University of Zurich[UZH]

**Department of Informatics**

University of Zürich
Department of Informatics
Binzmühlestr. 14
CH-8050 Zürich
Phone. +41 44 635 43 11
Fax +41 44 635 68 09
www.ifi.uzh.ch/dbtg

UZH, Dept. of Informatics, Binzmühlestr. 14, CH-8050 Zürich

**Prof. Dr. Michael Böhlen**
Professor
Phone +41 44 635 43 33
Fax +41 44 635 68 09
boehlen@ifi.uzh.ch

Zürich, 14. Dezember 2017

**MSc Basismodul: Study Stream Processing Platforms and develop a real time Data Analytics application**

Stream Processing is an ideal platform where data streams are produced continuously and requires real time processing. Over the past years such platforms have been developed which ensures distributed, high-performance, scalable, stateful and real time processing. Example of such platforms are Apache Flink [1], Apache Spark [2] etc.

**Tasks**

1. Literature study on Apache Flink [4] and Apache Spark [5], [6] Architecture.
2. Literature study on the Time Handling in Stream Environments: Event Times, Windows, Watermarks [7].
3. Get familiar with NYC Parking Tickets Dataset [3].
4. Implement streaming application using Apache Flink [1] and Apache Spark [2] and build the following streaming pipeline for the dataset [4].

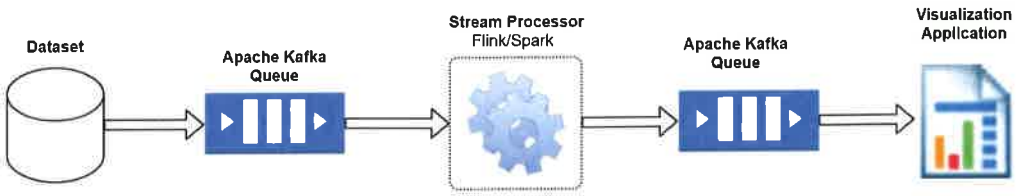


figure 1: Streaming Pipeline

5. Compare Flink and Spark by producing Latency and Throughput graphs for the ticket

count operation.

6. For visualization show real time graphs as the time windows moves on.
   (a) Tickets count per day, per month, per year.
   (b) Where are tickets most commonly issued?
   (c) What are the most common types of cars to be ticketed?

   Add a few more graphs of your own choice after analyzing data.
7. Summarize your findings in a report of atleast 10 pages.

**References**

1. www.flink.apache.org
2. www.spark.apache.org
3. www.kaggle.com/new-york-city/nyc-parking-tickets
4. Carbone, P., Katsifodimos, A., Ewen, S., Markl, V., Haridi, S. et al. (2015): Apache flink: Stream and batch processing in a single engine
5. https://dl.acm.org/citation.cfm?id=2934664
6. http://datastrophic.io/core-concepts-architecture-and-internals-of-apache-spark/
7. Introduction to Apache Flink: Stream Processing for Real Time and Beyond by Ellen Friedman and Kostas Tzoumas(Chapter 4 - Handling Time)

**Supervisor:** Muhammad Saad (saad@ifi.uzh.ch)
**Start Date:** Dec 15, 2017
**End Date/Oral Exam:** Feb 6, 2018 , 15:00 - 15:30 in room 2.E.13

University of Zurich
Department of Informatics

Prof. Dr. Michael Böhlen
Professor