Tietovaraston tiedon laadun hallinta

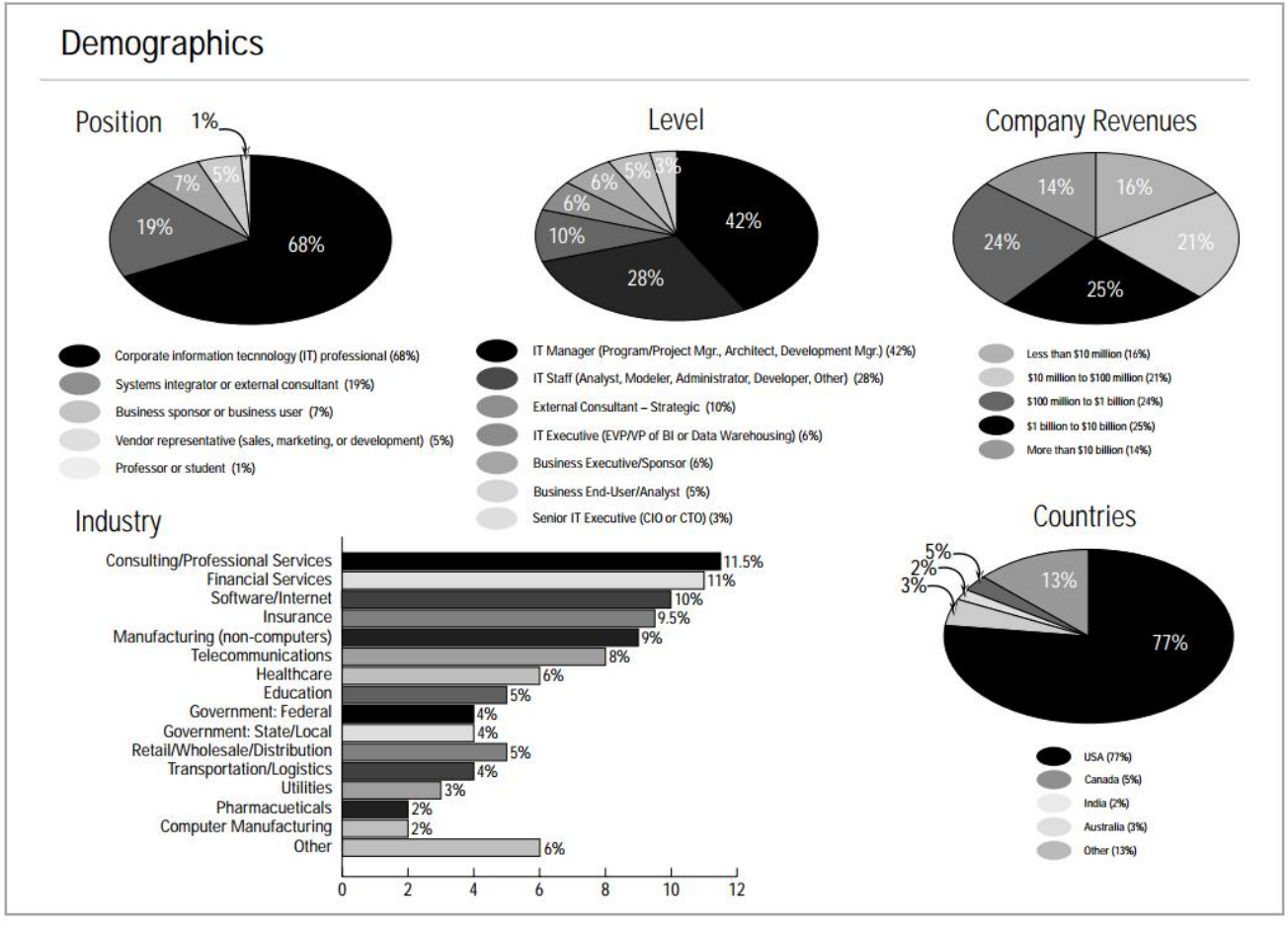
Arno Karatmaa

SUGIF laivaseminaari 15.3. -17.3.2016

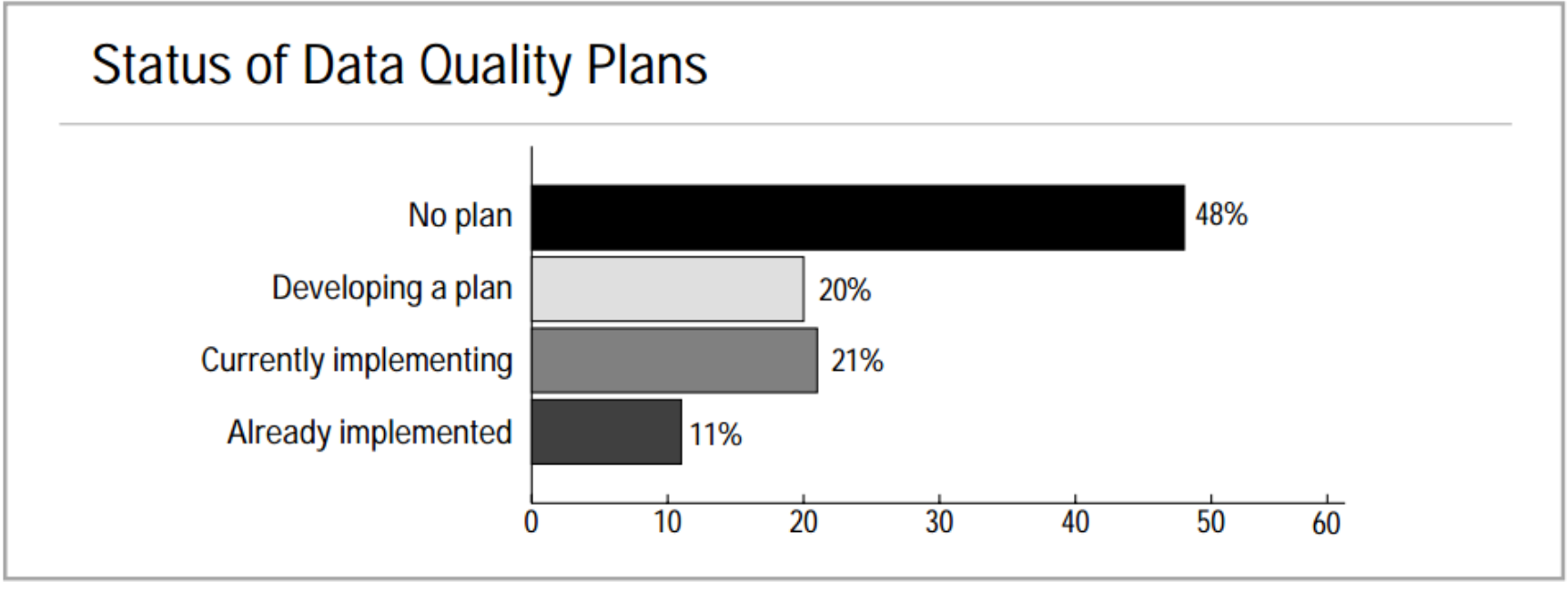
Mitä on tiedon laatu?

- Vastaavuus reaalimaailmaan
- Kattavuus, ajantasaisuus, ristiriidattomuus, tarkkuus
- Relevanssi, esitystapa, luotettavuus, käytettävyys
- Piirteet jotka tukevat annetun tarkoituksen täyttymistä
- Prosessit ja teknologiat jotka varmistavat hyväksymiskriteerien täyttymisen
- Standardoitu, aikaleimattu
- ISO 9000 (2015): vaatimusmäärittelyn näkökulma

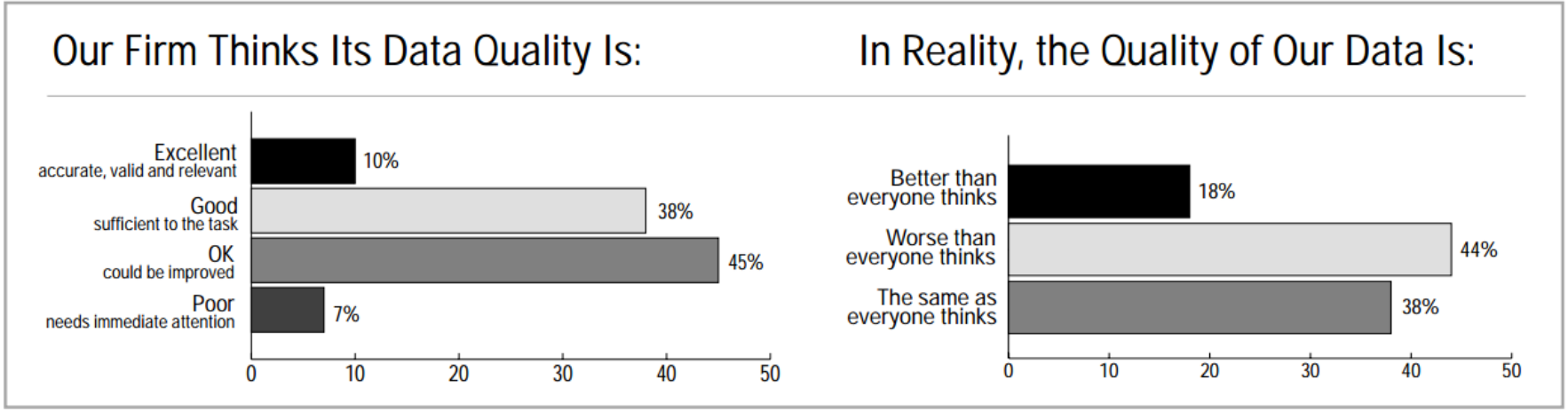
Liiketoimintanäkökulma



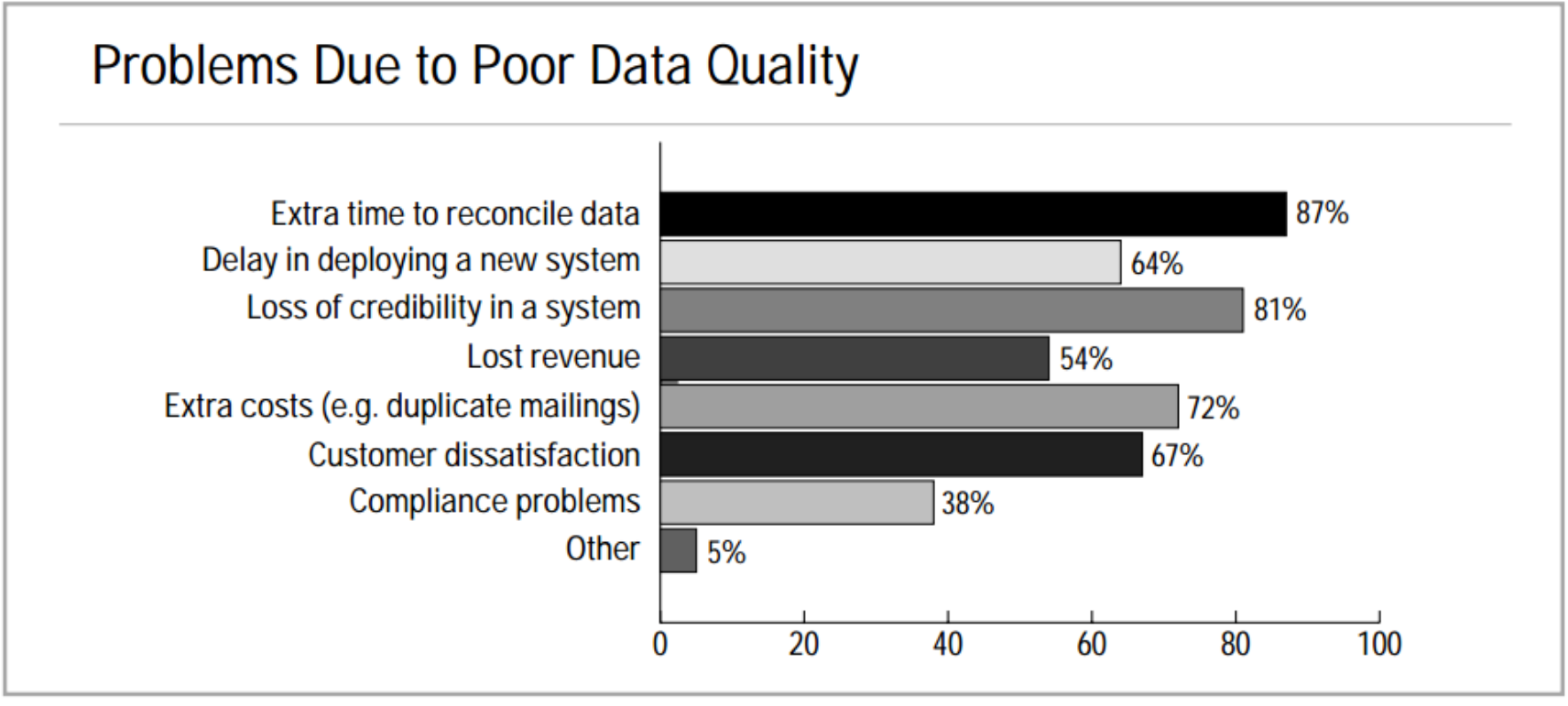
Liiketoimintanäkökulma



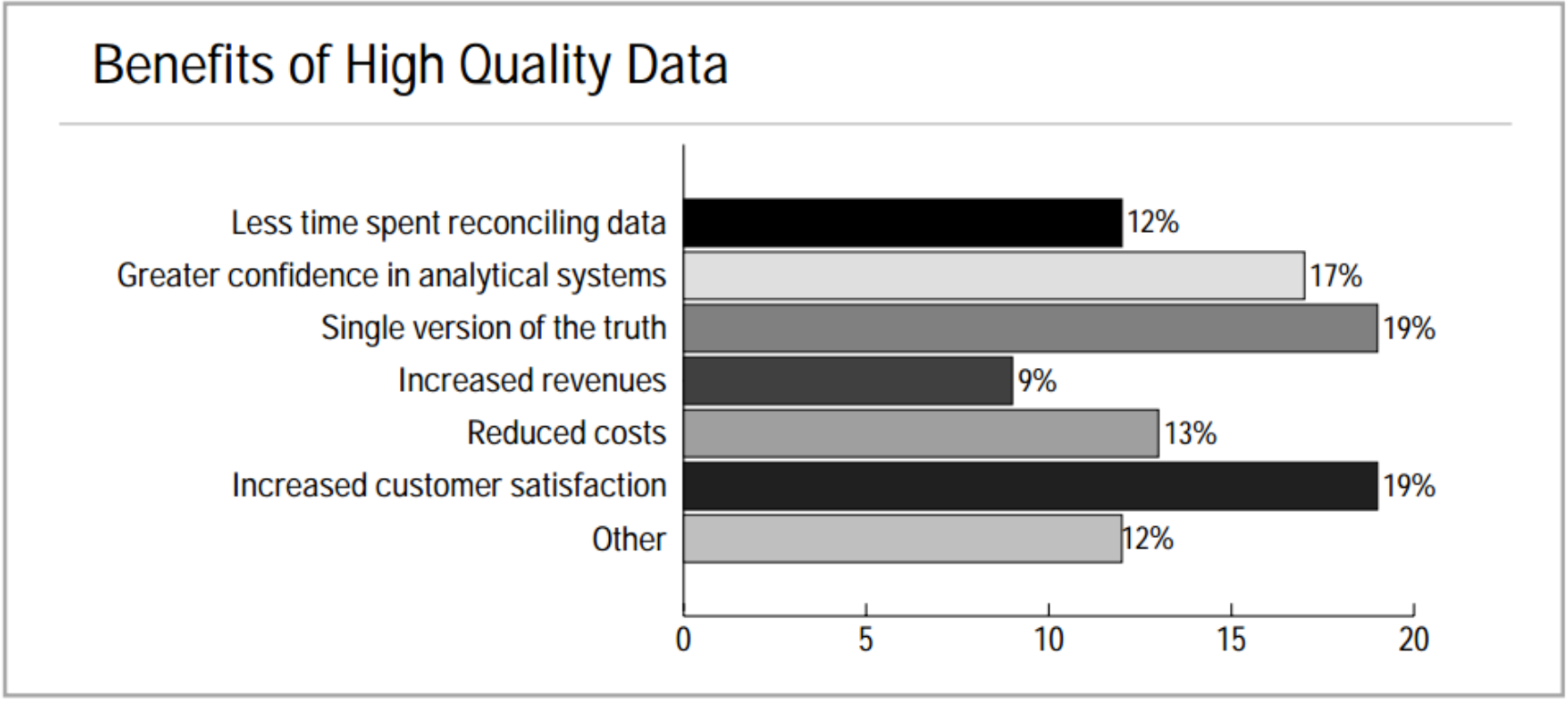
Liiketoimintanäkökulma



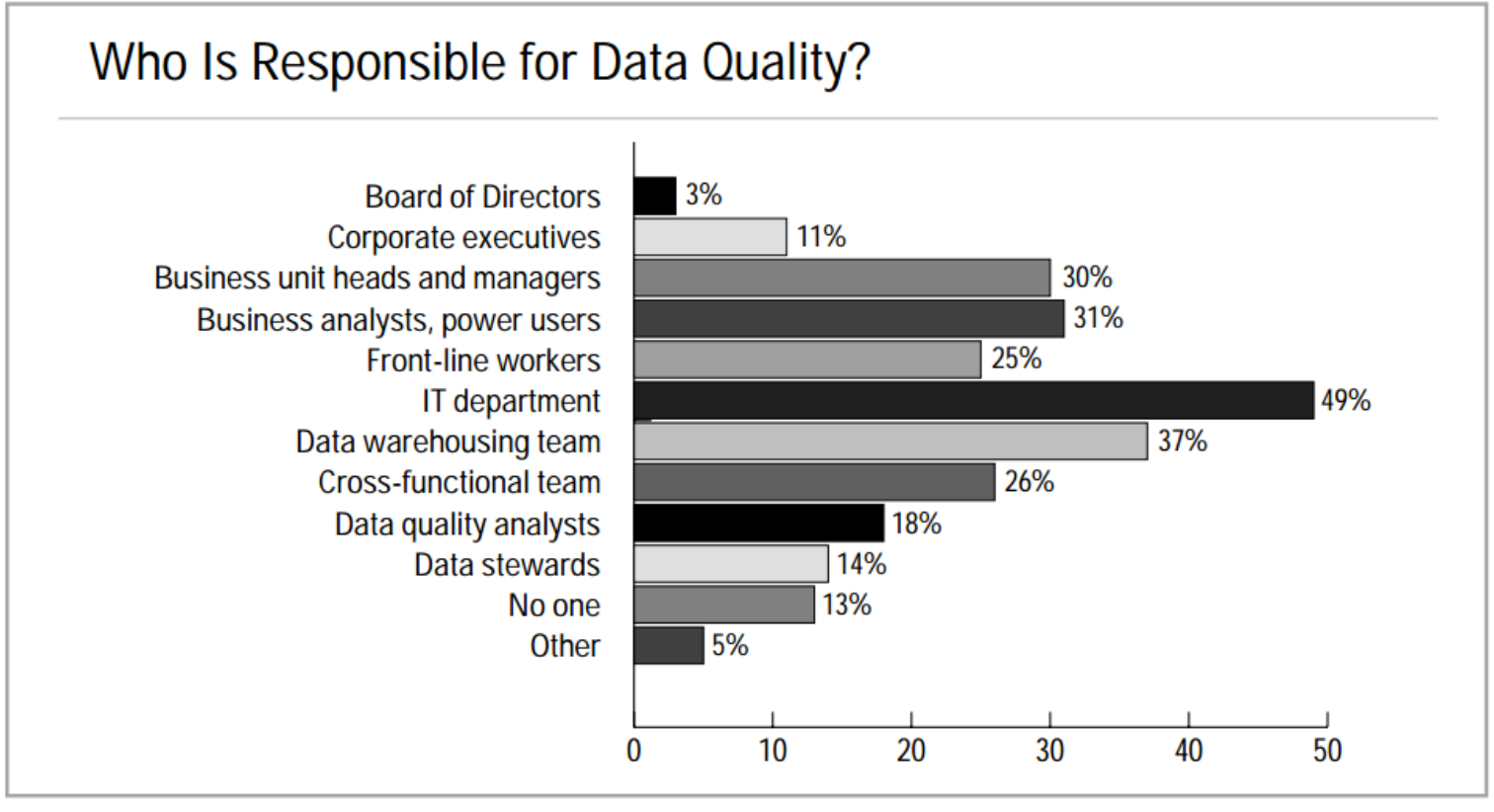
Liiketoimintanäkökulma



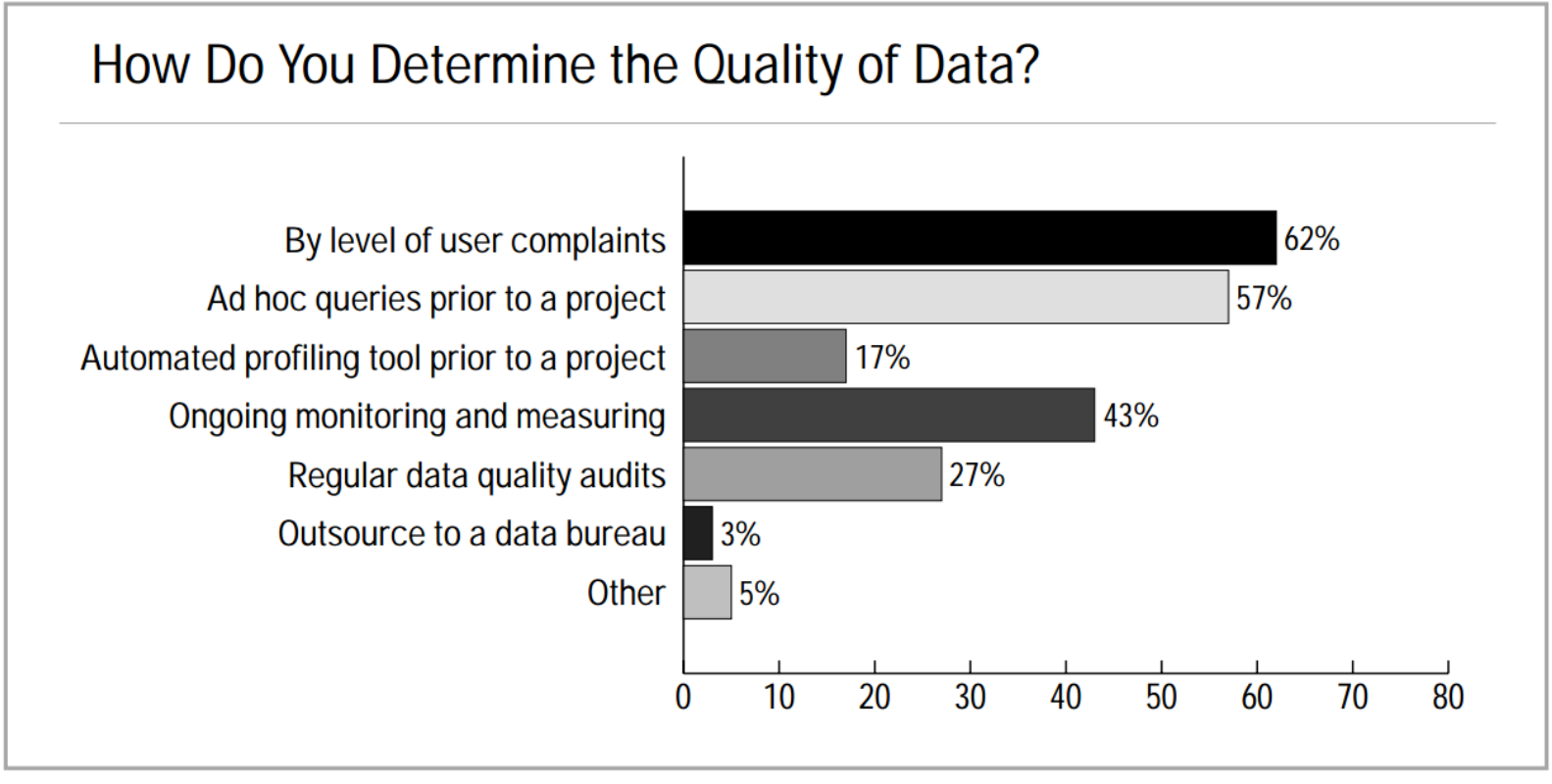
Liiketoimintanäkökulma



Liiketoimintanäkökulma



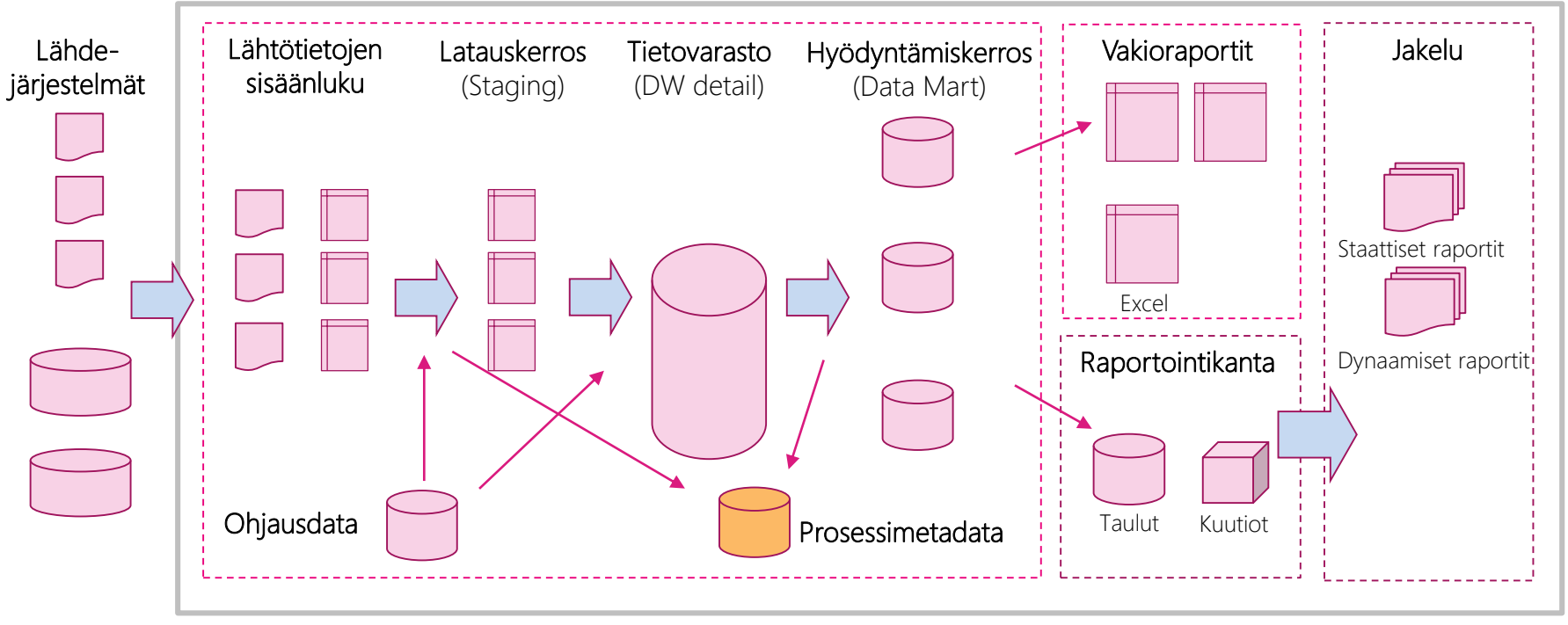
Liiketoimintanäkökulma



Datan laadun hinta

- Hyvän laadun hinta
- Huonon laadun hinta
- Datan laadun tavoitetaso

Tietovaraston arkkitehtuuri



Datavirheiden luokittelu

- Tietolähde
- Profilointi
- ETL-vaihe
- Tiedon mallinnus

Datavirheiden luokittelu - tietolähde

- 1 Inadequate selection of candidate data sources cause DQ Problems (sources which do not comply to business rules)
- 2 As time and proximity from the source increase, the chances for getting correct data decrease [4].
- 3 In adequate knowledge of inter dependencies among data sources incorporate DQ problems.
- 4 Inability to cope with ageing data contribute to data quality problems.[4]
- 5 Varying timeliness of data sources [6] [7].
- 6 Lack of validation routinesat sources causes DQ Problems.
- 7 Unexpected changes in source systems cause DQ Problems.
- 8 Multiple data sources generate semantic heterogeneity which leads to data quality issues [1][4]
- 9 The complexity of a data warehouse increases geometrically with the span of time of data to be fed into it
- 10 Usage of decontrolled applications and databases as data sources for data warehouse in the organizations.
- 11 Use of different representation formats in data sources.
- 12 Measurement errors [11].
- 13 Non-Compliance of data in data sources with the Standards.
- 14 Failure to update the sources in a timely manner causes DQ Problems.
- 15 Failure to update all replicas of data causes DQ Problems.
- 16 Presence of duplicate records of same data in multiple sources cause DQ Problems [6] [7] [11].
- 17 Approximations or surrogates used in data sources.
- 18 Contradictory information present in data sources cause data quality problems [6] [7].
- 19 Different encoding formats (ASCII, EBCDIC,....) [11]
- 20 Inadequate data quality testing on individual data source lead to poor data quality.

Datavirheiden luokittelu - tietolähde

- Lack of business ownership, policy and planning of
- the entire enterprise data contribute to data
- quality problems. [4]
- 22 Columns .having incorrect data values (for Eg.
- The AgeInYear Column for a person contain value
- 3 although the birthdate column is having value
- Aug, 14, 1967) [6] [7].
- 23 Having Inconsistent/Incorrect data formatting
- (The name of a person is stored in one table in the
- format "Firstname Surname" and in another table
- in the forma t "Surname, Firstname") [6] [7] [11].
- 24 System fields designed to allow free forms (Field
- not having adequate length).
- 25 Missing Columns (You need a middle name of a
- person but a column for it does not exist.) [6][7].
- 26 Missing values in data sources [2][11][12].
- 27 Misspelled data [11][12]
- 28 Additional columns [6] [7] [11].
- 29 Multiple sources for the same data ((For Eg.
- customer information is stored in three separate
- legacy databases)[11].
- 30 Various key strategies for the same type of entity
- (for Eg. One table stores customer information
- using the Social SecurityNumber as the key,
- another uses the ClientID as the key, and another
- uses a surrogate key) [6] [7].
- 31 Inconsistent use of special characters (for Eg. A
- date uses hyphens to separate the year, month, and
- day whereas a numerical value stored as a string uses hyphens to indicate negative numbers)[11] [20].

Datavirheiden luokittelu - tietolähde

- 32 Different data types for similar columns (A customer ID is stored as a number in one table and a string in another).
- 33 Varying default values used for missing data [6] [7].
- 34 Various representations of data in source data (The day of the week is stored as T, Tues, 2, and Tuesday in four separate Columns) [6] [7] [20].
- 35 Lack of record level validation in source Data.
- 36 Data values stray from their field description and business rules (Such as the Maiden Name column is being used to store a person's Hobbies, zip code into phone number box) [6] [7].
- 37 Inappropriate data relationships among tables.
- 38 Unrealized data relationships between data members.
- 39 Not specifying NULL character properly in flat files data sources result in wrong data.
- 40 Delimiter that comes as a character in some field of the file may represent different meaning of data than the actual one.
- 41 Wrong number of delimiters in the sources (Files) causes DQ Problems.
- 42 Presence of Outliers.
- 43 Orphaned or dangling data (Data Pointing to other data which does not exists)[11]
- 44 Data and metadata mismatch.
- 45 Important entities, attributes and relationships are hidden and floating in the text fields [6][7].
- 46 Inconsistent use of special characters in various sources [6][7].
- 47 Multi-purpose fields present in data sources.
- 48 Deliberate Data entry errors (Input errors) contribute to data quality issues [4][11].
- 49 Poor data entry training causes data quality problems in data sources [11] [24].
- 50 Wrongly designed data entry forms, allowing illegibility [11] [24].
- 51 Different business rules of various data sources creates problem of data quality.
- 52 Insufficient plausibility (comparison within the data set and within time)checks in operative systems (e.g. during data input). [20]

Datavirheiden luokittelu - profilointi

- 1 Insufficient data profiling of data sources is
- responsible for data quality issues.
- 2 Manually derived information about the data
- Contents in operational systems propagates poor
- data quality [8].
- 3 Inappropriate selection of Automated profiling tool
- cause data quality issues [8].
- 4 Insufficient data content analysis against external
- reference data causes data quality problems.
- 5 Insufficient structural analysis of the data sources
- in the profiling stage.
- 6 Insufficient Pattern analysis for given fields within
- each data store.
- 7 Insufficient column profiling, single table structural
- profiling, cross table structural profiling of the data
- sources causes data quality problems [9].
- 8 Insufficient range and distribution of values or
- threshold analysis for required fields.
- 9 Lack of analysis of counts like record count, sum,
- mode, minimum, maximum percentiles, mean and
- standard deviation.
- 10 Undocumented, alterations identified during
- profiling cause data quality problems.
- 11 Inappropriate profiling of the formats,
- dependencies, and values of source data
- 12 Inappropriate parsing and standardization of
- records and fields to a common format
- 13 Lack of identification of missing data relationships
- 14 Hand coded data profiling is likely to be incomplete
- and leave the data quality problems.
- 15 Unreliable and incomplete metadata of the data
- sources cause data quality problems [8].
- 16 User Generated SQL queries for the data profiling
- purpose leave the data quality problems.
- 17 Inability of evaluation of inconsistent business
- processes during data profiling cause data quality
- problems.
- 18 Inability of evaluation of data structure, data values
- and data relationships before data integration,
- propagates poor data quality.
- 19 Inability of integration between data profiling, ETL
- cause no proper flow of metadata which leave data
- quality problems.

Datavirheiden luokittelu - ETL

- 1 Data warehouse architecture undertaken affects the data quality (Staging, Non Staging Architecture).
- 2 Type of staging area, relational or non relational affects the data quality.
- 3 Different business rules of various data sources creates problem of data quality.
- 4 Business rules lack currency contributes to data quality problems [4].
- 5 The inability to schedule extracts by time, interval, or event cause data quality problems.
- 6 Lack of capturing only changes in source files [24].
- 7 Lack of periodical refreshing of the integrated data storage (Data Staging area) cause data quality degradation.
- 8 Truncating the data staging area cause data quality problems because we can't get the data back to reconcile.
- 9 Disabling data integrity constraints in data staging tables cause wrong data and relationships to be extracted and hence cause data quality problems [11].
- 10 Purging of data from the Data warehouse cause data quality problems [24].
- 11 Hand coded ETL tools used for data warehousing lack in generating single logical meta data store, which leads to poor data quality.
- 12 Lack of centralized metadata repository leads to poor data quality.
- 13 Lack of reflection of rules established for data cleaning, into the metadata causes poor data quality.
- 14 Inappropriate logical data map prepared cause data quality issues.
- 15 Misinterpreting/Wrong implementation of the slowly changing dimensions (SCD) strategy in ETL phase causes massive data quality problems.
- 16 Inconsistent interpretation or usage of codes symbols and formats [4].
- 17 Improper extraction of data to the required fields causes data quality problems [4].
- 18 Lack of proper functioning of the extraction logic for each source system (historical and incremental loads) cause data quality problems.
- 19 Unhandled null values in ETL process cause data quality problems.
- 20 Lack of generation of data flow and data lineage documentation by the ETL process causes data quality problems.

Datavirheiden luokittelu - ETL

- 21 Lack of availability of automated unit testing facility in ETL tools cause data quality problems.
- 22 Lack of error reporting, validation, and metadata updates in ETL process cause data quality problems.
- 23 Inappropriate handling of rerun strategies during ETL causes data quality problems.
- 24 Inappropriate handling of audit columns such as created date, processed date and updated date in ETL
- 25 Inappropriate ETL process of update strategy (insert/update/delete) lead to data quality problems.
- 26 Type of load strategy opted (Bulk, batch load or simple load) cause Data Quality problems. [24]
- 27 Lack of considering business rules by the transformation logic cause data quality problems.
- 28 Non standardized naming conventions of the ETL processes (Jobs, sessions,Workflows) cause data quality problems.
- 29 Wrong impact analysis of change requests on ETL cause data quality problems.
- 30 Loss of data during the ETL process (rejected records) causes data quality problems. (refused data records in the ETL process)
- 31 Poor system conversions, migration, reengineering or consolidation contribute to the data quality problems [4] [24].
- 32 The inability to restart the ETL process from checkpoints without losing data [14]
- 33 Lack of Providing internal profiling or integration to third-party data profiling and cleansing tools.[14]
- 34 Lack of automatically generating rules for ETL tools to build mappings that detect and fix data defects[14]
- 35 Inability of integrating cleansing tasks into visual workflows and diagrams[14]
- 36 Inability of enabling profiling, cleansing and ETL tools to exchange data and meta data[14]

Datavirheiden luokittelu -mallinnus

- 1 Incomplete or wrong requirement analysis of the project lead to poor schema design which further casue data quality problems.
- 2 Lack of currency in business rules cause poor requirement analysis which leads to poor schema design and contribute to data quality problems.
- 3 Choice of dimensional modeling (STAR, SNOWFLAKE, FACT CONSTALLATION) schema contribute to data quality.
- 4 Late identification of slowly changing dimensions contribute to data quality problems.
- 5 Late arriving dimensions cause DQ Problems.
- 6 Multi valued dimensions cause DQ problems.
- 7 Improper selection of record granularity may lead to poor schema design and thereby affecting DQ.
- 8 Incomplete/Wrong identification of facts/dimensions, bridge tables or relationship tables or their individual relationships contribute to DQ problems.
- 9 Inability to support database schema refactoring cause data quality problems.
- 10 Lack of sufficient validation, and integr schema contribute to poor data quality.

Tiedon laatu tietovarastossa

- Tyhjät kentät
- Virheelliset arvot
- Eheysongelmat
- Duplikaatit
- Käsittelysäännöt
- Toteutusvirheet

Tapaus 1

- data luvut;
  x = 1 ;
  y = . ;
  run;

- data luvut_testi1;
  set luvut;
  z = x + y;
  run;

  data luvut_testi2;
  set luvut;
  z = sum (x, y);
  run;

Tapaus 2

- data luvut;
  x = 1 ;
  output;
  x = 2 ;
  output;
  run;

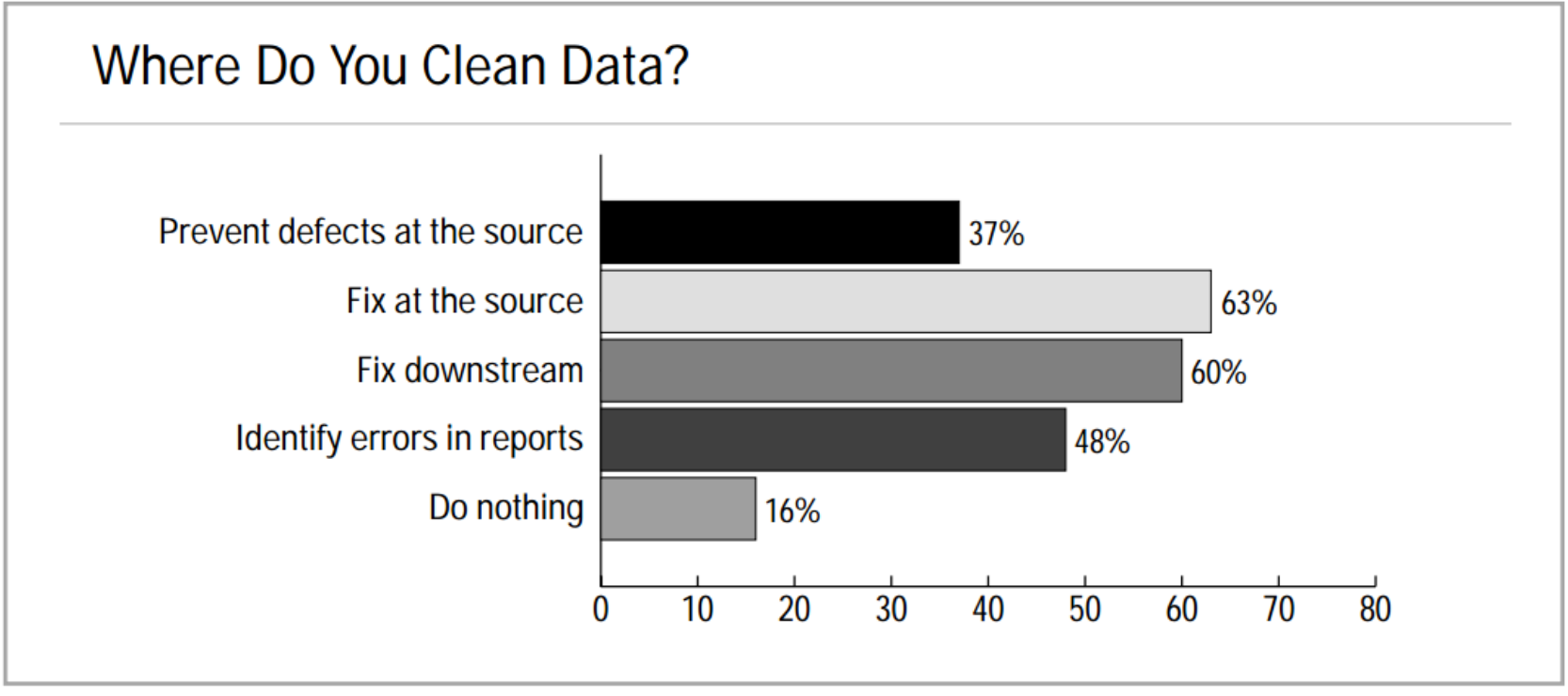
- data luvut_testi1;
  set luvut;
  if x ne 2;
  run;

  data luvut_testi2;
  set luvut;
  if x <> 2;
  run;

Dataongelmien käsittely

- Liittymätaso / rivitaso / kenttätaso
- Havaitseminen
- Puhdistus
- Korjaus
- Jälkipäivitys

Puhdistus käytännössä



Datan määrän vaikutus

- Ajoajat
- Selvitykset

Testausprosessi

- Yksikkötestaus
- Integraatiotestaus
- Tiedon validointi
- Loppukäyttäjätestaus
- Suorituskykytestaus
- Regressiotesaus
- Automatisointi

Tietovaraston lokitus

- Parametrit
- Ajot (vaiheet, virheet)
- Ajoajat
- Lokitustaso
- Work-taulut

Levytila

- Aktiivikäytössä oleva data
- Passiivikäytössä oleva data
- Poistettu data
- Liittymien talteenotto
- Tiedon palautukset
- Lasketut tiedot

DW:n suunnittelussa huomioitavaa

- Datan ymmärtäminen
- Toiminnallisuuden sijainti
- Muutosten ennakointi
- Kompleksisuuden hallinta
- Vaatimusmäärittelyt
- Dokumentointi
- Siirtyminen kehityksestä ylläpitoon

Inhimillinen tekijä

- Tiimi
- Prosessit

Kiitos!

Kysymykset & kommentit

AUREOLIS
BI for Better Information

Lähteet

- Wayne W. Eckerson:
  "DATA QUALITY AND THE BOTTOM LINE - Achieving Business Success through a Commitment to High Quality Data"
  http://download.101com.com/pub/tdwi/Files/DQReport.pdf
- Ranjit Singh, Kawaljeet Singh:
  "A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing"
  http://www.ijcsi.org/papers/7-3-2-41-50.pdf