Big Data Management
Knowledge objectives

1. Give a definition of Big Data
2. Name eight features of cloud databases
3. Give a definition of Distributed Database
4. Recognize the problem of impedance mismatch
5. Name different kinds of NOSQL databases
6. Recognize the main problems of NOSQL databases
Understanding Objectives

1. Estimate the cost of a distributed query
2. Transform the value in a schemaless database into a relational one
“Without data you are just another person with an opinion.”

William Edwards Deming

“It is a capital mistake to theorize before one has data.”

Sherlock Holmes (A Study in Scarlett)
The who, why and how of BIG DATA

WHO...

COMPANIES ARE SPENDING BIG ON BIG DATA

<table>
<thead>
<tr>
<th>Industry</th>
<th>Spending</th>
<th>Growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Financial Services</td>
<td>$6.4B</td>
<td>22%</td>
</tr>
<tr>
<td>Software/Internet</td>
<td>$2.8B</td>
<td>26%</td>
</tr>
<tr>
<td>Government</td>
<td>$2.8B</td>
<td>22%</td>
</tr>
<tr>
<td>Comms &amp; Media</td>
<td>$1.2B</td>
<td>40%</td>
</tr>
<tr>
<td>Energy/Utilities</td>
<td>$800M</td>
<td>54%</td>
</tr>
</tbody>
</table>

WHY...

THE COMPANIES THAT USE ANALYTICS BEST ARE...
It is estimated that in 2020 there will be more data than sand grains in the world (40 Zb)
Big Data definition

- Velocity
- Volume
- Variety
- ...
- Variability
- Validity/Veracity
- Value

From IBM “Understanding Big Data”
Bigbench

<table>
<thead>
<tr>
<th>Query processing type</th>
<th>Total</th>
<th>Percentage(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Declarative</td>
<td>10</td>
<td>33.3</td>
</tr>
<tr>
<td>Procedural</td>
<td>7</td>
<td>23.3</td>
</tr>
<tr>
<td>Mix of Declarative and Procedural</td>
<td>13</td>
<td>43.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data sources</th>
<th>Total</th>
<th>Percentage(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structured</td>
<td>18</td>
<td>60.0</td>
</tr>
<tr>
<td>Semi-structured</td>
<td>7</td>
<td>23.3</td>
</tr>
<tr>
<td>Un-structured</td>
<td>5</td>
<td>16.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Analytic techniques</th>
<th>Total</th>
<th>Percentage(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistics analysis</td>
<td>6</td>
<td>20.0</td>
</tr>
<tr>
<td>Data mining</td>
<td>17</td>
<td>56.7</td>
</tr>
<tr>
<td>Reporting</td>
<td>8</td>
<td>26.7</td>
</tr>
</tbody>
</table>
Big Data sources

- Structured
  - Created (i.e., business data)
  - Provoked (e.g., customer feedback)
  - Transacted
  - Compiled (e.g., demographics)
  - Experimental (e.g., sampling customers)

- Unstructured
  - Captured (e.g., search words)
  - User-generated (e.g., social networks)
## Types of Big Data Analyzed in Industry

<table>
<thead>
<tr>
<th></th>
<th>Manufacturing and Natural Resources</th>
<th>Media/Communications</th>
<th>Services</th>
<th>Government</th>
<th>Education</th>
<th>Retail</th>
<th>Banking</th>
<th>Insurance</th>
<th>Healthcare</th>
<th>Transportation</th>
<th>Utilities</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Transactions</strong></td>
<td>73%</td>
<td>62%</td>
<td>67%</td>
<td>67%</td>
<td>54%</td>
<td>93%</td>
<td>83%</td>
<td>81%</td>
<td>75%</td>
<td>79%</td>
<td>80%</td>
</tr>
<tr>
<td><strong>Log data</strong></td>
<td>44%</td>
<td>57%</td>
<td>58%</td>
<td>59%</td>
<td>54%</td>
<td>40%</td>
<td>66%</td>
<td>61%</td>
<td>33%</td>
<td>71%</td>
<td>60%</td>
</tr>
<tr>
<td><strong>Machine or sensor data</strong></td>
<td>53%</td>
<td>38%</td>
<td>35%</td>
<td>33%</td>
<td>31%</td>
<td>27%</td>
<td>27%</td>
<td>48%</td>
<td>42%</td>
<td>50%</td>
<td>40%</td>
</tr>
<tr>
<td><strong>Emails/documents</strong></td>
<td>27%</td>
<td>43%</td>
<td>43%</td>
<td>41%</td>
<td>46%</td>
<td>27%</td>
<td>34%</td>
<td>39%</td>
<td>17%</td>
<td>29%</td>
<td>20%</td>
</tr>
<tr>
<td><strong>Social media data</strong></td>
<td>32%</td>
<td>52%</td>
<td>39%</td>
<td>26%</td>
<td>54%</td>
<td>73%</td>
<td>27%</td>
<td>13%</td>
<td>-</td>
<td>50%</td>
<td>-</td>
</tr>
<tr>
<td><strong>Free-form text</strong></td>
<td>17%</td>
<td>24%</td>
<td>28%</td>
<td>30%</td>
<td>31%</td>
<td>20%</td>
<td>34%</td>
<td>35%</td>
<td>67%</td>
<td>21%</td>
<td>40%</td>
</tr>
<tr>
<td><strong>Geospatial data</strong></td>
<td>27%</td>
<td>14%</td>
<td>19%</td>
<td>19%</td>
<td>38%</td>
<td>27%</td>
<td>27%</td>
<td>26%</td>
<td>8%</td>
<td>29%</td>
<td>40%</td>
</tr>
<tr>
<td><strong>Images</strong></td>
<td>19%</td>
<td>24%</td>
<td>17%</td>
<td>11%</td>
<td>38%</td>
<td>13%</td>
<td>5%</td>
<td>16%</td>
<td>25%</td>
<td>7%</td>
<td>-</td>
</tr>
<tr>
<td><strong>Video</strong></td>
<td>8%</td>
<td>29%</td>
<td>12%</td>
<td>7%</td>
<td>31%</td>
<td>13%</td>
<td>-</td>
<td>6%</td>
<td>8%</td>
<td>7%</td>
<td>-</td>
</tr>
<tr>
<td><strong>Audio</strong></td>
<td>10%</td>
<td>19%</td>
<td>8%</td>
<td>4%</td>
<td>8%</td>
<td>-</td>
<td>-</td>
<td>6%</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td>8%</td>
<td>14%</td>
<td>13%</td>
<td>15%</td>
<td>8%</td>
<td>7%</td>
<td>10%</td>
<td>16%</td>
<td>42%</td>
<td>14%</td>
<td>-</td>
</tr>
</tbody>
</table>

**n =** 59 21* 127 27* 13* 15* 41 31 12* 14* 5*

Note: Highlighted cells indicate the top three data types by industry. Multiple responses allowed.

Source: Gartner (September 2013)
Big Data facets

- The Original
- as Technology
- as Data Distinctions
- as Signals
- as Opportunity
- as Metaphor
- as New Term for Old Stuff

Timo Elliott
Big Data related areas

- **Volume and Velocity**
  - Declarative querying
  - Query optimization

- **Variety and Variability**
  - Data quality
  - Data integration
  - Web and text mining
  - Information retrieval

- **Validity/Veracity**
  - Data consistency
  - Uncertainty
  - Statistical reasoning
  - Data linkage (provenance)

- **Value**
  - Analytics
    - Data mining
    - Algorithmics
    - Automatic learning
    - Simulation
    - Privacy
  - Biologists
  - Linguistics
  - Chemists
  - Sociologists
  - Engineers

- **Distributed Databases and Big Data**
- September 2013 Alberto Abelló & Oscar Romero
Key features of cloud databases

a) Quick/Cheap set up
b) Ability to horizontally scale
c) Ability to replicate & distribute (fragmentation)
d) Simple call level interface or protocol
   - No declarative query language
e) Weaker concurrency model than ACID
f) Efficient use of distributed indexes and RAM
g) Flexible schema
   - Ability to dynamically add new attributes
h) Multi-tenancy
Distributed Database

- A distributed database (DDB) is a database where data management is distributed over several nodes in a network.
  - Each node is a database itself
    - Potential heterogeneity
  - Nodes communicate through the network
Parallel database architectures

Shared nothing

Shared disc

Shared memory

Figure from D. Abady
Activity

- **Objective:** Recognize the benefits of distributing data

- **Tasks:**
  1. (5’) Individually solve one exercise
  2. (10’) Explain the solution to the others
  3. Hand in the three solutions

- **Roles for the team-mates during task 2:**
  a) Explains his/her material
  b) Asks for clarification of blur concepts
  c) Mediates and **controls time**
Impedance Mismatch

Of hammers and nails...

The Law of the Hammer

If the only tool you have is a hammer, everything looks like a nail.

Abraham Maslow - The Psychology of Science - 1966
Petra Selmer, Advances in Data Management 2012
Impedance Mismatch

The Law of the Relational Database

If the only tool you have is a relational database, everything looks like a table.

A Walk in Graph Databases - 2012
Petra Selmer, Advances in Data Management 2012
**Schemaless Databases**

CREATE TABLE Student (id int, name varchar2(50), surname varchar2(50), enrolment date);

- Insert into Student (1, 'Oscar', 'Romero', '01/01/2012', true, 'Lleida');
- Insert into Student (1, 'Oscar', 'Romero', NULL);
- Insert into Student (1, 'Oscar', 'Romero', '01/01/2012');

Consequences (?) – 2 mins to think of them
- Gain flexibility
- Lose semantics (also consistency)
- May reduce the impedance mismatch
- The data independence principle is lost (!)
- Applications can access and manipulate the database internal structures

The ANSI / SPARC architecture is not followed

**Distributed Databases and Big Data**

October 2013

Alberto Abelló & Oscar Romero

19
Different applications

Not Only SQL (different problems entail different solutions)

- OLTP
  - Object-Relational
    - Distributed databases
    - Parallel databases

- Scientific databases and other Big Data repositories
  - Key-value stores

- Data Warehousing & OLAP
  - MOLAP
  - Column stores
  - Multidimensional features

- Text / documents
  - Document databases
    - XML/JSON databases

- Stream processing
  - Stream processor

- Semantic Web and Open Data
  - Graph databases
Schemaless Databases

- NOSQL solution for the impedance mismatch
- Several new data models were introduced
  - Graph data model
  - Document-oriented databases
  - Key-value (~ hash tables)
  - Streams (~ vectors and matrixes)
- These *new* models lack of an explicit schema (defined by the user)
  - However, an implicit schema remains
Internal Structures

- Riak, MongoDB etc
- RAM
- Disk
- Hbase, Cassandra, RocksDB etc
- LSM

- Kafka
- Queues are Databases - 1995 Jim Gray

- Redshift etc. Parquet (Hadoop)

Ben Stopford
Progscon & JAX Finance 2015
Polyglot Systems

- Federate different kinds of storage systems

Martin Fowler

http://martinfowler.com/bliki/PolyglotPersistence.html
NOSQL drawbacks

- No ACID
- No standard
- Low-level query

Michael Stonebraker
The Problem is Not SQL

- Relational systems are too generic...
  - OLTP: stored procedures and simple queries
  - OLAP: ad-hoc complex queries
  - Documents: large objects
  - Streams: time windows with volatile data
  - Scientific: uncertainty and heterogeneity

- ... But the overhead of RDBMS has nothing to do with SQL
  - Low-level, record-at-a-time interface is not the solution

SQL Databases vS. NoSQL Databases
Michael Stonebraker
Communications of the ACM, 53(4), 2010
Brewery or bottled beer?

Do It Yourself
- Expensive
- Ad hoc development

Off the Shelf
- Economies of scale
- Concrete functionalities

Florian Waas analogy
Specific platforms

- **Google BigTable**
  - Published in 2006
  - Implemented by Hbase
    - Also Dynamo and Cassandra

- **Google MapReduce**
  - Published in 2004
  - Implemented by Hadoop

- **MongoDB**
  - Published in 2007

- **Neo4J/Sparksee**
  - Published in 2010/2008

- **SAP HANA**
  - Published in 2011
  - Prototyped in SanssouciDB
Summary

- Big Data definition
- Key features of cloud software (i.e., DBMS)
- Distributed Database definition
- Impedance Mismatch
- NOSQL main goals and features
Bibliography

- R. Cattell. *Scalable SQL and NoSQL Data Stores.* SIGMOD Record 39(4), 2010