Stream Processing in a Nutshell

Fabian Paul – HPI Master Student

Stream Processing systems compute over data as it enters the system. Instead of defining operations to apply to an entire dataset, stream processors define operations that will be applied to each individual data item as it passes through the system. Hence the results are immediately available and will be continually updated as new data arrives.
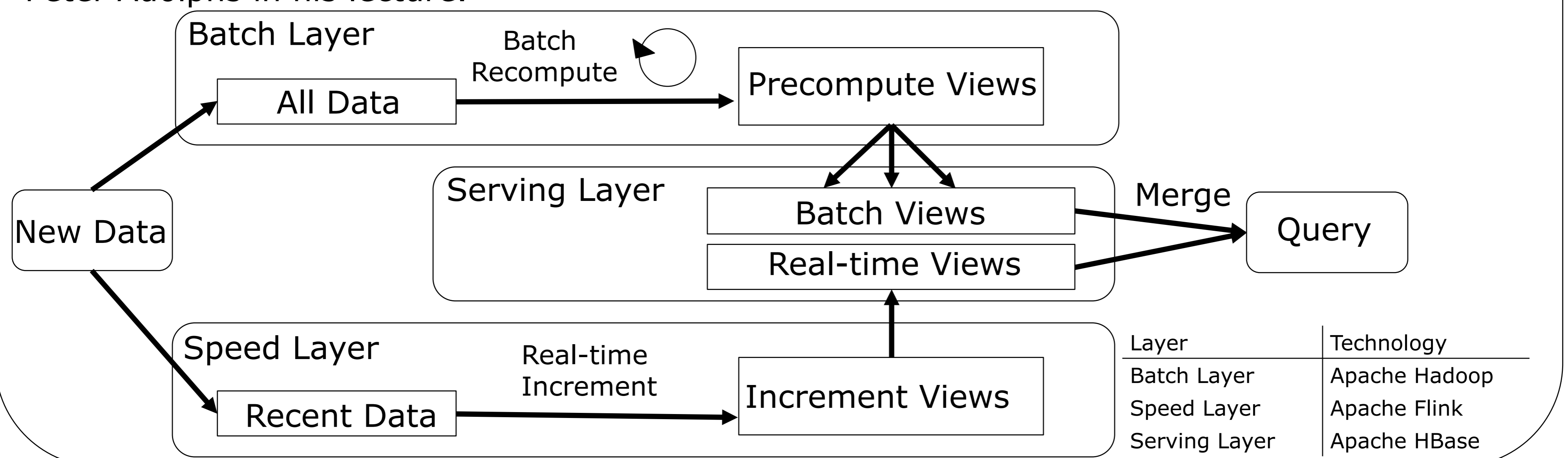
Event Data today is generated by almost every electronic device we are using in our daily life. Even though most of the time it happens without the user's notice. At the end of 2016 around

400 million IoT devices with cellular connections were emitting events on a daily basis. Another type of an event source is message logging: As mentioned by Fabian Hüske in his Flink talk Netflix's streaming system is processing **800 billion log messages** per day generated by their microservices.

Underlying Architectures

In the past 10 years two different theoretical architectures were proposed. Both are describing how a stream processing system could look like and what kind of technologies are used.

Lambda Architecture

This architecture was developed by Nathan Marz the creator of Apache Storm in 2011. Nowadays it is still used in several production systems e.g. the news feed system shown by Peter Adolphs in his lecture.

Layer	Technology
Batch Layer	Apache Hadoop
Speed Layer	Apache Flink
Serving Layer	Apache HBase

Kappa Architecture

In 2014 Jay Kreps argued that having multiple distributed systems, both layers inside the Lambda Architecture, would add too much operational complexity. So he proposed the Kappa Architecture.

Layer	Technology
Replay Layer	Apache Kafka
Processing Layer	Apache Flink
Serving Layer	Any database

Fabian Paul

Data Engineering in der Praxis WS 17/18
Hasso Plattner Institute, Potsdam, Germany

E-Mail: fabian.paul@student.hpi.de

HPI Hasso Plattner Institut