Sustainable Data Base

Program DATA-INFO-KNOWLEDGE, Talk 3

Prepared by: Katerina Zalamova, CEO

2 March 2018

Copyright ©2015, CREA IDEA LAB, Ltd. All rights reserved.

ABSTRACT

Nowadays, the data industry grows and proliferates steadily as many traditional industries digitalise their processes and create new products and services. Today, speaking about Big Data is a common topic. However, the industry still straggles the challenges to find meaningful business models. One of the basic characteristics of the data industry is the need of a data storage. The way, that need is satisfied, greatly affects the availability of a business model.

The great dependance between the viability of the business model and the data storage architecture comes from the cost of the data storage and the need for a cleaver optimization of the cost vs. performance ratio.

Today, the way how we recollect and use data and the type of data which we are storing in data bases influences directly over the infrastructure's needs and their cost.

In this article, we are going to discuss the key factors for building a sustainable data base. As a theoretical model of a sustainable data base, we will discuss the human brain. The analysis of the data organisational process of the human brain will be presented at a conceptual level without pretending to be supported by an experimental evidence.

INTRODUCTION

For the success of the data industry, there are a few challenges related to the treatment of the data. To our opinion, those challenges can be summarised in the following list:

- Exponential growth of the recollected data volume leading to an exponential growth of the needed infrastructure.
- Need for classification and organization of the data following a certain criteria which at the moment precise of human resources.
- Extraction of a valuable knowledge that can be monetised.

Those challenges directly impact the income vs. cost ratio and validate the existence of business models behind the data.

There is not any discussions about the first challenge in the above list as there is a commonly accepted opinion between the data base's developers that the infrastructure is cheap [From own market research]. However, nowadays, there is a large amount of evidence about failed businesses due to this issue [From own market research].

In this paper, we proposed a conceptual solution for this challenge through the so called "sustainable data base architecture". To visualise our proposition, we will use the human brain as a theoretical model. We will analyse how the human brain treats and organises data from a conceptual point of view without going to the details at celular level.

Copyright ©2015, CREA IDEA LAB, Ltd. All rights reserved.

SUSTAINABLE DATA BASE

First, we want to discuss what should be the data base to be called sustainable. Our definition is the following:

Sustainable data base should not grow or grow very little in time independently of the input data volume's growth.

This definition suggests that if we imagine a data base as a closed box, it should can store data without limit to the input data volume. To resolve this spacial-volume challenge, the only solutions is to overwrite the old data with the new one. Then, although, the input data can enter infinite in time, the recorded volume can be keep constant. In this case, we assume a constant data volume flow, but it becomes an issue if the input data volume grows overtime. Then, just overwriting won't be enough. In this second case, it is needed some way for compressing the data.

Then, a sustainable data base has two main mechanisms: overwriting the old data with the new one under a defined criteria and compressing the stored data.

THE HUMAN BRAIN AS A DATA BASE

Data processing of the human brain

Looking for an example of a sustainable data base that answers the previous definition, we found it to be the human brain. We believe the human brain is different from the animal one by its capacity to structure and to compress data.

To explain how the human brain process the data in the meaning of a data base, we use the following simple model:

First, we imagine a baby to be an empty data base with specific rules for data recollection, called instincts.

Once the baby is born, it starts to recollect data driven by those rules. Although, it is exposed to a huge amount of data, its sensor input's connections with the data base are not ready from the first moment and they develop over time. The feeding of the data base is gradual and all the basic input's connections to the data base become fully operative after the first year of the baby.

At this early stage, the brain basically is storing data and still does not need to compress it.

Second, with the first input of data, the brain starts to classify it. In our model, this classification is based on the principle of similarity. The brain starts to order the similar data in groups. At this stage, it starts for first time to organise the data, reserving different areas for different classes. Now, the data from each input connection is related to a specific area where it will be store.

At this level, the brain starts to recognise different stimulus because the data becomes classified.

Third, once the classification occurs, the brain start to interrelate different data classes. At this stage, the connections between different zones with classified data start to build. The objective of the brain at this

Copyright ©2015, CREA IDEA LAB, Ltd. All rights reserved.

moment is to resolve all possible combinations of relations between classes. The baby at this level can react simultaneously to different stimulus as the input data become interrelated.

Now, the brain as a data base becomes a structured and classified one. From this point further, the brain starts to analyse the storage data by classifying different areas of raw data with their connections by the type of input, and vice-versa, groups of inputs by a type of interrelation. Now, the baby starts to recognise the provider of the stimulus and it gets the understanding of "cause-effect" relationship.

At this stage, the brain as a data base evolutes to a classified base of information (classified data). The interrelation and analysis of the storage information are the principles of the learning process. The baby starts to learn objects, words, behaviour. In the light of the data base concept, the learning process is to feed the data base not with raw data but with information. At this stage, the brain as a data base has established first data analysis mechanisms on the data inputs and it stores a treated data.

To fully fill this new type of a data base- the informational one, the brain needs a certain time. At one point, there is a need to compress the data. It is in this moment when the brain starts to classified the data by importance.

This new classification creates a new type of a data base which has three main sections: raw data section, informational (operational) section and historical data (memory) section. Also, a compressing procedure is established to keep the brain as a sustainable data base. From this point further, the brain as a data base becomes independent to the input data volume and the growth of this volume in time.

We suggest for the compressing procedure to be in the following sequence:

1. Compressed raw data results in a specific information which is labeled again as data (classified data-feature, object, behaviour).

2. Compressing the classified data results in evolved information which we label as a conclusion (classified information).

3. Compressing the classified information results in higher evolved information which we called understanding (certain information or true).

Through this compressing procedure , it can be observed how the data efficiently transform to knowledge and at each level more volume of data is compressed and better knowledge is extracted. We proposed that the important mechanisms to compress the data are the data organization and the data classification which permit to assigned a new unit for the treated data volume. Then, through this process a volume is shrunk to an unit which resumes that volume.

In resume, an adult human brain is a data treatment machine with very efficient knowledge extraction process. The fundamental of this machine is the sustainable data base concept for the purpose of which the data is constantly classified and compressed to knowledge. In our model, the constant learning capacity of the human brain is represented as a data compressing process.

Also, we want to highlight that the key factor for the correct operation of the brain as a data treatment machine is the evolution of the data classification process by assigning a value of importance to the data. Thanks to this

Copyright ©2015, CREA IDEA LAB, Ltd. All rights reserved.

parameter the data can be directed either to the informational section or to the memory one. We suggest that the first data, which is stored in the memory section, is a collection of criteria built through a test-error procedure. We suggest that this set of criteria is used on a later stages in the first filter zone for a primarily data analysis.

Once the sustainable data base structure is settle, we propose that the data processing in the human brain is multidirectional. Then, a typical conventional data processing of the human brain can be presented as shown in the Figure 1.

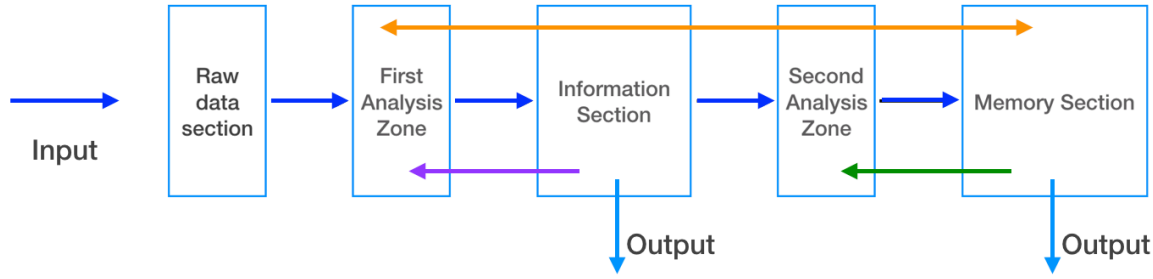

Figure 1. A schematic representation of the human brain as a sustainable data base. There are different paths for the data treatment which are pictured in different colours: blue, orange, purple and green.

Through the sensor system the data enters in the raw data section and it is treated through the first filter zone where a first compression occurs. Then, the compress data is store in the informational section. Here, we suggest that a few mechanisms can occur:

- The data is stored as a new information and it is analysed in the second analysis zone.
- The data overwrites similar old data.
- The data activates an output like a reaction.

From those mechanisms only the data classified as new and moved to be analysed in the second analysis zone can have a probability to be characterised as a very important and store in the memory section.

This possible treatment process is shown in blue color in the Figure 1.

By the orange and purple paths in Figure 1, we propose that the primarily treatment can be a verification process in which the new raw data is analysed against a set of criteria from the memory section or old data from the informational section.

Here, one can assume different scenarios. For example, the data is recognised and classified as known and it overwrites in the informational section. Other possibility is the data is not recognised, so, it is classified as new knowledge and it is stored in the memory section. Third possible scenario is the data seems to be known but cannot properly be classified, so, it is classified as a problem and it is store in the informational section for further treatment.

Copyright ©2015, CREA IDEA LAB, Ltd. All rights reserved.

We proposed that the data analysis in the human brain is based on crossing data from different sections in each analysis zone (the orange, purple and green paths). This makes the data treatment process multidirectional as the data flows trough different sections crossing the new input data with the stored one.

We suggest that the outputs from the data base comes from the informational section as reactions and from the memory section as knowledge and memories.

Sustainable data base architecture based on the human brain model

To create a sustainable data base which will not need to grow its infrastructure in time, we need to put some rules over its operation: the input, different sections, the output, etc. Replicating the basic conclusions over the data treatment process of the human brain, the sustainable data base has five components, shown in Figure 1:

- A raw data section, where the inputs are store as they are.
- First analysis zone, where the raw data is curated.
- An informational (operative) section, where the data can generate outputs.
- Second analysis zone, where the data is classified by their importance.
- A memory section, where only important data is stored.

These five components need to be well defined with rules for their operation. Today, the most common data bases are with the simple function of recording the inputs as they are. To evolve the concept of a common data base to a sustainable one, we need first to establish the first analysis zone as a set of rules which activates with the entrance of an input. These rules can be the following:

- Verify if a similar data has been recorded previously, if “Yes” then “ No recording”.
- If “No” on the previous statement then check for a class for the input data by a comparison to a set of criteria.
- If there is a class that fits the input data then label it and store it in the informational section.
- If there is not any class to fit the input data then label it as “for further study” and store it in the informational section.

Then with this, a common data base becomes an informational one as the store data is treated to a given level.

Another common fashion in the data industry is to record everything as a historical records for further applications like statistics and predictions. This opinion can threat our effort to create a sustainable data base as the infrastructure’s needs will increase in time. At this point, we need to establish a compressing mechanism to the informational data base and to split it by the creation of the memory section. The informational data base is split in two sections by the second analysis zone which again is a set of rules. These rules can be the following:

Copyright ©2015, CREA IDEA LAB, Ltd. All rights reserved.

- All data which define a class for first time is recorded in the memory section under a data set of classes.
- The data which operate over an output is overwritten in the informational section and the output is produced.
- The data which does not activate an output is verified if a similar data exists already in the memory zone. If "No", it is recorded.

At this point, we still do not compress the data as the human brain does. If we should imagine a similar compressing process, we need to establish a learning mechanism as a rule in our second analysis zone. This learning mechanism should has the capacity to recognise, in a volume of data, common features and label all the data volume with a single unit (word, symbol, etc.). For example, the analysed data volume contains features from localisation, position, durability, sound, colour. Then, the learning mechanism can establish that the studied data set represents a tea cup. Now, we can record only the data "tea cup" in place of all the data set.

If we can establish the proposed compressing mechanism or similar one, we can use it constantly or periodically in dependance with our needs of infrastructure. Then, our data base becomes sustainable.

We want to highlight that the learning mechanism as a compressing one does not mean that our data base is an intelligence itself. In our opinion, the intelligence can create outputs which are not establish previously by rules in the first and second analysis zones. The proposed sustainable data base can become an intelligence if we add a learning mechanism which creates outputs as a result. The discussion of this transformation is a topic of another paper and we will not continue it here.

CONCLUSIONS

In the presented paper, we have defined a sustainable data base as one with a constant infrastructure which is independent of the input data volume and its growth.

Then, we have analysed the human brain as a best theoretical model of a sustainable data base. We have suggested that the brain consists of five main sections: raw data section, first analysis zone, informational section, second analysis zone and memory section. In our model, we have define the learning process of the human brain as a data compressing one which converts the human brain in a very effective knowledge extraction machine where knowledge is understand as a compress data.

We have proposed that the data treatment is multidirectional process where the input data flows through different sections and is crossed with the stored one.

Lastly, we have proposed a sustainable data base architecture which follows the human brain model. We have suggested a few examples of rules for the first and second analysis zones.

Copyright ©2015, CREA IDEA LAB, Ltd. All rights reserved.