KNOWDIVE

KGE - Knowledge Graph Engineering

Integration of Streams

How to handle Streams

Alessio Zamboni, Fausto Giunchiglia

What is a Stream

What is a Stream?

Definition

Streams are the continuous surge of events that are happening in time and space.

Streams are complex, continuous objects. To handle them in a digital world those objects undergo a series of processes.

- **sampling**: we need to make sample of the events;
- **approximation**: we need to measure the event stream evolution through sensors, introducing errors:
  - *discretization* of values (e.g. due to the precision of the sensor, the resolution of the DAC, etc.)
  - *semantic approximation*: contexts themselves are multi-dimensional streams. If context is not fully aligned with the data there might be a misinterpretation.
- **windowing**: the stream is *unbounded*, but our memory is finite. We need to consider a *finite part* of the stream at each moment.

Stream datasets

We have an approximation of the real-world stream.

- Context does change / there are multiple contexts.
- The dataset is continuously updated.
- The purpose might change.

While very general, is **not possible** to create a lossless representation in knowledge graph.

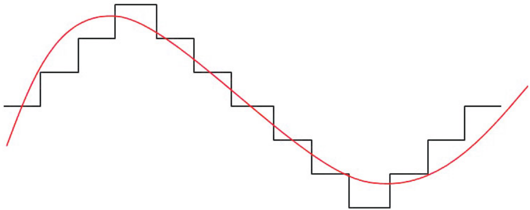


Figure: Real-word stream vs Stream dataset

Classic datasets

What we have in classic datasets is a "snapshot" in time and space of a stream.

- Context doesn’t change.
- The dataset doesn’t change.
- The purpose is fixed.

In this scenario, is **possible** create an exhaustive knowledge graph for that specific purpose.

Common issues

1. **Entity Resolution**: recognize entities and relationships (harder between datasets).
2. **Datatype alignment**: give homogeneous format and measure unit (e.g. date conversion).
3. **Conceptualization**: evaluate information in context, extracting knowledge (e.g. WSD to extract concept from text).

Stream issues

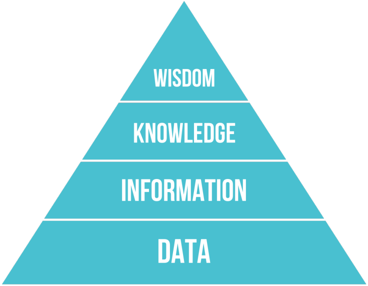
- Single Context, Single Actor
  1. **Sampling rate alignment**[1]
- Single Context, Multiple Actors
  1. **Sampling rate alignment**
  2. **Entity Resolution** (new value or new actor)
- Multiple Context, Single Actors
  1. **Sampling rate alignment**
  2. **Context Drift**[2][3]
- Multiple Context, Multiple Actors
  1. **Sampling rate alignment**
  2. **Entity Resolution** (new value or new actor)
  3. **Context Drift**

Note!

Since context and data are changing, also techniques need to adapt!

Other stream issues

- **Velocity**: generally the requirement is real-time (*online*) processing.
- **Volume**: generally is not possible to store all the values of the stream.
- Dealing more with low level **data** rather than **information**.



Project environment

- Your project lives in a controlled environment.
- Some of the mentioned problem are already solved for you.
- Data is not changing, but we will **simulate** it.
- Biggest problem will be **updates** and (maybe) **alignment**.

Sampling rate alignment

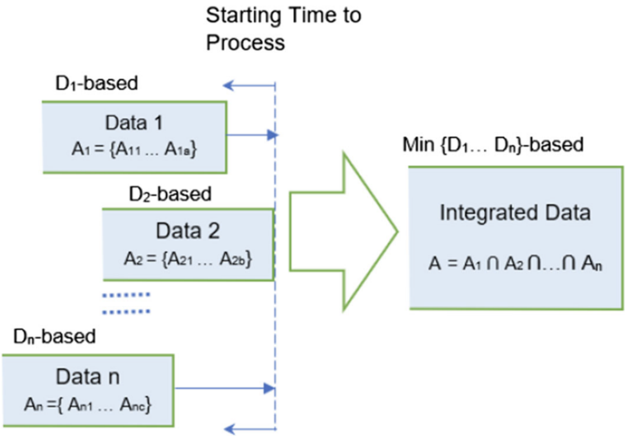


Fig. 1 A scenario of the IoT streaming data integration from multiple sources

References

- Tu, D.Q., Kayes, A.S.M., Rahayu, W. et al. IoT streaming data integration from multiple sources. Computing 102, 2299–2329 (2020). https://doi.org/10.1007/s00607-020-00830-9
- Cobb, Oliver, and Van Looveren, Arnaud. "Context-Aware Drift Detection." arXiv, 2022, https://doi.org/10.48550/arXiv.2203.08644.
- Bontempelli, A., Giunchiglia, F., Passerini, A. et al. Human-in-the-loop handling of knowledge drift. Data Min Knowl Disc 36, 1865–1884 (2022). https://doi.org/10.1007/s10618-022-00845-0

KGE - Knowledge Graph Engineering

Alessio Zamboni, Fausto Giunchiglia

Integration of Streams

How to handle Streams