

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

Distributed Computing with Apache Spark



Introduction to Big Data Concepts with Apache Spark

Apache Spark is an open-source cluster computing framework with in-memory processing to make analytics applications up to 100 times faster compared to technologies in wide deployment today. Developed in the AMPLab at UC Berkeley, Spark can help reduce data interaction complexity, increase processing speed and enhance data-intensive, real-time applications with deep intelligence.

Highly versatile in many environments, and with a strong foundation in functional programming, Spark is known for its ease of use in creating algorithms that harness insight from complex data. Spark was elevated to a top-level Apache Project in 2014 and continues to expand today.

Pagini

- [Calendar 2015](#)
- [Înscriere](#)
- [Workshops](#)

Workshops

- [Twelve days of functional programming](#)
- [Introduction to Big Data Concepts with Apache Spark](#)

The background of the image is a stone wall. The top portion of the wall is composed of dark, reddish-brown stones. Below this, there is a large, irregular white rectangular area that serves as a backdrop for the text. The bottom portion of the wall consists of lighter-colored, tan and grey stones.

Spark is one of the biggest reasons that Big Data is such an exciting area of work for technologists right now



Spark is hugely popular

**It has the most active
community of any open
source big data project
currently in development**

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, cream, and light brown. The stones are laid in a pattern that resembles a running bond or similar masonry style. The lighting is even, highlighting the textures and edges of the stones.

**This presentation
is an overview of Spark**

Disclaimer

I do not use Spark in a production setting

**The source material for this presentation
comes from my own Spark learning
experience as well as my own knowledge of
data engineering topics in general**




But first ...

What is Big Data?



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

The background of the image is a stone wall made of irregular, light-colored stones. A large, white rectangular area is centered on the wall, containing the text. The text is in a bold, black, sans-serif font and is arranged in five lines.

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

Big Data is defined by the three Vs


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

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 and reddish tones. The texture is rough and uneven.

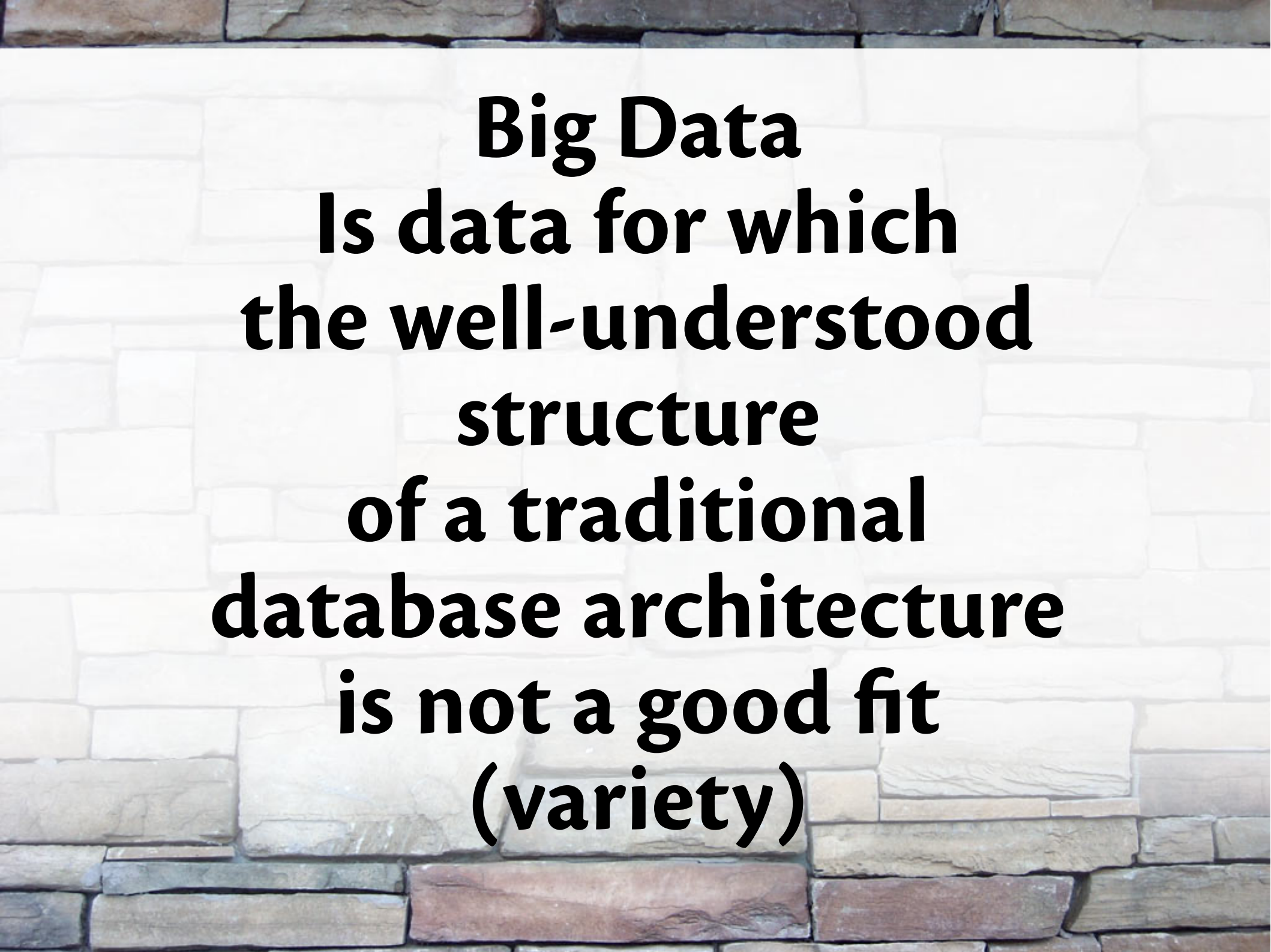
Big Data

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 shapes, with colors ranging from light beige to dark brown. The wall is composed of several horizontal courses of stones. In the center of the image, there is a large, bold, black text overlay.

**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 section features a row of dark, reddish-brown stones. Below this, the wall is composed of larger, light-colored (tan and cream) rectangular stones arranged in a regular pattern. The bottom section of the image shows a row of smaller, darker reddish-brown stones, similar to the top row.

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

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)**

**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**

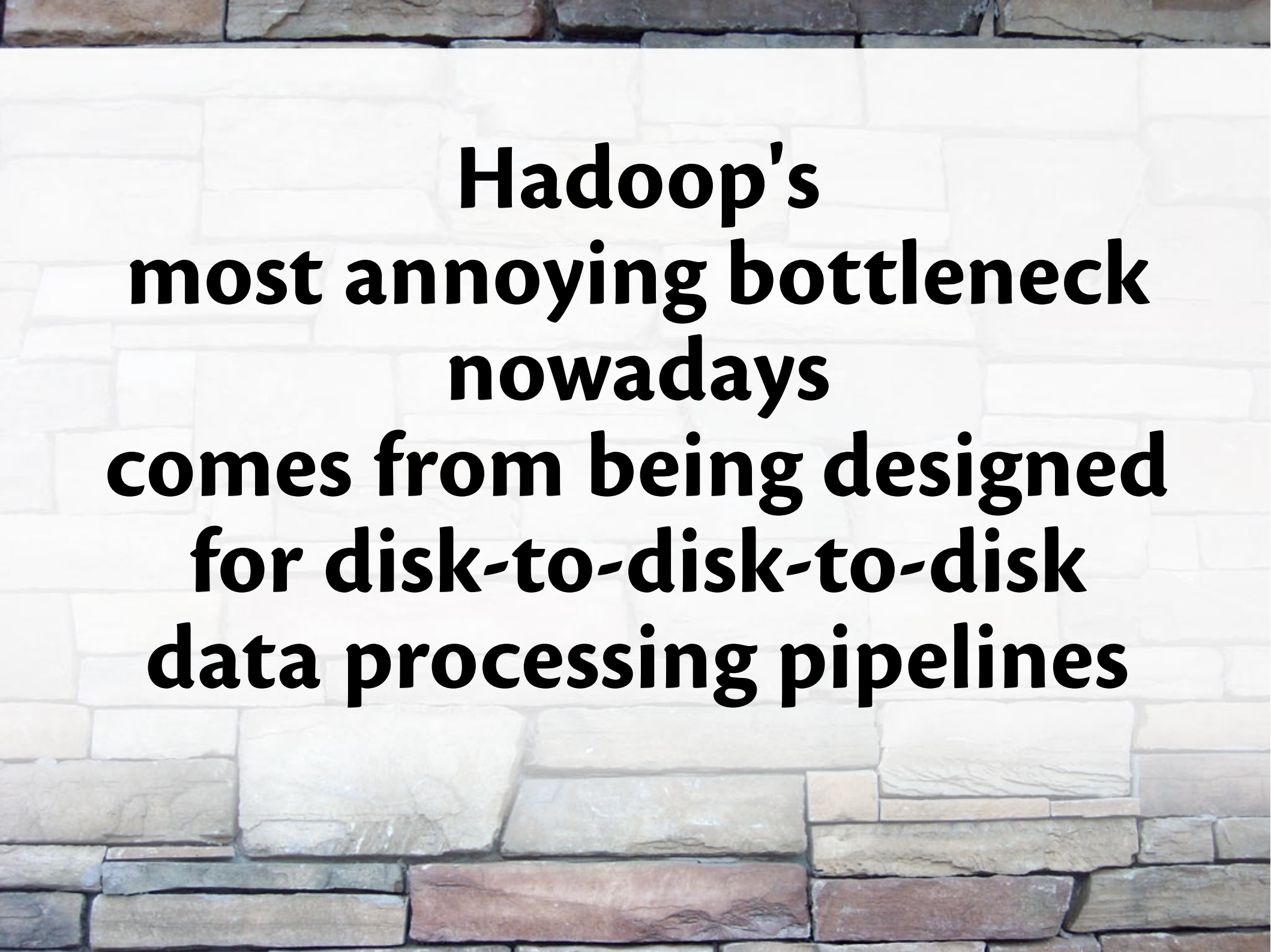


**Hadoop
became successful
when spinning media storage
became cheap enough**

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.

Year 2015:

**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**

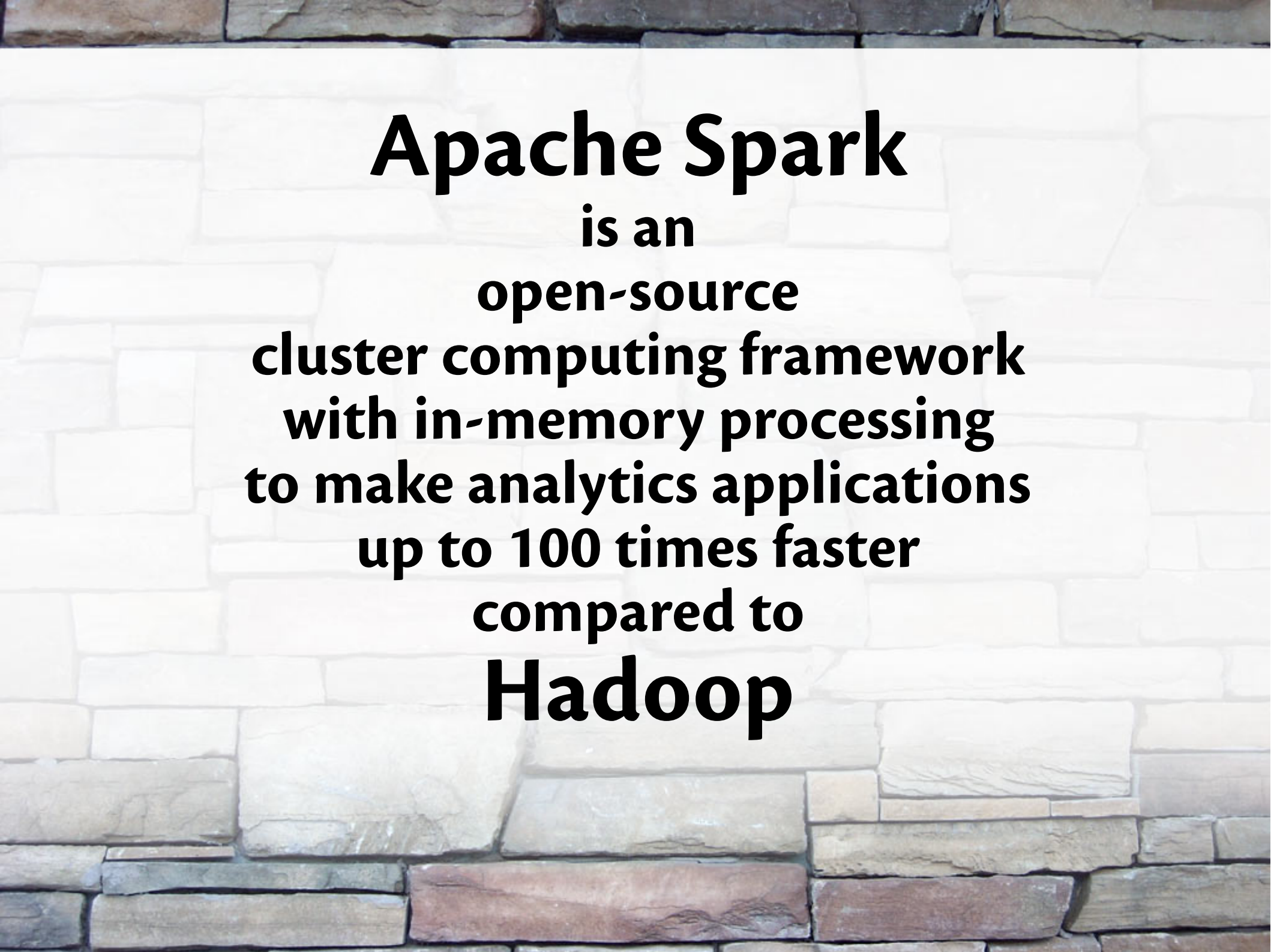
A background image of a stone wall with light-colored rectangular stones in the center and darker, more irregular stones at the top and bottom.

Obvious idea:

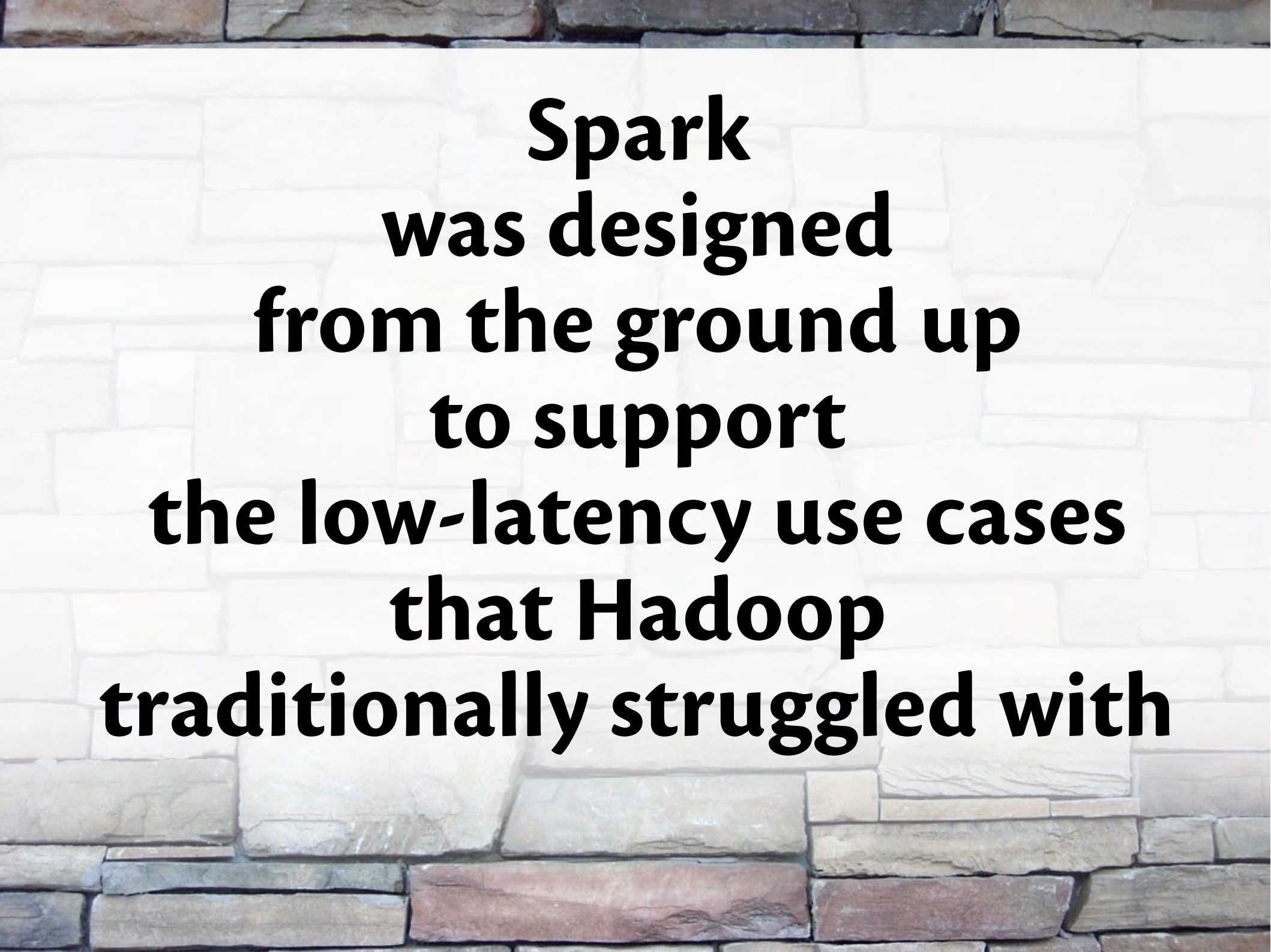
**How about a
RAM-to-RAM-to-RAM
pipeline?**



Meet Spark

The background of the slide is a close-up photograph of a stone wall. The stones are of various sizes and shades, including light beige, tan, and reddish-brown, arranged in a traditional masonry pattern.

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
provides
a distributed execution engine,
a fun-to-work-with programming model
and a set of libraries
that can help you
do cool and powerful things with
Big Data

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

**The 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 colors, including shades of tan, beige, and reddish-brown. In the center of the image, there is a 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!".

Demo time!

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.

Follow me on GitHub

github.com/dserban