Summary

This thesis addresses the modelling of streaming applications in the context of near real-time Big Data processing. The topic is motivated by todays' growing demand for in-time data analysis in various applications. The thesis is divided in two parts: Starting with a review of the *Lambda* architecture and the streaming components, we deal with the modelling and execution of streaming applications in part one. The second part focuses on real-world examples, demonstrating the use of our proposed streams framework in two real-world Big Data applications.

In the first part, we survey the current landscape of stream processing frameworks, focusing on the category of *general purpose streaming platforms*. These frameworks build the basis of a lot of modern Big Data streaming, providing a code-level interface to design domain specific applications. Following that, we introduce the streams framework, which builds a middle layer modelling approach for defining streaming applications in a platform independent way. This abstraction features the implementation of streaming functions that can easily be integrated into several runtime environments, providing a multi-platform and multi-paradim use of domain specific components.

The streams framework uses declarative XML specifications of an application's data flow graph, liberating users from a code-level formulation of their application. Based on this declarative approach, we investigate the modelling of applications by an interactive sketch-based user interface. For the sketch-based editing, we investigate the use of different machine learning approaches to detect and classify user gestures and map these to editor actions. We demonstrate the use of gestures in two prototype applications for the RapidMiner tool and an Android application to design streams applications.

In the second part, we demonstrate the use of streams in two Big Data applications, namely the online processing and analysis of data in Cherenkov astronomy and the near real-time extraction of viewership statistics in the context of an IP-TV platform.

For the astrophysical use case we demonstrate the multi-platform embedding of streams by running data preprocessing pipelines in a streaming fashion as well as mapping the same functionality to a massive parallel execution using the Apache Hadoop execution engine. By the demonstrated abstraction, we enable physicists to gain the processing power of modern Big Data platforms, while focusing on their application domain using the declarative modelling layer of the streams framework.

In the use case of the IP-TV platform, we build a streaming architecture for the EU project ViSTA-TV based on streams. We demonstrate the handling of heterogeneous data sources, as well as the integration of external data with high-speed lookups, as is often required in today's Big Data environments.

We conclude the thesis with a summary of the contributions made and the impact, which streams has to various projects, as well as an outlook to ongoing related work.