TECHNISCHE UNIVERSITÄT DRESDEN

DRESDEN concept

Center for Information Services and High Performance Computing (ZIH)

Continuous Information Processing Enabling Real-Time Capabilities: An Energy Efficient Big Data Approach

ICSEA 2021, Barcelona (Spain)

October 03 to October 07, 2021

Martin Zinner (martin.zinner1@tu-dresden.de)

Big Data

1. Big Data

- *Big Data* refers to enormous volumes of *data* that cannot be processed effectively with the traditional applications that exist.
  Graphic: https://unsplash.com/s/photos/big-data

- Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data, http://www.gartner.com/newsroom/id/1731916

- An article by techjury states that Data is growing faster than ever before and in the year 2020, about 1.7 megabytes of new information were created every second for every human being on the planet, https://techjury.net/blog/how-much-data-is-created-every-day/#gref

Small Data

2. Small Data

- Small Data, is nothing but the data that are small enough comprehensive for human in a volume and also for formatting, that makes it accessible, informative and actionable.
  Graphic: https://unsplash.com
- When Data volume grows beyond a certain limit traditional systems and methodologies are not enough to process data or transform data into a useful format. https://www.educba.com/small-data-vs-big-data/
- Most case in range of tens of hundreds of GB. Some case few TBs.
- Can be processed using on hand data management tools or traditional data processing applications https://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/#1f3ba65613ae

Small Data - cont.

3. How Small Data becomes bigger than Big Data

- In some cases processing small data becomes ineffective or impossible by using traditional applications.
- The impression is that we are dealing with inflated Small Data developing to Big Data.
- Some problems are difficult to solve even on small datasets.

Continuous manufacturing and assembly trends in the industry

Manufacturing

- Trend towards minimal stock (zero inventory) .
- The stock is on the highway (On-schedule delivery).
- Just in time delivery.

https://www.unleashedsoftware.com/blog/zero-inventory-primer-manufacturers

Civil engineering

- Assemble-to-order for prefabricated parts to be delivered.

https://www.investopedia.com/terms/a/assemble-to-order.asp

Current state-of-the-art in Business Intelligence

An example from the manufacturing industry having general validity

- Production runs 24x7x365, but reports regarding the previous day are posted at 8:00 a.m.

- Five Pain Points identified by Cisco regarding batch jobs in the Business Intelligence area are:

➢ *the race against time*; managing batch window time constraints,
➢ *cascading errors and painful recovery*; eliminating errors caused by improper job sequencing,
➢ *ad hoc reporting*; managing unplanned reports in a plan based environment,
➢ *service-level consistency*; managing service-level agreements,
➢ *resources*; ETL resource conflict management.

Real-time requirements

Motivation

- A real case at a semiconductor company regarding real-time requirements:

➢ Reports should reflect the *current state of the production*.
➢ *Preliminary measurement data*, hence reduced ramp-up time for new products.
➢ *Reduced computational effort*, hence smaller computers can be used.

Aim

- Give satisfactory answers to all Pains Points identified by Cisco except Pains Point 3).

Our strategy

Short description of our approach

- within the information flow, the process of information aggregation is started as early as possible, best when the data is still in memory.
- The aggregation phase takes place in parallel to the data collection phase.
- Intermediary aggregated results corresponding to the collection status should be available.
- Final aggregated values are available soon after the collection phase.

Challenge

- Redesign the aggregation algorithm.
- Some build-in function may not be available.

Exemplifying the strategy of continuous aggregation by using the standard deviation of a sample

— Formula (1) cannot be used with continuous aggregation strategy, it requires all of the data to be collected in order to calculate $\bar{x} := 1/N \sum_{i=1}^{N} x_i$

$$SD_N := \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}. \qquad (1)$$

— Representation (2) is suitable, at each step the values $S_n = n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2$ are calculated, $\{x_1, x_2, \ldots, x_N\}$ are the observed valued of the sample items.

$$SD_N = \frac{1}{N}\sqrt{\left|N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2\right|}. \qquad (2)$$

The advantages of continuous computing

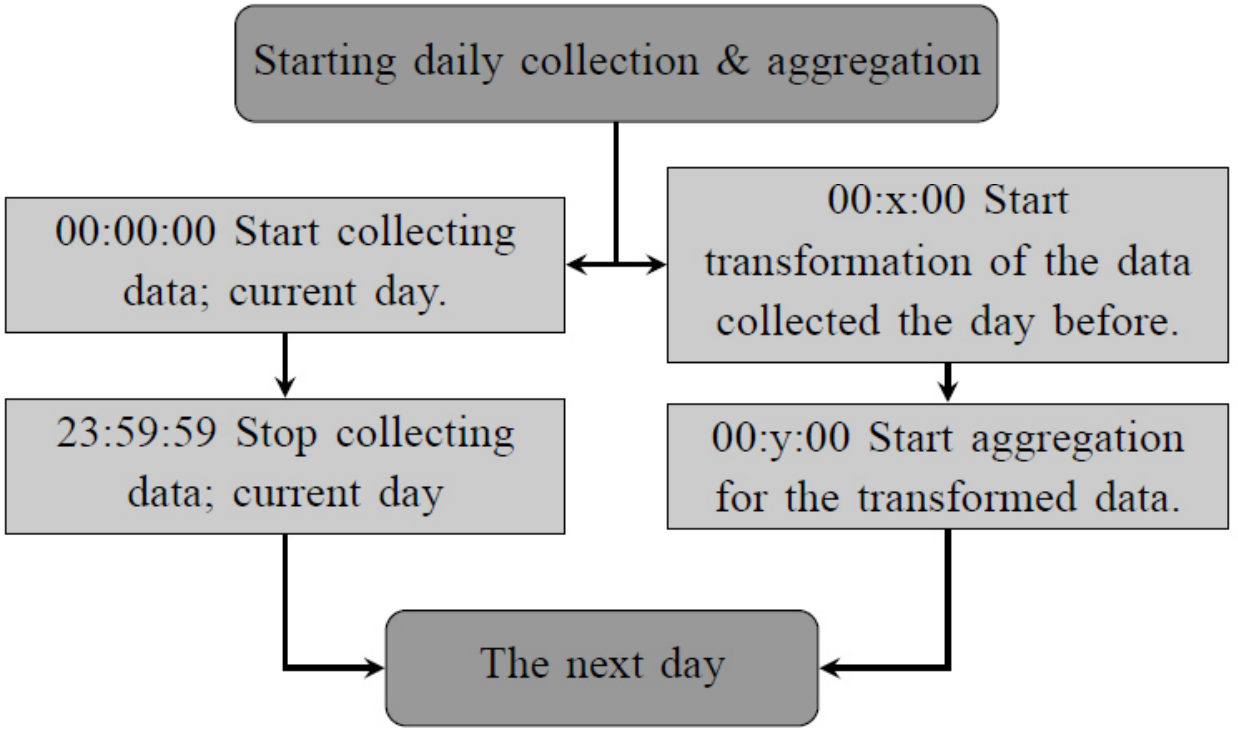
A particular selection

- *Real-time capabilities*; if time constraints can be met.
- *Aggregated values corresponding to the captured data*; i.e., reporting capabilities at any point in time during data collection.
- *Straightforward design strategies* due to clear, easy understandable architectural and implementation principles.
- *Uniform load of the underlying database system* due to the continuous aggregation principles.
- *Efficient recalculation of aggregated values*; in case erroneous data is collected.
- *Energy efficiency*; smaller computers can be used due to the fact that aggregation is performed during the whole data collection period.

Classical batch jobs aggregation strategy

Short description

- Data collection for a complete day, transformation & aggregation and can be started only after the whole data involved is fully retrieved.

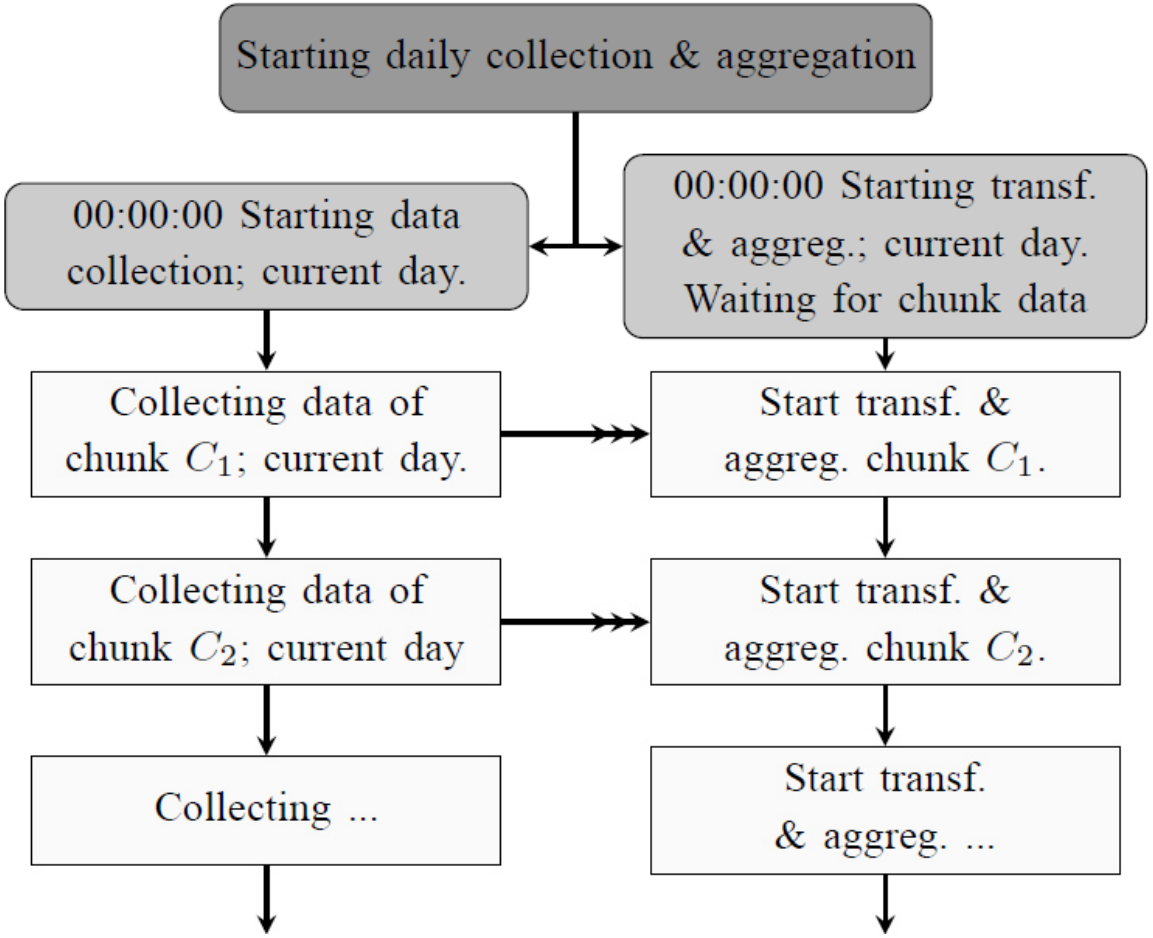


Continuous information processing strategy

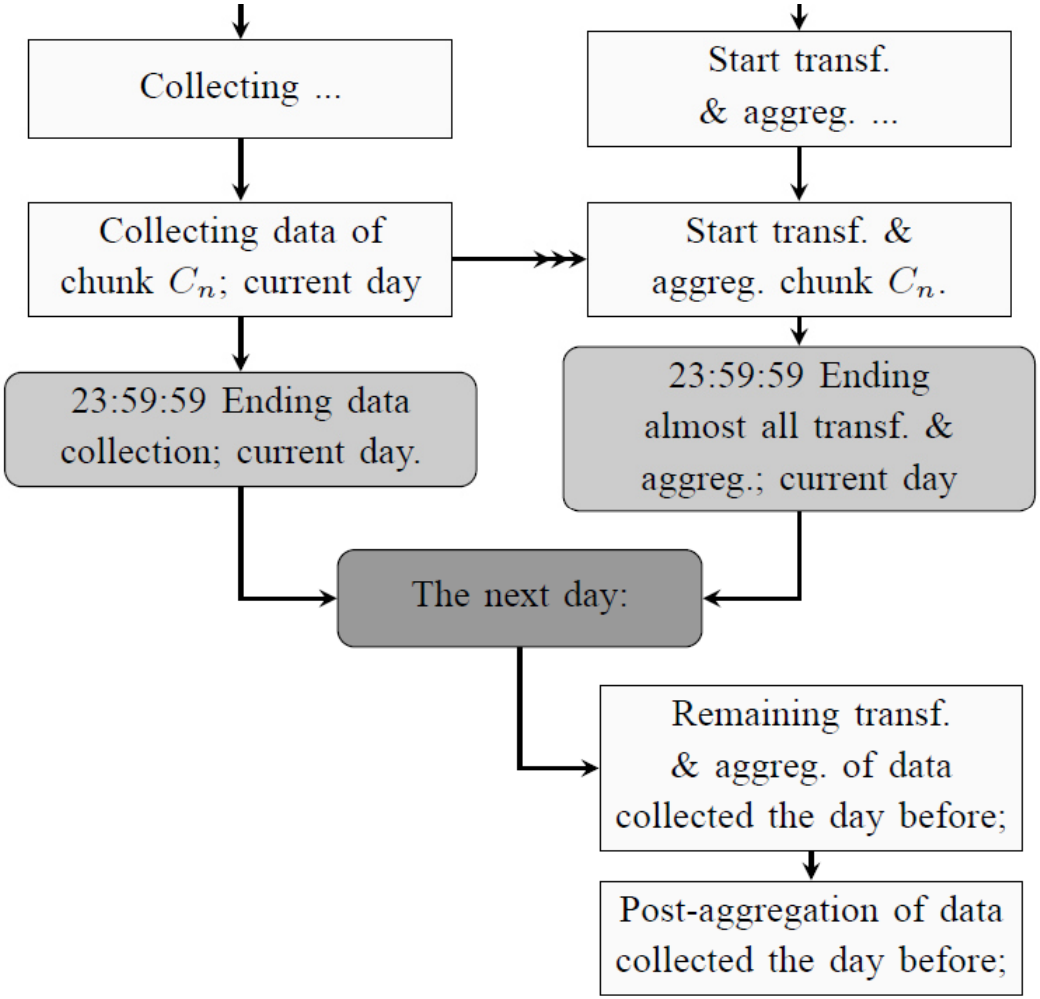
Short description

- Similarly, the data of one day is retrieved/transformed/aggregated.
- Carried out on small chunks of data, as long as the data is still in memory.
- While the data for a particular chunk is retrieved, the transformation and aggregation on the previous chunk is performed.
- After midnight, the remaining chunks of the previous day are retrieved/transformed/aggregated.
- Post-aggregation is performed, such that final values for evaluation are ready soon after midnight.

Continuous information processing strategy – cont. 01



Continuous information processing strategy – cont. 02



Lessons learnt

Continuous aggregation algorithm

- Advantageous if used from scratch.
- Needs a radical rethinking process of developers of algorithms and programmers.
- The overall design and implementation effort may be higher than using the classical bath jobs strategy.

The execution time and the quality of the result

- Minimal execution time due to optimal size of the chunks.
- Superior data quality due to transparent design strategies.

Transition from batch jobs legacy application to continuous aggregation

- Cumbersome and time consuming.
- Involves the redesign of the entire aggregation strategy and architecture.

Conclusion and Future Work

Main results

- Enables real-time capabilities of the system.
- Facilitates the paradigm shift from a subjective software construction activity, towards objectively verifiable straightforward strategies.

Outlook

- Pain Point No. 3 of Cisco's white paper "ad hoc reporting; managing unplanned reports in a plan-based environment" remains still unhandled and it is subject of future research.
- What is the optimal strategy regarding volatile versus persistent aggregation?

Thank you

Thank you for your attention

Questions?