Here comes the flood
Tools for Big Data analytics

Guy Chesnot - June, 2012

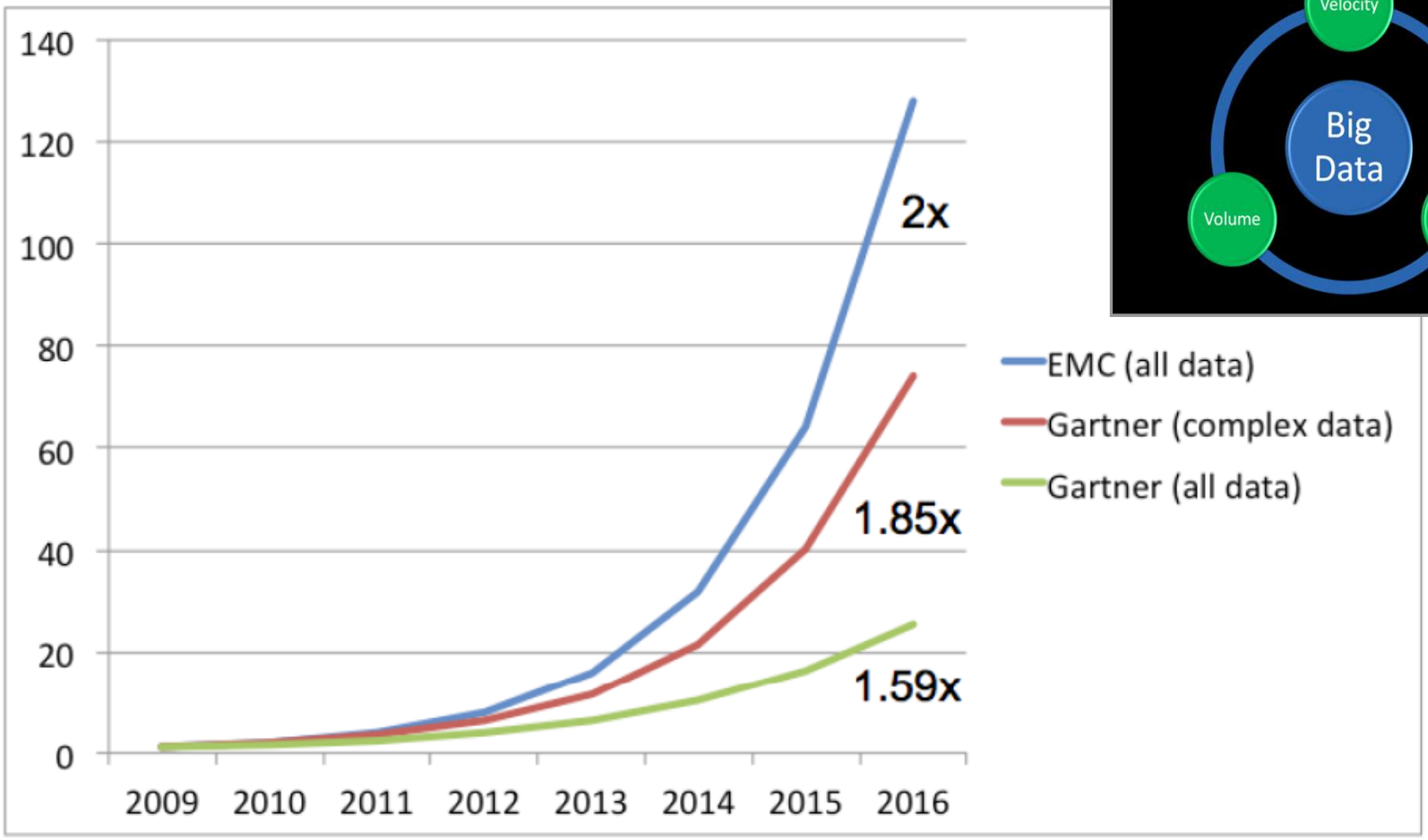
Agenda

- Data flood
- Implementations
- Hadoop
- Not Hadoop

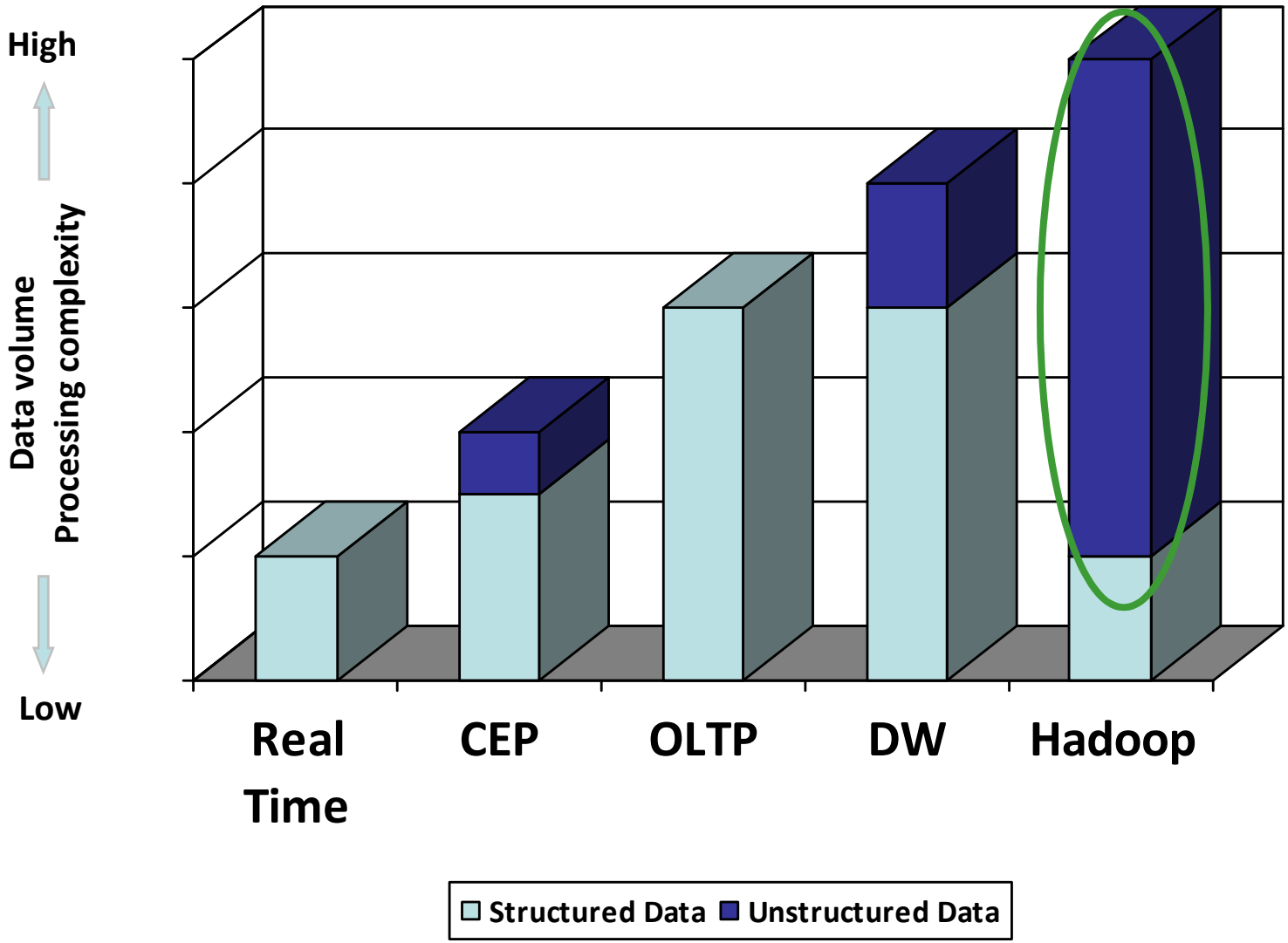
Agenda

- **Data flood**
- Implementations
- Hadoop
- Not Hadoop

Forecast Data Growth Rates



Computationally Intensive Distributed Data Analytics



Pride and Prejudice

Cloud ≡ Hadoop ≡ Big Data

Pride and Prejudice (cont.)

Cloud ≠ Hadoop ≠ Big Data

Agenda

- Data flood
- **Implementations**
- Hadoop
- Not Hadoop

Implementation

- **Several implementation levels**

- Application level
- Hardware: disk arrays
- Software layer, close to OS: « Cloud » OS, File system manager, in-between
  - Best choice
  - Efficiency
  - Feature rich
  - Not easy to develop

Implementation (cont.)

- Software layer very successful
  - Because of Open Source HADOOP
  - Other software products exist too
- Two main architectures at the file system manager level
  - Centralized metadata service (as in popular PFS) : Hadoop
  - Peer-to-peer model: metadata base is fully distributed

Hadoop is a widely used technology for Big Data processing

Advantages

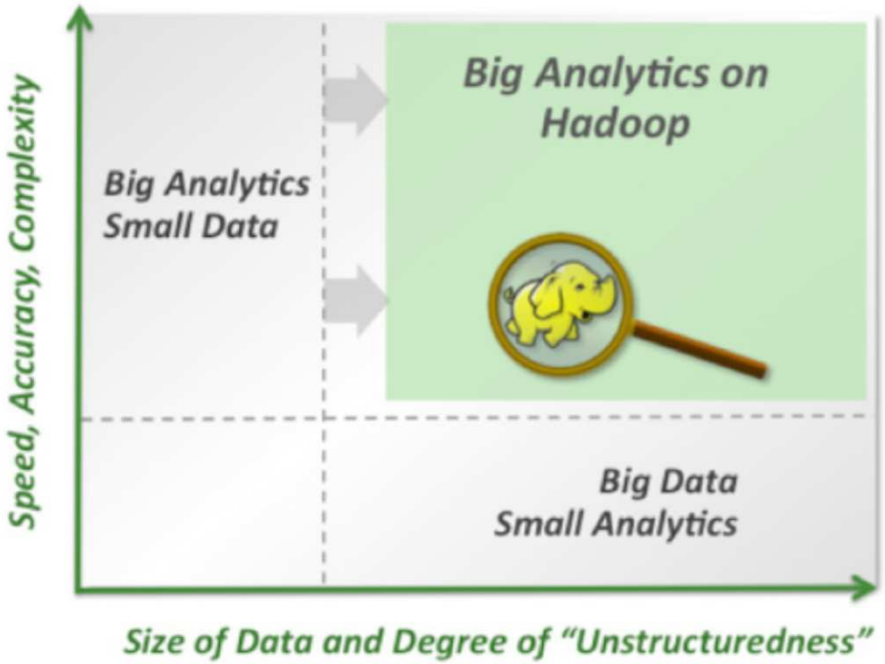
- Economics
- Flexibility
- Scalability

Challenges

- Raw Technology
- Complexity of deployment
- Requires significant resources
- No packaged Applications

Rapid Adoption

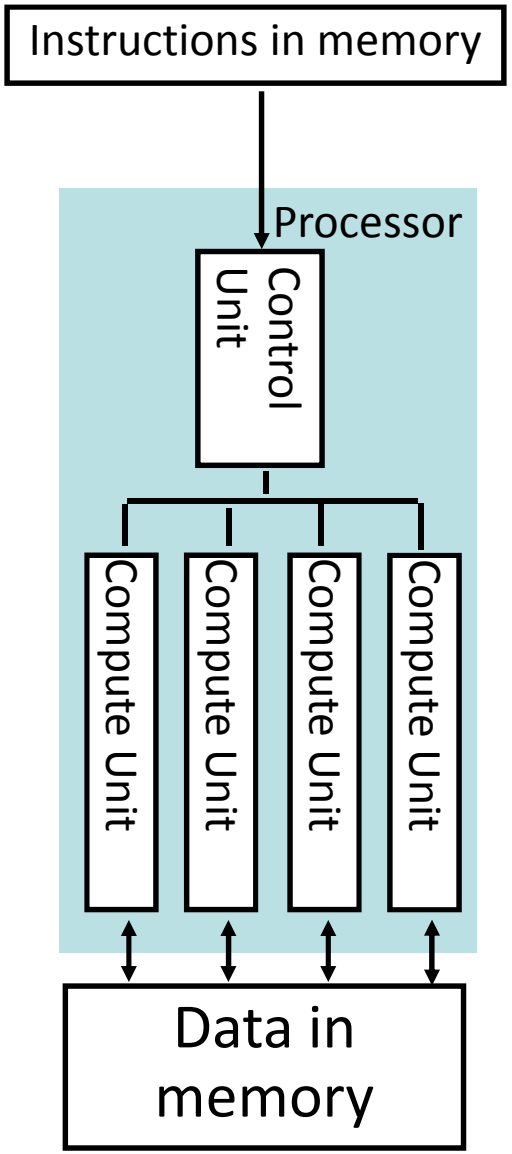
- Yahoo!, Facebook, eBay, Twitter
- JPMC, Schwab
- GAP, Walmart
- CIA
- Many more….

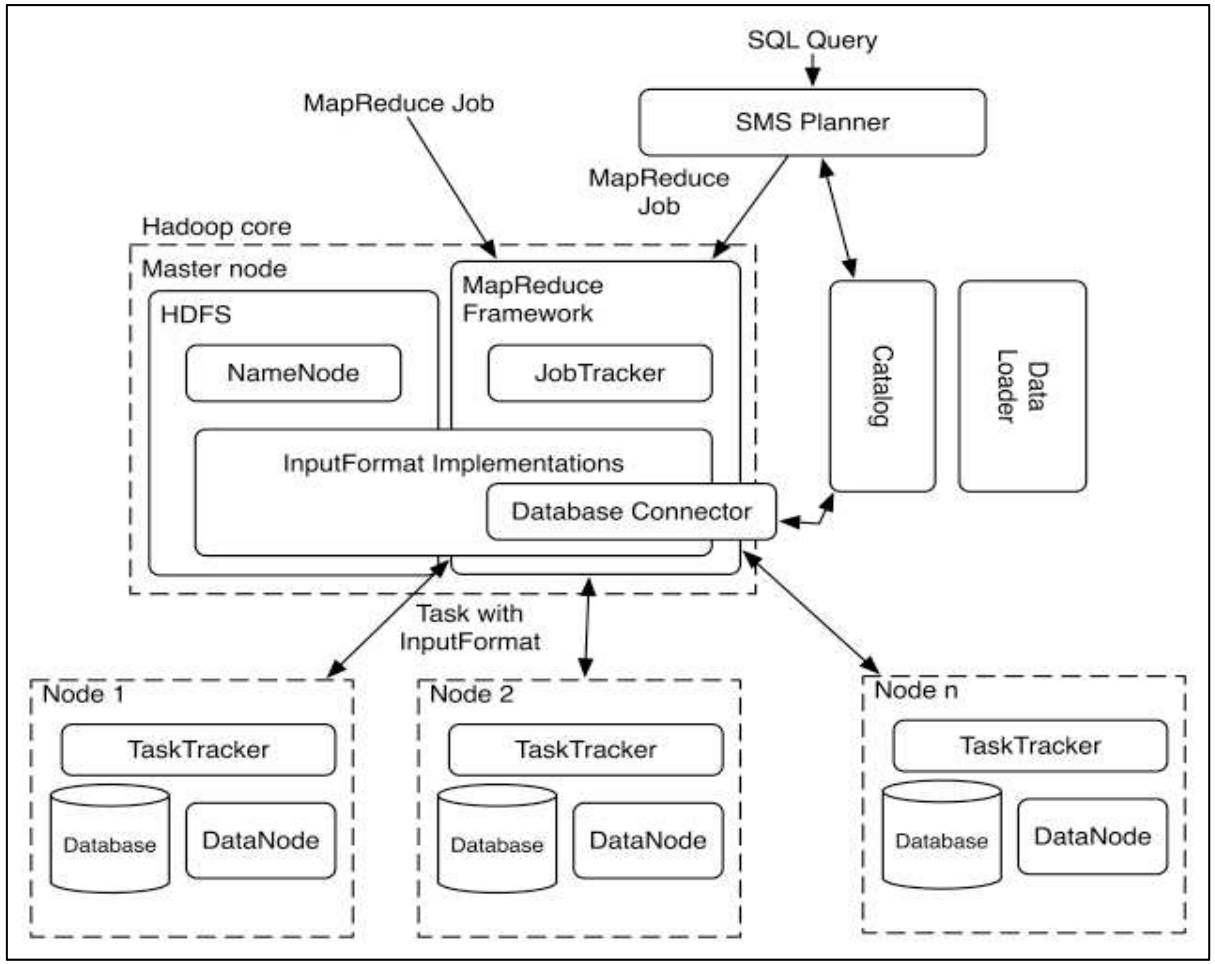


Hadoop adoption impetus is greatest when projects combine "Big Analytics" (fast, comprehensive analysis of complex data) and massive, unstructured data sets.

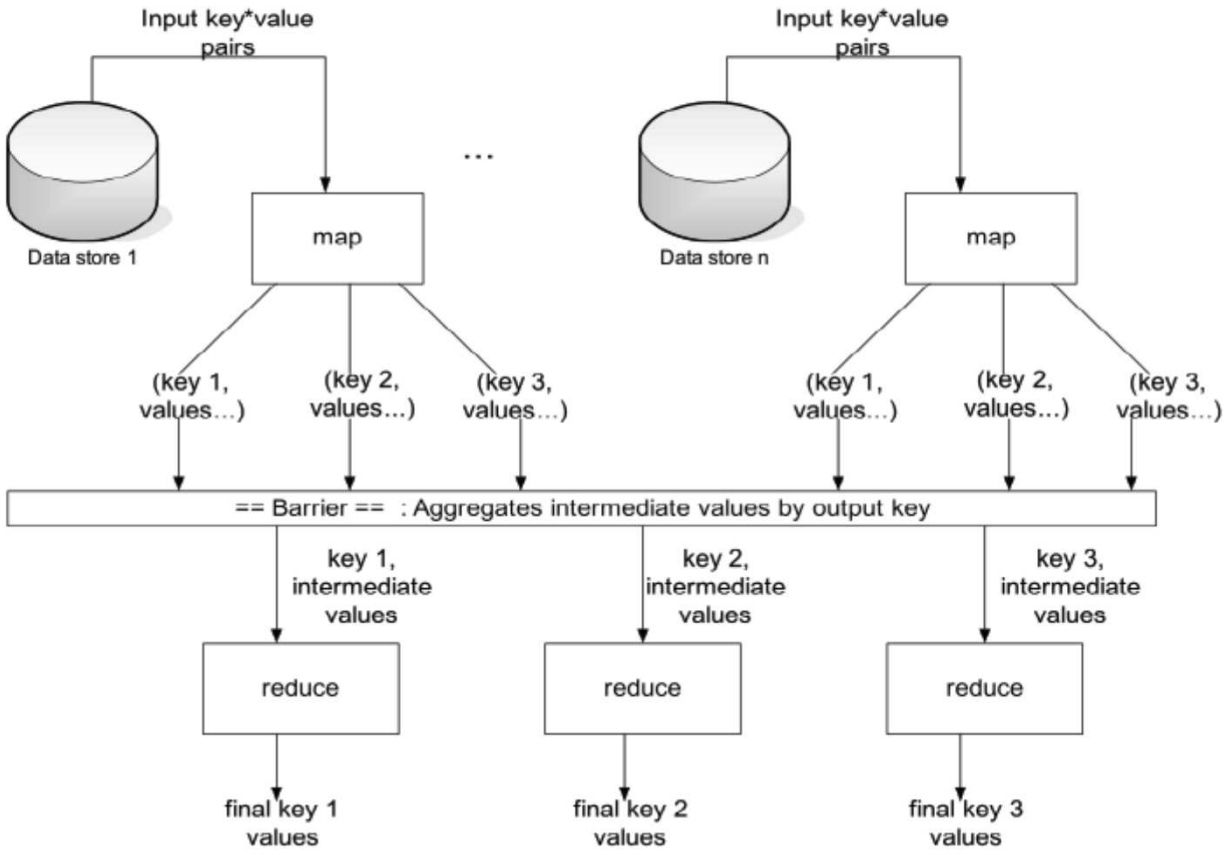
Source: Karmasphere & Booz Allen Hamilon

Hadoop is not a new model : Hadoop et SIMD





Hadoop uses MapReduce to bring processing to data



Hardware Hadoop implementation: Servers

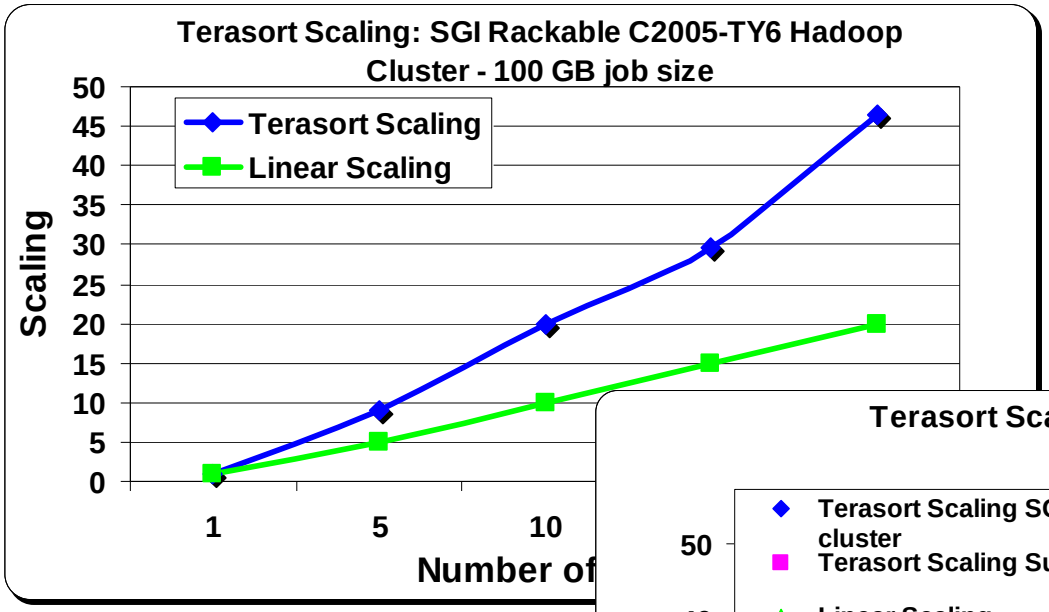
Excerpt from INTEL whitepaper:

« Optimizing Hadoop deployments »

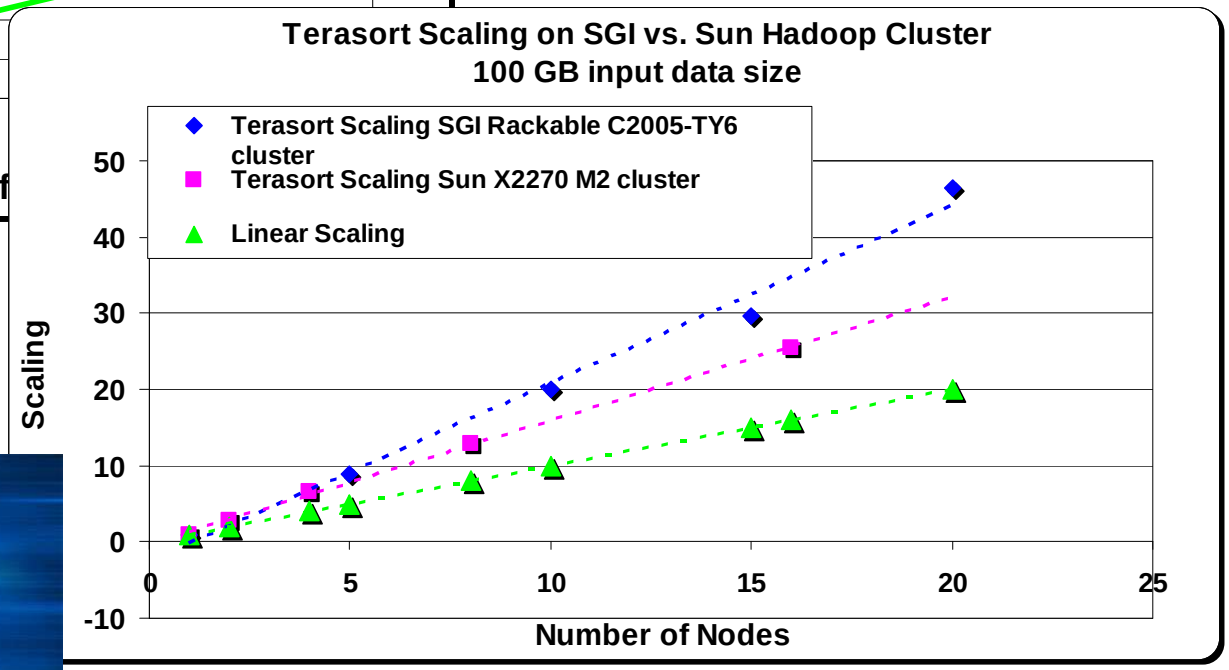
- *To maximize the energy efficiency and performance of a Hadoop cluster, it is important to consider that Hadoop deployments do not require many of the features typically found in an enterprise data centre server.*

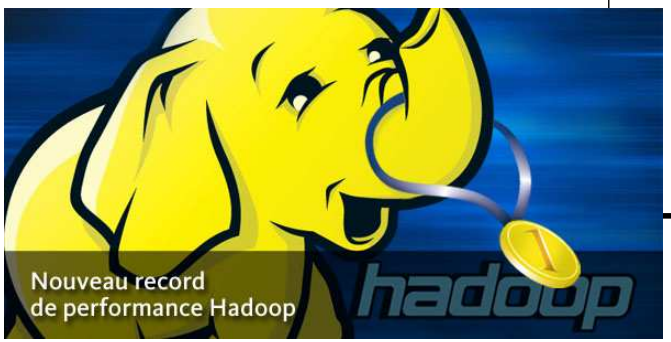
Hardware Hadoop implementation: Network

World Record Benchmark - SGI Hadoop Cluster running Terasort



Terasort @ 100GB scales ***super linearly*** on a 20-node SGI Rackable™ C2005-TY6 cluster running Cloudera distribution of Apache™ Hadoop™ (CDH3u0)





Agenda

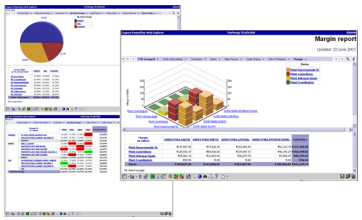
- Data flood
- Implementations
- **Hadoop**
- Not Hadoop

Choosing the right Application for Hadoop

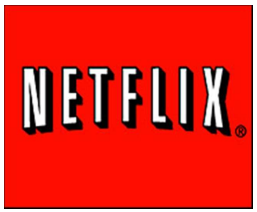
- Applications need to be ***written to scale*** to hundreds and thousands of nodes; and should *support tens of millions of files* in a single HDFS instance.
- Applications need ***streaming access*** to their data sets; designed more for batch processing rather than interactive use by users; need high throughput of data access rather than low latency of data access.
- Applications need a ***write-once-read-many access*** model for files.
- Applications need to be ***compatible*** to run in a Java MapReduce framework; need to be able to use HDFS interfaces to move themselves closer to where the data is located.
- Applications need the ability to ***process unstructured and semi-structured data*** or information.

Who is/Will be Using Hadoop

- Bioscience pharmacological trials produce massive amounts of data to validate complex interactions of molecular with experimental data.
- Financial services have larger volumes through smaller trading sizes, increased market volatility, and improvements in automated and algorithmic trading. Fraud detection analyzes otherwise unrecognizable patterns and data relationships.
- Science and research is increasingly being dominated by initiatives with large data volumes:
  - Large Hadron Collider [LHC] at CERN generates over 15 PB of data per year. The data must be distributed to be retained and processed.
  - Continental-scale experiments and environmental monitoring are both politically and technological feasible (e.g., Ocean Observatories Initiative [OOI], National Ecological Observatory Network [NEON], and USArray, a continental-scale seismic observatory)
  - Improving instrument and sensor technology (e.g., the Large Synoptic Survey Telescope [LSST] has a 3.2 Gpixel camera and will generate over 6 PB of image data per year)
- Retailers collect clickstream data from website interactions and data from traditional retailing operations for customer buying analysis and inventory management.
- Government and military agencies collect and process massive amounts of raw data from a wide variety of sources to arrive at actionable intelligence.

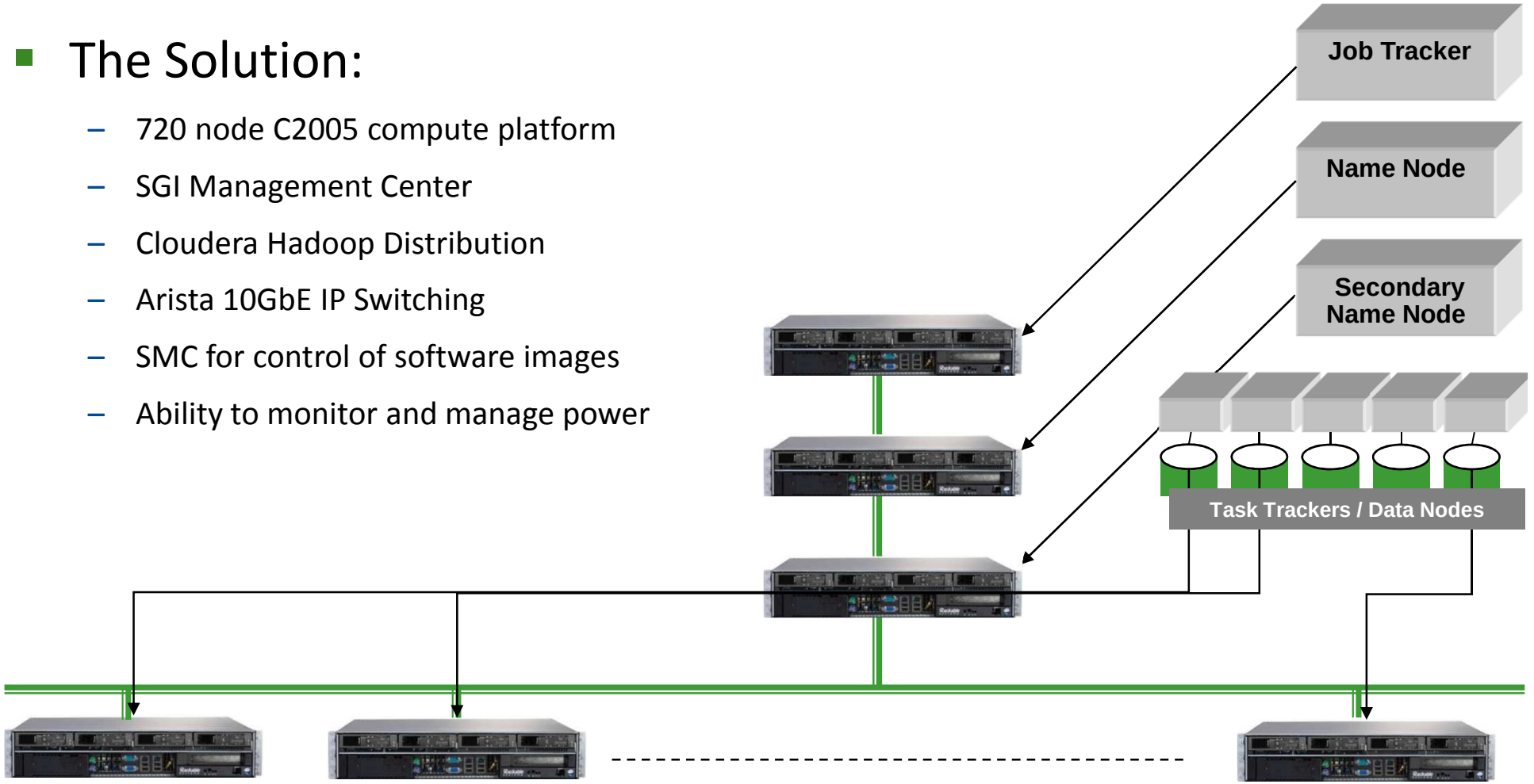


Some Hadoop users



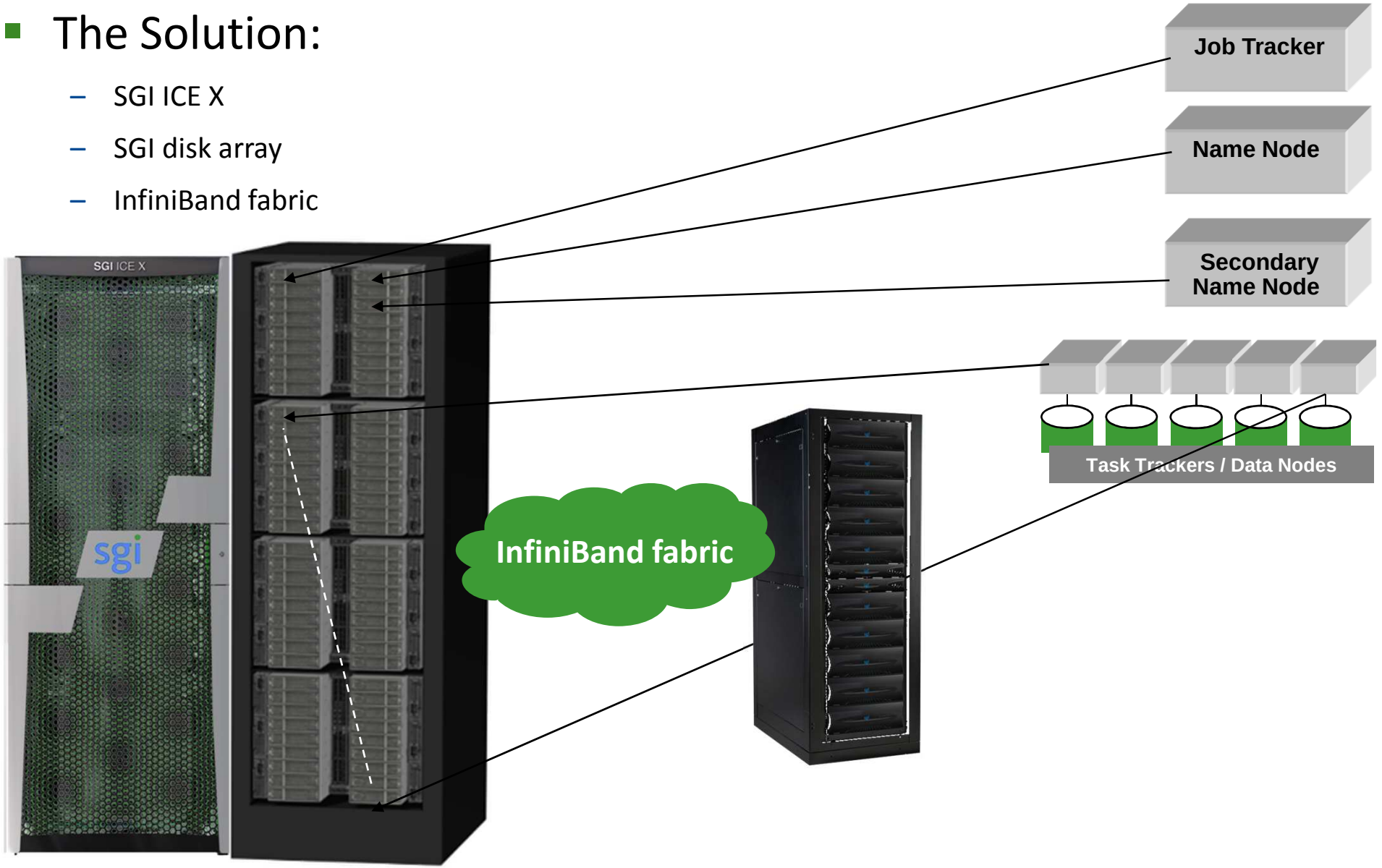
SGI Hadoop customer: regular configuration

- The Solution:
  - 720 node C2005 compute platform
  - SGI Management Center
  - Cloudera Hadoop Distribution
  - Arista 10GbE IP Switching
  - SMC for control of software images
  - Ability to monitor and manage power



SGI Hadoop customer:
not your vanilla Hadoop hardware architecture

- The Solution:
  - SGI ICE X
  - SGI disk array
  - InfiniBand fabric



Agenda

- Data flood
- Implementations
- Hadoop
- **Not Hadoop**

Big Data analytics without Hadoop

- Fraud detection

Istituto Nazionale della Previdenza Sociale

- Large memory server

Big Data analytics without Hadoop

- Wikipedia's view of: history, persons, categories, organizations
- Entire edition of English Wikipedia
- Metadata and data
  - 4 million pages
  - Connections among them

- Some kind of Google Earth view of Big Data

Big Data analytics without Hadoop (cont.)

Big Data analytics without data storage!

- IP packets analysis
- Real-time security enforcement
- Check and let the packets flow
- Extract relevant metadata for later analysis

Big Data analytics without data storage!

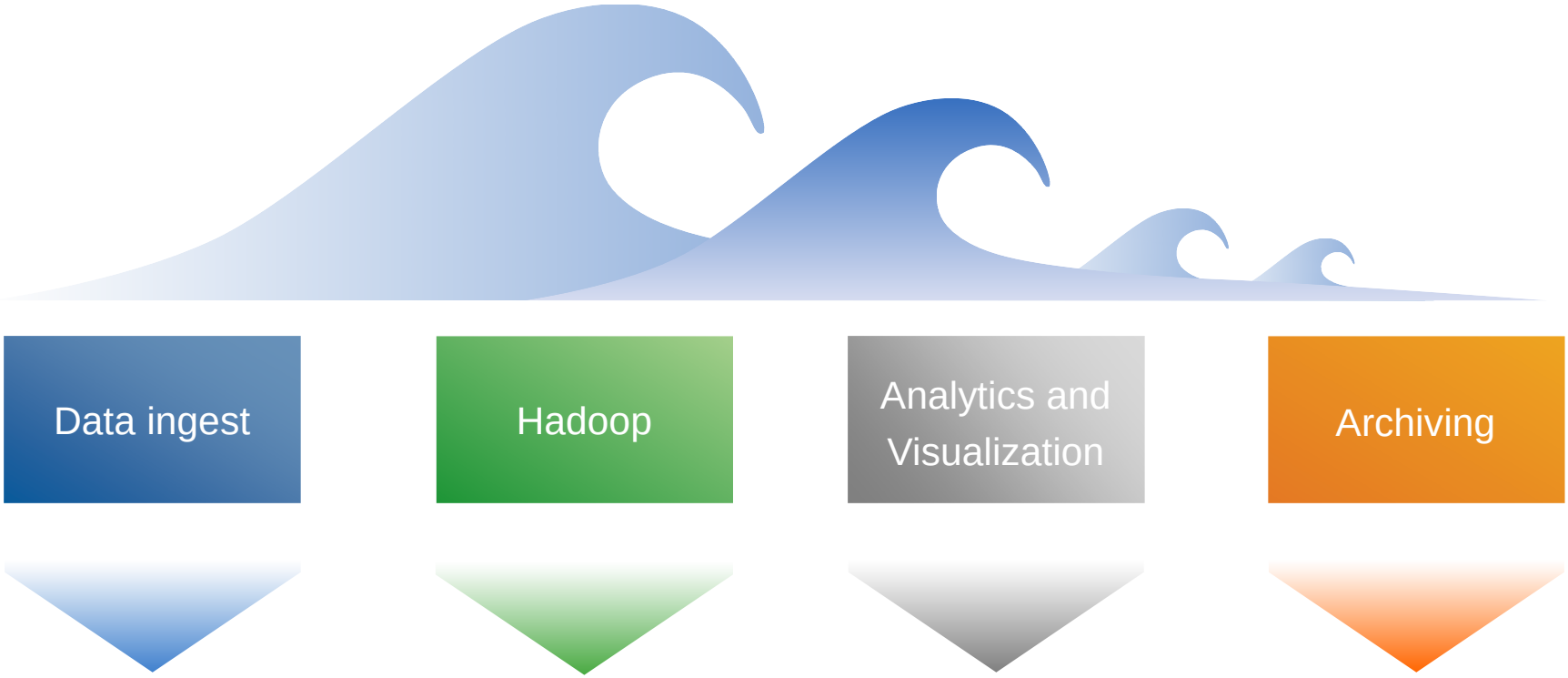
- Set top box events analysis
- Real-time Latency
- Analysis performance

Big Data and Cloud with data management

- Cloud for backup/restore, archival

- HDFS only
- Scalability
- Low cost
  - Purchase
  - TCO

The *Data Wave*



- The wave arrives
- **Single large memory server**
- **Numerous regular servers**

- Focus the wave
- **Hadoop Clusters**

- Processing eddies
- **Misc servers**

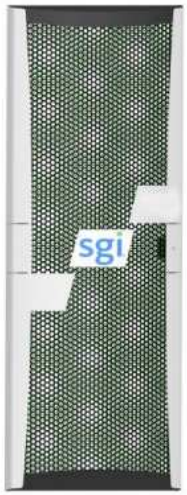
- Store the value
- **Dense disk arrays**
- **Archival solution**

Big Data Dream Team

Rackable

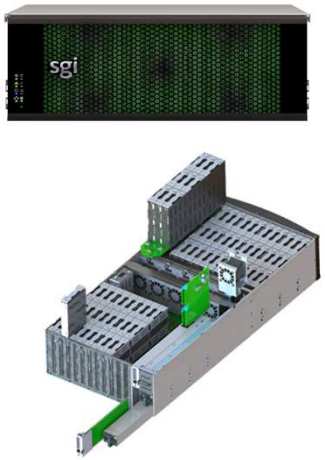
Unstructured data

SGI UV



Structured data

SGI MIS



Cloud storage

ArcFiniti

Archival


sgi
accelerating results