Big Enterprise Data

David Trastour
3rd July 2012

Agenda

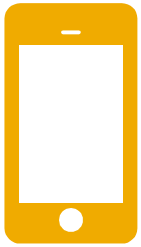
Context and challenges around Big Data

Fast Data

Big Analytics

Conclusion

Big Data Trends



Smart phones growing 20% y/y

30M networked sensors nodes growing 30% y/y

5B Mobile Phones in Use

Population of 7B in 2012

48 hours of video uploaded/minute

800M active users 30B pieces of content shared/month

Information Explosion for Enterprises

Increasingly beyond the perimeter of enterprises

Web

Open Data

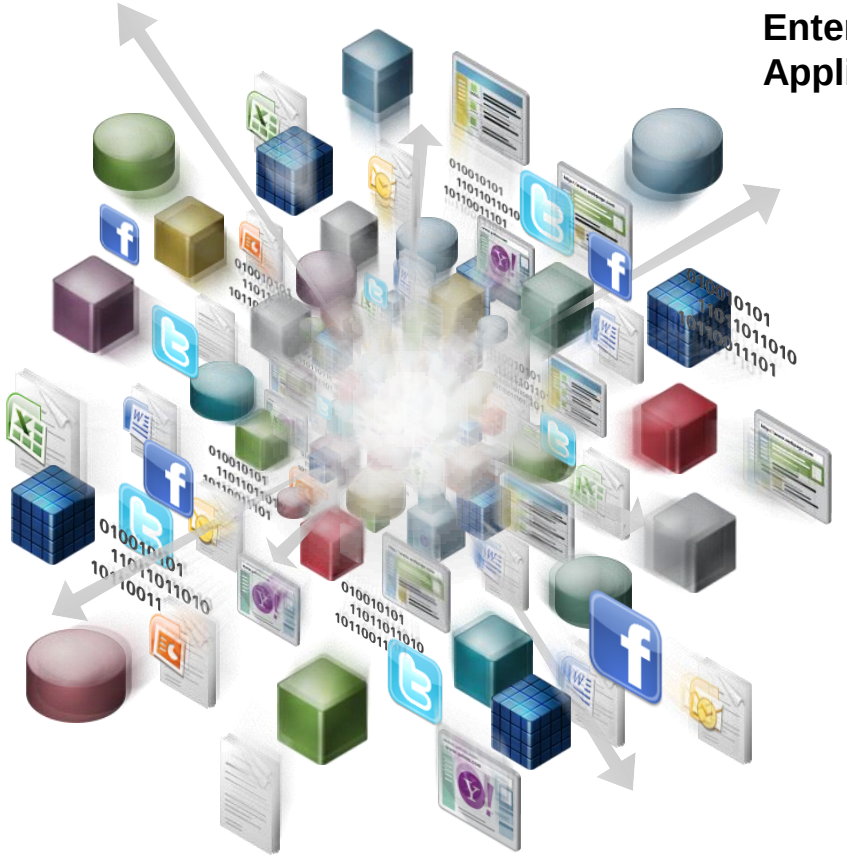
Twitter

Email

Facebook

Spreadsheets

Events



Enterprise Applications

Data Marts

Data Warehouses

Databases

Data Doubles Every 18 months

80% of Enterprise Data Is Unstructured

Information Is a Strategic Corporate Asset

The 3 V's of Big Data

VELOCITY

Real-time access to information enables new applications.

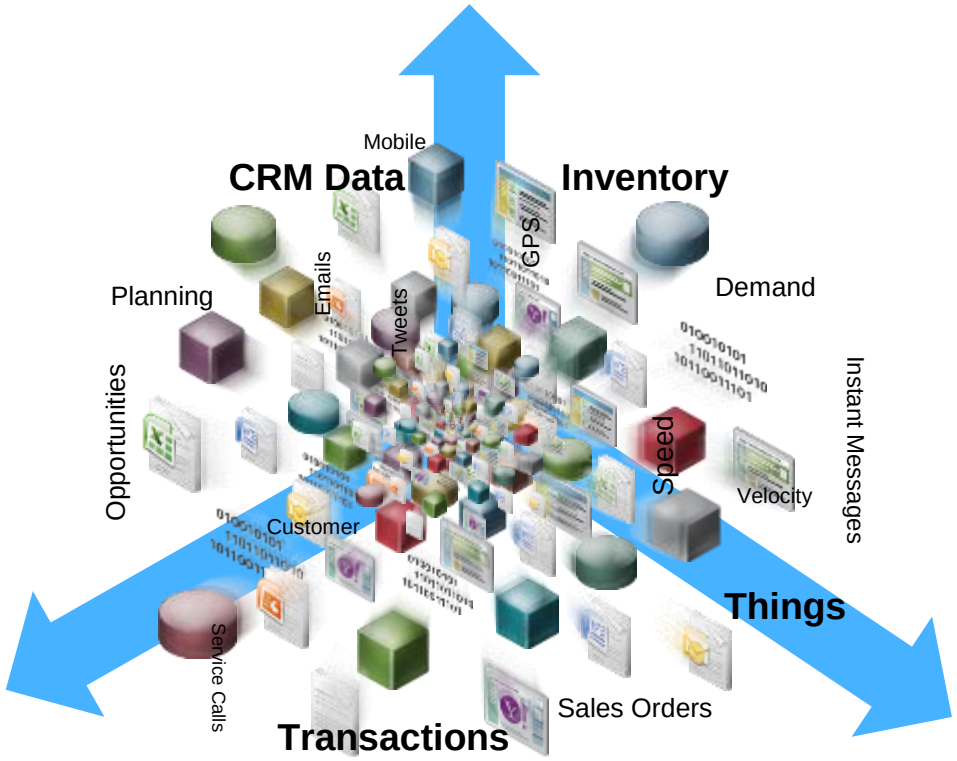
- Batch
- Near time
- Real time
- Streaming

- Terabytes
- Records
- Transactions
- Tables, files

VOLUME

Worldwide digital content will **double every 18 months.**

IDC



VARIETY

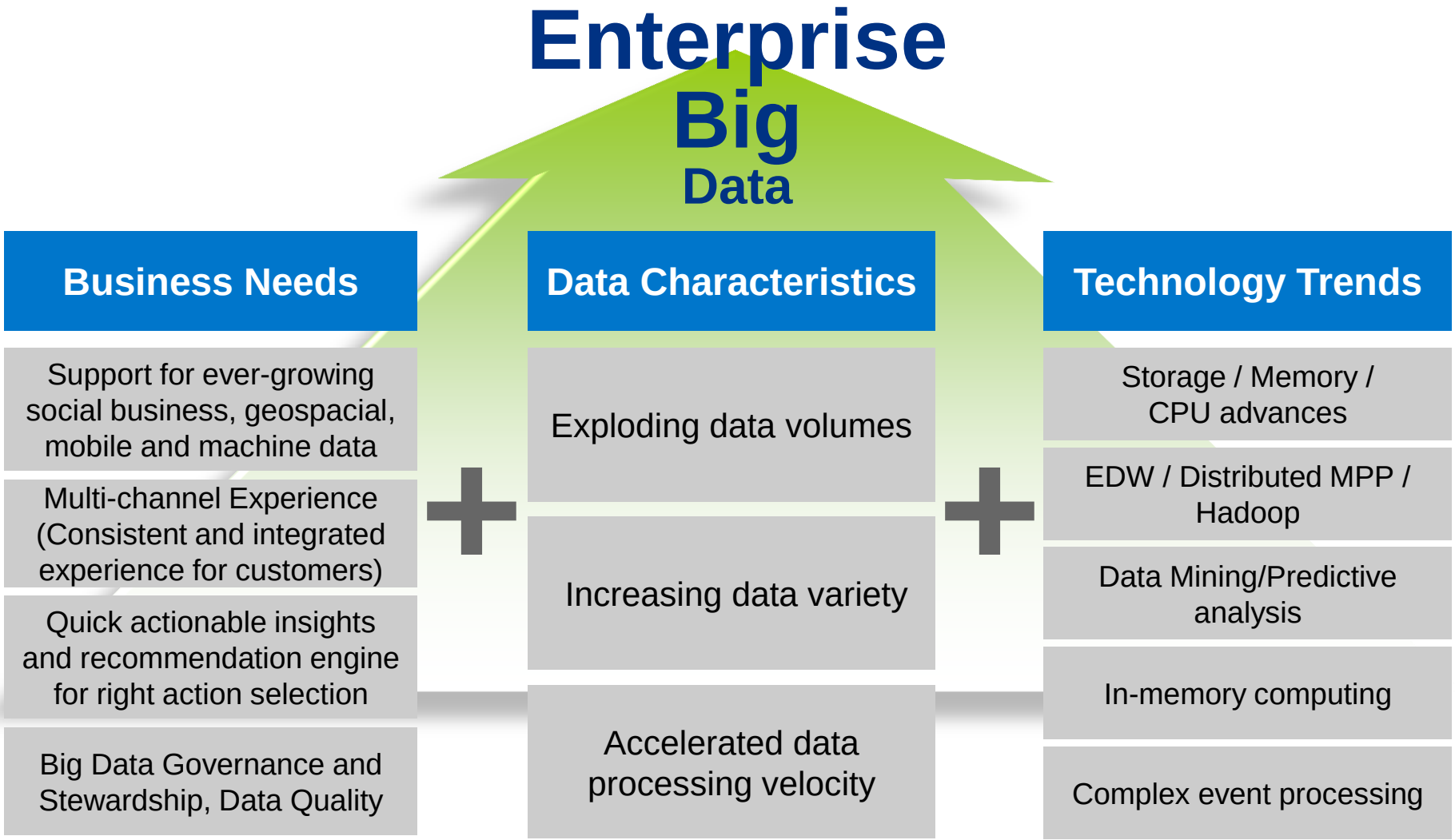
80% of enterprise data will be unstructured spanning traditional and non traditional sources.

Gartner

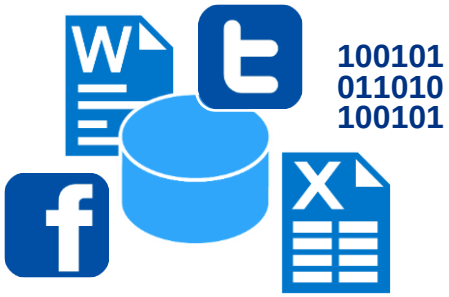
- Structured
- Semi-structured
- Unstructured

The Big Data Phenomenon

Big Data is not a single technology, but a platform for extremely scalable analytics



Big Data Challenges



All Relevant Big Data

Challenges

- Cost of store vs. cost to process data considerations
- Diversity of data formats makes it difficult to analyze

Quick Insight

Challenges

- Pressure to gain insight quickly from data
- Need to have disparate technologies interoperate across enterprise

Right Action

Challenges

- Determine the right action among many valuable insights across variety, volume, velocity and technology is difficult

Big Data Creates New Opportunities

Deliver enhanced insights and enable new applications that were not feasible (or cost-effective) before

Value Scenario Example

Out-of-stock predictive analysis, product affinity insights, sales forecast

A large CPG company running a country-wide promotion

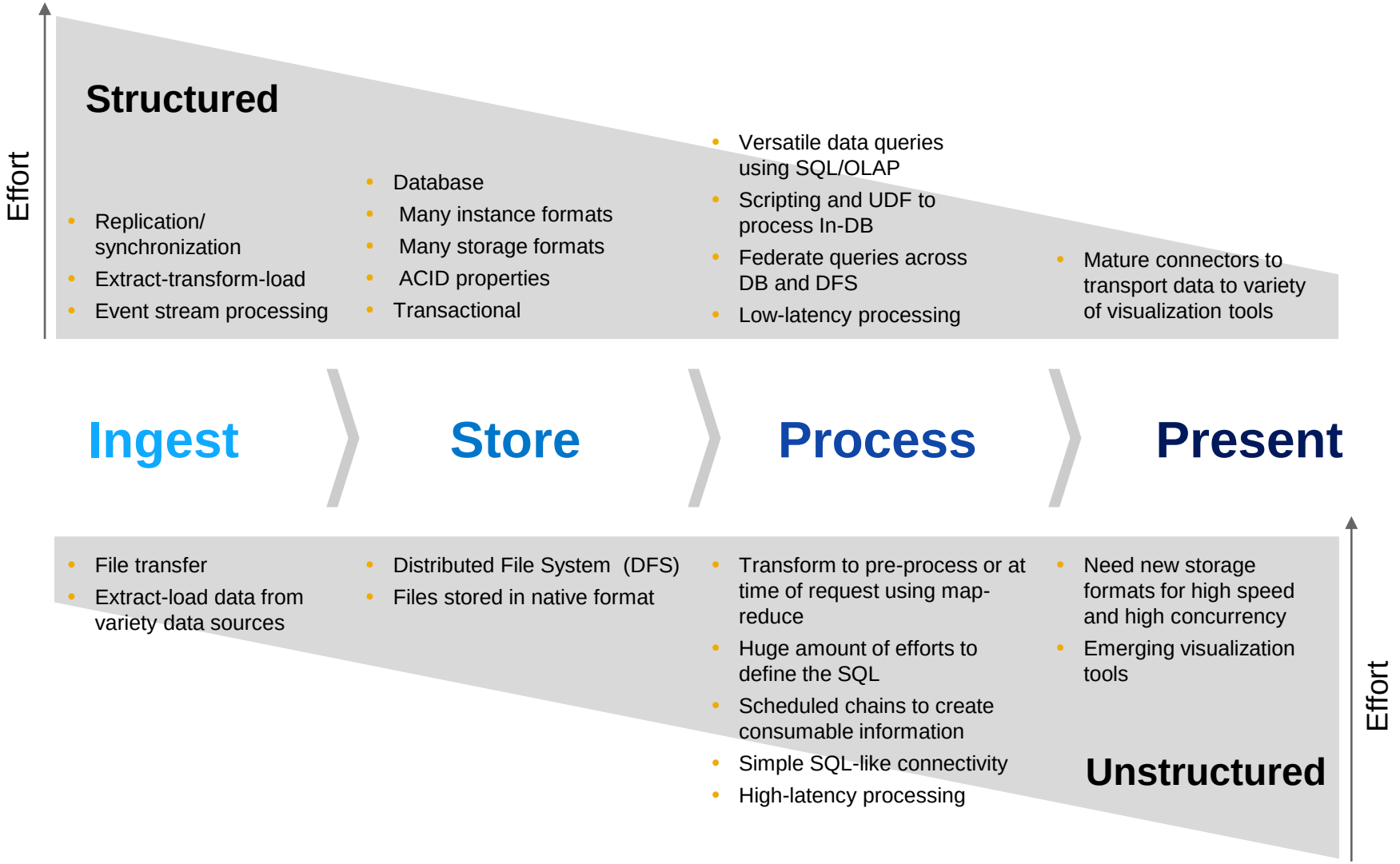
20x faster analysis with **200x** better price/performance ratio on large volumes of POS data

PB+ of web log histories and social media feeds

Big Data Value

- Shelf turnaround: 5 to 2 days
- Eliminate out of stock scenario during promotion

Understand Big Data through Information Management Lifecycle



Fast Data

The NoSQL approaches

In-memory databases

Physical Limits

Current relational database technology can’t handle the volume, velocity and variety of all your data

Driving Forces for NoSQL approaches

ACID vs. Relaxed Consistency

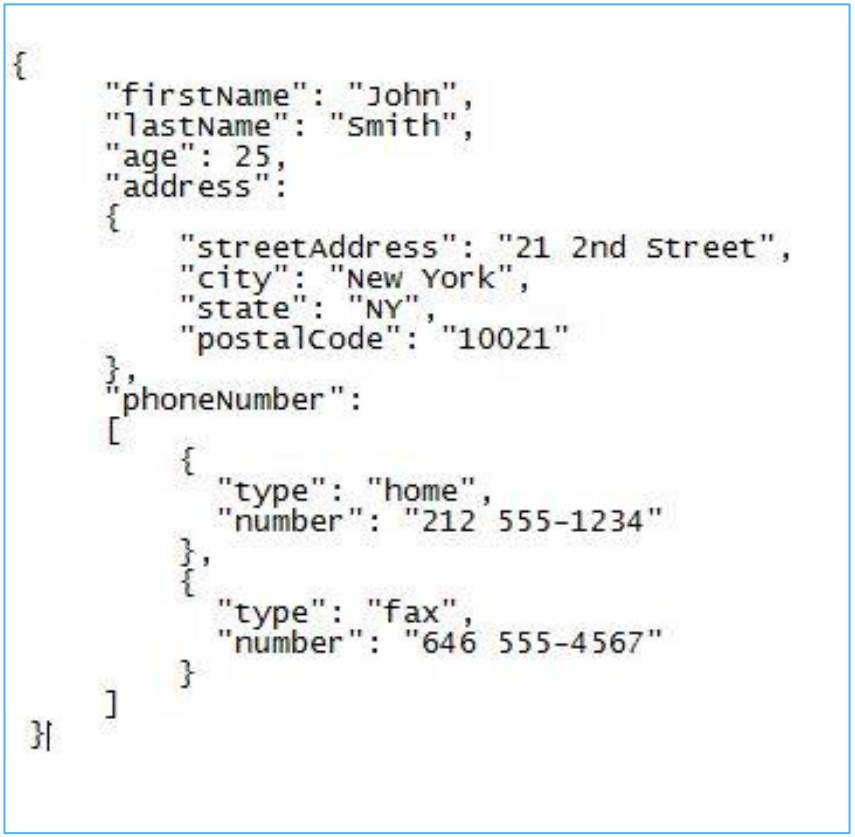
- Eventual Consistency

Data Model: Relational Schema vs. Semi-Structured and Unstructured Data Stores

- Schema and Data Definition (DDL) vs. Schema-Freedom
- Key-Values and Map-of-Maps
- Typed fixed length versus untyped variable length

Analysis of Large Data Sets

- Column-stores are usually more efficient
- Social networks data are naturally modeled as graphs



The CAP Theorem

Consistency of data: each client always has the same view of the data

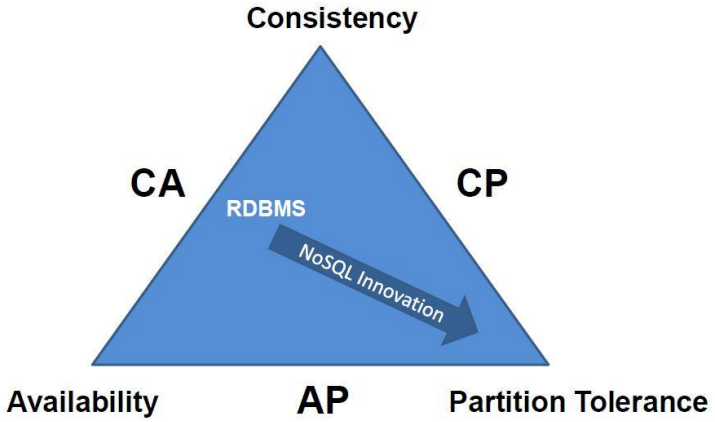
- Total order of all operations such that it appears as if each operation completed atomically. In particular, Reads see Writes that are previous in the total order

Availability: all clients can always read and write

- Pinging a live node should produce results

Partition Tolerance: the system works well across physical network partitions

- A request cannot be blocked forever



Theorem: In a distributed system you have to give up one of the CAP

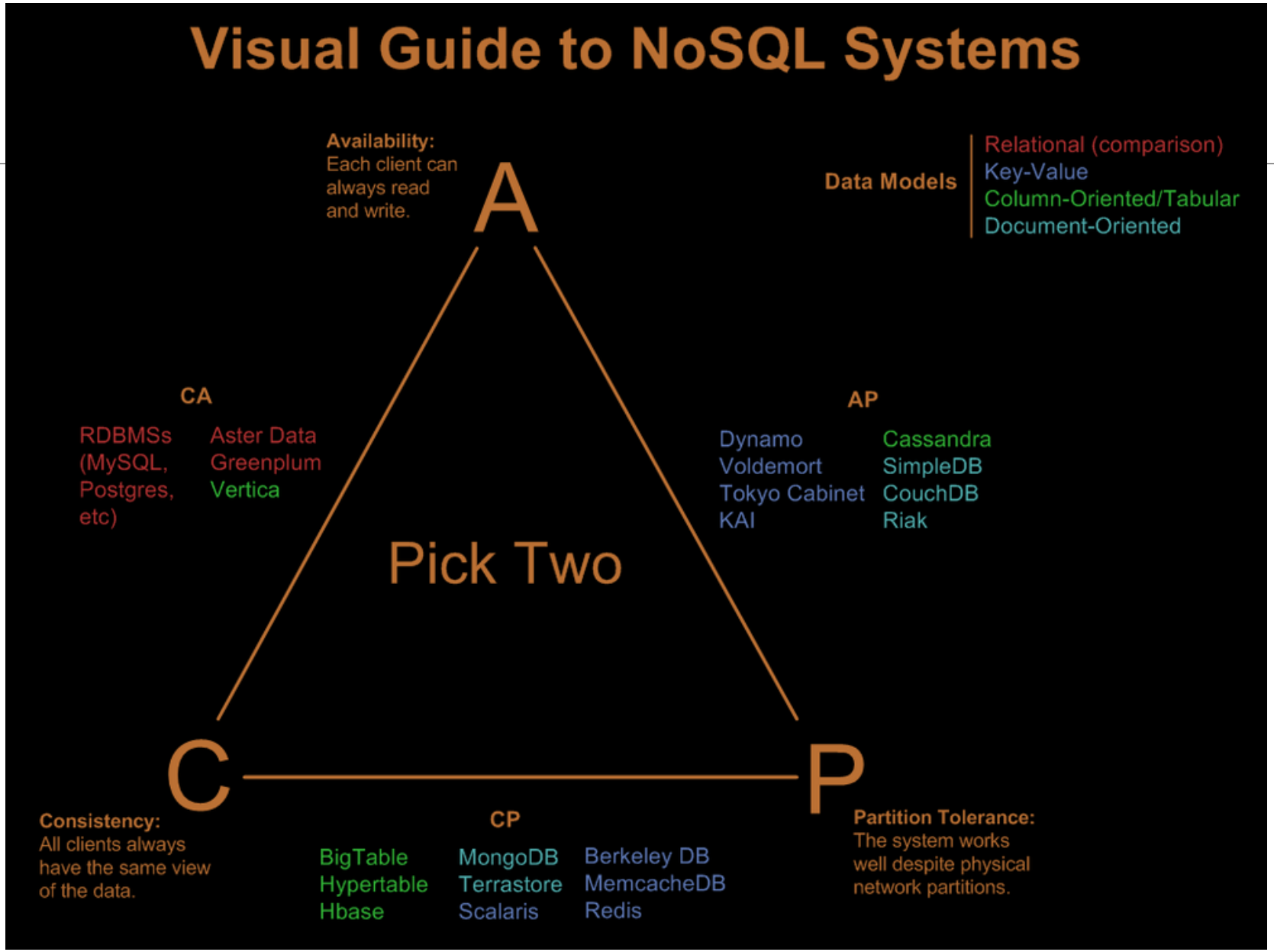
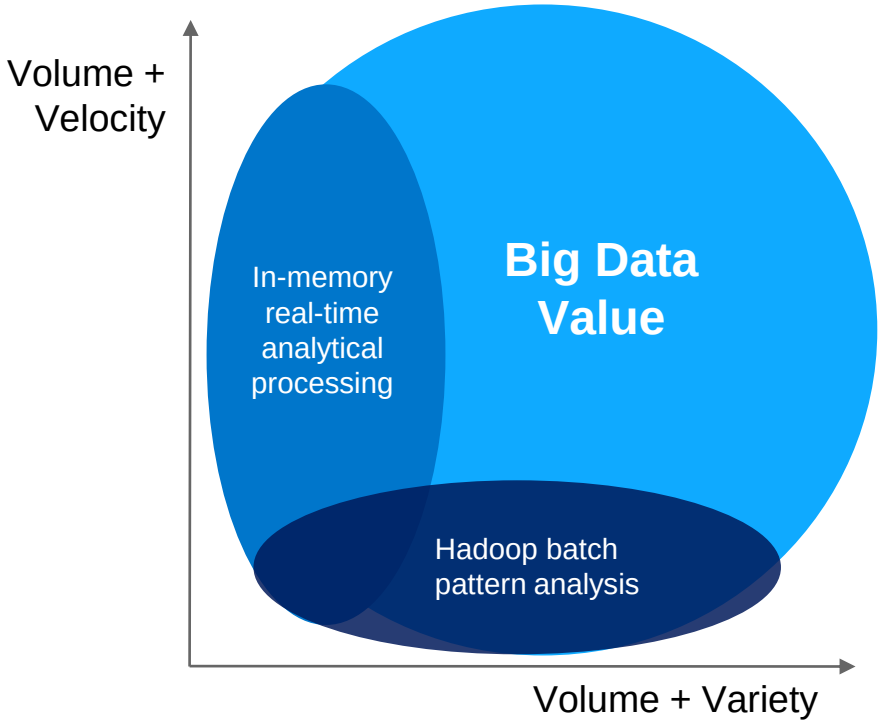


Diagram taken from http://blog.nahurst.com/visual-guide-to-nosql-systems

Two Examples



Understanding the Business Value of Unstructured Data

Some key challenges are

- Bringing meaningful structure into unstructured data
- Relating unstructured, complex data to structured information can provide completely new insights but is very hard
- Identify the meaningful information in the data and reduce waste
- Expenses in storing extreme amounts of data, while keeping the accessible

Example – Public Product Reviews

13 of 22 people found the following review helpful:

☆☆☆☆☆ **Good features, lousy image quality**, June 9, 2008

By [redacted] (State College, PA United States) - See all my reviews

REAL NAME

This review is from: **Samsung S860 8.1MP Digital Camera with 3x Optical Zoom (Black) (Electronics)**

The S860 has a decent feature set for a cheap camera, but beware the image quality. At ISO80 in bright daylight, the S860 managed to introduce immense noise into reds. I took 8 photos of a building with a red banner and had to run each of them through Noise Ninja to make them presentable.

Better than a camera phone? Sure, but worse then just about anything else.

Source: www.amazon.com

STANDARD_FORM	TYPE	CONVERTED_TEXT
Polaroid	PRODUCT	Very disappointed. Because of other duties, I did
Quicktime	PRODUCT	June 12, 2008 I bought this camera yesterday ar
Samsung S860	PRODUCT	Samsung S860 8.1MP Digital Camera with 3x Opt
Samsung S860	PRODUCT	Samsung S860 8.1MP Digital Camera with 3x Opt
Samsung S860	PRODUCT	I bought one of the Kodak C813's and returned it
Samsung S860	PRODUCT	I bought a pink Samsung S860. I used to use cam
Samsung S860	PRODUCT	I own a Canon Power Shot that works perfectly t
Samsung S860	PRODUCT	The Samsung S860 is a low priced camera that is
Samsung S860	PRODUCT	Admittedly, I am a novice. But this camera is perf
Samsung S860	PRODUCT	Admittedly, I am a novice. But this camera is perf
Samsung S860	PRODUCT	I purchased a pink Samsung S860 for my mother.
Samsung S860	PRODUCT	I purchased a pink Samsung S860 for my mother.
Samsung S860	PRODUCT	Let me start off my saying, I am not a camera ex
Samsung S860	PRODUCT	Let me start off my saying, I am not a camera ex
Samsung S860	PRODUCT	The Samsung S860 has features I wanted. It doe
Samsung S860	PRODUCT	I needed a digital camera to take pictures of item
Sony A700	PRODUCT	I also own a Sony A700 but it is a big camera to l

Apache Hadoop/HIVE

Apache Hadoop addresses some of the key challenges mentioned, but leaves some wishes unanswered

- Open-source project administered by the Apache Software Foundation
- Allows for scalable and accessible storage of massive data amounts (structured and unstructured) on commodity hardware clusters
- Designed for **non-real time analysis** of both structured data and complex data

Key Hadoop/HIVE Services:

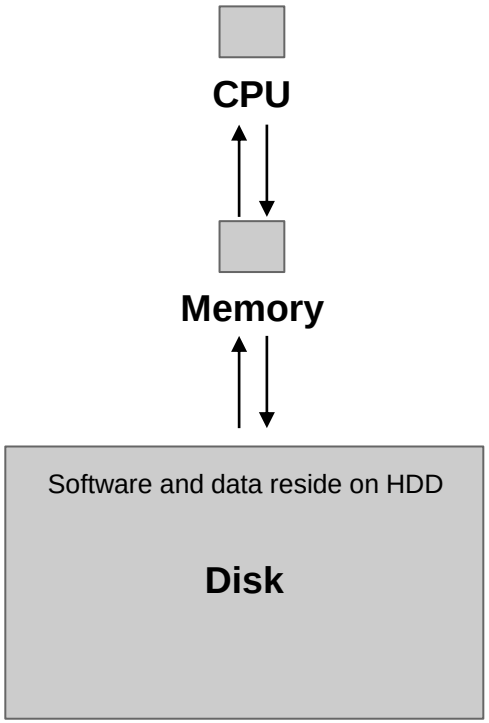
- Reliable data storage using the Hadoop Distributed File System (HDFS) – structured and unstructured
- HIVE is a data warehousing solution on top of Hadoop – direct access to HDFS and Hbase
- Parallel data processing and query execution using MapReduce

Companies starting to adopt Apache Hadoop

- Originally developed and employed by dominant Web companies like Yahoo and Facebook
- Today used in finance, technology, telecom, media and entertainment, government, research institutions and other markets with significant data

In-Memory computing
(e.g. SAP HANA)

Yesterday



- IO constraint
- Support many platforms
- Optimized for None

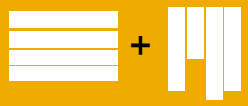
Today

CPU

Multi-Core
Massively Parallel
SingleOptimized Platform

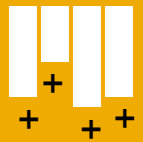
64-bit address space
supports 2TB RAM
100GB/s throughput

Memory

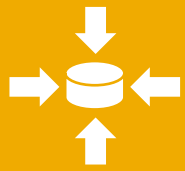


Row and Column Store
No aggregates

Partitioning



Insert Only on Delta



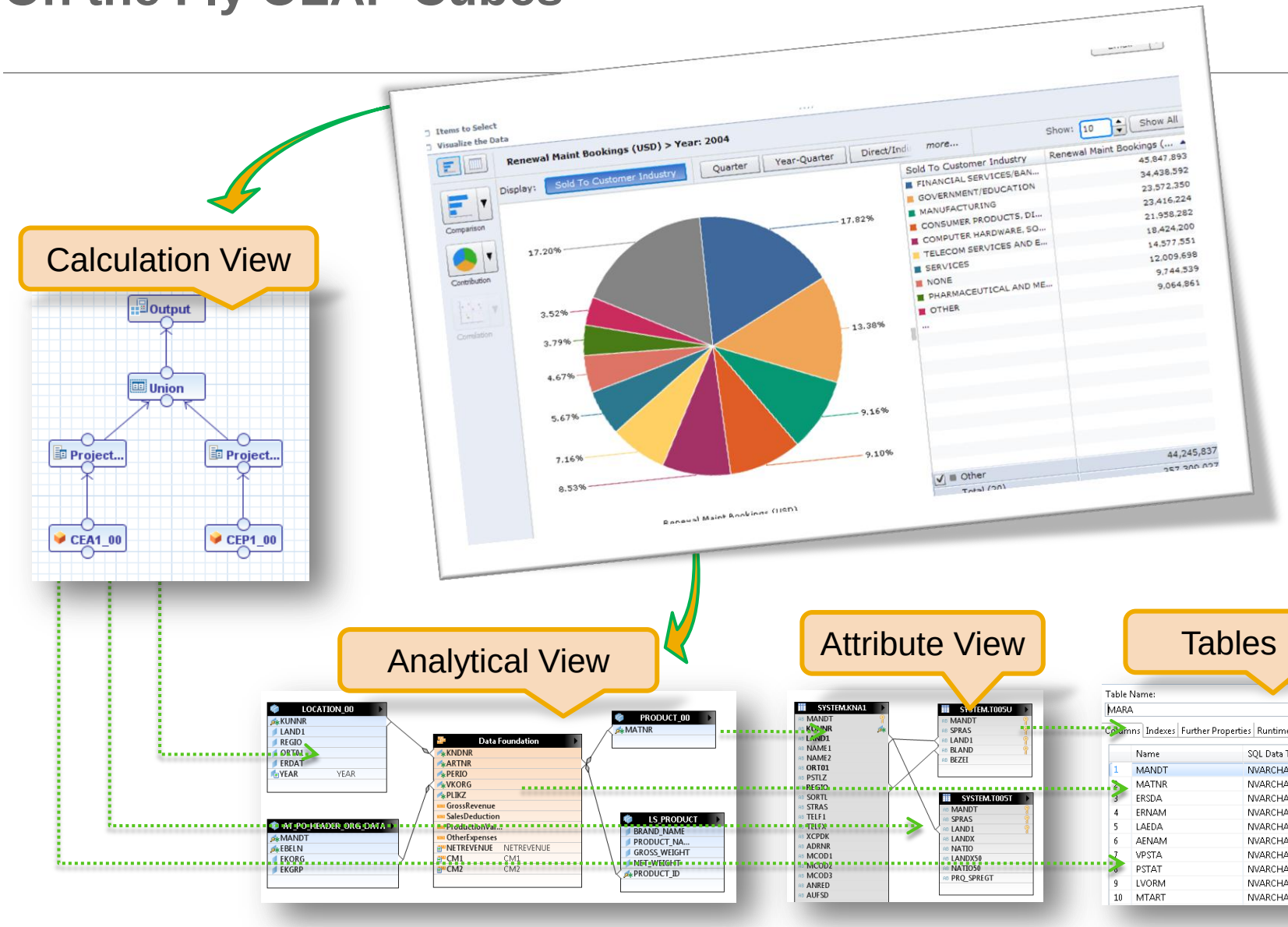
Compression

Disk

Logging and Backup –
Solid State / Flash / HDD

- Take advantage of latest advances in hardware
- Minimum IO time
- Optimized for x86 platform

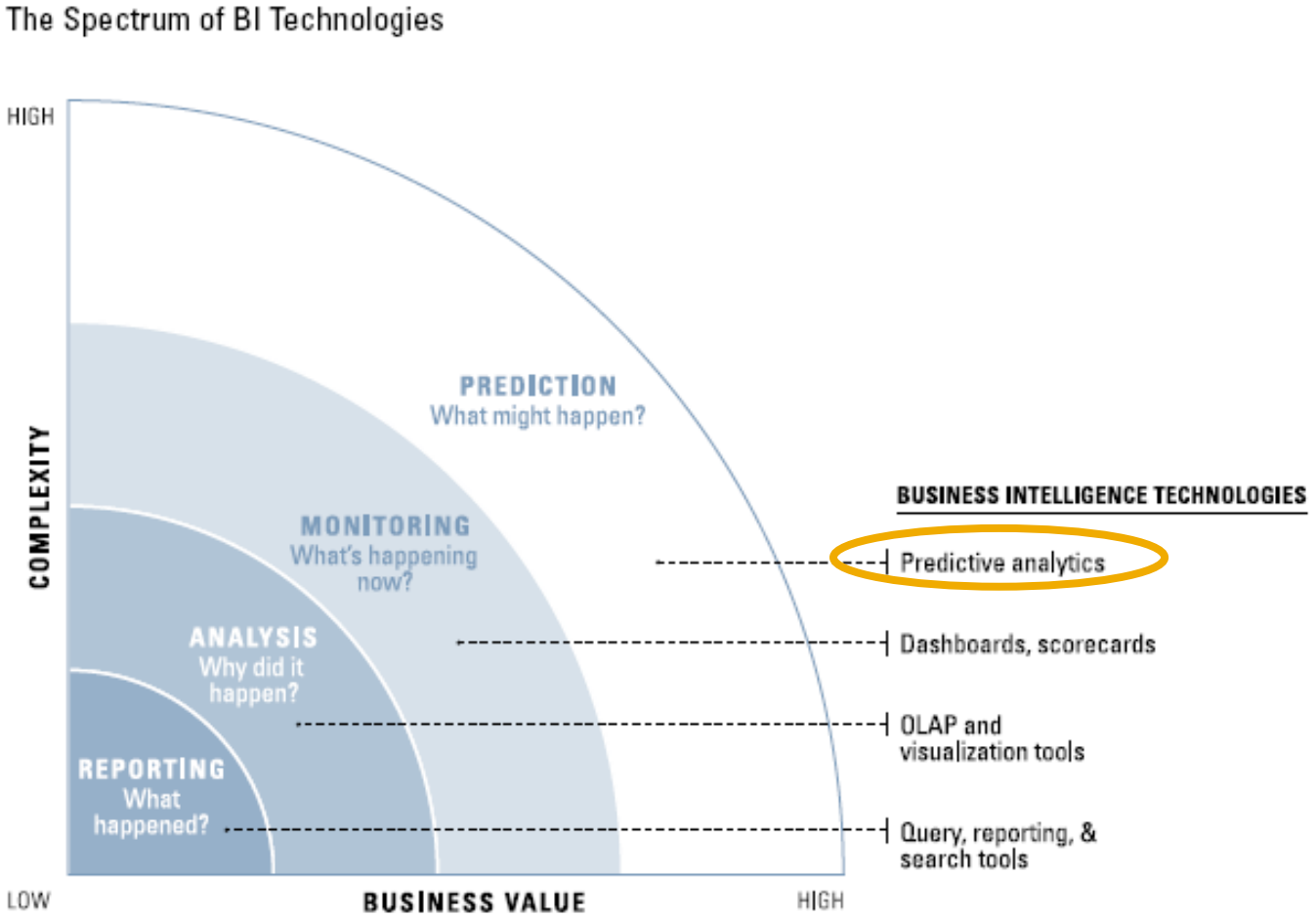
On the Fly OLAP Cubes



Processing and Analysis

The Data Warehouse Institute

"...prediction provides the most business value"

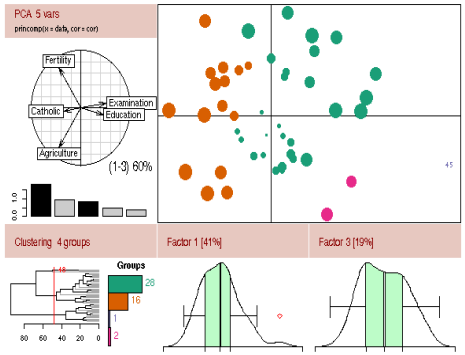


Among business intelligence disciplines, prediction provides the most business value but is also the most complex. Each discipline builds on the one below it—these are additive, not exclusive, in practice

R

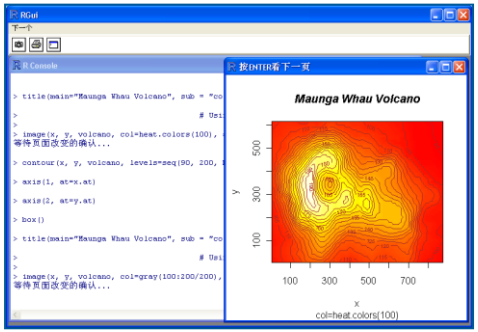
R is a software environment for statistical computing and graphics

- Open Source, programming language plus a run-time environment
- Over 3,500 add-on packages; ability to write your own functions
- Widely used for a variety of statistical methods: linear and non-linear models, statistical tests, time series analyses, classification and clustering, predictive, etc.
- More algorithms and packages than SAS + SPSS + Statistica



Who is using it?

- Growing number of data analysts in industry, government, consulting, and academia
- Cross-industry use: high-tech, retail, manufacturing, CPG, financial services , banking, telecom, etc.

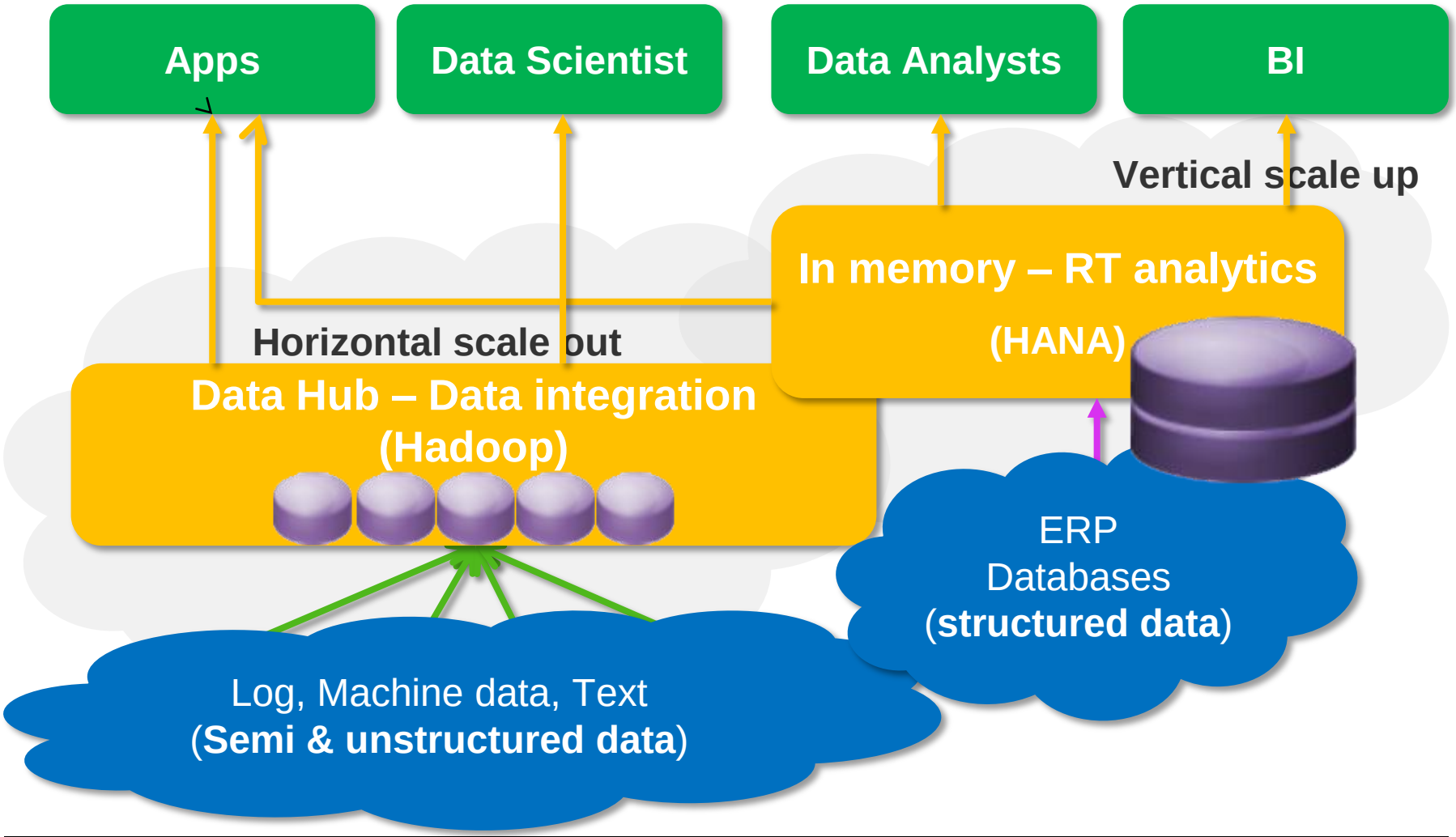


Why are they using it?

- Free, comprehensive, and many learn it at college/university
- Offers rich library of statistical and graphical packages
- Integrated with major analytics offerings (IBM, SAP, Oracle, Cloudera)

Hybrid Analytics Architecture

Horizontal scale out & vertical scale up



Conclusion

How to Capitalize on the Big Data Opportunity and Address Big Data Technical Challenges?

To deploy an integrated data processing framework

Optimize data management in each phase of the information lifecycle process

Regardless of data source, processing technologies, latency challenges, number of user demands

To enable real-time, actionable insights in business process context

Marry business process insights from structured data analysis with deep pattern, behavior analysis of unstructured data

Enable decision making based on multi-factor considerations, not just instinct/experience

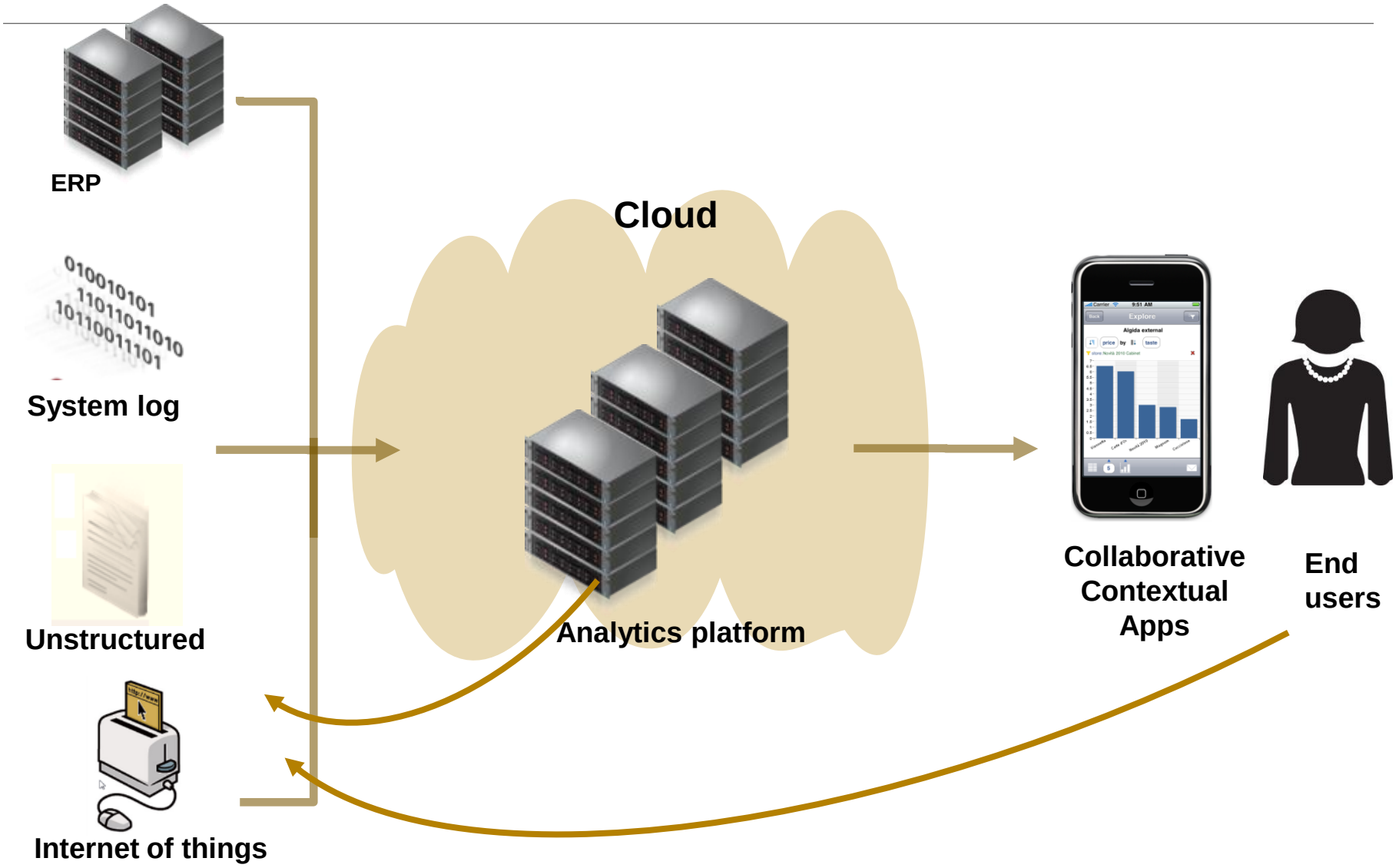
To derive new value from data

Focus on deriving new value from data by enabling new business and technology use cases previously not feasible

Augment existing business scenarios with new data insights to enable better decision

Consumerized & Contextual Applications

Closing the loop - end users and M2M generated data



Thank You!

Contact information:

David Trastour
+33 4 9228
david.trastour@sap.com