eng.understanding_data_storage

Alexandr Kirilov (https://github.com/alexandrkirilov)

Understanding data storage.

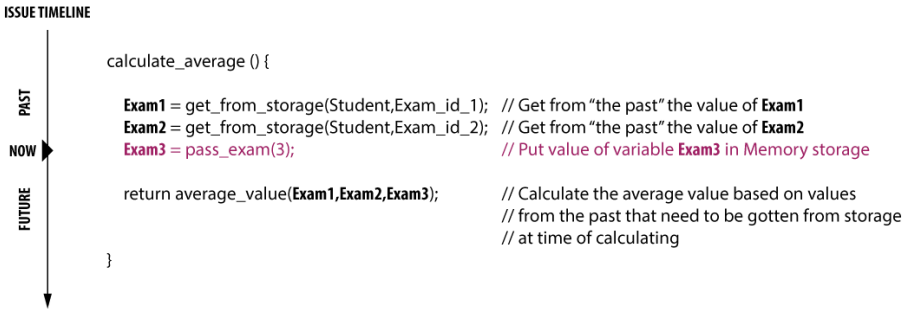
In order for projecting and developing data storage need to look for the causes of "Why we need to store data?". The cause itself will require of how this data important, how long this data valid, how it need to be used, how rapid this data should be extracted from storage and etc.

When describing data storage developing process using the terms "past", "present" and "future". It has no any hidden meaning at all. It's only simplest way to describe the actions sequence, like 1-2-3. From the data storage point of view:

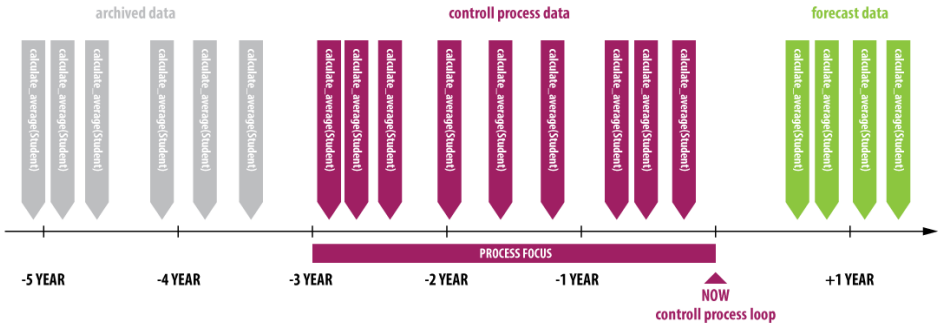
- 1 - past: data stored
- 2 - present (now): the data is under the process of actions or they are in stored state for future usage
- 3 - future: the data will be under the process of actions

In all of any cases, the reason of why you are storing data ONLY ONE - something unresolved or unfinished in THE REAL WORLD ("in the real world" should be pointed especially) and will require to transfer data from the "past" through the "present" (now) by the actions for the "future" result. Look on examples:

- Getting value of graduation on any exam: the unfinished or unresolved issue in the real world is getting average value for diploma or certification where the value of current exam will be part of calculation and therefore you should somehow to store it for future action. Where "the time of storing" is restricted by the time from examination itself to the time of calculation average value and this time is "the data focus".



- The system of Education Quality Control: unresolved and unfinished issue - the looped process of control education quality by analysing results of examination. The time line for this process is "while the process alive itself", the data focus is, for example - 3 years in the past, mean from current time towards minus 3 years in the past is time of storing for this process.



- The accounting system for goods in supermarket: has at least 2 process that requiring one data - paying taxes and strategical planing for supermarket future actions. All of this process using data from the past and by calculating dow defining actions in future - to pay defined sum for taxes and amount of goods that need to be restored in.
- The Facility Security system: controlling access to the facility and has looped process of controlling and storing data for future analysis of possible breaches
- etc

The unresolved or unfinished issues might be limited by time (in IT we are naming it like functions or procedures) or be unlimited (in IT we are naming it like service, daemon or process). The processes itself might be contained the subprocess, subprocess might be contained another subprocesses. The article "Understanding Blockchain" describing the principle of structure organising.

The time of data-alive-limit is always biggest value from any process or subprocess that related to this data. This is the reason why is so important to understand the structure of it in real world. Based on it you will project the storage size, data storage cleaning procedures - as result of it the price for building and maintaining data storage. The actors that involved into the process will require you to develop access to the data and how rapid this data should be extracted from storage out.

```
Author notice.
```

```
The Big Data problem not a problem itself. It's the result of
the problem of Big List of Unresolved or Unfinished issues in
the real world. The understanding Big Data Storage started from
the point where you are considering carefully about how you
organised process and why this process is stubbing you by this
amount of data.
May be the problem of solving Big Data storages should be
started from "Why do you have so much data?" and optimise it
first in real world?
```

Based on experience of developing data storages you always have to add kind of reserved size for ability to handle avalanche-like data size growing. Usually it 30%-40% for one-unit storage and for distributed storage it's 3%-25%. The reserve size is depend on company ability to adopt the size of storage for the level of increased data size, if you need more time for reaction - add more reserve to storage.

The process of developing of the data storage structure usually contain 90% of issues do not related to the IT at all. In most cases it looks like cooperation between two persons: one IT specialist another the highly experienced person from the field where this storage going to be used. There are a lot of issues "in general" that is always using in any field, but one small special issue from one special field might to make almost impossible to use "in general" data storage solution.