Eötvös Loránd University (ELTE)
Faculty of Informatics (IK)
Pázmány Péter sétány 1/c
1117 Budapest, Hungary

Introduction to the Stream mining (SM) course

Stream mining (SM)

Imre Lendák, PhD, Associate Professor
Péter Kiss, PhD candidate

2020
Budapest, Hungary

Outline

- About the course
- Draft list of lecture topics & references
- Streaming intro

Stream mining

ABOUT THE COURSE

About the course

Theory and labs

- Lectures according to schedule
  - Initially on Teams only
  - PDF slides posted on Canvas
- Occasional assignments during the semester for extra points
- Course team:
  - Imre Lendák, Associate Professor
  - Péter Kiss, PhD candidate

Exam

- 60% for the projects
  - 5-member teams
  - Joint project for the OST and SM courses
- 40% for the theory
  - Oral examination
  - Defended project is entry criteria
- Note: both exam elements are obligatory

Stream mining

LECTURE LIST (DRAFT)

Lectures list (draft from 2019)

Part I: Stream mining intro

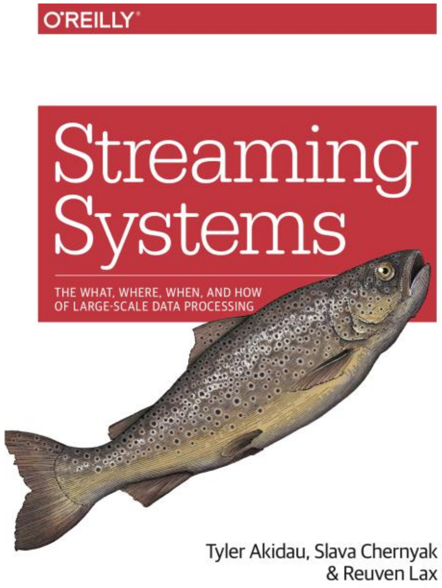
- Introduction, i.e. these slides
- Basics of stream mining
- Basics of stream mining
- Concept drift

Part II: Stream analysis

- Clustering
- Frequent pattern mining
- Novelty detection
- Time series
- Distributed stream mining

References

- Akidau T., Chernyak S., Lax R. Streaming Systems: The What, Where, When, and how of Large-scale Data Processing, O'Reilly Media, 2018.
  - http://streamingsystems.net
  - Streaming 101 & 102: The world beyond batch
- Gama, J. Knowledge discovery from data streams. Chapman and Hall/CRC, 2010.
- Aggarwal, C. C., ed. Data streams: models and algorithms. Vol. 31. Springer Science & Business Media, 2007.



Stream mining

INTRO TO STREAMING

Why streaming?

- Businesses need timely (i.e. immediate) insights into their data
- Massive, unbounded datasets are increasingly common in different business domains
- Processing data as it arrives spreads workloads more evenly over time → consistent and predictable consumption of computing resources (e.g. if we rent cloud-based resources)
  - This is the opposite of hoarding large amounts of data and periodically running high CPU/memory use analyses

Kinds of streams

- Click streams
- Sensor measurements
- Satellite imaging data
- Power grid electricity distribution
- Banking/e-commerce transactions
- Security monitoring data

Necessary definitions

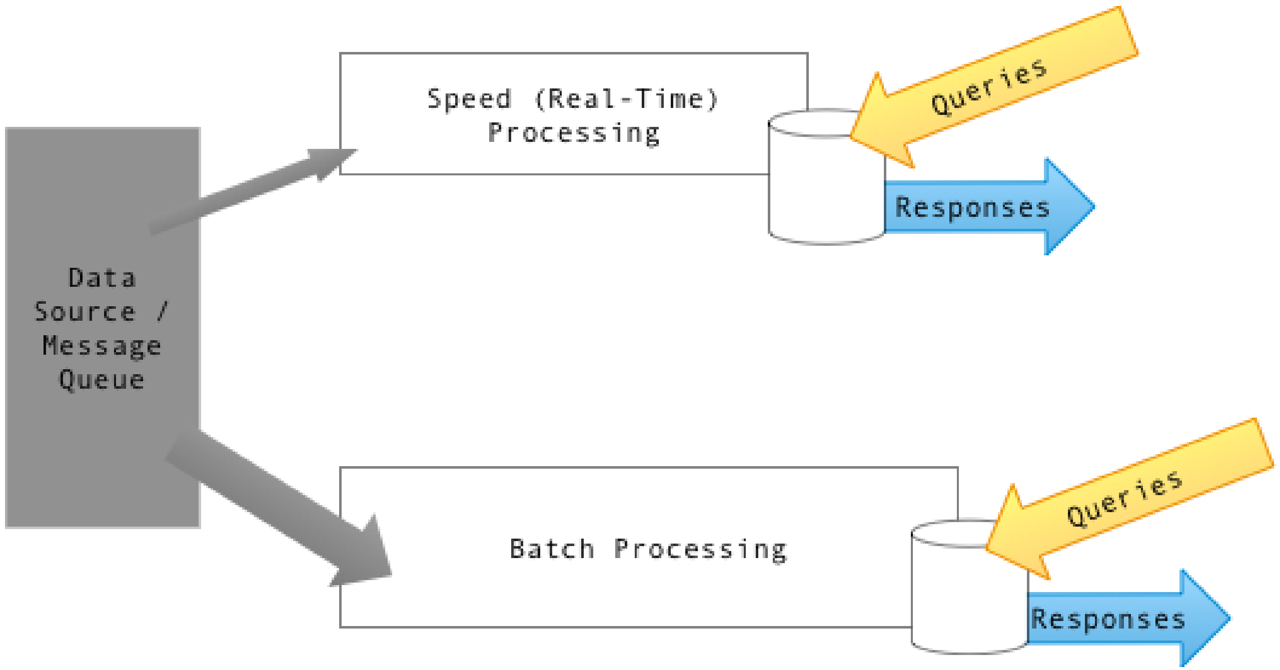
- Dataset cardinality
  - DEF: **Bounded data** is finite in size.
  - DEF: **Unbounded data** is infinite in size.
- Data constitution
  - DEF: A **table** represents a holistic (~complete) view of a dataset at a specific point in time.
  - DEF: A **stream** is an element-by-element view of the evolution of a dataset over time.
    - Alternate definition: A data stream is an ordered (not necessarily always) and potentially infinite sequence of data points (e.g. numbers, words, sequences).
- DEF: A **streaming system** is a data processing engine designed for handling infinite (unbounded) datasets.

Traditional streaming

- Characteristics of traditional streaming systems:
  - The good: low latency
  - The bad: inaccurate, i.e. lack of consistency → non-deterministic

- **DEF:** Batch systems are deterministic as they provide eventually correct results, i.e. once all relevant data is acquired and analyzed

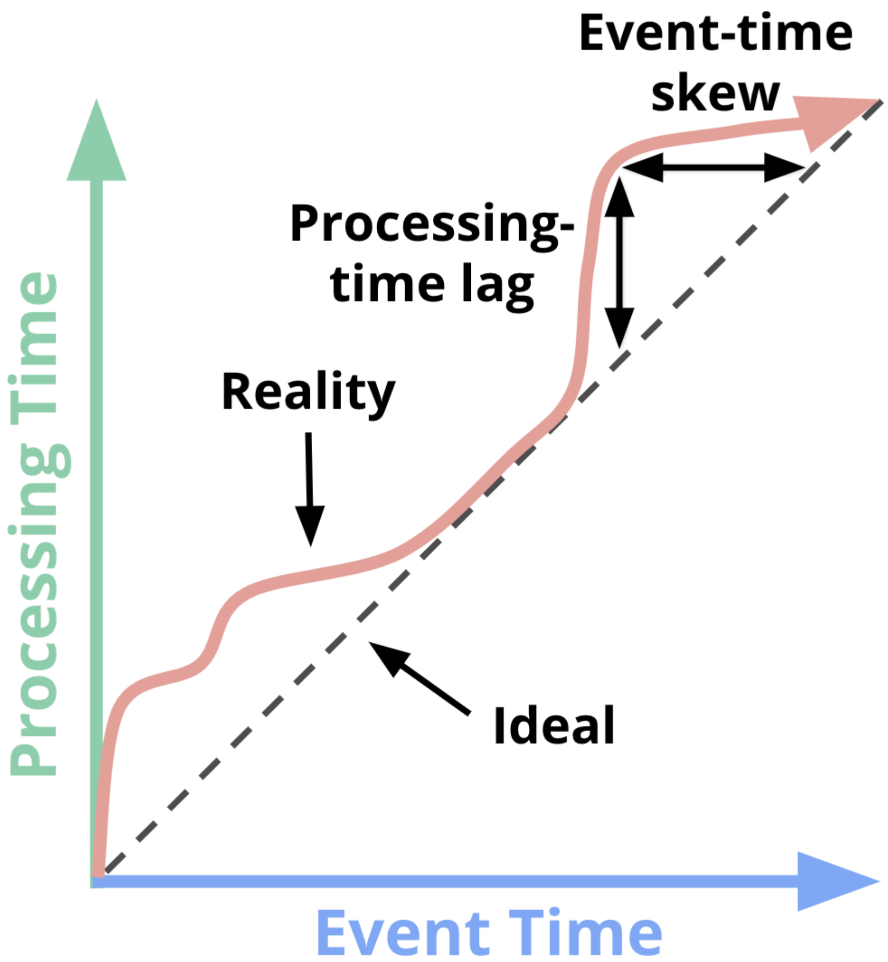
Lambda architecture



https://en.wikipedia.org/wiki/Lambda_architecture

- Lambda architecture: running a traditional streaming system in parallel with a batch data analysis solution
  - The good: low-latency (though inaccurate) results from the streaming element, correct results from the batch subsystem
  - The bad: hassle to implement and maintain (2 systems!)

Event time vs processing time



http://streamingsystems.net/fig/1-1

- X axis → event-time completeness in the system → the time X in event time up to which all data with event times less than X have been observed.
- Y axis → the progress of processing time → normal clock time as observed by the data processing system as it executes.

‘Modern’ streaming

1. Correctness
   - Consistent storage
   - Exactly-once processing
2. Reasoning about time
   - Techniques for reasoning about time in the presence of unbounded, unordered data of varying event time skew

Up next…

- Streaming basics
- Concept drift
- Stream analysis

Thank you for your attention!