

The background of the slide is a close-up photograph of a stone wall. The stones are irregular in shape and size, with a mix of light beige, tan, and reddish-brown hues. The texture is rough and natural, with visible mortar lines between the stones.

Distributed Computing with Apache Spark

What is Spark?

**Spark is
a better implementation
of the
MapReduce paradigm
for Big Data**

The background of the slide is a close-up photograph of a stone wall. The stones are of various sizes and shapes, with colors ranging from light beige to dark brown. The texture is rough and uneven.

Why Spark?

Oh wait...

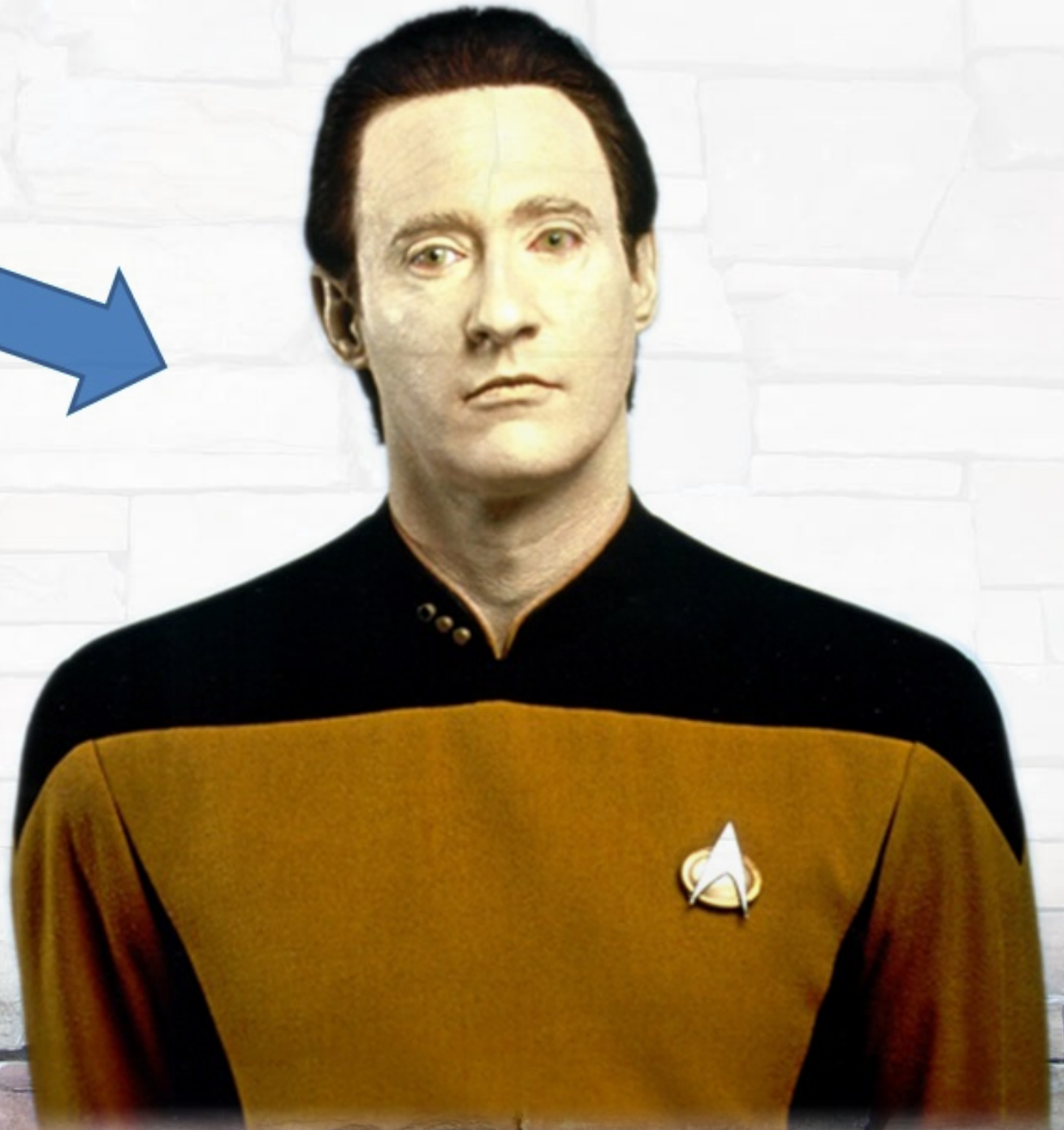
What is Big Data?
What is MapReduce?




What do people mean when they say the words "Big Data" ???

(Bad joke coming up on the next slide.)

BIG
DATA



The background of the image is a close-up of a stone wall. The stones are of various sizes and shapes, with colors ranging from light beige to dark brown. A large, white rectangular area is superimposed on the wall, centered horizontally and vertically. Inside this white area, the text "Big Data is data that exceeds the processing capacity of conventional database systems" is written in a bold, black, sans-serif font. The text is arranged in five lines, with "Big Data" on the first line, "is data that exceeds" on the second, "the processing capacity" on the third, "of conventional" on the fourth, and "database systems" on the fifth.

Big Data
is data that exceeds
the processing capacity
of conventional
database systems

The background of the slide is a close-up photograph of a stone wall. The stones are irregular in shape and size, with a mix of light beige, tan, and reddish-brown hues. The texture is rough and natural.

Big Data is defined by the three Vs


- **volume**
- **velocity**
- **variety**



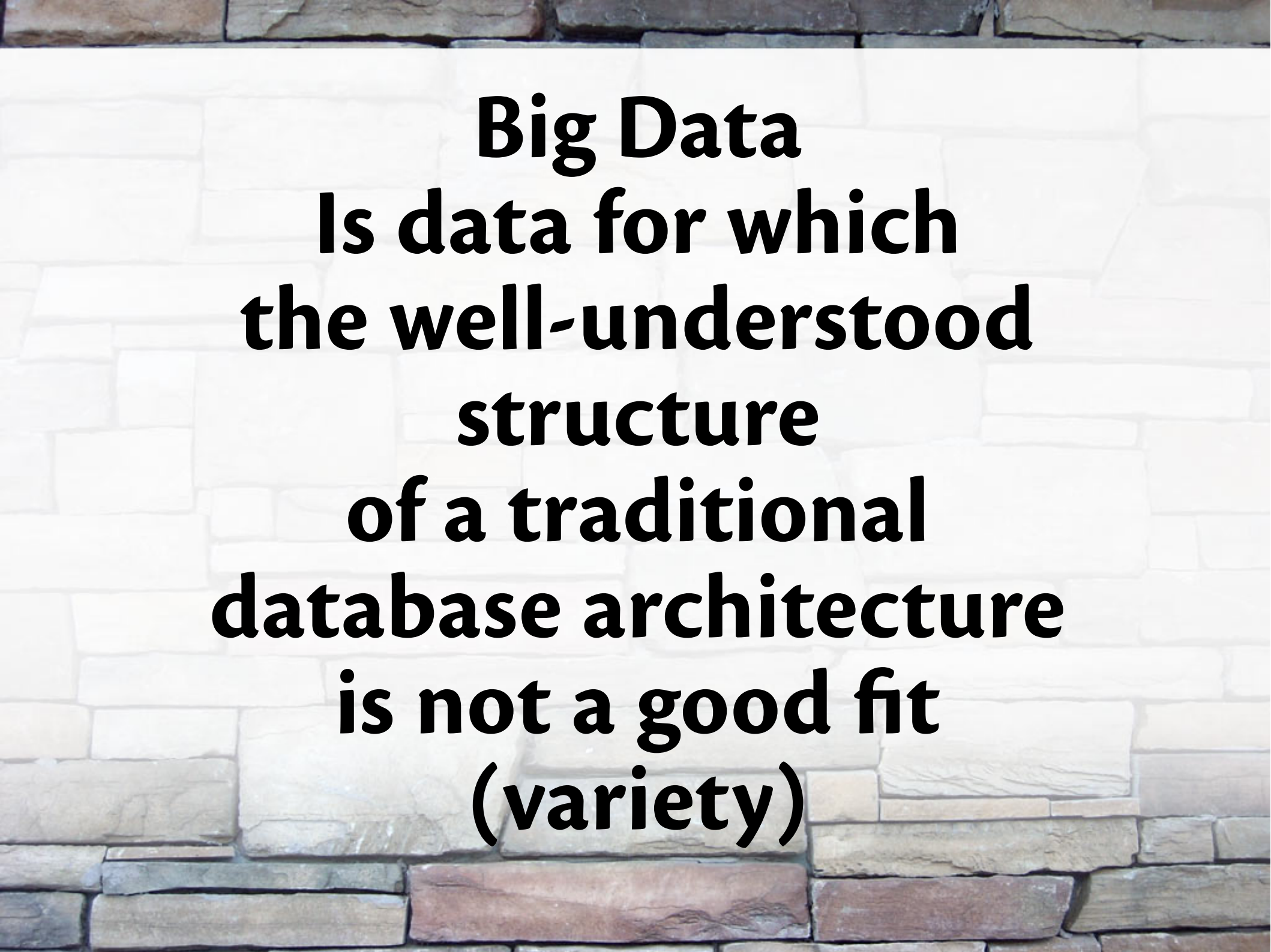
Big Data

There is overwhelmingly too much of it


**To the point where the solution
you developed for small datasets
performs very poorly
(volume)**



**Big Data
Comes to you in the form of
an unrelenting stream at
firehose speeds
(velocity)**



Big Data
Is data for which
the well-understood
structure
of a traditional
database architecture
is not a good fit
(variety)

The background of the image is a close-up of a stone wall. The stones are of various sizes and colors, including shades of tan, beige, and reddish-brown. The wall has a rough, textured appearance. In the center of the image, there is a white rectangular area that serves as a backdrop for the text.

**Therefore,
to gain value from Big Data,
you must choose
an alternative way
to process it**

A background image of a stone wall. The top portion of the wall is composed of light-colored, rectangular stones. The bottom portion features a mix of darker, reddish-brown and grey stones, some of which are rectangular while others are more irregular and layered.

**Year 2003:
MapReduce to the rescue!**

Data analytics at scale was enabled by two big ideas:

**a framework for
distributed storage (HDFS)**

**a framework for
distributed computing (MapReduce)**

The background of the slide is a close-up photograph of a stone wall. The wall is composed of irregularly shaped stones in various shades of beige, tan, and light brown. The stones are laid in a traditional pattern, with some larger flat stones and some smaller, more angular pieces. The lighting is even, highlighting the textures and colors of the stone.

Knowledge vs. Insight

**Knowledge comes by taking
things apart
(map)**

Knowledge vs. Insight

**Insight comes
by inventing ways
to put those pieces
back together
(reduce)**

**MapReduce
should have been called
ExplodeRecombine
(because that's what it is)**

The background of the image is a close-up of a stone wall. The stones are of various sizes and shades of tan, beige, and light brown. A large, vertical white rectangle is centered on the wall, serving as a backdrop for the text. The text is written in a bold, black, sans-serif font and is arranged in five lines, centered within the white rectangle.

**Until recently,
Hadoop was
the standard tool
for distributed computing
across really large data sets**

Hadoop

is an operating system for Big Data

**is a rich ecosystem of
tools and techniques**

**allows engineers to build clusters
from commodity hardware
and do computing
at supercomputer scale**

**Big Data processing
went from
prohibitively expensive
in the pioneering days,
to feasible for even
the smallest garage startups,
who can cheaply rent server
time in the cloud**

A background image of a stone wall with light-colored rectangular stones in the center and darker, more irregular stones at the top and bottom.

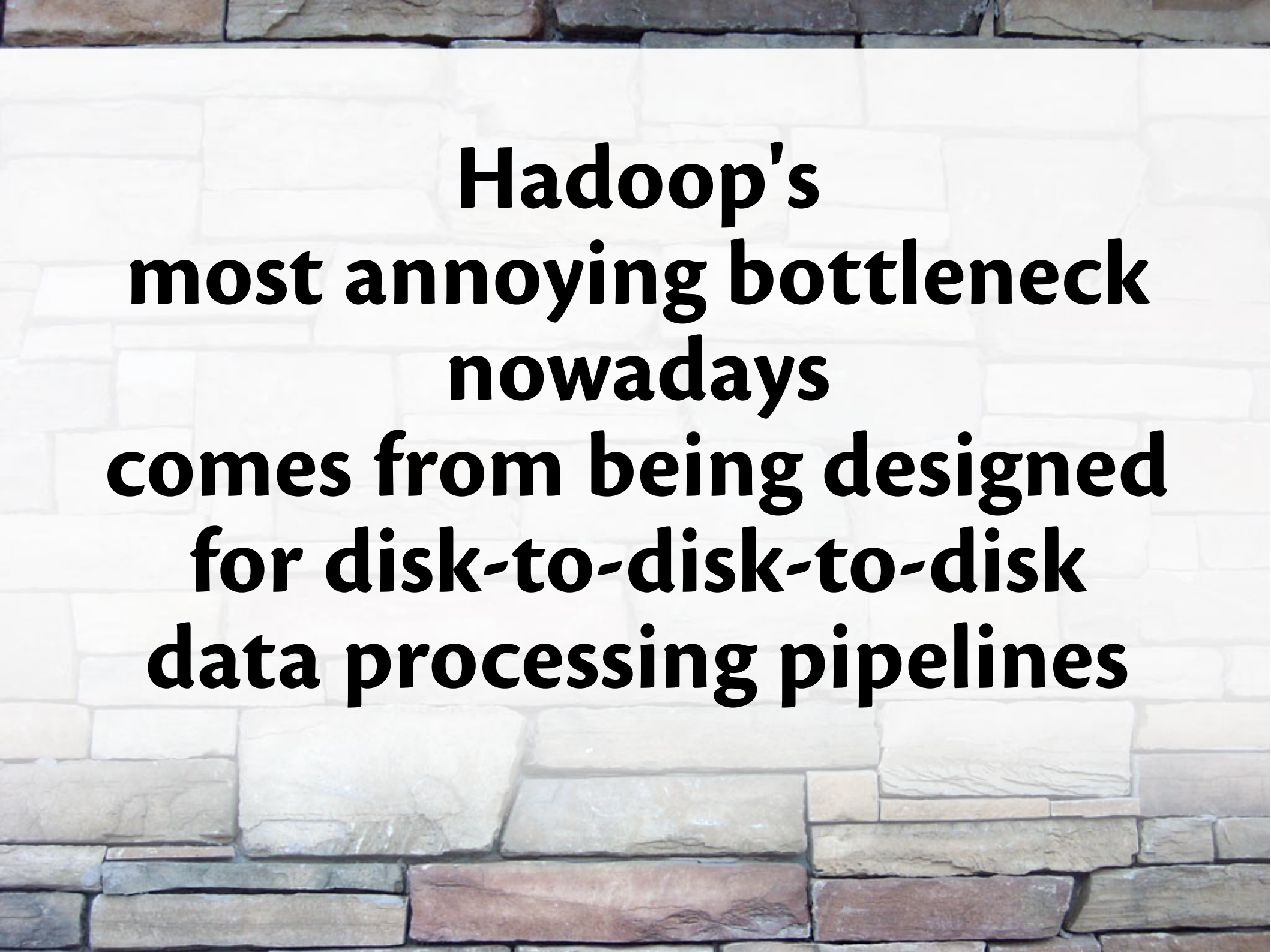
**Hadoop
became successful
when spinning media storage
became cheap enough**

**Powerful force #1:
research has shifted focus
toward generalizations
of distributed computation,
expanding on the ideas
first imagined
in MapReduce**

The background of the slide is a close-up photograph of a stone wall. The stones are light-colored, possibly limestone or sandstone, with a rough, irregular texture. They are arranged in a pattern that looks like a traditional stone masonry. The lighting is even, highlighting the natural grain and color variations of the stones.

Powerful force #2:

RAM
has also become
cheap enough



**Hadoop's
most annoying bottleneck
nowadays
comes from being designed
for disk-to-disk-to-disk
data processing pipelines**

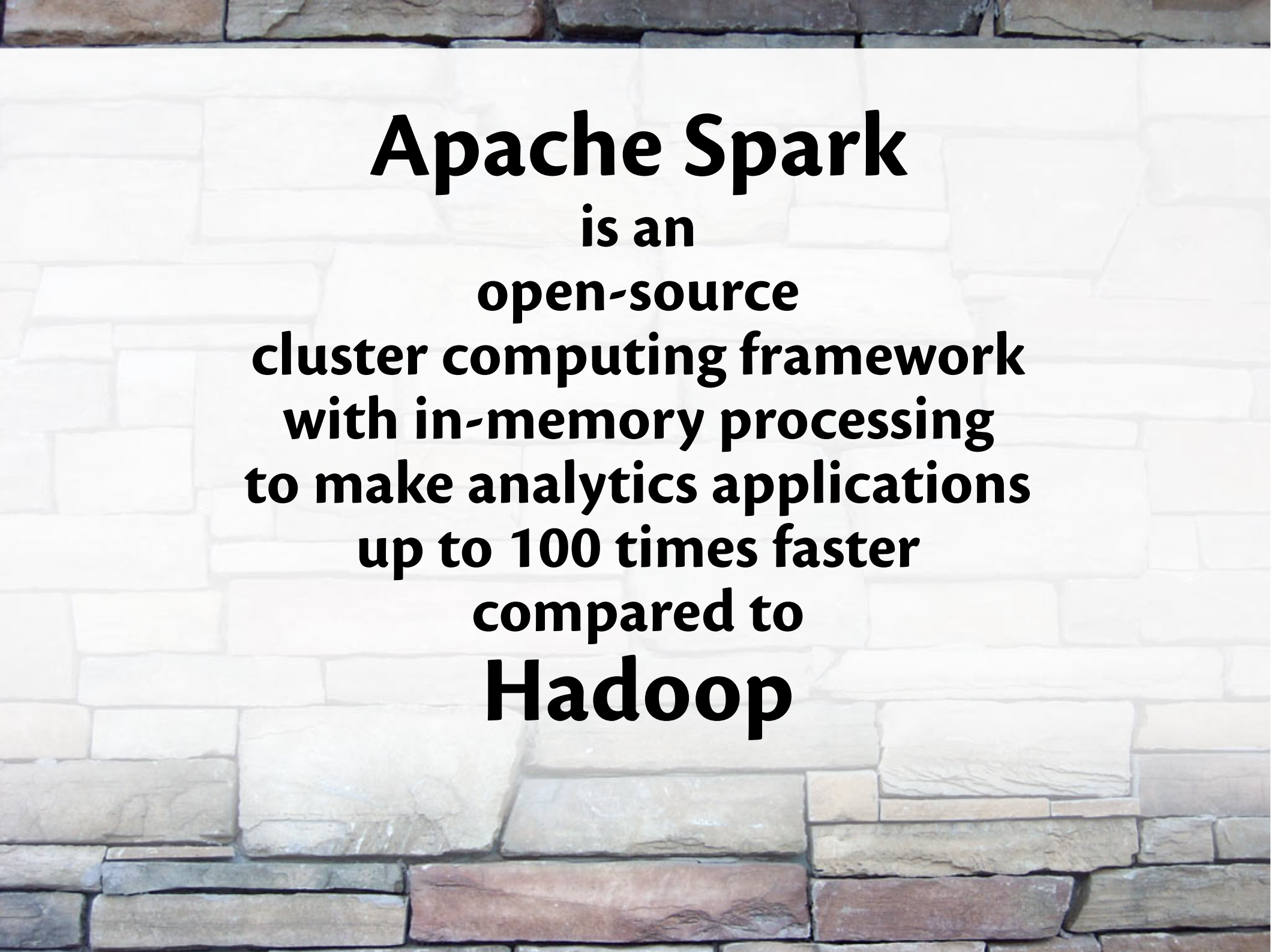
The background of the slide is a close-up photograph of a stone wall. The stones are of various sizes and shapes, with colors ranging from light beige to dark brown. The texture is rough and natural.

Obvious idea:

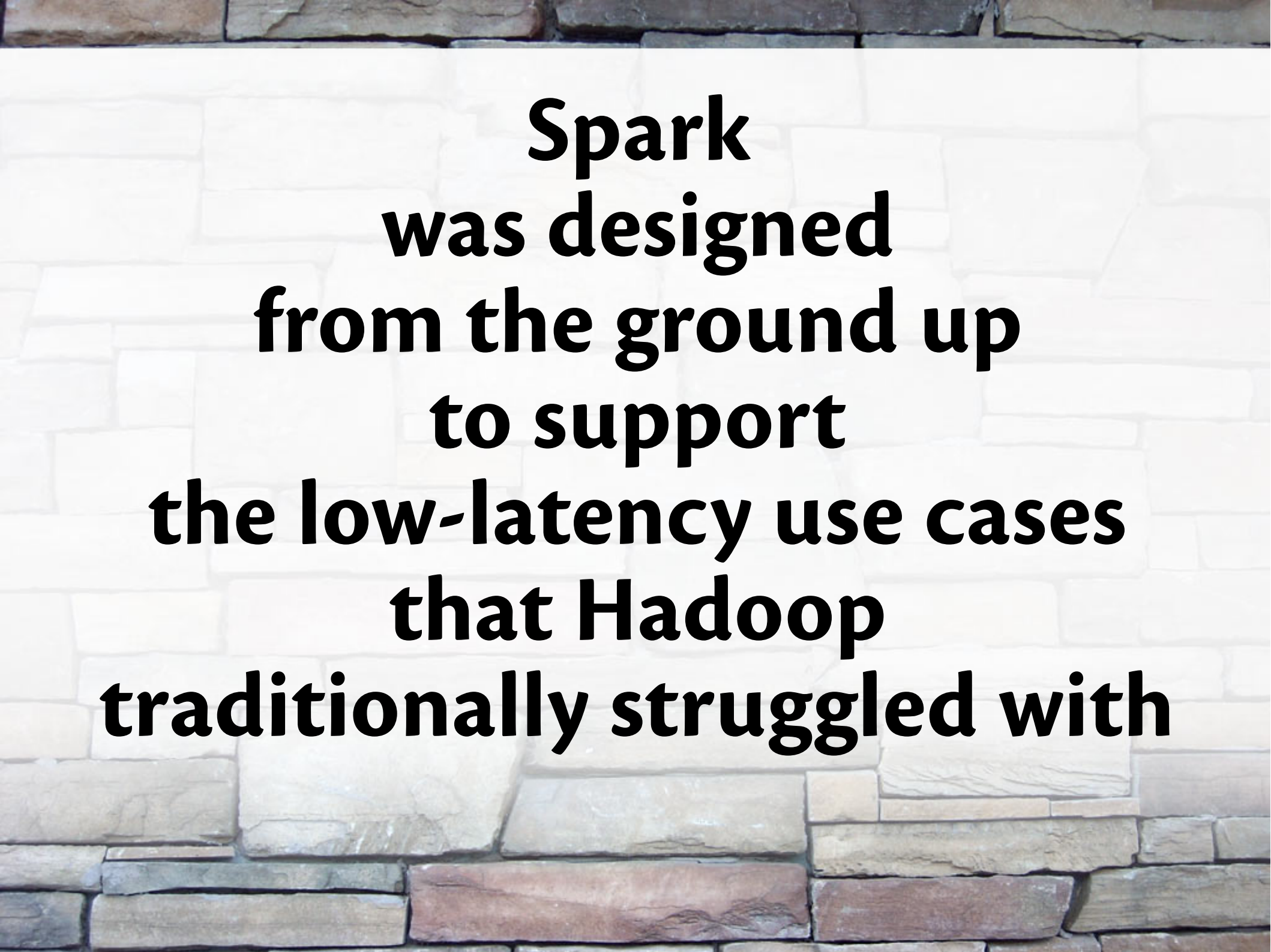
**How about a
RAM-to-RAM-to-RAM
pipeline?**

A background image of a stone wall. The wall is composed of irregularly shaped stones in various shades of beige, tan, and brown. The stones are arranged in a roughly horizontal pattern, with some larger flat stones and some smaller, more angular pieces. The lighting is even, highlighting the textures of the stone surfaces.

Meet Spark

The background of the slide is a close-up photograph of a stone wall. The stones are of various sizes and shades of tan, beige, and light brown, with darker, reddish-brown stones interspersed near the bottom. The texture is rough and natural.

Apache Spark
is an
open-source
cluster computing framework
with in-memory processing
to make analytics applications
up to 100 times faster
compared to
Hadoop



**Spark
was designed
from the ground up
to support
the low-latency use cases
that Hadoop
traditionally struggled with**



Spark
combines
an engine for distributing programs
across clusters of machines
with
an elegant programming model
for writing programs on top of it
and a set of libraries
that can help you
do cool and powerful things with
Big Data

Spark

**is finally an open source framework
that allows a data scientist
to be productive
with large data sets
by making distributed programming
truly accessible**



Spark's promise

**is to make writing
distributed programs
feel like writing
regular programs**



**Spark's execution engine
achieves
near linear scalability**

Linear scalability

**As the data size increases,
we can throw more computers at it
and see jobs complete
in the same amount of time**

Spark

**is resilient to the fact
that failures which
occur rarely on a single machine
occur all the time
on clusters of thousands of machines**

Spark


**breaks up work
into small tasks
and can gracefully accommodate
task failures
without compromising
the job to which they belong**

Spark

**is well suited
for highly interactive
data applications
that quickly respond
to user queries
by scanning
large in-memory data sets**

Spark's fundamental abstraction

**the Resilient Distributed Dataset (RDD)
allows developers
to materialize
any point in a processing pipeline
into memory
across the cluster**



**In this way,
future steps
that want to deal
with the same data set
do not need
to recompute it
or reload it from disk**

An RDD

**is essentially
a read-only collection of objects
that are partitioned
across machines**

An RDD

**can be operated on
in a parallel manner
using Spark's built-in set of
over 80 high-level operators,
enabling users
to express computation steps
in a natural fashion,
using a functional programming mindset**

A background image of a stone wall with light-colored, irregularly shaped stones in the center and darker, reddish-brown stones at the top and bottom.

An RDD

can be created at the shell prompt

An RDD

**can be interactively queried,
sliced and diced at the shell prompt
using more than 80 operators**

**Did I mention it's fun to
program in Spark?**

**The RDD API
is highly focused
on the needs of its users
(the data engineers)**

The background of the image is a close-up of a stone wall. The stones are of various sizes and shades of brown, tan, and grey, arranged in a rough, natural pattern. In the center of the image, there is a large, white rectangular area that serves as a backdrop for the text.

**Spark's vision
is to make the transition
from working
on a single machine
to working on a cluster,
a seamless experience**

Spark
SQL

Spark
Streaming

MLlib
(machine
learning)

GraphX
(graph)

Apache Spark

The image features a background of a stone wall. The wall is composed of various sized and shaped stones. The top and bottom sections of the wall are made of darker, reddish-brown stones. The central portion of the wall is made of lighter, cream-colored stones. A large, white rectangular area is superimposed over the center of the wall, containing the text "Demo time!" in a bold, black, sans-serif font.

Demo time!



Follow me on GitHub

github.com/dserban