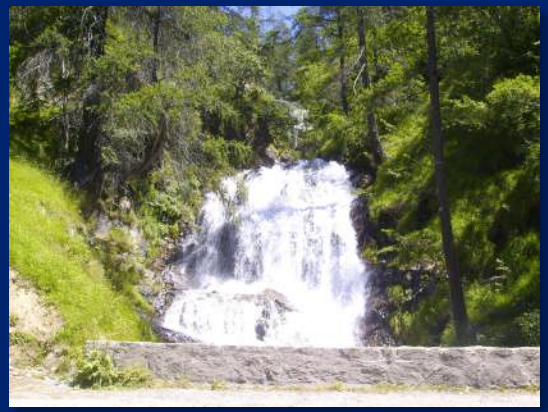
Summary of Introduction to Big Data

After this video you will be able to…

- Recall what started the big data era and the three main big data sources
- Summarize the volume, variety, velocity and veracity issues related to each source
- Explain the 5-step data science process to gain value from big data
- Remember the main elements of the Hadoop stack

Big Data Era



Data Torrent

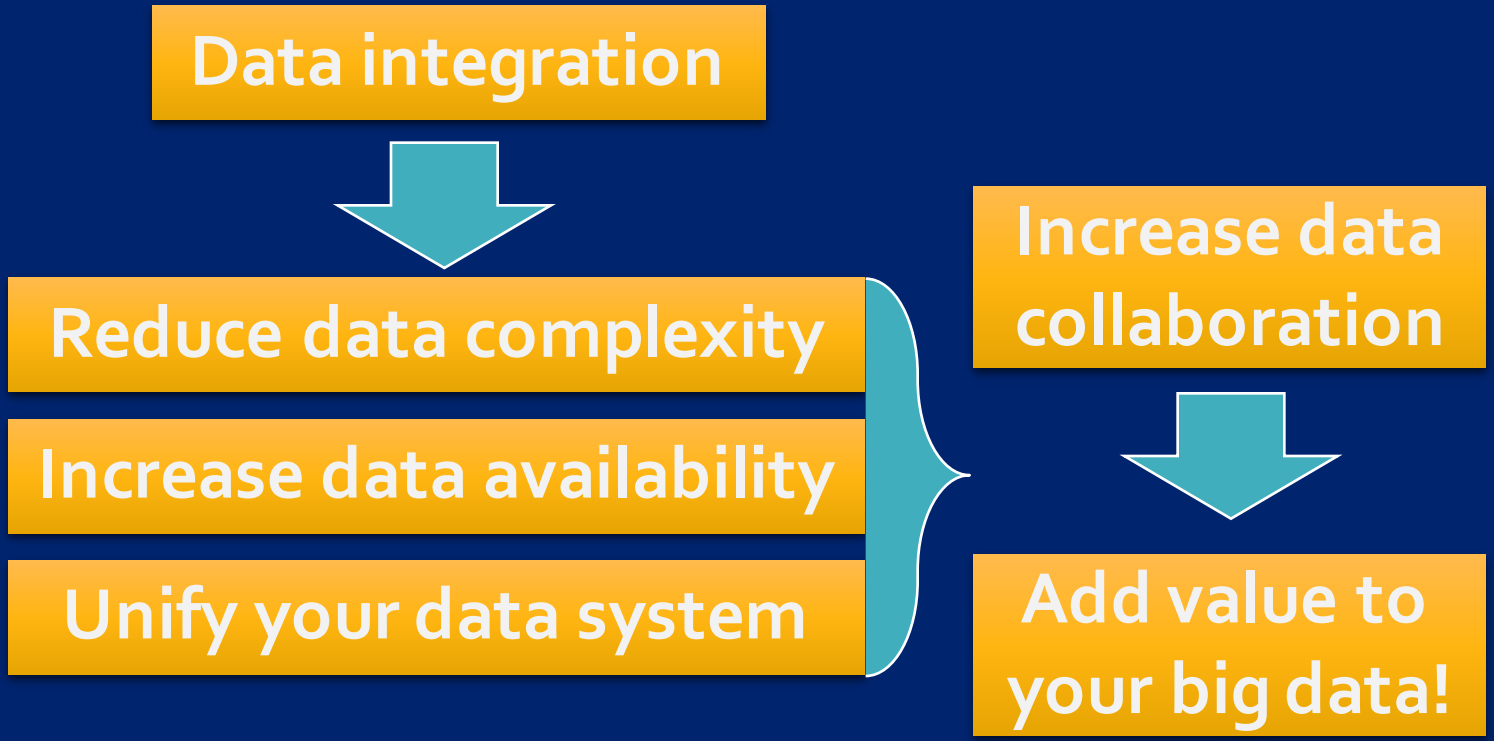
Computing
Anytime, Anywhere

Three major sources of big data

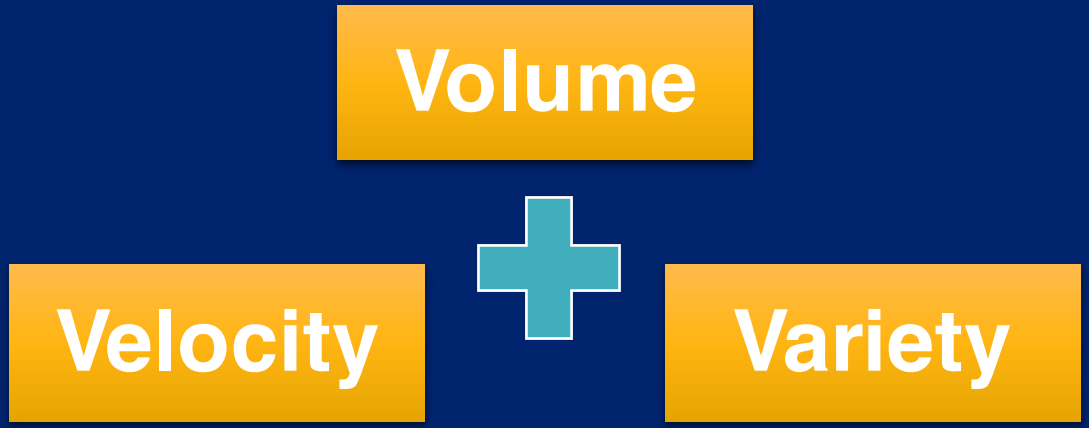
- Machines
- People
- Organizations

Getting Value from Big Data

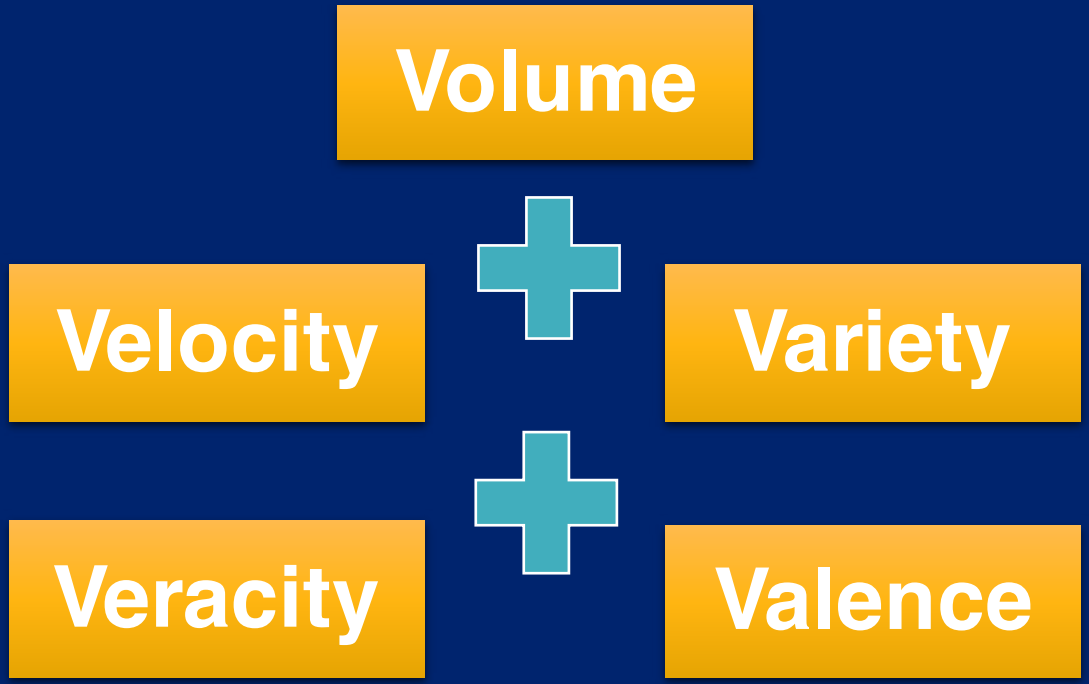
Value comes from integrating different types of data sources

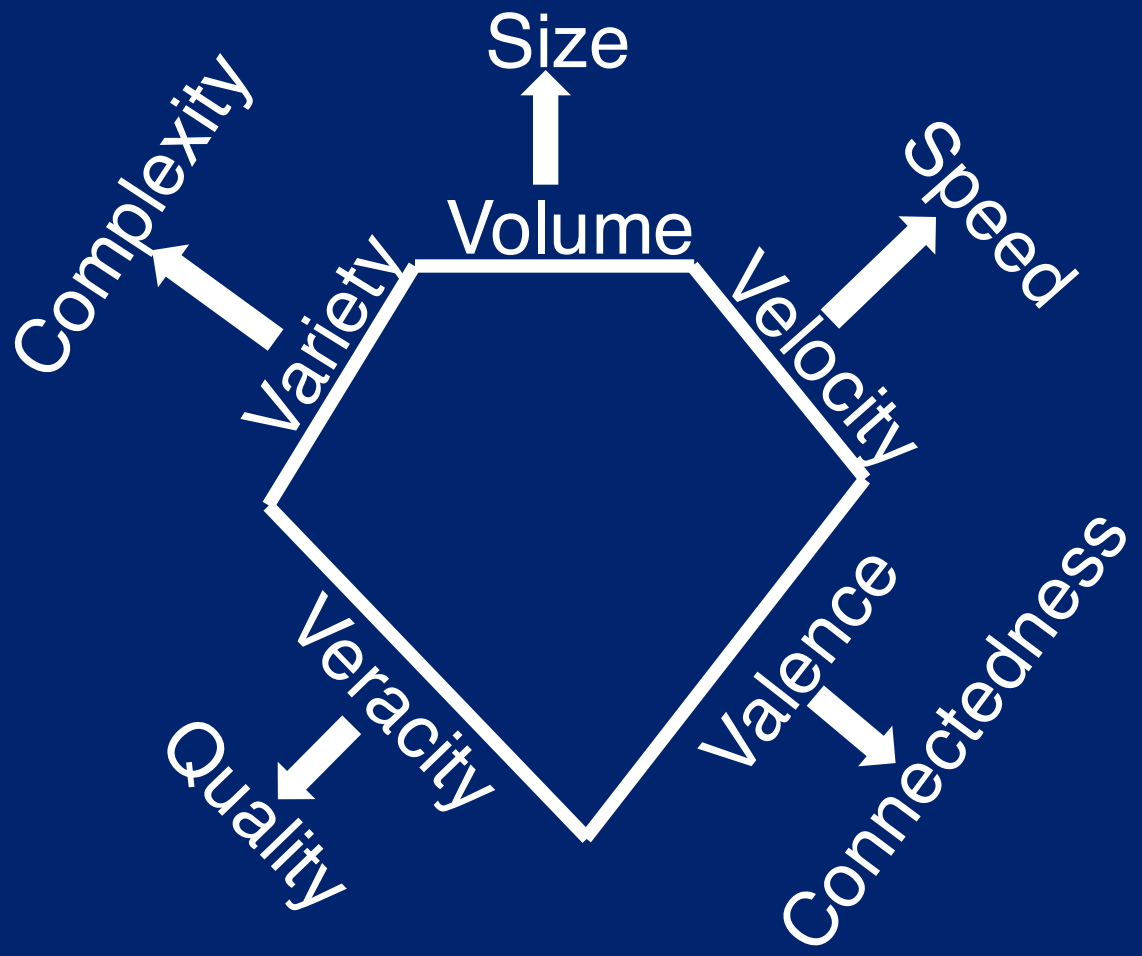

Data integration
Reduce data complexity
Increase data availability
Unify your data system
Increase data collaboration
Add value to your big data!

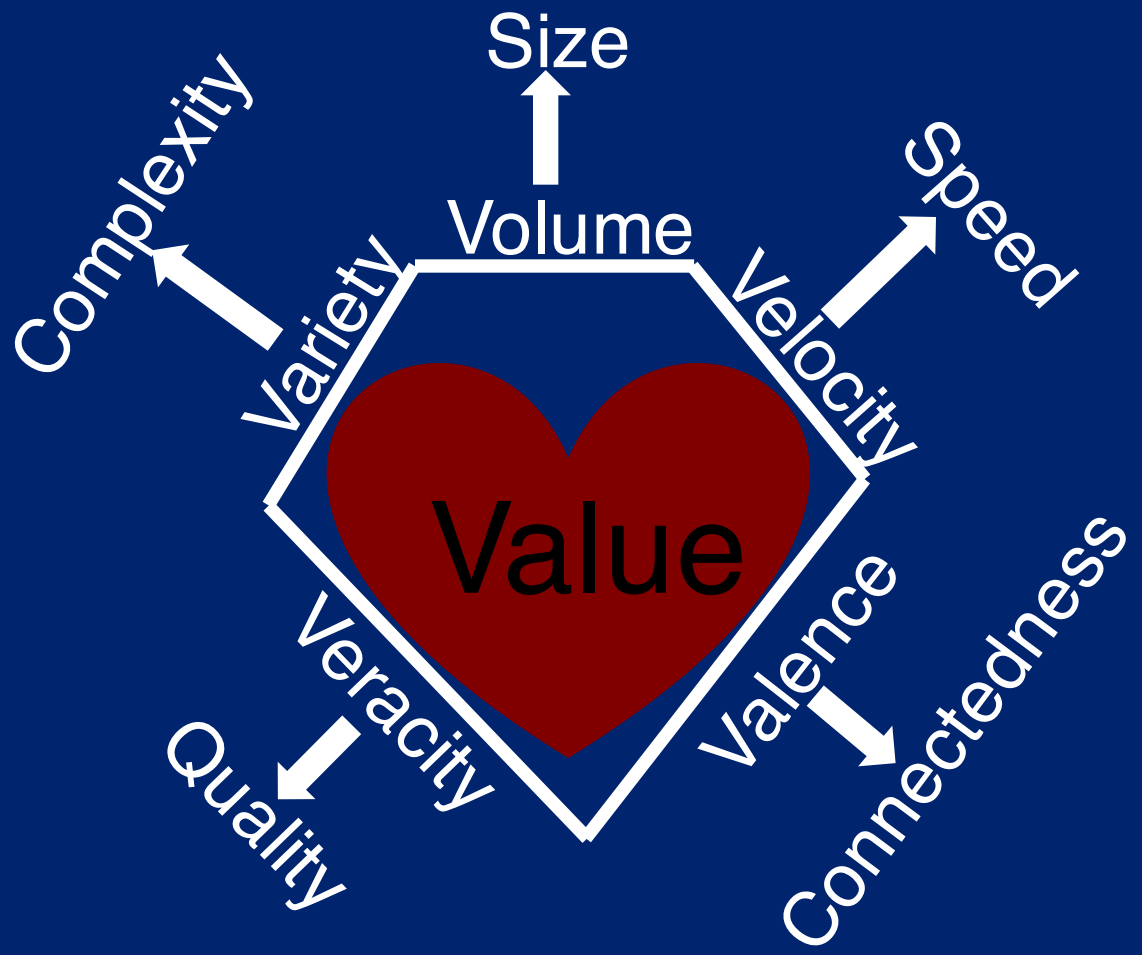
Characteristics of Big Data



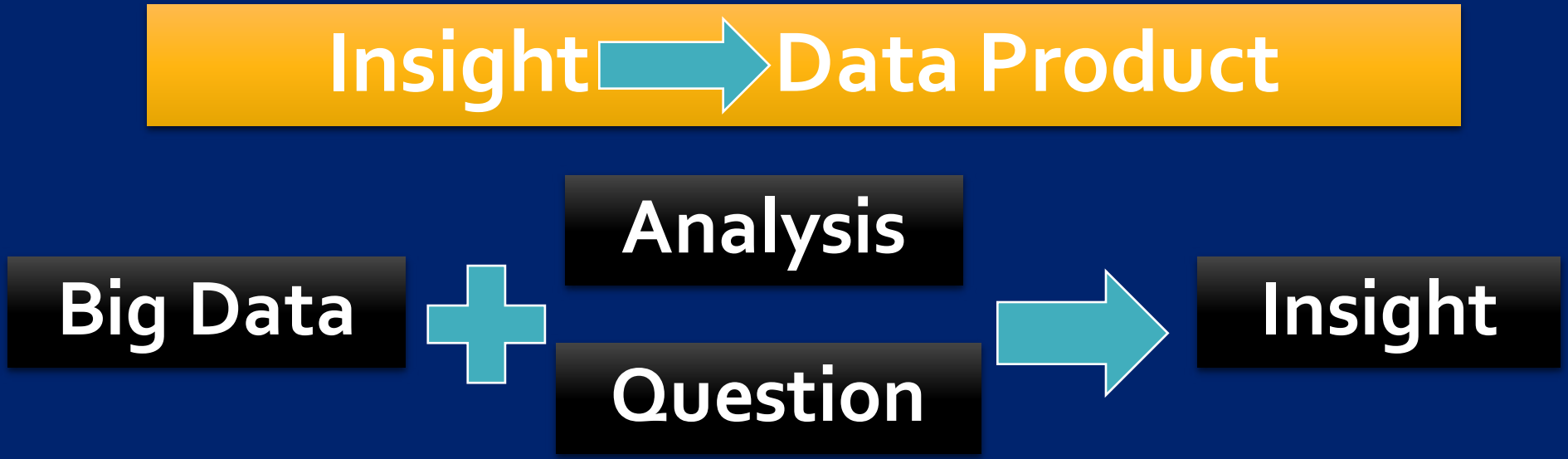
Characteristics of Big Data



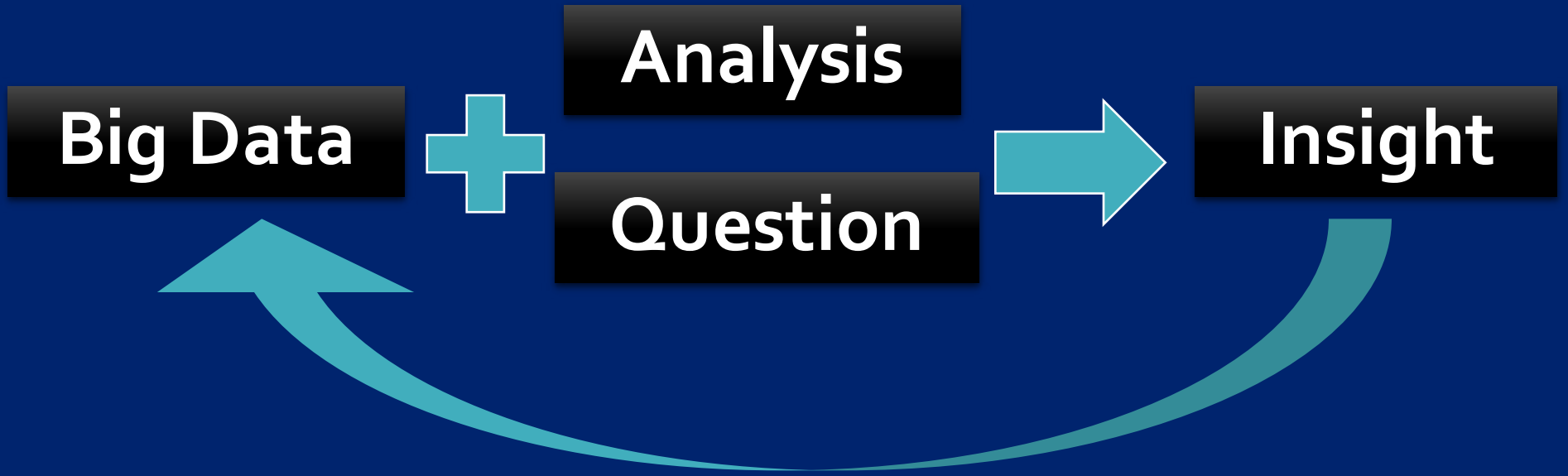

Size
Volume
Complexity
Variety
Velocity
Speed
Veracity
Quality
Valence
Connectedness


Size
Volume
Speed
Velocity
Complexity
Variety
Value
Veracity
Quality
Valence
Connectedness

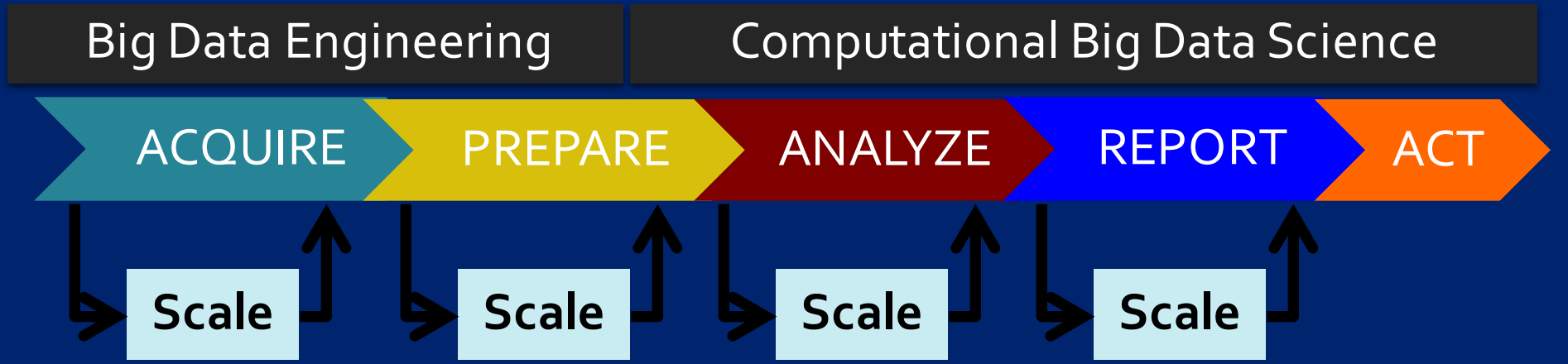

Big Data
Insight
Action
Data Science


Insight
Data Product
Big Data
Analysis
Question
Insight

Insight → Data Product




Big Data Engineering
Computational Big Data Science
ACQUIRE
PREPARE
ANALYZE
REPORT
ACT


Big Data Engineering
Computational Big Data Science
ACQUIRE
PREPARE
ANALYZE
REPORT
ACT
Scale
Scale
Scale
Scale

ACQUIRE PREPARE ANALYZE REPORT ACT

Step 1: Acquire Data

- Identify data sets
- Retrieve data
- Query data

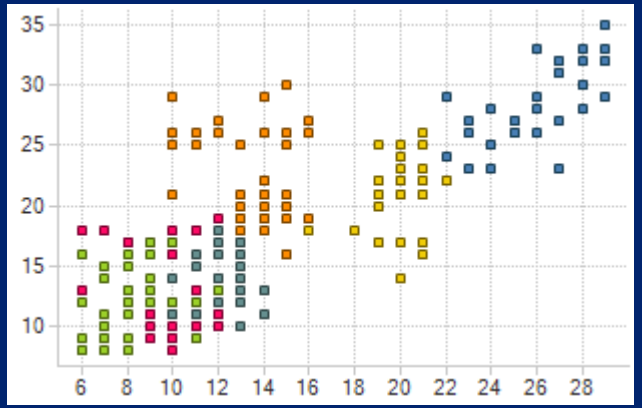
Step 2: Prepare Data

Step 2-A: Explore

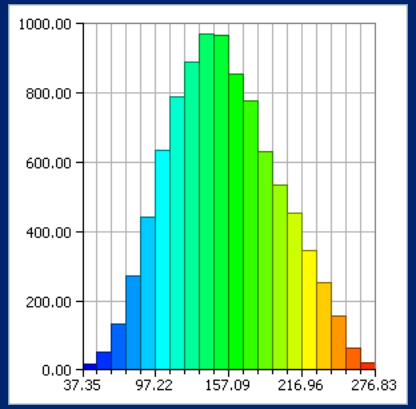
Step 2-B: Pre-process

ACQUIRE | **PREPARE** | ANALYZE | REPORT | ACT

Step 2-A: Explore Data

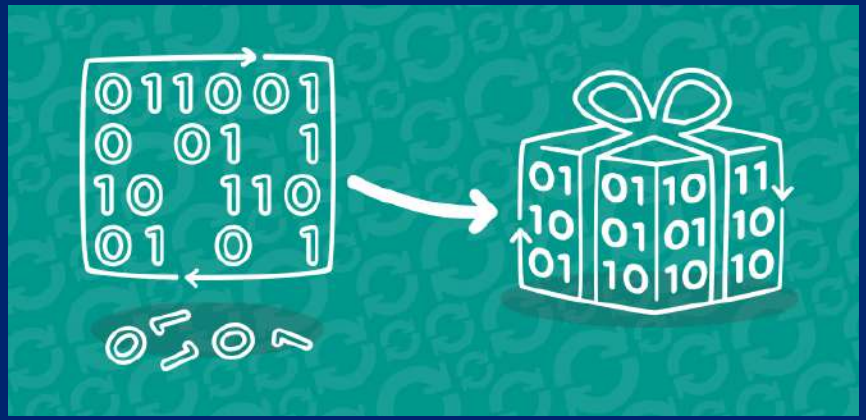


Understand nature of data



Preliminary analysis

Step 2-B: Pre-process Data



Clean

Integrate

Package

ACQUIRE | PREPARE | **ANALYZE** | REPORT | ACT

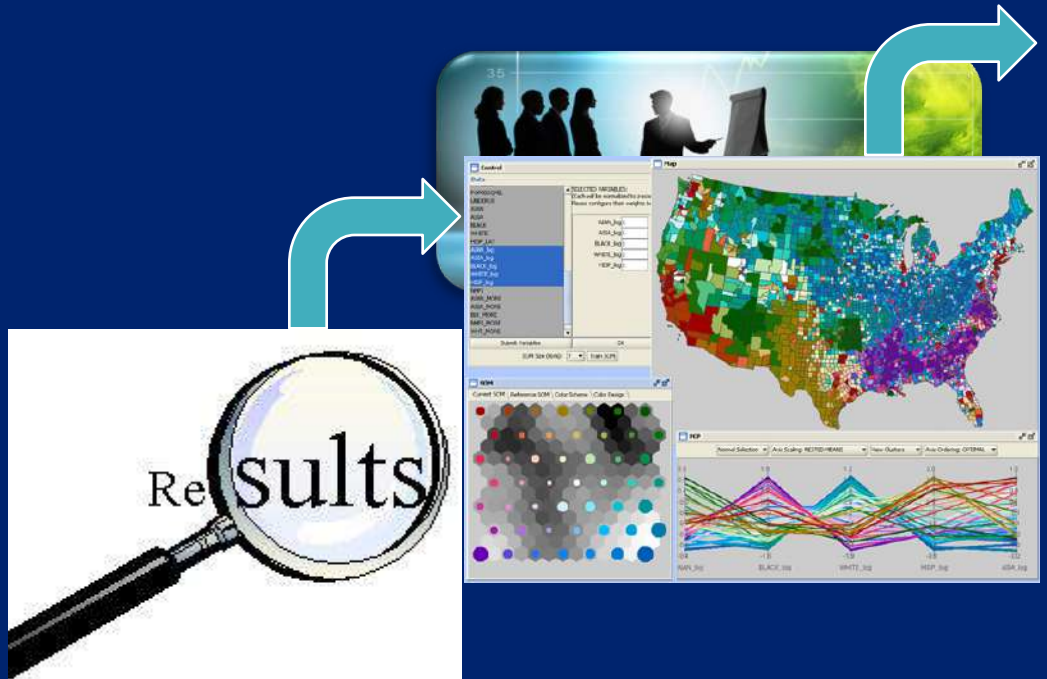
Step 3: Analyze Data

Select analytical techniques

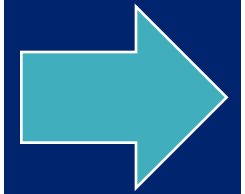
Build models

ACQUIRE | PREPARE | ANALYZE | **REPORT** | ACT

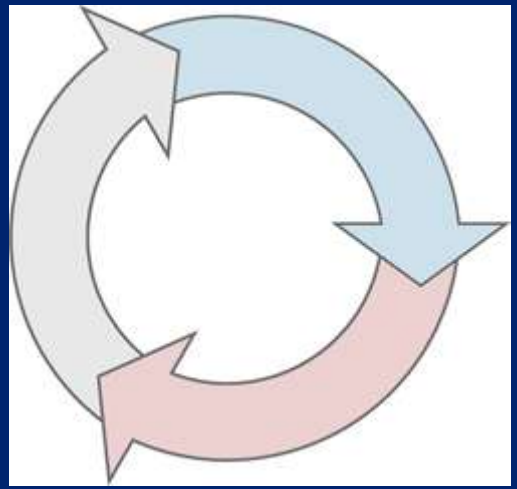
Step 4: Communicate Results



Step 5: Apply Results



ACQUIRE PREPARE ANALYZE REPORT ACT

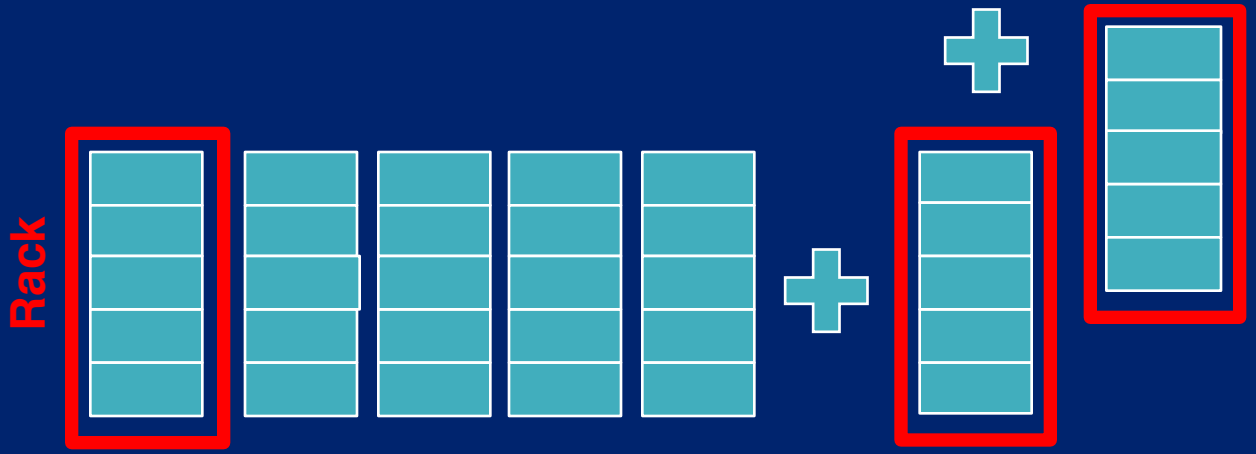


Iterative process

Scalable tools

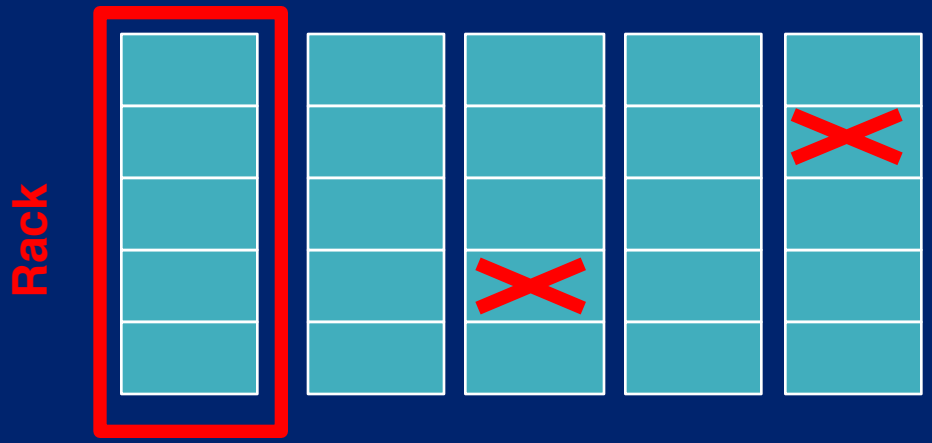
1. Enable Scalability

Commodity hardware is cheap

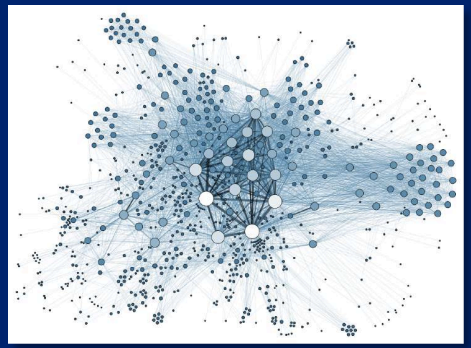


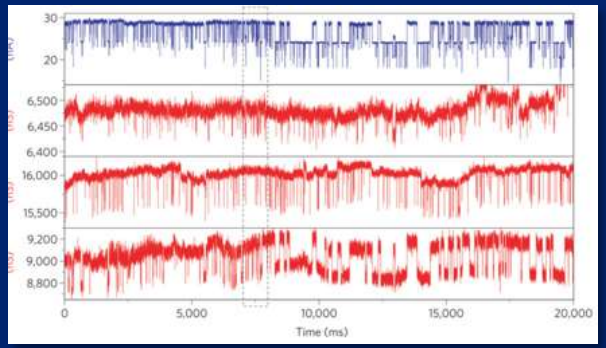
2. Handle Fault Tolerance

Be ready: crashes happen



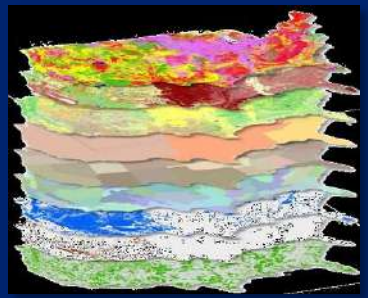
3. Optimized for a Variety Data Types



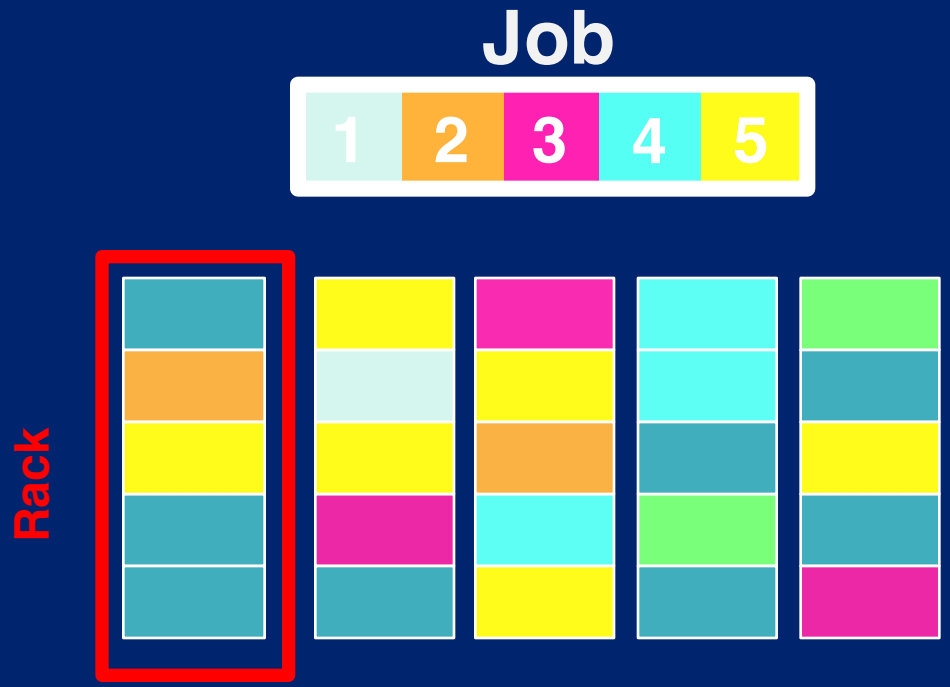


Cars marketplace

vendor	Model	Price	Mileage	VIN Code
Chevrolet	Corvette	17226	25965.0	ILLAKAWAZDZ
Chevrolet	Corvette	34229	46429.0	RCPNSRYGXCN
Chevrolet	Corvette	27982	50209.0	NWLGCEVEHGI
Chevrolet	Corvette	51825	72998.0	NGVZSCIZGSV
Chevrolet	Corvette	52945	34364.0	PSDRJYYDUG
Chevrolet	Malibu	37874	37273.0	VLFPQPWNEFD
Chevrolet	Malibu	15600	71441.0	EXLJGDWOZSA
Chevrolet	Malibu	52447	46700.0	NLWGJZAKBRC
Chevrolet	Malibu	27129	36254.0	OPFUBILEHSM
Chevrolet	Malibu	28846	77162.0	WRCOOFREZLL
Chevrolet	Malibu	46165	60590.0	HUFTTHQHSFJ
Chevrolet	Malibu	18263	37790.0	JLUHNAFSHVD



4. Facilitate a Shared Environment

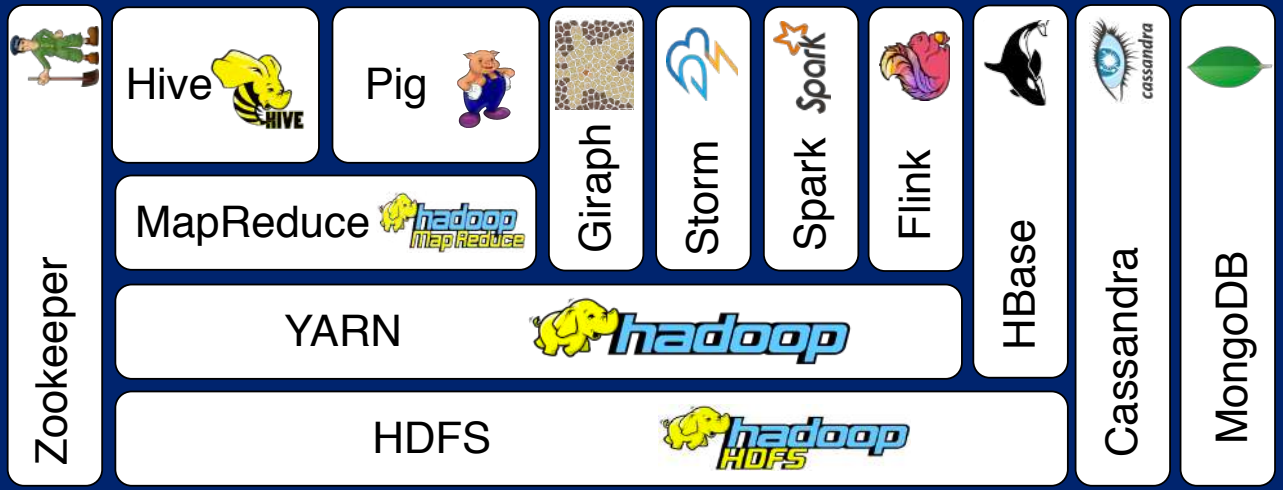


5. Provide Value

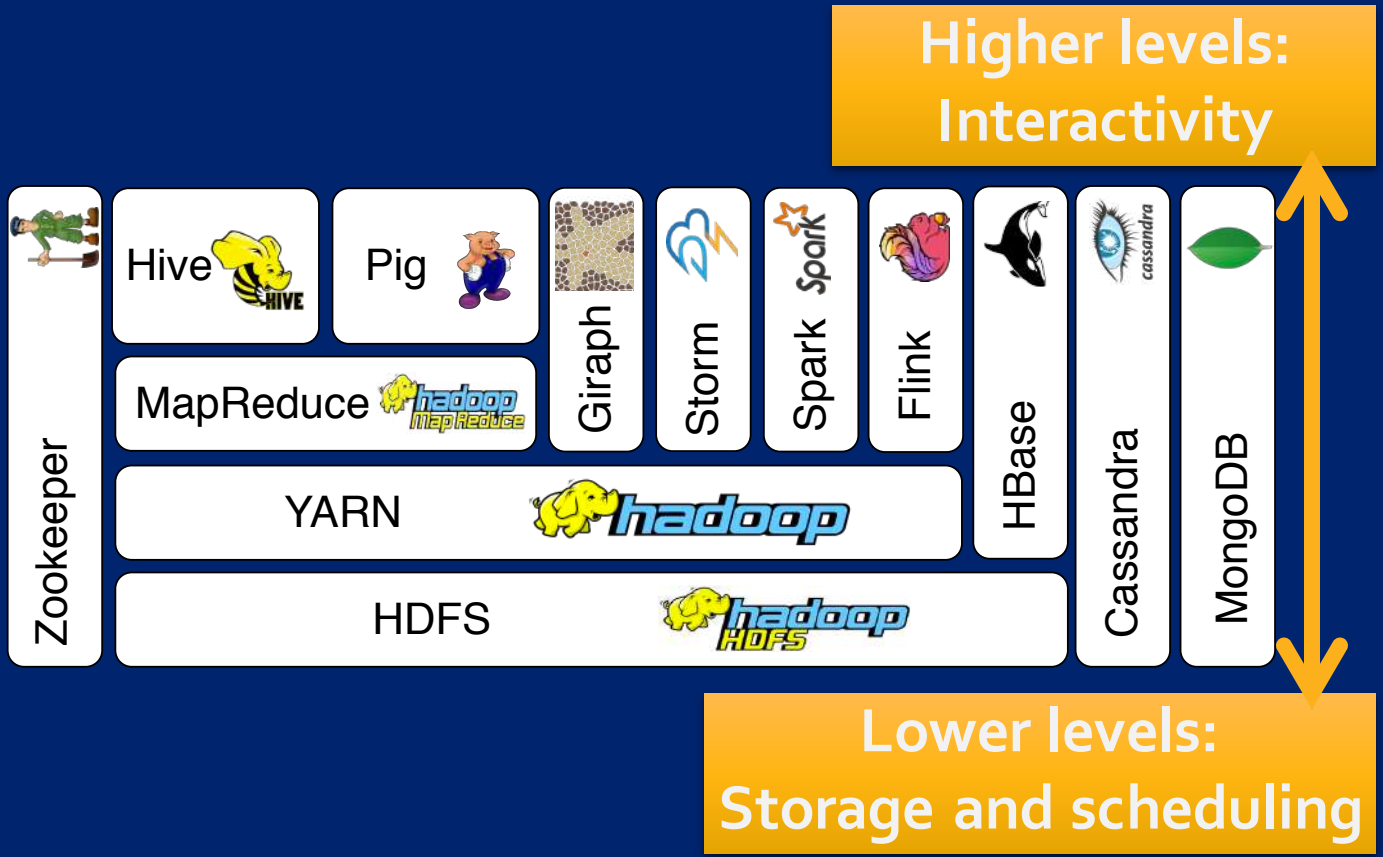
Community-supported

Wide range of applications

One possible layer diagram for Hadoop Ecosystem



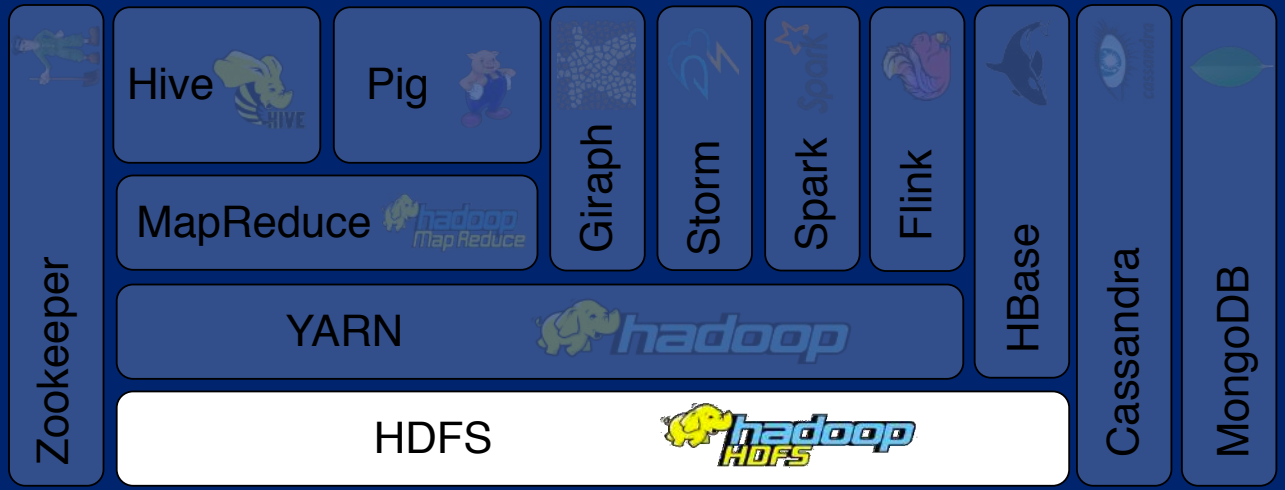
One possible layer diagram for Hadoop



Distributed file system as foundation

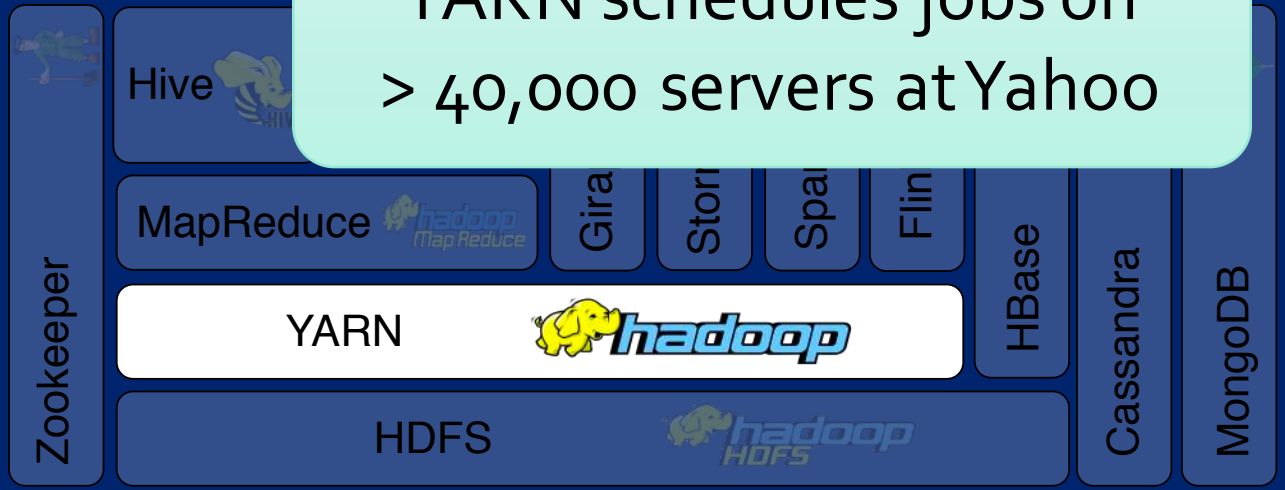
Scalable storage

Fault tolerance



Flexible scheduling and resource management

YARN schedules jobs on
> 40,000 servers at Yahoo



Simplified programming model

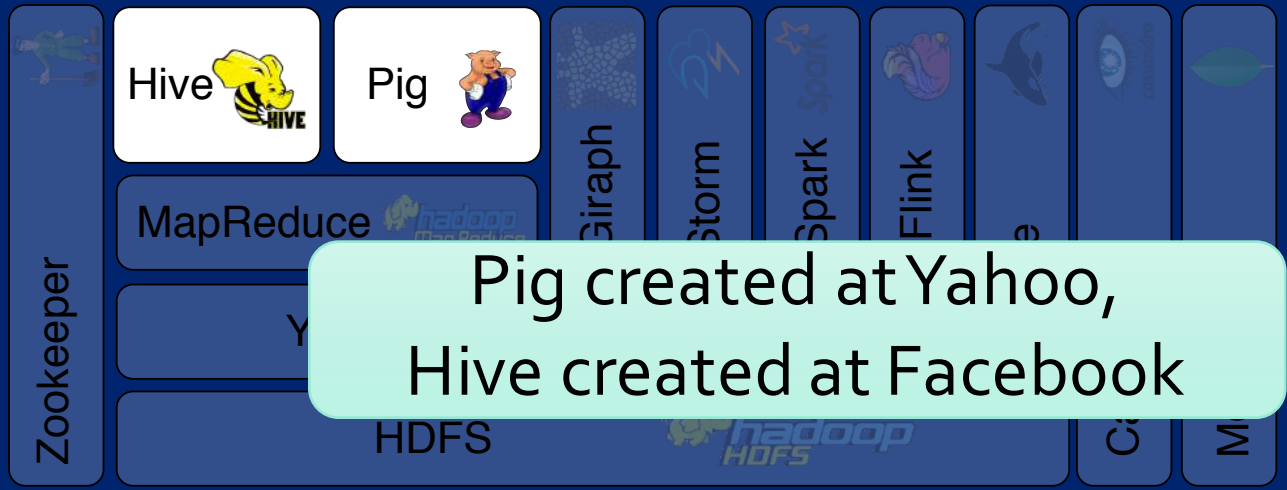
Map → apply()

Reduce → summarize()

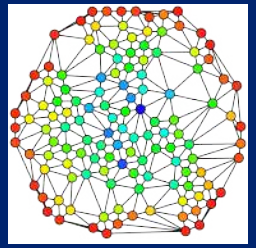
Higher-level programming models

Pig = dataflow scripting

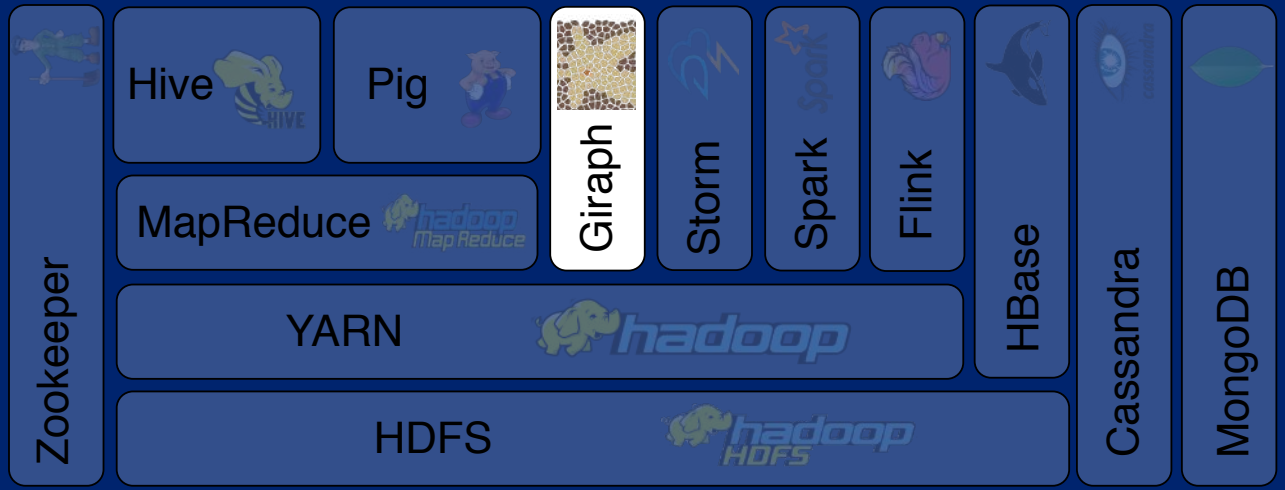
Hive = SQL-like queries



Specialized models for graph processing

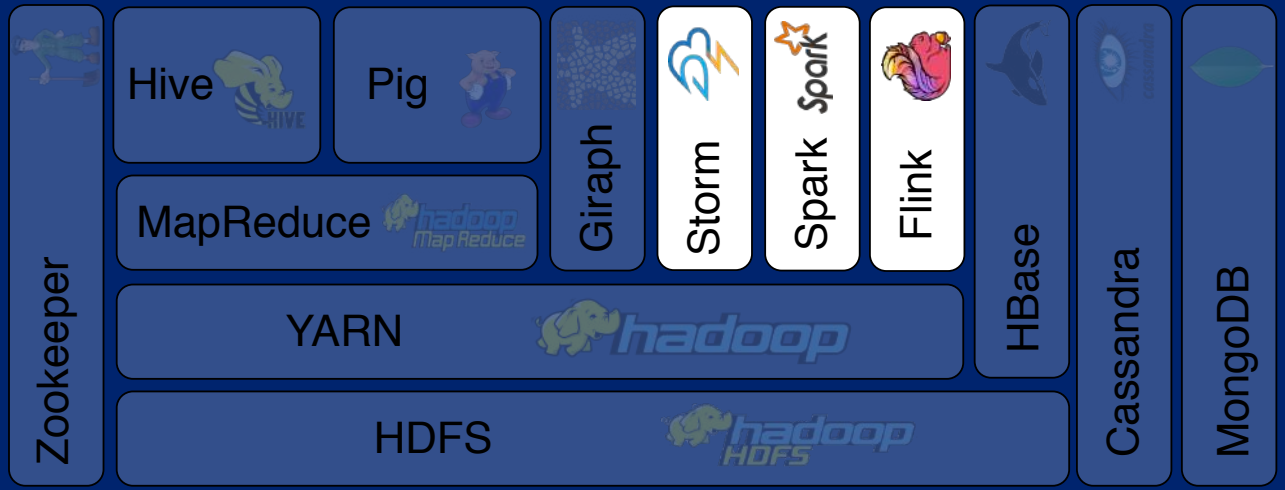


Giraph used by Facebook to analyze social graphs



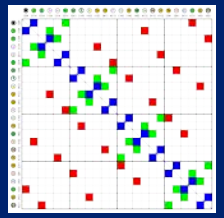
Real-time and in-memory processing

In-memory → 100x faster for some tasks



NoSQL for non-files

Key-values



Sparse tables

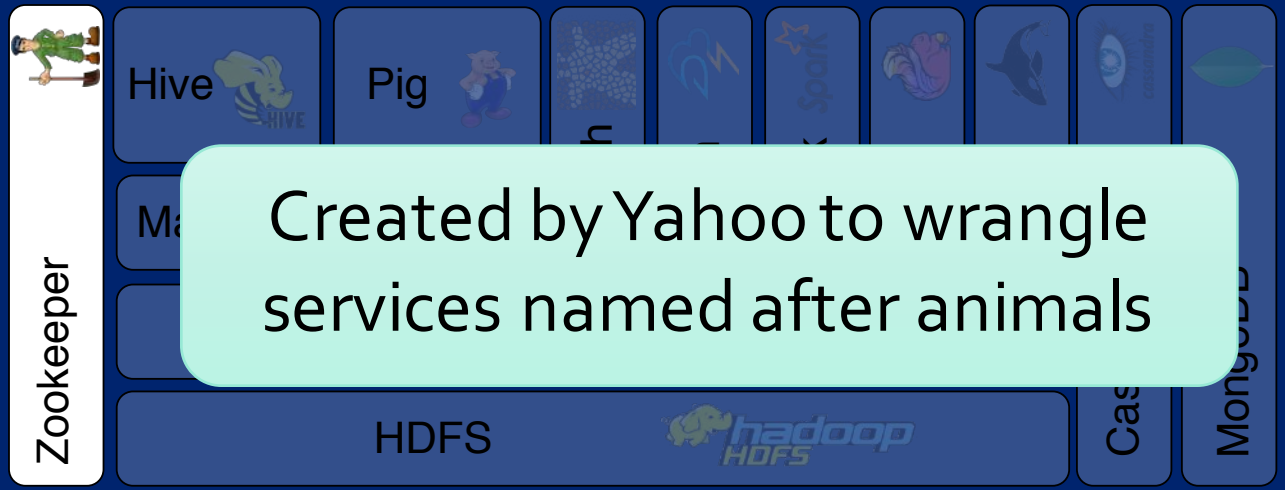
HBase used for Facebook's Messaging Platform

Zookeeper for management

Synchronization

Configuration

High-availability



Created by Yahoo to wrangle services named after animals

Many Big Data Modeling and Management Challenges

Big Data Platforms and Management Systems