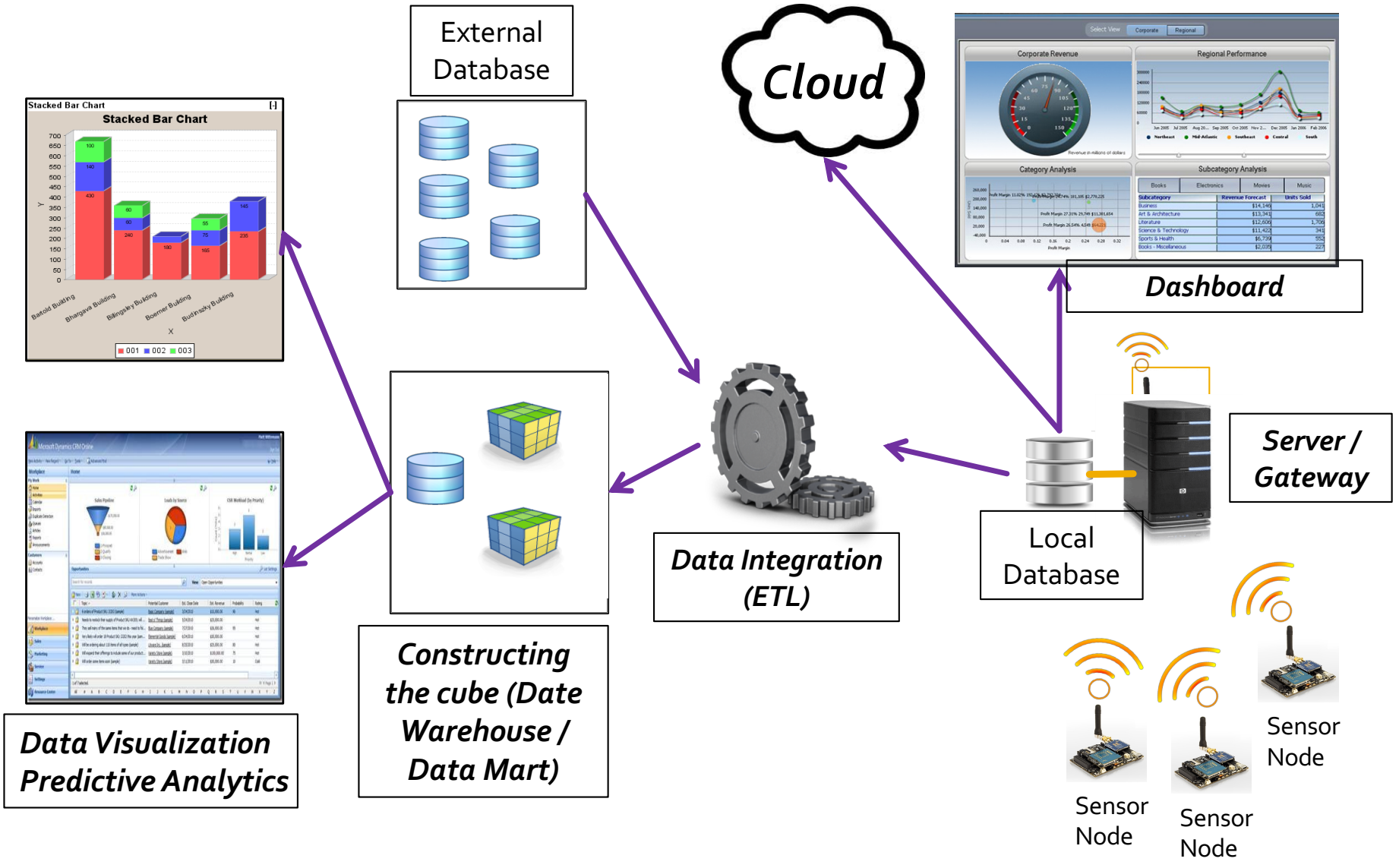
Lo8 Introduction To Big Data & Analytics

IT3779
Smart Object Technologies

Agenda

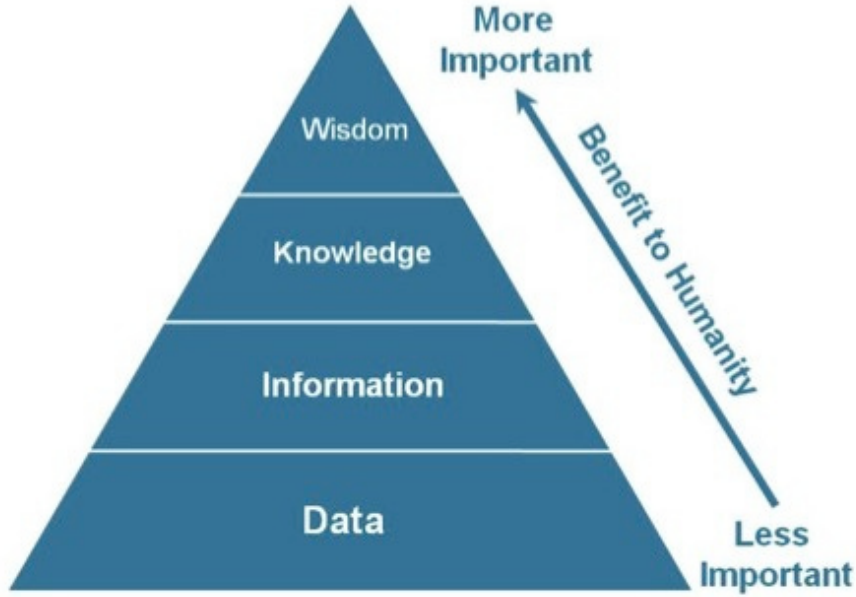
- Define business analytics (BA)
- Compare BA with business intelligence (BI)
- Understand the difference levels of business analytics
- Understand how BA applications can help to create business values

Overview IoT System Architecture



The Intelligence (DIKW) Hierarchy

- **Data** comes in the form of raw observations, e.g. temperature, scores.
- **Information** adds context to data; it is created by analyzing relationships and connections between the data.
- **Knowledge** is created by using the information for action. Knowledge answers the question "how".
- **Wisdom** is created through the use of knowledge and through reflection



Source: Cisco IBSG, April 2011

IoT - Data and Wisdom

- Typical IoT application involves thousands of smart devices.
- Huge amount data are generated and stored.
- What can these data do?
- How can we exploit these data?
- Can we use past rainfall and canal's water level data to forecast flood in Orchard road?
- Can we predict if an elderly will suffer heart attack if we scrutinize their past ECG and heart beat data?

What is Business Analytics (BA)?

Analytics (Thomas)

- is the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions.

Business Analytics (Wikipedia)

- refers to the skills, technologies, applications and practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning.

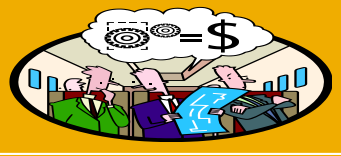
Source: http://en.wikipedia.org/wiki/Business_analytics

What is Business Analytics (BA)?

- Comparison with Business intelligence (BI)

Business Analytics

Focuses on developing new insights and understanding of business performance



Business Intelligence includes

-Data access

-Reporting

-Analytics

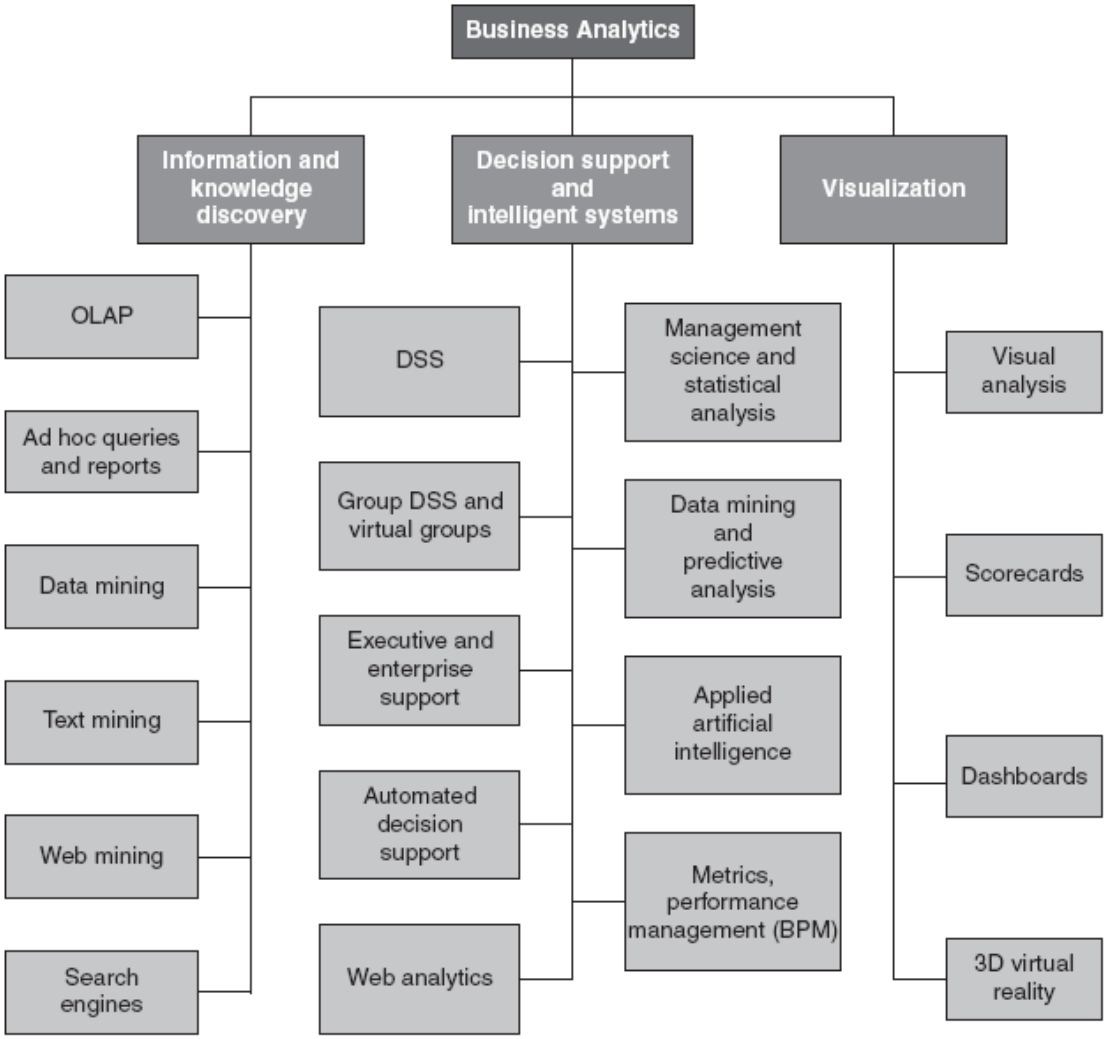
- **Business Analytics is a subset of Business Intelligence**

Business Analytics (BA) Definition

Business Analytics (BA) is the use of analytical methods, either manually or automatically, to **derive relationships from data**.

BA provides the **models** and the **analysis procedures** to Biz Intelligence data. It also involves tracking data and then analyzing them for competitive advantage.

Category of Business Analytics



Levels of Business Analytics

Competitive Advantage increases with degree of intelligence

Advanced Business Analytics

Basic BA: Access and reporting

Standard Reports

- Answer the questions:
  - What happened? When did it happen?
- Example
  - Monthly or quarterly weather reports
    - Generated on regular basis
    - Describe just "what happened" in a particular area
    - Useful for short-term, not for making long-term decision

Monthly Data Report for 2010

Notes on Data Quality.

VANCOUVER INT'L A
BRITISH COLUMBIA

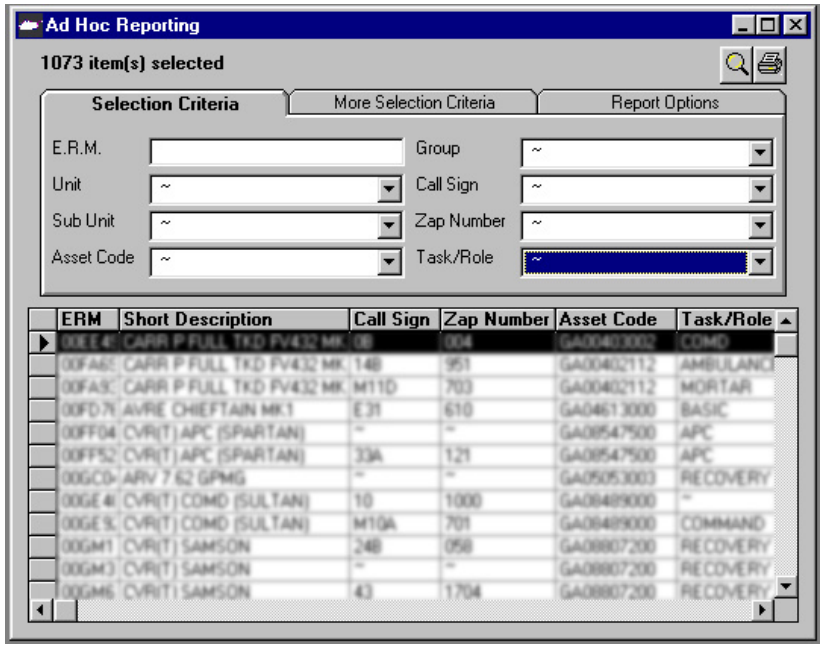
Latitude: 49°11'42.000" N Longitude: 123°10'55.000" W Elevation: 4.30 m
Climate ID: 1108447 WMO ID: 71892 TC ID: YVR
Previous Year Next Year

Monthly Data Report for 2010

Month	Mean Max Temp °C	Mean Temp °C	Mean Min Temp °C	Extr Max Temp °C	Extr Min Temp °C	Total Rain mm	Total Snow cm	Total Precip mm	Snow Grnd Last Day cm	Dir of Max Gust 10's deg	Spd of Max Gust km/h
Sum						M	M	M			
Avg	M	M	M								
Xtrm				M	M					M	M
Summary, average and extreme values are based on the data above.											
Jan	9.9	7.2	4.5	14.1	-2.7	182.8	0.0	182.8	0	13E	82E
Feb	10.3	7.1	3.9	13.1	-0.4S	102.2	0.0	102.2	0	13E	54E
Mar	11.2	7.7	4.2	16.4	0.2	108.2	0.0	108.2	0	29E	76E
Apr	13.1	9.6	6.0	20.4	1.0	88.0	0.0	88.0	0	29E	93E
May	15.8	12.0	8.2	21.1	2.6	54.2	0.0	54.2	0	30E	67E
Jun	18.5	15.0	11.4	23.1	6.3	48.4	0.0	48.4	0	28E	54E
Jul	22.5	18.1	13.7	28.1	10.0	0.6	0.0	0.6	0	28E	59E
Aug	22.4	18.2	14.0	29.8	10.4	69.6	0.0	69.6	0	11E	46E
Sept											
Oct	14.4	11.3	8.2	21.8	3.7	76.2	0.0	76.2	0	12E	59E
Nov	7.8	5.1	2.4	15.8	-9.5	126.4	16.8	143.0	0	26E	70E
Dec	7.6	4.9	2.2	11.8	-6.3	167.4	T	167.4	0	19E	59E

Ad Hoc Reports

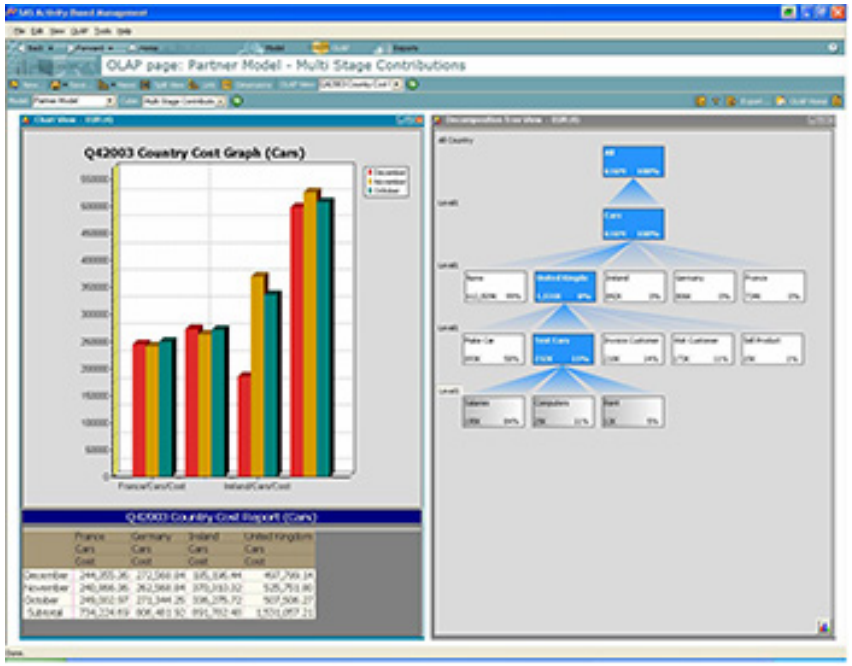
- Answer the questions:
  - How many? How often? Where?
- Example
  - Customer reports that describe the number of hospital patients for every diagnosis code for each day of the week
    - Let you ask the questions and request a couple of custom reports to find the answers



Query Drilldown (or OLAP)

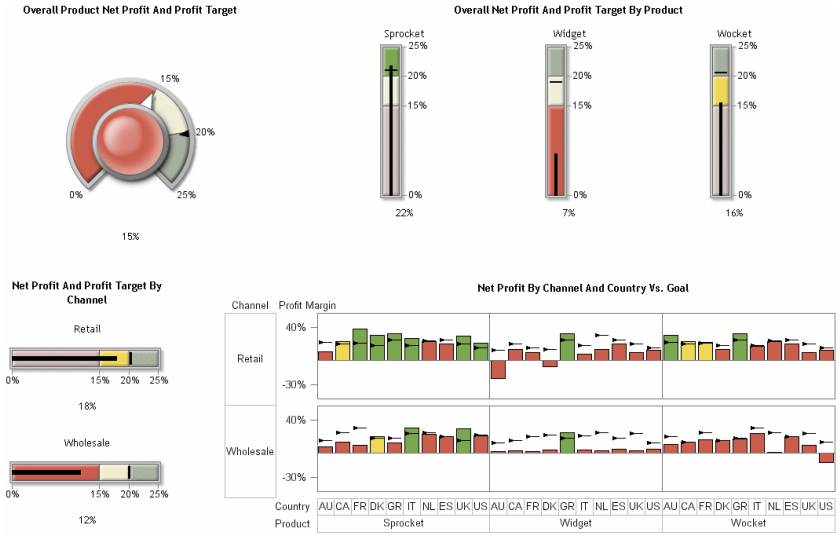
OLAP: OnLine Analytical Processing

- Answer the questions:
  - Where exactly is the problem? How do I find the answers?
- Example
  - Sort and explore data about different types of cell phone users and their calling behaviours
    - Query drilldown allows a little bit of discovery
    - OLAP lets you manipulate the data yourself to find out how many, what colour and where



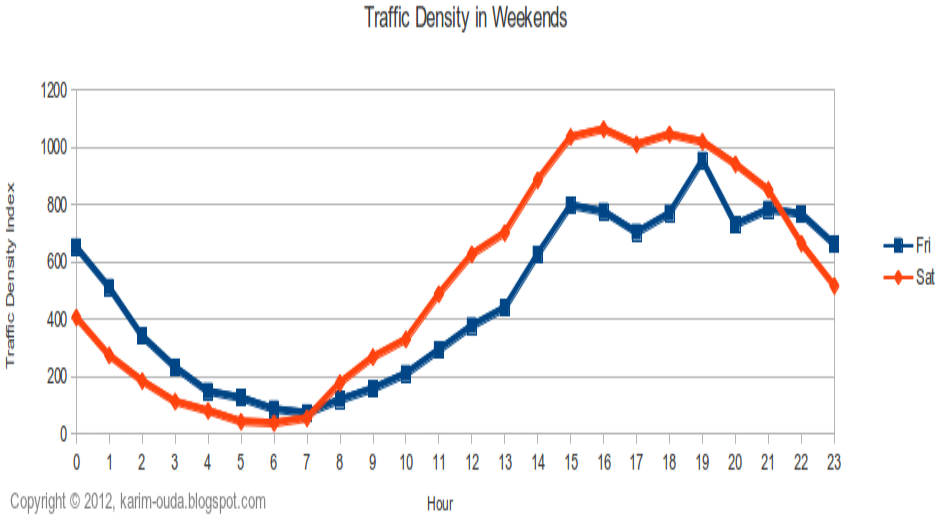
Alerts

- Answer the questions:
  - When should I react? What actions are needed now?
- Example
  - Purchasing executives receive alerts when inventory are low
    - With alerts, you learn when you have a problem and be notified when something similar happens again in the future
    - Alerts can appear via email, sms or as red dials on a scorecard or dashboard



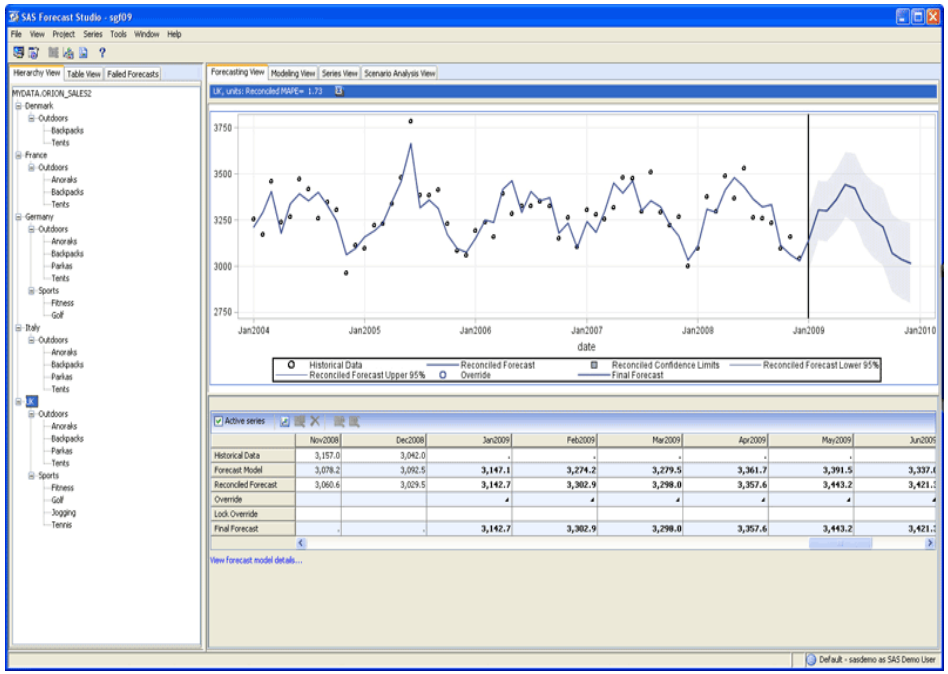
Statistical Analysis

- Answer the questions:
  - Why is this happening? What opportunities am I missing?
- Example
  - Transport authority can discover why there traffic jam is more common in weekend evening
    - Can run complex analytics like frequency models and regression analysis



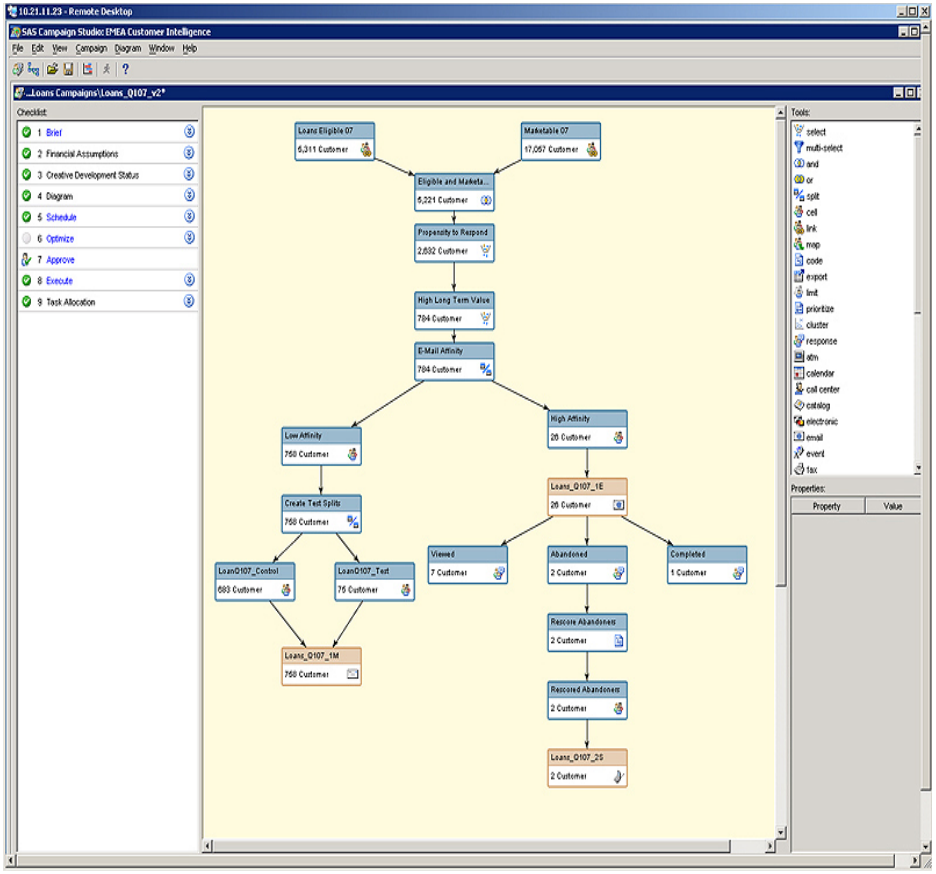
Forecasting

- Answer the questions:
  - What if these trends continue? How much is needed? When will it be needed?
- Example
  - Retailers can predict how demand for individual products will vary from store to store
    - Forecasting demand helps supply just enough inventory



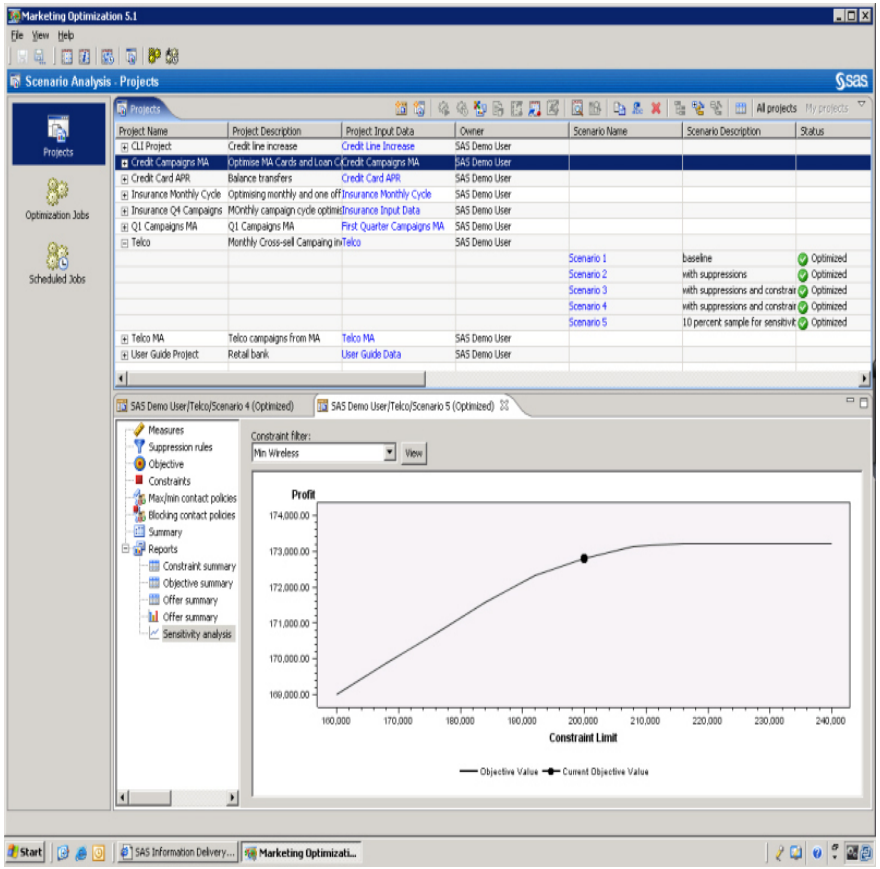
Predictive Modeling

- Answer the questions:
  - What will happen next? How will it affect my business?
- Example
  - Hotels and casinos can predict which VIP customers will be more interested in particular packages
    - With 10 million customers, who's most likely to response to a marketing campaign?
    - How do you segment that group?

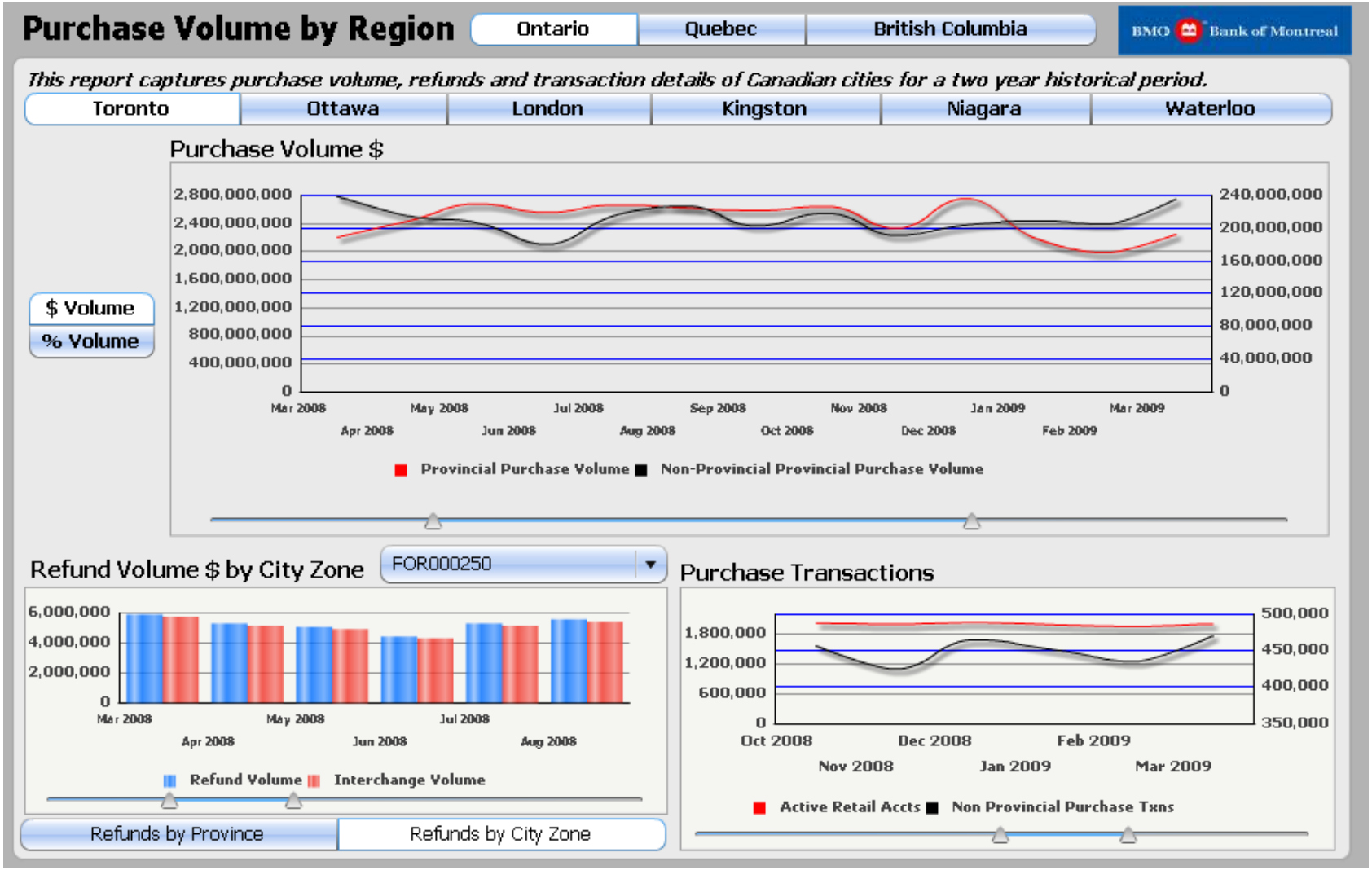


Optimization

- Answer the questions:
  - How do we do things better? What is the best decision for a complex problem?
- Example
  - Given business priorities, resource constraints and available technology, determine the best way to optimize your IT platform to satisfy the needs of every user
    - Optimization supports innovation

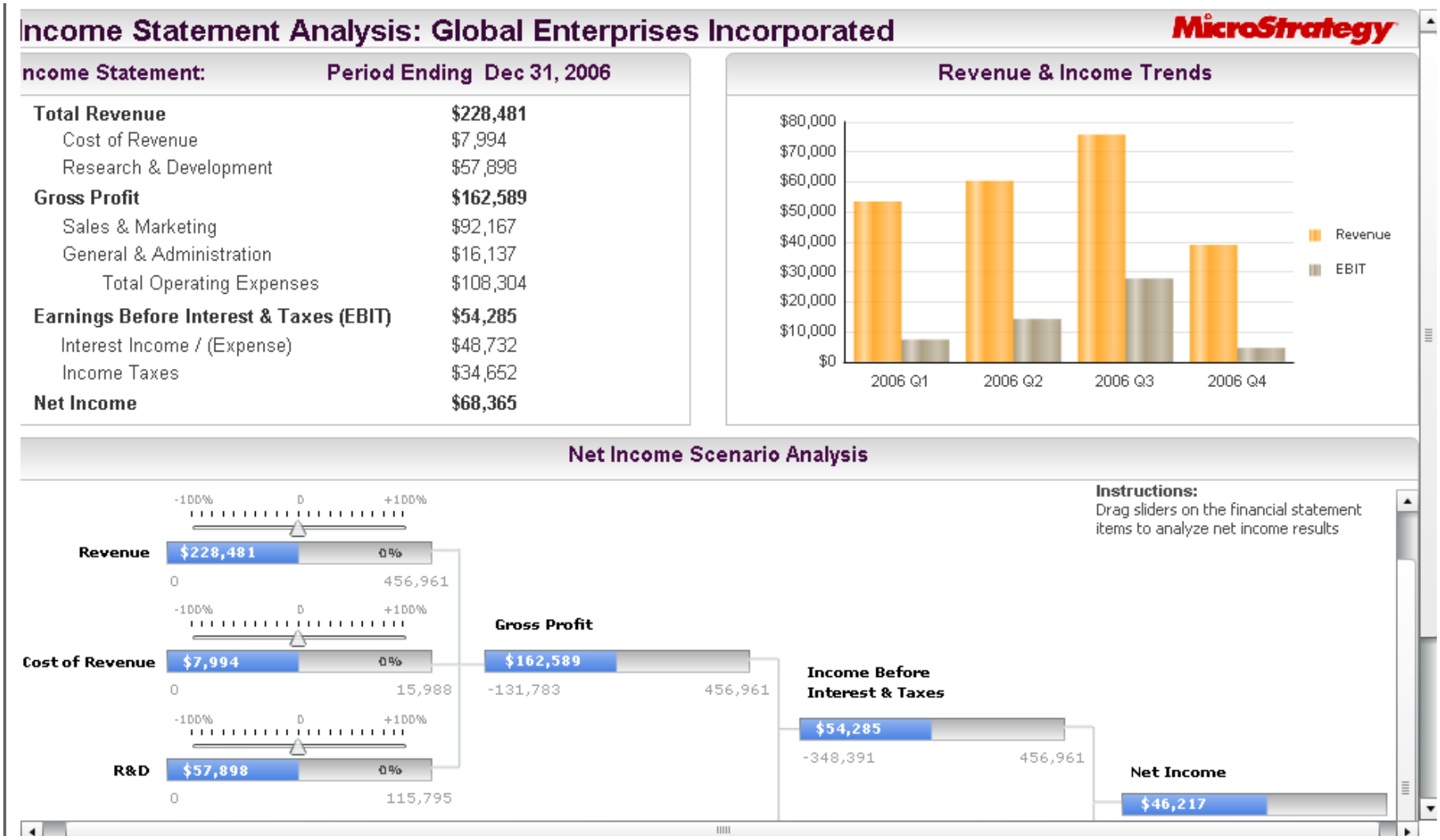


Example: Which Level of BA is this?



http://www.microstrategy.com/DashboardGallery/Dashboards/customers/bmo/

Example: Which Level of BA is this?



http://microstrategy.com/DashboardGallery/Dashboards/IncomeStatementSilver.htm

BA Example Scenario: Retention in the mobile phone industry

- Marketing manager realizes a large number of customers discontinuing their services.
- Customer attribution or churn is a critical factor for companies in the service industries.
- Given a budget adequate to pursue a customer retention campaign at 2000 individuals out of a customer base of 2 million people
  - How to find these 2000 customers to optimize the effectiveness of the campaign?
  - Find an estimate probability that a single customer will discontinue service.
  - Target this group of customers to maximize retention and reduce churning.

BA Example Scenario: Credit Risk Management at Bankcards

- Who is making money for the organization versus who is risky?
  - Evaluate different aspects of customers
  - Profile project / customer segmentation
- How to manage the credit card life cycle of a customer?
  - Identify those get the cards but do not want them
  - Target the right market segments
  - Retention program to address customer needs at later stage of life cycle
- How to avoid risk later in customer's life cycle?
  - Evaluate different customer segments
  - Differentiate treatments for different segments

Business Values of BA Applications

No.	BA Applications	Business Values
1.	Customer segmentation • What market segments do my customers fall into? • What are their characteristics?	Personalize customer relationships for higher satisfaction & retention.
2.	Tendency to buy • What customers are most likely to respond to my promotion?	Target customers based on their need to increase their loyalty to your product line. Increase campaign profitability by focusing on those most likely to buy.

Business Values of BA Applications

No.	BI Analytics	Business Value
3.	Customer profitability • What is the lifetime profitability of my customer?	Make individual business interaction decisions based on the overall profitability of customers.
4.	Customer attrition • Which customer is at risk of leaving?	Prevent lost of high-value customers and let go of low-value customers.
5.	Channel optimization • What is the best channel to reach my customer in each segment?	Interact with customers based on their preference and your need to manage cost.

Business Values of BA Applications

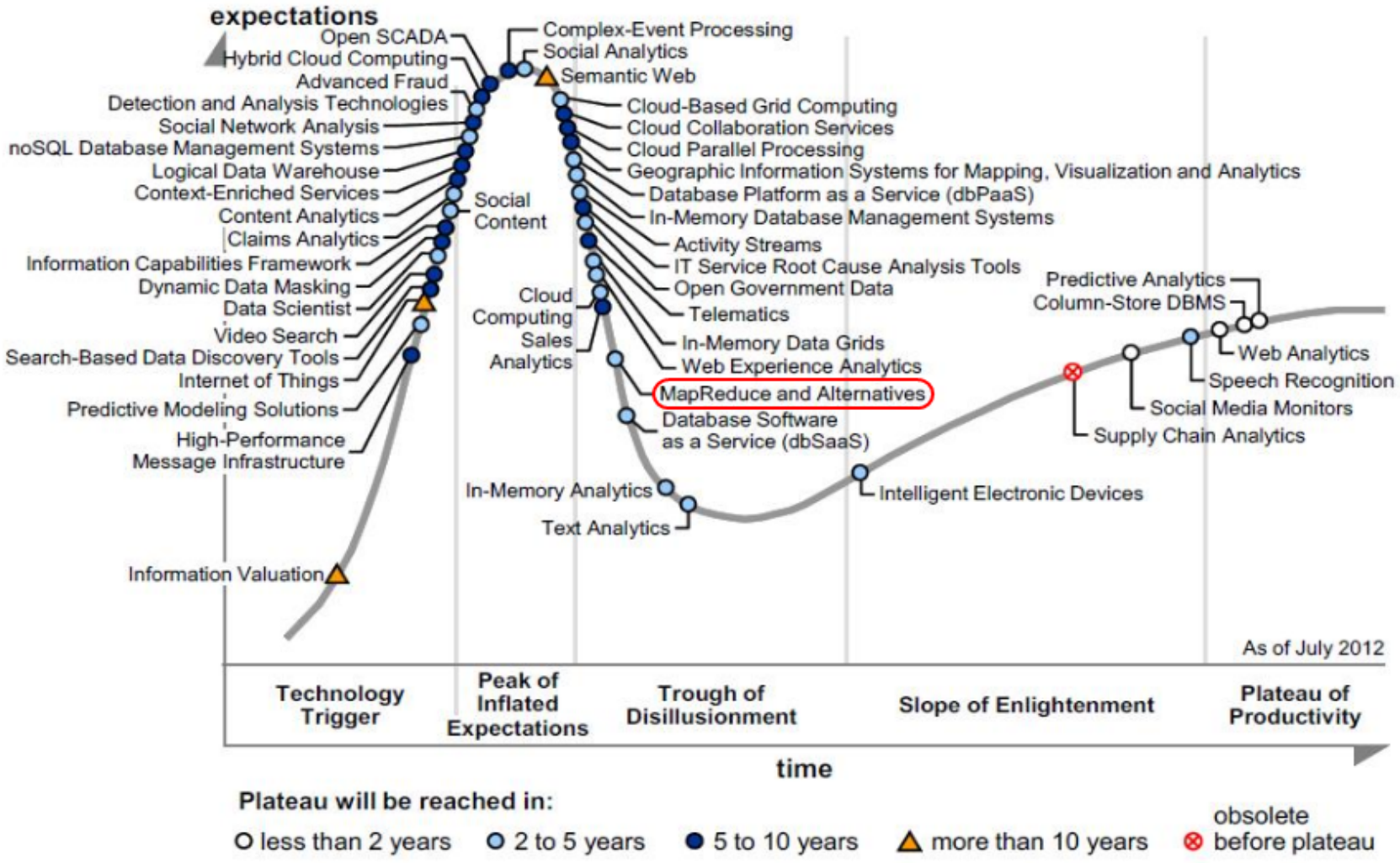
No.	BI Analytics	Business Value
6.	Fraud detection • How can I tell which transactions are likely to be fraudulent?	Quickly determine fraud and take immediate action to minimize cost.
7.	Credit scoring • Which customer will successfully repay his loan? • Which customer will not default on his credit card payment?	Prevent lost to organization due to default.

New Trend Data Analytics …

Big Data

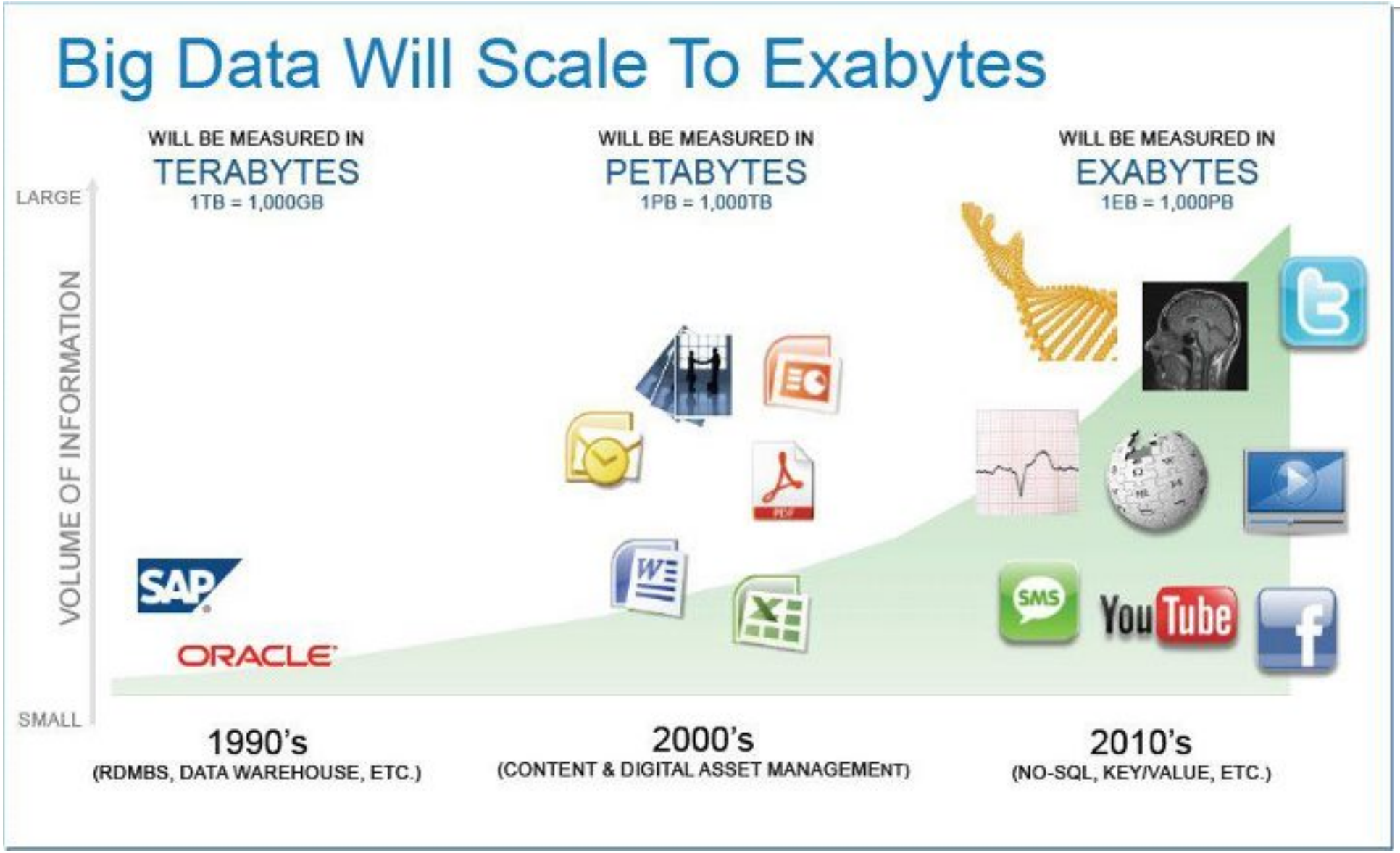
Big Data

Figure 1. Hype Cycle for Big Data, 2012

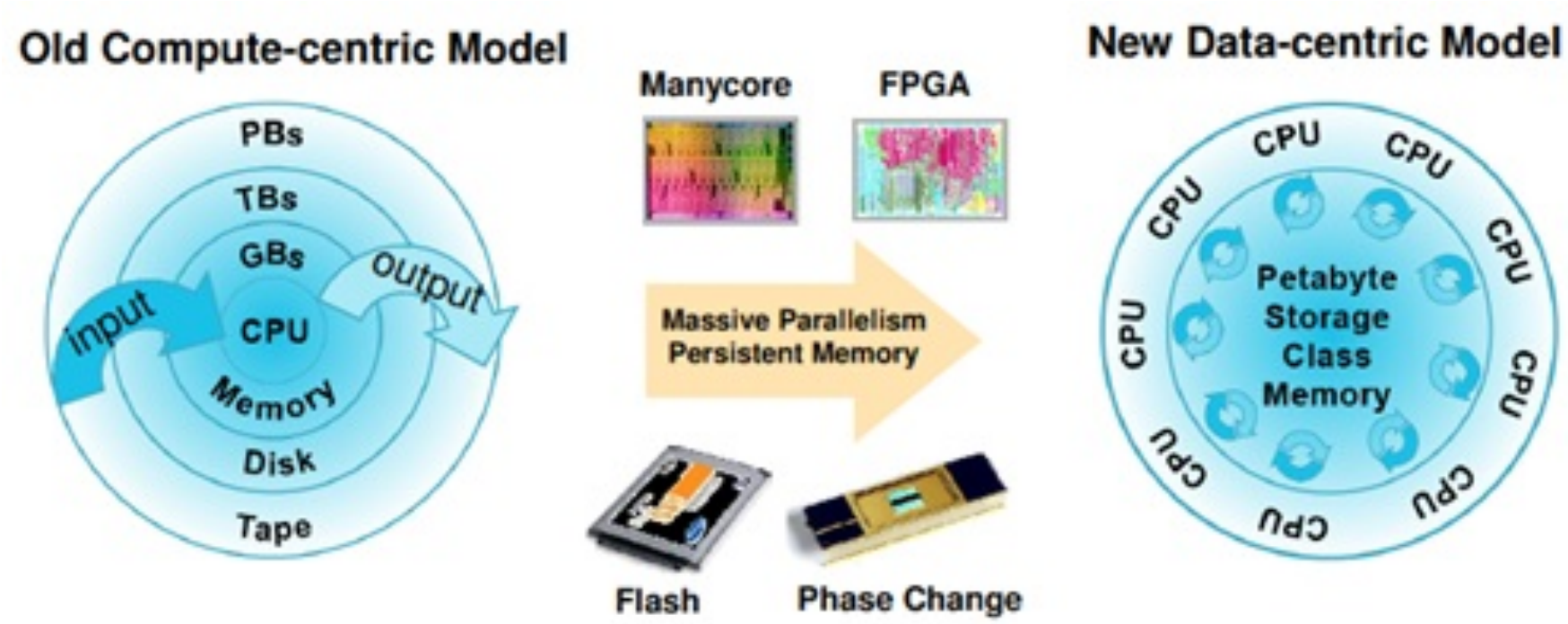


Source: Gartner (July 2012)

Big Data

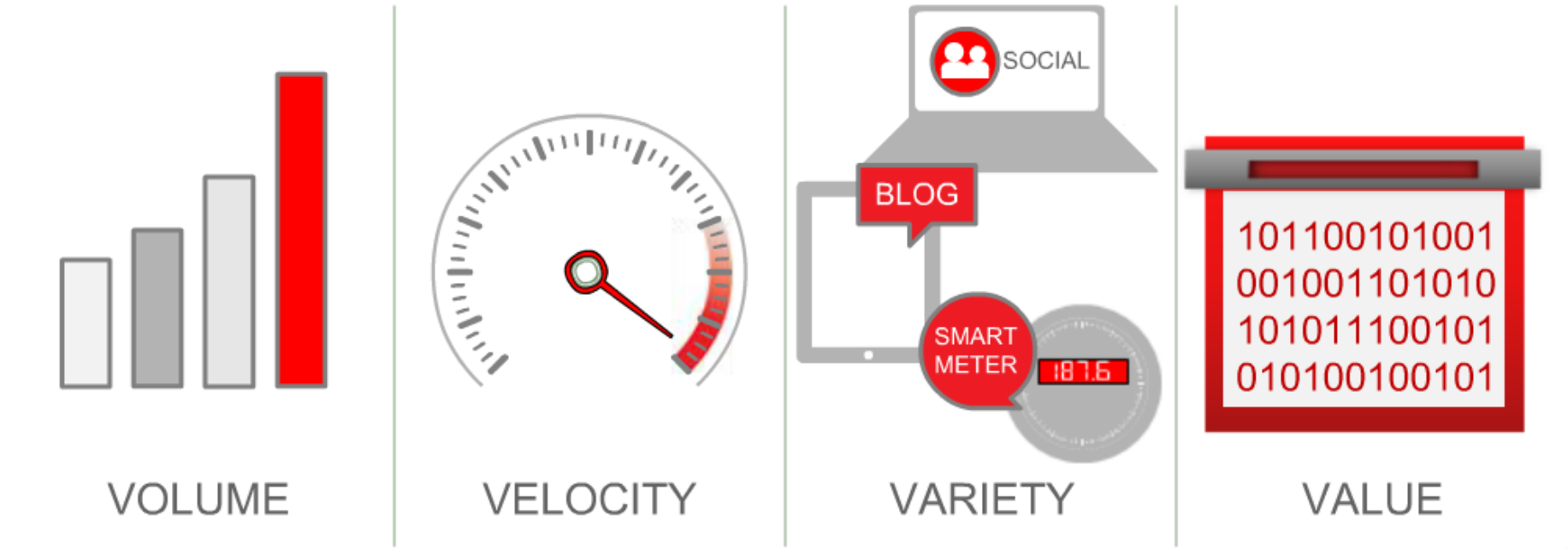

Big Data Will Scale To Exabytes
WILL BE MEASURED IN
TERABYTES
1TB = 1,000GB
WILL BE MEASURED IN
PETABYTES
1PB = 1,000TB
WILL BE MEASURED IN
EXABYTES
1EB = 1,000PB
LARGE
VOLUME OF INFORMATION
SMALL
SAP
ORACLE
SMS
YouTube
1990's
(RDMBS, DATA WAREHOUSE, ETC.)
2000's
(CONTENT & DIGITAL ASSET MANAGEMENT)
2010's
(NO-SQL, KEY/VALUE, ETC.)

Big Data


Old Compute-centric Model
PBs
TBs
GBs
input
CPU
output
Memory
Disk
Tape
Manycore
FPGA
Massive Parallelism
Persistent Memory
Flash
Phase Change
New Data-centric Model
CPU
Petabyte
Storage
Class
Memory

Big Data

What Makes it Big Data?

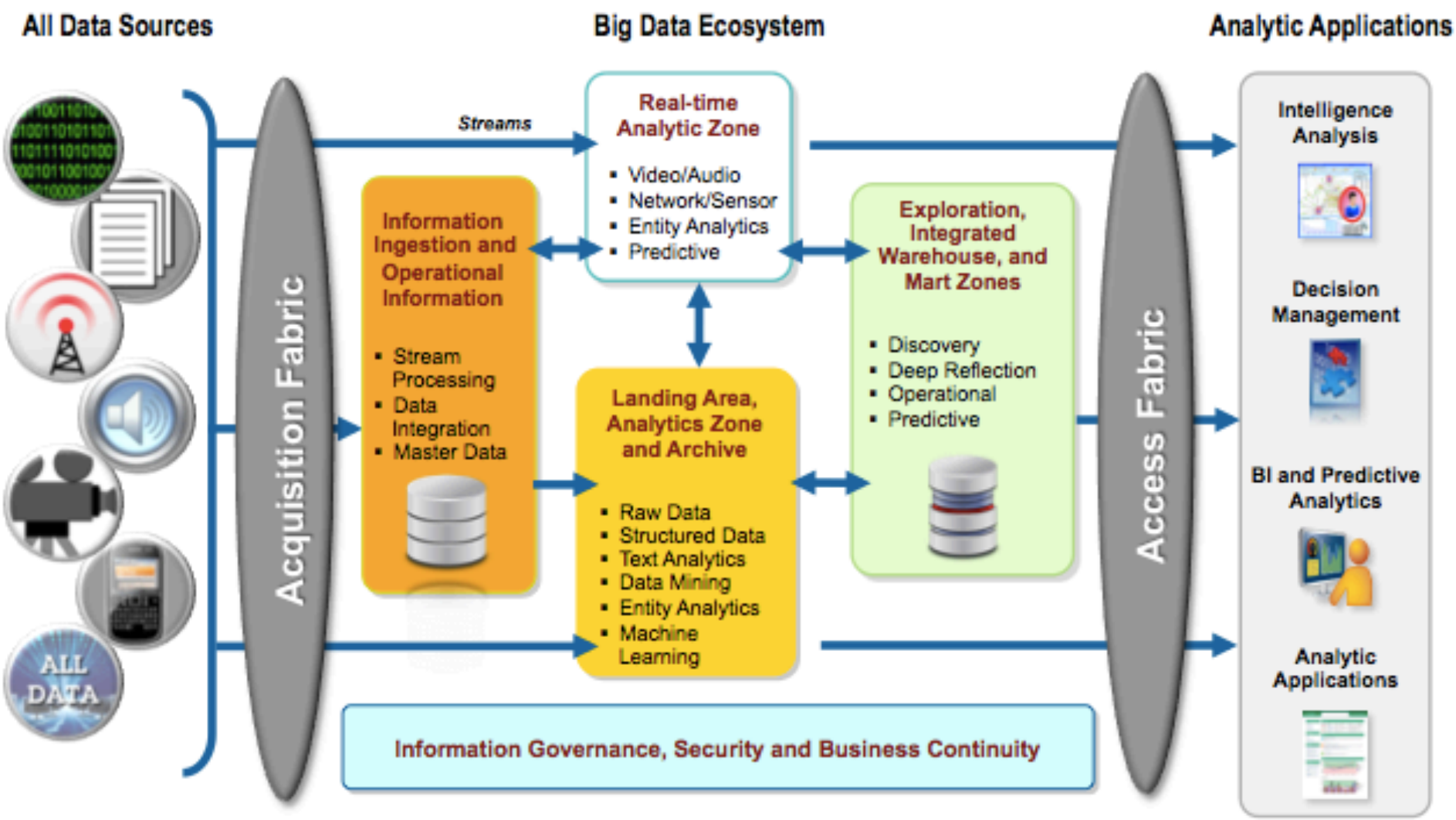


ORACLE

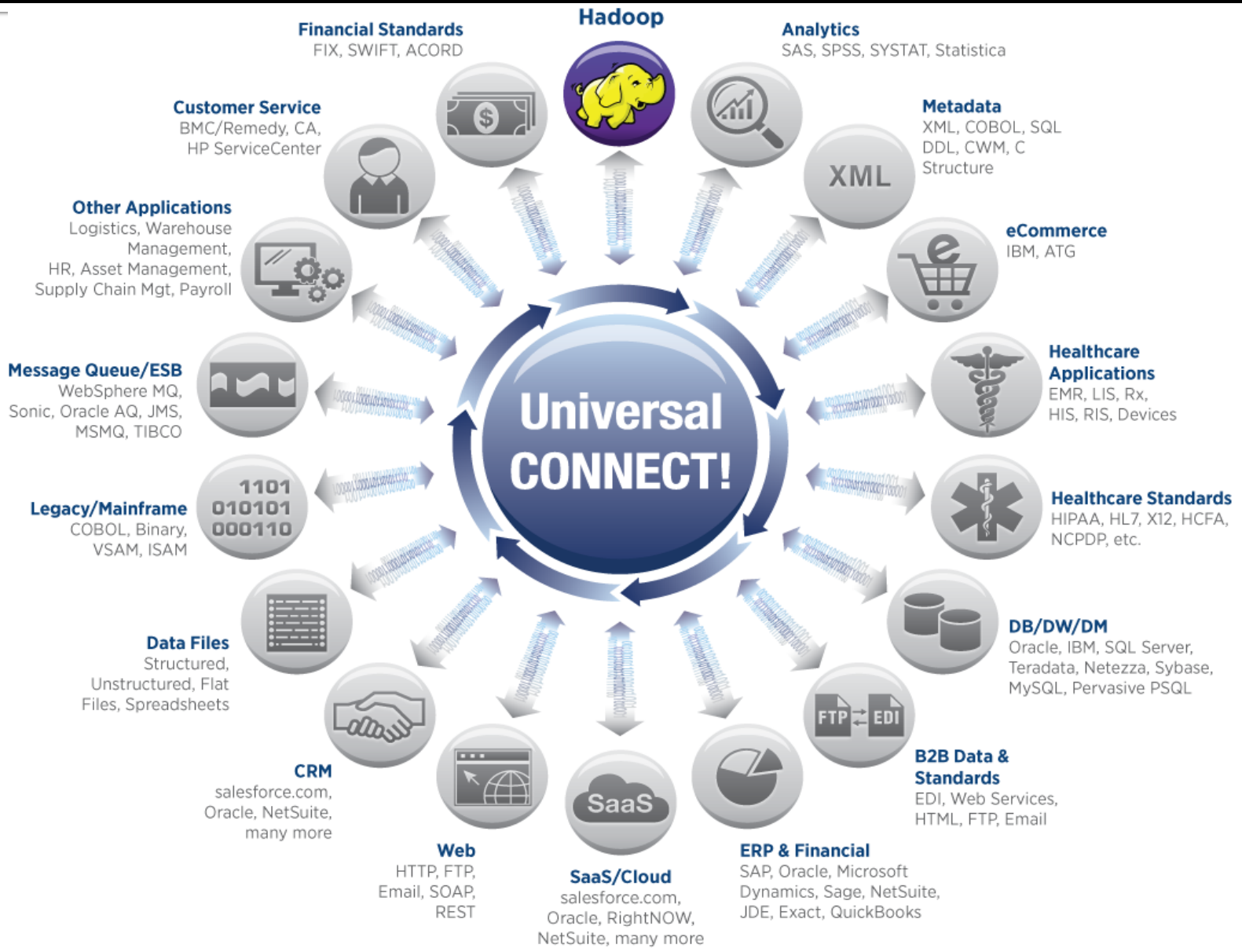
Sources of Data for Big Data

- Documents
- Existing relational databases
  (CRM, ERP, Accounting, Billing)
- E-mails and attachments
- Imaging data (graphs, technical plans)
- Sensor or device data
- Internet search indexing
- Log files
- Social media (FB, TW, INSTA, etc)
- Telephone conversations
- Videos
- Pictures
- Clickstreams (clicks from users on web pages)

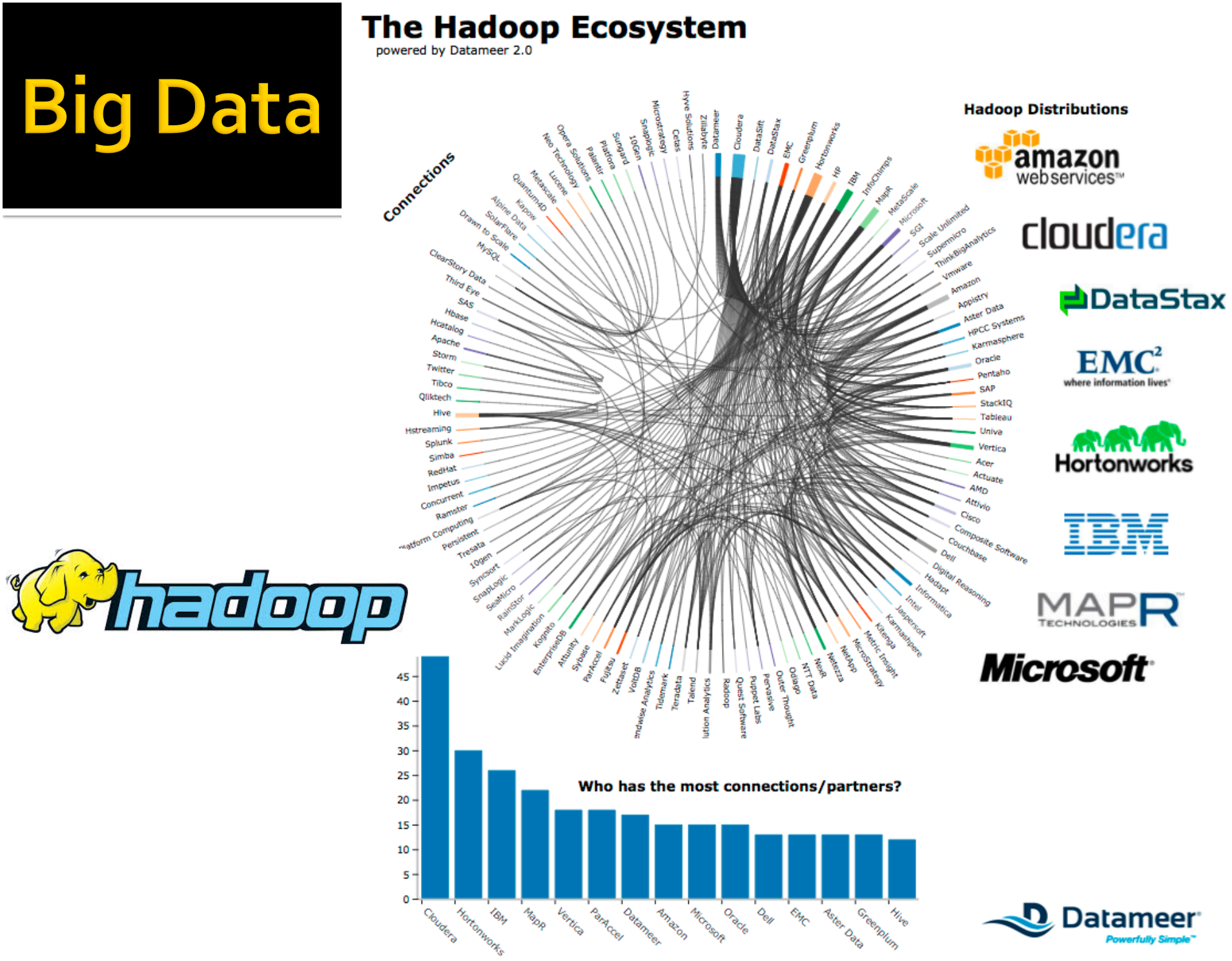
Big Data


All Data Sources
Big Data Ecosystem
Analytic Applications
ALL DATA
Acquisition Fabric
Streams
Real-time Analytic Zone
Video/Audio
Network/Sensor
Entity Analytics
Predictive
Information Ingestion and Operational Information
Stream Processing
Data Integration
Master Data
Landing Area, Analytics Zone and Archive
Raw Data
Structured Data
Text Analytics
Data Mining
Entity Analytics
Machine Learning
Exploration, Integrated Warehouse, and Mart Zones
Discovery
Deep Reflection
Operational
Predictive
Access Fabric
Intelligence Analysis
Decision Management
BI and Predictive Analytics
Analytic Applications
Information Governance, Security and Business Continuity

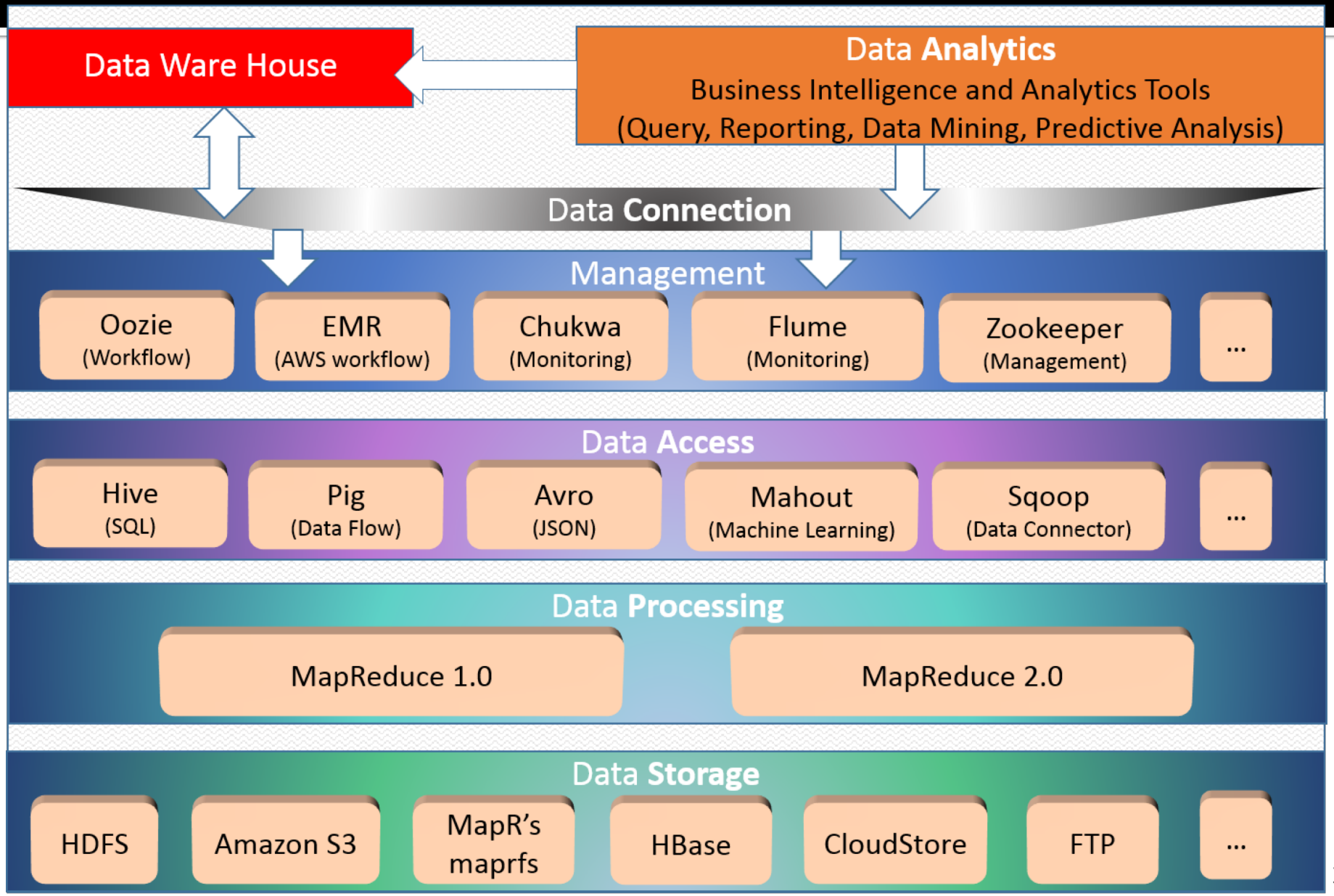
Big Data - Hadoop



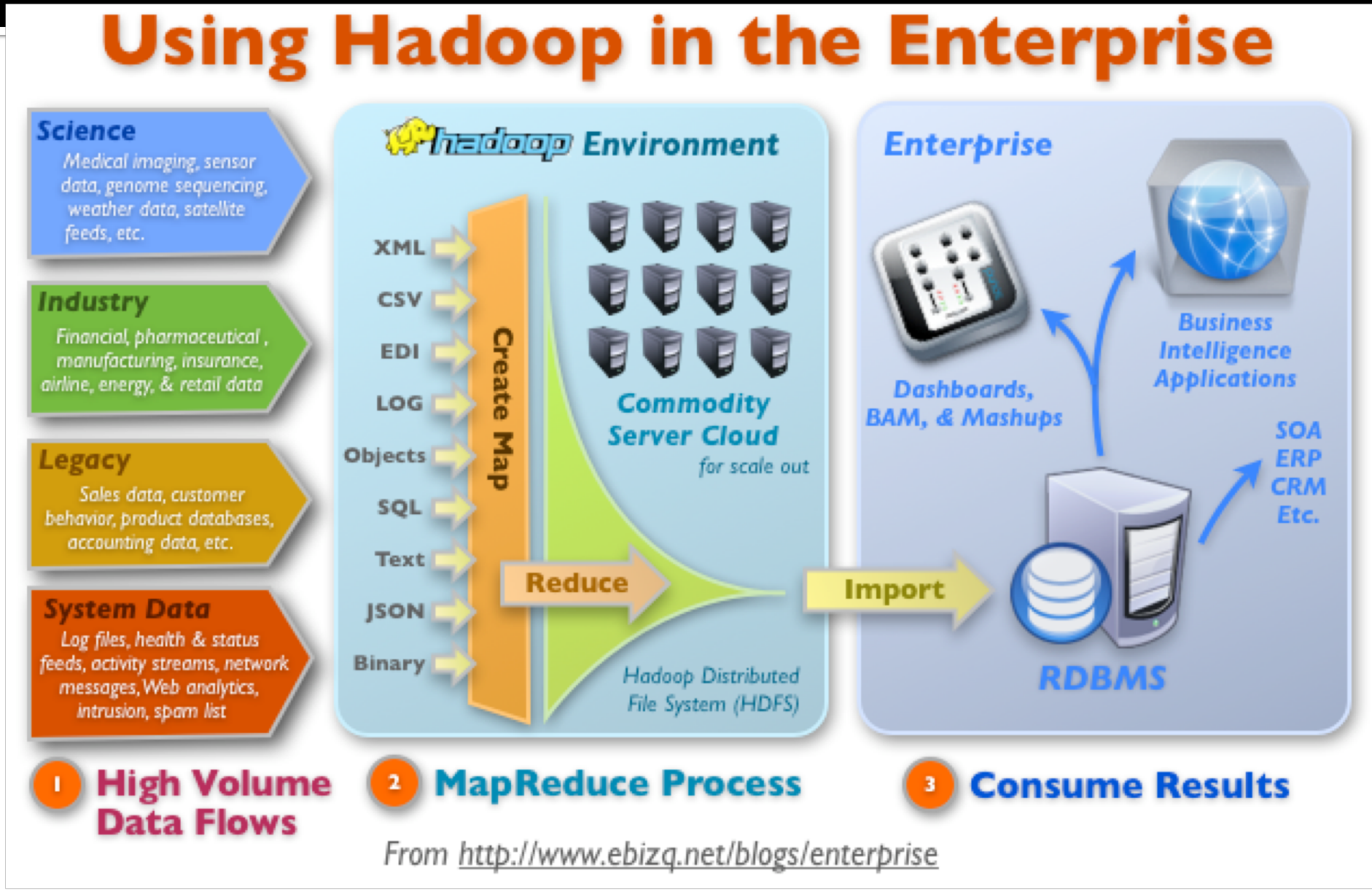
Big Data


The Hadoop Ecosystem
powered by Datameer 2.0
Connections
Hadoop Distributions
amazon web services
cloudera
DataStax
EMC²
where information lives
Hortonworks
IBM
MAPR
TECHNOLOGIES
Microsoft
hadoop
Who has the most connections/partners?
0
5
10
15
20
25
30
35
40
45
Cloudera
Hortonworks
IBM
MapR
Vertica
ParAccel
Datameer
Amazon
Microsoft
Oracle
Dell
EMC
Aster Data
Greenplum
Hive
Datameer
Powerfully Simple

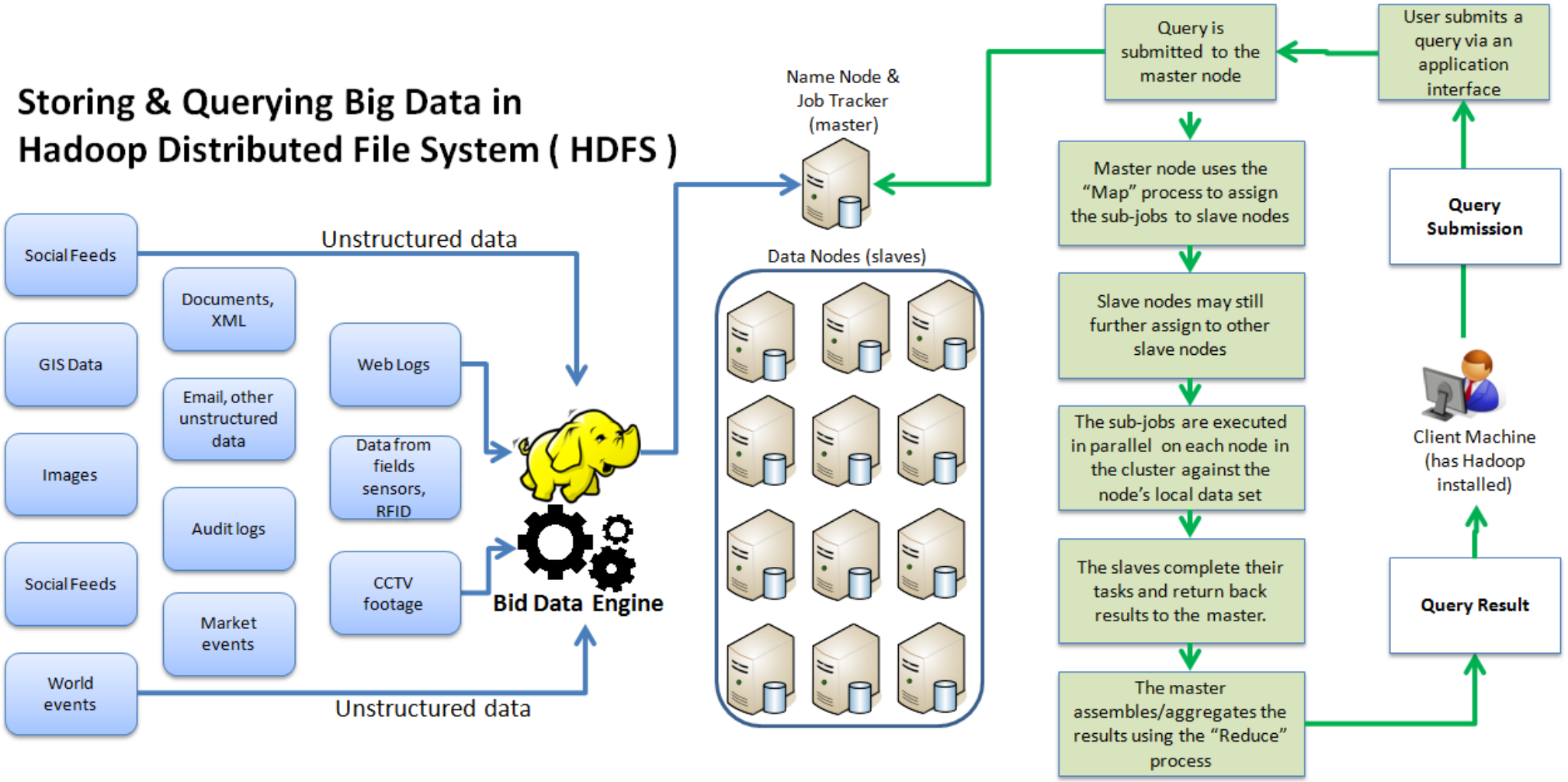
Big Data - Hadoop



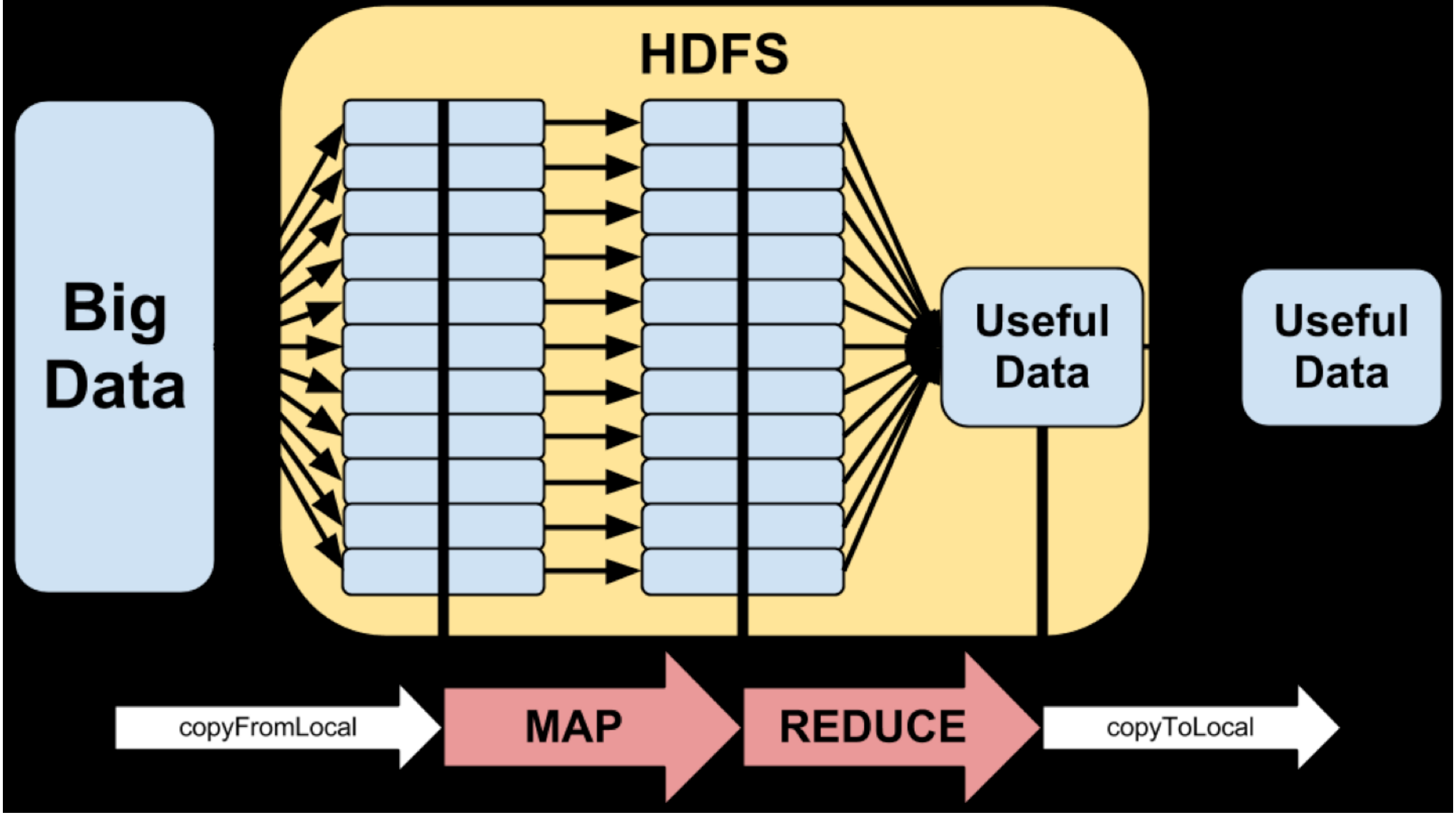
Big Data - Hadoop



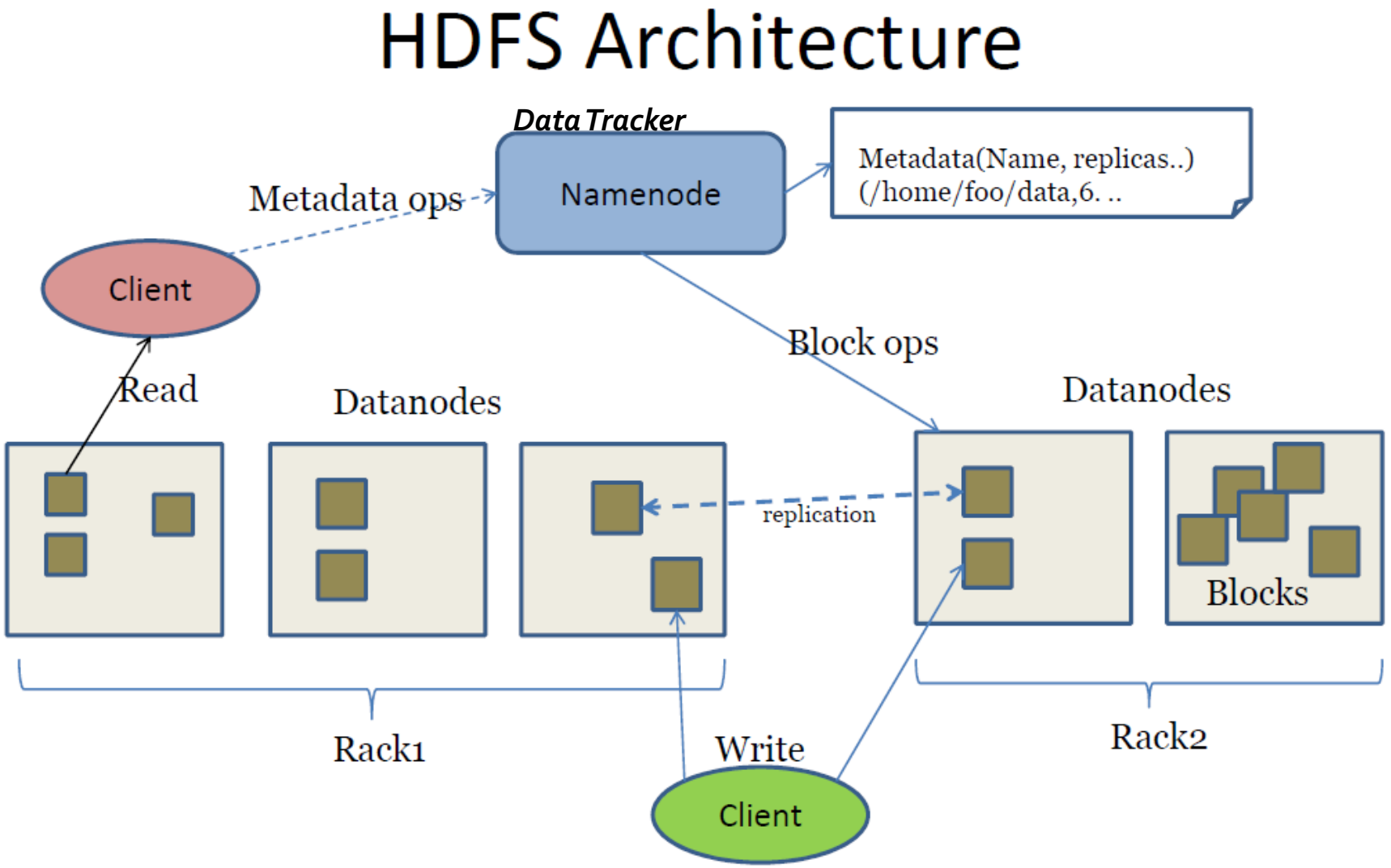
Big Data - Hadoop



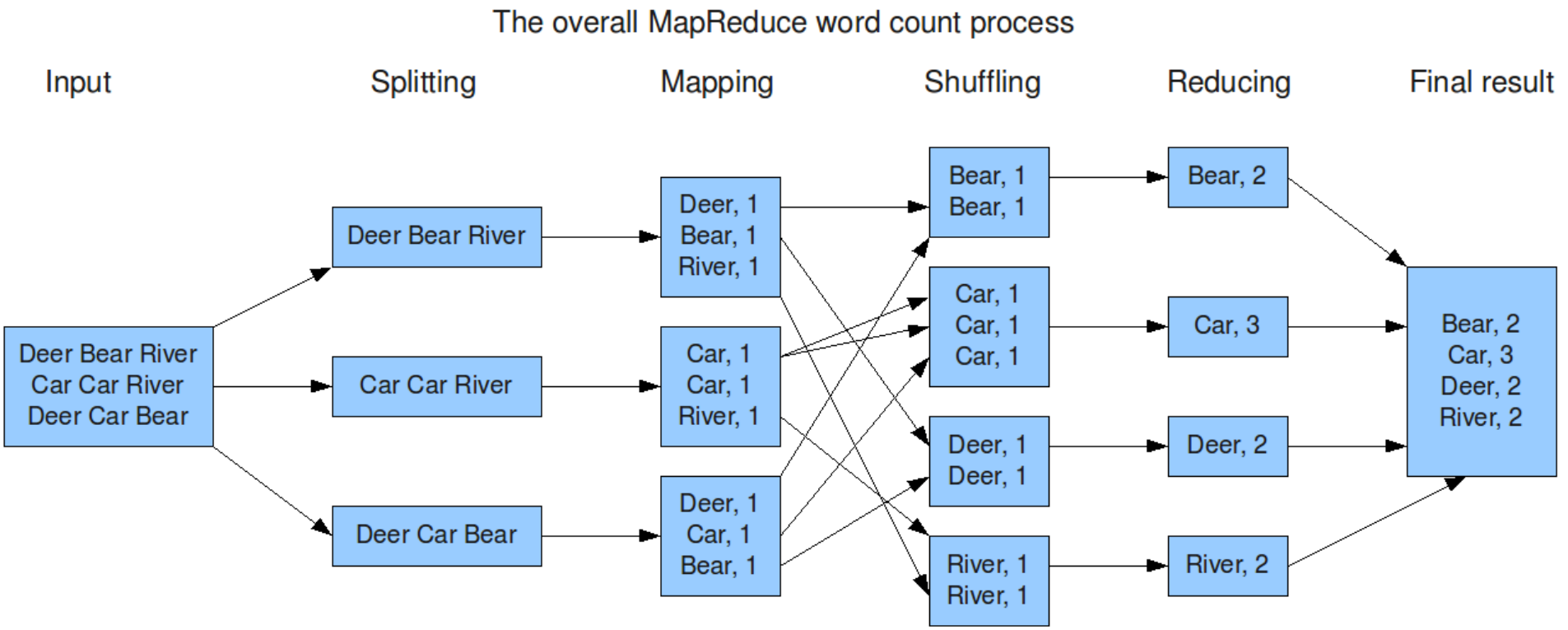
Overview of Hadoop MapReduce



Big Data - Hadoop



Big Data - MapReduce



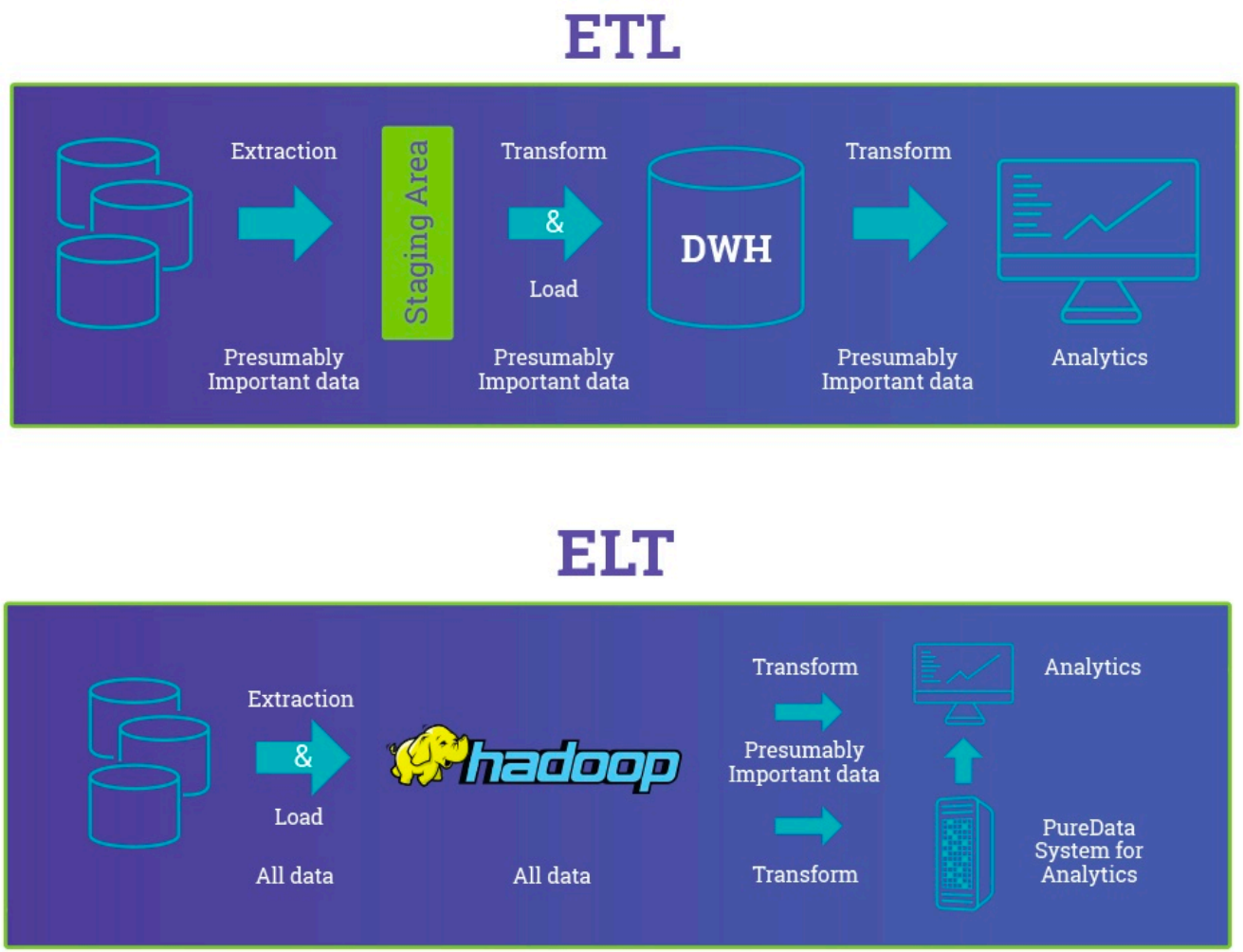
ETL vs ELT

- For the last couple of decades ETL (extract, transform, load) has been the traditional approach for data warehousing and analytics. The ELT (extract, load, transform) approach changes the old paradigm.

ELT

- In the ELT approach, after you've extracted your data, you immediately start the loading phase - moving all the data sources into a single, centralized data repository. With today's infrastructure technologies such as Hadoop, or cloud storage, systems can now support large storage and scalable compute. Therefore, a large, expanding data pool and fast processing is virtually endless for maintaining all the extracted raw data.

ETL vs ELT

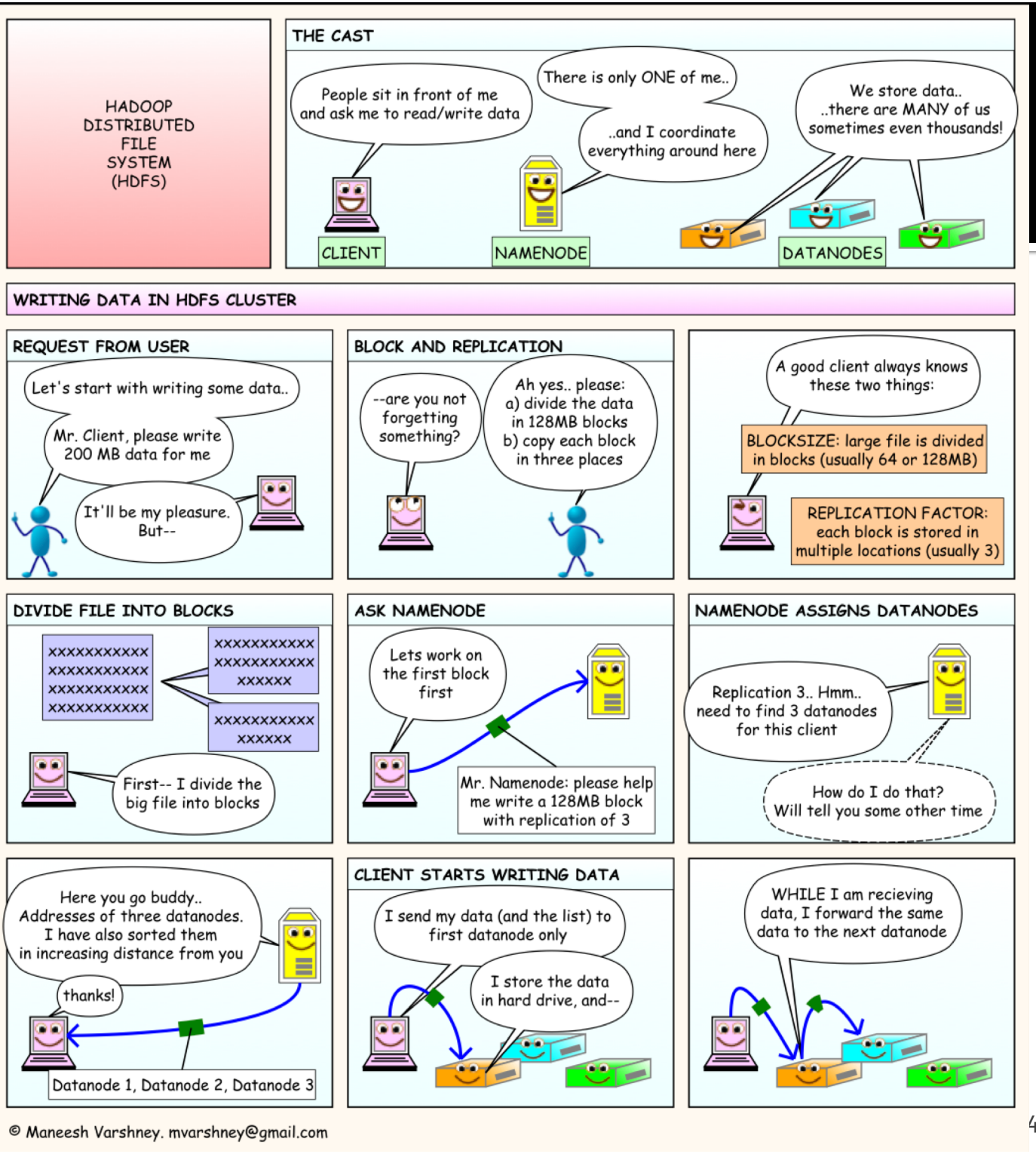

ETL
Extraction
Staging Area
Transform
&
Load
DWH
Transform
Analytics
Presumably Important data
Presumably Important data
Presumably Important data
ELT
Extraction
&
Load
All data
hadoop
All data
Transform
Presumably Important data
Transform
Analytics
PureData System for Analytics

Why ELT instead of ETL?

- **When ingestion speed is important.** Because ELT doesn't have to wait for the data to be worked off-site and then loaded, (data loading and transformation can happen in parallel) the ingestion process is much faster, delivering raw information considerably faster than ETL.

- **When more intel is better intel.** The advantage of turning data into business intelligence lay in the ability to surface hidden patterns into actionable information. By keeping all historical data on hand, organizations can mine along timelines, sales patterns, seasonal trends, or any emerging metric that becomes important to the organization. Since the data was not transformed before being loaded, you have access to all the raw data.

Big Data


HADOOP DISTRIBUTED FILE SYSTEM (HDFS)
THE CAST
People sit in front of me and ask me to read/write data
There is only ONE of me..
..and I coordinate everything around here
We store data.. ..there are MANY of us sometimes even thousands!
CLIENT
NAMENODE
DATANODES
WRITING DATA IN HDFS CLUSTER
REQUEST FROM USER
Let's start with writing some data..
Mr. Client, please write 200 MB data for me
It'll be my pleasure. But--
BLOCK AND REPLICATION
--are you not forgetting something?
Ah yes.. please: a) divide the data in 128MB blocks b) copy each block in three places
A good client always knows these two things:
BLOCKSIZE: large file is divided in blocks (usually 64 or 128MB)
REPLICATION FACTOR: each block is stored in multiple locations (usually 3)
DIVIDE FILE INTO BLOCKS
First-- I divide the big file into blocks
ASK NAMENODE
Lets work on the first block first
Mr. Namenode: please help me write a 128MB block with replication of 3
NAMENODE ASSIGNS DATANODES
Replication 3.. Hmm.. need to find 3 datanodes for this client
How do I do that? Will tell you some other time
Here you go buddy.. Addresses of three datanodes. I have also sorted them in increasing distance from you
thanks!
Datanode 1, Datanode 2, Datanode 3
CLIENT STARTS WRITING DATA
I send my data (and the list) to first datanode only
I store the data in hard drive, and--
WHILE I am recieving data, I forward the same data to the next datanode
© Maneesh Varshney. mvarshney@gmail.com