EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Advanced Database Systems

Introduction

Strategic project of TBU in Zlín, reg. no. CZ.02.2.69/0.0/0.0/16_015/0002204

Zdenka Prokopova
TBU in Zlín

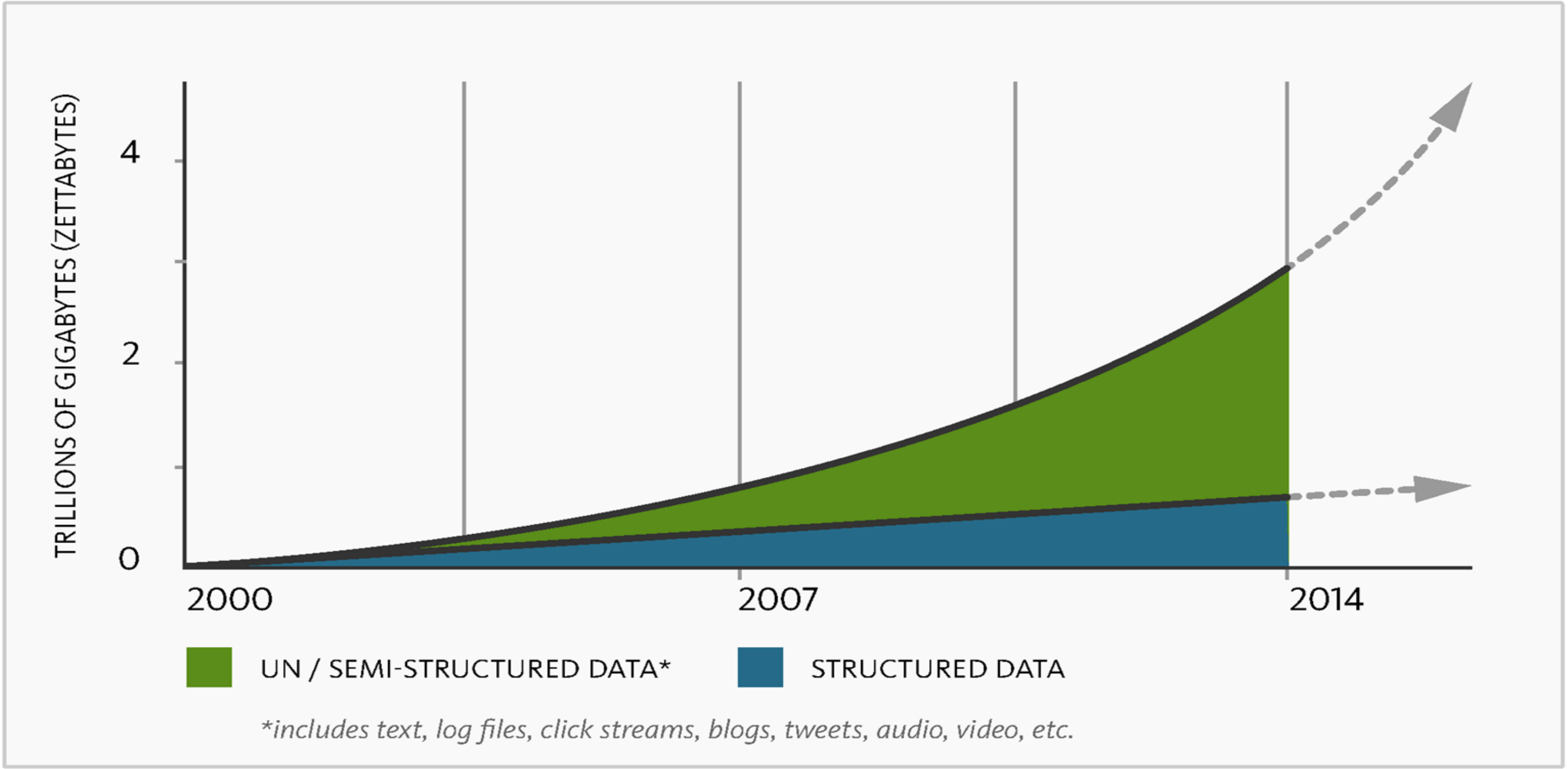
Content

- Current trends in data management & computing
- Big Data
- Relational vs. NoSQL databases
    - the value of relational databases
    - new requirements and NoSQL features
    - flexible data models
- Types of NoSQL databases
    - key-value stores, document databases, column-family databases, graph databases
    - principles and examples

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Current Trends: Big Data



*includes text, log files, click streams, blogs, tweets, audio, video, etc.

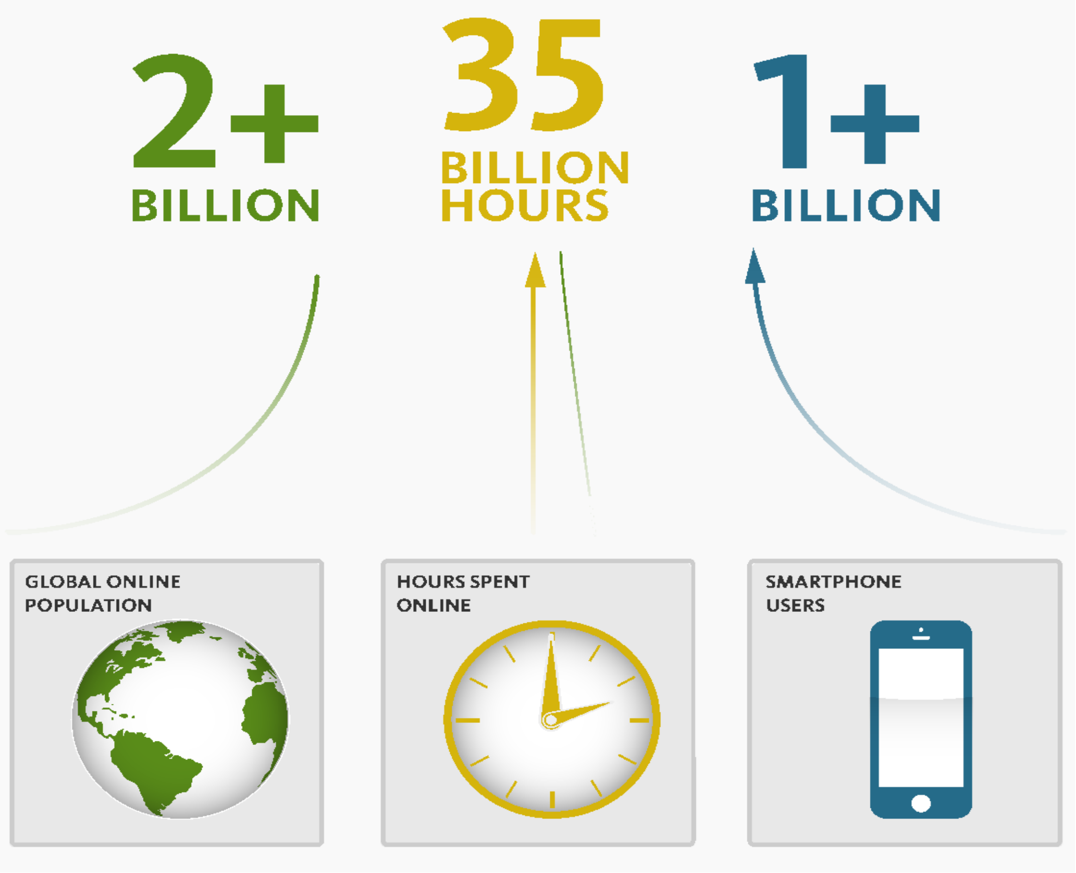
source: http://www.couchbase.com/sites/default/files/uploads/all/whitepapers/NoSQL-Whitepaper.pdf

Zdenka Prokopova
TBU in Zlín

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

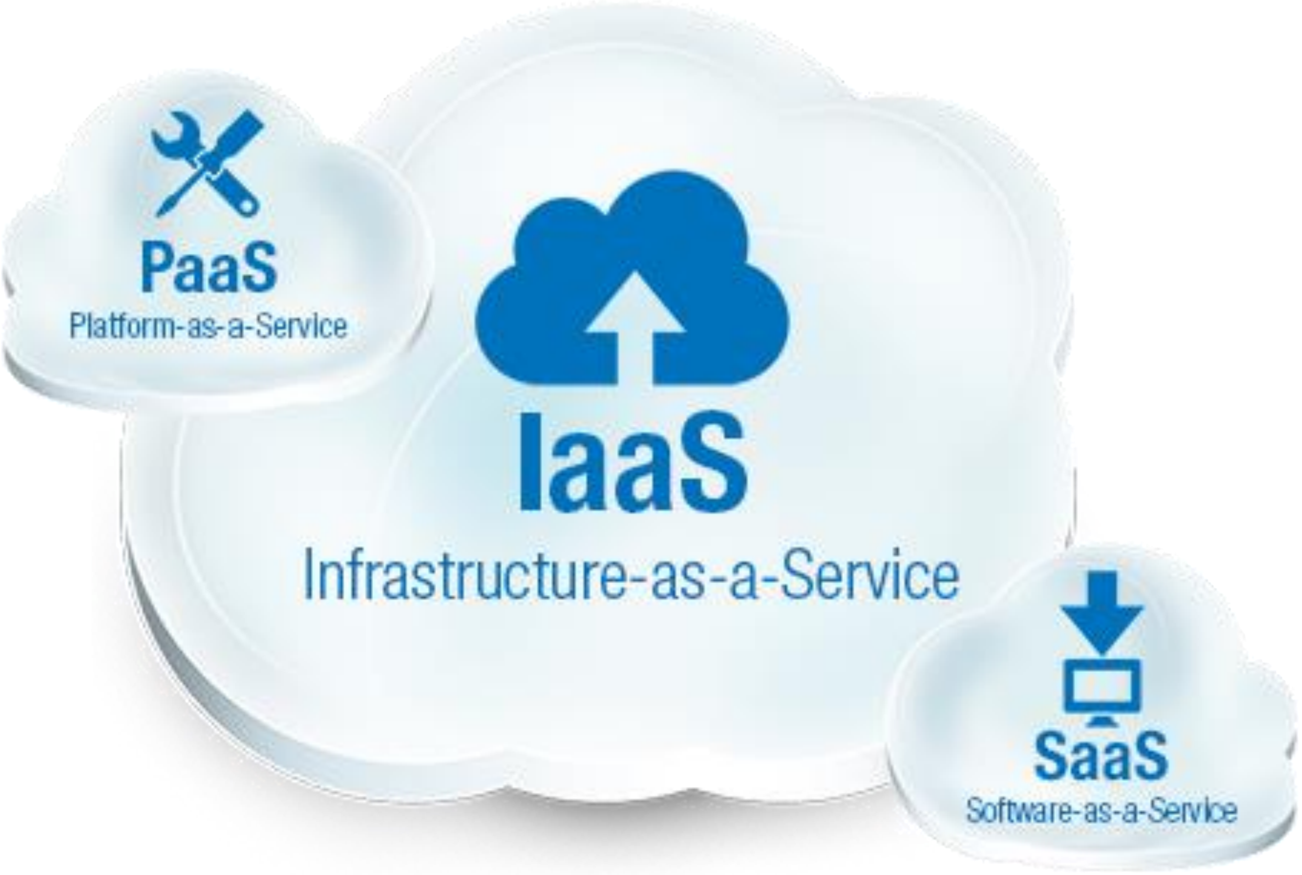
Current Trends: Big Users



Zdenka Prokopova
TBU in Zlín

source: http://www.couchbase.com/sites/default/files/uploads/all/whitepapers/NoSQL-Whitepaper.pdf

Current Trends: Cloud Computing

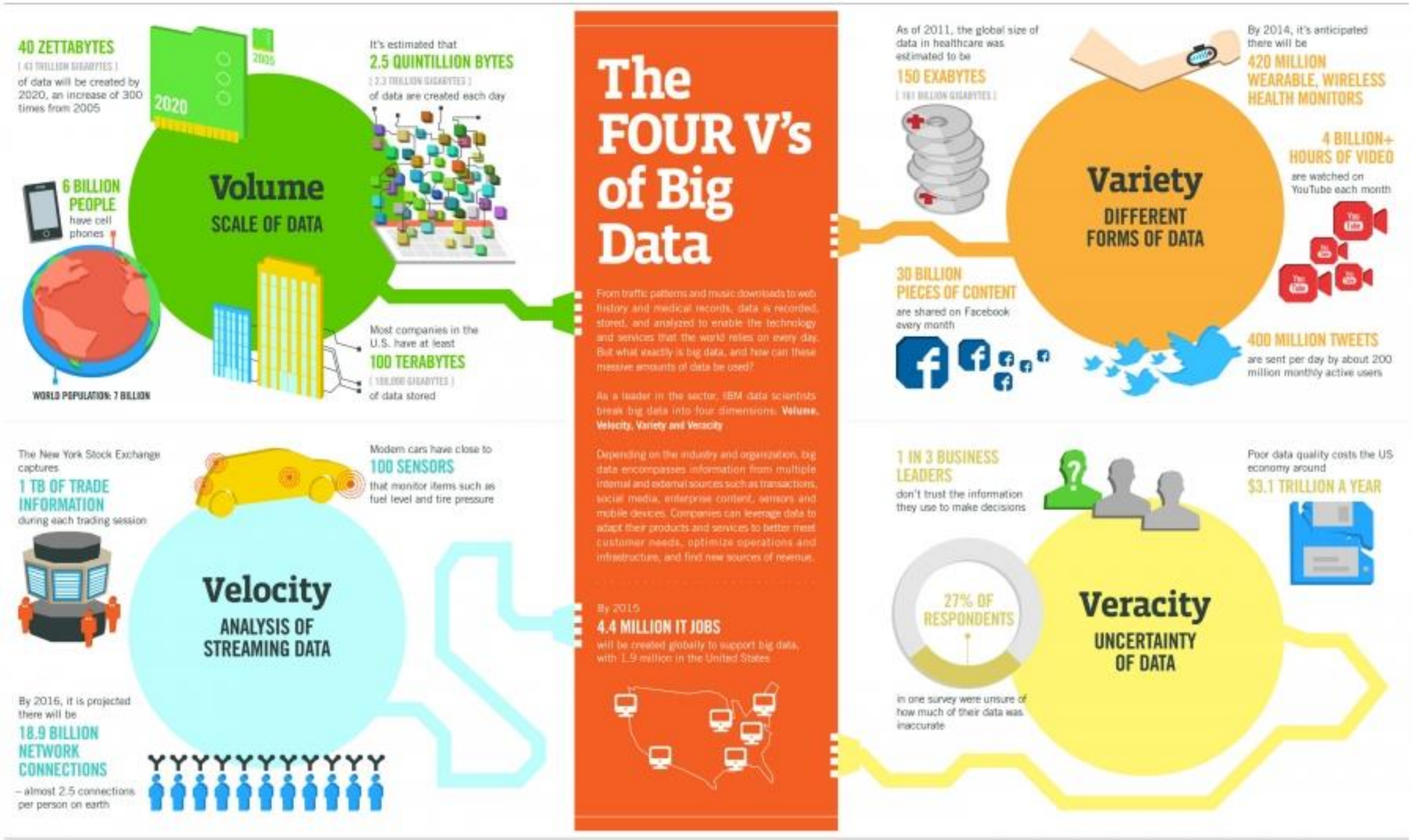


Zdenka Prokopova
TBU in Zlín

source: http://www.profitbricks.com/what-is-iaas

Big Data

“Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.” (Gartner, 2012)



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

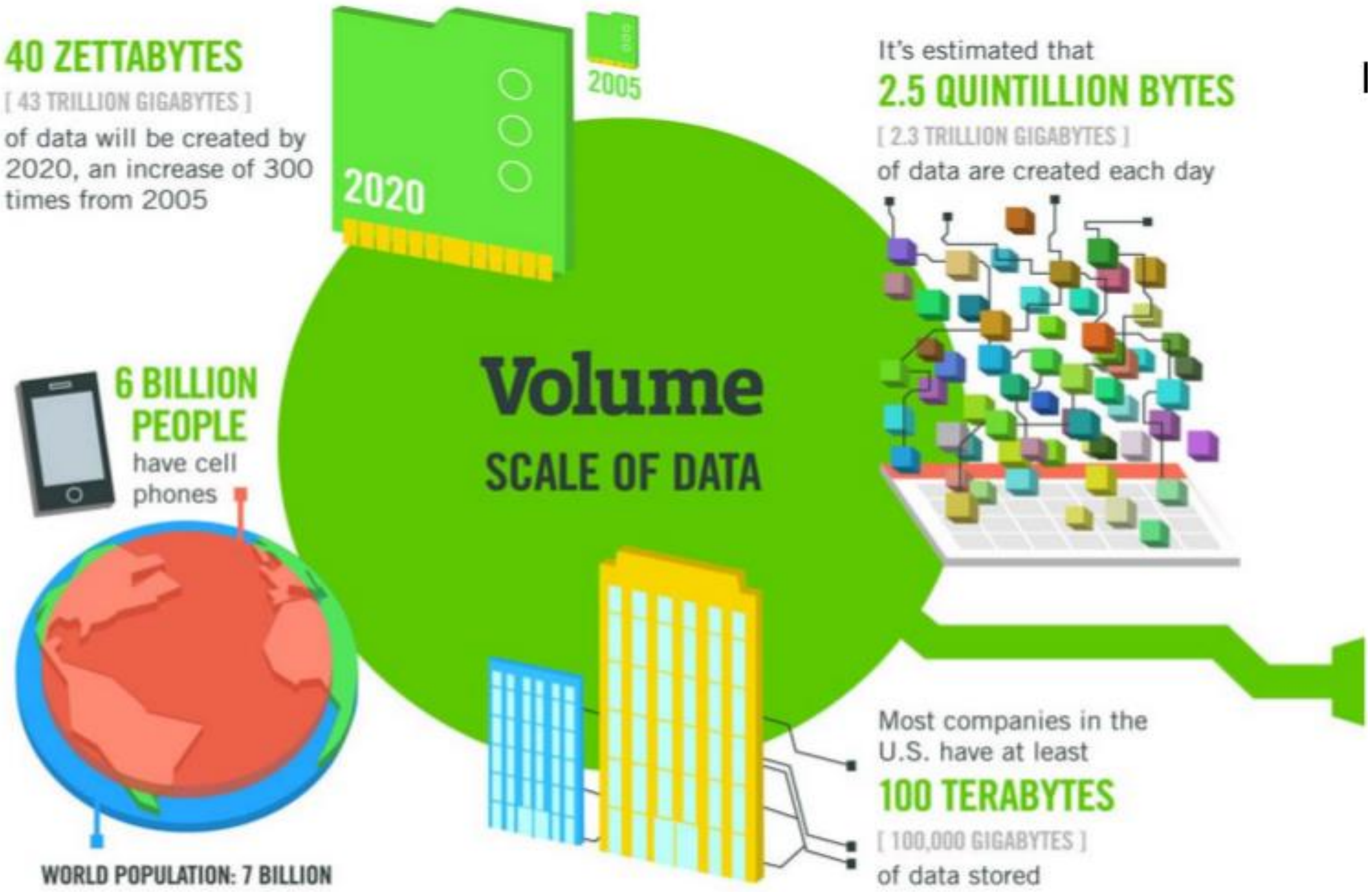
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

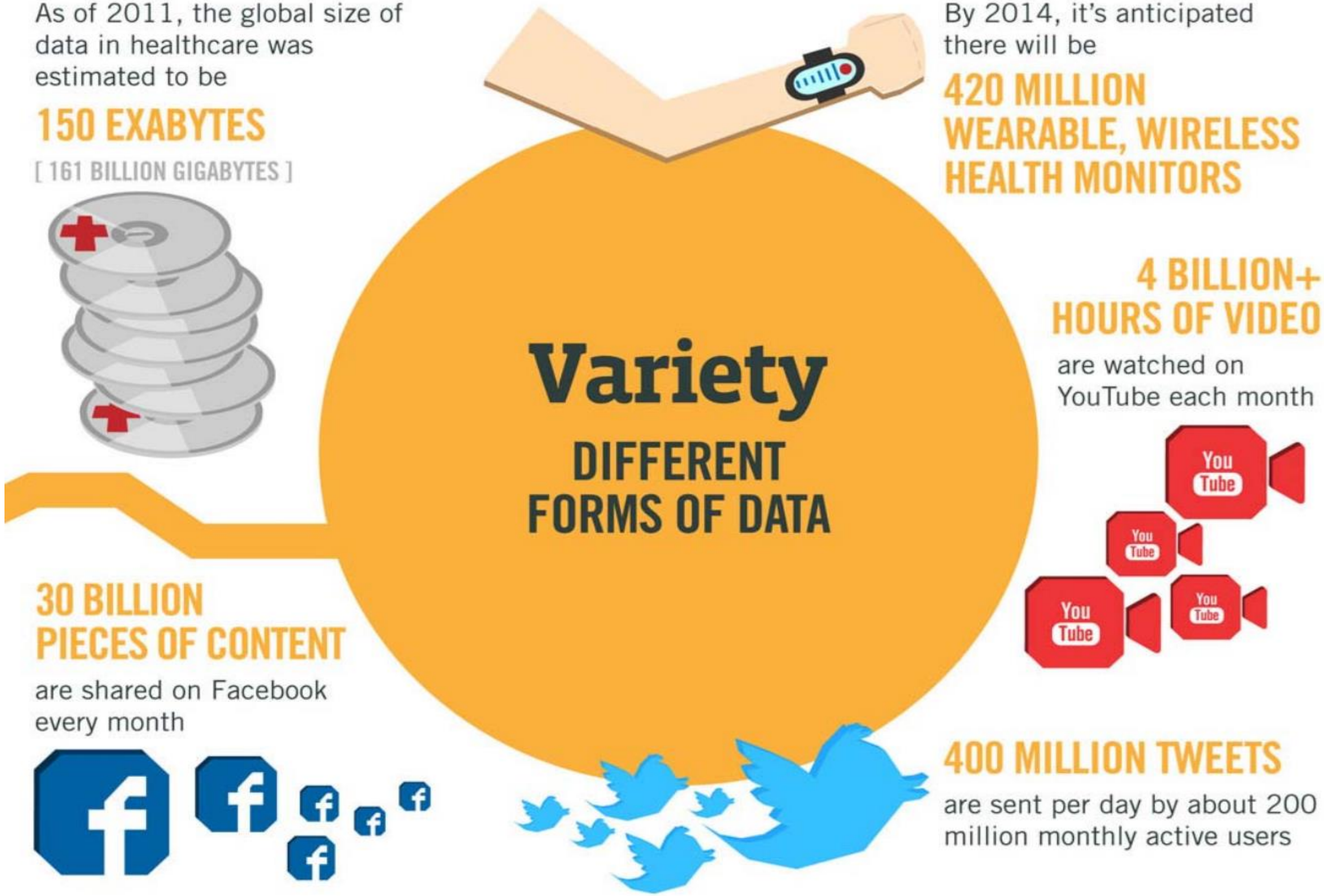
source: http://www.ibmbigdatahub.com/infographic/four-vs-big-data

IBM

Data volume is increasing exponentially


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2020
2005
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
Volume
SCALE OF DATA
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Various data types, formats and structures



As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

Variety

DIFFERENT FORMS OF DATA

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month

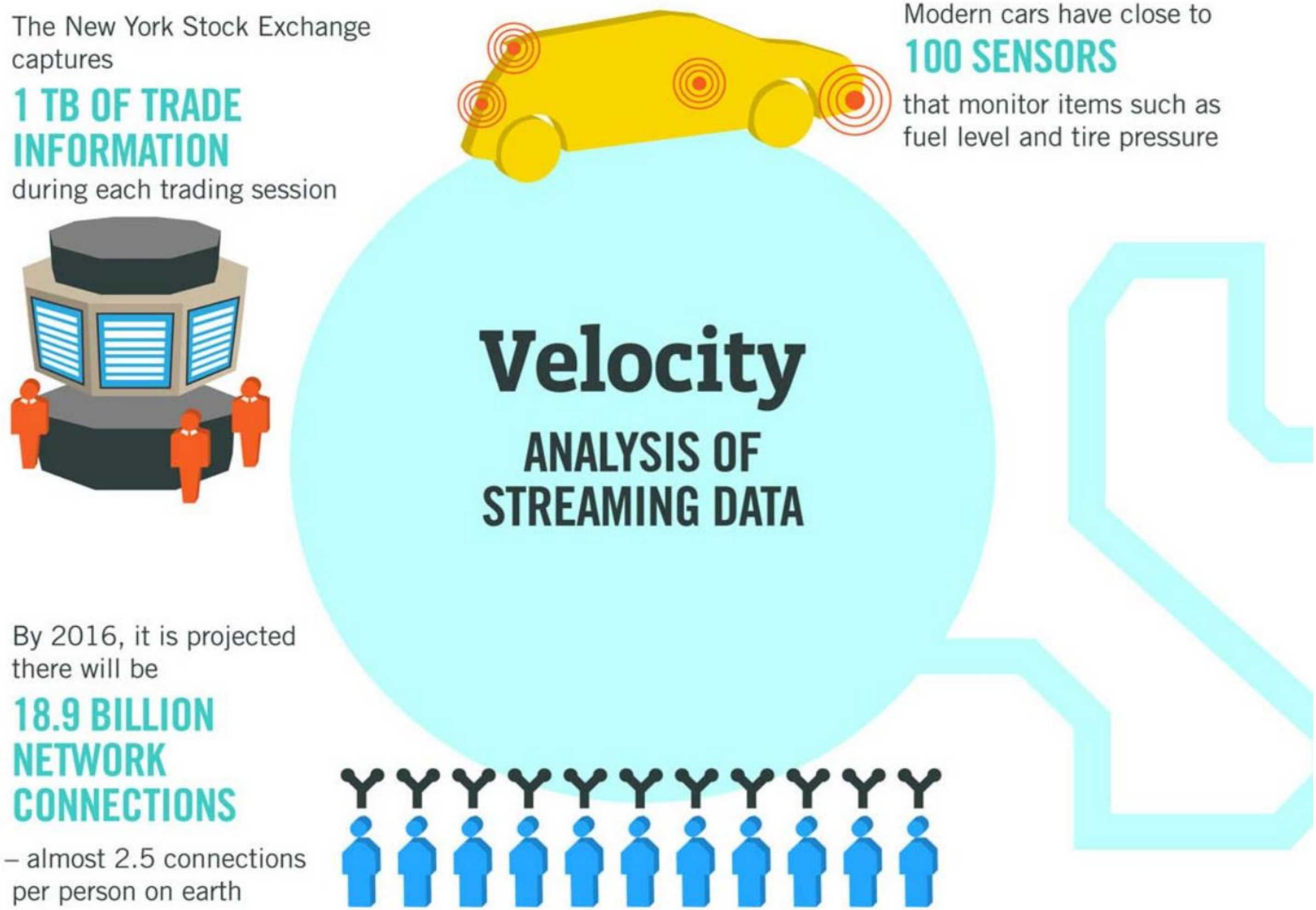
30 BILLION PIECES OF CONTENT

are shared on Facebook every month

400 MILLION TWEETS

are sent per day by about 200 million monthly active users

Data is being generated fast and need to be processed fast



Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures **1 TB OF TRADE INFORMATION** during each trading session

Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS** – almost 2.5 connections per person on earth

Processing (Big) Data

- OLTP: Online Transaction Processing (DBMSs)
  - Database applications
  - Storing, querying, multi-user access
- OLAP: Online Analytical Processing (Warehousing)
  - Answer multi-dimensional analytical queries
  - Financial/marketing reporting, budgeting, forecasting, …
- RTAP: Real-Time Analytic Processing (Big Data Architecture & Technology)
  - Data gathered & processed in real-time (streaming)
  - Real-time and history data combined

Technologies for Big Data

- Distributed file systems (GFS, HDFS, etc.)
- MapReduce
    - and other models for distributed programming
- NoSQL databases
- Grid computing, cloud computing
- Large-scale machine learning

Relational Database Management Systems

- RDBMS are predominant database technologies
  - first defined in 1970 by Edgar Codd of IBM's Research Lab

- Data modeled as relations (tables)
  - object = tuple of attribute values
  - tables contain objects of the same type
  - tables interconnected via foreign keys

- Relational calculus, SQL query language

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

The Value of Relational Databases

- A (mostly) standard data model
- Many well developed technologies
  - physical organization of the data
  - search indexes: B^+-Trees, hash indexes
  - query optimization, search operator implementations
- Good concurrency control (ACID)
  - transactions: atomicity, consistency, isolation, durability
- Many reliable integration mechanisms
  - “shared database integration” of applications

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MŠMT
MINISTRY OF EDUCATION,
YOUTH AND SPORTS

New Requirements on Data Management

Trends

- **Volume** of data
- **Cloud** comp. (IaaS)
- **Velocity** of data
- **Big** users
- **Variety** of data

Requirements

- Real data **scalability**
  - massive database distribution
  - dynamic resource management
  - horizontally scaling systems
- Frequent **update** operations
- Massive **read** throughput
- **Flexible** database schema

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

NoSQL Databases

- What is "NoSQL"?
  - term used in late 90s for a different type of technology: Carlo Strozzi: http://www.strozzi.it/cgi-bin/CSA/tw7/I/en_US/NoSQL/
  - "Not Only SQL"?
    - but many RDBMS are also "not just SQL"

"NoSQL is an accidental term with no precise definition"

- first used at an informal meetup in 2009 in San Francisco (presentations from Voldemort, Cassandra, Dynomite, HBase, Hypertable, CouchDB, and MongoDB)

[Sadalage & Fowler: NoSQL Distilled, 2012]

NoSQL Databases

- NoSQL: Database technologies that are (mostly):
    - Not using the relational model (nor the SQL language)
    - Designed to run on large clusters (horizontally scalable)
    - No schema - fields can be freely added to any record
    - Open source
    - Based on the needs of 21st century web estates

[Sadalage & Fowler: NoSQL Distilled, 2012]

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

NoSQL Databases

- Other characteristics (often true):
    - easy replication support (fault-tolerance, query efficiency)
    - simple API
    - eventually consistent (not ACID)

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Assumptions about Data and Usage

	RDBMS	**NoSQL**
integrity	is mission-critical	OK as long as most data is correct
data format	consistent, well-defined	unknown or inconsistent
data	is of long-term value	is expected to be replaced
growth	predictable, linear growth	unpredictable growth (exponential?)
querying	non-programmers writing queries	only programmers writing queries
fault tolerance	regular backups	automatic data replication
distribution	access through master server	data sharding (partitioning)

The End of Relational Databases?

- Relational databases are not going away
- Many projects would use RDBMS also because of:
  - maturity/stability,
  - available support
  - familiarity

We should see RDBMS as one option for data storage

Polyglot persistence – using different data stores in different circumstances

[Sadalage & Fowler: NoSQL Distilled, 2012]

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Data Model: Aggregates

- The model by which the database organizes data
- Each NoSQL type has a different data model
  - Key-value, document, column-family, graph
  - First three are oriented on aggregates

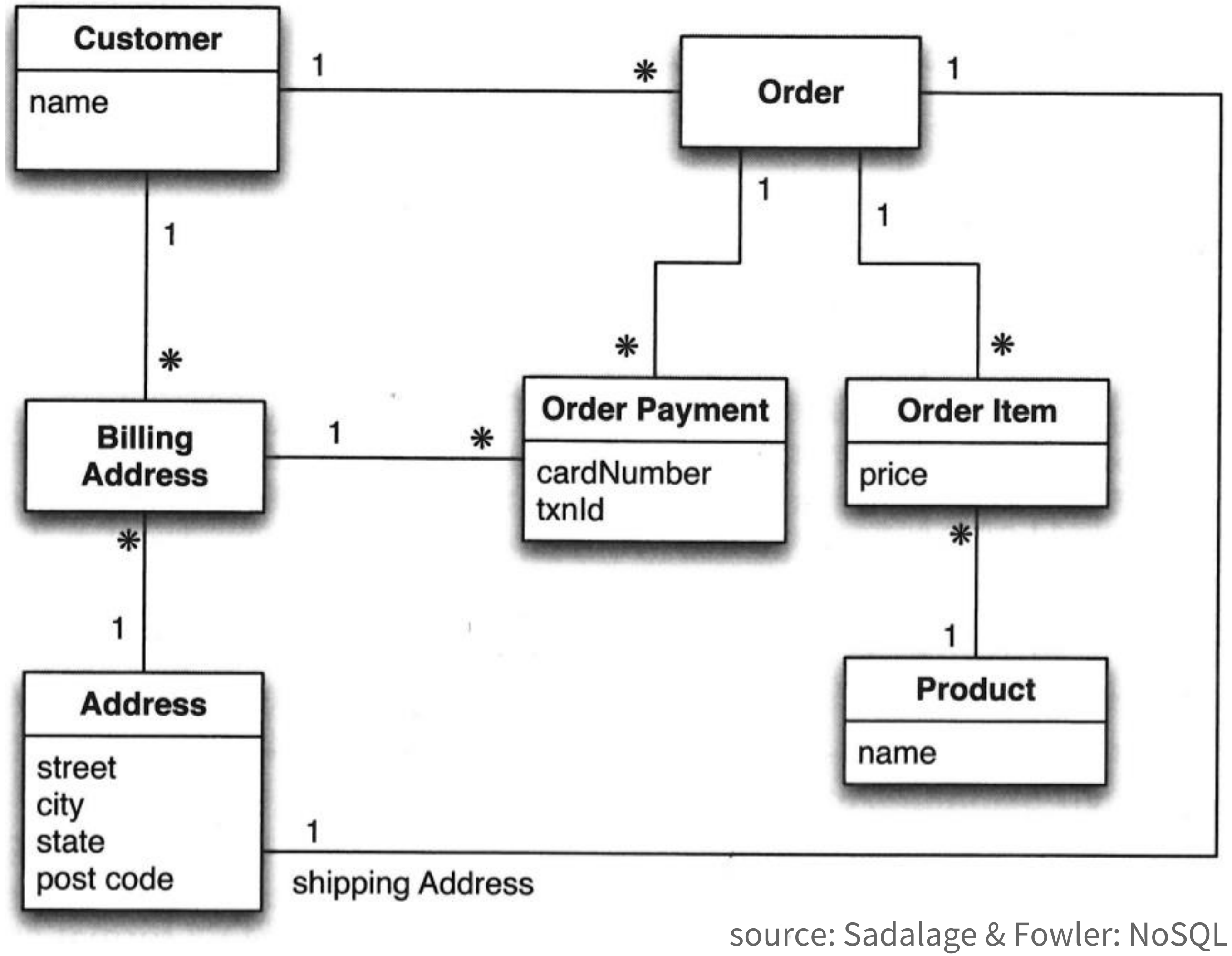
Aggregate

- A data unit with a complex structure
  - Not simply a tuple like in RDBMS
- An aggregate is a collection of related objects that we wish to treat as a unit
  - unit for data manipulation and management of consistency

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Example: UML Model of an e-shop



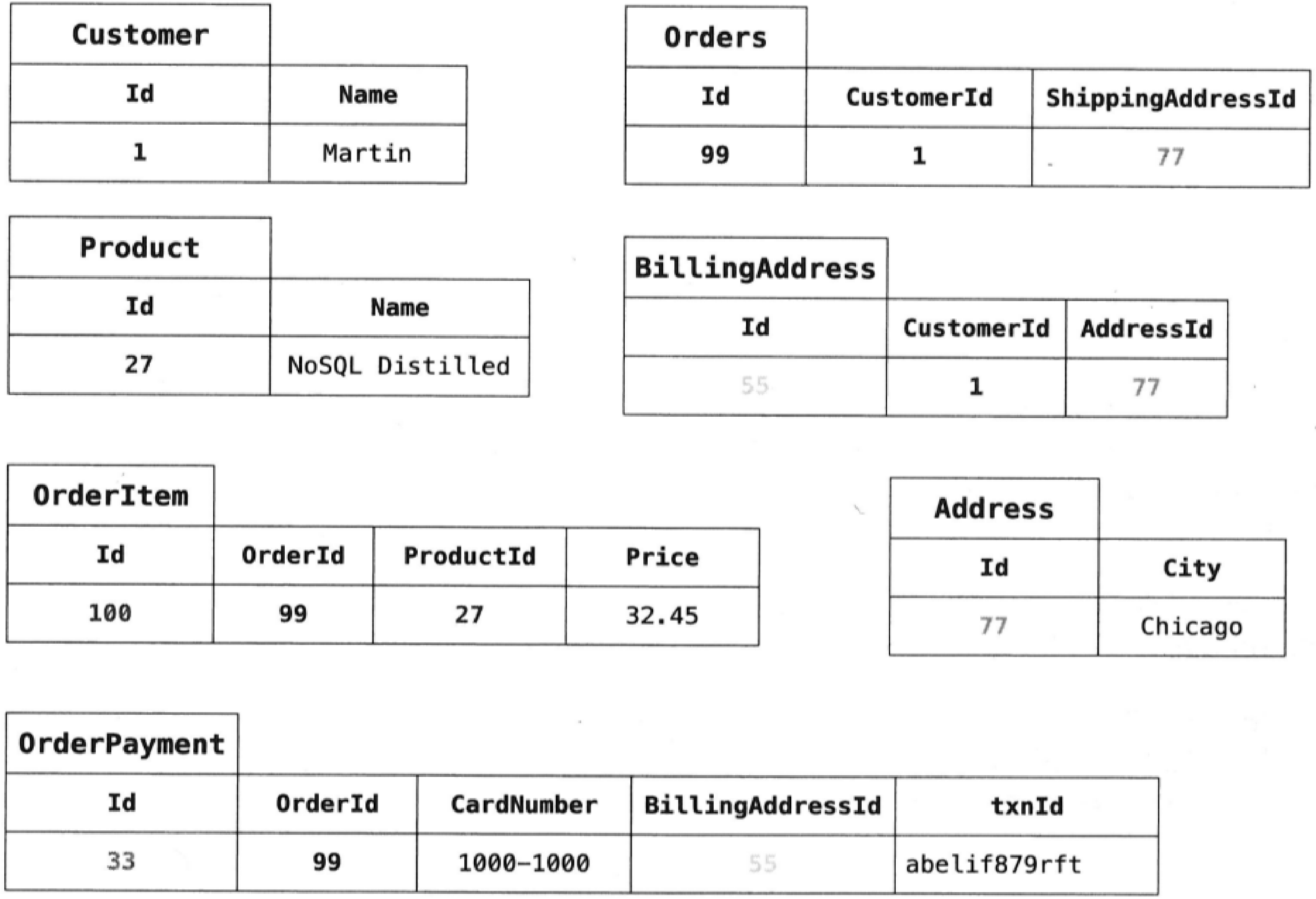
source: Sadalage & Fowler: NoSQL Distilled, 2012

Zdenka Prokopova
TBU in Zlín

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Example: Relational Model



Customer	
Id	Name
1	Martin

Orders		
Id	CustomerId	ShippingAddressId
99	1	77

Product	
Id	Name
27	NoSQL Distilled

BillingAddress		
Id	CustomerId	AddressId
55	1	77

OrderItem			
Id	OrderId	ProductId	Price
100	99	27	32.45

Address	
Id	City
77	Chicago

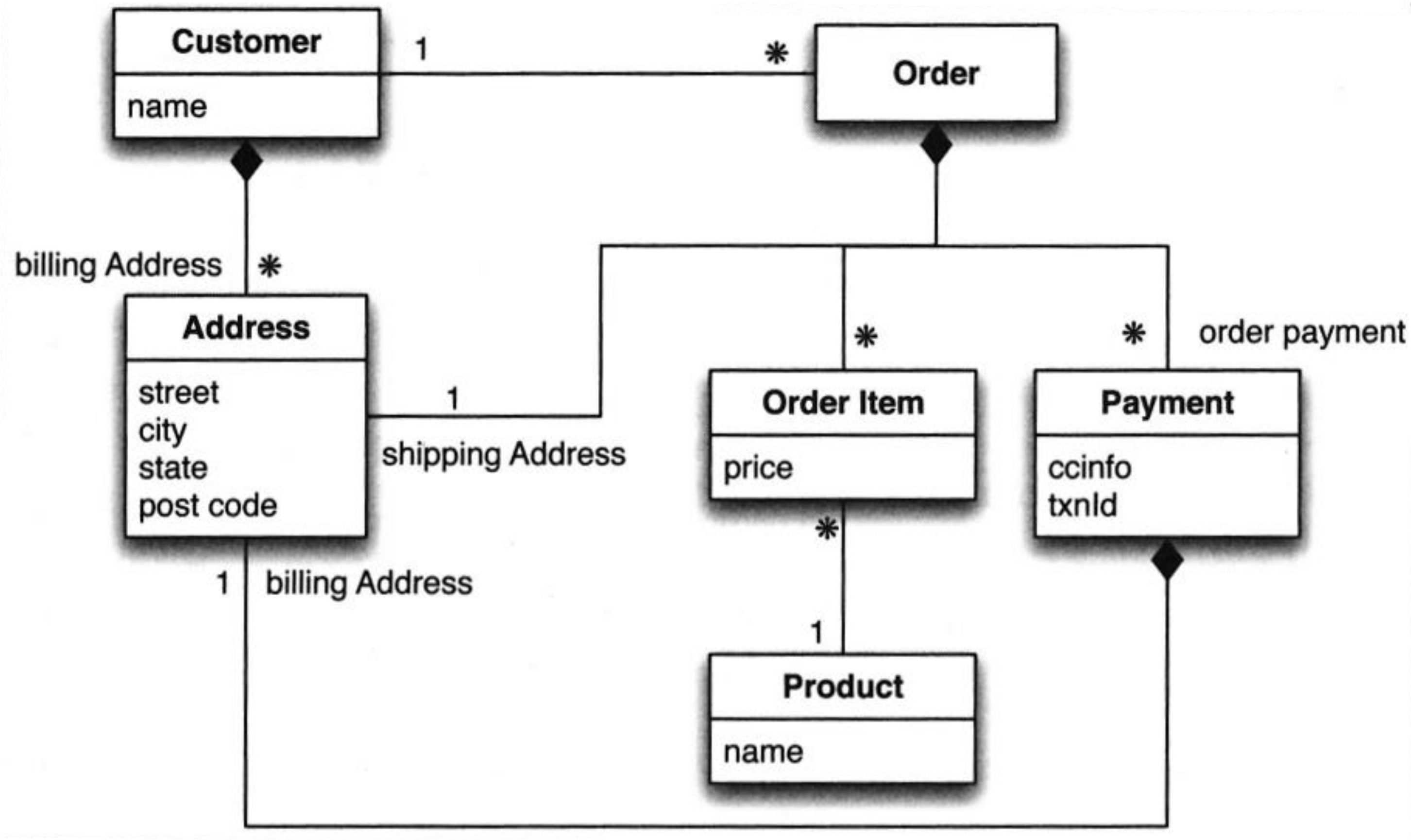
OrderPayment				
Id	OrderId	CardNumber	BillingAddressId	txnId
33	99	1000-1000	55	abelif879rft

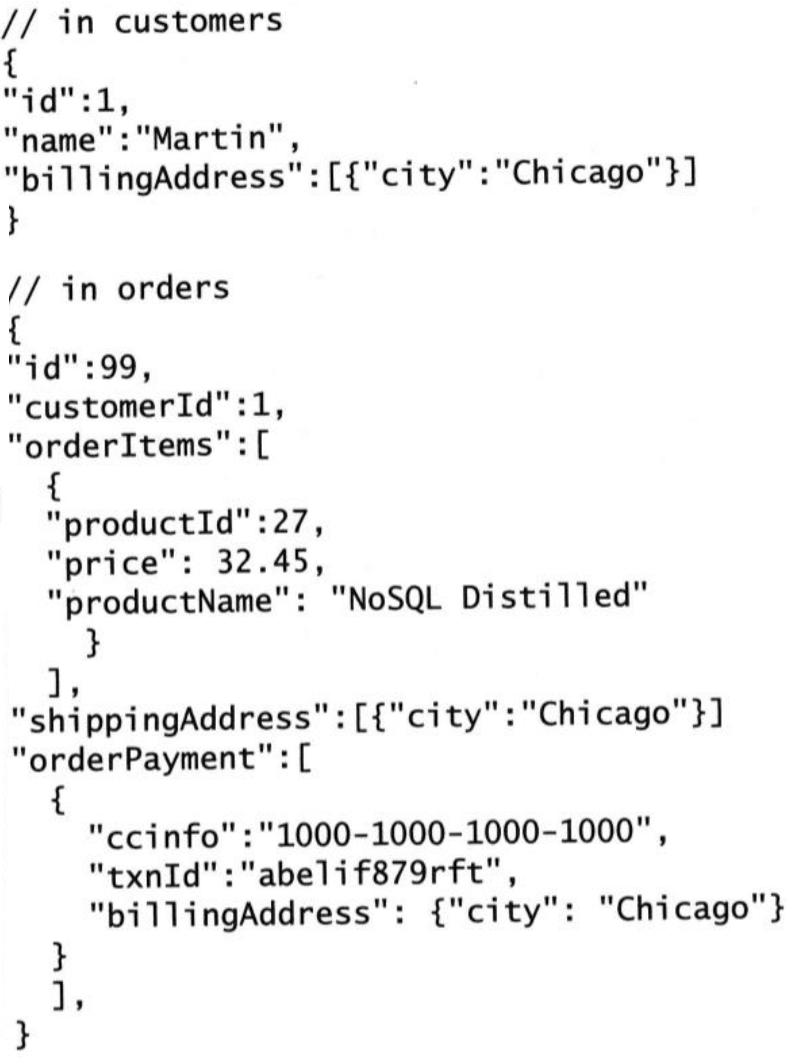
source: Sadalage & Fowler: NoSQL Distilled, 2012

Relational Model: Aggregate Ignorant

- Relational databases are aggregate-ignorant
    - It is not a bad thing, it is a feature
    - Allows to easily look at the data in different ways
    - Best choice when there is no primary structure for data manipulation

Example: NoSQL Solution





```
// in customers
{
"id":1,
"name":"Martin",
"billingAddress":[{"city":"Chicago"}]
}

// in orders
{
"id":99,
"customerId":1,
"orderItems":[
  {
  "productId":27,
  "price": 32.45,
  "productName": "NoSQL Distilled"
    }
  ],
"shippingAddress":[{"city":"Chicago"}]
"orderPayment":[
  {
    "ccinfo":"1000-1000-1000-1000",
    "txnId":"abelif879rft",
    "billingAddress": {"city": "Chicago"}
  }
  ],
}
```

source: Sadalage & Fowler: NoSQL Distilled, 2012

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

NoSQL Databases: Aggregate-oriented

- NoSQL databases are typically either:
    - schemaless (with implicit schema maintained by application)
    - or aggregate-oriented (more or less explicit schema)

Aggregate-oriented:

- There is no general strategy to set aggregate boundaries
- Aggregates give the database information about which bits of data will be manipulated together
    - Which should be stored on the same node

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Aggregate-oriented

Aggregates

- Helps greatly with running on a cluster of nodes
  - Minimize the number of nodes accessed during a search

- Impact on concurrency control:
  - NoSQL databases typically support atomic manipulation of a single aggregate at a time

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Four Basic Types of NoSQL Databases

- Key-value stores
- Document databases
- Column-family stores
- Graph databases

Zdenka Prokopova
TBU in Zlín

Key-value Stores: Representatives

redis

riak

MEMCACHED

MapDB

LevelDB

ORACLE BERKELEY DB 12c

ORACLE NOSQL DATABASE

Infinispan

amazon DynamoDB

Project Voldemort

Ranked list: http://db-engines.com/en/ranking/key-value+store

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Document Databases: Basics

- Basic concept of data: *Document*
- Documents are self-describing pieces of data
  - Hierarchical tree data structures
  - Nested associative arrays (maps), collections, scalars
  - XML, JSON (JavaScript Object Notation), BSON, …
- Documents in a collection should be “similar”
  - Their schema can differ
- Documents stored in the value part of key-value
  - Key-value stores where the values are examinable
  - Building search indexes on various keys/fields

Document Databases: Representatives

Apache
CouchDB
relax

terrastore

Ranked list: http://db-engines.com/en/ranking/document+store

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

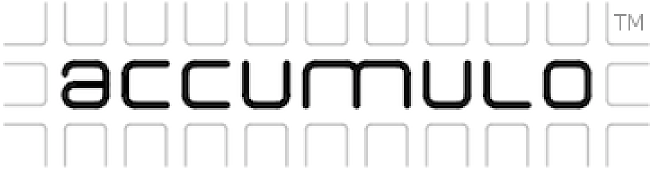
MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Column-family Stores: Basics

- wide-column, columnar
- Data model: rows that have many columns associated with a row key
- Column families are groups of related data (columns) that are often accessed together
  - e.g., for a customer we typically access all profile information at the same time, but not customer's orders

Zdenka Prokopova
TBU in Zlín

Column-family Stores: Representatives



Ranked list: http://db-engines.com/en/ranking/wide+column+store

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

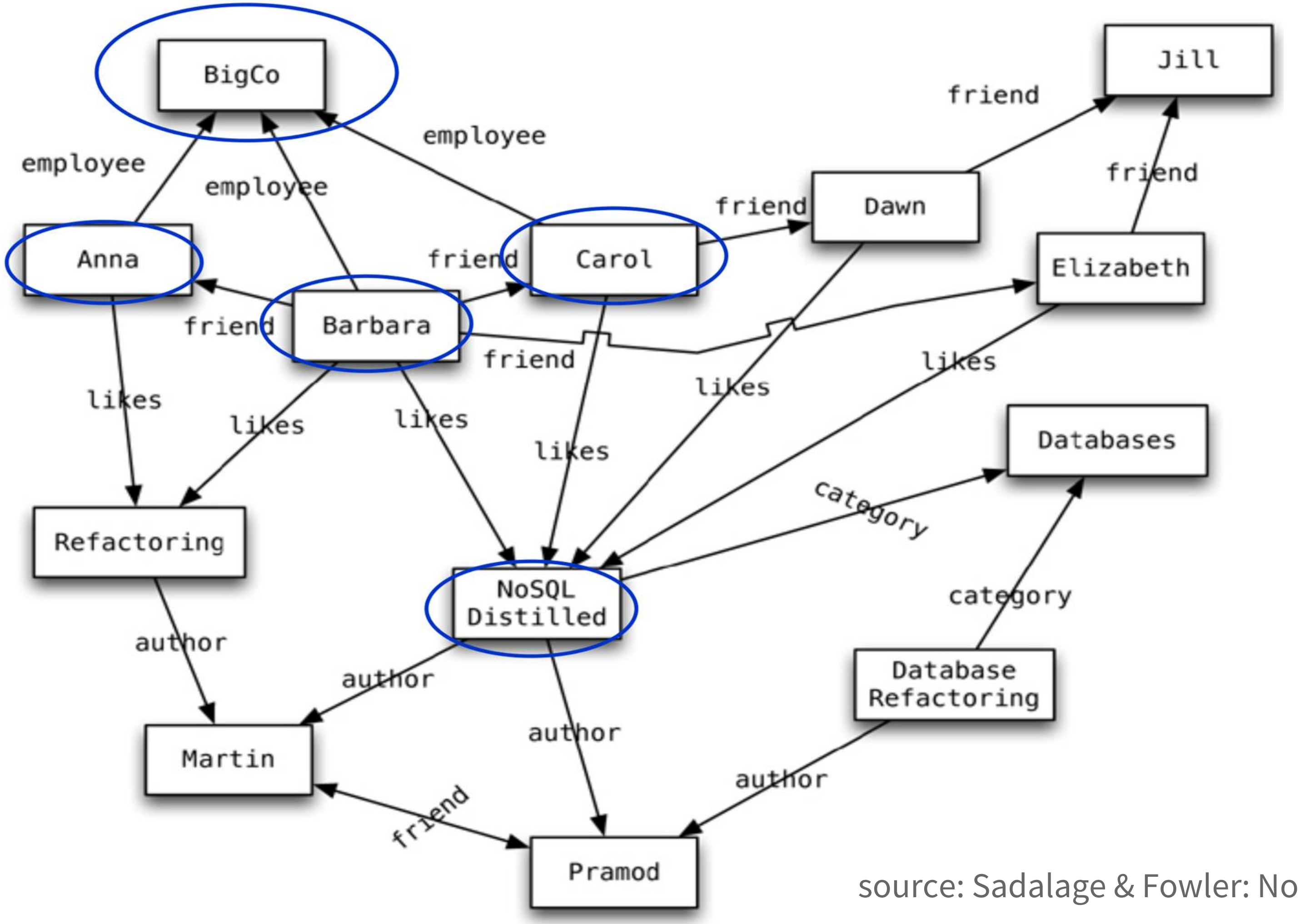
Graph Databases

- To store entities and relationships between them
  - Nodes are instances of objects
  - Nodes have properties, e.g., name
  - Edges have directional significance
  - Edges have types e.g., likes, friend, …
- Nodes are organized by relationships
  - Allow to find interesting patterns
  - example: Get all nodes that are “employee” of “Big Company” and that “likes” “NoSQL Distilled”

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Graph Databases: Example



Zdenka Prokopova
TBU in Zlín

source: Sadalage & Fowler: NoSQL Distilled, 2012

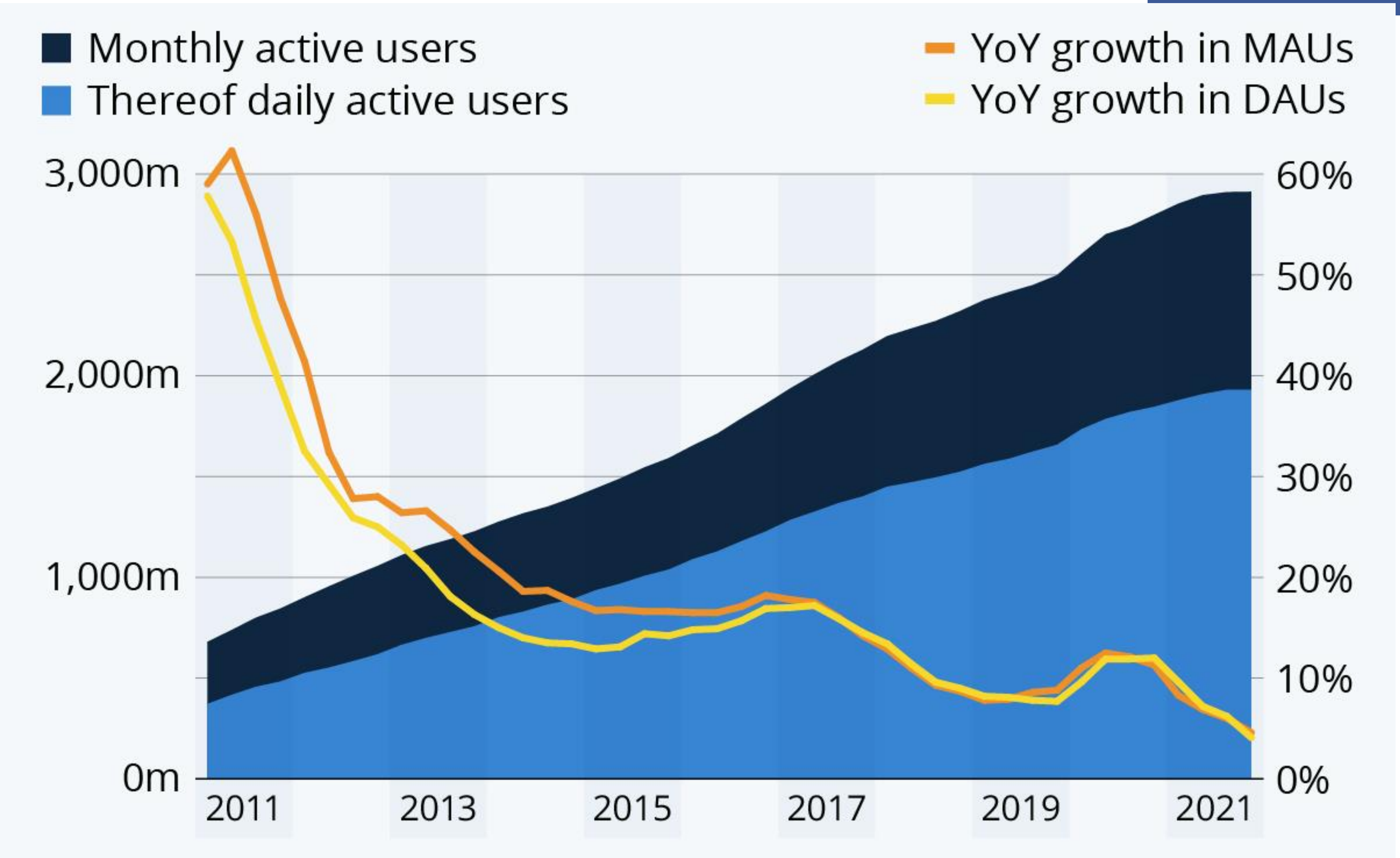
Graph Databases: Representatives

Ranked list: http://db-engines.com/en/ranking/graph+dbms

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

One Example of NoSQL Usage



Source: Facebook

Zdenka Prokopova
TBU in Zlín

Facebook: Database Technology Behind

Apache Hadoop http://hadoop.apache.org/

- Hadoop File System (HDFS)
  - over 100 PB in a single HDFS cluster
- an open source implementation of MapReduce:
  - Enables efficient calculations on massive amounts of data

Apache Hive http://hive.apache.org/

- SQL-like access to Hadoop-stored data
- integration of MapReduce query evaluation

sources: http://goo.gl/SZ6jia http://royal.pingdom.com/2010/06/18/the-software-behind-facebook/

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Facebook: Database Technology Behind

Apache HBase http://hbase.apache.org/

- a Hadoop column-family database
- used for e-mails, instant messaging and SMS
- replacement for MySQL and Cassandra

Memcached http://memcached.org/

- distributed key-value store
- used as a cache between web servers and MySQL servers since the beginning of FB

Zdenka Prokopova
TBU in Zlín

sources: http://goo.gl/SZ6jia http://royal.pingdom.com/2010/06/18/the-software-behind-facebook/

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Facebook: Database Technology Behind

Apache Giraph http://giraph.apache.org/

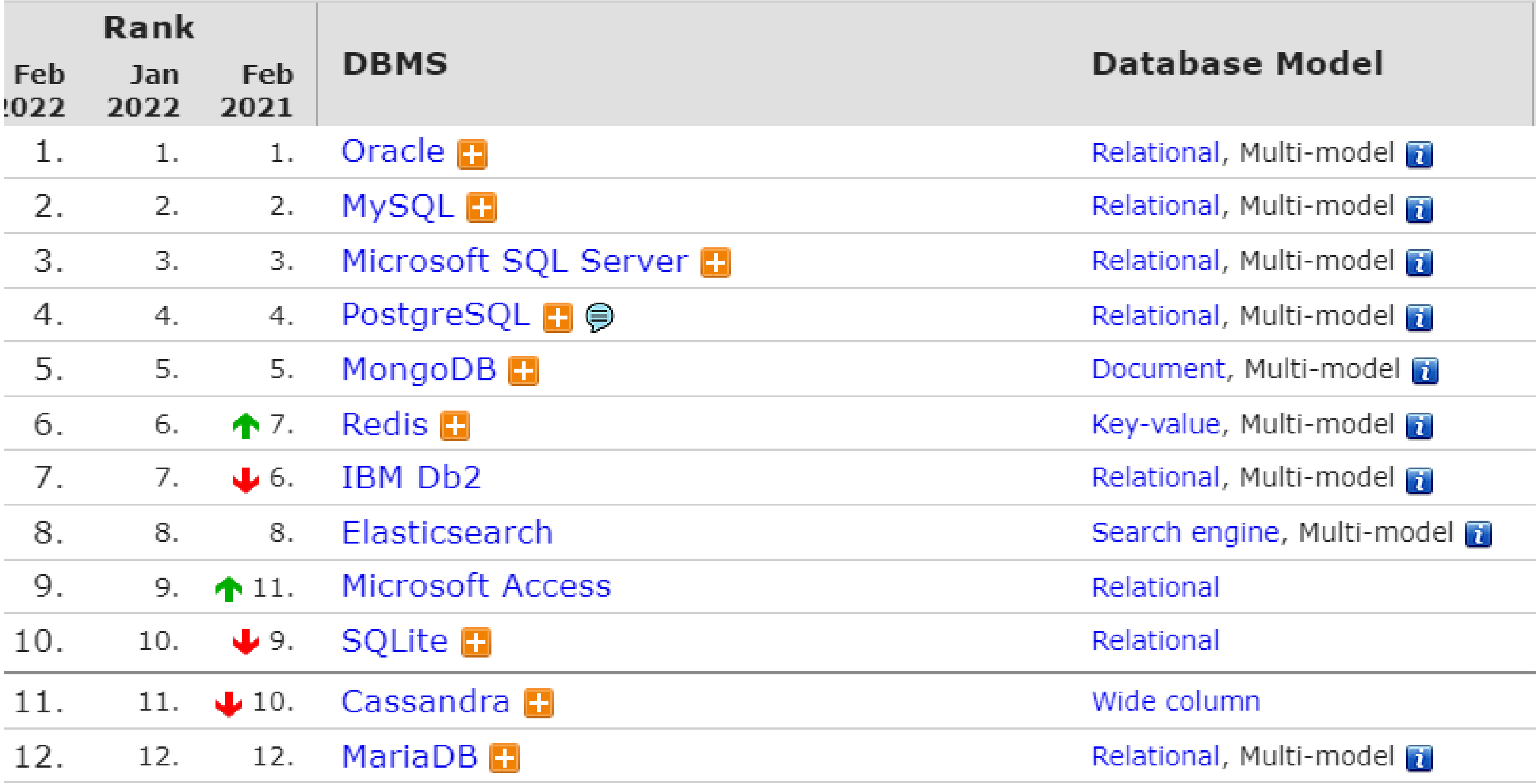
- graph database
- facebook users and connections is one very large graph
- used since 2013 for various analytic tasks

RocksDB http://rocksdb.org/

- high-performance key-value store
- developed internally in FB, now open-source

sources: https://code.facebook.com/posts/509727595776839/scaling-apache-giraph-to-a-trillion-edges/ http://goo.gl/XNtG6p

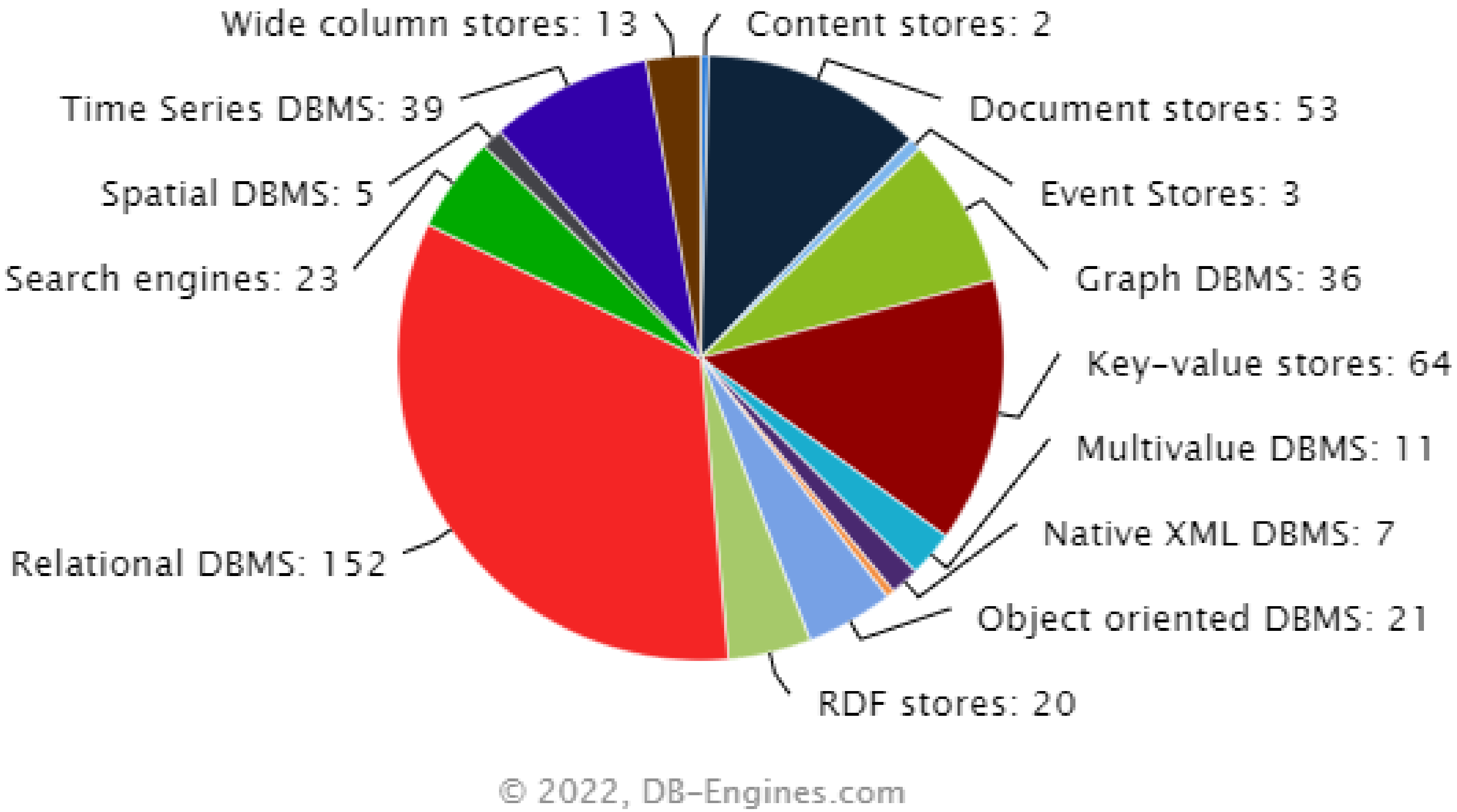
DB-Engines Ranking



Rank			DBMS	Database Model
Feb 022	Jan 2022	Feb 2021		
1.	1.	1.	Oracle	Relational, Multi-model
2.	2.	2.	MySQL	Relational, Multi-model
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model
4.	4.	4.	PostgreSQL	Relational, Multi-model
5.	5.	5.	MongoDB	Document, Multi-model
6.	6.	↑ 7.	Redis	Key-value, Multi-model
7.	7.	↓ 6.	IBM Db2	Relational, Multi-model
8.	8.	8.	Elasticsearch	Search engine, Multi-model
9.	9.	↑ 11.	Microsoft Access	Relational
10.	10.	↓ 9.	SQLite	Relational
11.	11.	↓ 10.	Cassandra	Wide column
12.	12.	12.	MariaDB	Relational, Multi-model

DB-Engines Ranking. (2022). Retrieved from https://db-engines.com/en/ranking

DBMS popularity per category



DB-Engines Ranking. (2022). Retrieved from https://db-engines.com/en/ranking

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

References

- Erl, T., Khattak, W., & Buhler, P. (2016). *Big Data Fundamentals: Concepts, Drivers \& Techniques*: Prentice Hall Press.
- Sadalage, P. J., & Fowler, M. (2013). *NoSQL distilled: a brief guide to the emerging world of polyglot persistence*: Pearson Education.
- Strauch, C., Sites, U.-L. S., & Kriha, W. (2011). NoSQL databases. *Lecture Notes, Stuttgart Media University, 20*.
- DB-Engines Ranking. (2019). Retrieved from https://db-engines.com/en/ranking
- Gain the insights, advice and tools (2019). Retrieved from https://www.gartner.com/en

References

- *RNDr. Irena Holubova, Ph.D. MMF UK course PA195: NoSQL Databases*
- *Data science and big data analytics*. (2018). New York, NY: Springer Berlin Heidelberg.
- Deka, G. C. (2017). *NoSQL : database for storage and retrieval of data in cloud*. Boca Raton, FL: CRC Press, Taylor & Francis Group.
- Perkins, L., Redmond, E., & Wilson, J. R. (2018). *Seven databases in seven weeks : a guide to modern databases and the NoSQL movement* (Second edition. ed.). Raleigh, North Carolina: The Pragmatic Bookshelf.
- Wiese, L. (2015). *Advanced data management : for SQL, NoSQL, cloud and distributed databases*. Berlin ; Boston: De Gruyter, Oldenbourg.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Questions?

Strategic project of TBU in Zlín, reg. no. CZ.02.2.69/0.0/0.0/16_015/0002204