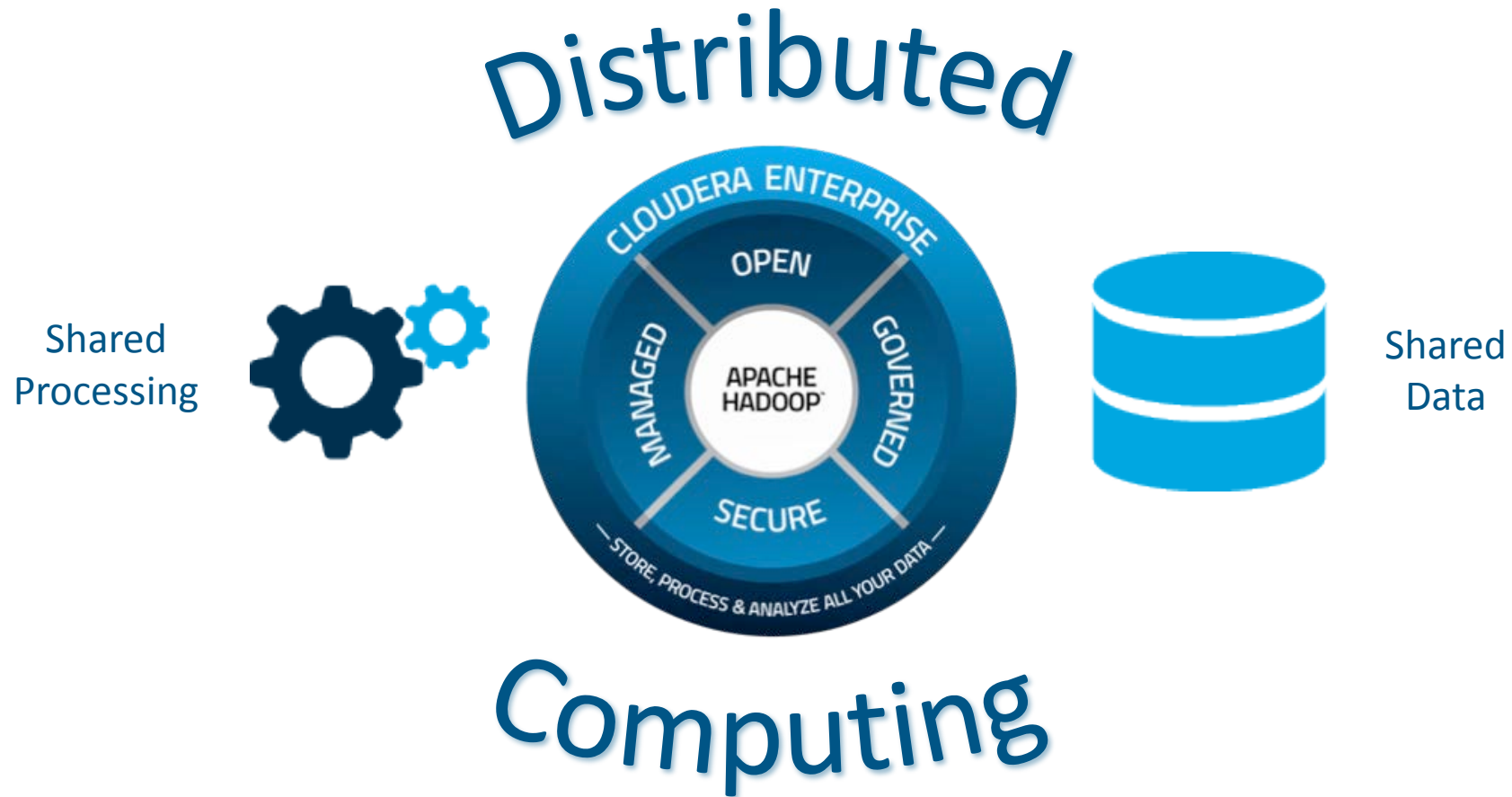


# What is Hadoop?

How does this stuff work?.



# Cloudera Technology

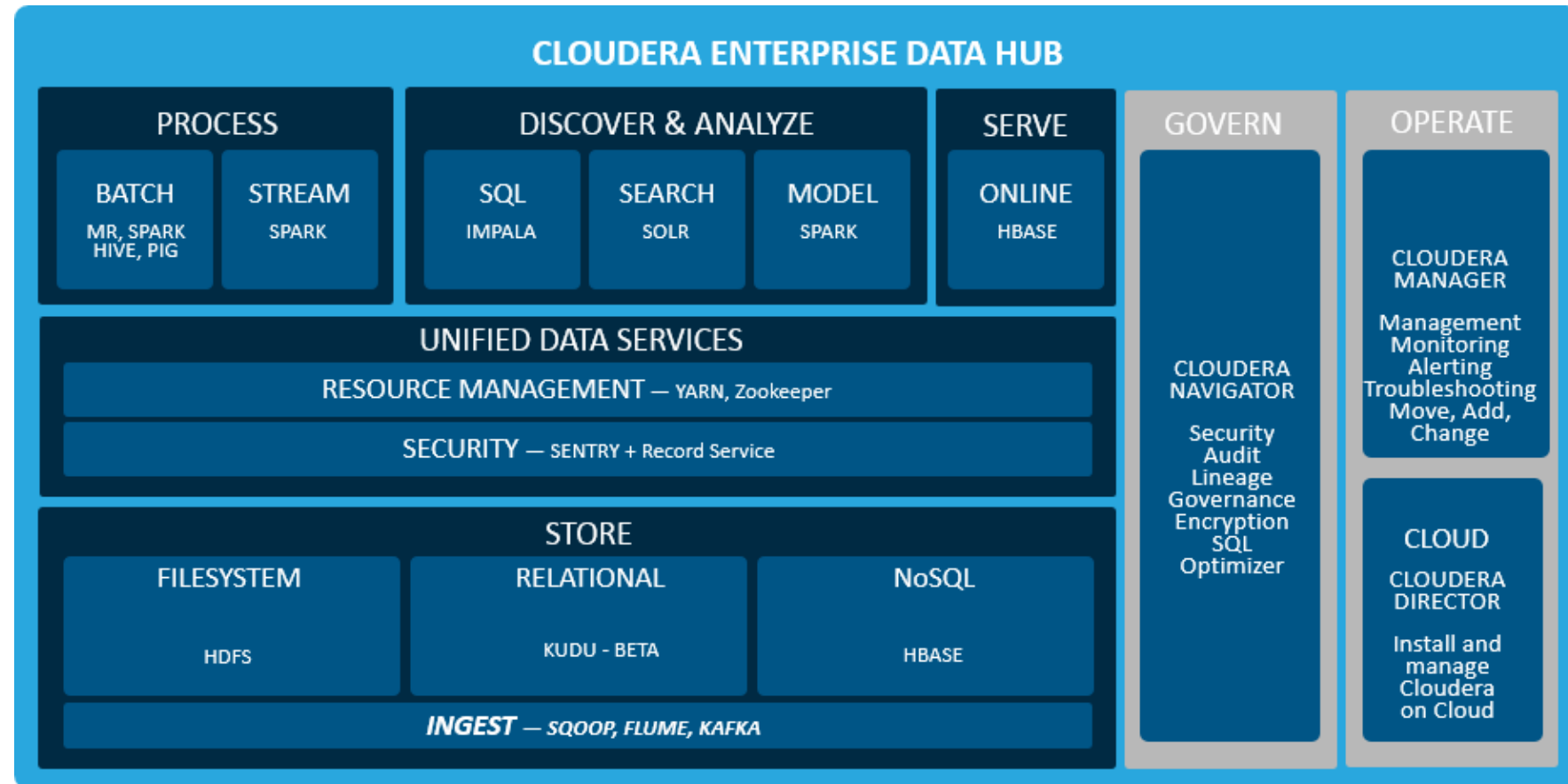
Making Hadoop Fast, Easy, and Secure for the Modernized Architecture

Hadoop is a new kind of data platform.

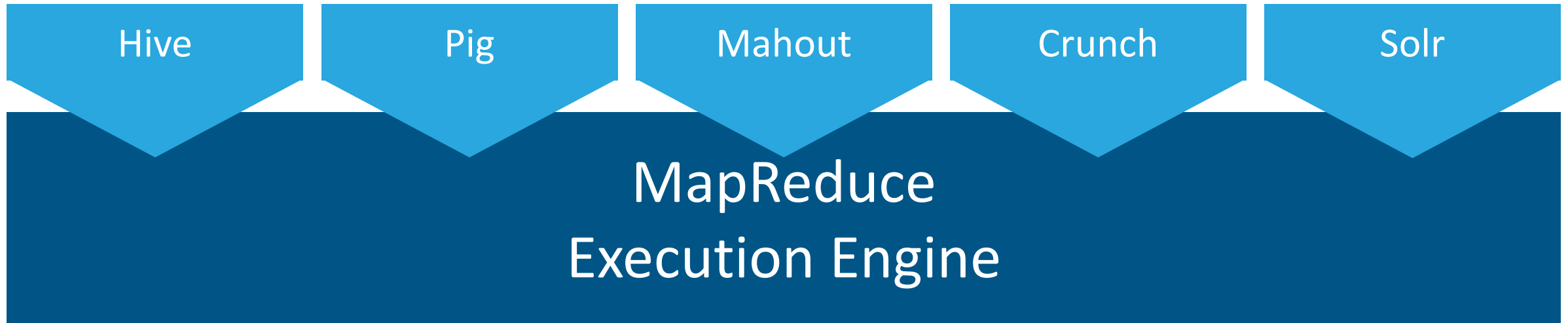
- One place for unlimited data
- Unified data access

Cloudera makes it:

- **Fast** for business
- **Easy** to manage
- **Secure** without compromise



# MapReduce: A great tool for its day



The original scalable, general, processing engine of Hadoop ecosystem

- Useful across diverse problem domains
- Fueled initial ecosystem explosion

# Enter Apache Spark

Flexible, in-memory data processing for Hadoop

## Easier Development

- Rich APIs for Scala, Java, and Python
- Interactive shell

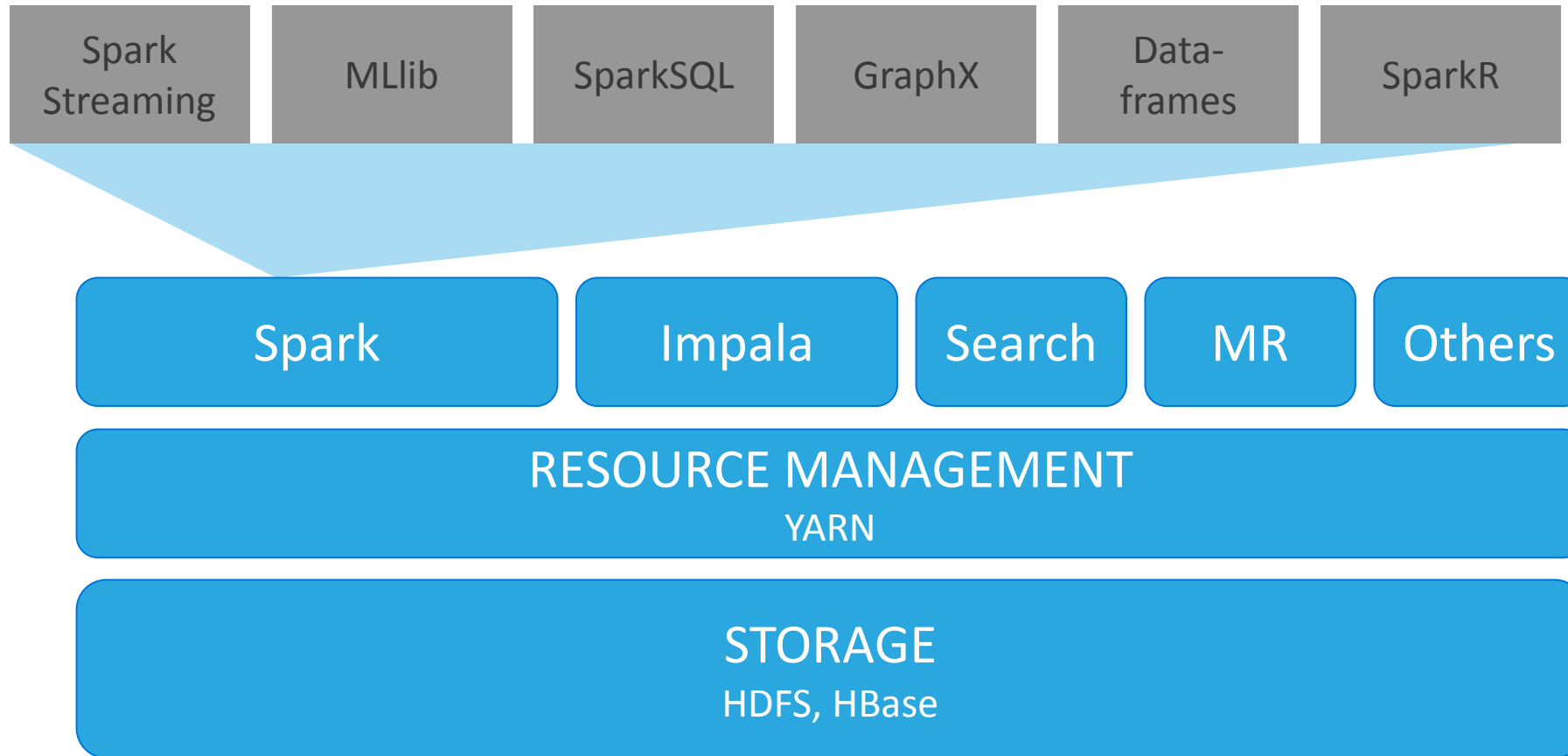
## Flexible, Extensible API

- APIs for different types of workloads:
  - Batch
  - Streaming
  - Machine Learning
  - Graph

## Faster Processing (Batch & Streaming)

- In-Memory processing and caching

# The Spark Ecosystem & Hadoop

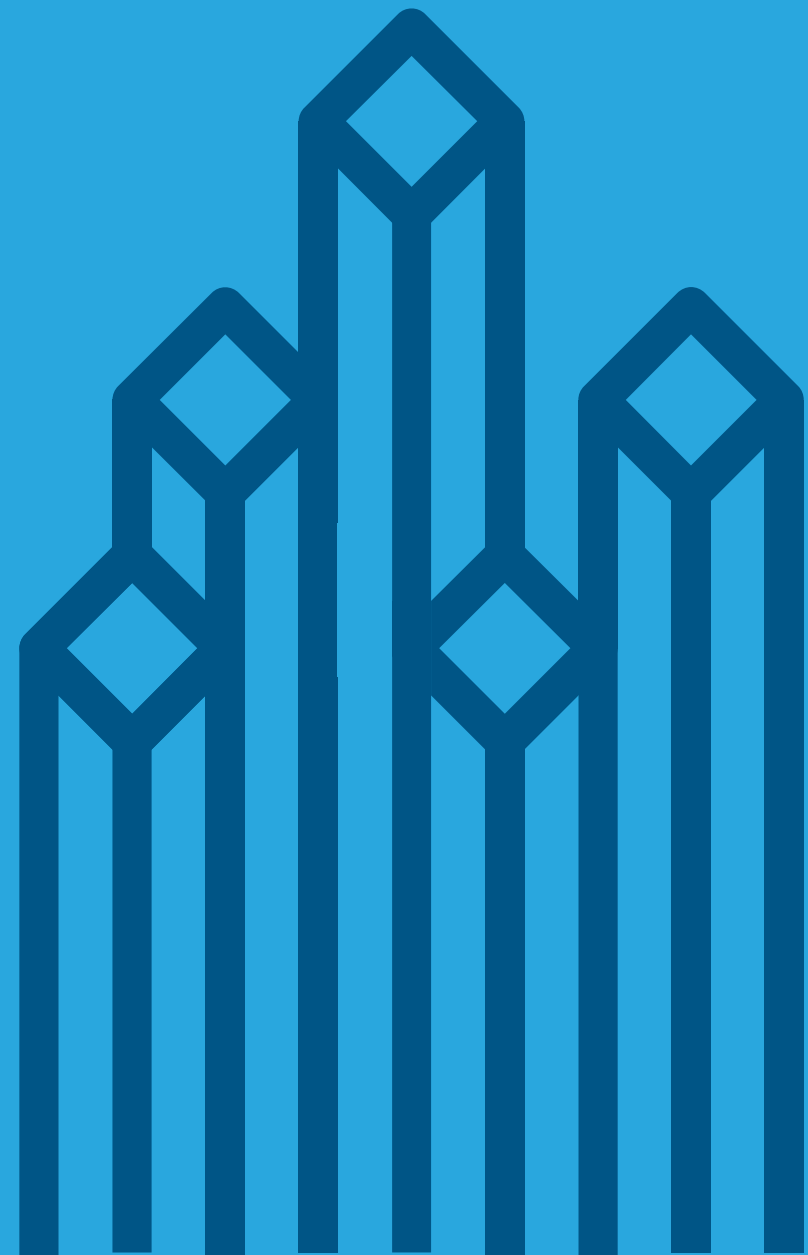




# Back Slides Hadoop and Spark Essentials and a few other topics

John Hope – Senior Solutions Engineer

[hope@cloudera.com](mailto:hope@cloudera.com)



Our relationship with data  
is **changing - forever.**

**Data** can be a  
powerful strategic asset

**...only if...**

data helps achieve  
your **business vision**.

# Data is Transforming Business

Drive Customer  
Insights +revenue



Improve Product & Services  
Efficiency -costs



Lower Business  
Risk



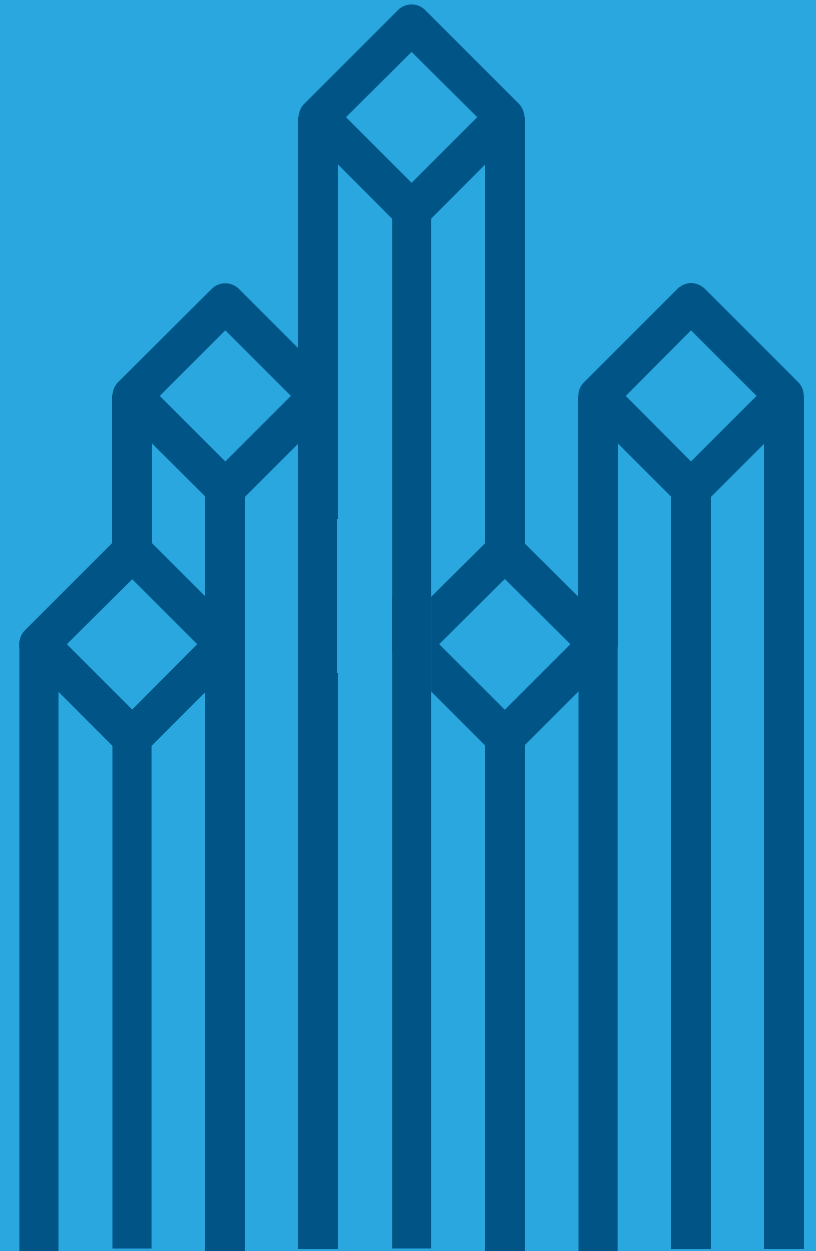
# Exploring Use Cases

- ETL Offload
  - Too much data, too little time, too costly
- Active Archive
  - Save all data vs. moving to slow archival storage
- Mainframe Migration
  - Move costly CPU loads
- Real-time streaming
  - Fraud detection, patient care, transactions
- Data Discovery
- Search All Types of Data
- Predictive Analytics
- Scalable BI
- 360 View of xyz
  - Eliminate siloed data
- Anomaly detection
- PB-scale platform for cyber security
- Behavioral analytics
- Multi-tenant / shared resources



# Data Management

Yesterday and Today

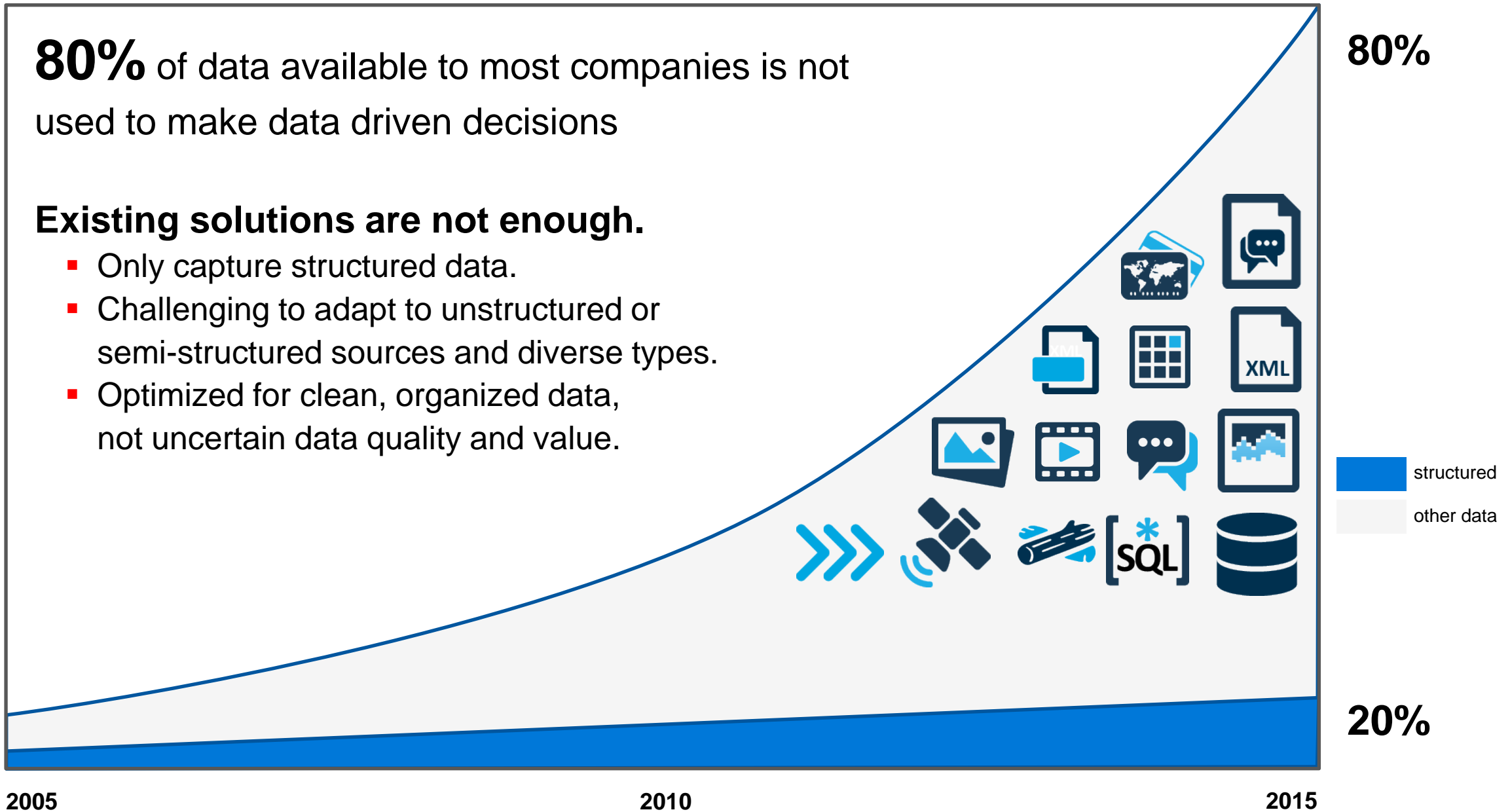


**80%** of data available to most companies is not used to make data driven decisions

**Existing solutions are not enough.**

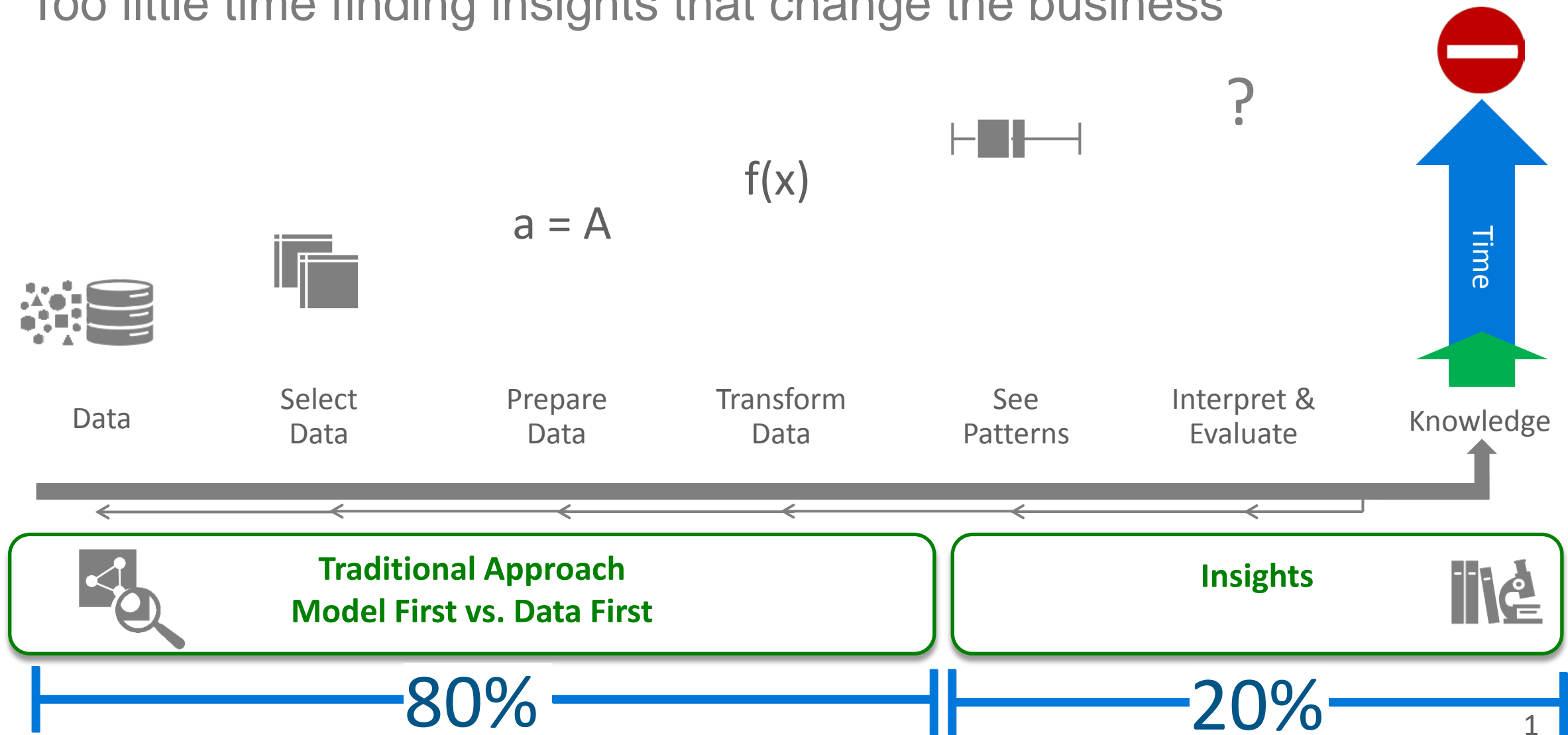
- Only capture structured data.
- Challenging to adapt to unstructured or semi-structured sources and diverse types.
- Optimized for clean, organized data, not uncertain data quality and value.

Volume of Data Generated



# Discovery & Data Management Lifecycle

Too little time finding insights that change the business



# Inverting Data Access Cycles

What if we could make data preparation 20% of the effort so you can focus 80% of your time on executing and improving your business?



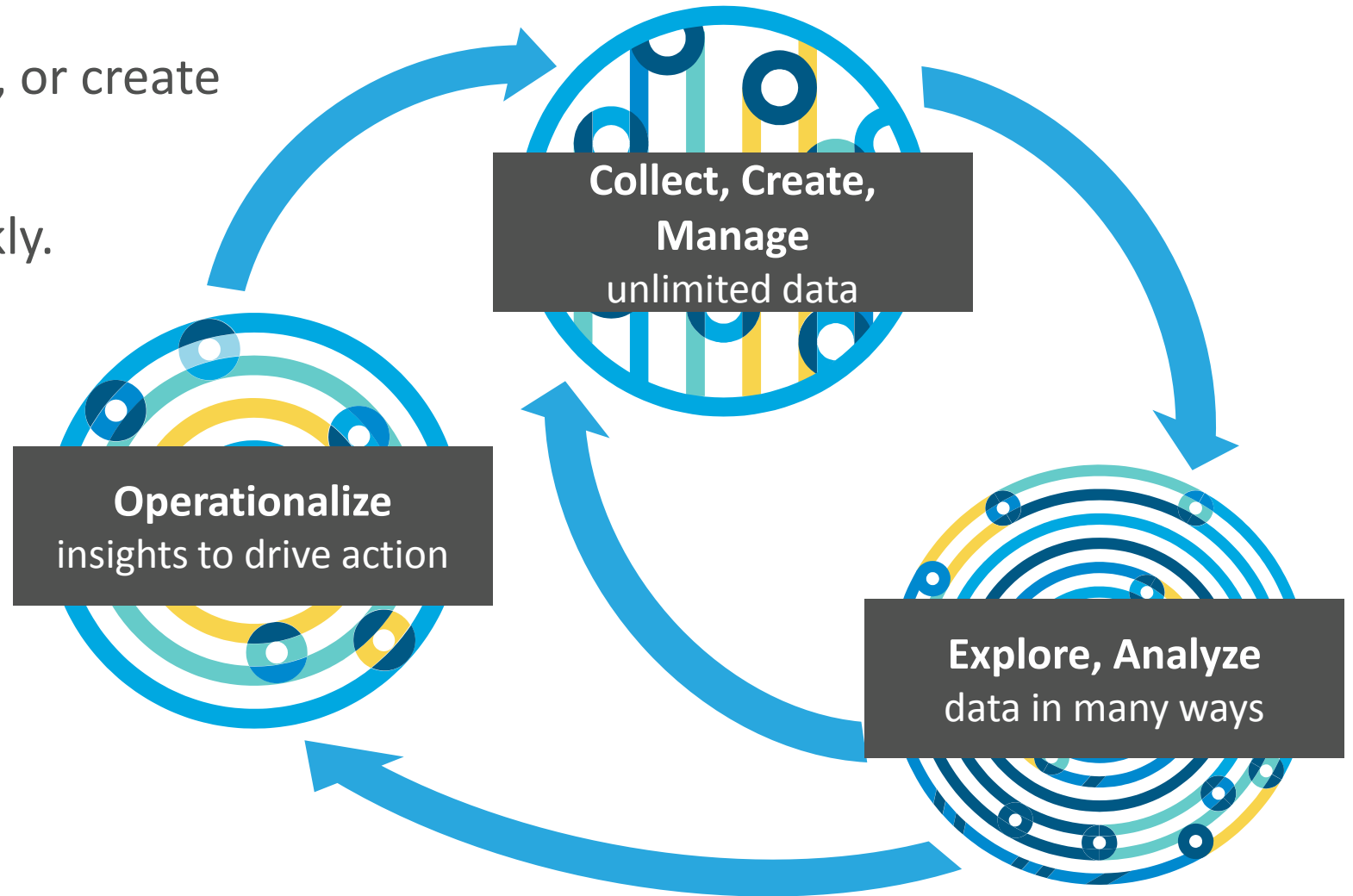
# Adopt an Agile Approach

Successful projects start small, fail often, and iterate to success

1. **Get data** you already have, or create new data.
  2. **Explore and analyze**, quickly.
  3. **Deploy** your application.
- ...and repeat

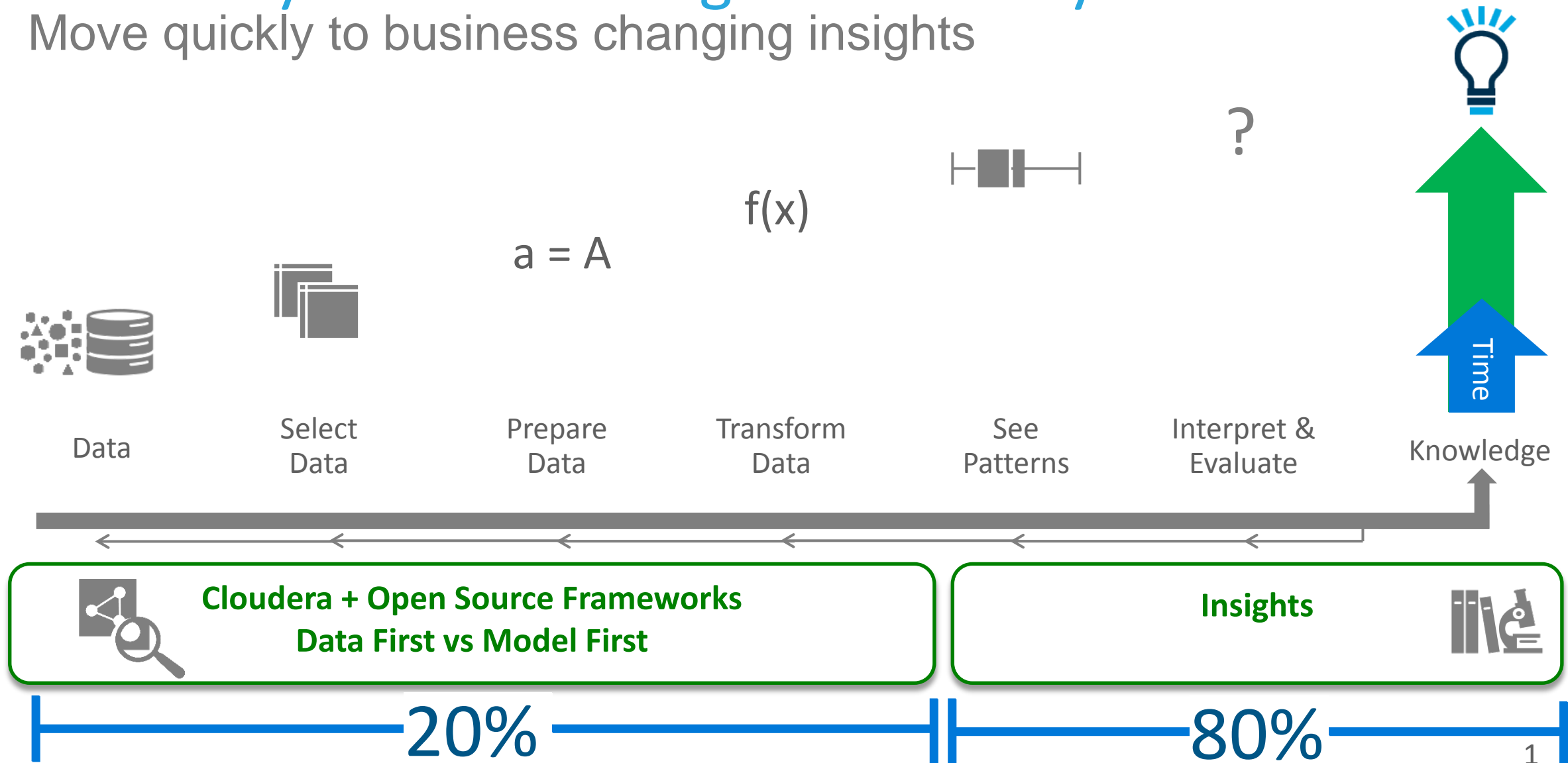
## Add:

new data sources, more users, more use cases, more complex analytics, go real-time



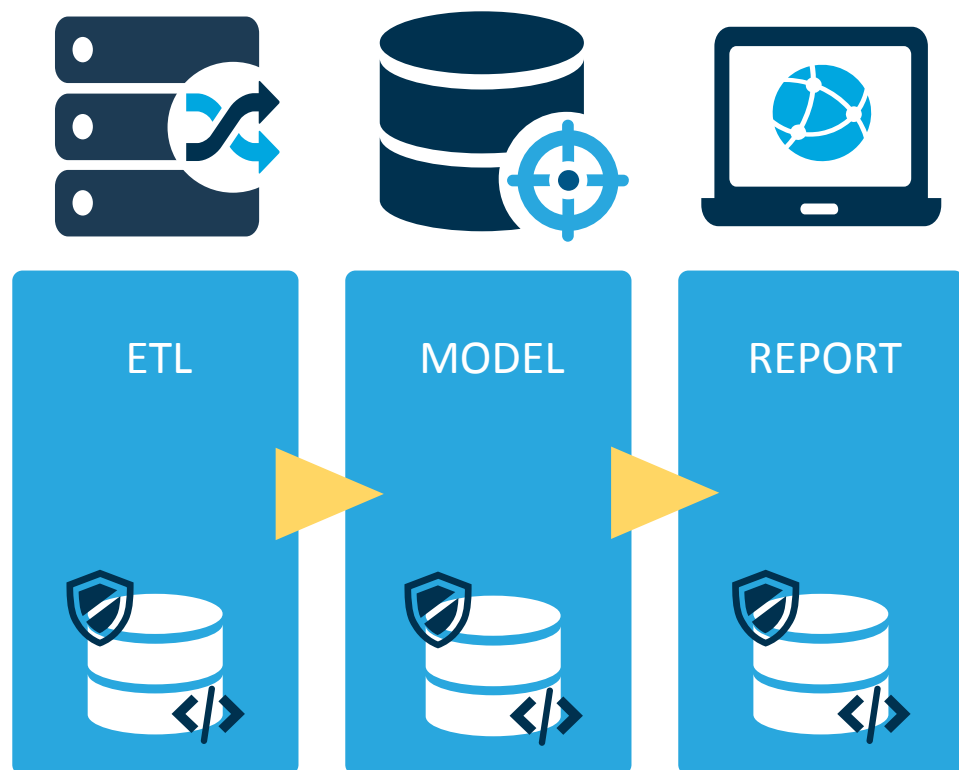
# Discovery & Data Management Lifecycle

Move quickly to business changing insights



# Current Data Management Architectures

Limited data. Single access. Platform silos.



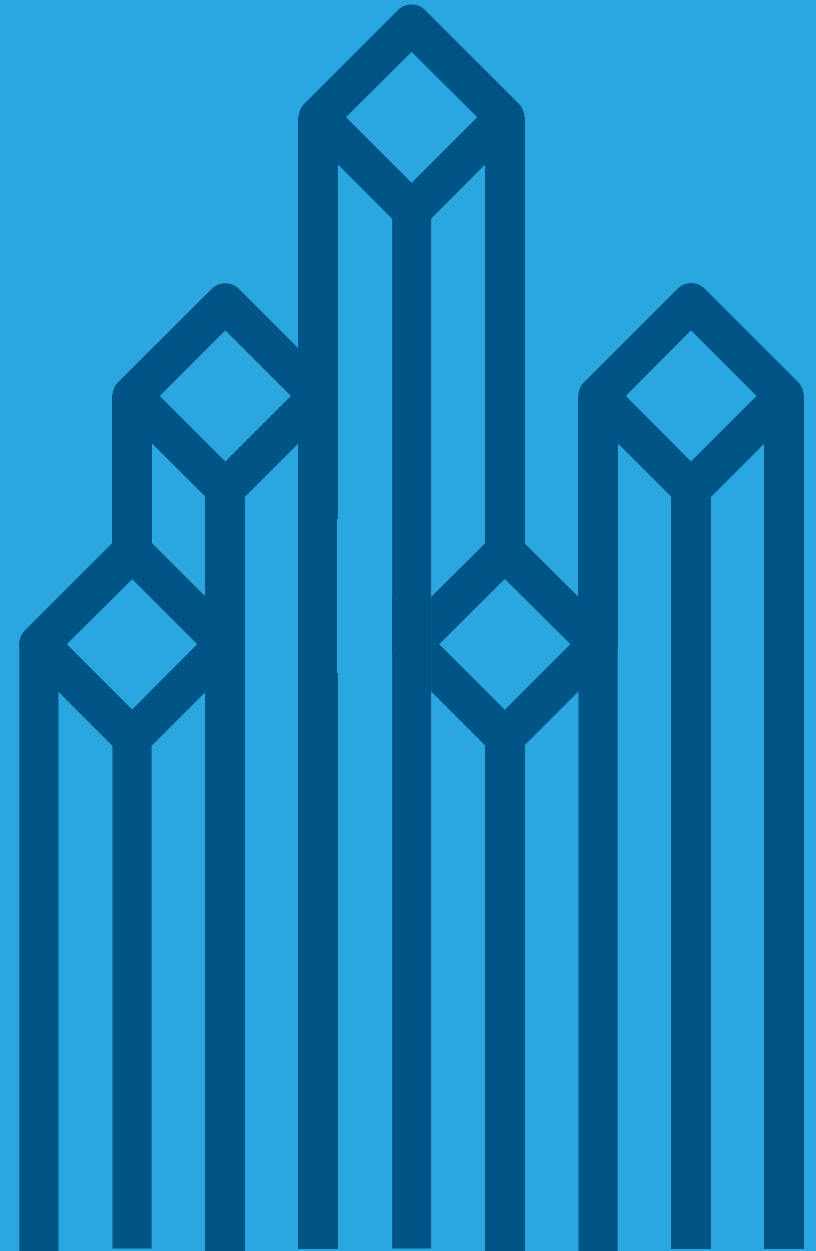
*20% of Available Data Used Today*

- ❌ Limited Data
- ❌ Structured Only
- ❌ Slow Performance
- ❌ Restricted Use
- ❌ Difficult Redundancy
- ❌ Sometimes SPOF
- ❌ Not Real-Time
- ❌ Many more.....



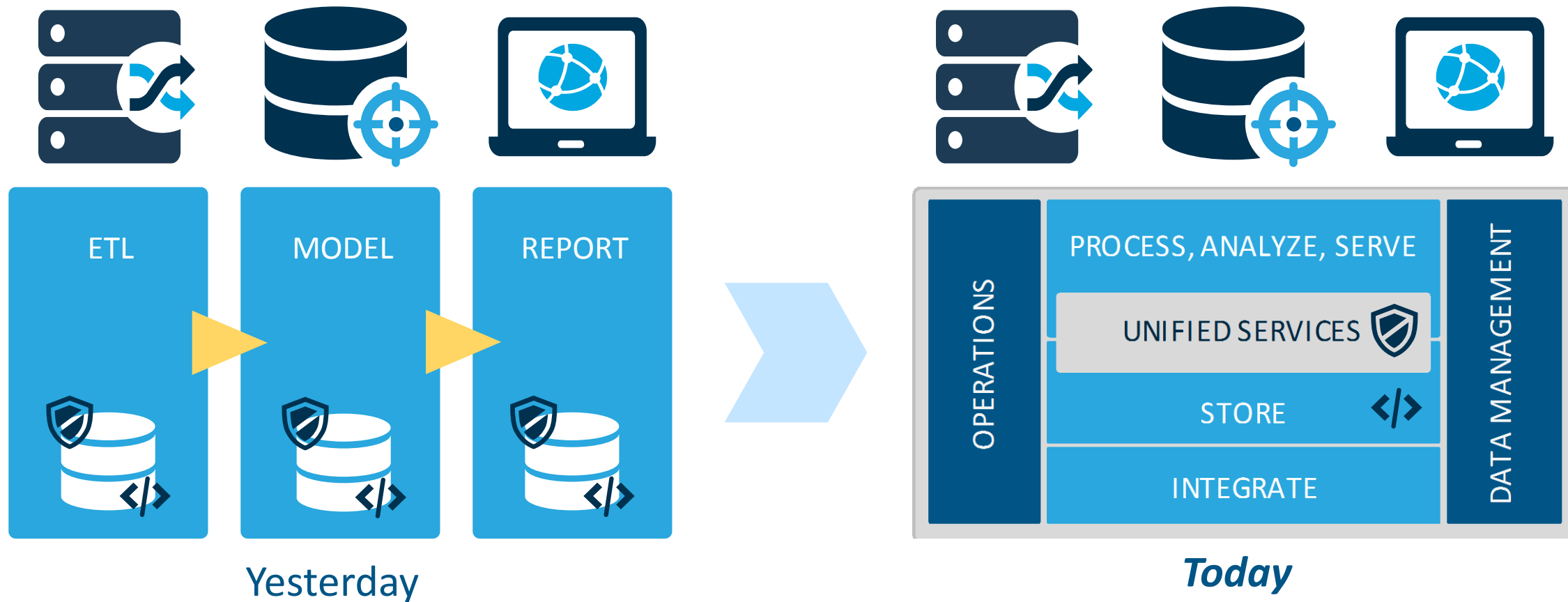
# The Cloudera Solution

Fast, Easy and Secure



# Cloudera Enterprise Data Hub

Unlimited data. Diverse access. One platform.



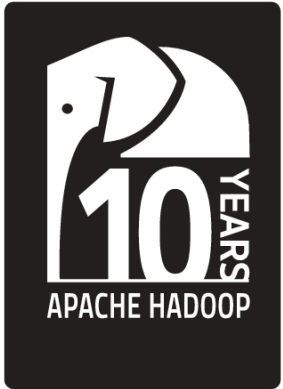
# Cloudera Enterprise Data Hub

Unlimited data. Diverse access. One platform.

All Data  
Sources

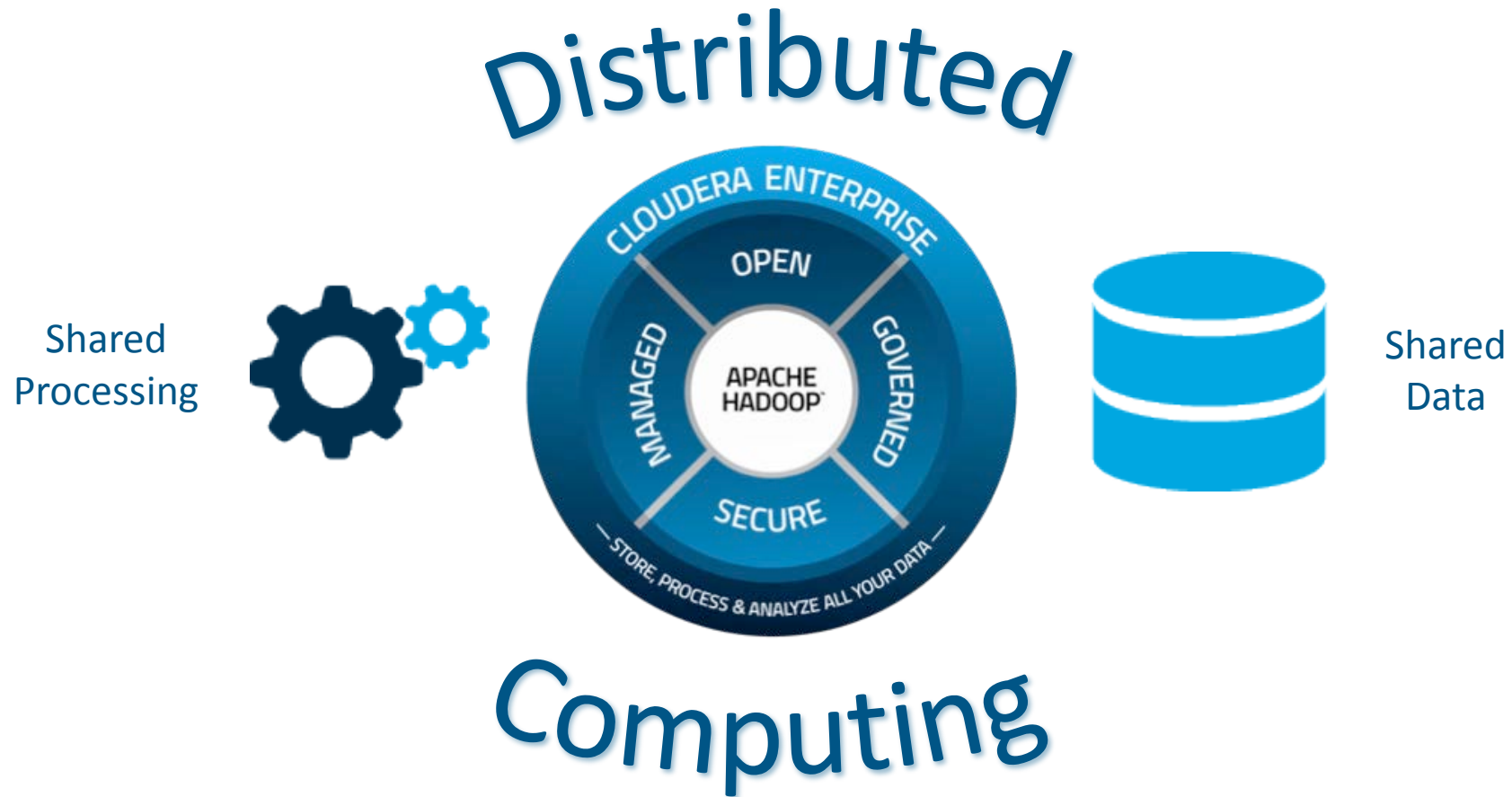


100% of Data +  
Unlimited History



# Cloudera Enterprise Data Hub

How does this stuff work?.



# Cloudera Enterprise Data Hub

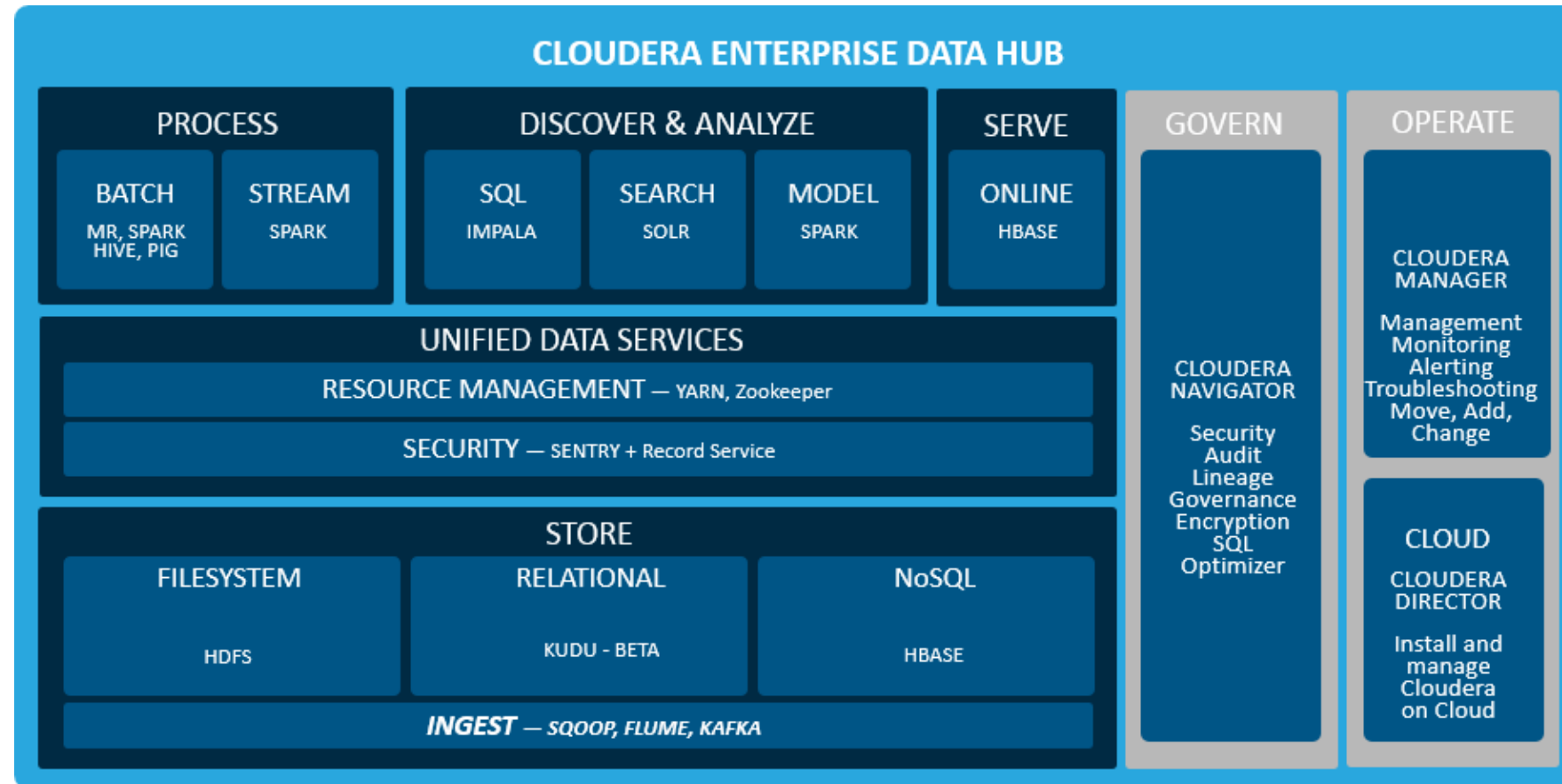
Making Hadoop Fast, Easy, and Secure for the Modernized Architecture

Hadoop is a new kind of data platform.

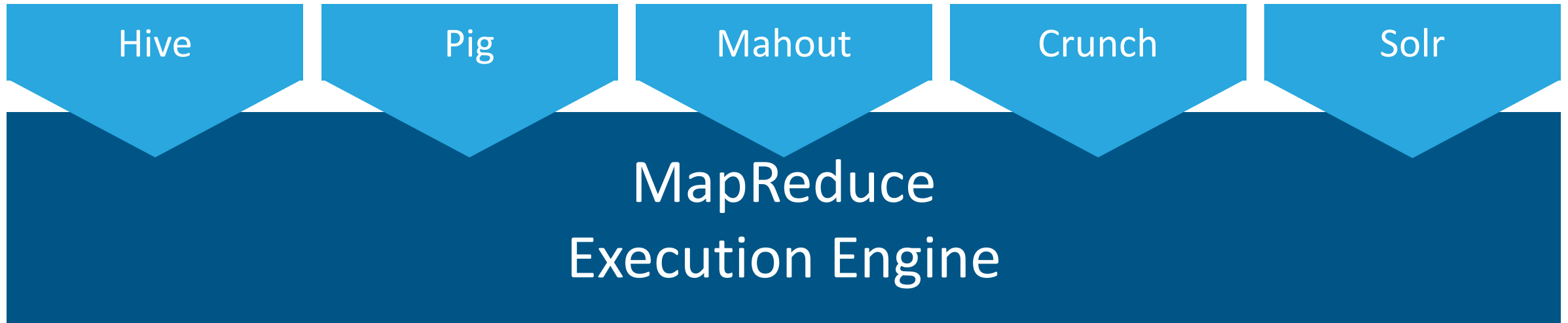
- One place for unlimited data
- Unified data access

Cloudera makes it:

- **Fast** for business
- **Easy** to manage
- **Secure** without compromise



# MapReduce: A great tool for its day



The original scalable, general, processing engine of Hadoop ecosystem

- Useful across diverse problem domains
- Fueled initial ecosystem explosion

# Enter Apache Spark

Flexible, in-memory data processing for Hadoop

## Easier Development

- Rich APIs for Scala, Java, and Python
- Interactive shell

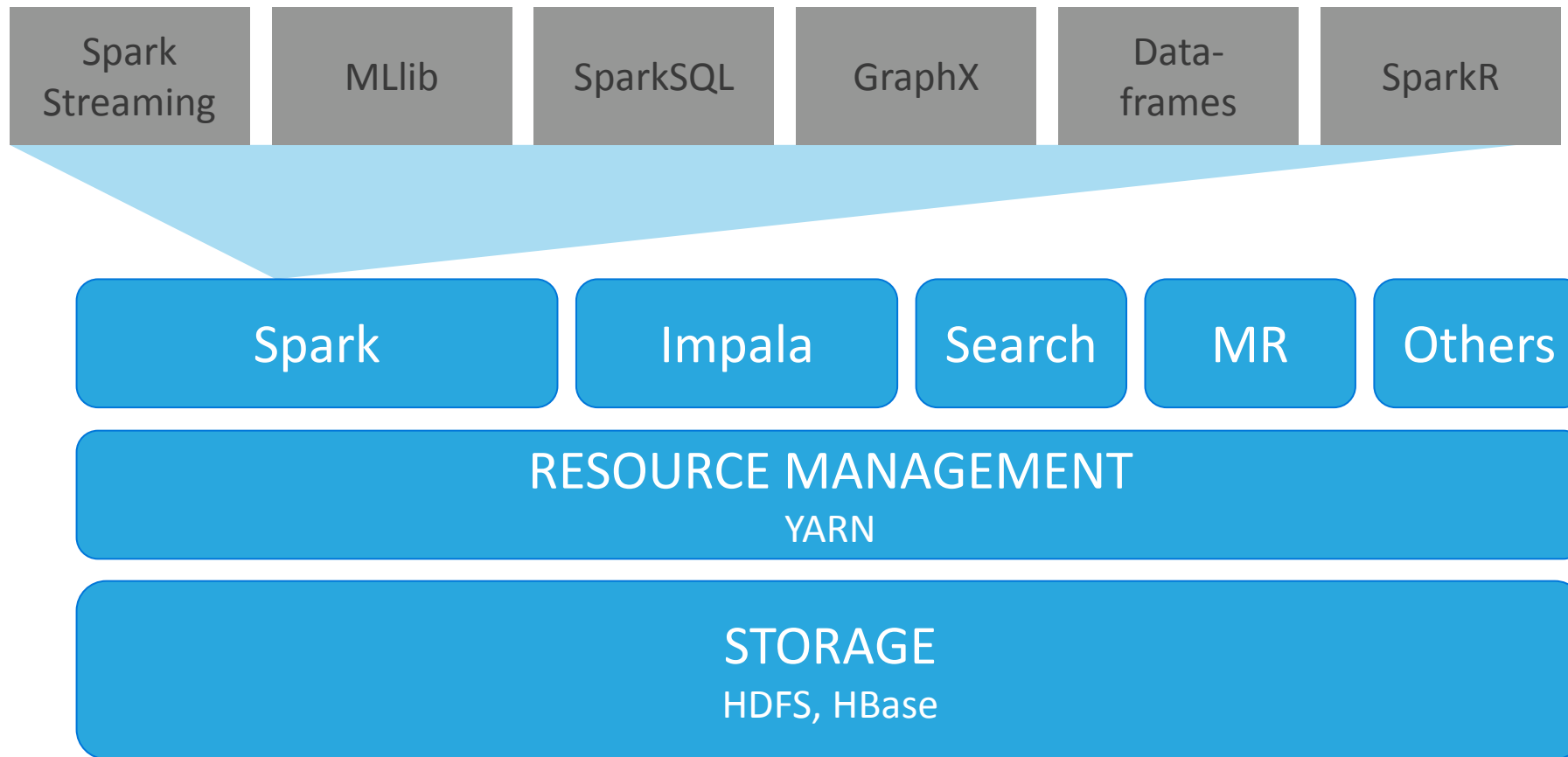
## Flexible, Extensible API

- APIs for different types of workloads:
  - Batch
  - Streaming
  - Machine Learning
  - Graph

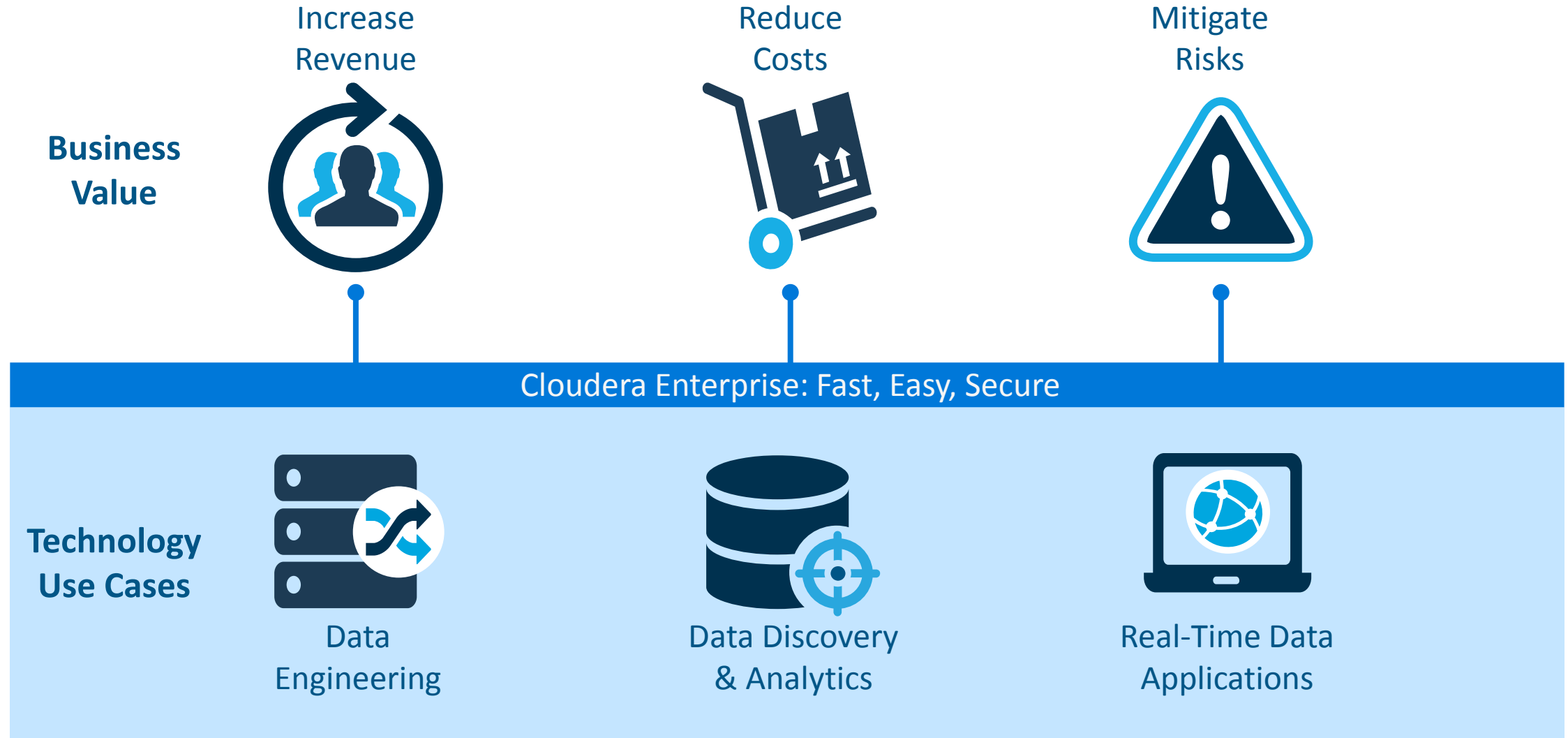
## Faster Processing (Batch & Streaming)

- In-Memory processing and caching

# The Spark Ecosystem & Hadoop



# One platform. Many applications. Future Proof.



# Apache Hadoop

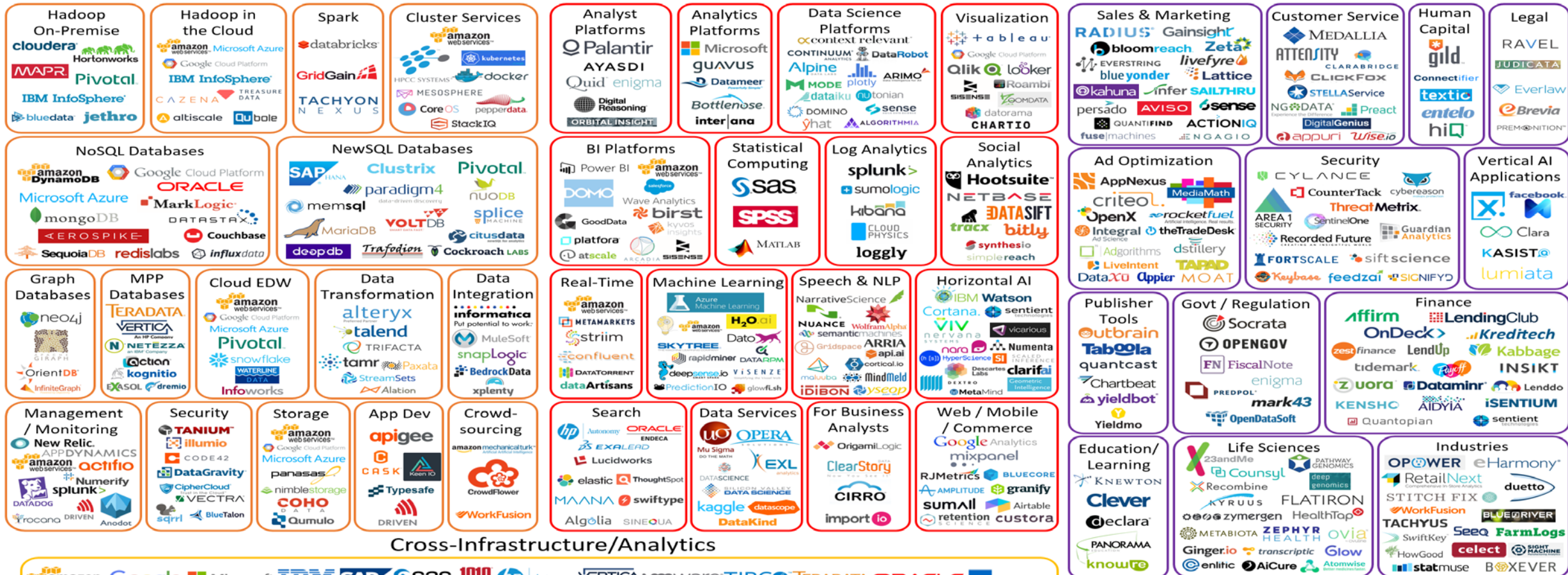
... This will change before we leave the room.

# Big Data Landscape 2016 (Version 3.0)

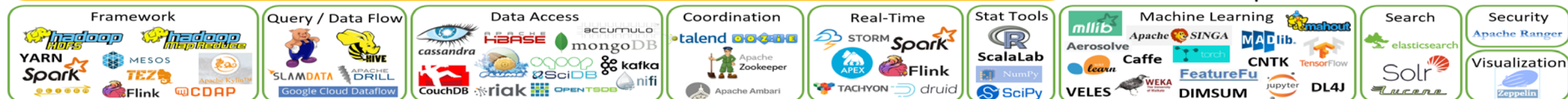
## Infrastructure

## Analytics

## Applications



## Open Source




## Data Sources & APIs



# Hadoop Isn't Just Hadoop Anymore



# Hadoop Isn't Just Hadoop Anymore

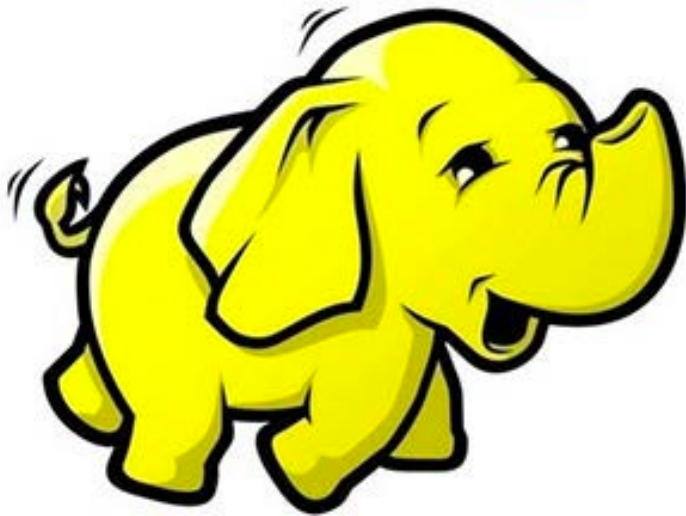


Year	Core Hadoop (HDFS, MR)	2008	2009	2010	2011	2012	Present
2006-07	Core Hadoop (HDFS, MR)						
2008		HBase ZooKeeper					
2009			Hive Mahout				
2010				Sqoop Whirr Avro Hive Mahout			
2011					Flume Bigtop Oozie MRUnit HCatalog Sqoop Whirr Avro Hive Mahout HBase ZooKeeper		
2012						Spark Tez Impala Kafka Flume Bigtop Oozie MRUnit HCatalog Sqoop Whirr Avro Hive Mahout HBase ZooKeeper	
Present							Kudu RecordService Spark SparkSQL Parquet Sentry SparkSQL Tez Impala Kafka Flume Bigtop Oozie MRUnit HCatalog Sqoop Whirr Avro Hive/HoS Mahout HBase ZooKeeper

# Apache Hadoop

---

# ***hadoop***



Scalable

---

Flexible

---

Open (future proof)

---

Cost-Effective

