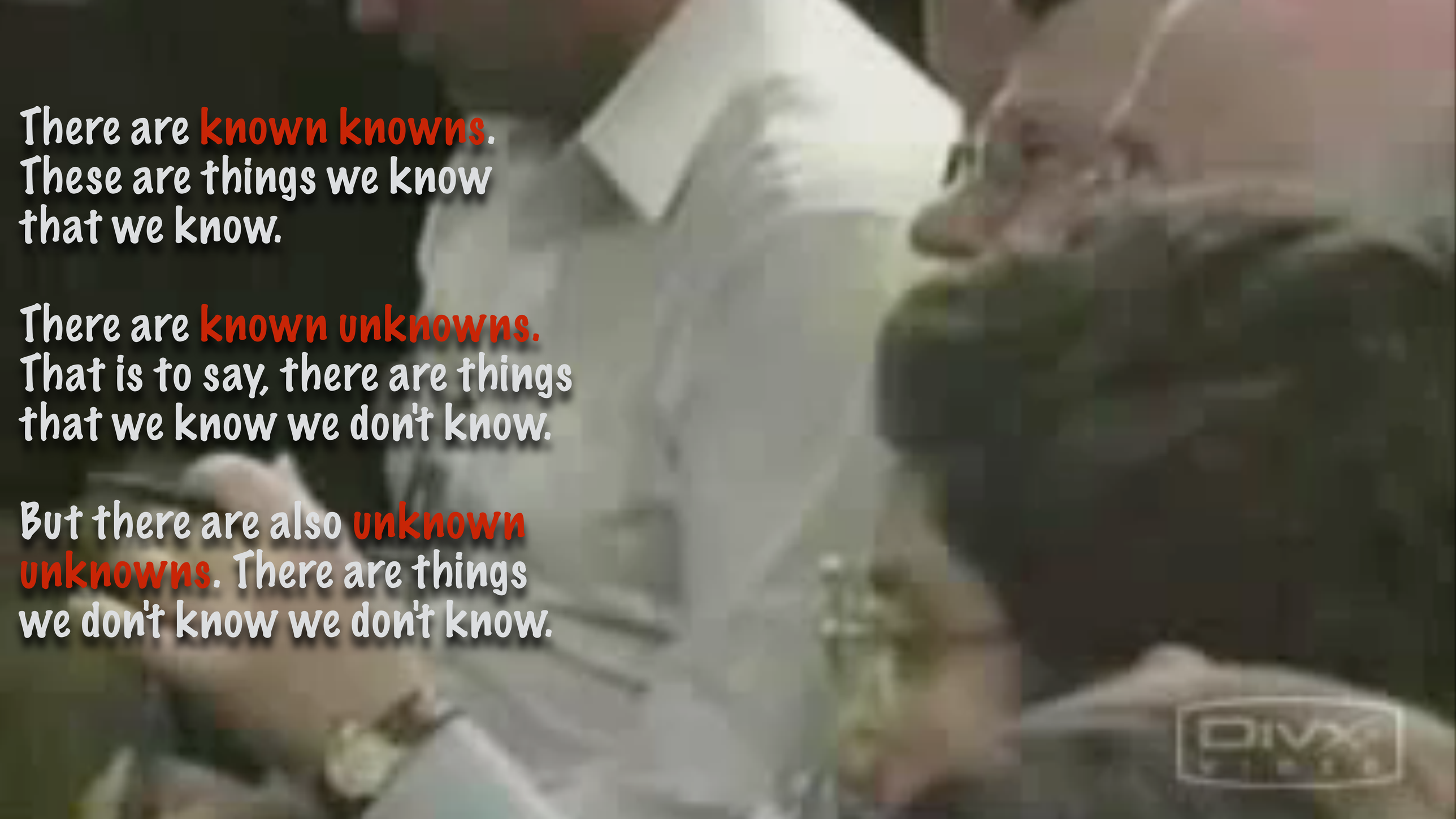
SPARK INTRO

US Secretary of Defense

Donald Rumsfeld

was once asked a tricky question


There are known knowns.
These are things we know
that we know.
There are known unknowns.
That is to say, there are things
that we know we don't know.
But there are also unknown
unknowns. There are things
we don't know we don't know.

There are **known knowns**. These are things we know that we know.

There are **known unknowns**. That is to say, there are things that we know we don't know.

But there are also **unknown unknowns**. There are things we don't know we don't know.

-Donald Rumsfeld

What seemed to be a clever evasion

There are **known knowns**. These are things we know that we know.

There are **known unknowns**. That is to say, there are things that we know we don't know.

But there are also **unknown unknowns**. There are things we don't know we don't know.

-Donald Rumsfeld

..went on to be recognized as a profound truth

Known knowns

Known unknowns

Unknown unknowns

This is a framework that can be applied to many things in life

Known knowns

Known unknowns

Unknown unknowns

In Behavioral psychology,

it represents a famous model about personal awareness

Personal Awareness

Known knowns	You know what you like and dislike
Known unknowns	You know that you don't understand rocket science and Greek
Unknown unknowns	You have an ability or weakness that you are unaware of

Personal Awareness

Known knowns	You know what you like and dislike
Known unknowns	You know that you don't understand rocket science and Greek
Unknown unknowns	You have an ability or weakness that you are unaware of

Things you need to seek feedback about

Known knowns

Known unknowns

Unknown unknowns

In Project Management

it represents a way to classify risks that need to be managed

Known knowns

Known unknowns

Unknown unknowns

In Data Analysis

it represents a way to classify types of insights we can get from the data

Data Analysis

Known knowns

Facts that you know and can confirm

Known unknowns

Business working hours and holidays

Unknown unknowns

Planned marketing events

Available production capacity

Data Analysis

Known knowns

Known unknowns

Unknown unknowns

Questions that need to be answered on a regular basis

How much revenue did we earn yesterday?

How many customers visited our store?

What were our sales by product type?

Data Analysis

Known knowns

Known unknowns

Unknown unknowns

Relationships and drivers that you are unaware of

Searches for vacation destinations are accompanied by searches for diet tips

(Folks want to look good when they go to the beach)

Data Analysis

Known knowns

Known unknowns

Unknown unknowns

Relationships and drivers that you are unaware of

Phones take much longer to be packed and shipped compared to Books

(Packaging instructions for Phones are too complicated)

Data Analysis

Known knowns

Known unknowns

Unknown unknowns

Unknown unknowns in data can only be identified through exploration and investigation

Data Analysis

Known knowns

Known unknowns

Unknown unknowns

But they represent a
huge opportunity

Data Analysis

Known knowns

Known unknowns

Unknown unknowns

They can help

Shape Marketing plans

Build recommendation systems

Ex: Promote healthy recipe cook books during summer vacations

Data Analysis

Known knowns

Known unknowns

Unknown unknowns

They can help

Identify process bottlenecks

Ex: Make the packaging instructions as simple as possible for all product types

Data Analysis

Known knowns

Known unknowns

Unknown unknowns

They can help

Develop intelligent systems

Ex: Spam detection,
Fraud detection

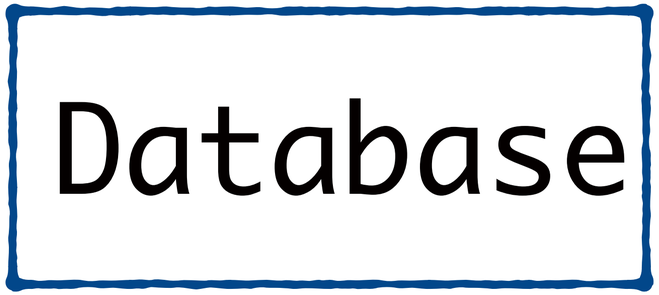
Data Analysis

Unknown unknowns

Let's understand how these are identified and utilized traditionally

Data Analysis (Traditionally)

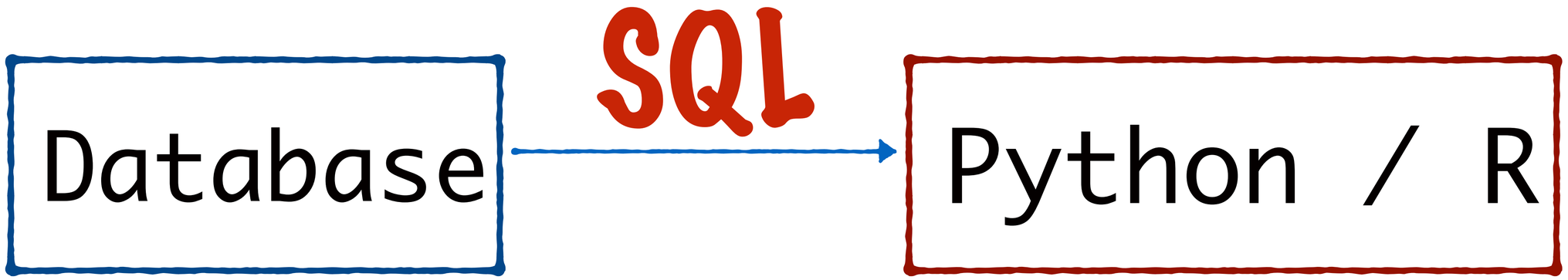
Data usually resides in a database



The data is accessed using

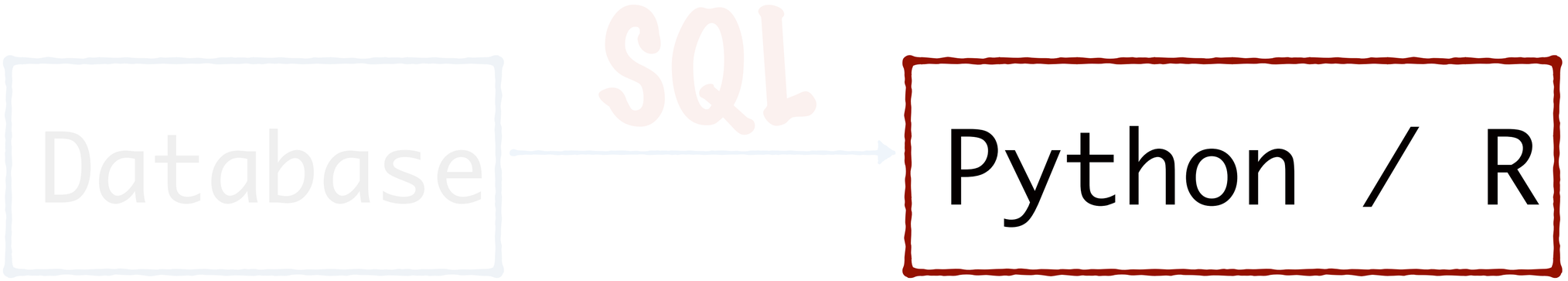
SQL

Data Analysis (Traditionally)



The data is explored in an interactive programmatic environment

Data Analysis (Traditionally)



The interactive environment allows for exploration and iteration

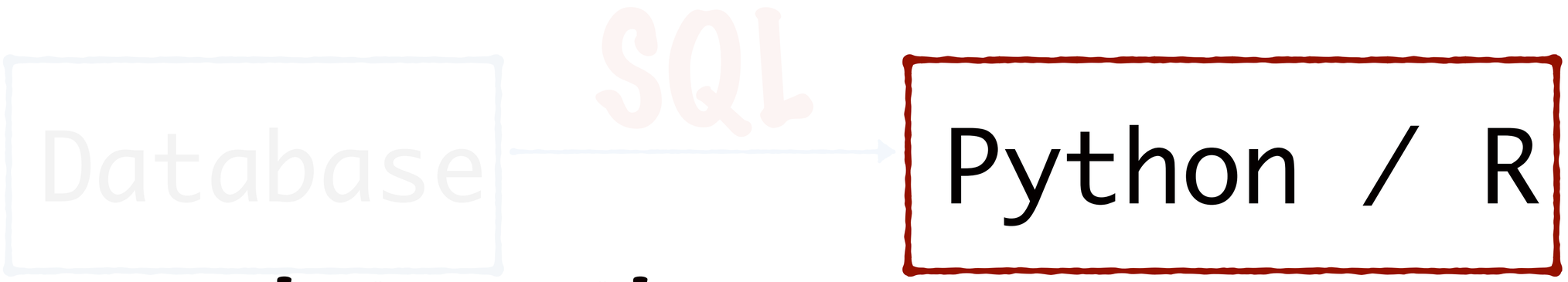
Data Analysis (Traditionally)

Within this environment, you would

→ Python / R

- Represent the data in a form suitable for analysis
- Explore and identify hidden patterns
- Confirm and quantify using Statistical models, machine learning
- **Iterate**

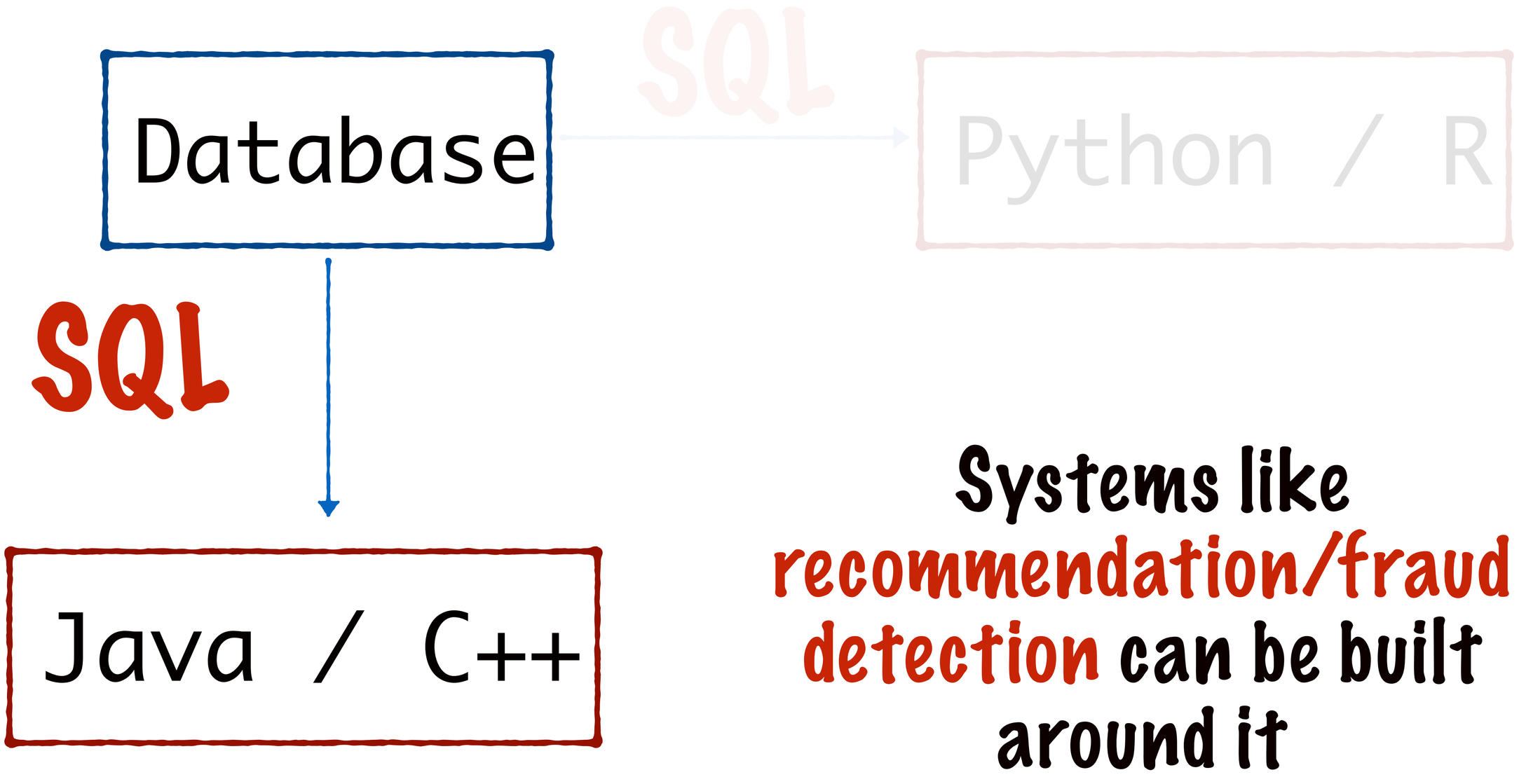
Data Analysis (Traditionally)



Once an interesting and useful insight/ model is found

Systems like recommendation/fraud detection can be built around it

Data Analysis (Traditionally)



Systems like recommendation/fraud detection can be built around it

Data Analysis (Traditionally)

In addition to SQL, we usually require 2 distinct systems for exploring vs operational use

Python / R

Java / C++

Data Analysis (Traditionally)

Python / R

Java / C++

Read-Evaluate-Print Loop

Python and R have **REPL environments** that are useful for **exploration and iteration**

Java and C++ are used to build **stable and performant** systems with **strict SLAs**

Traditionally

Python / R

Java / C++

APACHE SPARK

REPL environments
Exploration and iteration

Stable and performant
Strict SLAs

APACHE SPARK

Spark is a general-purpose engine
for data processing and analysis

APACHE SPARK

Spark is a general-purpose engine
for data processing and analysis

One engine that does the jobs of
SQL, Python/R and Java/C++

APACHE SPARK

Spark was built using Scala

But it provides APIs in
Scala, Python and Java

APACHE SPARK

Spark provides interactive REPL environments for Python and Scala

These are perfect for exploration and iteration

APACHE SPARK

Once you have identified
useful models

Productionize them within
Spark itself!

APACHE SPARK

Spark is a part of the
Hadoop ecosystem

APACHE SPARK

It's engine is capable of
Distributed Computing

i.e processing data stored
across a cluster of machines

APACHE SPARK

Spark was built to overcome some of the limitations of Hadoop's MapReduce engine

Limitations of Hadoop's MapReduce engine

Everything has to be expressed as a chain of map and reduce tasks

APACHE SPARK

Operations can be expressed in a very intuitive way

Limitations of Hadoop's MapReduce engine

No Interactive environment

APACHE SPARK

Interactive shells are available for Python and Scala

Limitations of Hadoop's MapReduce engine

Disk writes occur at the end of each intermediate map-reduce task

APACHE SPARK

Data is kept in-memory and can be passed directly to the next step

Limitations of Hadoop's MapReduce engine

Can only do batch processing

ie. files stored on disk

APACHE SPARK

Can do stream processing

ex: a stream of status messages from a web service

APACHE SPARK

Spark was built to overcome some of the limitations of Hadoop's MapReduce engine

We'll dig deeper into Spark vs MapReduce later

APACHE SPARK

Spark is made up of a few different components

APACHE SPARK

Spark Core

Spark Core contains the basic functionality of Spark

APACHE SPARK

Spark Core

It provides an API called the RDD
(Resilient Distributed Dataset) API

APACHE SPARK

Spark Core

RDDs are the main programming abstraction in Spark

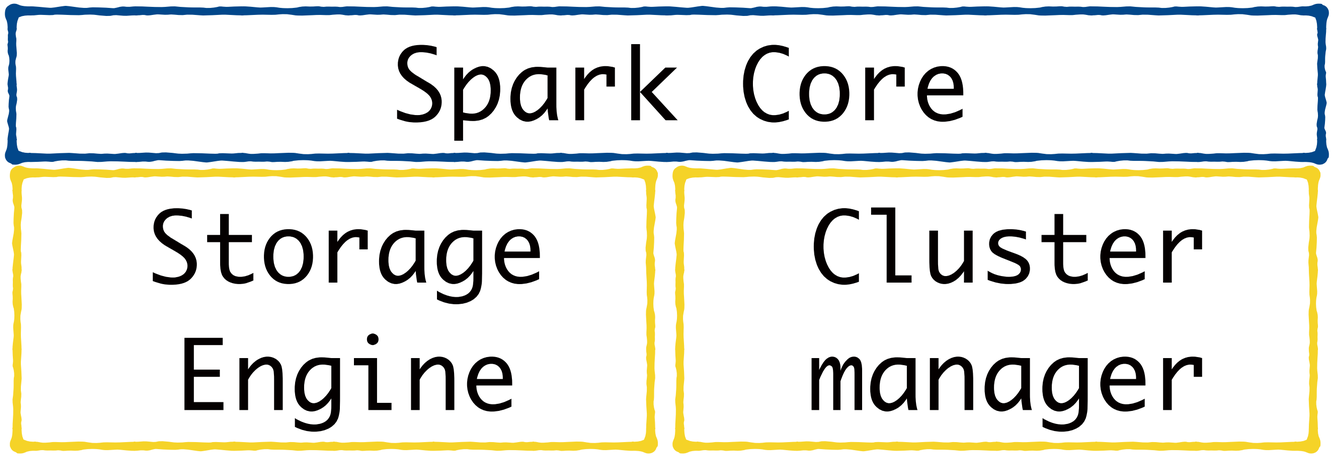
APACHE SPARK

Spark Core

RDDs are in-memory objects and
all data is processed using them

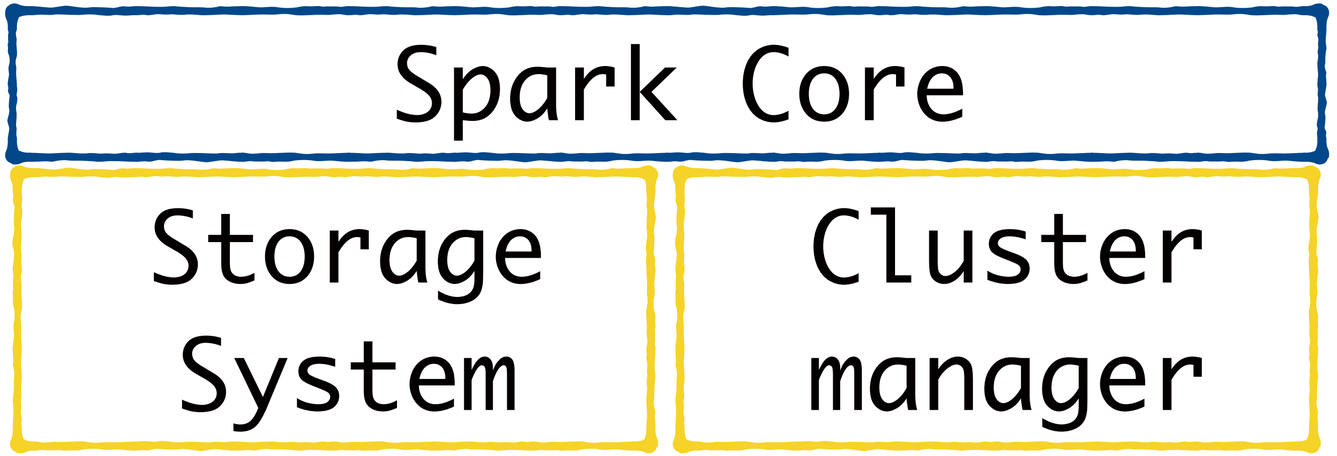
More on this later...

APACHE SPARK



Spark Core needs a Storage system and a Cluster manager to interact with

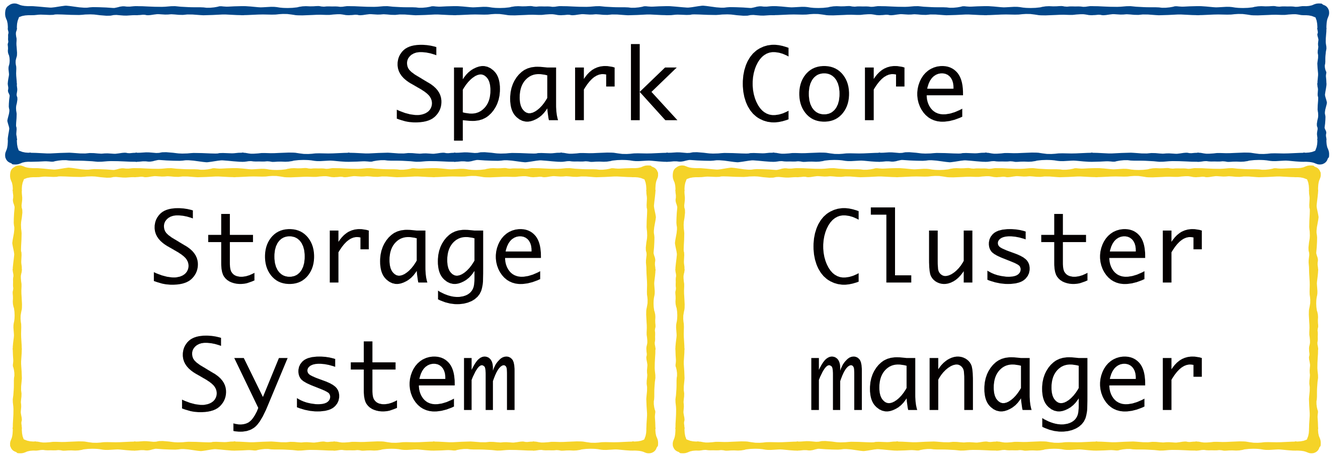
APACHE SPARK



The storage system stores the data to be processed

The cluster manager helps Spark run across a cluster of machines

APACHE SPARK



Both of these are plug and play components

APACHE SPARK

Spark Core

Storage System

Cluster manager

Could be a Local file system, HDFS, Hive, HBase, Cassandra etc

APACHE SPARK

The cluster manager schedules tasks and manage resources across the cluster

APACHE SPARK

Spark comes with a built-in Cluster manager

APACHE SPARK

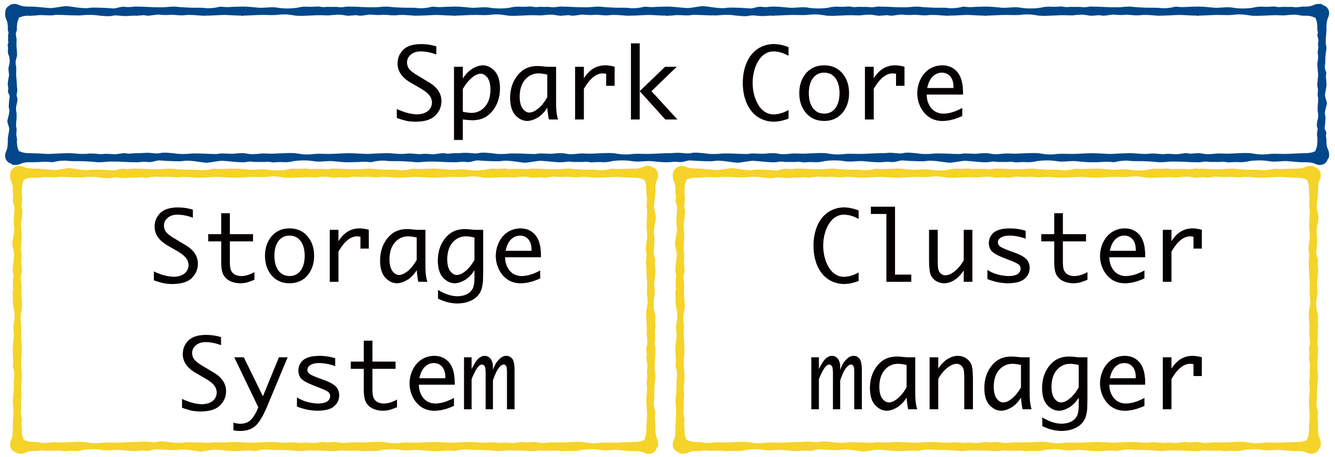
If you already have a Hadoop cluster, you can use Hadoop's YARN as the cluster manager

APACHE SPARK

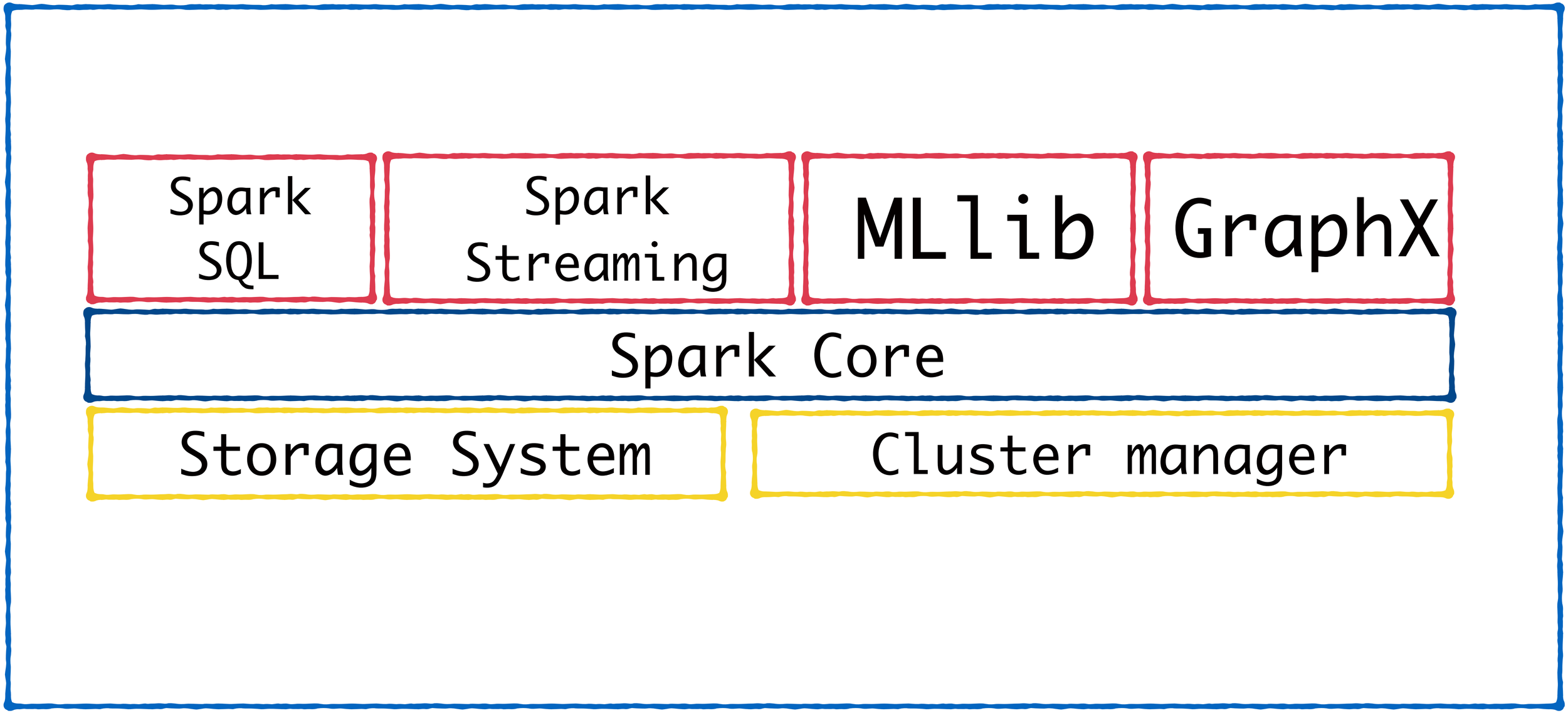
Apache Mesos is also available as an option for the Cluster Manager

APACHE SPARK

Spark comes with some additional packages that make it truly general - purpose



APACHE SPARK



APACHE SPARK

Spark SQL

Spark SQL provides an SQL interface for Spark

APACHE SPARK

Spark SQL

Data is loaded into memory as an RDD

APACHE SPARK

Spark SQL

This RDD can be manipulated using Python, Java, Scala or SQL

APACHE SPARK

Spark SQL

Spark allows programmers to mix SQL manipulations with Python/Java data manipulations within a single program

APACHE SPARK

Spark SQL

Spark Streaming

MLlib

GraphX

Spark Core

Storage System

Let's say you have a live stream of data

APACHE SPARK

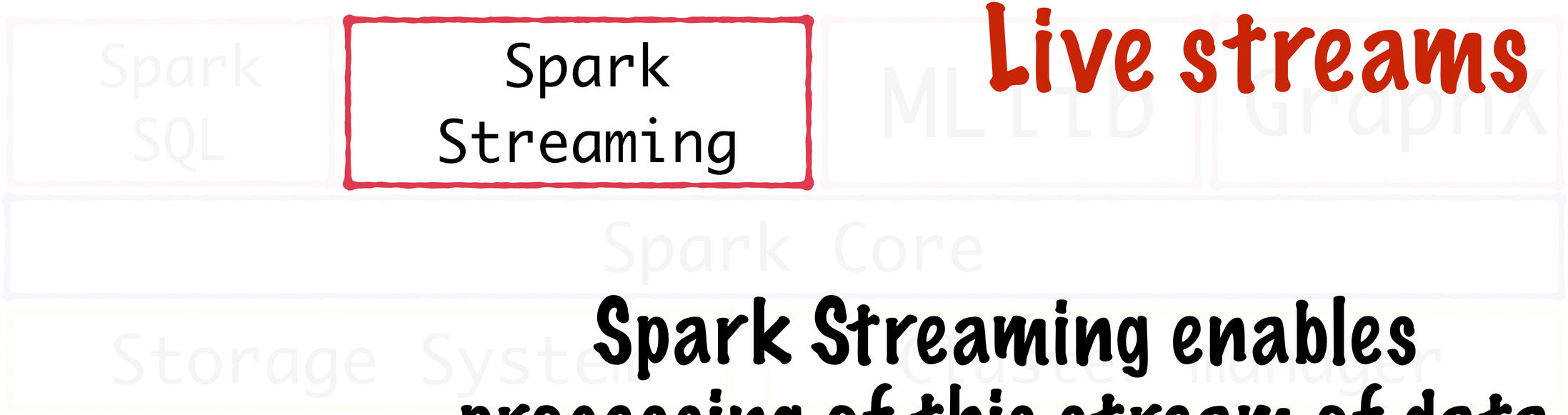
Spark Streaming

Live streams

Logs generated by a web server

Status updates posted by users

APACHE SPARK



Live streams

Spark Streaming enables processing of this stream of data in (near) real time

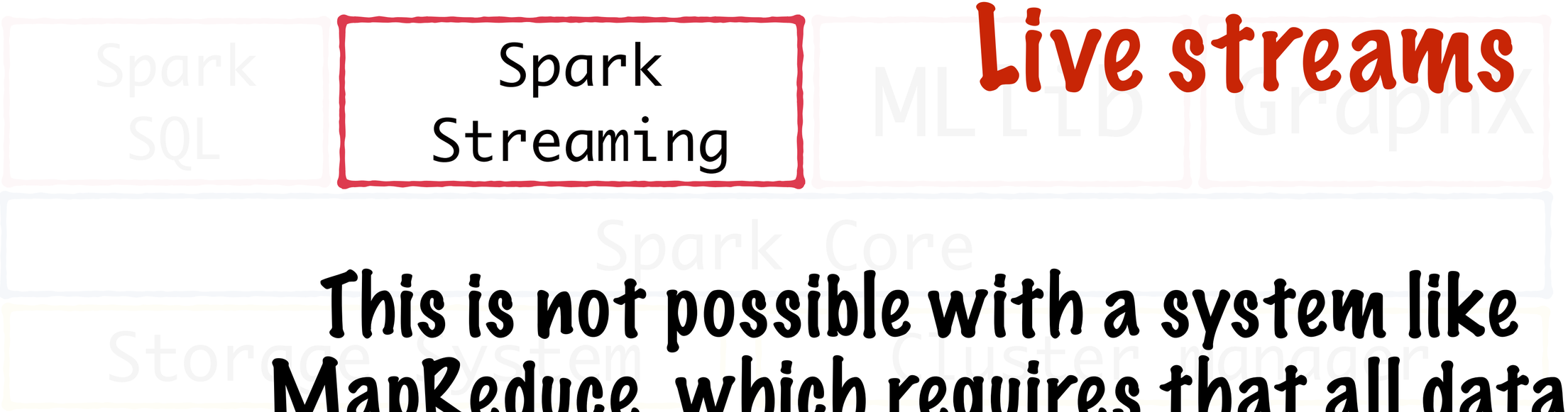
APACHE SPARK

Spark Streaming

Live streams

You can process logs for reporting, monitoring and react to them in real time

APACHE SPARK



This is not possible with a system like MapReduce, which requires that all data be written to disk before it's processed

APACHE SPARK

Spark SQL | Spark Streaming | MLlib | GraphX

Spark Core

Storage System | Cluster Manager

MLlib provides built-in Machine Learning functionality

APACHE SPARK

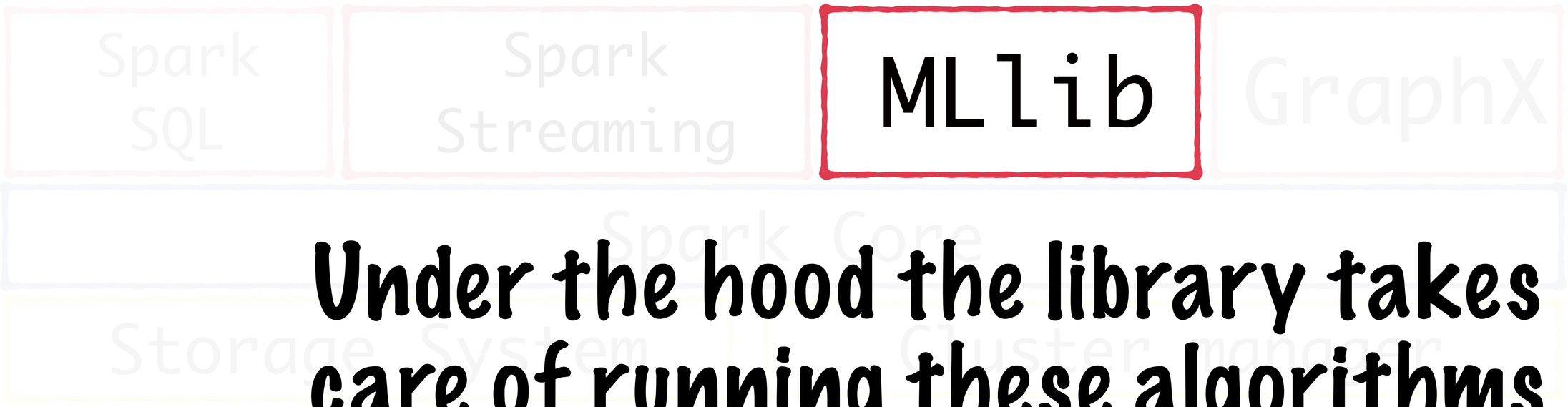
Spark SQL | Spark Streaming | MLlib | GraphX

Spark Core

Storage System | Cluster manager

Built-in methods for Classification, regression, clustering etc algorithms are provided

APACHE SPARK



Under the hood the library takes care of running these algorithms across a cluster

APACHE SPARK

Spark SQL | Spark Streaming | MLlib | GraphX

Spark Core

Storage System | Cluster Manager

GraphX is a library for graph algorithms

APACHE SPARK

Spark SQL | Spark Streaming | MLlib | GraphX

Spark Core

Storage Systems | Cluster Managers

Many interesting datasets can be represented as graphs

APACHE SPARK

Spark SQL | Spark Streaming | MLlib | GraphX

Spark Core

Storage System | Cluster Manager

Social networks, linked webpages etc

APACHE SPARK

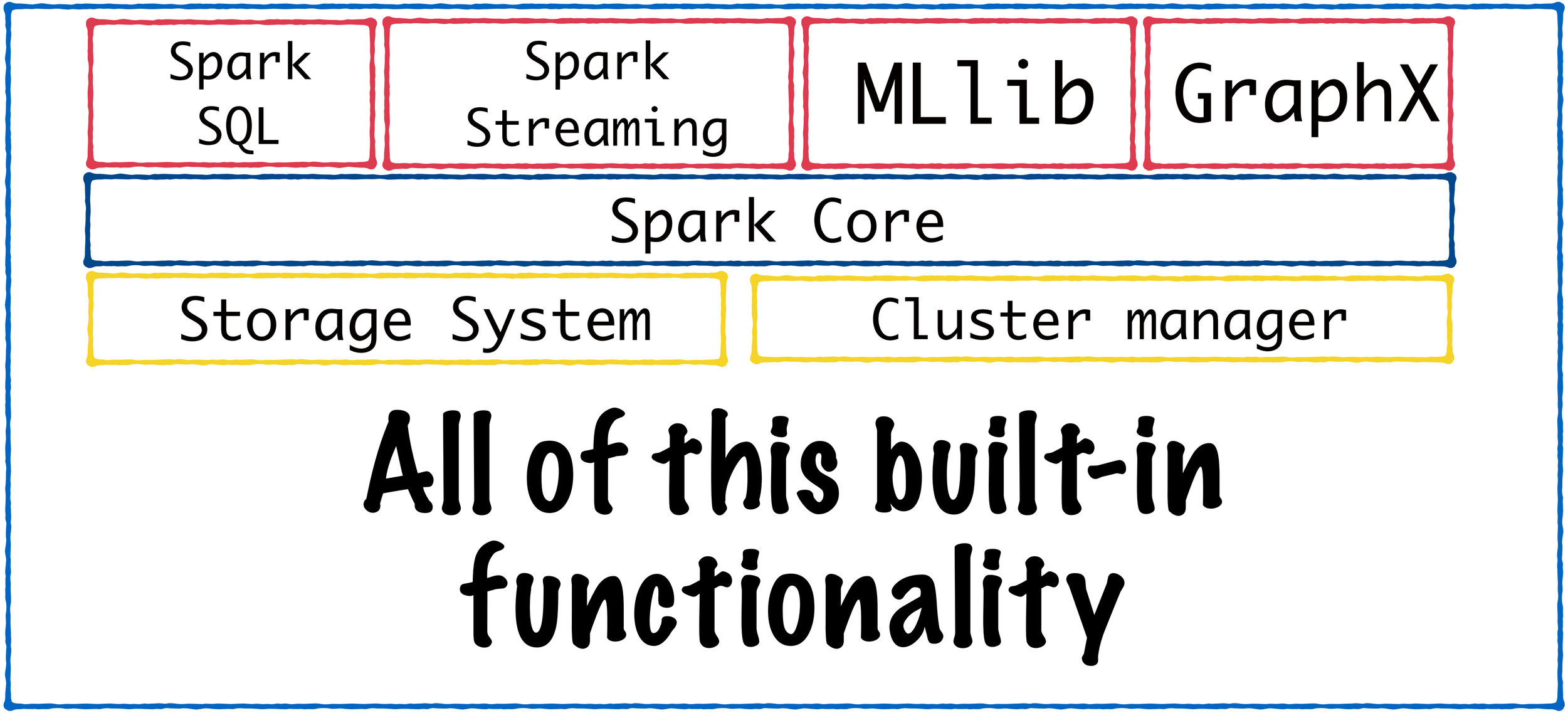
Spark SQL | Spark Streaming | MLlib | GraphX

Spark Core

Storage System | Cluster manager

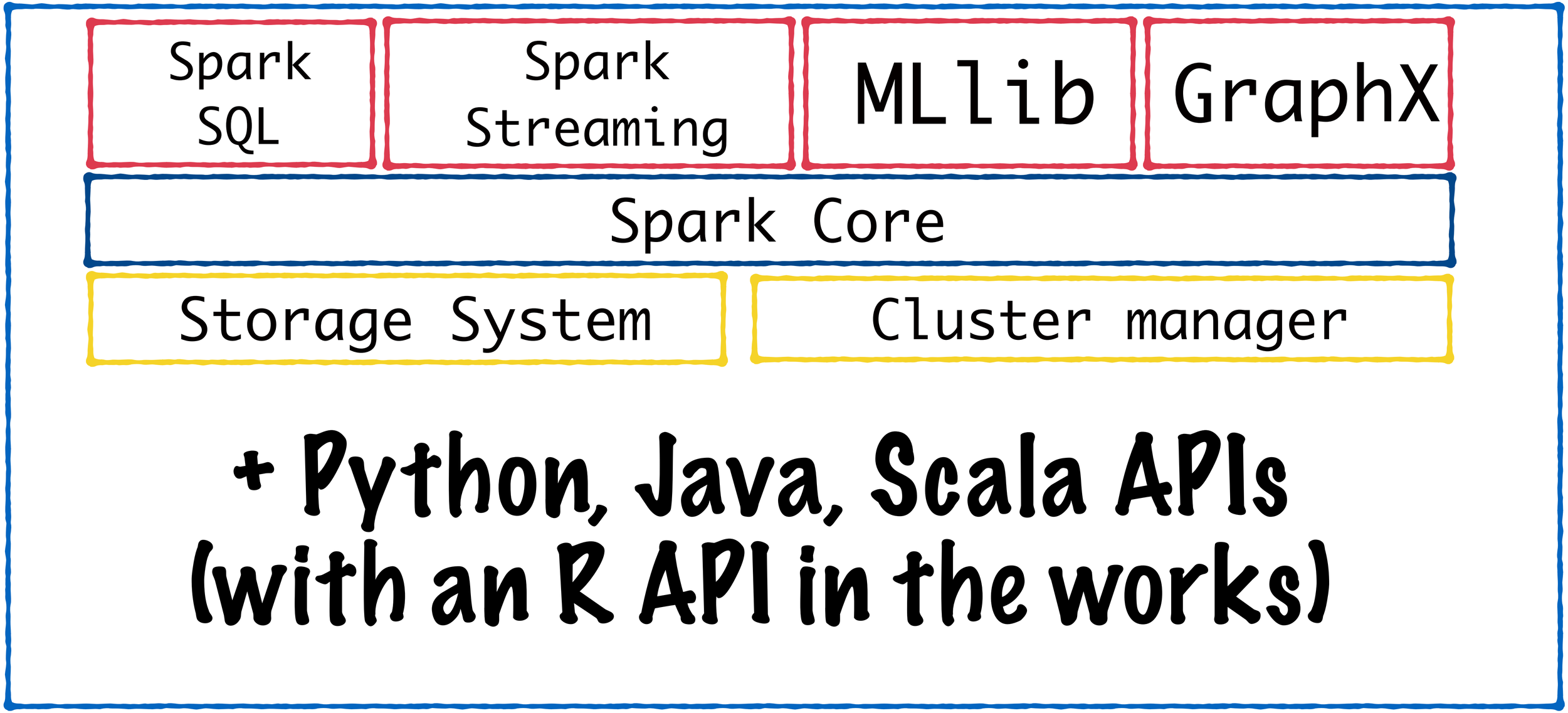
With GraphX you can represent and then perform computations across these datasets

APACHE SPARK



All of this built-in functionality

APACHE SPARK



+ Python, Java, Scala APIs
(with an R API in the works)