Preparing your data sets to be more open

The Why, What, and How

Al Domingo, Director, Sales Engineering

cloudera

Our relationship with data is **changing**.

cloudera

Data is Transforming Government

Increase Transparency

Reduce Waste

Manage Security, Risk & Compliance

Opportunities Across Pillars

Transportation

Predictive modeling,
preventative maintenance

Education

Data-driven instruction,
understand trends sooner

Public Safety

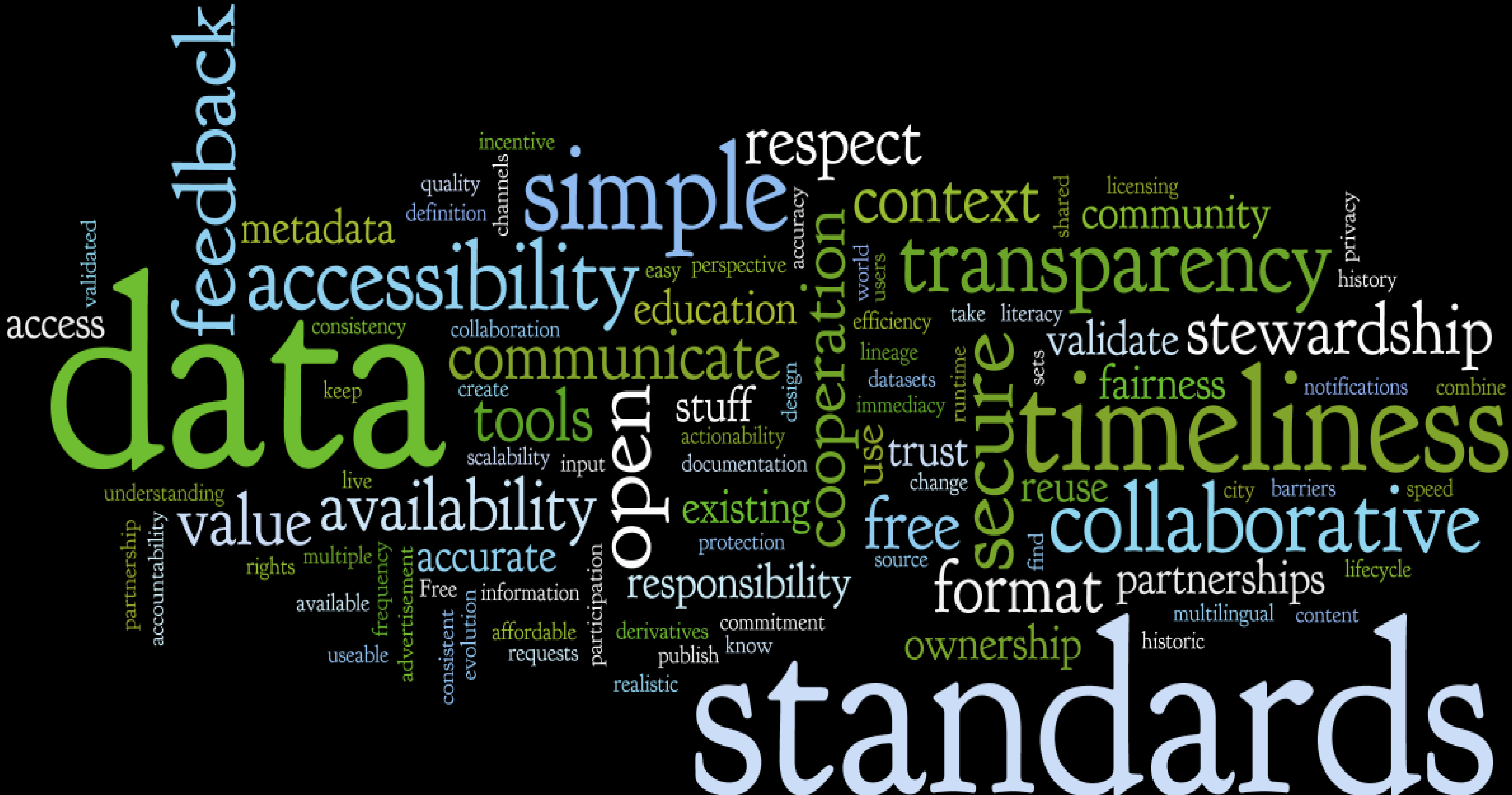
Prevent and solve crime,
Fire risk profiling

Department of Revenue

Increase collections,
maximize revenue

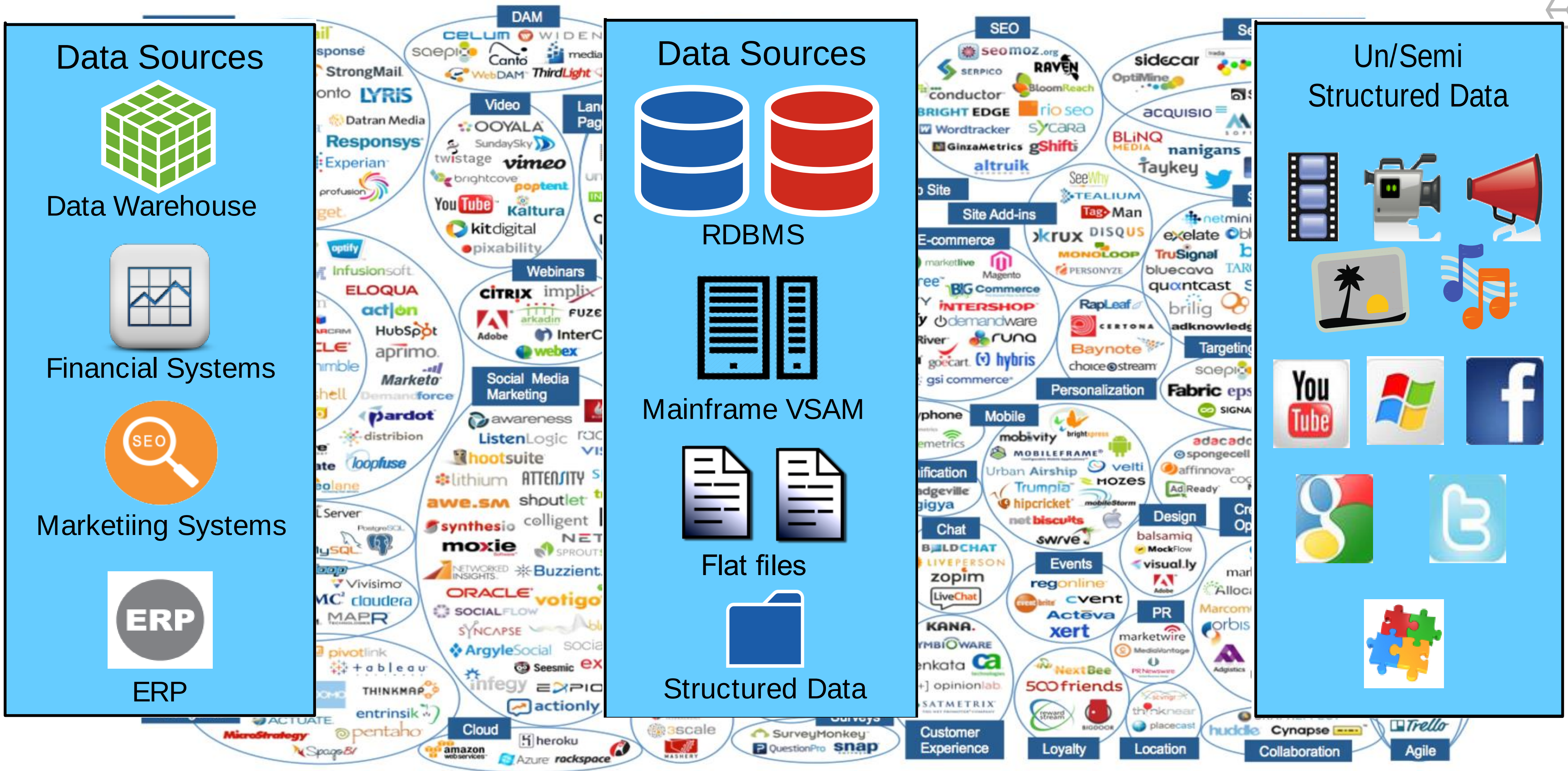
Human Services

Citizen 360, improve services,
reduce fraud


feedback
incentive
respect
quality
channels
simple
definition
accuracy
context
shared
licensing
community
privacy
metadata
validated
accessibility
easy
perspective
cooperation
world
users
transparency
history
access
data
consistency
collaboration
education
efficiency
take
literacy
validate
stewardship
communicate
lineage
runtime
sets
secure
keep
create
design
datasets
fairness
notifications
combine
tools
stuff
immediacy
timeliness
open
actionability
use
trust
scalability
input
documentation
live
change
understanding
reuse
city
barriers
speed
value
availability
existing
free
collaborative
partnership
accountability
rights
multiple
frequency
advertisement
accurate
protection
find
source
lifecycle
participation
responsibility
format
partnerships
Free
available
information
consistent
evolution
multilingual
content
affordable
derivatives
commitment
historic
useable
requests
publish
know
ownership
realistic
standards

The Challenges

Data is everywhere



Data Sources

Data Warehouse

Financial Systems

Marketiing Systems

ERP

Data Sources

RDBMS

Mainframe VSAM

Flat files

Structured Data

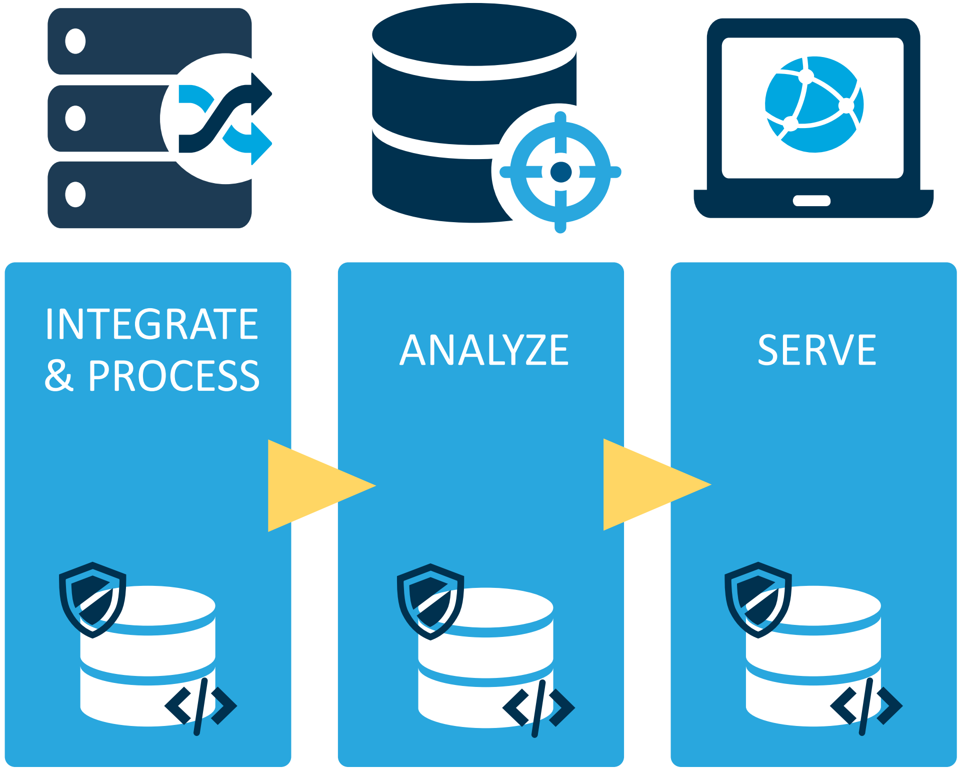
Un/Semi Structured Data

by Scott Brinker @chiefmartec http://www.chiefmartec.com

cloudera

Current Data Architectures

Limited data. Single access. Platform silos.



cloudera

Holistic Insights are Difficult

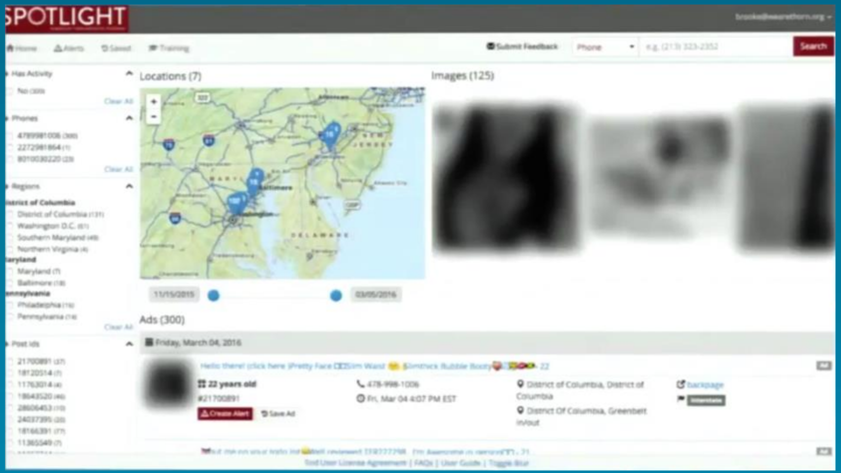
Disconnected & Heterogeneous Systems of Record

Offline Analysis in Multiple Systems

Long Delays to Insights

cloudera

Thorn: Drive Tech Innovation to Fight Child Sexual Exploitation



The Challenge:

- Predators exploiting technology to exploit vulnerable children
- Too much data to investigate
- Disrupt platforms that enable abuse
- Accelerate Victim Identification , get them help

The Solution: **SPOTLIGHT**

- Tool to provide intelligence and leads on suspected human trafficking networks
- Leverage technology to analyze online classified ads to ID crimes and victims

The Results:

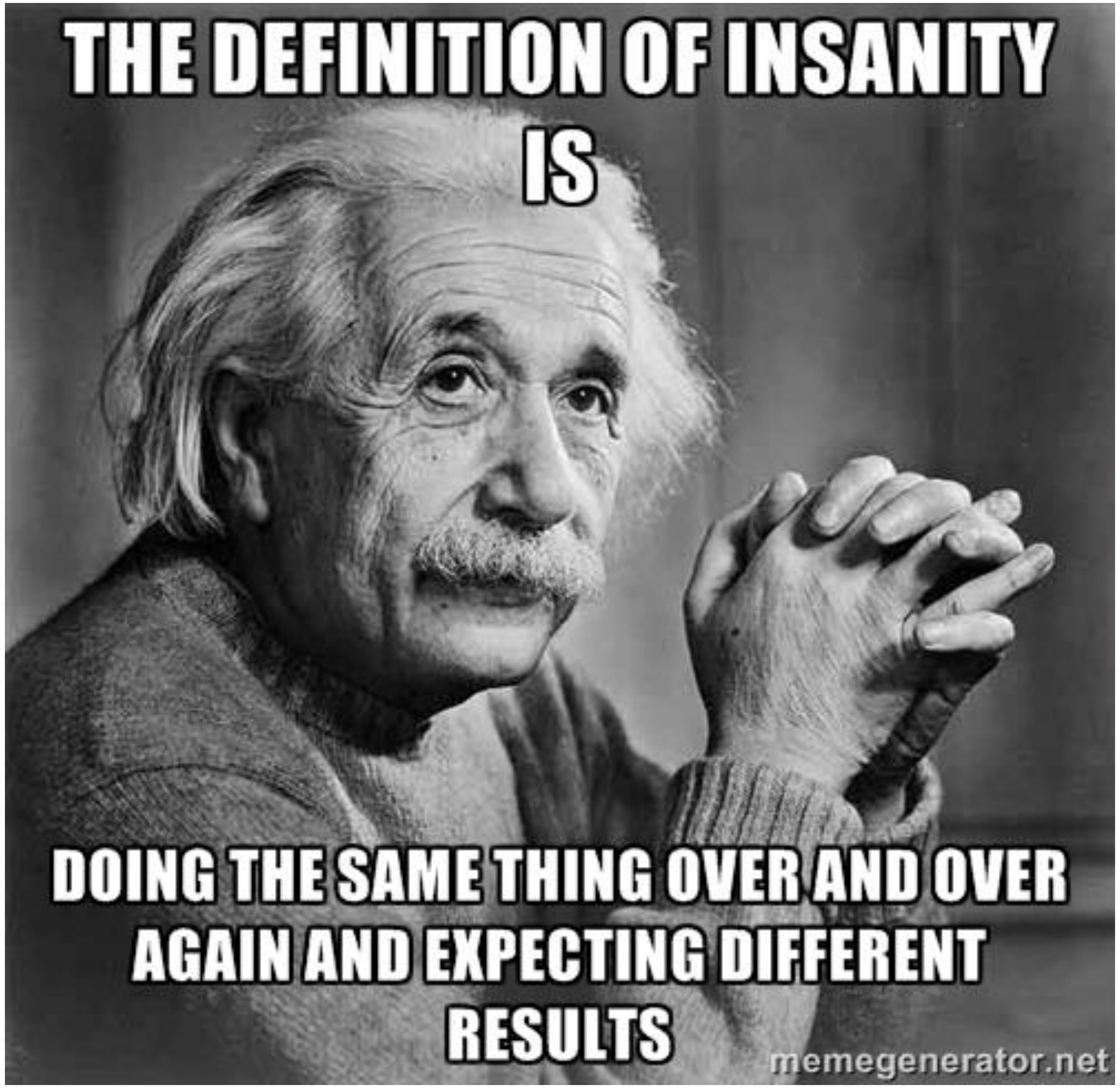
- Used in 860 trafficking cases
- Identify over 300 victims include 50 children
- 1,300 Law Enforcement across 46 states

cloudera

Where Do I Start?

Get the right architecture!

Insanity


THE DEFINITION OF INSANITY
IS
DOING THE SAME THING OVER AND OVER
AGAIN AND EXPECTING DIFFERENT
RESULTS
memegenerator.net

What is Required?

Extreme performance and efficiency

1. Economically feasible to store more data
2. Powered to predictably process large data sets
3. Ability to build your data asset at linear scale

Analytic agility

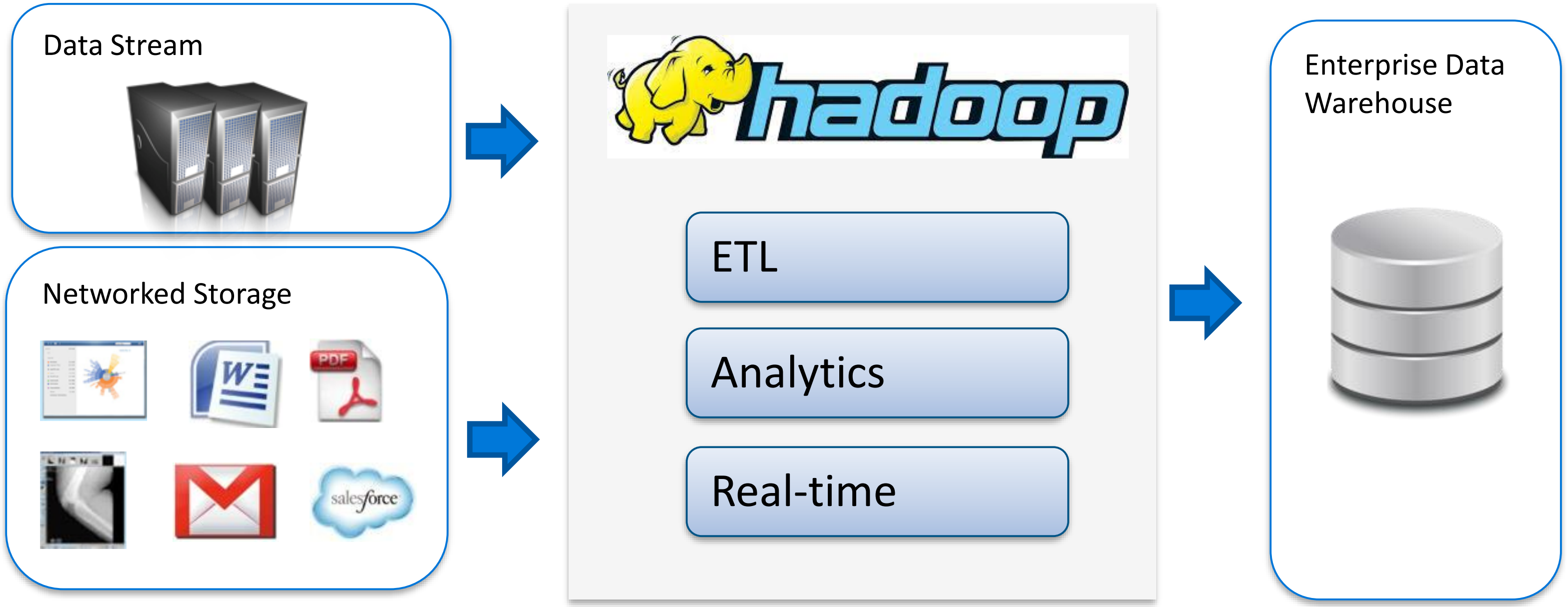
4. Collect data in native format – enables agility
5. Build history of activity by collecting data prior to its use
6. You can have near real-time access to data, plus a view of history
7. Create community data by sharing across Agencies, Jurisdictions and Orgs
8. Out-of-the-box thinking and fail-fast increases innovation

cloudera

Apache Hadoop

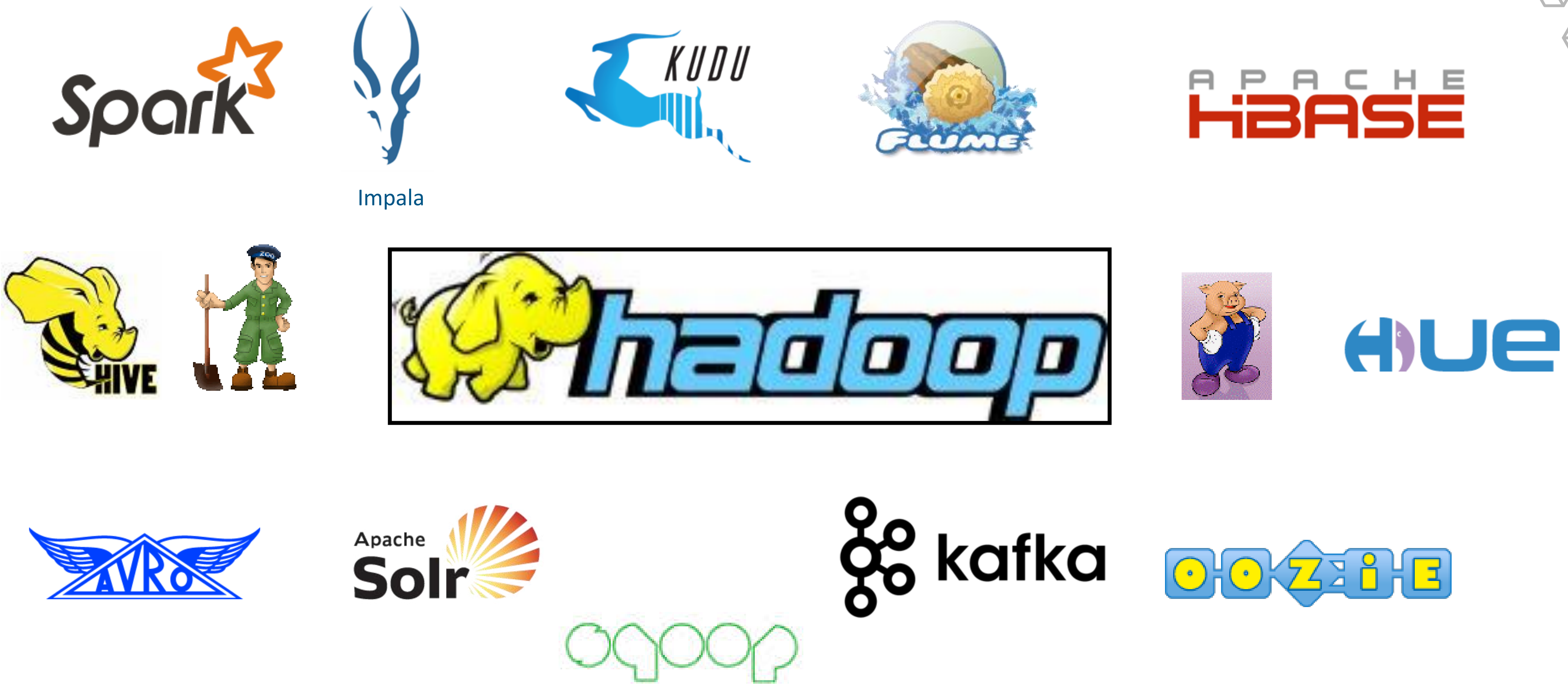
- Hadoop is a software framework for storing, processing, and analyzing “big data”
  - Handle Many Data Formats
  - Distributed
  - Scalable
  - Fault-tolerant
  - Open source
- Hadoop is based on work done at Google in the late 1990s/early 2000s

Apache Hadoop is the Platform



- Hadoop cluster is used for centralizing all data
- Structured and unstructured data is moved into the cluster
- Once processed, data can be analyzed in Hadoop or moved to the EDW

A Large (and Growing) Ecosystem



Impala

Hadoop in the Real World

cloudera

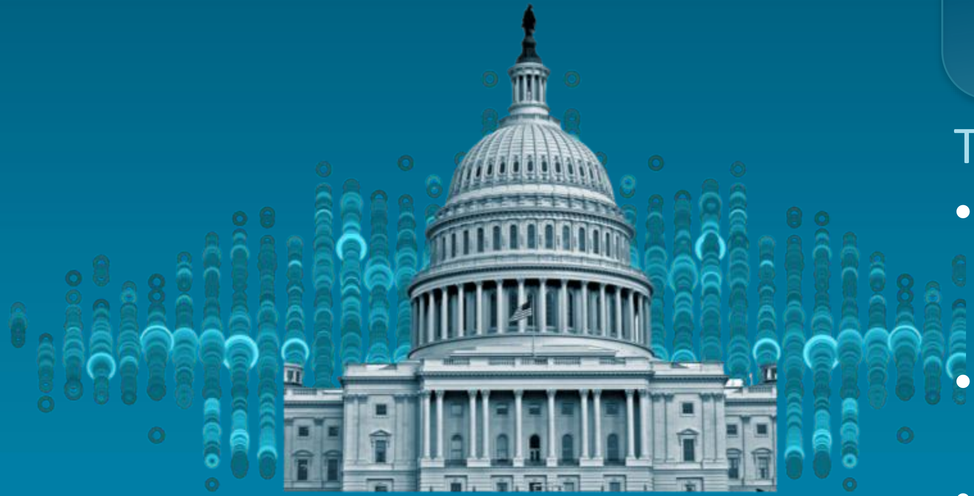
Popular Use Cases

- Data Processing
- Business Intelligence – Query & Analysis
- Predictive Analytics
- Enterprise Data Hub
- Low-cost Storage of Large Data Volumes
- Cyber Security

Cloudera is driving new insights

The Challenge:

- Suspicious activity across global web must be identified & made available to 700 commercial & federal organizations
- Database is meeting scalability & performance limitations



National security organization offers real-time information, warnings and guidelines that strengthen our ability to protect against cyber attacks.

The Solution:

- Cloudera Enterprise + Sherpasurfing: PB-scale platform for cyber security analytics
- Integrated: HP ArcSite, IBM Netezza, Tableau, Centrifuge
- Offering real-time data & warnings

Government Revenue Service

The Challenge:

- Estimated that 7% of 506B Pounds are not collected£506B annual collection data is spread across 12 data warehouses
- Unacceptably slow data processing, access and analysis

The Solution:

- Enterprise Data Hub that will support 1 PB of Data for Analytics and Digital ambitions to improve yield and productivity across the Government Revenue Service.

The Results:

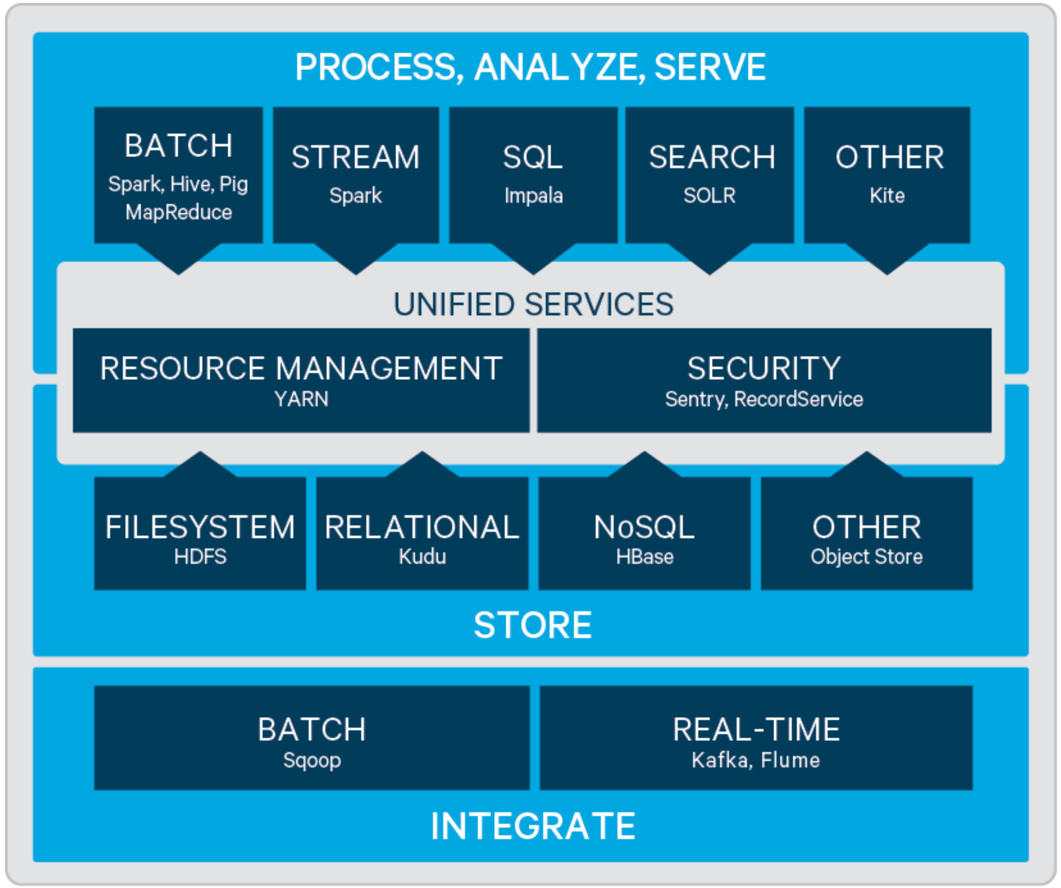
- Supported Data Analytics and Digital ambitions to improve yield and productivity.
- View the complete taxpayer journey
  - How long did it take us to respond to that letter?
  - Creates ability to pre-populate tax returns

Why Cloudera?

- The leader in Apache Hadoop-based software and services
- Founded in 2008 by leading experts on Hadoop
  - Over 1000 employees
  - Global operations spanning over 20 countries
- Provides support, consulting, training, and certification for Hadoop users
- Employs committers to virtually every significant Hadoop-related project
- Many authors of industry standard books on Apache Hadoop projects
  - Tom White, Ted Malaska, Kathleen Ting, etc.

CDH

- CDH (Cloudera's Distribution, including Apache Hadoop)
- Open source, enterprise-ready distribution of Apache Hadoop and related projects
- The most complete, tested, and widely-deployed distribution of Hadoop
- Integrates all the key Hadoop ecosystem projects



cloudera

Vendor Integration

Applications

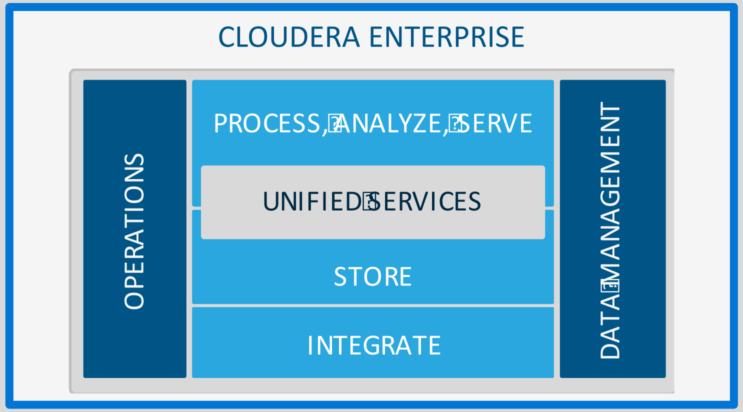
CASK · Datameer · Microsoft · MicroStrategy · ORACLE · pentaho · platfora · Qlik · SAP Partner · sas · Spotfire TIBCO Software · tableau · TRIFACTA · ZOOMDATA

Operational Tools

informatica · ORACLE · pentaho · talend open data solutions · syncsort

Data Systems

IBM · hp · mongoDB · ORACLE · SAP Partner · TERADATA


CLOUDERA ENTERPRISE

OPERATIONS

PROCESS, ANALYZE, SERVE

UNIFIED SERVICES

STORE

INTEGRATE

DATA MANAGEMENT

Infrastructure

amazon webservices · CenturyLink Technology Solutions · cisco · DELL · EMC² · FUJITSU · IBM SOFTLAYER · intel · hp · Google Cloud Platform · lenovo · Microsoft · NetApp · ORACLE · redhat · T··Systems· · vmware

System Integration

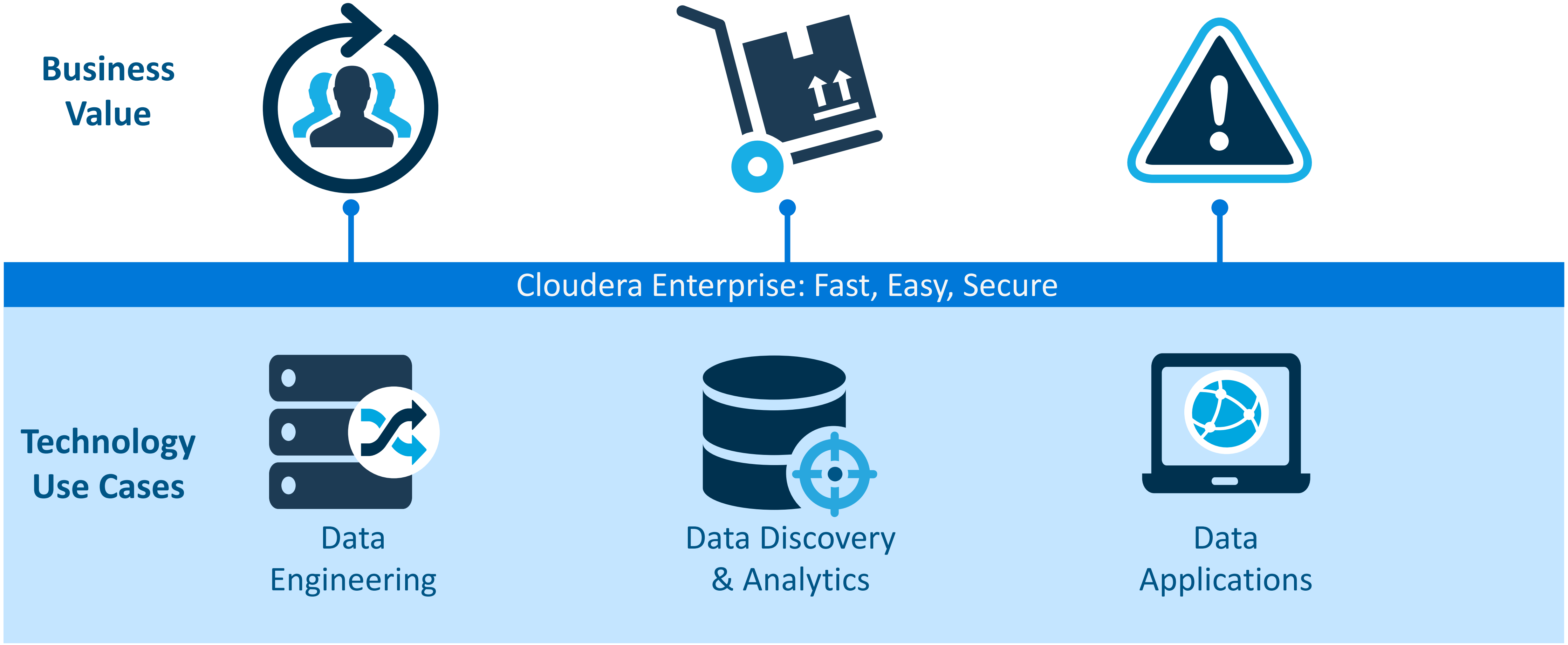
accenture · Capgemini · Cognizant · CSC · Deloitte. · IBM · Infosys · pwc · TATA · WIPRO

intel

cloudera

cloudera

One Platform. Many Applications.



Business Value

Drive Citizen & Mission Insights

Improve Product & Services Efficiency

Lower Risks

Cloudera Enterprise: Fast, Easy, Secure

Technology Use Cases

Data Engineering

Data Discovery & Analytics

Data Applications

cloudera

Summary

- Data is Transforming Government
- Many Opportunities across Pillars to Share and Leverage Data
- Hadoop is the Right Architecture
- Hadoop easily scales to store and handle all of your data
- More data means bigger questions, better answers
- Hadoop integrates with your existing datacenter components
- Cloudera is Open. Your data is Open.

More Information & Next Steps

Get Started

- Download C5.7
  - www.cloudera.com/downloads
- Release Notes
  - www.cloudera.com/documentation/enterprise/latest/topics/rg_release_notes.html
- Demo Videos
  - Hive-on-Spark: https://youtu.be/morvk4pI5OM
  - Cloudera Manager Cluster Utilization Reporting:
    - YARN https://youtu.be/szr7bUZ_kn8
    - Impala https://youtu.be/KYNgHbI04DY

Thank You