Big Data

What it is and what is isn't…

Vincent Staropoli

CloudDevelop Conference
October, 23 2015 – Columbus, OH

About Me

Microsoft CERTIFIED
Solutions Expert
Business Intelligence

Microsoft CERTIFIED
Solutions Expert
Data Platform

http://linkedin.com/in/vincentstaropoli

http://twitter.com/vstaropoli

http://stackoverflow.com/users/387041/vinnie

http://github.com/thevinnie

http://cardinalsolutions.com/profiles/vstaropoli

The Story of Hadoop

- Google and Yahoo got together, got drunk and made a baby elephant named Hadoop
- It was so big, they had to wrap a map around it's flume to reduce the splunk on mahout
- Personally, I'd rather learn pig latin from a guy named ambari than sqoop the hive of hbase that elephant leaves around
- But, when you're in a storm spinning a yarn like this one, there's only one place you wish you could go in a flash
- The Cloud

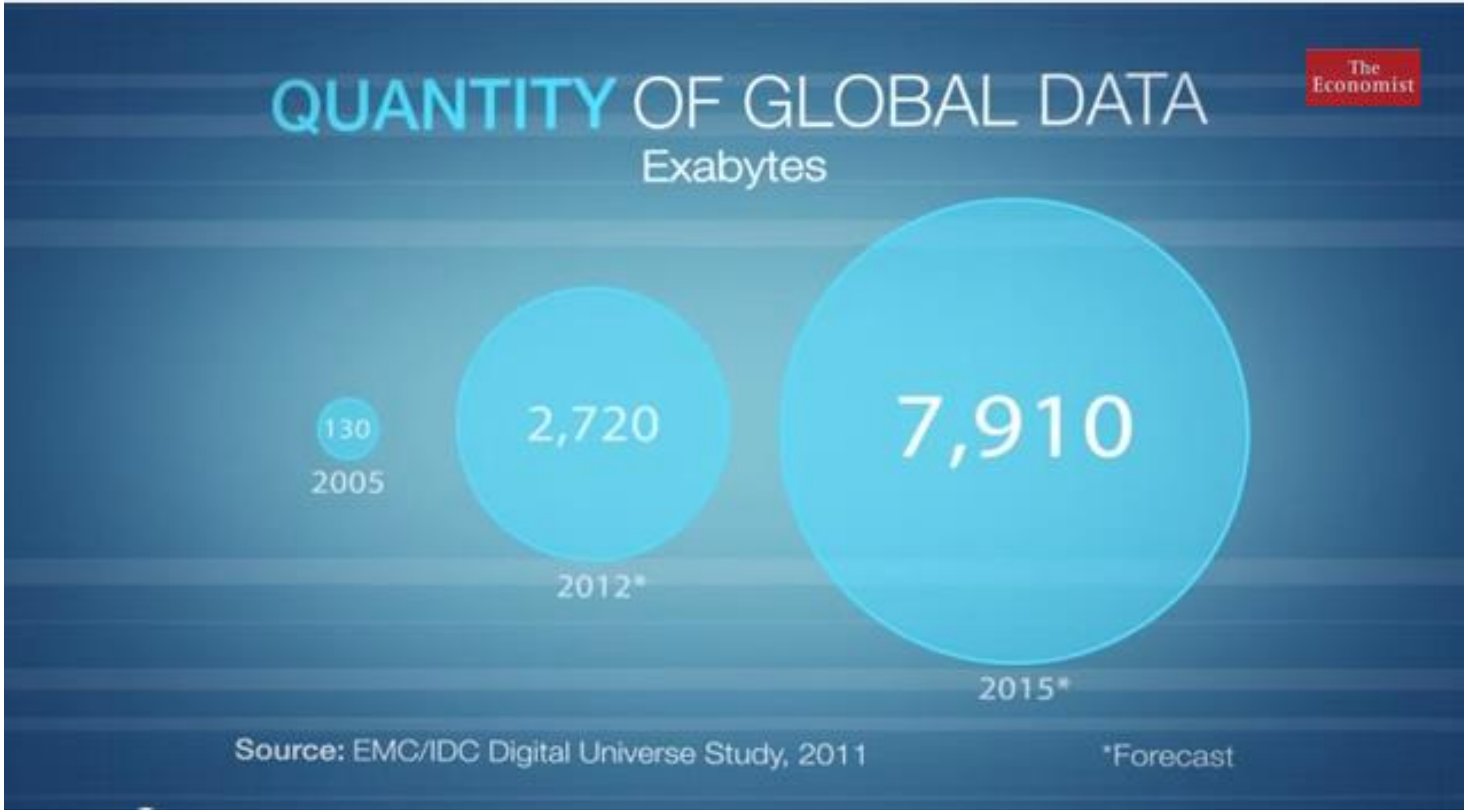
Why Discuss Big Data?

- Not just a technology topic
- Adjective as well as a noun, occasionally a verb
- Tool in the toolkit for developers and businesses
- No magic, no silver bullet and won’t change culture
- Changing *RAPIDLY*

Our Journey…

- Data, data, data
- Systems & technology
- Data-driven
- “Welcome to real life”

Data Getting Bigger



QUANTITY OF GLOBAL DATA

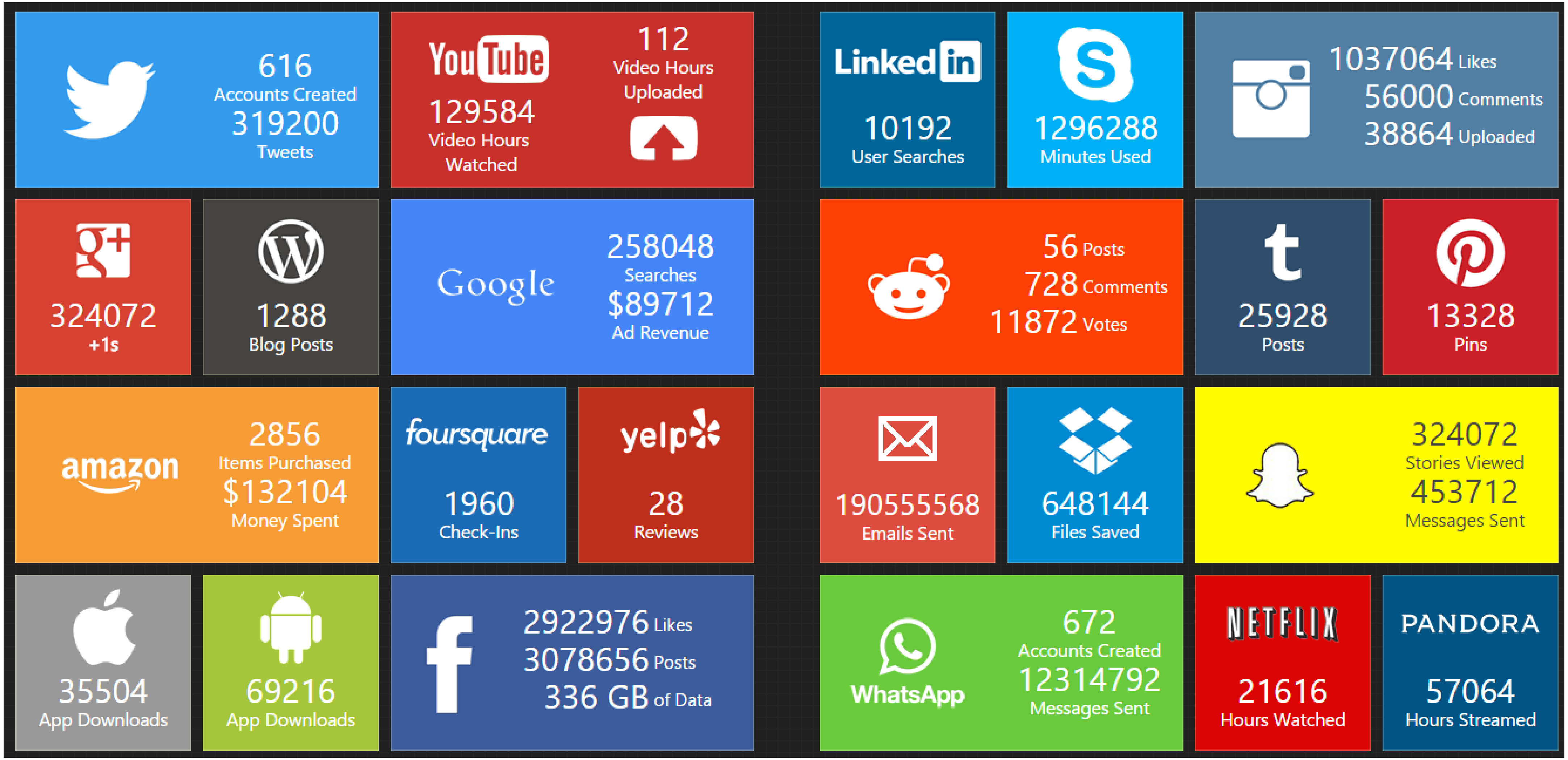
Exabytes

Source: EMC/IDC Digital Universe Study, 2011

*Forecast

Big Data Sources

http://pennystocks.la/internet-in-real-time/



616 Accounts Created
319200 Tweets

YouTube
112 Video Hours Uploaded
129584 Video Hours Watched

LinkedIn
10192 User Searches

1296288 Minutes Used

1037064 Likes
56000 Comments
38864 Uploaded

324072 +1s

1288 Blog Posts

Google
258048 Searches
$89712 Ad Revenue

56 Posts
728 Comments
11872 Votes

25928 Posts

13328 Pins

amazon
2856 Items Purchased
$132104 Money Spent

foursquare
1960 Check-Ins

yelp
28 Reviews

190555568 Emails Sent

648144 Files Saved

324072 Stories Viewed
453712 Messages Sent

35504 App Downloads

69216 App Downloads

2922976 Likes
3078656 Posts
336 GB of Data

WhatsApp
672 Accounts Created
12314792 Messages Sent

NETFLIX
21616 Hours Watched

PANDORA
57064 Hours Streamed

“The biggest reason that investments in big data fail to pay off is that most companies don’t do a good job with the information they already have.”

-"*You May Not Need Big Data After All"*
https://hbr.org/2013/12/you-may-not-need-big-data-after-all

Big Data Systems

- Framework of technologies to store and analyze data
- Exceeds the processing capacity of conventional database systems
- Characterized by **volume**, **velocity** and **variety**
- Significant effort to find the signal within the noise

Big Data Systems

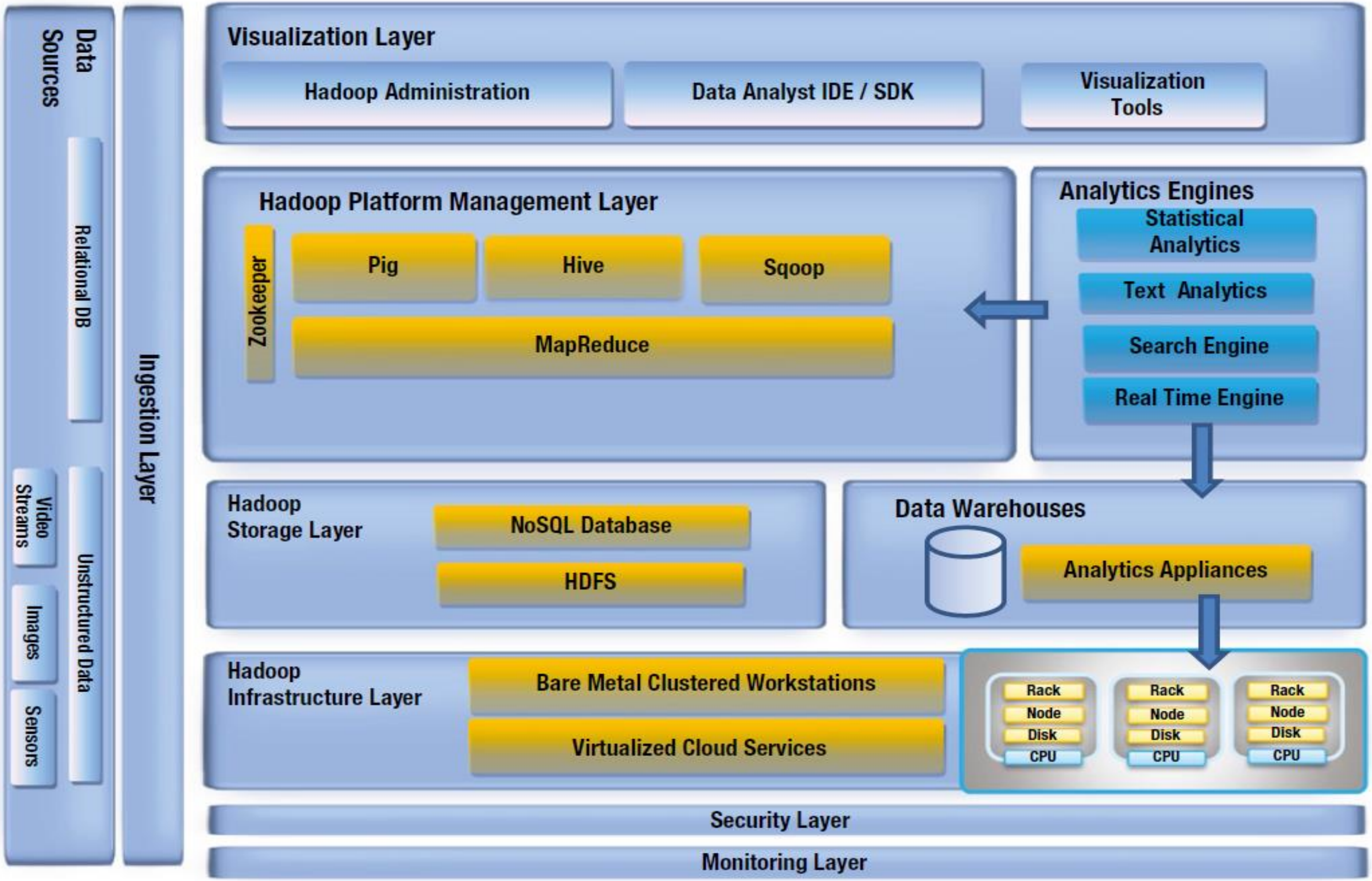
- Core Technologies
  - Hadoop
  - MapReduce
- New technologies to manage huge amounts of data
  - Store
  - Access
  - Analyze
- Monetize the benefits of owning huge amounts of data
  - Capability to process massive amounts of data
  - Efficient, cost-effective, and timely

Big Data Systems

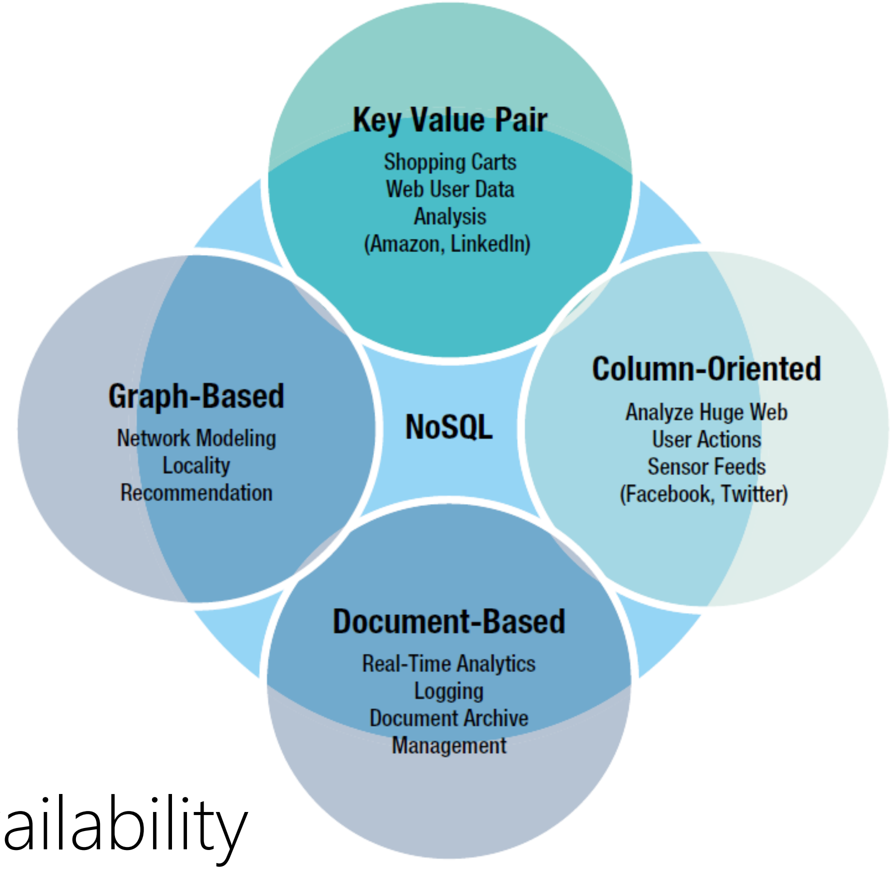
- Storage and infrastructure
- Platform management
- Ingestion
- Visualization

Big Data Systems

Reference Architecture

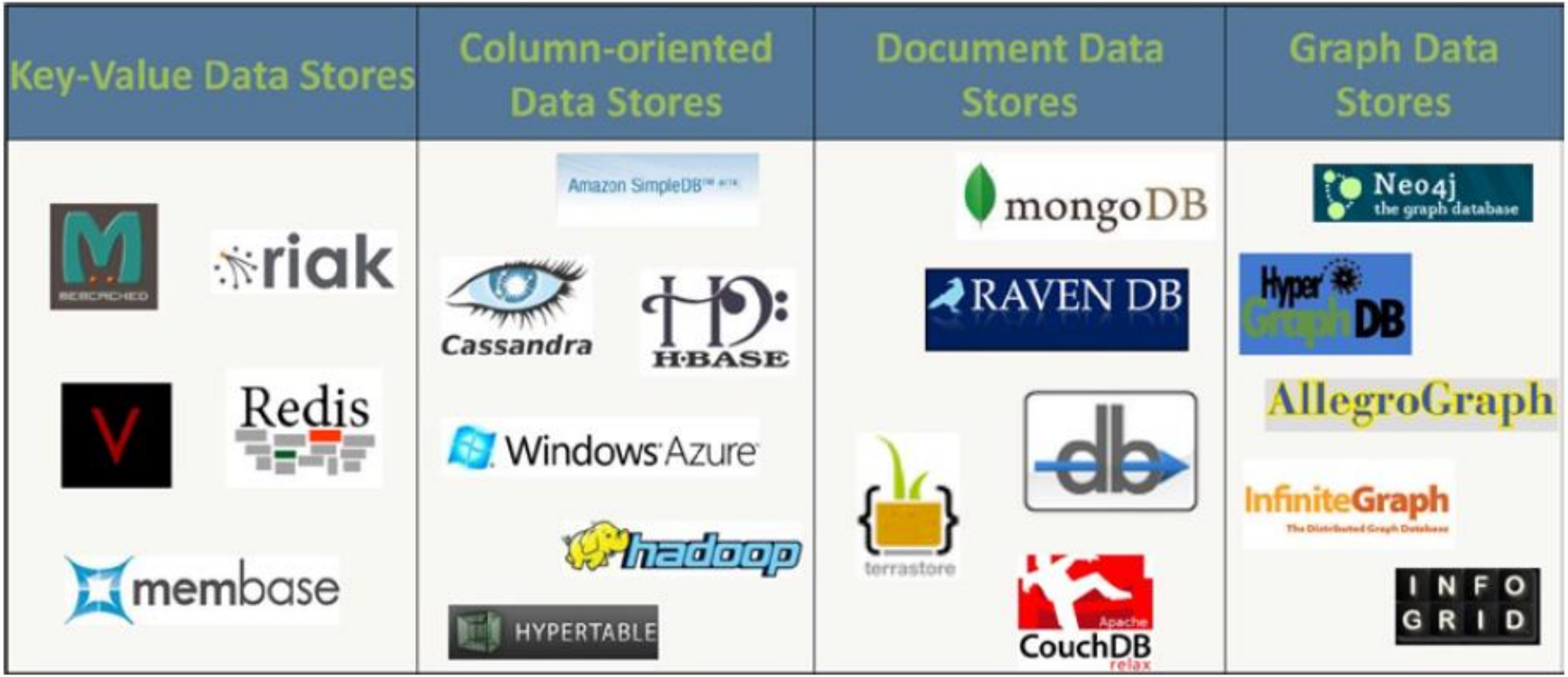


Big Data and NoSQL



- “Not Only SQL”
- Different solutions for different applications
- Must relax guarantees around consistency, availability and partition tolerance (the CAP Theorem)
- Likely have a combination of relational and NoSQL databases

NoSQL Database Systems



Key-Value Data Stores	Column-oriented Data Stores	Document Data Stores	Graph Data Stores
MEMCACHED riak V Redis membase	Amazon SimpleDB™ BETA Cassandra H·BASE Windows Azure hadoop HYPERTABLE	mongoDB RAVEN DB terrastore db Apache CouchDB relax	Neo4j the graph database HyperGraphDB AllegroGraph InfiniteGraph The Distributed Graph Database INFO GRID

MapReduce

- Map function
  - Applies a function on every key/value pair in the collection
  - Generates a new collection

- Reduce function
  - Works on the new generated collection
  - Applies an aggregate function to compute a final output

MapReduce

Sample Data:

```
[{ “94303”: “Tom”}, {“94303”: “Jane”}, {“94301”:
“Arun”}, {“94302”: “Chen”}]
```

Get the names of all those who reside in a particular zip code:

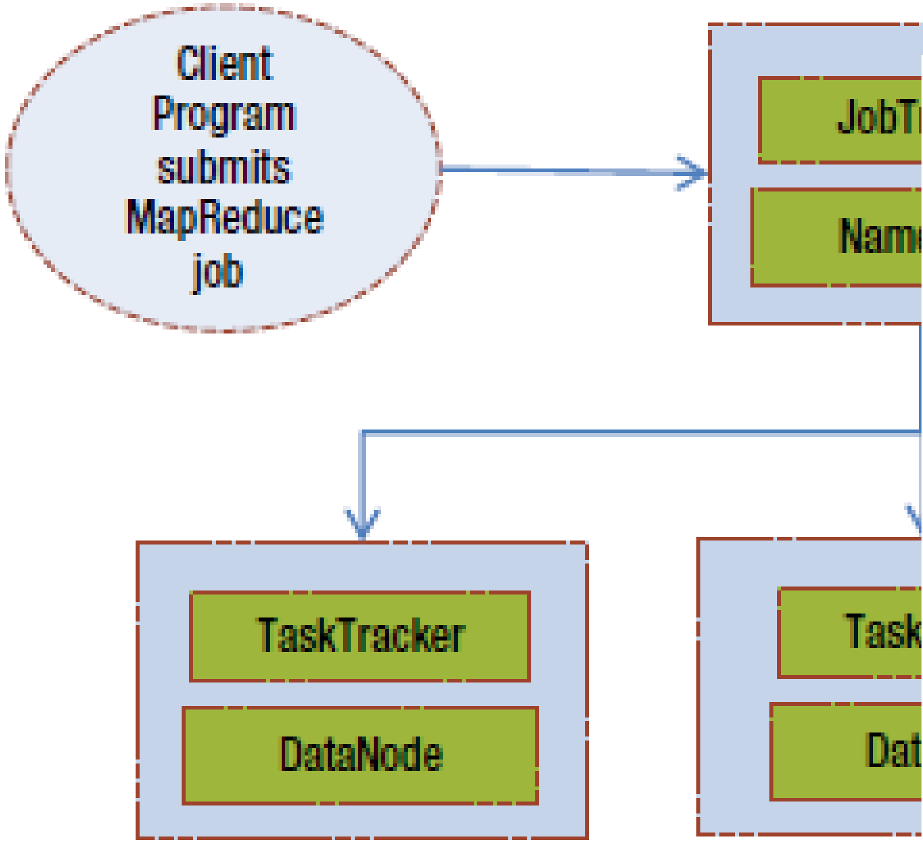
```
[{“94303”:[“Tom”, “Jane”]}, {“94301”:[“Arun”]},
{“94302”:[“Chen”]}]
```

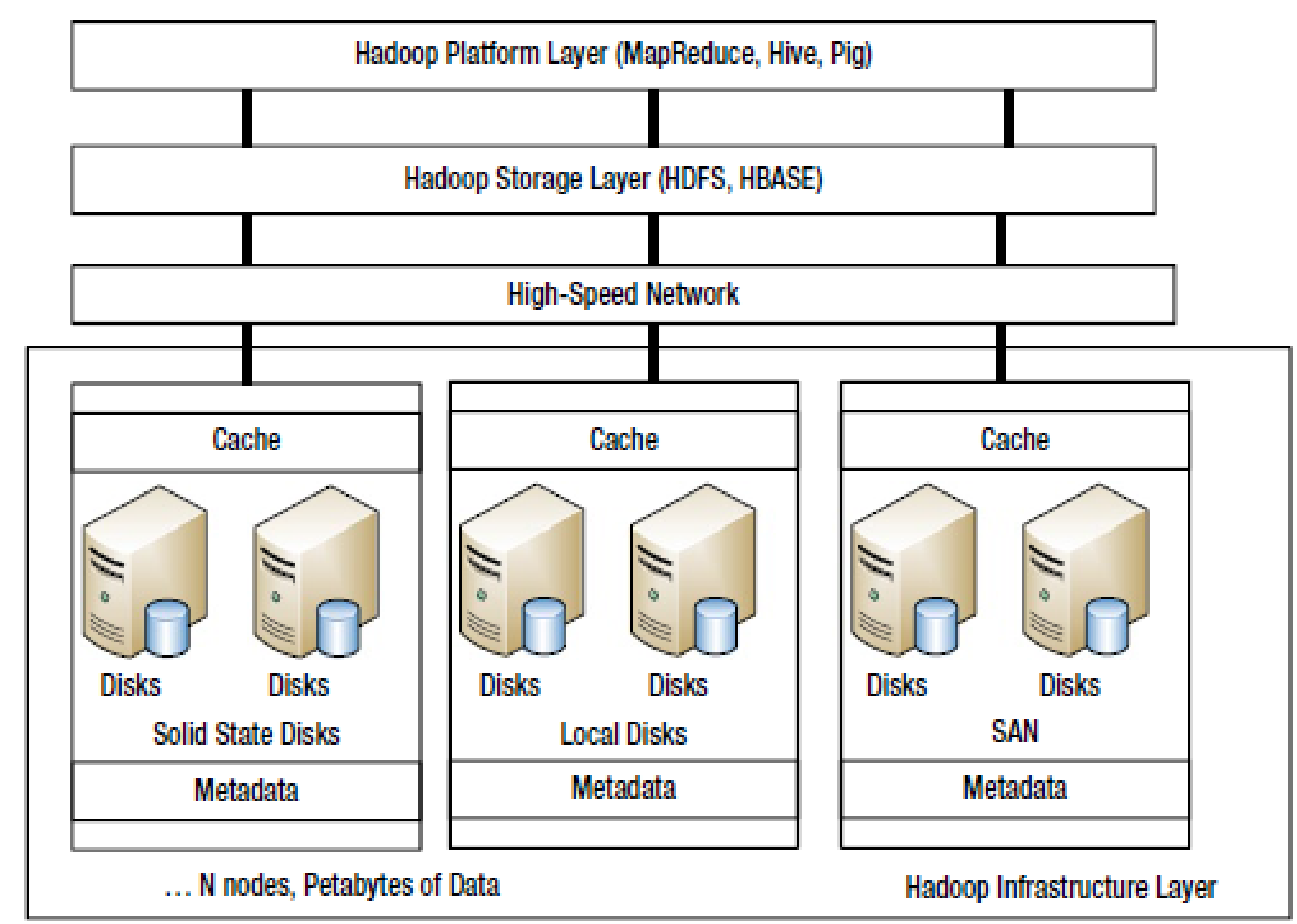
Reduce function output to count the number of people by zip code:

```
[{“94303”: 2}, {“94301”: 1}, {“94302”: 1}]
```

Hadoop Clustering

Conceptual Architecture

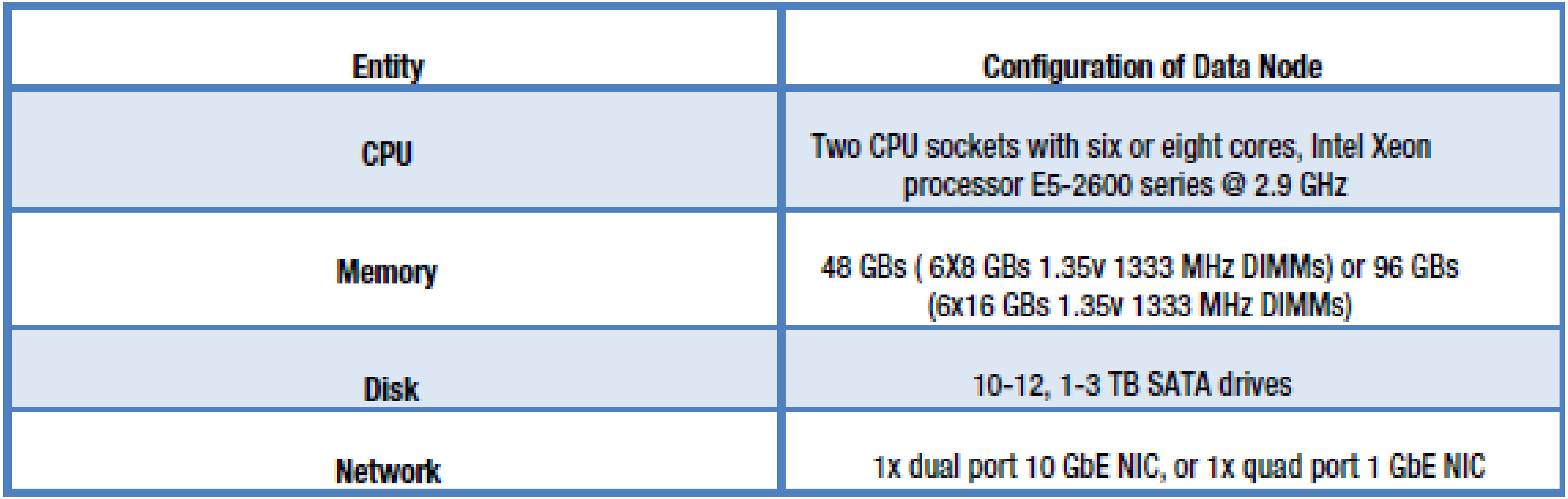




Hadoop Platform Layer (MapReduce, Hive, Pig)
Hadoop Storage Layer (HDFS, HBASE)
High-Speed Network
Cache
Disks
Disks
Solid State Disks
Metadata
Cache
Disks
Disks
Local Disks
Metadata
Cache
Disks
Disks
SAN
Metadata
… N nodes, Petabytes of Data
Hadoop Infrastructure Layer

Hadoop Clustering

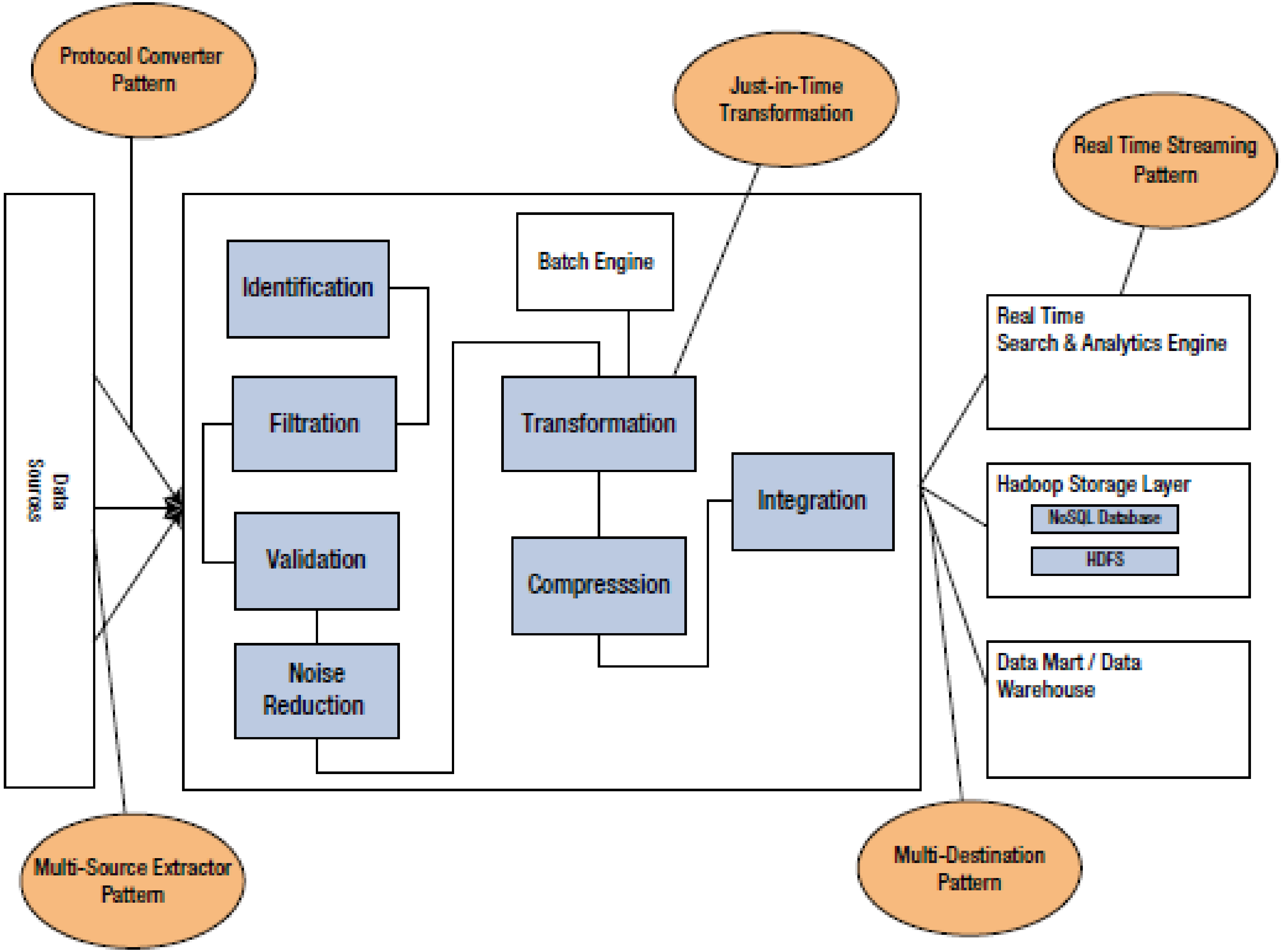
Commodity Hardware



Entity	Configuration of Data Node
CPU	Two CPU sockets with six or eight cores, Intel Xeon processor E5-2600 series @ 2.9 GHz
Memory	48 GBs (6X8 GBs 1.35v 1333 MHz DIMMs) or 96 GBs (6x16 GBs 1.35v 1333 MHz DIMMs)
Disk	10-12, 1-3 TB SATA drives
Network	1x dual port 10 GbE NIC, or 1x quad port 1 GbE NIC

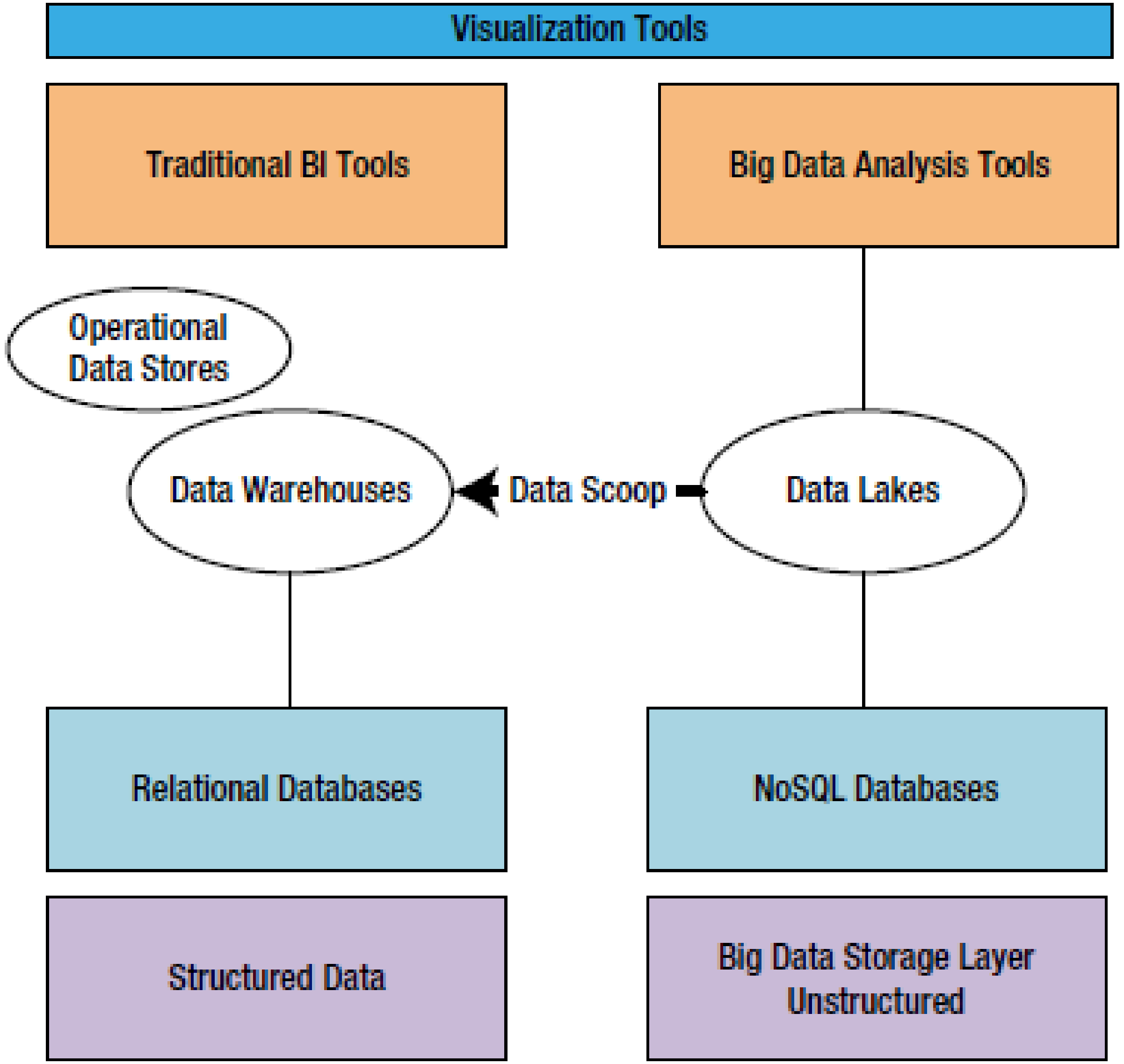
Hadoop Clustering

Ingestion Engine


Protocol Converter Pattern
Just-in-Time Transformation
Real Time Streaming Pattern
Data Sources
Identification
Filtration
Validation
Noise Reduction
Batch Engine
Transformation
Compresssion
Integration
Real Time Search & Analytics Engine
Hadoop Storage Layer
NoSQL Database
HDFS
Data Mart / Data Warehouse
Multi-Source Extractor Pattern
Multi-Destination Pattern

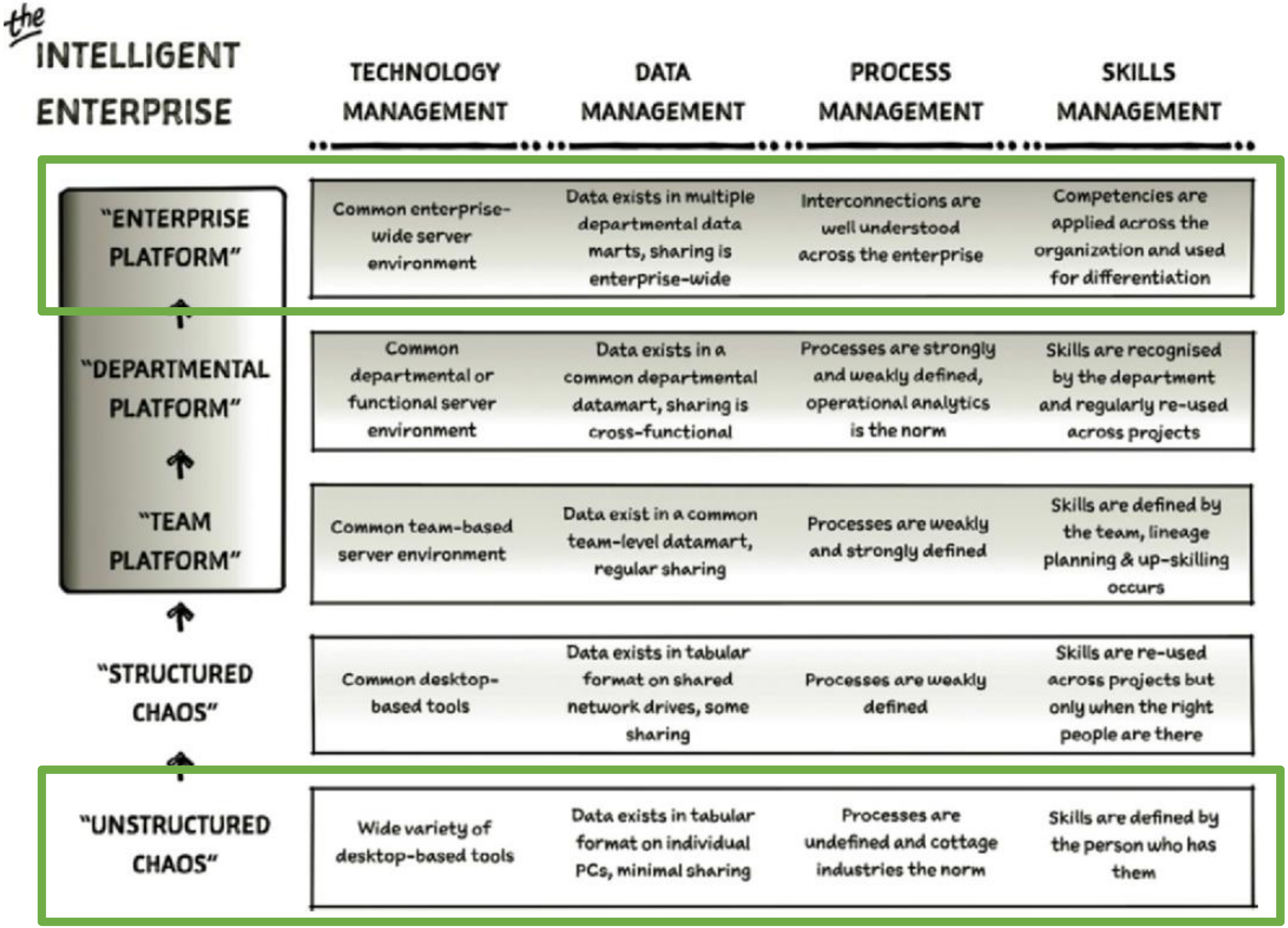
Hadoop Clustering

Scalability



Data Driven Organizations

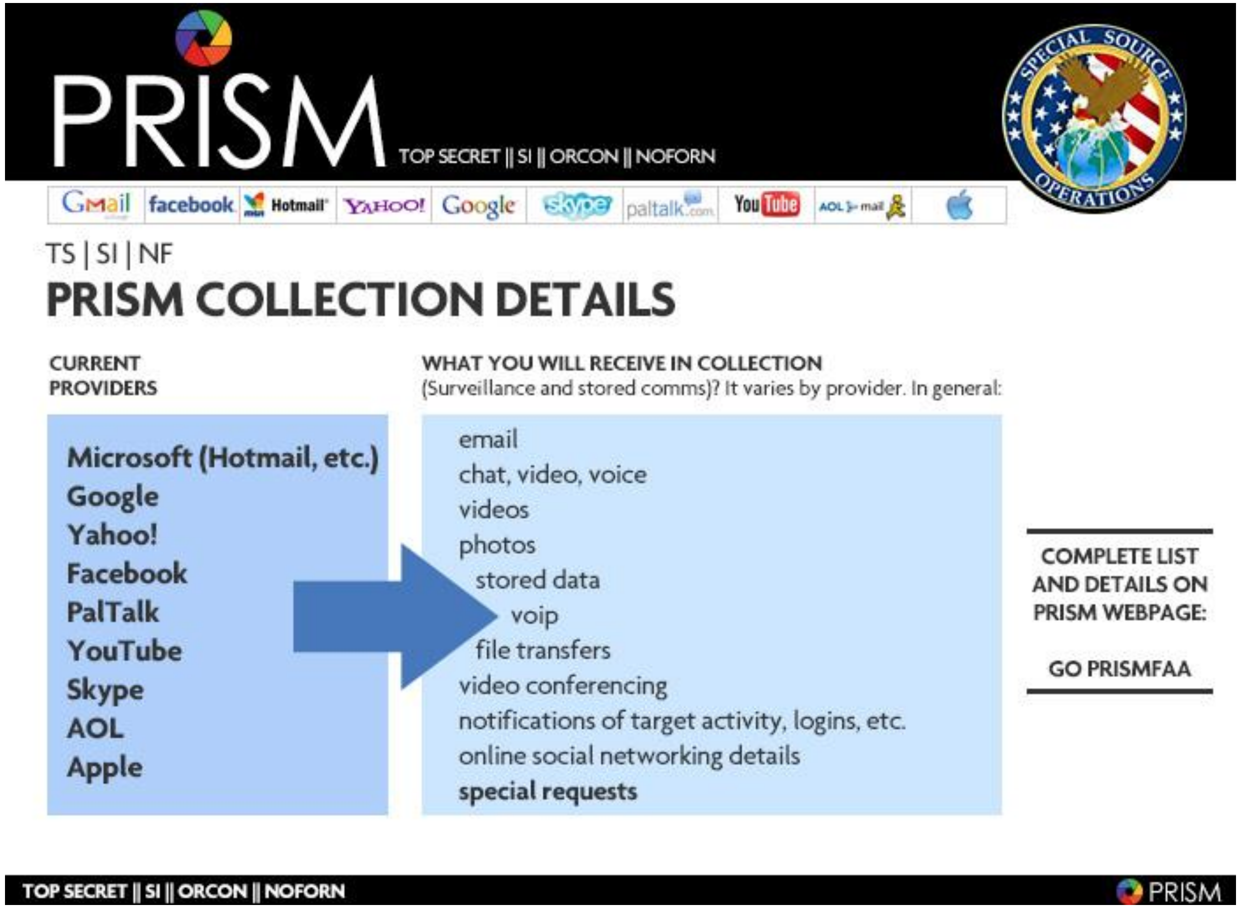
- Establish one undisputed source
- Near-real-time feedback
- Articulate business rules (and regularly update them)
- High-quality coaching to employees



the INTELLIGENT ENTERPRISE	TECHNOLOGY MANAGEMENT	DATA MANAGEMENT	PROCESS MANAGEMENT	SKILLS MANAGEMENT
"ENTERPRISE PLATFORM"	Common enterprise-wide server environment	Data exists in multiple departmental data marts, sharing is enterprise-wide	Interconnections are well understood across the enterprise	Competencies are applied across the organization and used for differentiation
"DEPARTMENTAL PLATFORM"	Common departmental or functional server environment	Data exists in a common departmental datamart, sharing is cross-functional	Processes are strongly and weakly defined, operational analytics is the norm	Skills are recognised by the department and regularly re-used across projects
"TEAM PLATFORM"	Common team-based server environment	Data exist in a common team-level datamart, regular sharing	Processes are weakly and strongly defined	Skills are defined by the team, lineage planning & up-skilling occurs
"STRUCTURED CHAOS"	Common desktop-based tools	Data exists in tabular format on shared network drives, some sharing	Processes are weakly defined	Skills are re-used across projects but only when the right people are there
"UNSTRUCTURED CHAOS"	Wide variety of desktop-based tools	Data exists in tabular format on individual PCs, minimal sharing	Processes are undefined and cottage industries the norm	Skills are defined by the person who has them

Big Data Systems

- Why is this so powerful?
  - A more complete picture
  - Machine learning, data mining
  - Patterns, patterns, patterns
  - Tradeoffs?


Gmail
facebook
Hotmail
YAHOO!
Google
skype
paltalk.com
YouTube
AOL mail
SPECIAL SOURCE OPERATIONS

PRISM TOP SECRET || SI || ORCON || NOFORN

TS | SI | NF

PRISM COLLECTION DETAILS

CURRENT PROVIDERS

- **Microsoft (Hotmail, etc.)**
- **Google**
- **Yahoo!**
- **Facebook**
- **PalTalk**
- **YouTube**
- **Skype**
- **AOL**
- **Apple**

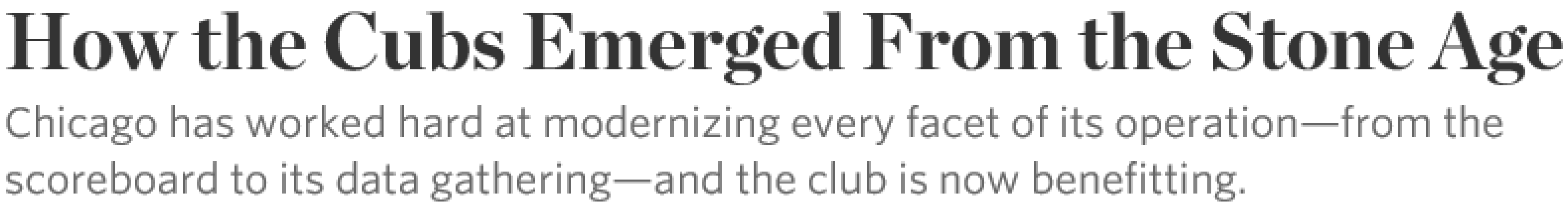
WHAT YOU WILL RECEIVE IN COLLECTION
(Surveillance and stored comms)? It varies by provider. In general:

- email
- chat, video, voice
- videos
- photos
- stored data
- voip
- file transfers
- video conferencing
- notifications of target activity, logins, etc.
- online social networking details
- **special requests**

COMPLETE LIST AND DETAILS ON PRISM WEBPAGE:

GO PRISMFAA

TOP SECRET || SI || ORCON || NOFORN PRISM



How the Cubs Emerged From the Stone Age

Chicago has worked hard at modernizing every facet of its operation—from the scoreboard to its data gathering—and the club is now benefitting.

“These were not revolutionary advances within the industry. But for the Cubs, **it felt like the space age**. The team’s previous information hub was a lone secretary who kept player contracts in file cabinets.”

Making the Case for the ‘Long-Tail’ of Big Data

“Instead, the **future of data management lies in “data curation,”** which he describes as being “aimed directly at the ‘long tail’.”

“The long tail refers to the hundreds or thousands of data silos not captured within the traditional data warehouse, and **which can only be captured and integrated at scale by applying automation and machine-learning** based on statistical patterns.”

http://blogs.wsj.com/cio/2015/08/27/making-the-case-for-the-long-tail-of-big-data/

Do you consider analytics in your applications?

Are you helping create a data driven culture?