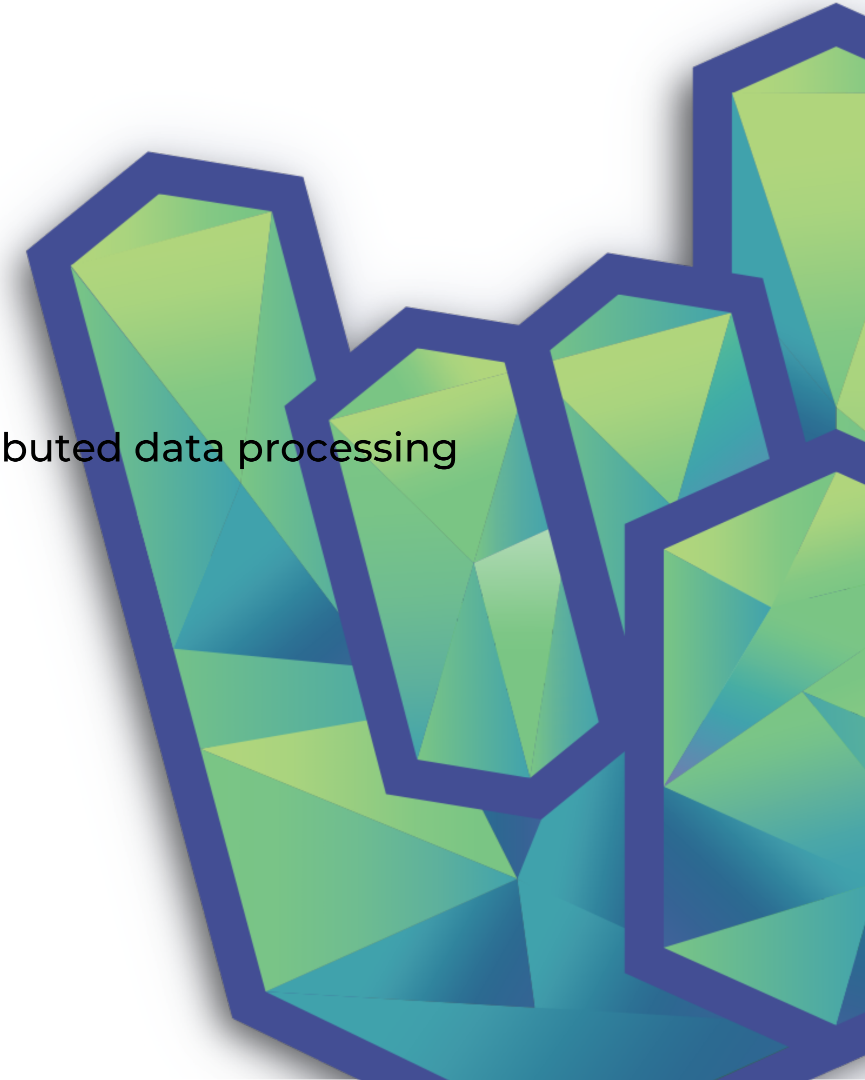
# First Principles

# What Spark Is

Unified computing engine & libraries for distributed data processing

Big data = cannot fit on a standard computer



# Unified Computing Engine

Spark supports a variety of data processing tasks

- data loading
- SQL queries
- machine learning
- streaming

Unified

- consistent, composable APIs in multiple languages
- optimizations across different libraries

Computing engine = detached from data storage & I/O

Libraries

- standard: Spark SQL, MLlib, Streaming, GraphX
- hundreds of open-source third-party libraries

# Context of Big Data

Computing vs data

- CPUs are only incrementally faster
- data storage keeps getting better and cheaper
- gathering data keeps getting easier and cheaper and more important

Data needs to be distributed and processed in parallel

Standard single-CPU software cannot scale up

Spark was born.

# Motivation for Spark

## A 2009 UC Berkeley project by Matei Zaharia et al

- MapReduce was the king of large distributed computation
- inefficient for large applications and ML
- each step required another data pass, written as separate application

## Spark phase 1

- a simple functional programming API
- optimize multi-step applications
- in-memory computation and data sharing across nodes

## Spark phase 2

- interactive data science and ad-hoc computation
- Spark shell and Spark SQL

## Spark phase 3

- same engine, new libraries
- ML, Streaming, GraphX
- structured data & optimizations

# Why Spark is Important

The most popular data processing engine

- used at hundreds of companies
- well maintained and documented

Career boost

- one of the most in-demand technologies
- salaries > 120k
- consulting projects – sky is the limit
- timeless skills

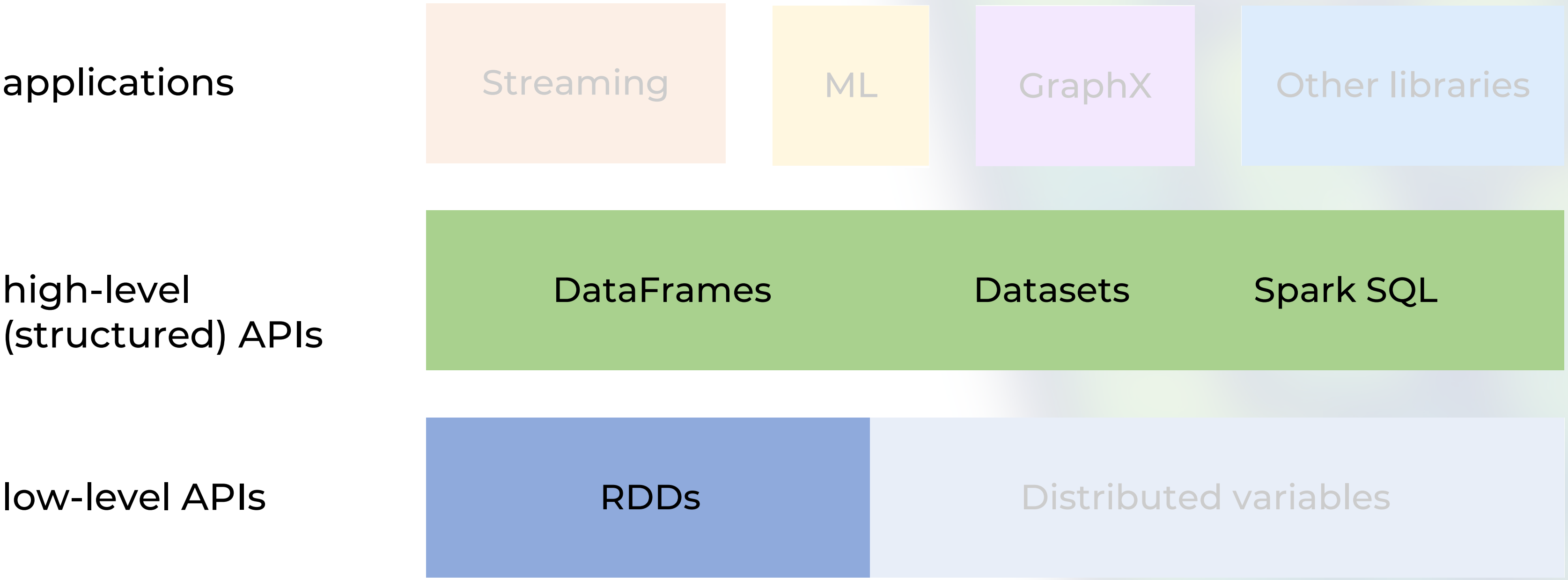
# Spark Misconceptions

Spark is not concerned with data sources

- files
- Azure
- S3
- Hadoop/HDFS
- Cassandra
- Postgres
- Kafka

Spark is not part of Hadoop

# Spark Architecture



applications

Streaming

ML

GraphX

Other libraries

high-level
(structured) APIs

DataFrames

Datasets

Spark SQL

low-level APIs

RDDs

Distributed variables

# Spark rocks