NORTAL

How to use Big Data in Industry 4.0 implementations

LAURI ILISON, PhD

Head of Big Data and Machine Learning

Big Data definition?

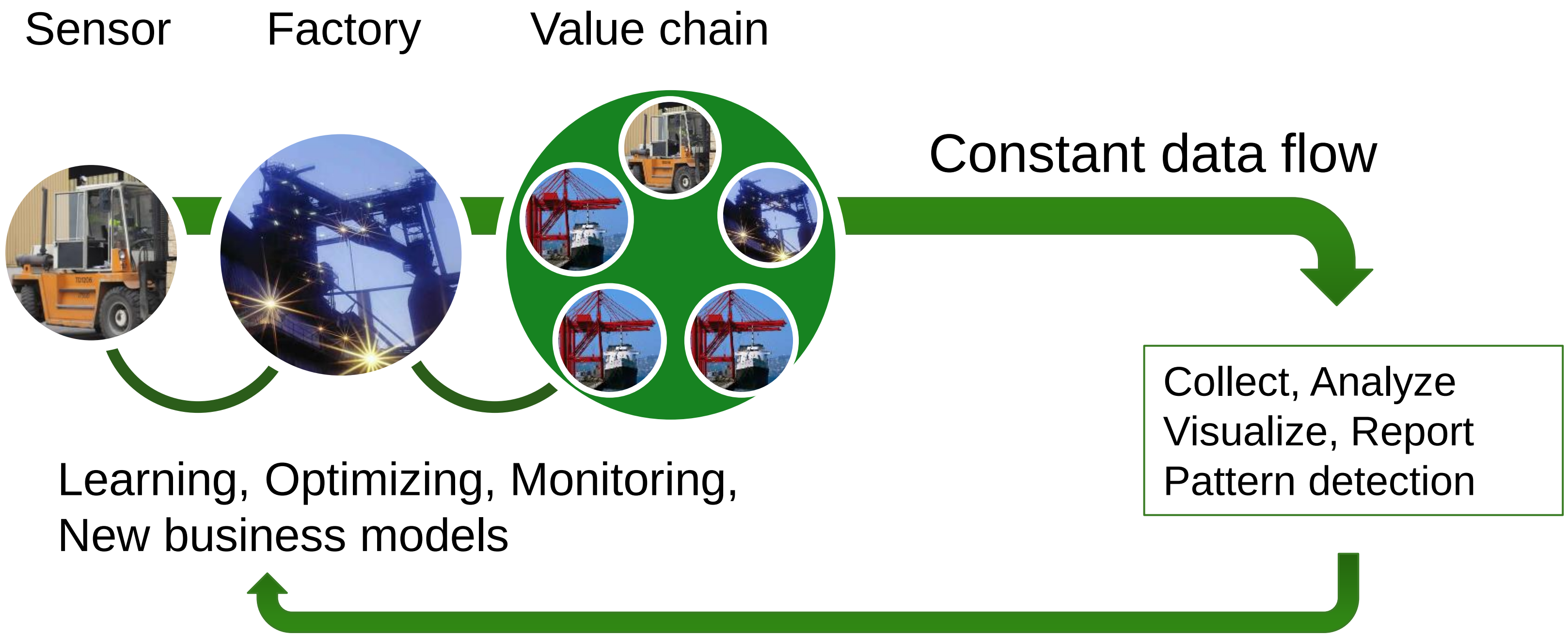
Big Data is about structured vs unstructured data

Big Data is about Volume of data in 100000 Terabytes

Volume
Variety
Velocity

No single and clear definition for BIG DATA

Big Data in Industry 4.0



Sensor

Factory

Value chain

Constant data flow

Collect, Analyze
Visualize, Report
Pattern detection

Learning, Optimizing, Monitoring,
New business models

Google™

In 2003, 2004, 2005 Google released three academic papers describing Google’s technology for massive data processing

1. Google File System (GFS)
Google storing all web content

2. Map-Reduce
Google calculating PageRank and web search index

3. BigTable
Google storing Crawling data Analytics, Earth and Personalized Search in columnar database

HADOOP

- In 2004/5 Doug Cutting developed Nutch open source web search engine struggling with huge data processing issues

- Doug implemented Google File System analog and named it HADOOP

- From 2006 HADOOP is an Apache Foundation project

HADOOP has been adopted!

2003

Google Whitepaper

2004 2005

Google file system reimplementation

2006

YAHOO!

INTERNET ARCHIVE

2007

Powerset

last.fm

2008

Google
ablegrape
ImageShack
Cascading
IBM
facebook
ENORMO
A9
The University of Edinburgh
krugle
rackspace HOSTING
Lookery Control freaks welcome
The New York Times
Joost
Zvents Discover Things To Do
Information Sciences Institute
News Corporation
Cornell University Computing and Information Science
Visible Measures
LOTAME
NetSeer
parc Palo Alto Research Center
DNS Security Enhanced Domain Name System
veoh

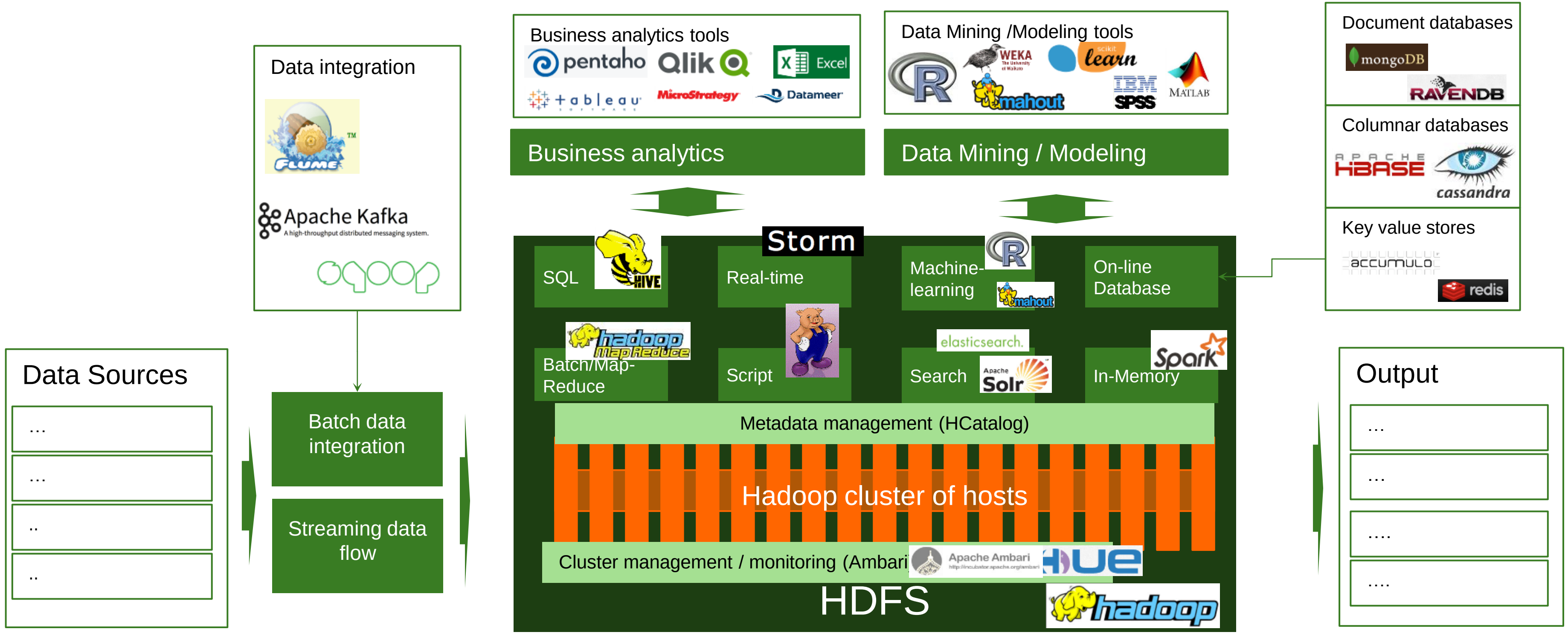
2009

AOL
amazon web services
deepdyve
cooliris
eyealike
TEXTMAP The entity search engine
PSG College of Technology
iterend
tailsweep
hulu
RapLeaf
USCMS
Ning
quantcast
cloudera
pressflip
detikSearch
WorldLingo
Systems@ETH Zürich
VK SOLUTIONS Global Solutions Provider
TARAGANA
HOSTING HABITAT
HOLA
Terrier
adknowledge
stampede beta

2010

SAMSUNG
rubicon
BERKELEY LAB
VISIBLE TECHNOLOGIES
APOLLO GROUP
ADSDAQ
rackspace HOSTING
RapLeaf
wordnik
MOBIGEN
comScore
trulia real estate search
Accela Communications
Forward3D
LinkedIn
Microsoft
Infochimps Find the world's data
Pharm2Phork
ADMELD
gumgum
Pronux The Datagraph Blog
NETFLIX
mobileanalytics.tv
markt24.de
twitter
media6degrees
BEEBLER
SLC Security
ebay

Big Data technical stack



Data Sources

…

…

..

..

Data integration

Batch data integration

Streaming data flow

Business analytics tools

Business analytics

Data Mining /Modeling tools

Data Mining / Modeling

SQL

Real-time

Machine-learning

On-line Database

Batch/Map-Reduce

Script

Search

In-Memory

Metadata management (HCatalog)

Hadoop cluster of hosts

Cluster management / monitoring (Ambari

HDFS

Document databases

Columnar databases

Key value stores

Output

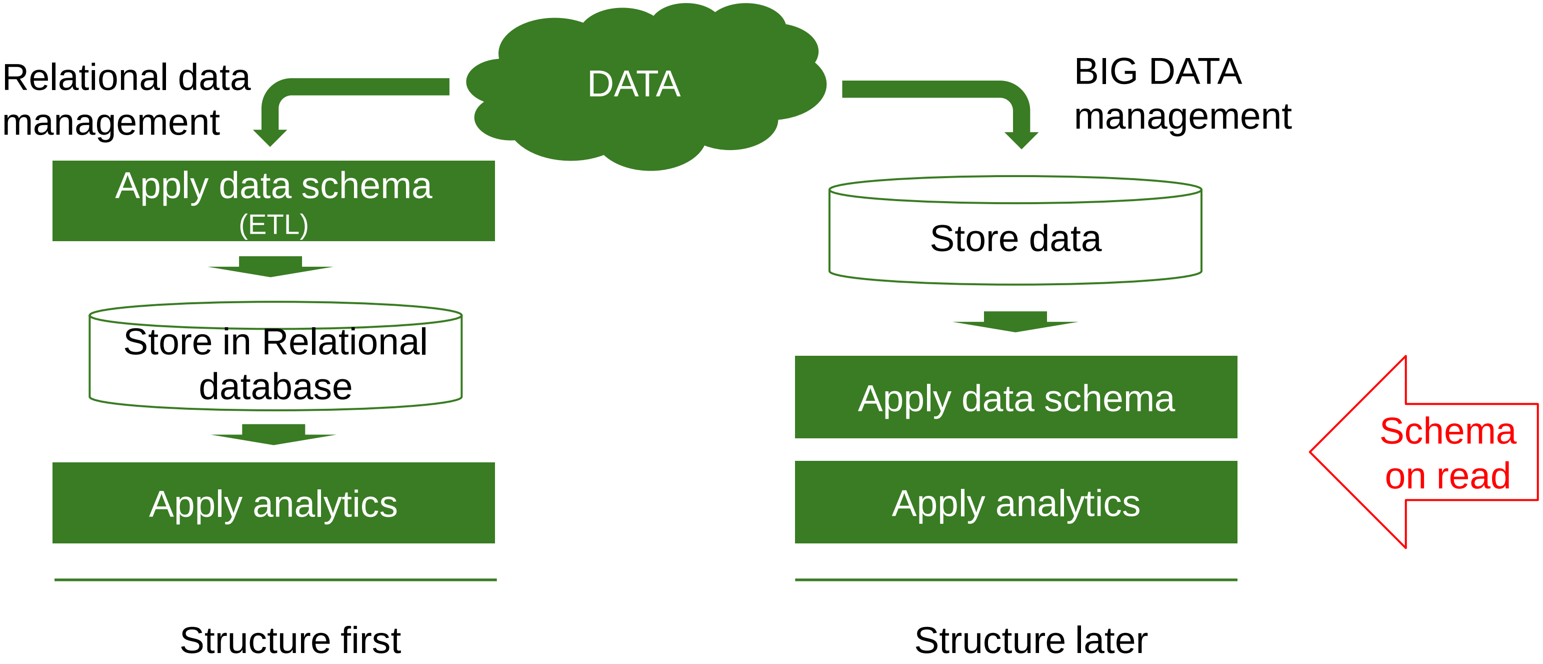
…

…

….

….

Relational Data vs BIG DATA



DATA

Relational data management

Apply data schema (ETL)

Store in Relational database

Apply analytics

Structure first

BIG DATA management

Store data

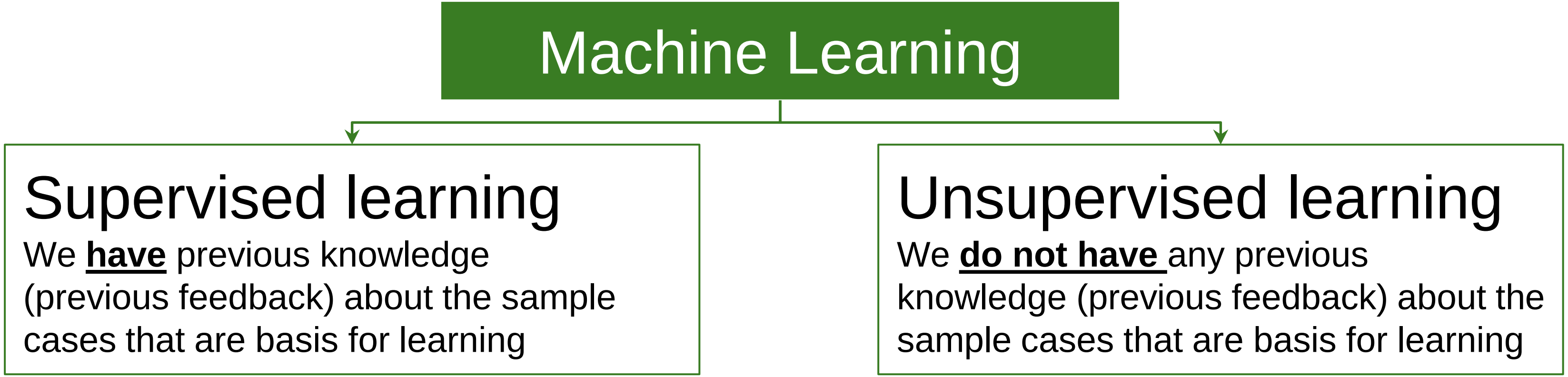
Apply data schema

Apply analytics

Schema on read

Structure later

How to find the patterns?



Machine Learning

Supervised learning

We **have** previous knowledge (previous feedback) about the sample cases that are basis for learning

Unsupervised learning

We **do not have** any previous knowledge (previous feedback) about the sample cases that are basis for learning

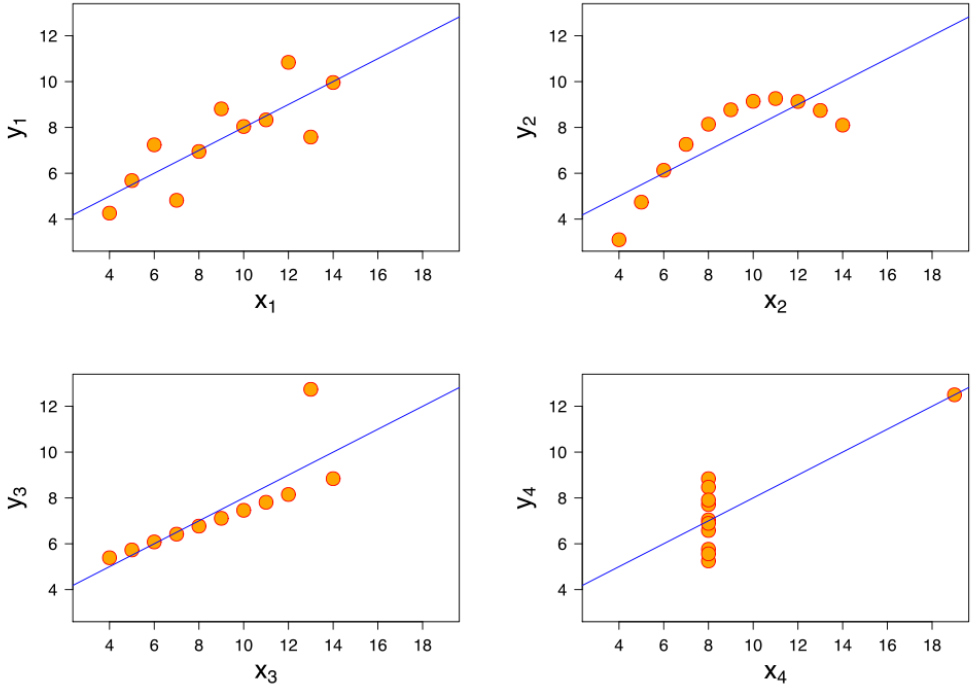
Algorithms

- Classification
- Regression
- Decision Trees
- Neural Networks
- Deep Learning

Algorithms

- Clustering
- Hidden Markov Chains
- Dimensionality reduction

Pattern recognition required!



Examples have same

- Mean x = 9
- Variation x = 11

- Mean y = 7.5
- Variance y = 4.12

- Correlation = 0.816
- Same linear regression

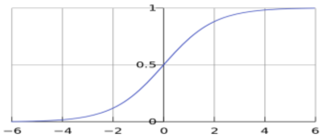
We need algorithms that look into the data

Example: Event failure scoring

TASK: Find the probability of event failure

1. Logistic function

$$f(x) = \frac{1}{1+e^{-x}}$$

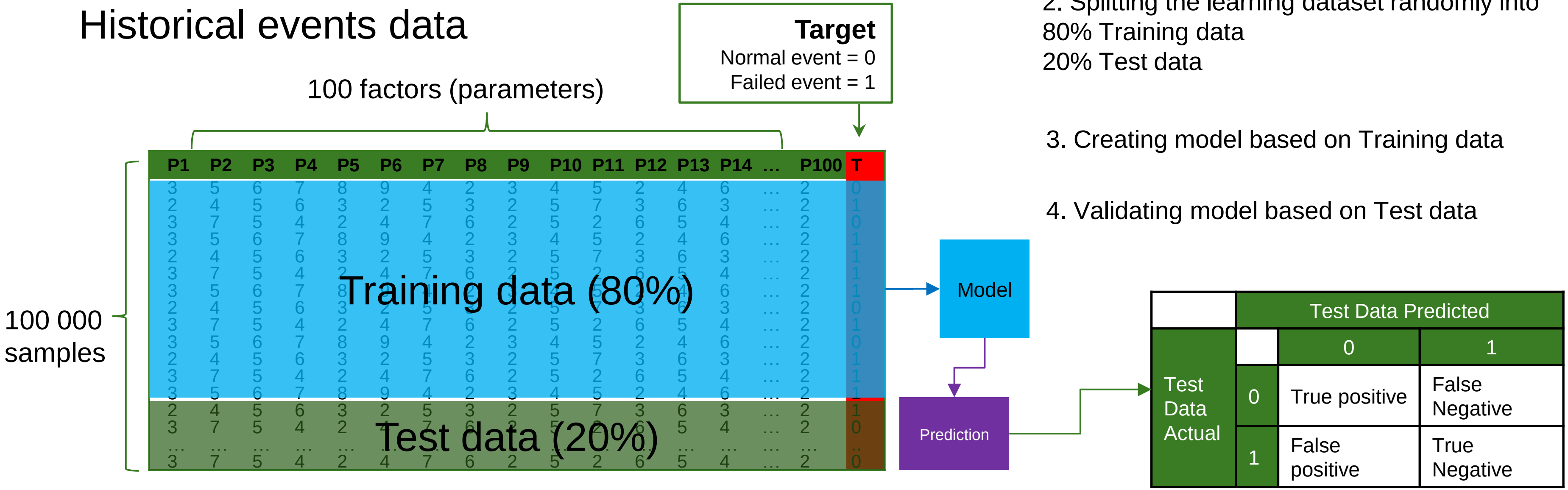


2. Splitting the learning dataset randomly into
80% Training data
20% Test data

3. Creating model based on Training data

4. Validating model based on Test data

Historical events data



Target
Normal event = 0
Failed event = 1

100 factors (parameters)

100 000 samples

Training data (80%)

P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	…	P100	T
3	5	6	7	8	9	4	2	3	4	5	2	4	6	…	2	0
2	4	5	6	3	2	5	3	2	5	7	3	6	3	…	2	1
3	7	5	4	2	4	7	6	2	5	2	6	5	4	…	2	0
3	5	6	7	8	9	4	2	3	4	5	2	4	6	…	2	1
2	4	5	6	3	2	5	3	2	5	7	3	6	3	…	2	1
3	7	5	4	2	4	7	6	2	5	2	6	5	4	…	2	1
3	5	6	7	8	9	4	2	3	4	5	2	4	6	…	2	1
2	4	5	6	3	2	5	3	2	5	7	3	6	3	…	2	0
3	7	5	4	2	4	7	6	2	5	2	6	5	4	…	2	1
3	5	6	7	8	9	4	2	3	4	5	2	4	6	…	2	0
2	4	5	6	3	2	5	3	2	5	7	3	6	3	…	2	1
3	7	5	4	2	4	7	6	2	5	2	6	5	4	…	2	1
3	5	6	7	8	9	4	2	3	4	5	2	4	6	…	2	1
2	4	5	6	3	2	5	3	2	5	7	3	6	3	…	2	1
3	7	5	4	2	4	7	6	2	5	2	6	5	4	…	2	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	..
3	7	5	4	2	4	7	6	2	5	2	6	5	4	…	2	0

Test data (20%)

		Test Data Predicted	
		0	1
Test Data Actual	0	True positive	False Negative
	1	False positive	True Negative

Example: Heavy industry manufacturer

Problem: Unhide the manufacturing information for products faults discovery and quality control

About the case

- Sophisticated manufacturing processes
- Data is generated in all steps by machines
- Data usage for quality and error discovery
- Historical data should be used for detecting errors and failures

Nortal Solution:

- Streaming data collection, analytics
- Historical data access for trend and pattern discovery

Business impact:

- Improved manufacturing quality as data is fully used
- Predictive maintenance will decrease production line stops

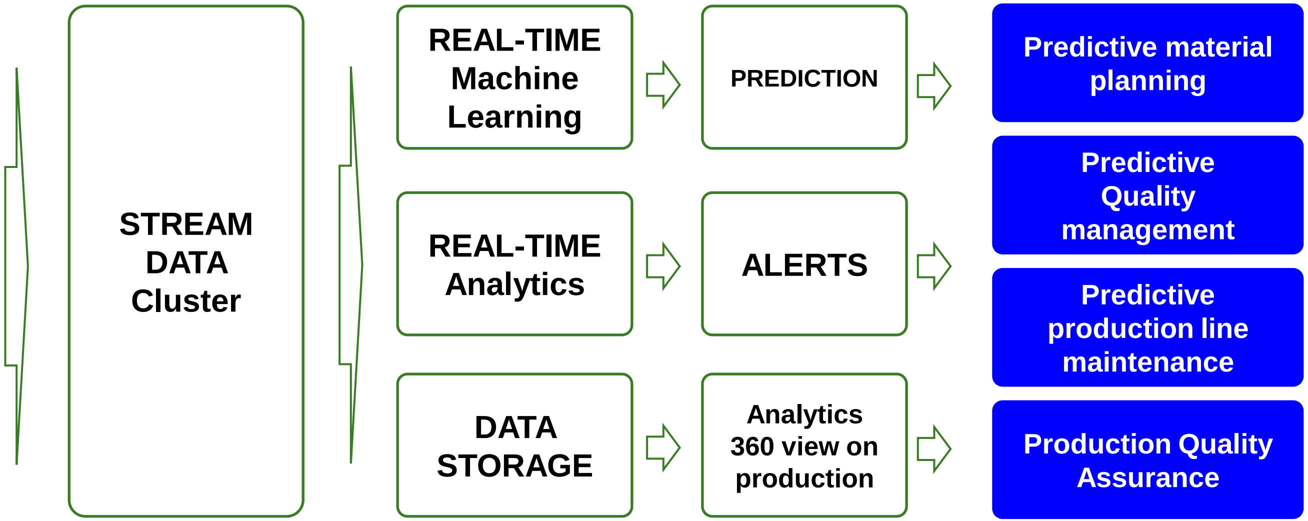
Setup architecture

Speed capacity:
15000 events per sec

INPUT

BIG DATA CLUSTER

OUTPUT



STREAM DATA Cluster

REAL-TIME Machine Learning

REAL-TIME Analytics

DATA STORAGE

PREDICTION

ALERTS

Analytics 360 view on production

Predictive material planning

Predictive Quality management

Predictive production line maintenance

Production Quality Assurance

Example: Telecom

Problem: Improve data warehouses performance and capabilities to store and analyze all telecom data

About the case

- Telecom has existing data warehouse that consists only part of the original data
- Costs are linearly increasing (1TB = 20’000 EUR)

Nortal solution:

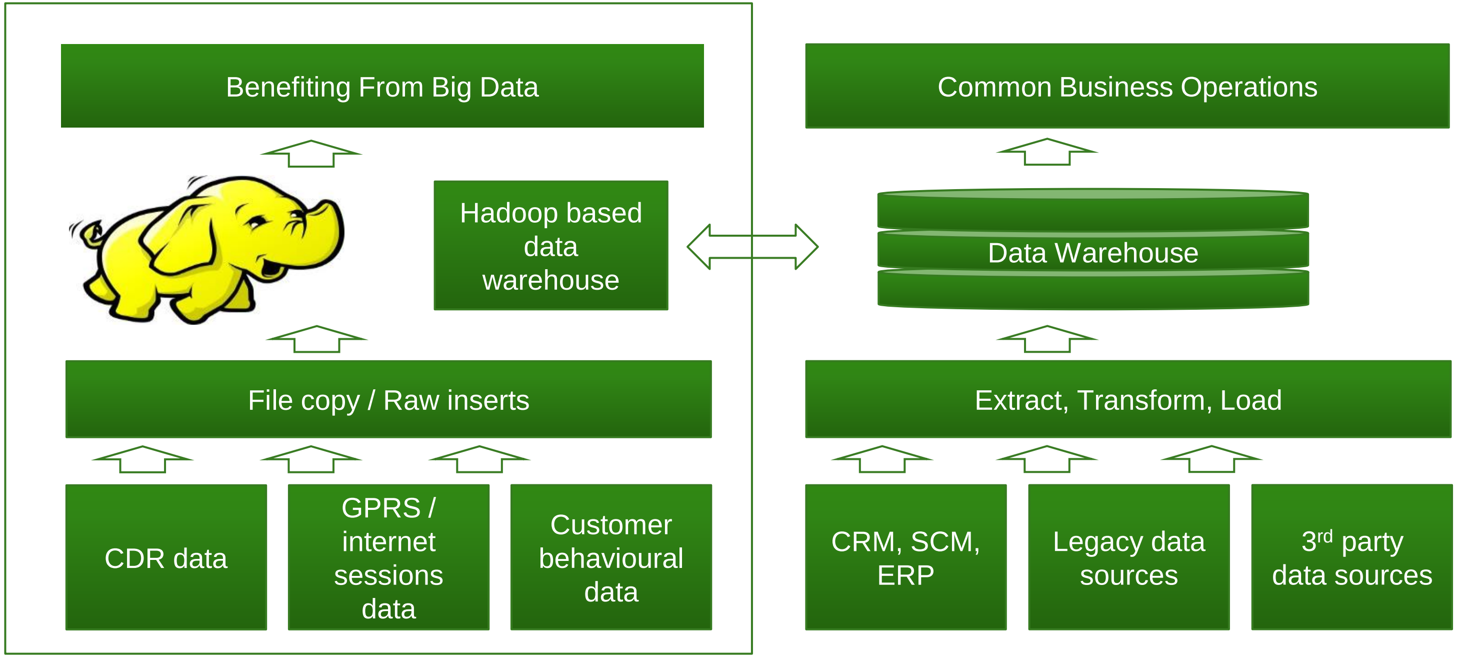
- Big Data technology based data warehouse
- All historical raw data could be stored and accessed

Business impact:

- 95% Cost improvement per TB of data (1TB = 1’000 EUR)
- All raw and historical data accessible
- Complex data analytics available

Offloading existing Data Warehouse

Big Data Warehouse and Traditional Data Warehouse working back-to-back



Benefiting From Big Data

Hadoop based data warehouse

File copy / Raw inserts

CDR data

GPRS / internet sessions data

Customer behavioural data

Common Business Operations

Data Warehouse

Extract, Transform, Load

CRM, SCM, ERP

Legacy data sources

3rd party data sources

Big Data values for Industry 4.0

Efficient technology

- Decreased costs for IT
- Linear cost increase
- Small scale POC-s available
- Commodity hardware
- Fast time to market

Data capture and analysis

- Data stored in one place
- Data fully accessible
- Data fully analyzed
- Pattern detection on all data

New opportunities

- Improved manufacturing processes
- Improved customer services and experience
- Cost optimization
- New services based on data

Thank you!

Nortal Big Data and Machine Learning Solutions

Lauri Ilison, PhD
Head of Big Data and Machine Learning
Tel: + 372 5111 003
Email: lauri.ilison@nortal.com