From batch to stream processing of massive data.

Dario Colazzo
dario.colazzo@dauphine.fr
Laboratoire d'Analyse et Modélisation de Systèmes pour l'Aide à la Décision (LAMSADE) – Université Paris-Dauphine, Centre National de la Recherche Scientifique : UMR7024 – Place de Lattre de Tassigny
75775 PARIS CEDEX 16, France

Recent years have seen the rapid diffusion of paradigms and systems for processing large datasets, relying on shared-nothing distributed programming paradigms. While first applications manipulating massive datasets were based on batch processing, there is currently a growing interest in processing large-scale data collections in a streaming fashion, still relaying on distribution and parallelism.

This talk will present the basic principles and systems behind large scale data processing, with a first focus on batch processing and a second focus on streaming. The attention will be on paradigms behind mainstream technologies, namely MapReduce Hadoop, Flink and Spark.

Dario Colazzo. Dario Colazzo is Full Professor at Université Paris Dauphine. His research interests focus on databases and programming languages, and include cloud databases and type systems for safe and efficient processing of semi-structured data. Past research interests included type systems for the lambda-calculus and pi-calculus.