DataBench

Evidence Based Big Data Benchmarking to Improve Business Performance

Virtual BenchLearning
Success Stories on Big Data & Analytics

Chiara Francalanci
Politecnico di Milano
May 28, 2020

IDC
ANALYZE THE FUTURE

Atos

Institut "Jožef Stefan", Ljubljana, Slovenija

SINTEF

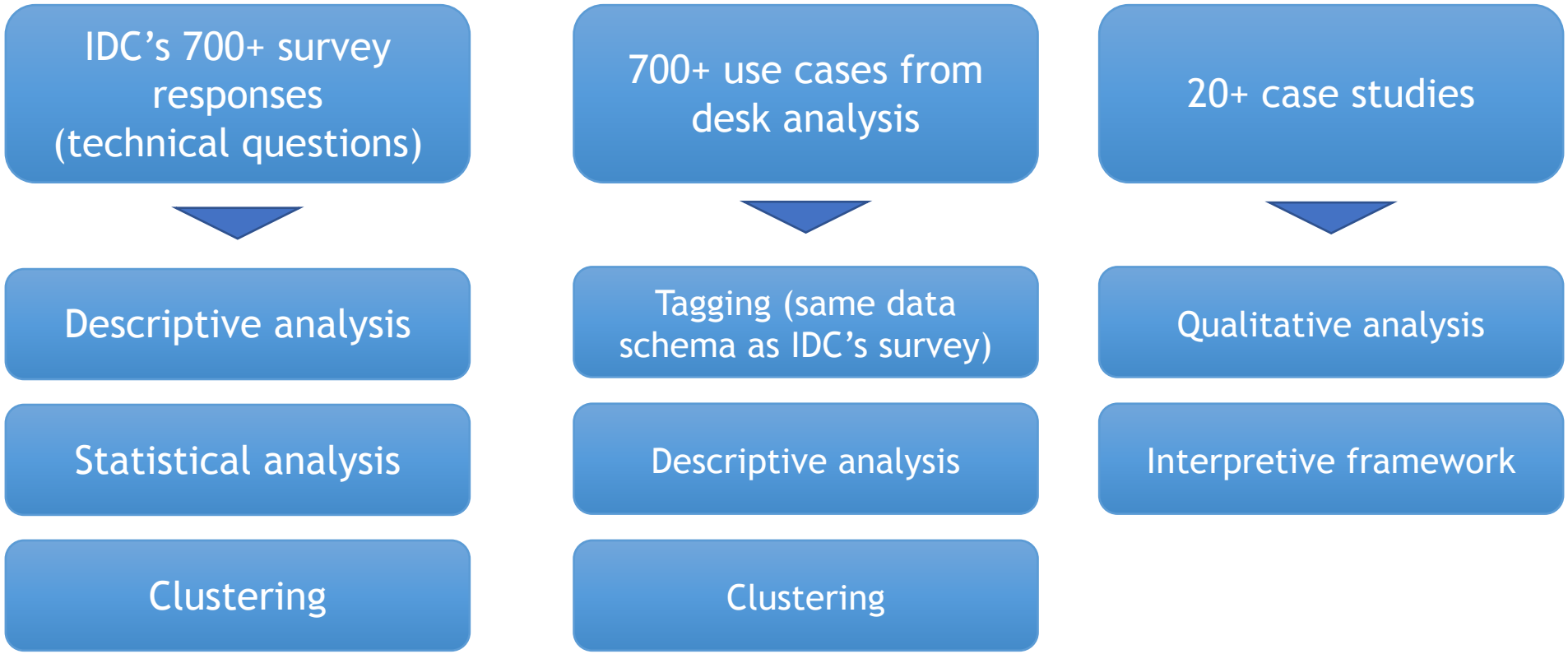
POLITECNICO
MILANO 1863

Frankfurt Big Data Lab
GOETHE UNIVERSITÄT

Research objectives

- The main objective of our research in Databench is to evaluate the impact of big data technologies (BDTs) on business performance in key use cases.
- A fundamental output of our work is the industrial and business performance benchmarks of the use of advanced BDTs in representative use cases.
- We are also providing insights on how technical benchmarking can help to make informed decisions and maximize benefits, as a input to the design of the Databench Toolbox and as an aid to the definition of an interpretative framework of the relationship between technical and business performance.
- We are writing a handbook describing the main industrial and business performance benchmarks targeted at industrial users and European technology developers.

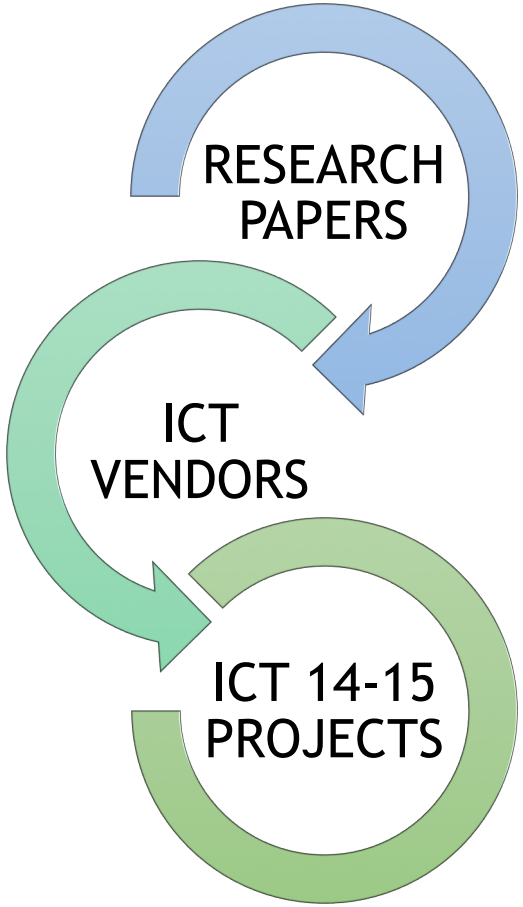
Summary of data collection and analysis



Insights from the survey – descriptive analytics

- Companies mainly analyze and store **gigabytes and terabytes** of data, while a small number of companies (less than 10%) deal with petabytes and exabytes.
- **Tables and structured data** seem to play a prominent role, followed by structured-text and graph data.
- Currently, **descriptive and diagnostic** analytics are the most popular types of analytics among European companies.
- The **batch processing approach is most common**, and only 16% of companies are pioneering the management and exploitation of real-time data.
- In the future, companies are planning to move to prescriptive and predictive analytics.
- These results highlight the emerging need to integrate heterogenous data to effectively exploit all the information gathered by companies.
- The most adopted technical performance metric is **data quality**.

Desk analysis contribution



- ✓ 703 use cases in total
- ✓ 58 use cases per industry on average

Main sources of information: scientific literature, Web sites/white papers from IT providers, public documentation from ICT 14-15 projects.

Comparing data from the survey with data from the desk analysis provides **mainstream vs. innovation insights**.

A quali-quantitative analysis (tagging) can be found at:

http://131.175.15.19/databench/desk-analysis_17_2_2020.xlsx

Insights from the desk analysis

- Use cases from the desk analysis mainly deal with **terabytes** of data.
- Most use cases are mainly processing data in **streaming**, as well as iterative/in-memory processing.
- The most widely used analytics type is by far **predictive** analytics, while **prescriptive**, **descriptive** and **diagnostic** analytics are adopted in approximately the 30% of use cases.
- The most widely adopted performance metric seems to be the **throughput**.
- Data types are primarily tables and structured data, including structured legacy data, graph and linked data and text and semi-structured data.
- Use cases store and process highly heterogenous data, thus stressing the growing need and potential for data integration.

Most common use-cases by industry

Agriculture	Crops monitoring	Equipment optimization	Precision agriculture
Automotive	Predictive maintenance	Self driving	Smart services
Financial Services	Fraud detection	Risk assessment	Targeting
Healthcare	Diagnostic	Patient monitoring	Preventive systems
Manufacturing	Predictive maintenance	Smart manufacturing	R&D optimization/ Smart design
Retail	Assortment optimization/ Intelligent fulfilment	Price optimization/ Promotions	Targeting
Telecommunication	Churn prediction/ Promotions	Network capacity optimization	Targeting
Transport & logistics	Churn prediction/ Promotions	Fleet management	Network capacity optimization
Utilities	Churn prediction/ Promotions	Network capacity optimization	Personalized fares

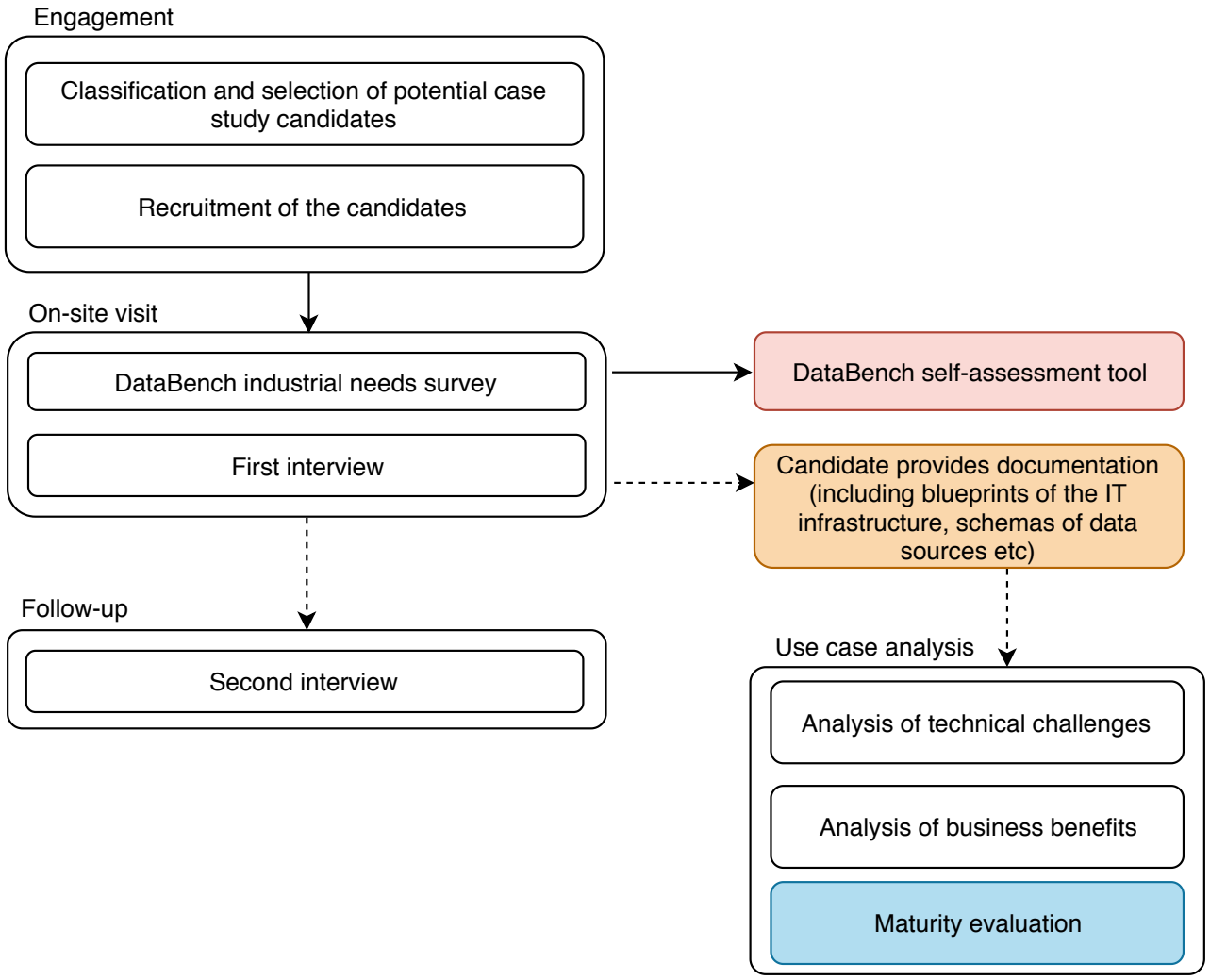
Most frequent Business KPI by use-case (1)

Agriculture	*Crops monitoring*: Costs = -10%	*Equipment optimization*	*Precision agriculture*
Automotive	*Predictive maintenance*	*Self driving*	*Smart services*: Costs = -80%
Financial Services	*Fraud detection*: Operational Ex. = -80%	*Risk assessment*	*Targeting*: Marketing costs = -35% TCO costs = -80% Conversion rate = 10x
Healthcare	*Diagnostic*	*Patient monitoring*	*Preventive systems*
Manufacturing	*Predictive maintenance*: Maintenance costs = -30%	*Smart manufacturing*: Utilities costs = -20% Cust. retention = +110%	*R&D optimization/* *Smart design*

Most frequent Business KPI by use-case (2)

Retail	*Assortment optimization/ Intelligent fulfilment*	*Price optimization/ Promotions*: Conversion rate = 50% Cust. retention = +14%	*Targeting*: Conversion rate = +85% TCO costs = -15%
Telecommunication	*Churn prediction/ Promotions*	*Network capacity optimization*	*Targeting*: Conversion rate = +130%
Transport & logistics	*Churn prediction/ Promotions*	*Fleet management*	*Network capacity optimization*: TCO costs = -90%
Utilities	*Churn prediction/ Promotions*	*Network capacity optimization*: Costs = -20% Cust. Expenses = -30%	Personalized fares: Marketing costs = -50% TCO costs = -50%

Case study analysis methodology



The methodology was tested in the Whirlpool pilot case study that allowed us to improve the interview template to serve important areas of in-depth analysis:

- Analysis of technical challenges
- Analysis of business benefits

As a consequence of the pilot, we introduced a second interview to collect additional information and/or involve other company profiles.

Interview template

Date / interviewer(s) / interviewee(s)	
Company description	
Case study description	
Data characteristics	Volume, velocity, variety and variability
Data sharing and exchange platform use	
Data anonymization and privacy needs	
Data processing and analytics characteristics	Volatility, veracity, monetary value, visualization, storage, processing, analytics and machine learning/AI
Big Data specific challenges	Short term, long term
Technical benchmark adoption	Current, short term, long term
Relevant technical performance metrics	
Expected benefits (including business KPIs)	Current (measured), short term, long term

General insights (1/2)

- We have evidence of business KPIs for a subset of case studies, evidence is aligned with results from the survey (business impact is in the 4-8% range).
- From the evidence that has been collected so far, an important lesson learnt is that most companies believe that **technical benchmarking requires highly specialized skills and a considerable investment**. We have found that very few companies have performed an accurate and extensive benchmarking initiative. In this respect, using cloud solutions grants them with an easier access to a broader set of technologies that they can experiment with.
- On the other hand, they acknowledge the variety and complexity of technical solutions for big data and envision the following **risks**:
  - The risk of realizing that they have chosen a technology that proves non scalable over time, either technically or economically.
  - The risk of relying on cloud technologies that might create a lock in and require a considerable redesign of software to be migrated to other cloud technologies.
  - The risk of discovering that cloud services are expensive, especially as a consequence of scalability, and that technology costs are higher than business benefits (edge vs. cloud decisions).

General insights (2/2)

- From a technical benchmarking perspective, it is important that benchmarking is supported with tools that **reduce complexity** by guiding users along predefined user journeys towards the identification and execution of benchmarks.
- **Results from previous benchmarking** initiatives are also very useful.
- It is important to have cost estimates of individual technologies and **end-to-end solutions**, on premises and in cloud to support edge vs. cloud solutions.
- There exists a growing number of **tools** designed for a specific use case (e.g. recommendation systems, markdown optimization systems, sensor data normalization, etc.), which **represent end-to-end off-the-shelf solutions** that are typically outside of the scope of benchmarking.

Business KPIs for case studies

- Intelligent fulfilment: +5% margin
- Recommendation systems: +3/4% margin
- Markdown: -20% promotional investment, +7% revenues, +5% margin
- Yield prediction in agriculture: +10% precision in yield prediction corresponding to +0.3% profit increase from trading
- Rail transport quality of service: +10% logistic efficiency
- Manufacturing production quality: +5% reduction of quality issues
- New business models in manufacturing: cloud-based prediction maintenance service

Yield prediction, Agriculture

- A company specialized on earth observation services has designed an innovative yield prediction machine learning algorithm based on Sentinel and Landsat high-resolution satellite information.
- The approach was tested with reference to the needs of a company operating in the financial industry, providing predictions as a support to trading decisions.
- Machine learning algorithms have demonstrated roughly 10% more accurate, supporting better investment decisions.
- The focus was the production of soy beans and corn in the US.
- The export market of soy beans and corn is aroung 21 billion dollar.
- The delta between projected and actual price was around 3.2%. This figure has been reduced by 10%, with a corresponding gain from trading around 0.32% on traded volumes.

Intelligent fulfilment, Retail

- Retail (grocery) industry
- Based on the idea of using machine learning to **optimize** assortment selection and automated fulfilment **at an individual shop level**
- Complex AI system including machine learning (sales prediction)
- Piloted in one shop: 5% increase of margins (equivalent to roughly 5 million/year)
- Run on Spark in cloud on Amazon, 100 euro per run, per category, per shop
- There are hundreds of categories, shops, and runs …
- Full deployment of project is currently on hold due to the economic scalability of IT.

Online recommendation system

- Retail (grocery) industry
- Based on the idea of **personalizing recommendations at an individual customer level**
- System designed to increase margins with up-sell and cross-sell recommendations
- Measured 3-4% impact on margin
- Cross-sell tables too large to be deployed on a hardware appliance on premises have been simplified to the client segment level
- Less personalized recommendations have been found to generate 1/10th of the clicks

A dangerous shift towards mass market

- Economic scalability issues cause business superficiality
- In turn, business superficiality drives a shift towards mass market
- For example, off-the-shelf recommendation systems suggest products that are «most frequently purchased», that is mass market, shifting customer behaviour towards purchasing choices that are less profitable for the company (the opposite of up-selling)
- Simplifying recommendations by adopting an off-the-shelf recommendation system is likely to result in a loss of competitiveness

Benchmarking can help reduce costs

Examples:

- Choice of the top performing DB can reduce costs by an order of magnitude
- AI benchmarking
- Choice of the most convenient machine in cloud

Choice of optimal cloud instances

Comparison between the cheapest optimal machine and the most expensive one for two different use-cases.

- Intelligent fulfilment

AWS instance	vCPU	Memory	# instances	Price	EMR Price	Total cost
c5.9xlarge	36	72	2	$ 1.728	$ 0.270	$ 1 166 832.00
r5.xlarge	4	32	3	$ 0.252	$ 0.063	$ 275 940.00

- 76 %

- Targeting

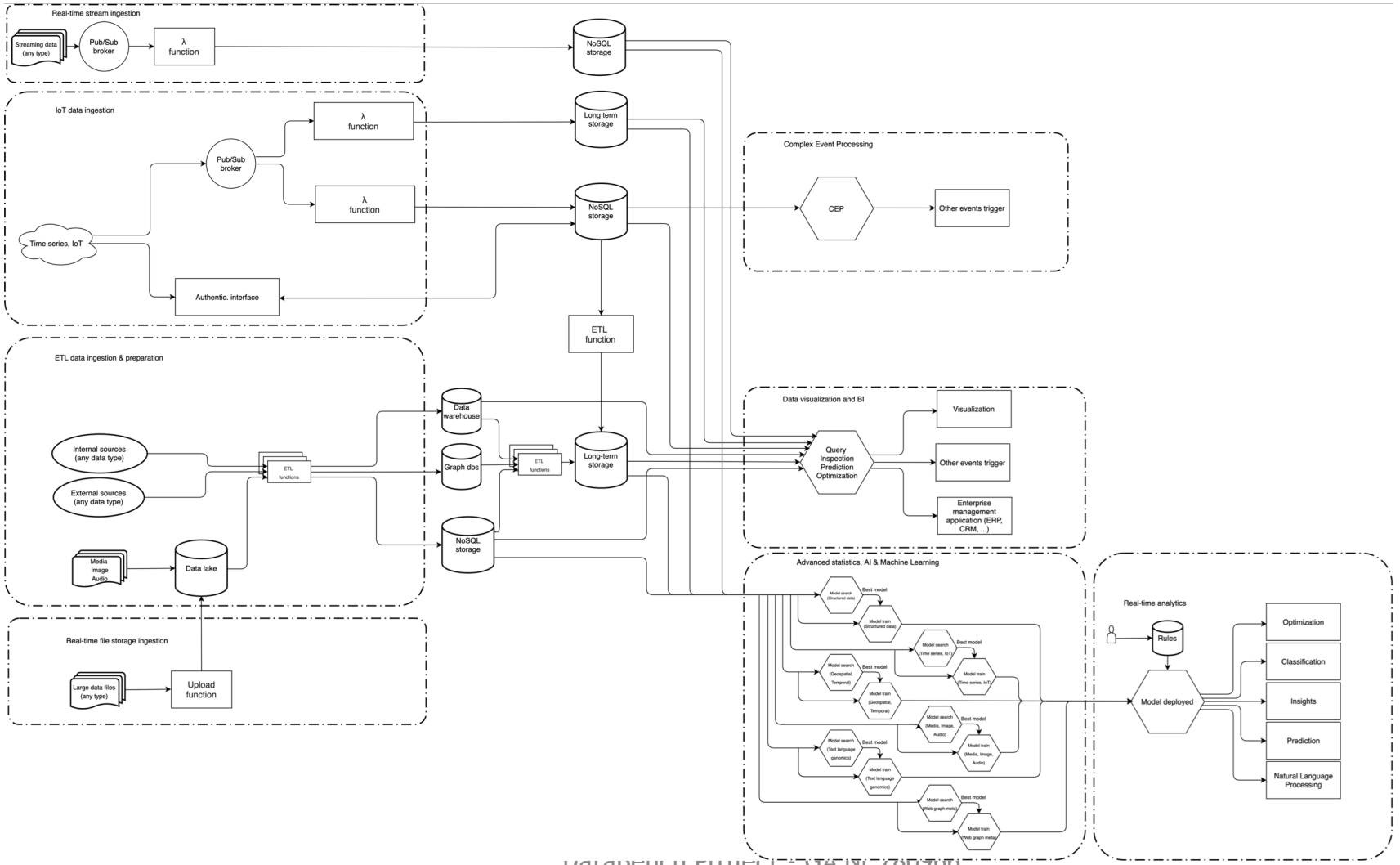
AWS instance	vCPU	Memory	# instances	Price	EMR Price	Total cost
c5.4xlarge	16	32	45	$ 0.768	$ 0.170	$ 52 678.08
r5.24xlarge	96	768	2	$ 6.048	$ 0.270	$ 15 769.73

- 70 %

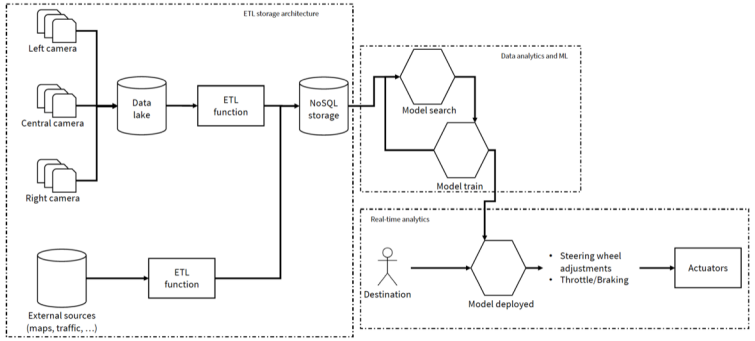
Is the cost issue general?

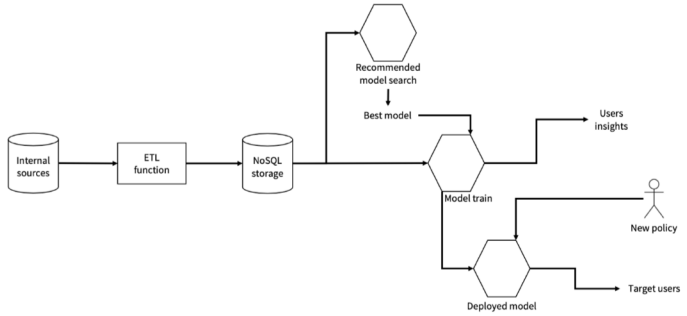
- Framing big data architecture by defining building blocks
- Design a per-use case architecture by combining building blocks
- Estimating data size and computation time
- Estimating cost in cloud
- We can tell where the cost issue is and what component should be benchmarked
  - Prediction model/AI algorithm (e.g. intelligent fulfilment)
  - Data preparation/query (e.g. targeting in telephone companies)
  - Data management/database massive insert (IoT)
  - Data storage (e.g. Satellite data)

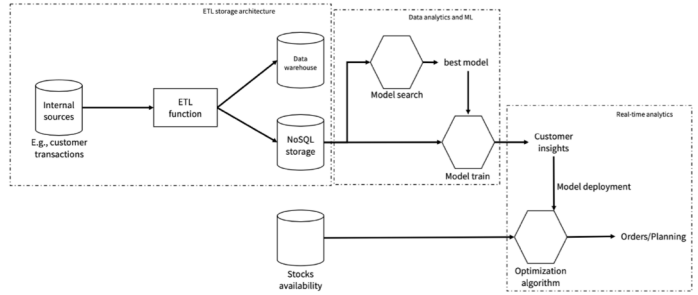
Integrated Architectural blueprint…

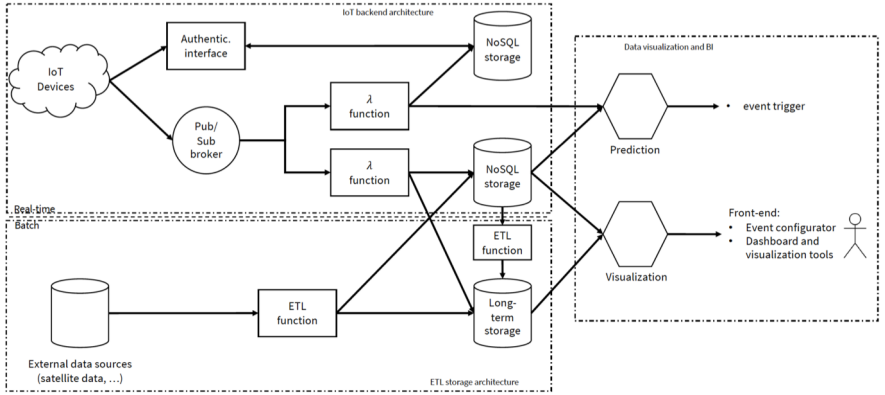


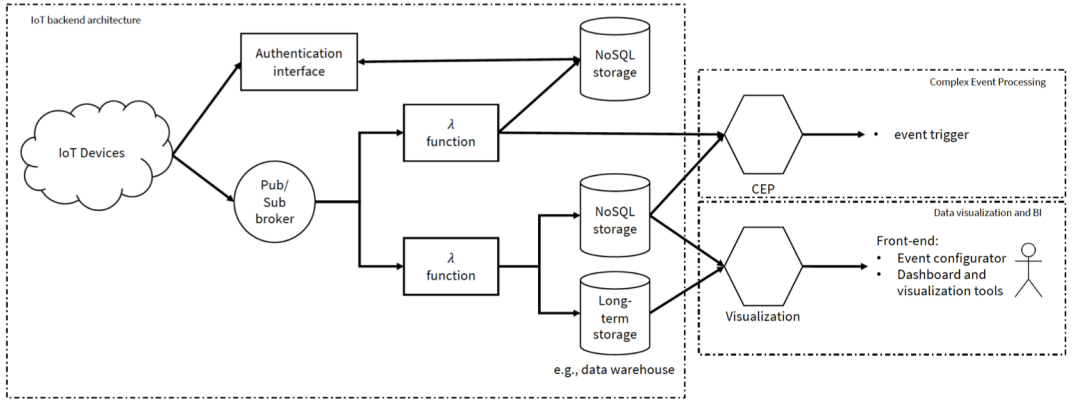
…as a unified view of 27 use-case specific blueprints



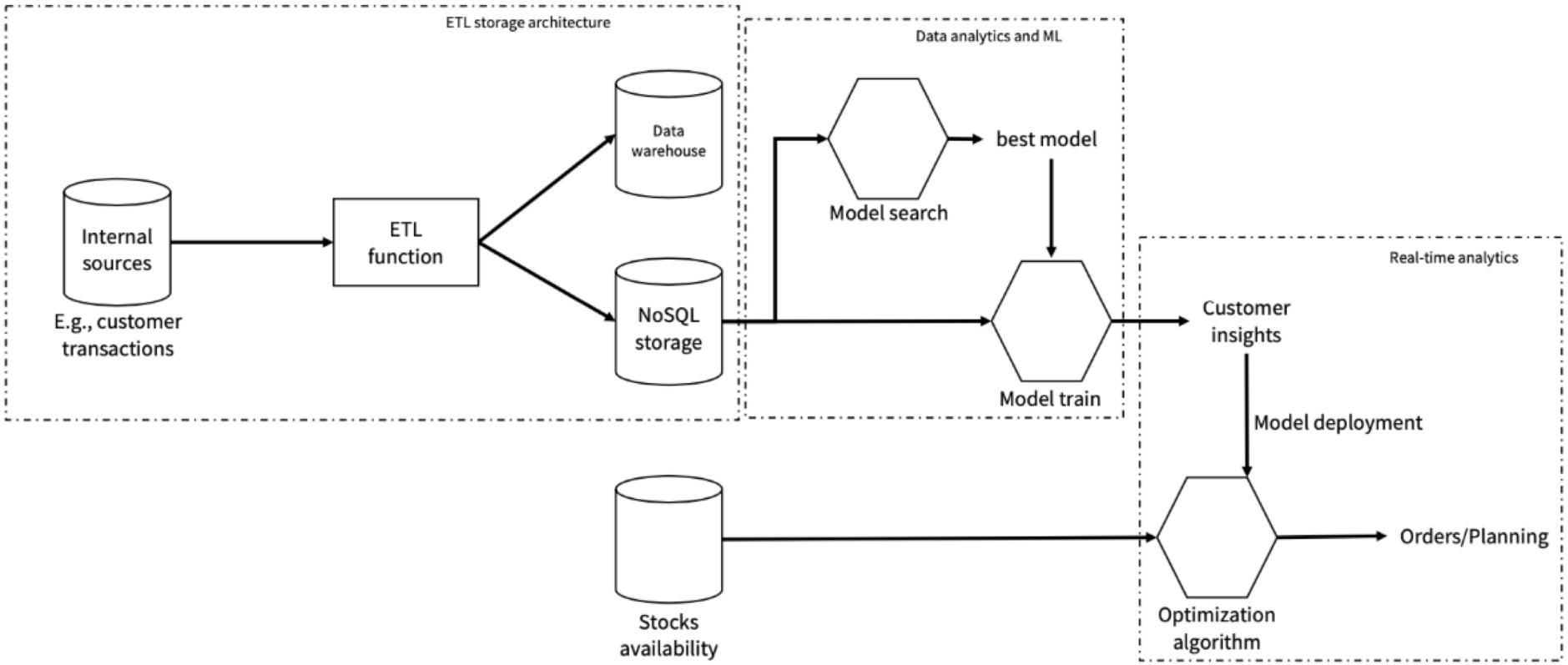








Example: Intelligent fulfilment blueprint



Intelligent fulfilment: sizing

- Receipts per day: 200’000
- Average products per receipt: 7
- This results in 511’000’000 products sold per year in all shops
- Which means approximately 65 GB/year
- Suppose using 5 years of receipts to get better results, we would need at least 325 GB of memory

Intelligent fulfilment: simulation

- *Four* periods considered: weekdays with and without promotions, weekends with and without promotions.
- Elapsed time for computing the optimal reorder point for each period: around *60 minutes*.
- Executions on Amazon AWS, prices refer to Amazon Ireland data centre.
- The following tables express the annual cost, by using respectively the last one year and last five years datasets.

Intelligent fulfilment: costs (1 year)

AWS instance	vCPU	Memory	# instances	Price	EMR Price	Total cost	x 200 shops
c5.4xlarge	16	32	3	$ 0.768	$ 0.170	$ 4 108.44	$ 821 688.00
c5.9xlarge	36	72	2	$ 1.728	$ 0.270	$ 5 834.16	$ 1 166 832.00
c5.18xlarge	72	144	1	$ 3.456	$ 0.270	$ 5 439.96	$ 1 087 992.00
m5.2xlarge	8	32	3	$ 0.428	$ 0.096	$ 2 295.12	$ 459 024.00
m5.4xlarge	16	64	2	$ 0.856	$ 0.192	$ 3 060.16	$ 612 032.00
m5.12xlarge	48	192	1	$ 2.568	$ 0.270	$ 4 143.48	$ 828 696.00
r5.xlarge	4	32	3	$ 0.252	$ 0.063	$ 1 379.70	$ 275 940.00
r5.2xlarge	8	64	2	$ 0.504	$ 0.126	$ 1 839.60	$ 367 920.00
r5.4xlarge	16	128	1	$ 1.008	$ 0.252	$ 1 839.60	$ 367 920.00

- Cheapest combination of instances: $ 275 940.00
- Most expensive combination of instances: $ 1 166 832.00

→ **x 4.2**

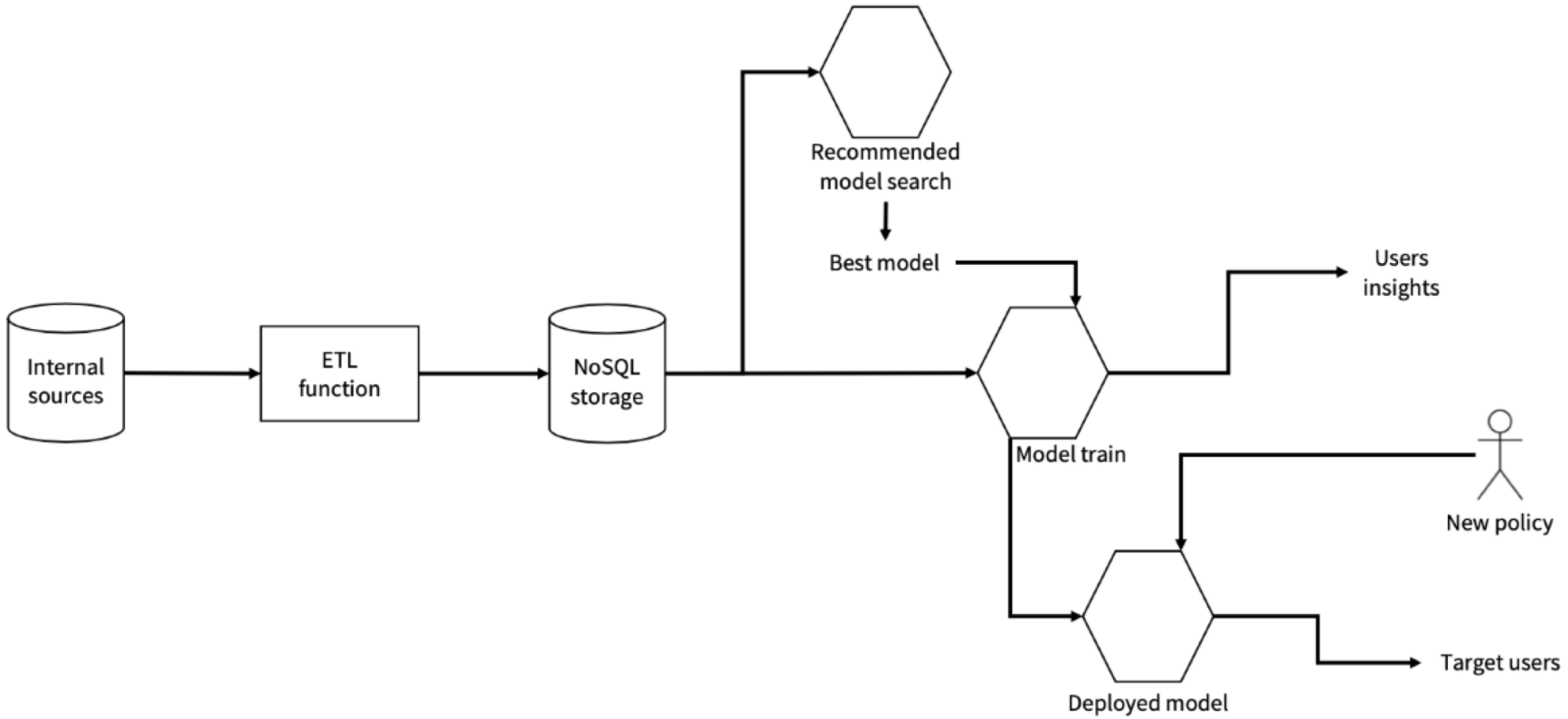
Intelligent fulfilment: costs (5 years)

AWS instance	vCPU	Memory	# instances	Price	EMR Price	Total cost	x 200 shops
c5.4xlarge	16	32	15	$ 0.768	$ 0.170	$ 20 542.20	$ 4 108 440.00
c5.9xlarge	36	72	7	$ 1.728	$ 0.270	$ 20 419.56	$ 4 083 912.00
c5.18xlarge	72	144	4	$ 3.456	$ 0.270	$ 21 759.84	$ 4 351 968.00
m5.2xlarge	8	32	15	$ 0.428	$ 0.096	$ 11 475.60	$ 2 295 120.00
m5.4xlarge	16	64	8	$ 0.856	$ 0.192	$ 12 240.64	$ 2 448 128.00
m5.24xlarge	96	384	2	$ 5.136	$ 0.270	$ 15 785.52	$ 3 157 104.00
r5.xlarge	4	32	15	$ 0.252	$ 0.063	$ 6 898.50	$ 1 379 700.00
r5.2xlarge	8	64	8	$ 0.504	$ 0.126	$ 7 358.40	$ 1 471 680.00
r5.16xlarge	64	512	1	$ 4.032	$ 0.270	$ 6 280.92	$ 1 256 184.00

- Cheapest instances combination: $ 4 351 968.00
- Most expensive instances combination: $ 1 256 184.00

→ **x 3.5**

Targeting in telecom industry: schema



Targeting: results

- Targeting advantages are:
  - Higher campaigns conversion rate
    - Up-selling strategies to targeted customers
  - Lower customer fatigue
- Targeting reduce the number of calls to customers, hence reducing the number of call centre operators from 380 to 10.
  - It results in cost reduction by over 90%. (From 15M€ to 400K€ yearly)
  - This cost reduction enables new marketing models with higher redemption rate.

Targeting: sizing

- Number of customers: 5’000’000 (Big companies: 30M)
- Average SMS/MMS per day per customer: 10
- Average Calls per day per customer: 7

- Weekly processing of data to produce aggregated profiles
- 24 hours of processing: 12h for data preparation, 12h for targets

Targeting: costs

AWS instance	vCPU	Memory	# instances	Price	EMR Price	Total cost
c5.4xlarge	16	32	45	$ 0.768	$ 0.170	$ 52 678.08
c5.18xlarge	72	144	10	$ 3.456	$ 0.270	$ 46 500.48
m5.2xlarge	8	32	45	$ 0.428	$ 0.096	$ 29 427.84
m5.24xlarge	96	384	4	$ 5.136	$ 0.270	$ 26 986.75
r5.xlarge	4	32	45	$ 0.252	$ 0.063	$ 17 690.40
r5.8xlarge	32	256	6	$ 2.016	$ 0.270	$ 17 117.57
r5.24xlarge	96	768	2	$ 6.048	$ 0.270	$ 15 769.73

- Cheapest instances combination: $ 15 760.73
- Most expensive instances combination: $ 52 678.08

→ **x 3.3**

On-going work

- Mapping 300 market technologies on blueprint components
- Mapping 60 benchmarks on market technologies
- Organize Databench knowledge around this knowledge base (in cooperation with Sintef)

Contacts

chiara.francalanci@polimi.it

paolo.ravanelli@polimi.it

gianmarco.ruggiero@polimi.it

giulio.costa@polimi.it