Navigation

Frankfurt School of Finance & Management

3 Big Data

Limitations of the Data Warehouse Approach

Frankfurt School of Finance & Management

Traditional DWH solutions are designed to provide a single point of truth. Important aspects are:

- Merge and unify data from multiple data sources
- High data quality
- Proper historization of the data
- Data Governance and Compliance

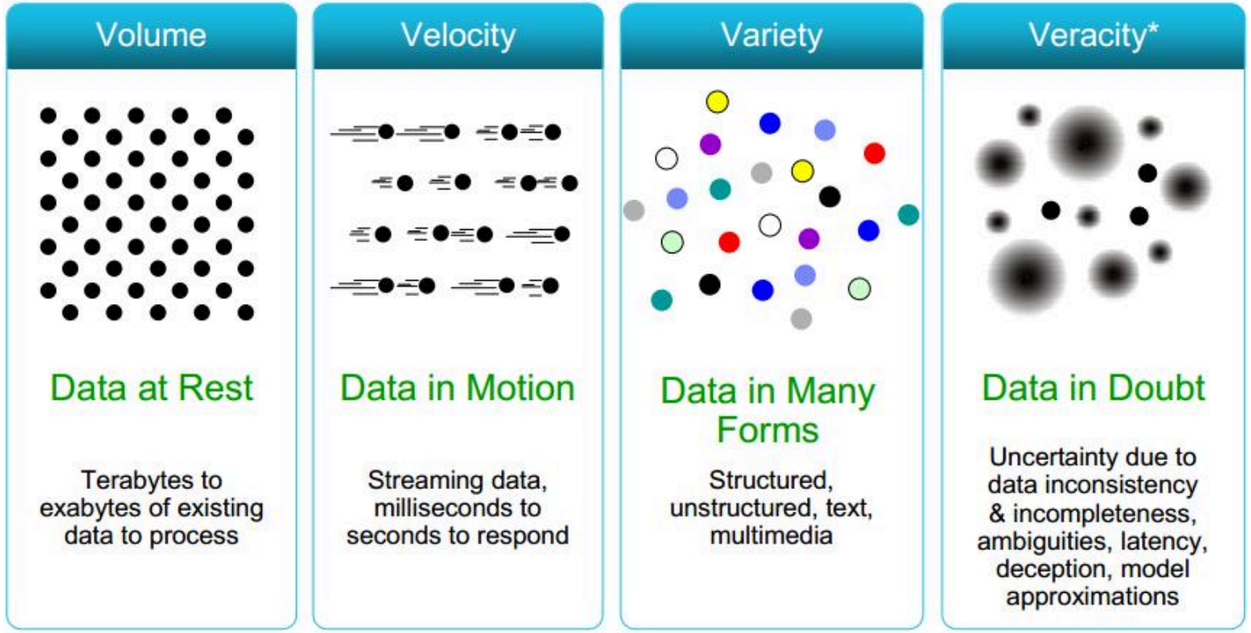
What is Big Data?

The basic idea behind the phrase ‘Big Data’ is that everything we do is increasingly leaving a digital trace (or data), which we (and others) can use and analyze.

– Bernard Marr

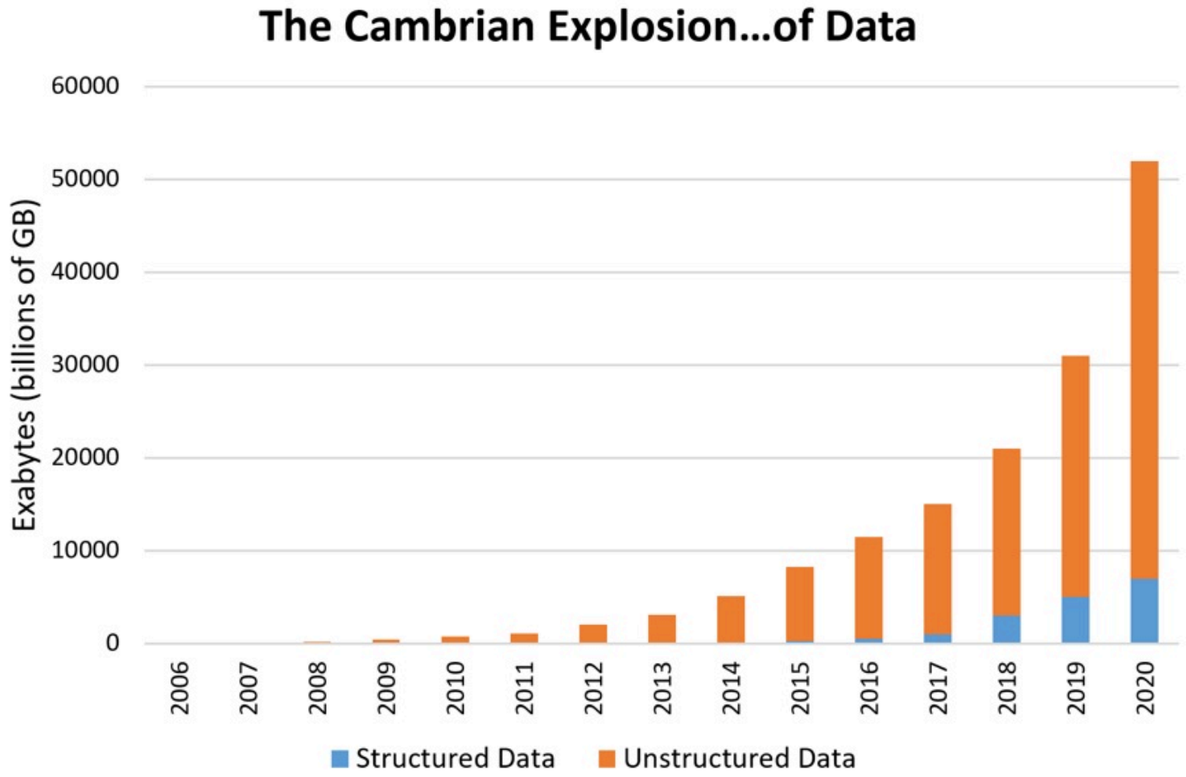
Big Data is the frontier of a firm’s ability to store, process, and access (SPA) all the data it needs to

4V's of Big Data



Source: www.cs.kent.edu/~jin/BigData/Lecture1.pptx

Volume (Scale)



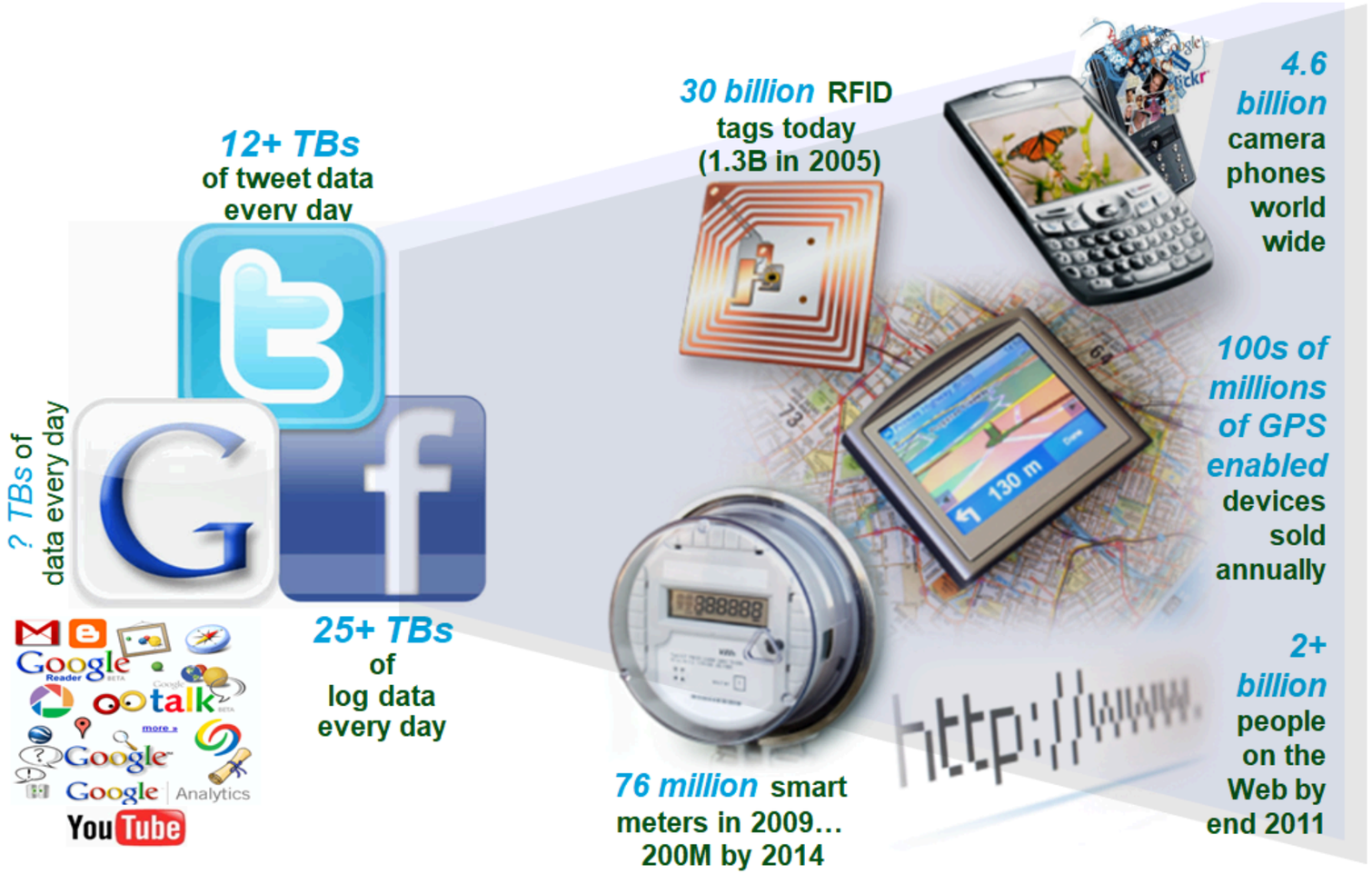
The Model of Generating and Consuming Data has Changed

Old Model: Few companies are generating data, all others are consuming data

New Model: all of us are generating data, and all of us are consuming data

Source: www.cs.kent.edu/~jin/BigData/Lecture1.pptx

Collecting Data



Source: www.cs.kent.edu/~jin/BigData/Lecture1.pptx

Types of Data People are Creating (I)

Activity Data

Simple activities like listening to music or reading a book are now generating data. Digital music players and eBooks collect data on our activities. Your smartphone collects data on how you use it and your web browser collects information on what you are searching for. Your credit card company collects data on where you shop and the shops collect data on what you buy. It is hard to imagine any activity that does not generate data.

Conversation Data

Our conversations are now digitally recorded. It all started with

Types of Data People are Creating (II)

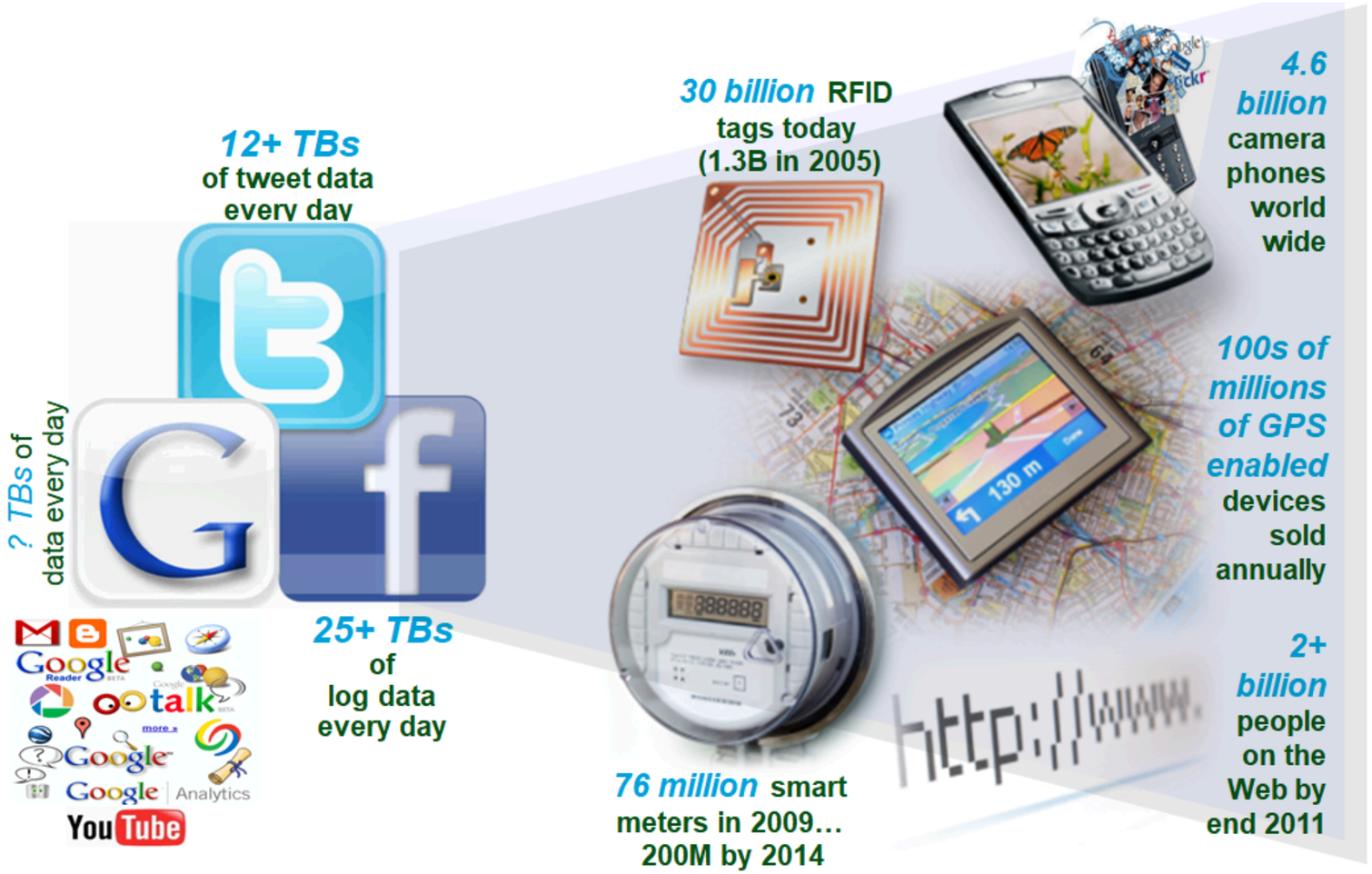
Sensor Data

We are increasingly surrounded by sensors that collect and share data. Take your smart phone, it contains a global positioning sensor to track exactly where you are every second of the day, it includes an accelometer to track the speed and direction at which you are travelling. We now have sensors in many devices and products.

The Internet of Things Data

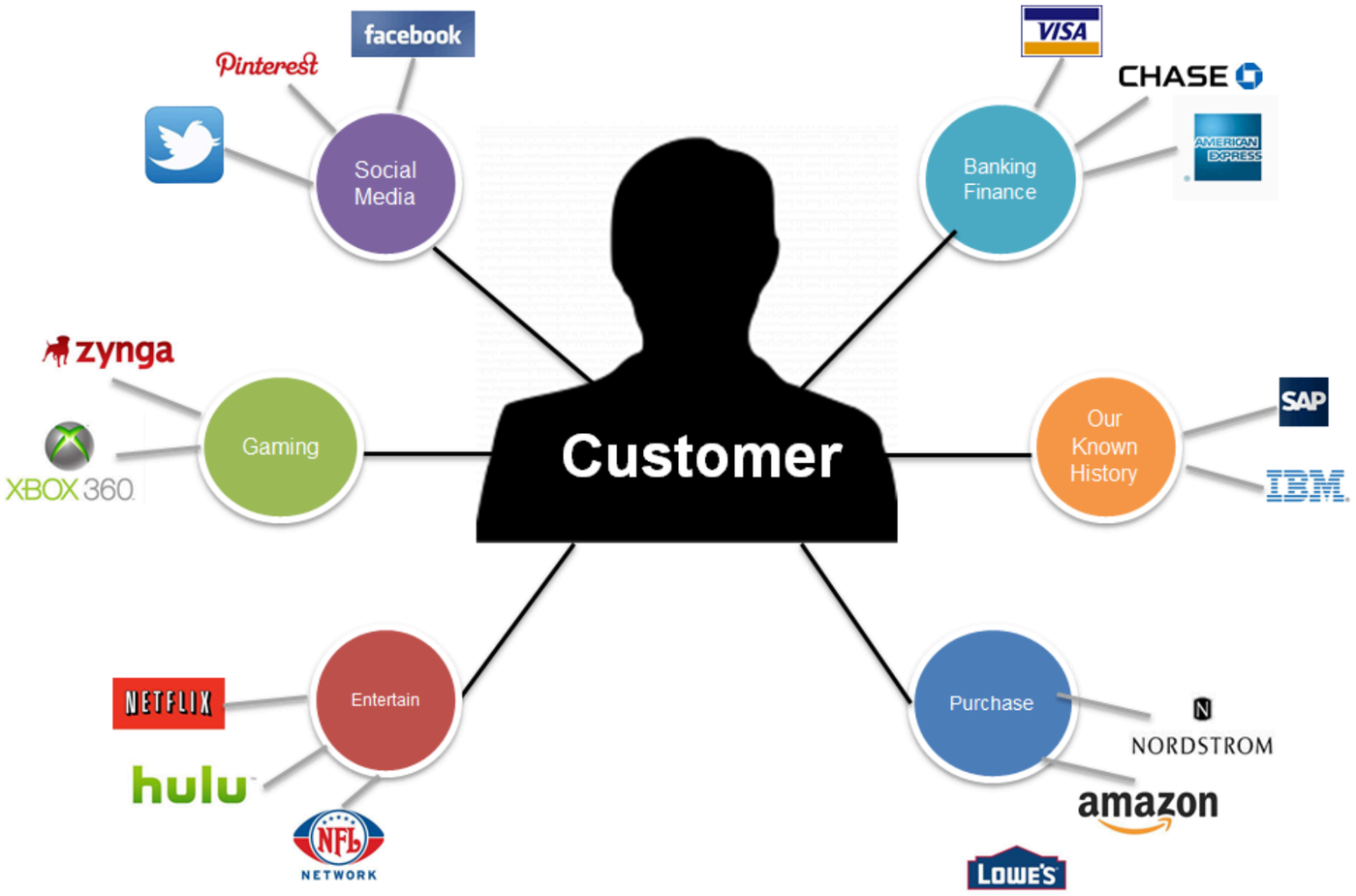
We now have smart TVs that are able to collect and process

Variety (Complexity)



Source: www.cs.kent.edu/~jin/BigData/Lecture1.pptx

A Single View to the Customer

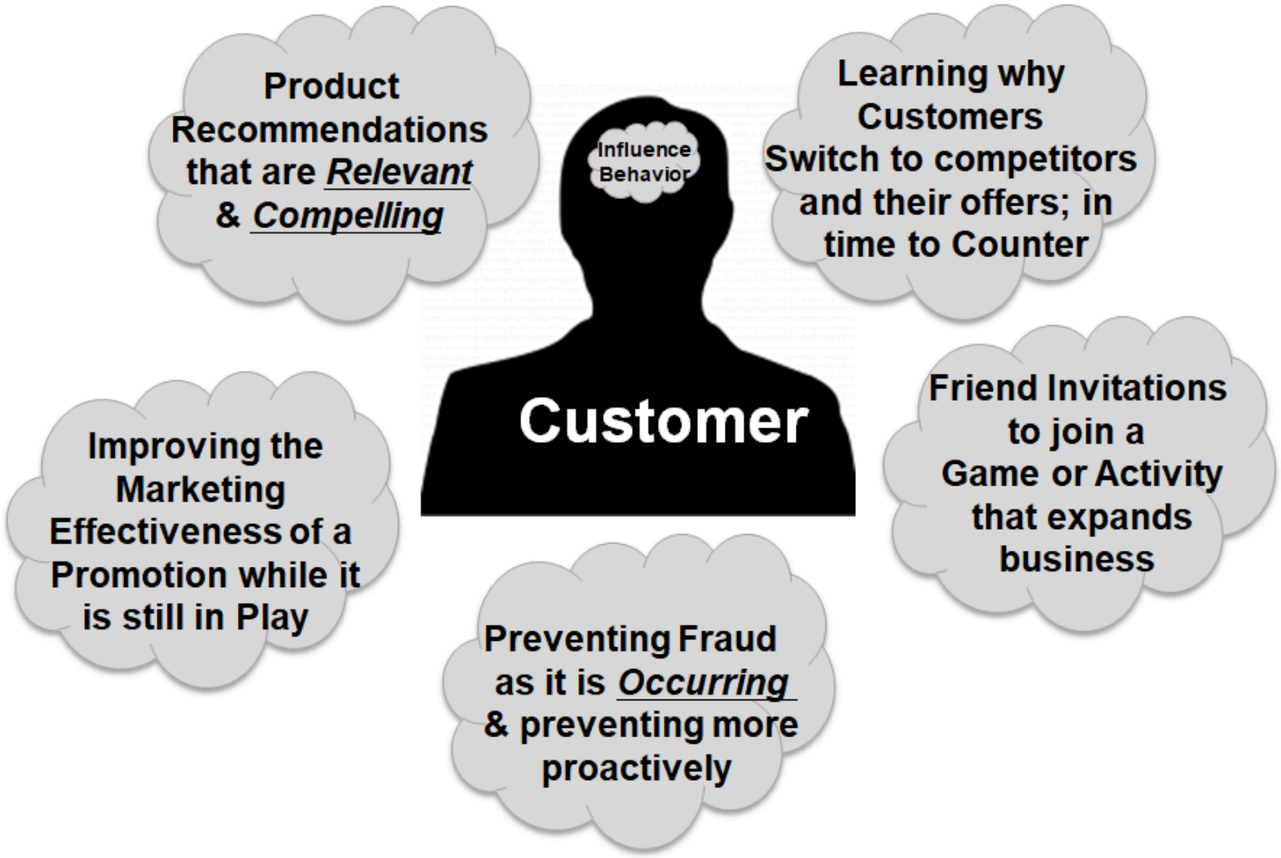


Source: www.cs.kent.edu/~jin/BigData/Lecture1.pptx

Velocity (Speed)

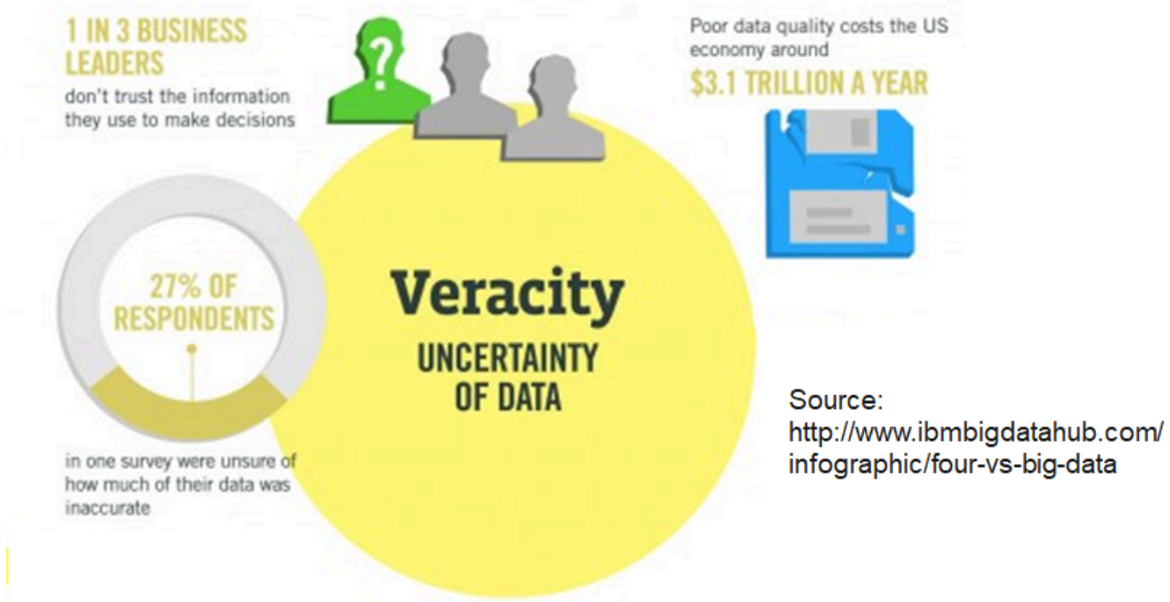
- Data is being generated fast and needs to be processed fast
- Late decisions result in missing opportunities

Real-Time Analytics



Source: www.cs.kent.edu/~jin/BigData/Lecture1.pptx

Veracity (Uncertainty)



Organizations must now analyze both structured and unstructured data that is uncertain and imprecise.

In many cases, it is not known whether the data is correct (e. g. fake news) or representative (e. g. biased expressions of opinion in forums).

It may be prudent to assign a Data Veracity score and ranking for specific data sets to avoid making decisions based on analysis of uncertain and imprecise data.

How is Big Data actually used?

- ● **Better understand and target customers**

  Companies expand their traditional data with social media data, browser, text analytics or sensor data to get a more complete picture of their customers.

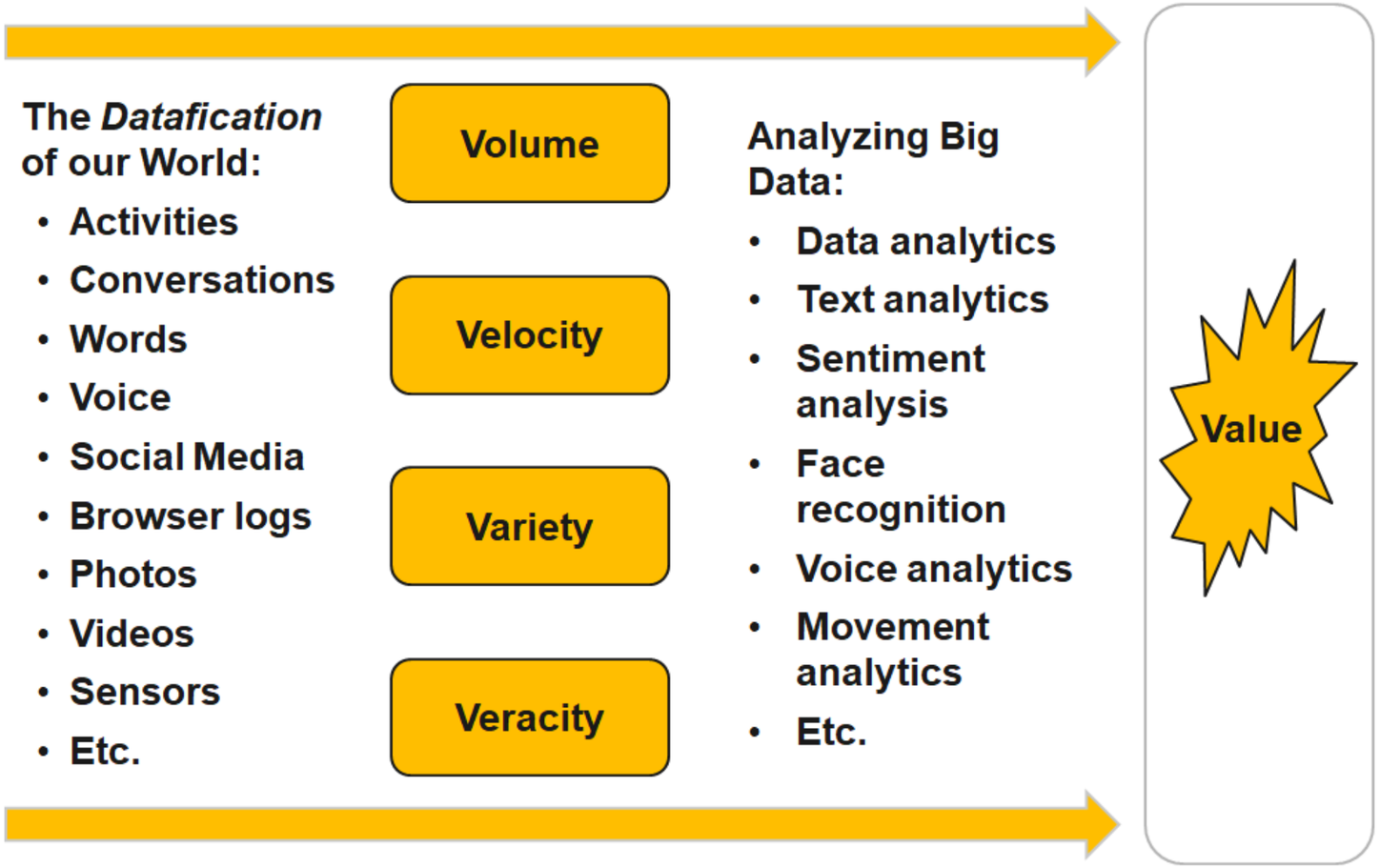
- ● **Understand and Optimize Business Processes**

  Retailers are able to optimize their stock based on predictive models generated from social media data, web search trends, weather forecasts…

- ● **Improving Health**

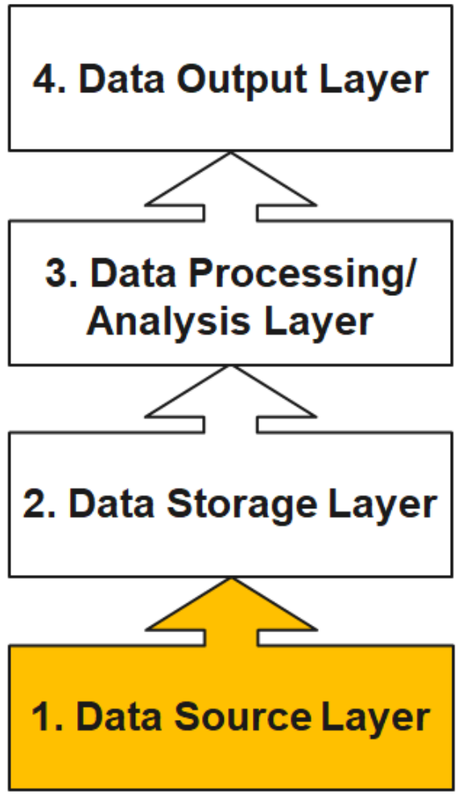
  Use the data from smart watches, wearable devices, Google Trends,

Turning Big Data into Value



Source: http://de.slideshare.net/BernardMarr/140228-big-data-slide-share/3-The_basic_idea_behind_the

The Four Layers of Big Data (I)



Data Source Layer

This is where the data arrives at the organization. It includes everything from sales records, customer database, feedback, social media channels, marketing list, email archives etc.

Source: http://de.slideshare.net/BernardMarr/140228-big-data-slide-share/3-The_basic_idea_behind_the

Identify and Prioritize Data Sources

Frankfurt School of Finance & Management

Key: Worst... Best

Data Source	Increase Store Traffic	Increase Shopping Bag Revenue	Increase # Corporate Events	Increase Promotional Effectiveness	Improve NPI Effectiveness
Point of Sales Transactions	●	●	●	●	●
Market Baskets	●	●	◑	●	●
Store Demographics (Zip Code)	◕	◕	◔	◕	◕
Local Competitive Stores	◑	◑	◑	◑	◑
Store Manager Demographics	◔	◔	◕	◔	◔
Consumer Comments	◕	◕	◕	◕	◑
Social Media	◑	◔	◔	◕	◕
Weather	◕	◔	◔	◔	◔
Local Events	●	◑	◔	◑	◔
Traffic	◕	◔	◔	◑	◔
Zillow	◔	◑	◑	◑	◑

Business value of potential data sources

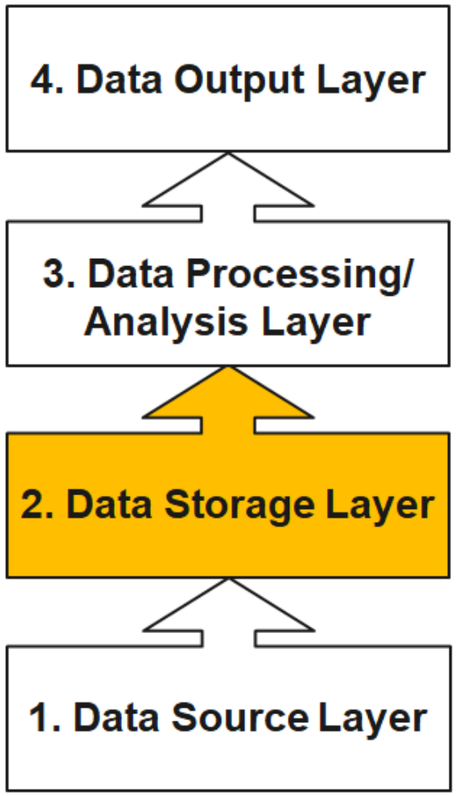
Source: Schmarzo (2016): Big Data MBA, p. 49

The Four Layers of Big Data (II)

Frankfurt School of Finance & Management

Data Storage Layer

This is where Big Data is stored, once it is gathered from the sources. As the volume of data stored has started to explode, new database technologies have been developed — such as NoSQL database systems.

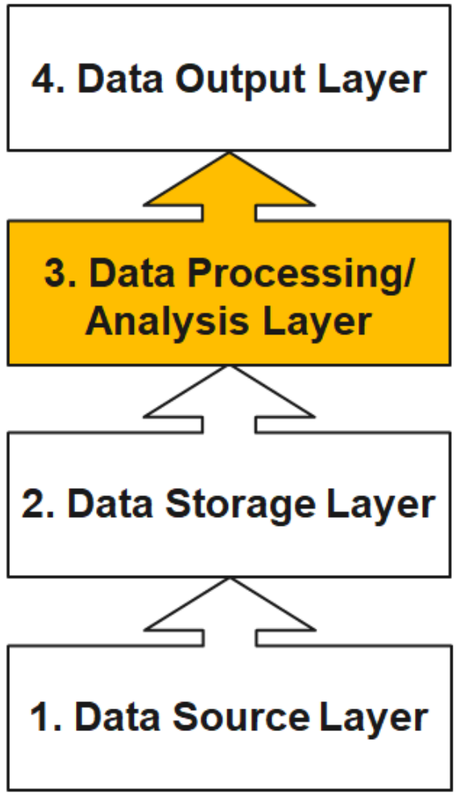


Source: http://de.slideshare.net/BernardMarr/140228-big-data-slide-share/3-The_basic_idea_behind_the

The Four Layers of Big Data (III)

Data Processing/Analysis Layer

When you want to use the data you have stored to find out something useful, you will need to process and analyze it.

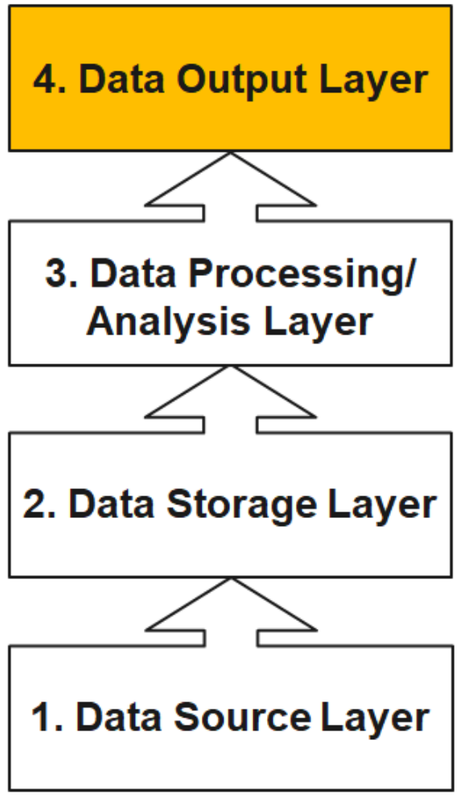


Source: http://de.slideshare.net/BernardMarr/140228-big-data-slide-share/3-The_basic_idea_behind_the

The Four Layers of Big Data (IV)

Data Output Layer

This is how the insights gleaned through the analysis is passed on to the people who can take action to benefit from them.



Source: http://de.slideshare.net/BernardMarr/140228-big-data-slide-share/3-The_basic_idea_behind_the

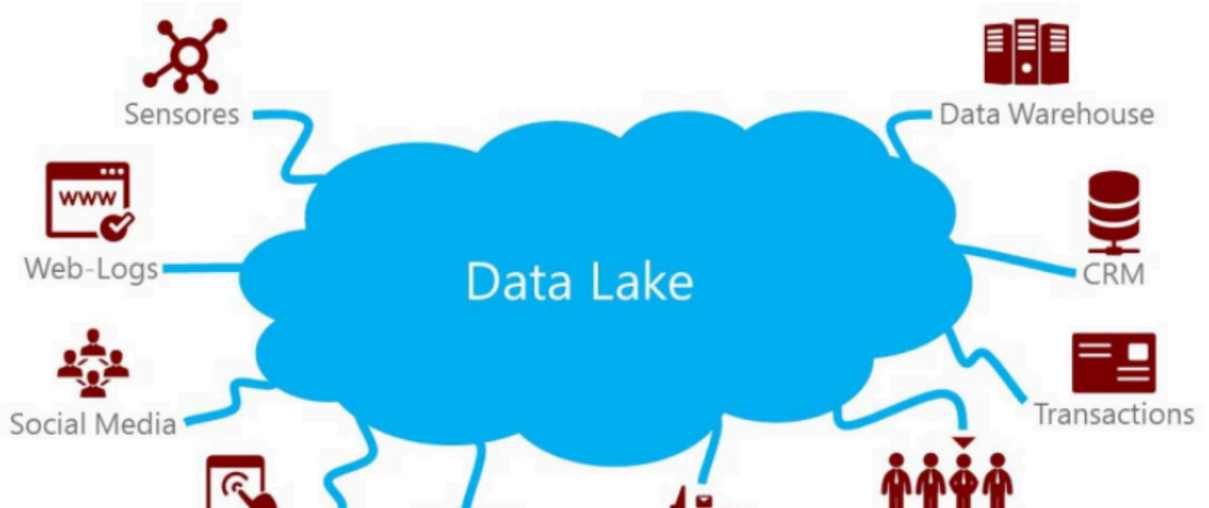
From Data Warehouse to Data Lake

Instead of recording millions of transactions, todays organizations are recording billions of interactions. Companies are capturing more and more data that can open business opportunities and unlock new sources of value for organizations.

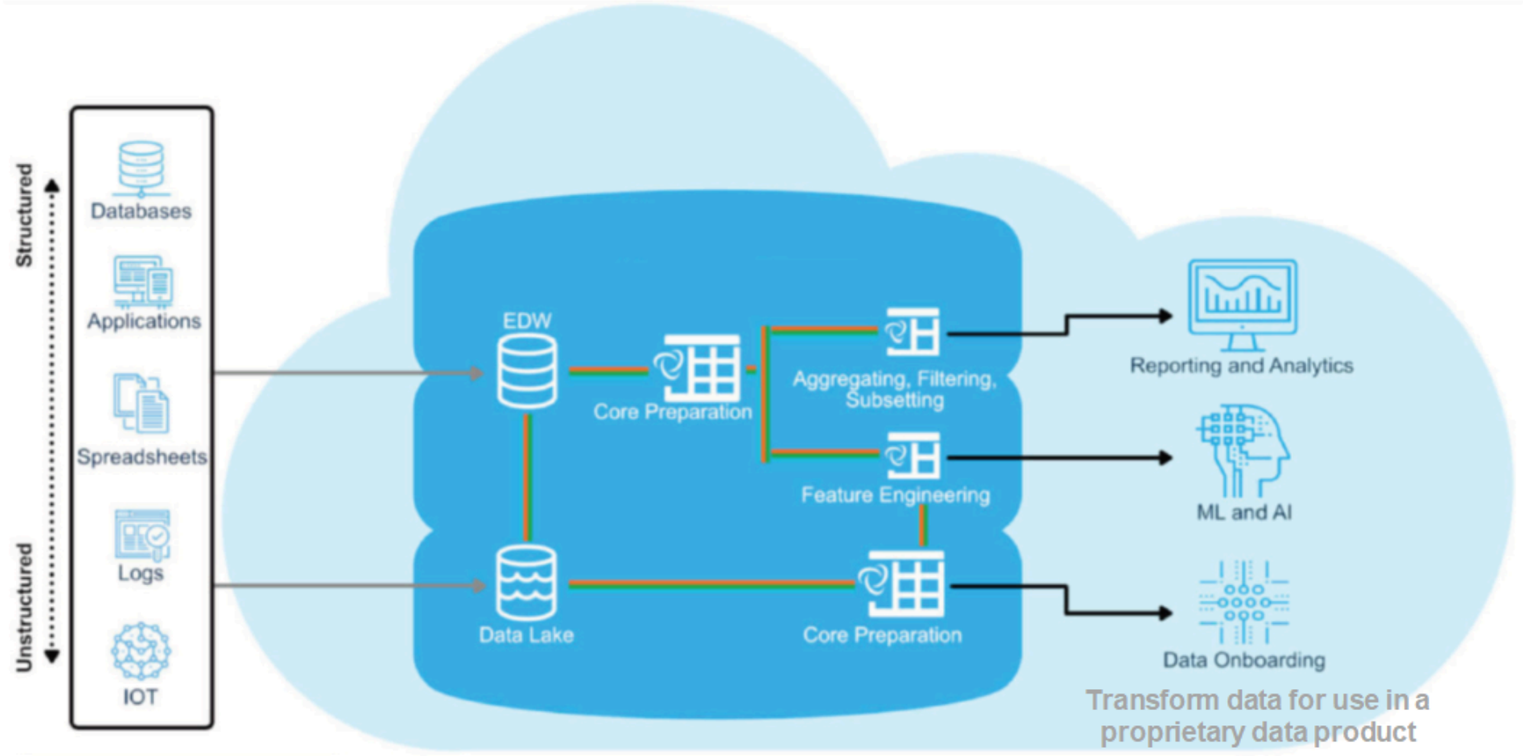
Companies are not able to store this data in data warehouses because it is of high volume, mostly raw and often not structured. As consequence, data lakes have emerged as an alternative approach. The intent is to capture enterprise data and load it in its raw form into a

The Data Lake

A data lake is a method of storing data within a system in its natural format, that facilitates the collocation of data in various schemata and structural forms. The idea of data lake is to have a single store of all data in the enterprise ranging from raw data to transformed data which is used for various tasks including reporting, visualization, analytics and machine learning. [Wikipedia]



Modern Data Analytics Architecture



Source: Trifacta: EOL for ETL?

Logical Data Warehouse

A “logical data warehouse” provides analytical company data without first physically moving it to a physical data warehouse.

As in a classic data warehouse, uniform views are provided for analysis purposes.

While the data in the classic data warehouse comes from a “well-defined” physically uniform database, the “logical data