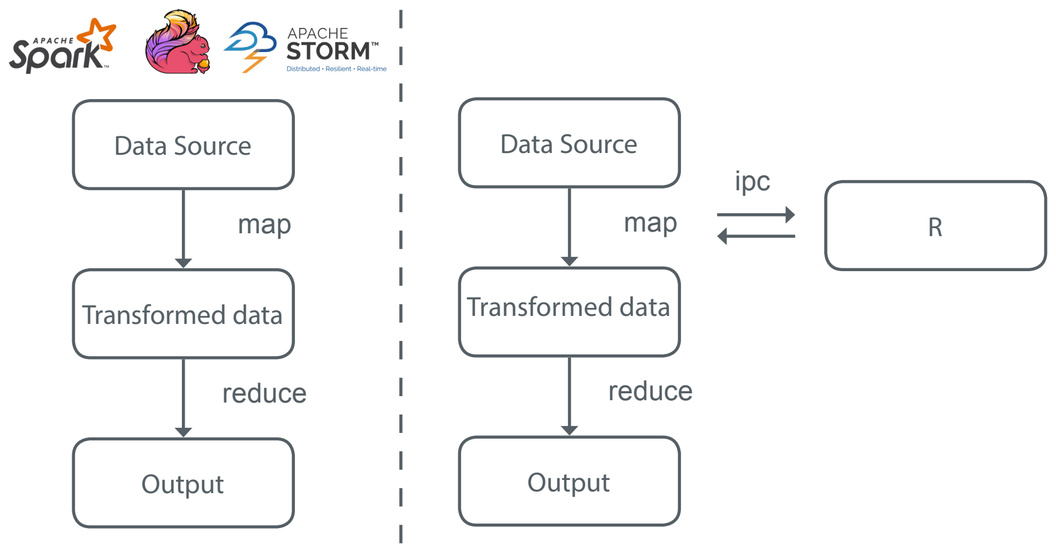
Borderless Data Streams

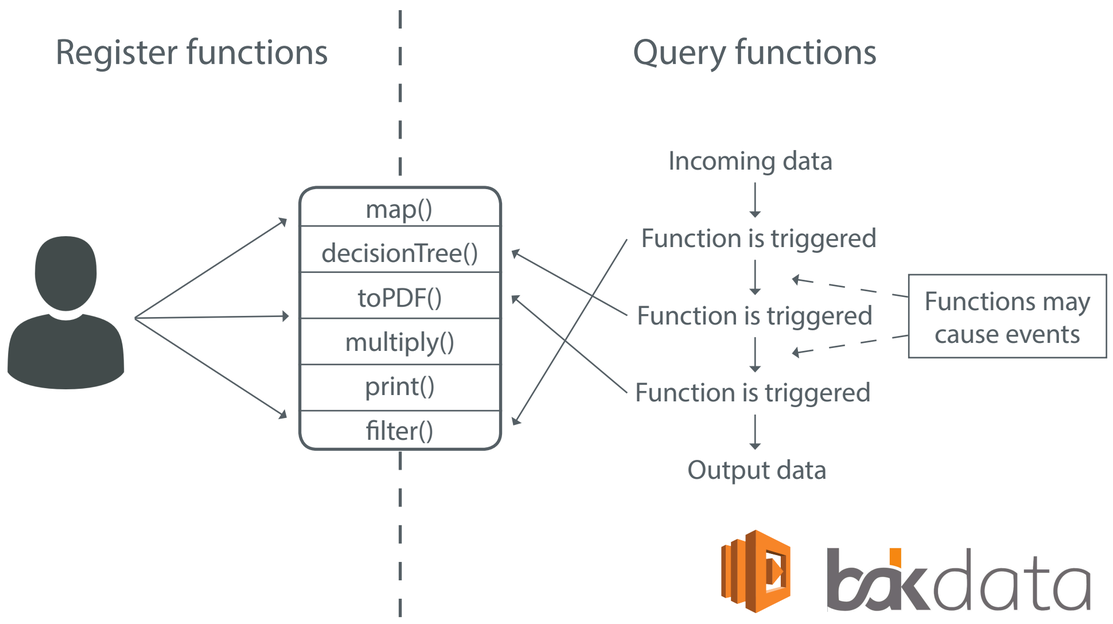
A vision of building reactive data flow graphs across fields of expertise

In a traditional company people in various roles and with diverging expertise work closely together. For example, a statistician may use R or other statistic tools to solve a problem, whereas a software engineer may use another language to query and transform data, which is then used by the statistician. Today's streaming applications, such as Apache Flink or Apache Spark support the latter, but are usually limited to a small set of languages.

Defining data flow graphs

With applications like Apache Flink or Apache Spark software engineers are able to design and implement data flows as directed, acyclic graphs. These graphs consist of concatenations of functions that specify operations on this data. However, these operations are usually written in a language that must be supported by the framework, thus mostly JVM languages or Python. Statistic frameworks like R are generally not supported.



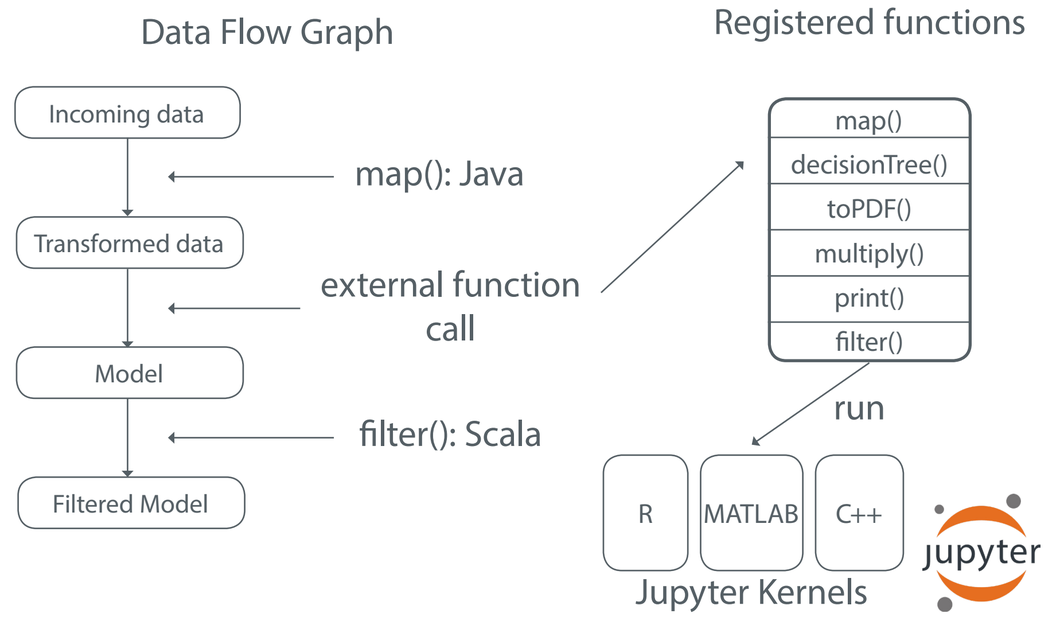


Data flow graphs across system borders

While Amazon Lambda allows chaining functions via events, it lacks support for statistic frameworks. A more general approach as proposed by Bakdata is to use, e.g., Jupyter Kernels that exist for various languages. However, compared to the existing frameworks both miss features, such as scheduling and distribution of tasks. The next generation of data processing engines should guarantee both, extensibility for further languages and powerful optimizing.

Serverless computing in the cloud

Another trend that arose in the recent years is the trend of outsourcing computation. Platforms like Amazon Lambda or Bakdata's Data Science API allow users to register functions that compute results based on a given input. These stateless functions can be triggered by events or REST calls returning values that can be used in subsequent function calls. These platforms usually scale with the number of function calls.



Sven Lehmann

sven.lehmann@student.hpi.de

Master IT-System Engineering

Data Engineering in der Praxis WS 17/18

Icon made by Freepik

from www.flaticon.com

HPI Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam