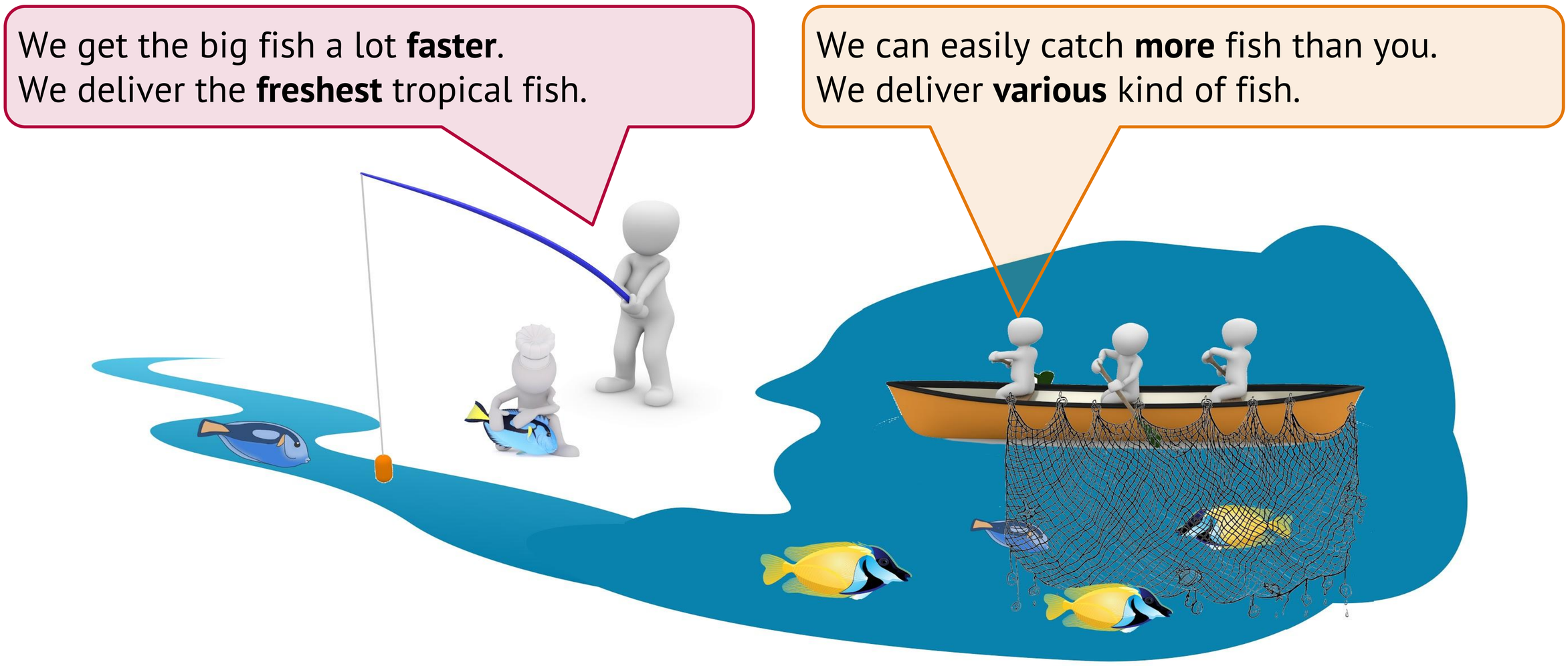
The Rise of Streaming
Is Streaming the end of Batch Processing?

Stream Processing vs. Batch Processing

In the Fishing Industry



In the World of Data Engineering

Data changes fast

Query changes slowly

Data changes slowly

Query changes fast

use cases

In a fast changing world, where new data is continuously generated, streaming looks like the right approach to support real-time data analytics. Streaming can be used for event-driven and machine learning model based decisions making, when a fast response time is crucial. Example scenarios are fraud detection, news feeds analysis[1] or visualising ongoing events in a hospital[2].

technical limitations

But streaming brings some disadvantages: Since a stream is infinite over time, streaming can just emit a result, while the traditional batch processing has access to all data and returns a complete result.[3] To compensate that, streaming applications need to save a state for some specific scenarios.

conclusion

Traditional Data Warehouse Systems are designed for batch processing. Nowadays, many companies build a streaming layer and batch processing layer in their architecture[1]. One of the rising streaming platforms is Apache Flink.[3] Apache Flink supports both: input from historic data stored in database tables and data streams. With the capabilities of Apache Flink the stream processing technology will spread to much more areas than just real-time analytics. But its aim is not to replace batch processing. Streaming enables new options to process data. Apache Flink extends existing data workflows. The question is not: Should I build a streaming or batch processing architecture? It should be: Do I leverage the streaming or the batch processing capabilities for my particular application?
To decide whether to use streaming of batch processing for a specific application ask *What changes faster? Data or Query?* And enjoy the best of both worlds.

Pascal Crenzin (Masterstudent)
Ringvorlesung Data Engineering in der Praxis
Hasso Plattner Institute, Potsdam, Germany

E-Mail: pascal.crenzin@student.hpi.de

[1] as presented in *Eating News from the Web: Our implementation of the lambda architecture for scalable text analytics* by Peter Adolphs, Neofonie
[2] as presented in *Queue Mining - Analysis of Clinical Pathways based on Sensed Data* by Matthias Weidlich, Humboldt Universität zu Berlin
[3] as presented in *Modern stream processing and real-time event-driven applications with Apache Flink* by Fabian Hueske, data Artisans

HPI Hasso Plattner Institut